TRAITEMENT BIOINFORMATIQUE DE DONNÉES RNA-Seq

Plan

- Mercredi après midi (14h-17h)
  - Rappels biologiques
  - Mode d'étude du transcriptome
  - Pipeline d'analyse RNA-seq
  - Vérification de la qualité
- Jeudi matin (9h- 12h)
  - Algorithmes d'alignement
  - Visualisation
  - Quantification des gènes
- Jeudi après midi (14h – 17h)
  - Quantification des transcrits
  - Découverte de nouveaux transcrits

_01

Rappels biologiques

Rappels biologiques

Qu’est-ce qu’un gène ?

Rappels biologiques

Qu'est-ce qu'un gène ?

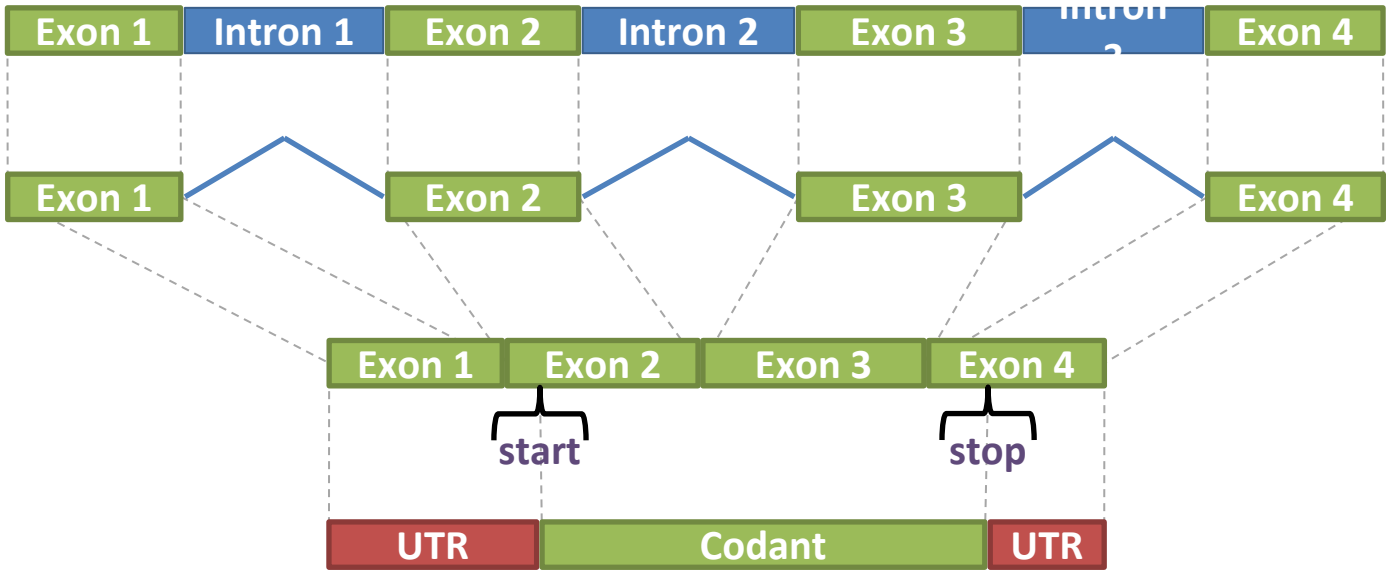
- **Gène :** unité fonctionnelle de l'ADN qui contient les instructions nécessaires à la création d'un produit fonctionnel

- **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 ?

- **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’epissage ?

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**

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

Rappels biologiques

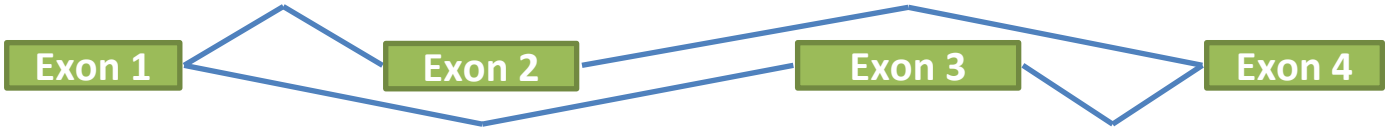
Epissage alternatif et isoformes

- Excision d'exon

- Rétention d'intron

- TSS alternatif

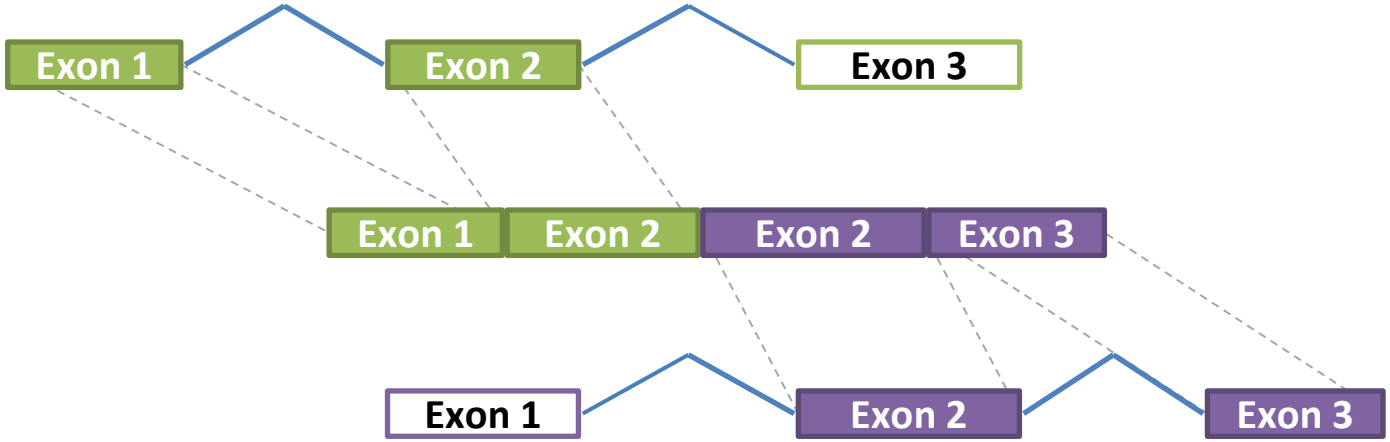
- Exons exclusifs



Rappels biologiques

Et plus encore ?

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

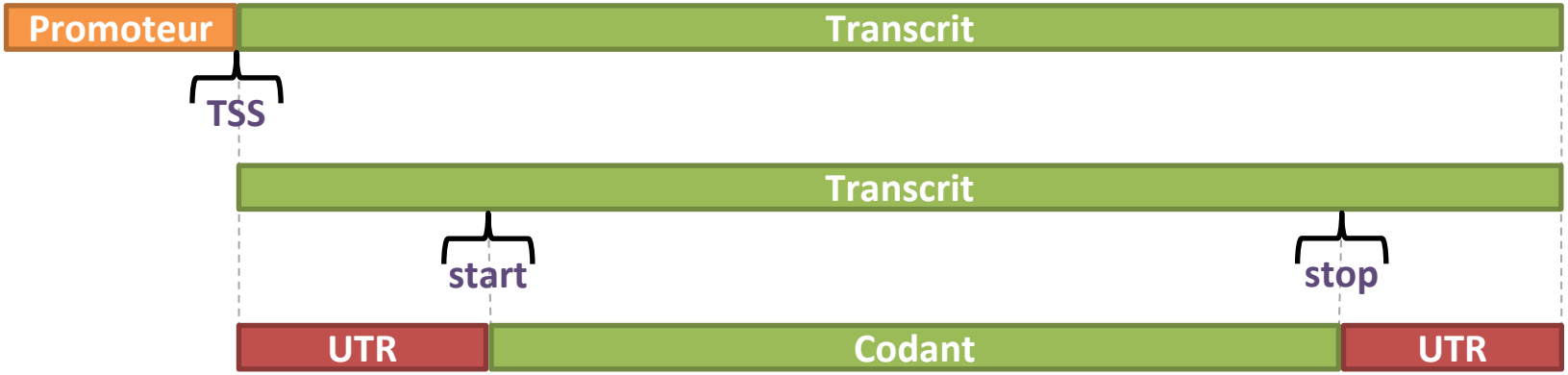


- **Chimère biologique**

Rappels biologiques

Gène procaryote / gène eucaryote

- Pas d'intron chez les procaryotes



Modes d’étude du transcriptome

Quels sont les modes d’étude du transcriptome ?

Modes d'étude du transcriptome

EST, rt-PCRq, puce d'expression…

- **EST :** séquençage bas débit de transcrits
  - (+) "longues" séquences, découverte d'épissage
  - (-) méthode **historique** (Sanger), **non quantitative**

- **rt-PCRq : quantification PCR** d'**ADNc**
  - (+) très **quantitatif**
  - (-) nécessite des **armorces spécifiques par gène**

- **Puce d'expression :** hybridation de **gènes connus**
  - (+) **quantitatif**, expression différentielle
  - (-) connaissance au minimum des **séquences des gènes**

Modes d'étude du transcriptome

… tiling array et RNA-Seq

- **<u>Tiling array</u> : hybridation le long de l'ensemble du génome**
  - (+) **quantitatif**, expression différentielle, nouveaux transcrits
  - (-) connaissance de l'ensemble du génome, petit génome (bactérie)

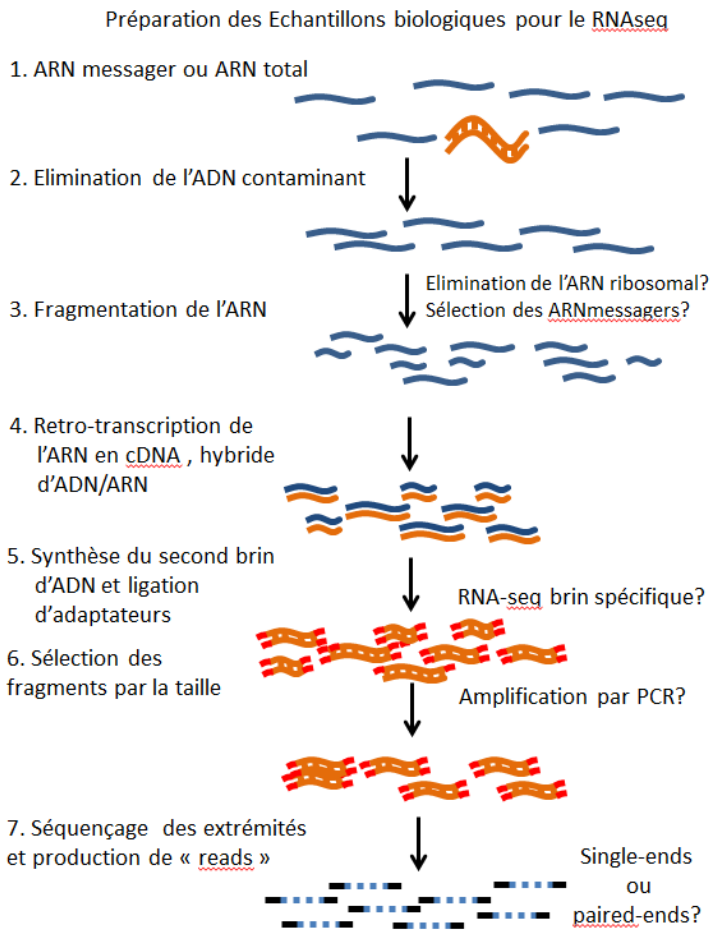
- **<u>RNA-Seq</u> : séquençage de l'ensemble des transcrits**
  - (+) séquence du génome pas forcément nécessaire
  - (-) encore en développement

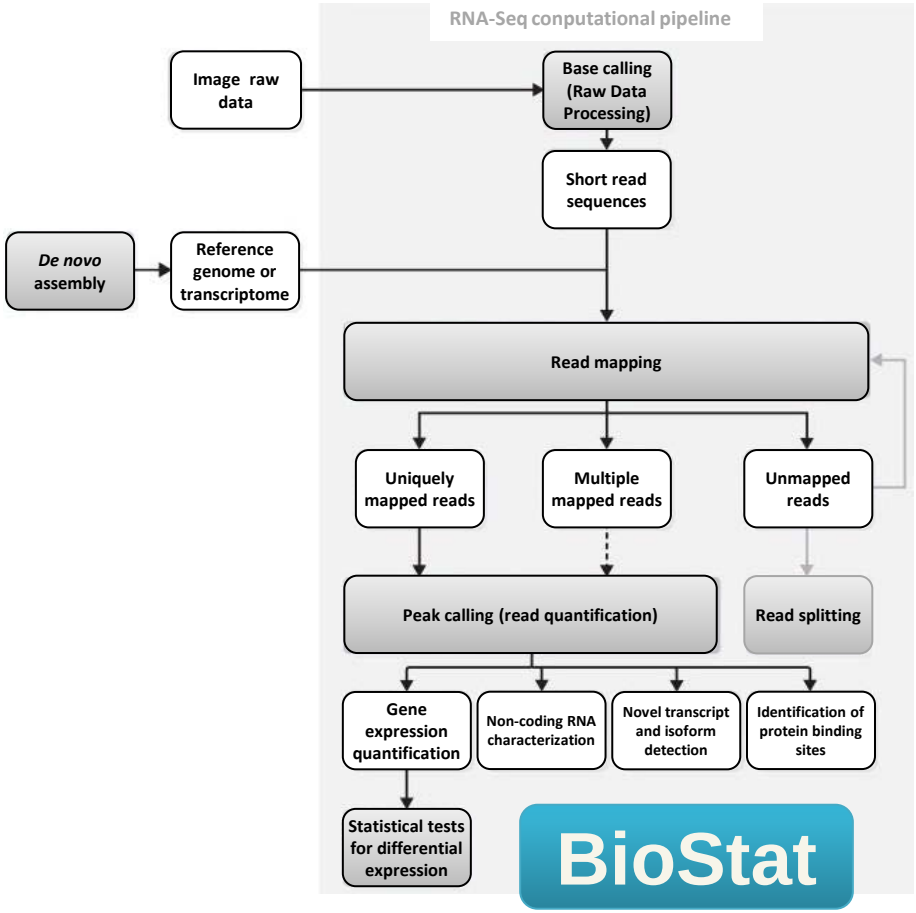
Une expérience de RNA-seq de A à Z

BioStat

Bio

BioInfo





Mutz 2013, Current Opinion in Biotechnology

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)

A quelles questions biologiques NE PEUT PAS répondre le RNA-seq ?

- Étude spécifique de :
  - **quelques gènes**
  - **gènes connus**
  - **gènes** présents sur une **puce commerciale** ou **dédiée à un organisme**

- **Coupler** en **UNE SEULE expérience de RNA-seq** :
  - expression des transcrits d'un génome
  - étude des gènes faiblement exprimés
  - découverte des petits ARN et leurs niveaux d'expression

Quels choix quand on fait du RNA-Seq ?

Les technologies de séquençage

Plateforme	454 Roche Titanium	HiSeq2500 Illumina	Ion PGM Life Technologies
Caracteristiques	- Titanium chemistry - Pyroséquençage - Amplification PCR	- Séquençage par synthèse basé sur la polymérase - Amplification PCR - Multiplexage	- Séquençage basé sur la ligation des bases - Amplification PCR
Applications	- Séquençage *de novo* - Petits genomes - Transcriptome	- Resequençage - Transcriptome - Epigenomique - Petits ARN - Séquençage allèle spécifique	- Séquençage *de novo* - Resequençage - Transcriptome - Epigenomique - Petits ARN
Mb / run	450 Mb	1 Tb	2.2 Gb
Taille des lectures	600 bp	125 bp	400 bp
Biais connus	- Long homopolymères saturent le signal - Duplication de lectures	- Régions riches en AT ou GC sous-représentées durant l'amplification - Plus d'erreurs en fin de cycle	- Duplication de lectures

Quels choix quand on fait du RNA-Seq ?

- **Déplétion / enrichissement** :
  - déplétion des ARNr (eucaryote ou procaryote)
  - sélection des transcrits poly-A (eucaryote)

- Séquençage en tenant compte du **sens du brin** :
  - utile pour l'étude des expressions anti-sens

- **Multiplexage** :
  - ajout de **séquences tags** pour regrouper **plusieurs échantillons** à séquencer sur une **même piste** de séquençage

Quels choix quand on fait du RNA-Seq ?

- Directives du consortium ENCODE en 2011
- RNAseq n'est pas mature!
- Equilibre **profondeur /** nombre de **répétitions** :
  - **plus de deux répétitions biologique**
  - **Corrélation de Pearson de 0.92 à 0.98 entre 2 echantillons**
  - **Si corrélation < 0.9 , cela doit être répétée ou refait.**
- Entre 30M et 100M de lecture par échantillons selon l'étude.

❖**NB. Guidelines for the information to publish with the data.**

Quels choix quand on fait du RNA-Seq ?

- **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 *vs* 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

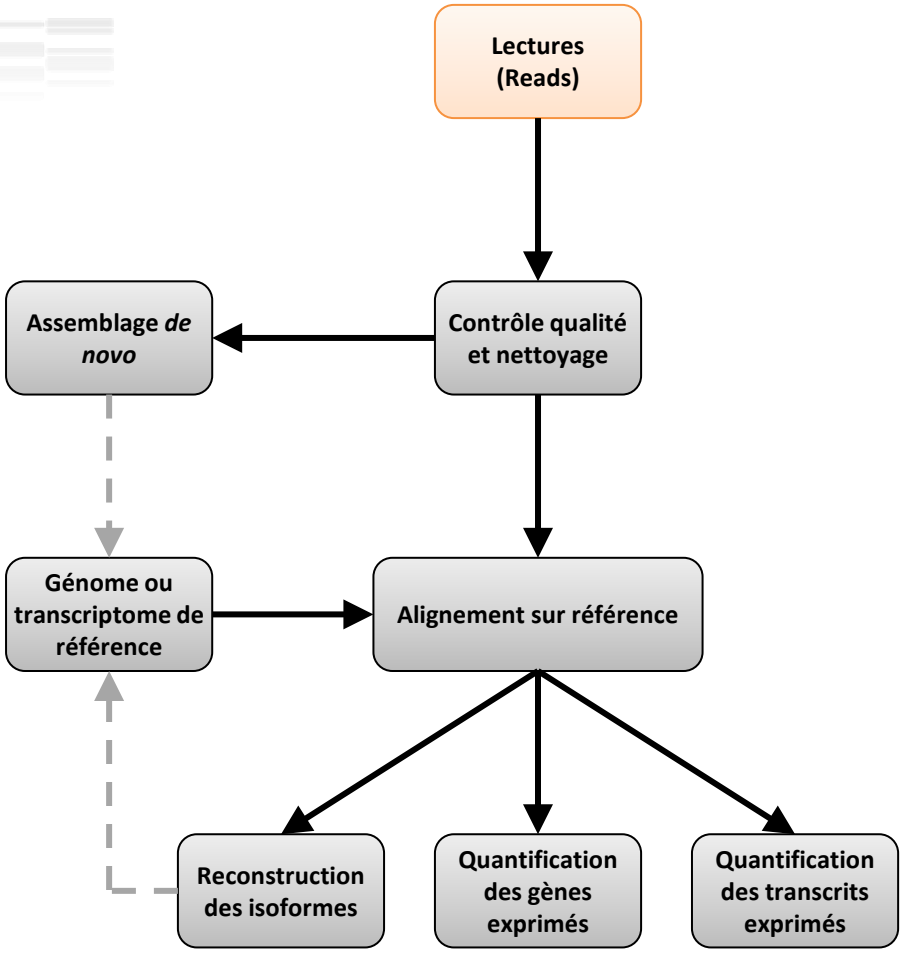
Stratégie d'analyse en fonction des données disponibles

- **De novo :**
  - Pas de génome/transcriptome de référence
  - Outils en évolution permanente
  - Ressources (cpu/disque) +++
- **Transcriptome de 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**
- **Nettoyage** des lectures
  - (**suppression** des **adaptateurs** de **multiplexage**)
  - **suppression** des **adaptateurs** de **séquençage**
  - **tronquer** les **extrémités** de **mauvaise qualité** des lectures
- **Alignement** des **lectures sur** la **référence**
  - gènes ou génome complet
- **Découverte de nouveaux transcrits**
- **Comptage** des gènes / transcrits

Workflow d'analyse RNA-Seq



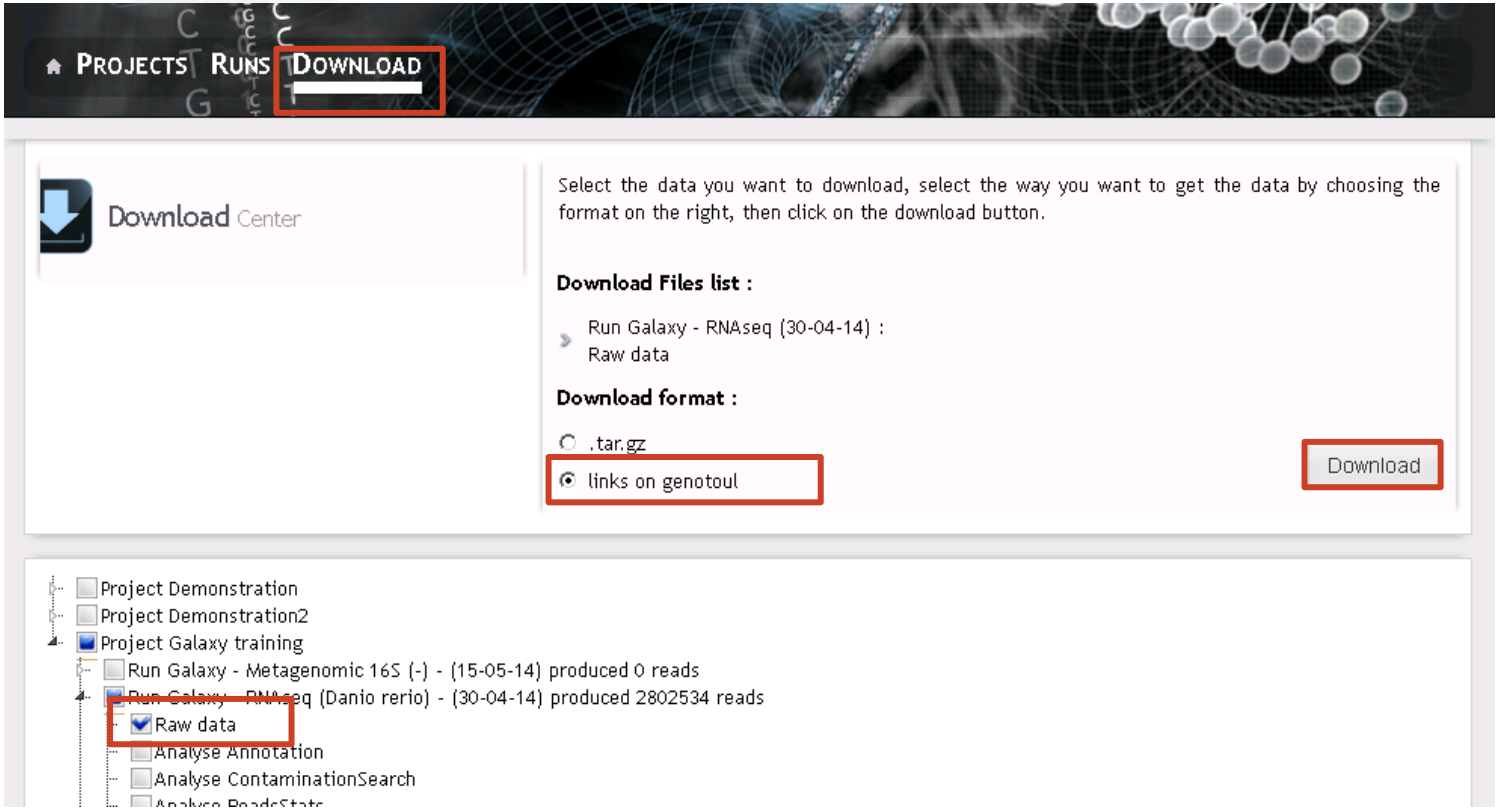
Travaux pratiques

Présentation des objectifs

- **Aborder** les **différentes étapes** indispensables au **traitement bioinformatique** de **données RNA-Seq** à travers un **exemple** issu de **données réelles :**
  - expérience chez **Danio Rerio (chr22)**
  - données correspondantes aux runs **ERR022486** et **ERR022488**
  - explorer sur l'ENA ces jeux de données (http://www.ebi.ac.uk/ena/)
- Pour le TP :
  - données réduites dans ng6.toulouse.inra.fr

Travaux pratiques

Depuis ng6 télécharger les données sur genotoul



Depuis galaxy « Uploader » les fichiers de genotoul

02

DONNÉES BRUTES

Obtenir des séquences de qualités

INRA SCIENCE & IMPACT

Biais spécifiques au RNA-Seq

Biais spécifiques:

- Influence du mode de préparation de la banque
  - amplification hexamèrique aléatoire (**Random hexamer priming**)

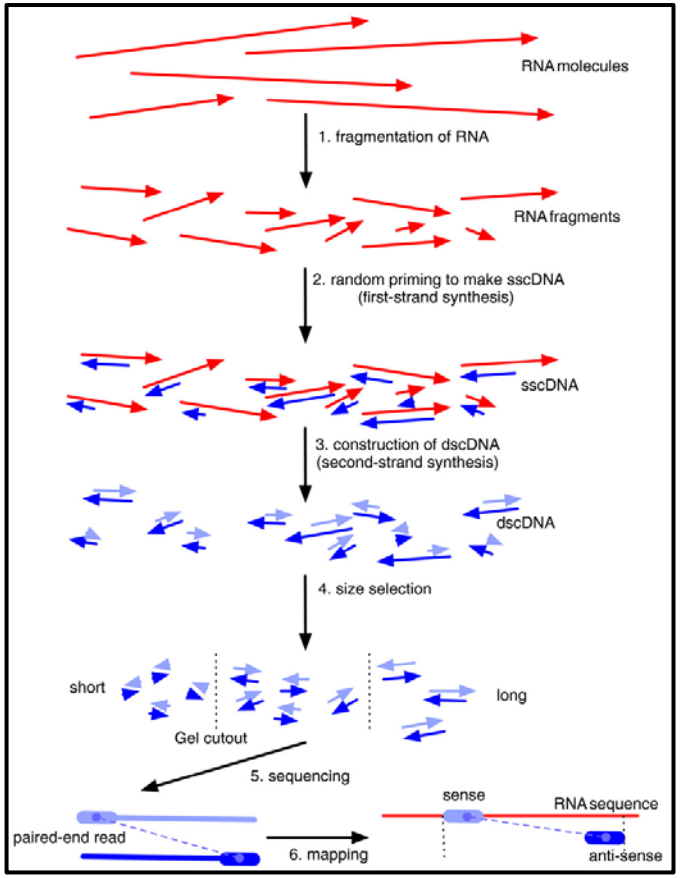
- Influence du séquençage
  - biais de position, de composition en séquence (contenu en GC)
  - influence de la longueur des transcrits

- « Mapabilité » du génome/transcriptome

Préparation de la banque

Étapes de 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

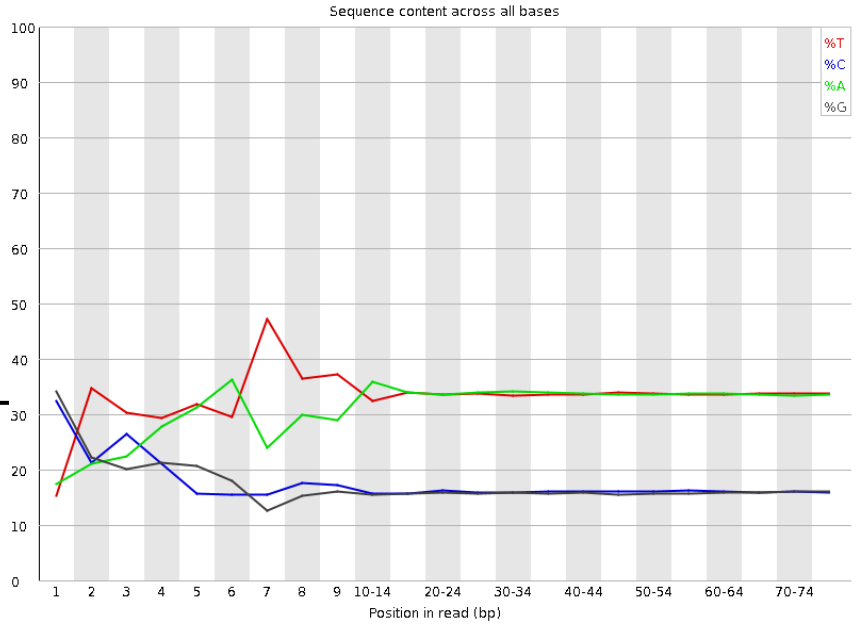
Published online 14 April 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 12 e131*
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

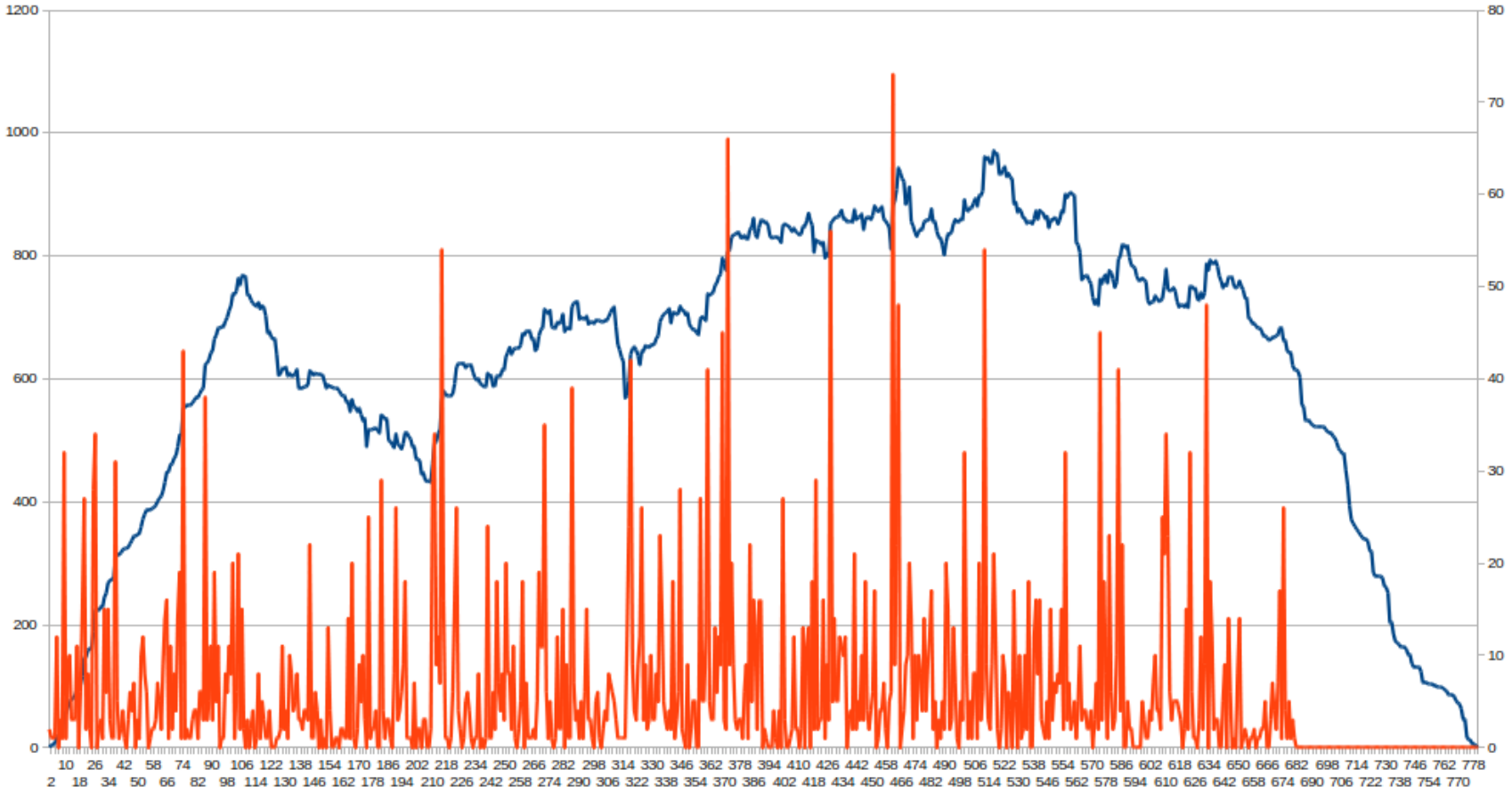
Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

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*



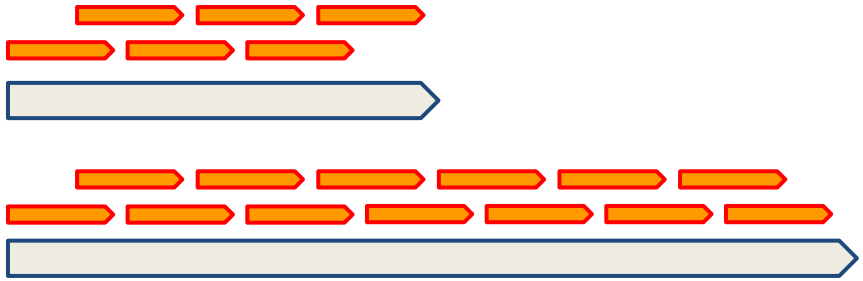
Orange = reads start sites

Blue = coverage

Biais : 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**.

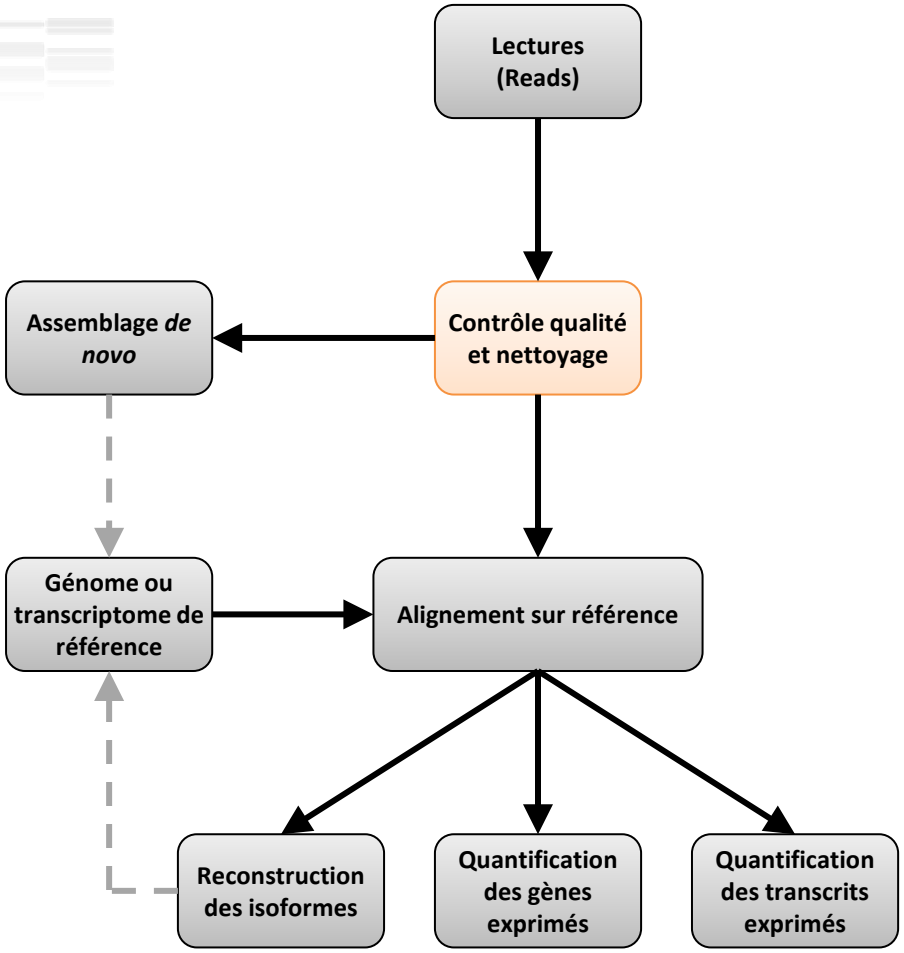
- Pour un **même gène** ayant **deux isoformes**, l'une faisant la moitié de l'autre, exprimé en **même abondance dans deux conditions différentes** :
  - L'isoforme la plus courte sera deux fois moins « comptée » que la plus longue



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



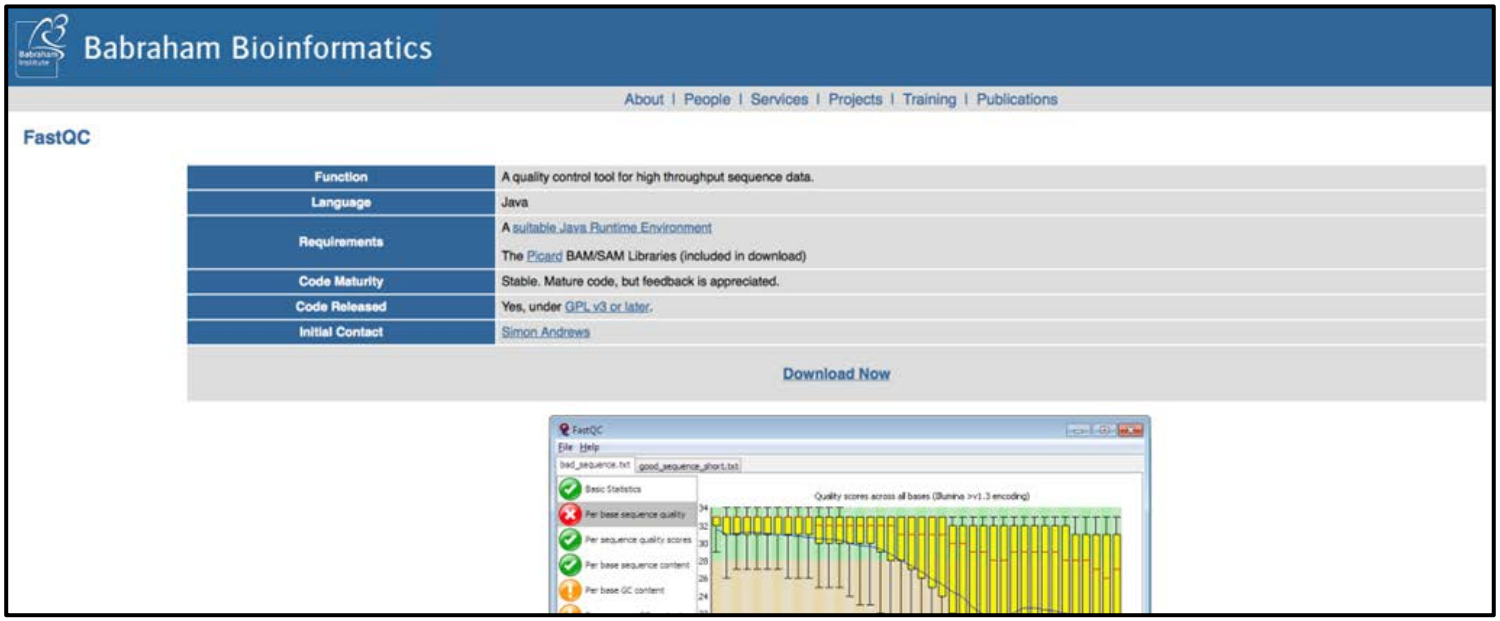
Rappel du format Fastq

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

Contrôle qualité

Outils :

- **FastQC**
  - orienté DNA-Seq



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

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

Nettoyage des données

- **Adaptateurs ou Tags**
  - Cutadapt

- **Lectures de mauvaise qualité**
  - Prinseq
  - Sickle

- **Nettoyer les lectures avec les paramètres suivants :**
  - Minimal length of 20
  - Minimal mean quality of 20
  - No N in seq
  - No 5’ trimming

Travaux pratiques

Contrôle de la qualité

- Lancer FastQC sur les 2 fastq téléchargés

- **Quels sont les biais présentés que vous pouvez identifier ?**

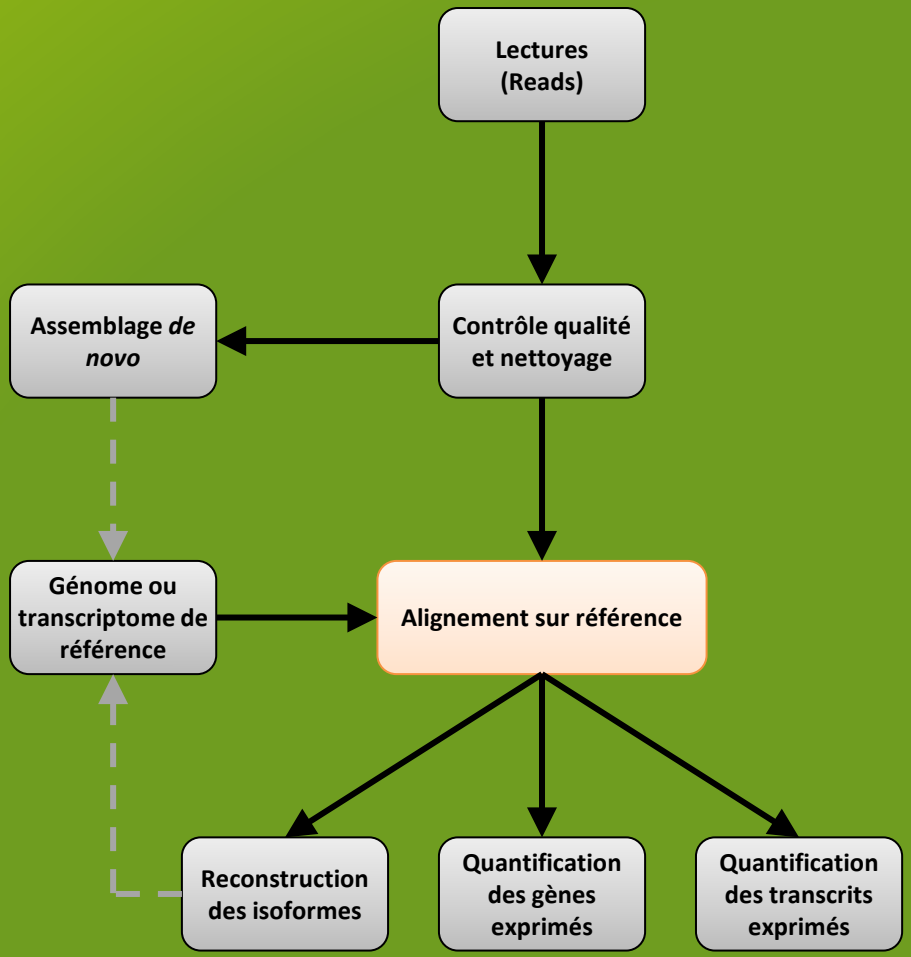
Nettoyage des données

- Exécuter les programme Sickle sur tous les fichiers fastq, et lancer FastQC pour évaluer le nettoyage réalisé par Sickle
- **Ne pas oublier de renommer les sorties**.

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**

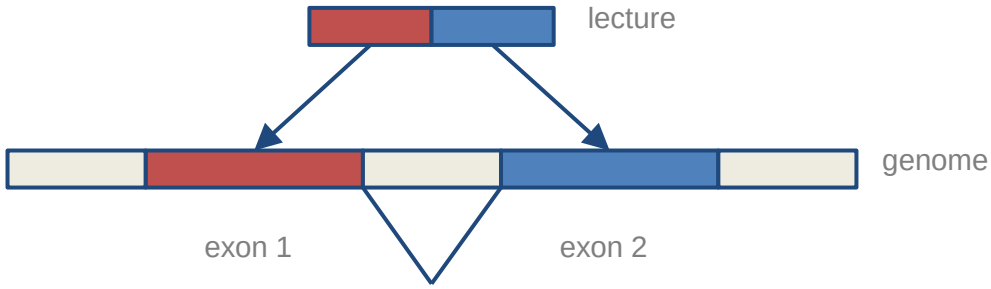
- Être capable d'**exploiter** les liste des **jonctions exons-exons connues**, mais également d'en **détecter** de **nouvelles**

- 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) au format SAM/BAM.

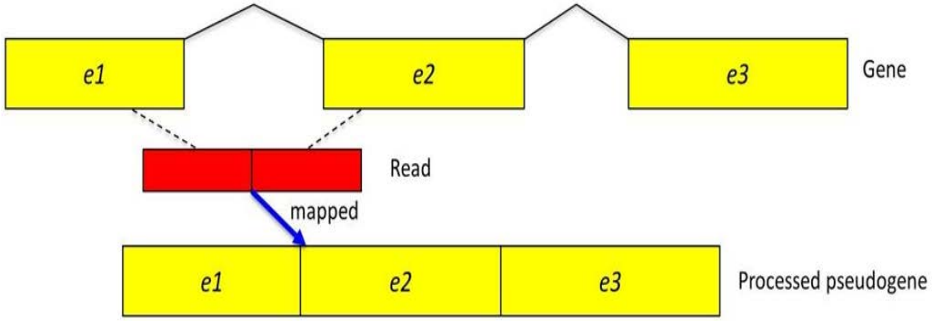


Mapping ARN ≠ Mapping ADN
Mapping ≠ Alignement

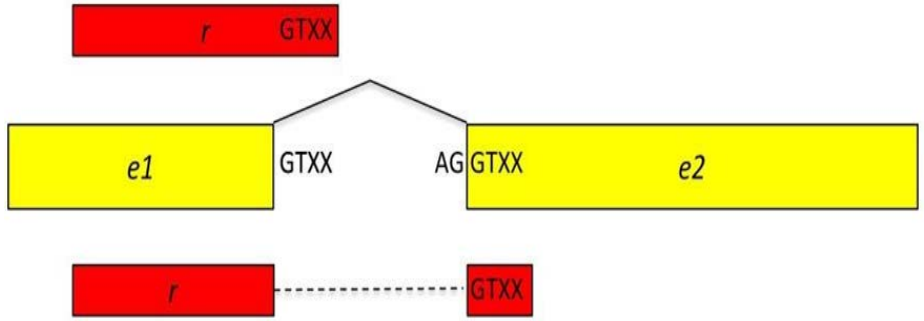
Les outils de mapping font de mauvais alignements (sauf aux jonctions).

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

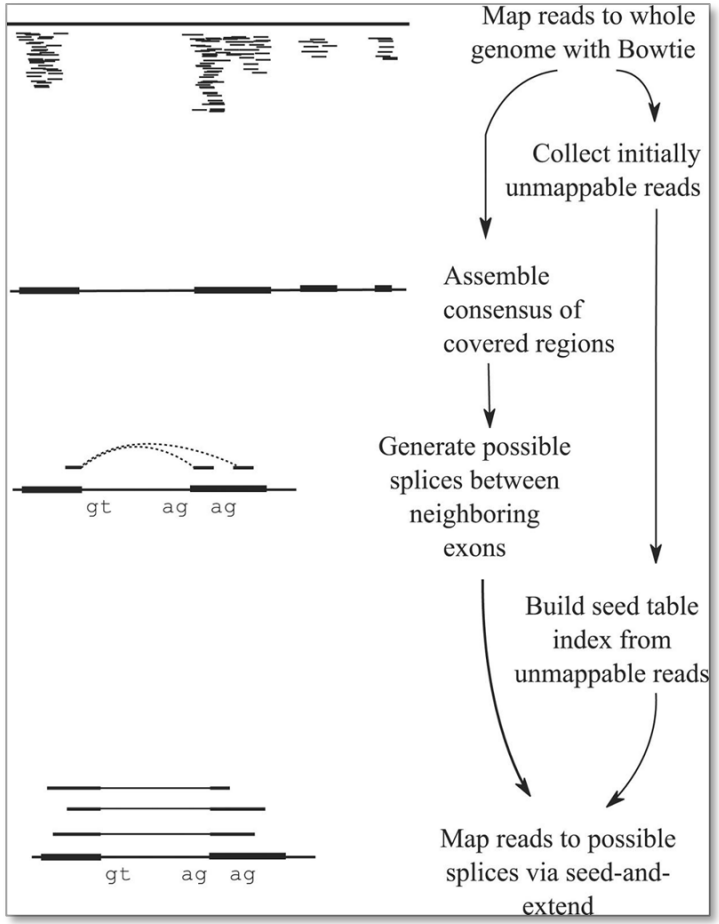
Tophat

TopHat: discovering splice junctions with RNA-Seq.

Trapnell C[1], Pachter L, Salzberg SL.

Étapes de mapping :

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

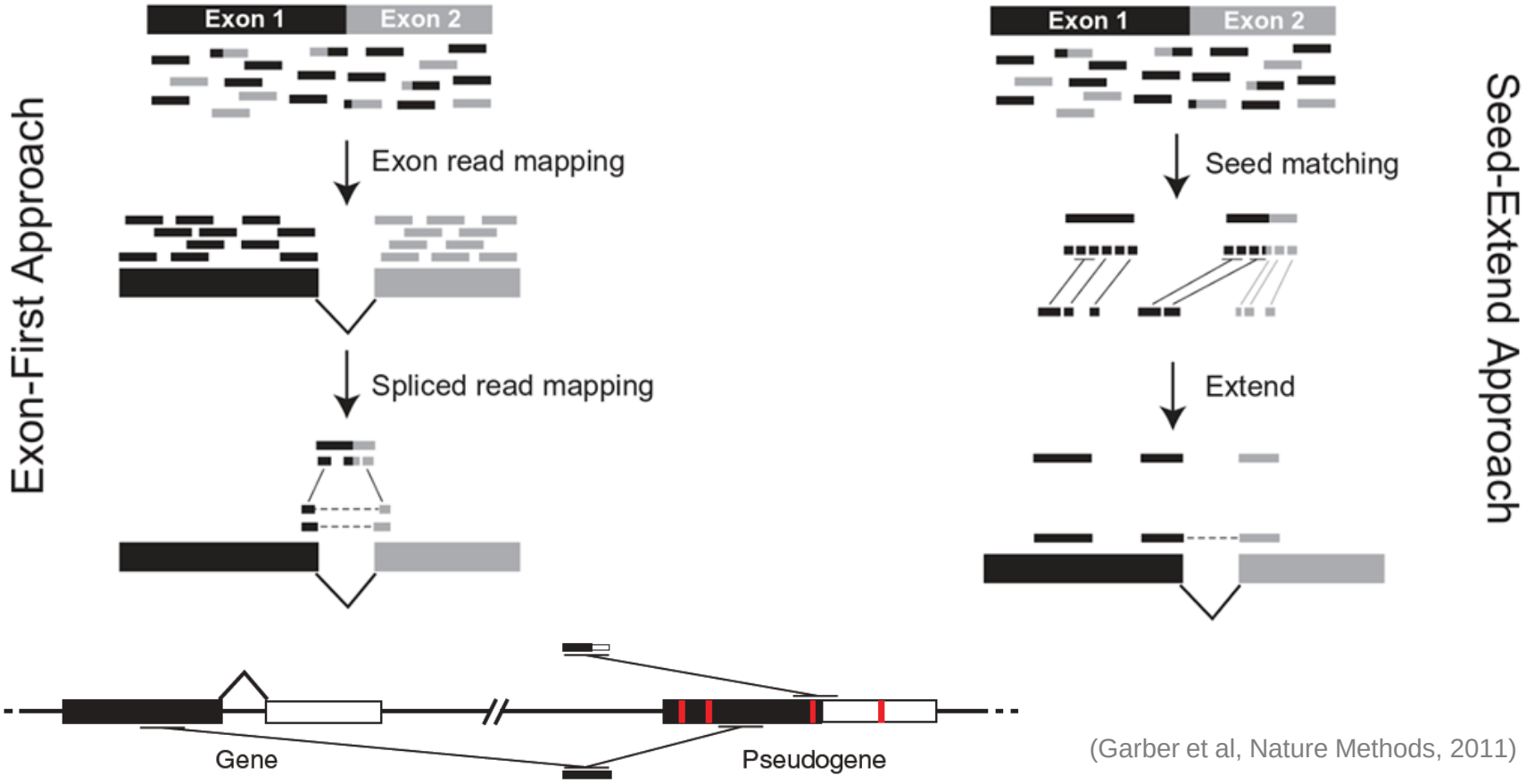


Utilise Bowtie pour mapper les lectures sur le génome.

Problème pour les pseudo-gènes !

(Trapnell et al, Bioinformatics, 2009)

2 types d'algo

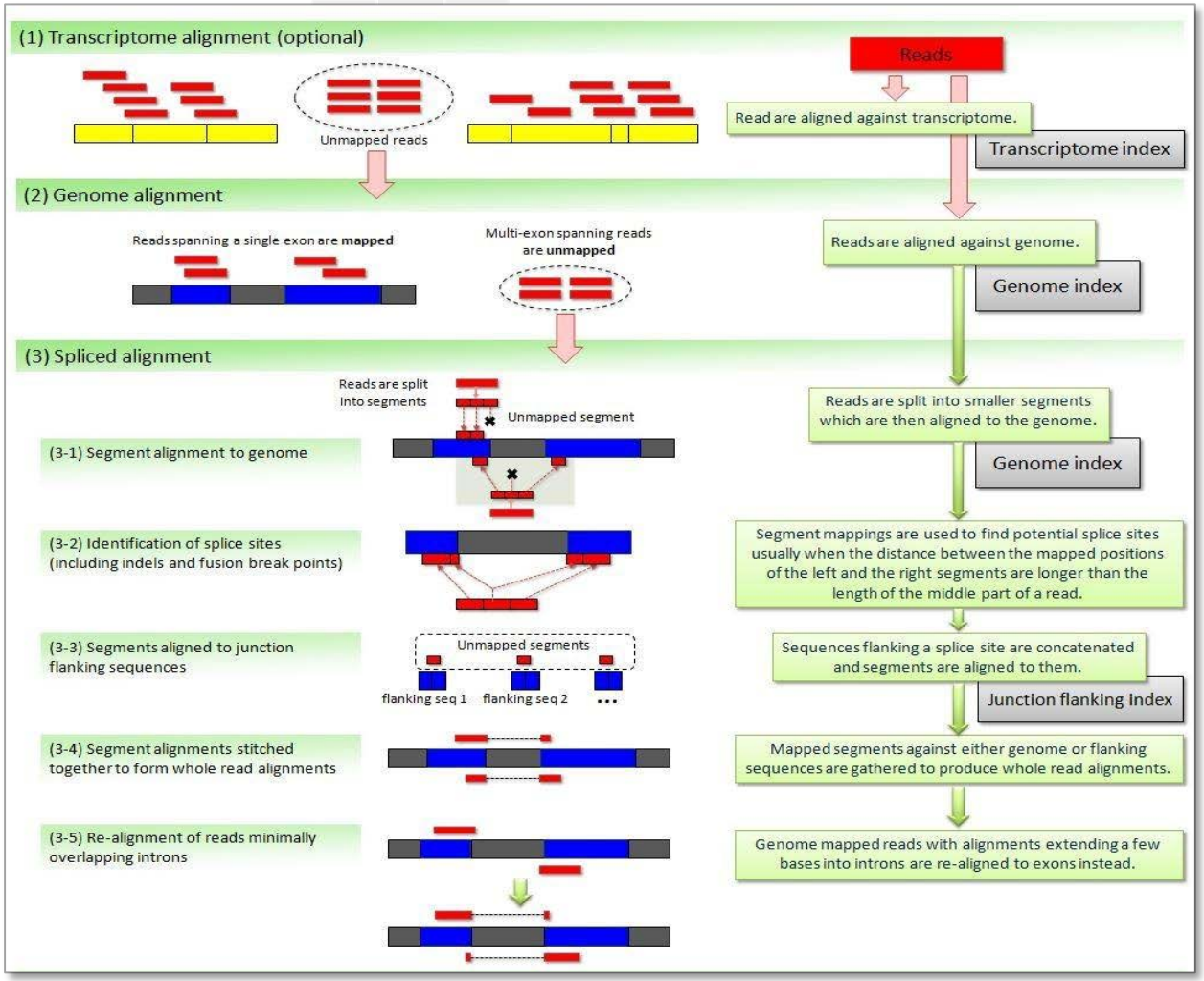


(Garber et al, Nature Methods, 2011)

Tophat2

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL.



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

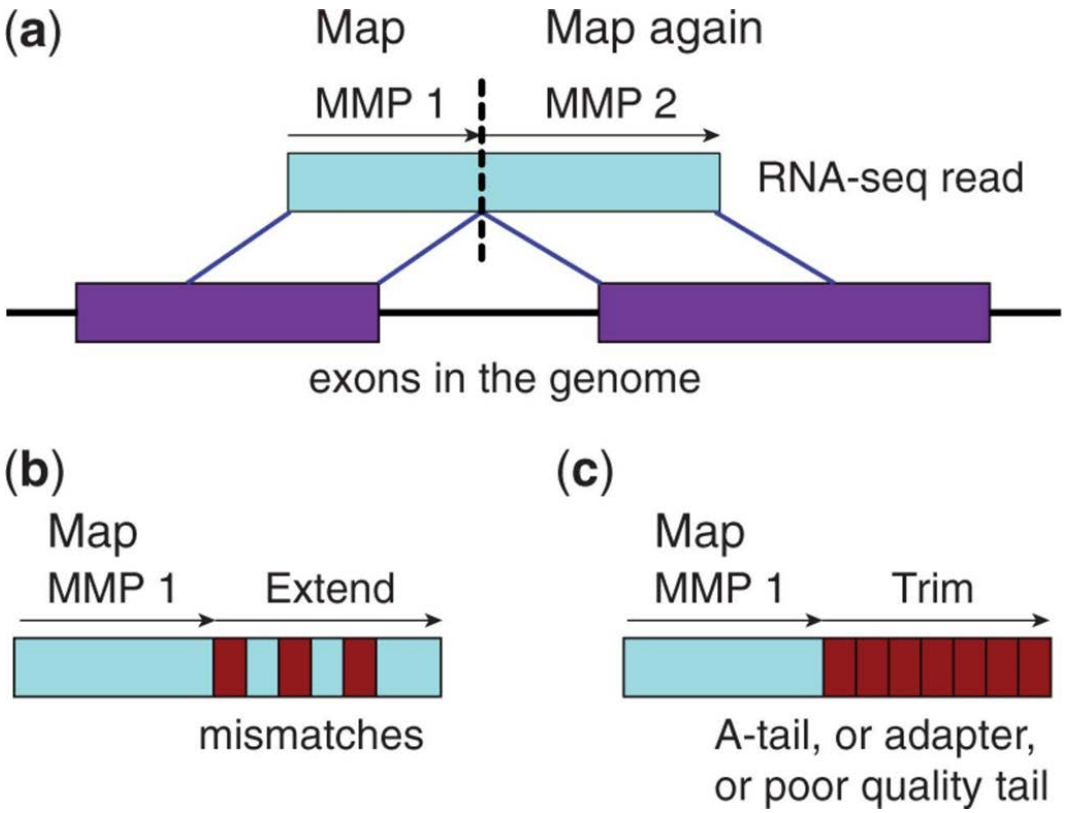
Chaque étape contient des heuristiques dont les paramètres sont à fixer.

(Kim et al, Genome Biology, 2013)

(RNA-)STAR

STAR: ultrafast universal RNA-seq aligner.

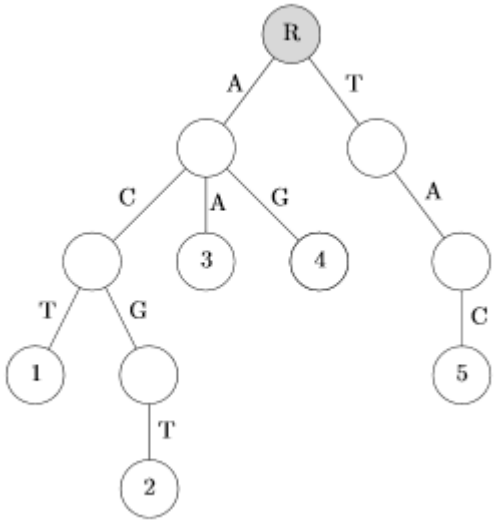
Dobin A[1], Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.



(Dobin et al, Bioinformatics, 2011)

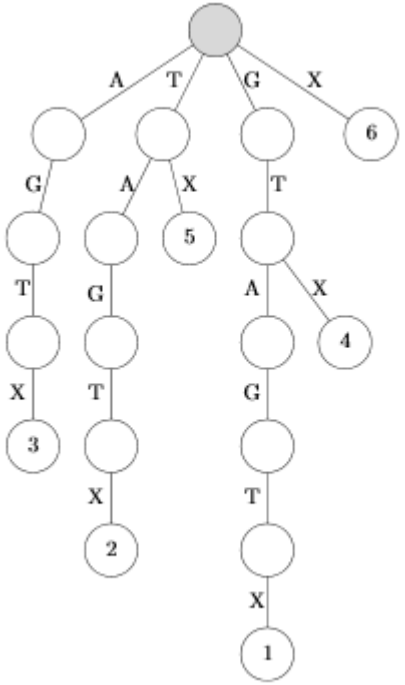
(RNA-)STAR

Aligneurs index BWT (BWA, Bowtie, SOAP)



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

STAR

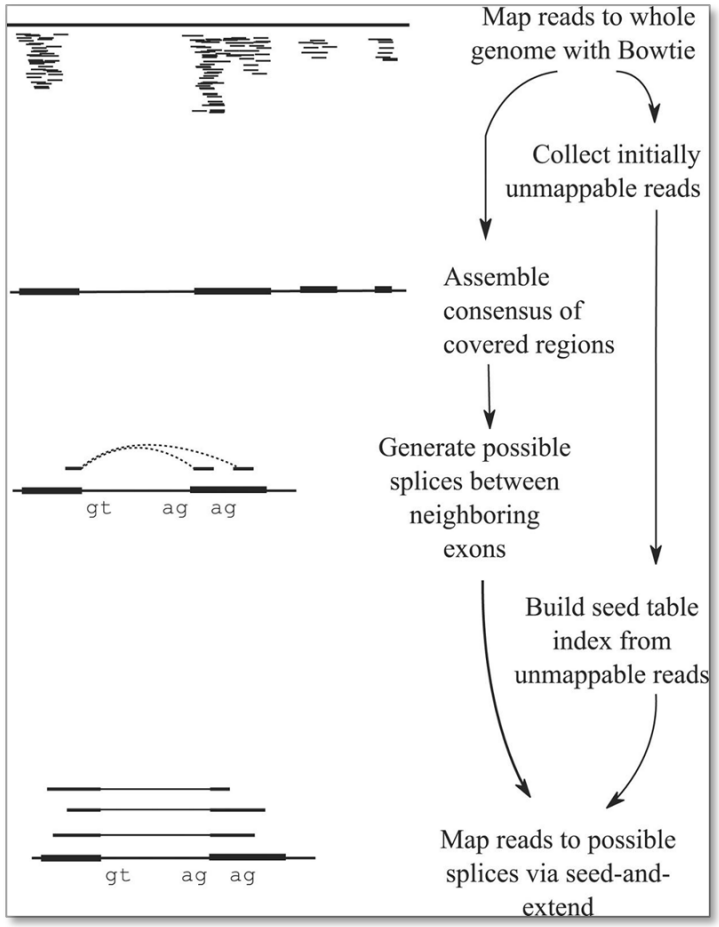


GTAGT.

Tophat1

Étapes de mapping :

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

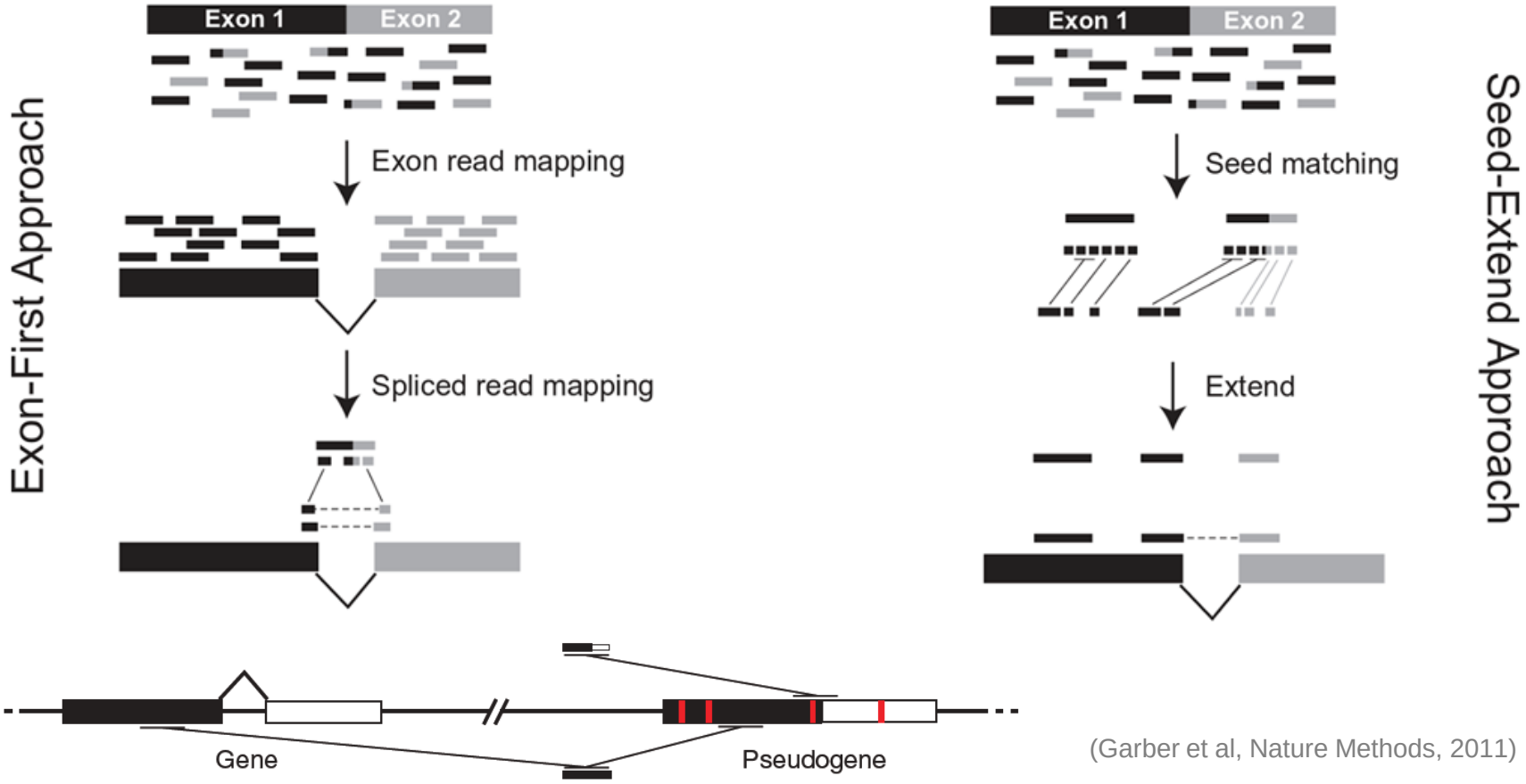


Utilise Bowtie pour mapper les lectures sur le génome.

Problème pour les pseudo-gènes !

(Trapnell et al, Bioinformatics, 2009)

2 types d'algo



(Garber et al, Nature Methods, 2011)

Outils

- Tophat2 (le plus utilisé, le plus suivi)
- STAR (runner-up)
- Crac (français !)
- GSNAP
- Subread
- MapSplice
- …

Outils

La plupart des outils

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

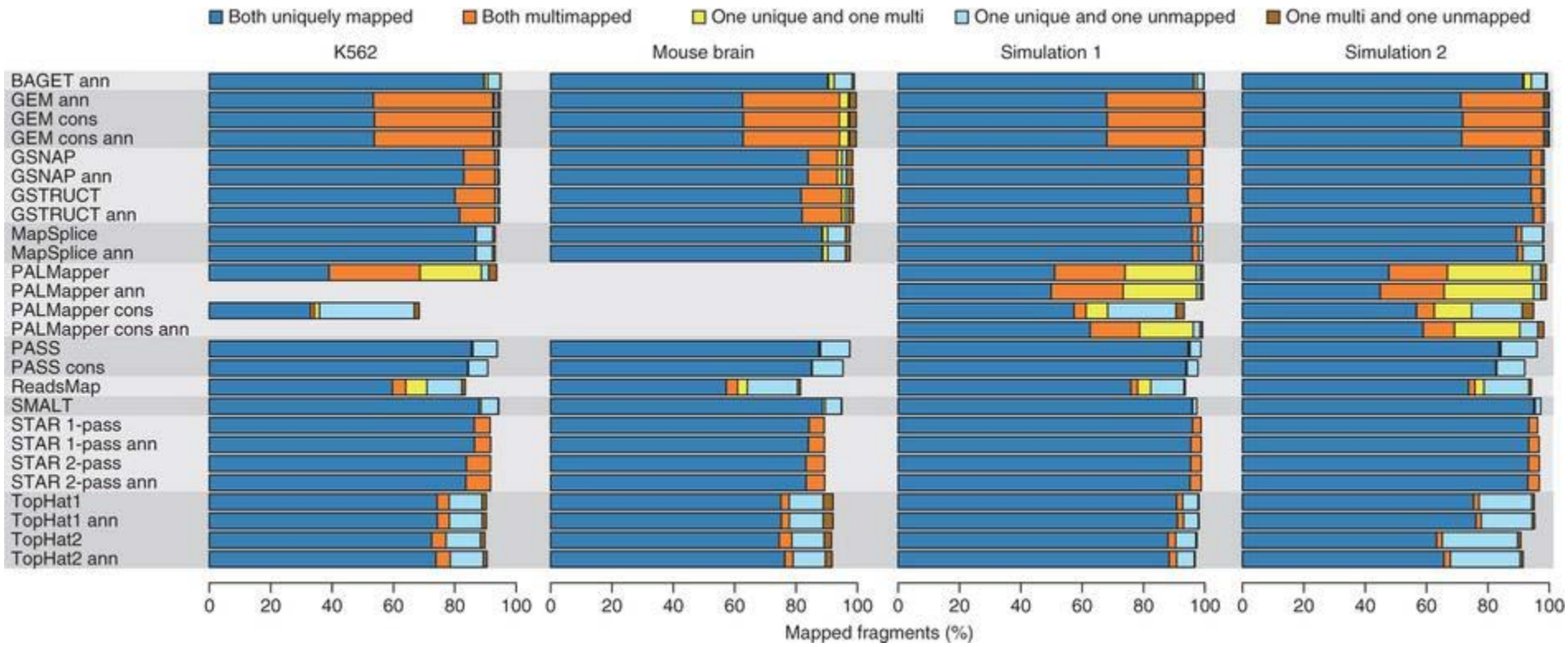
Comment évaluer un outil ?

- 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.

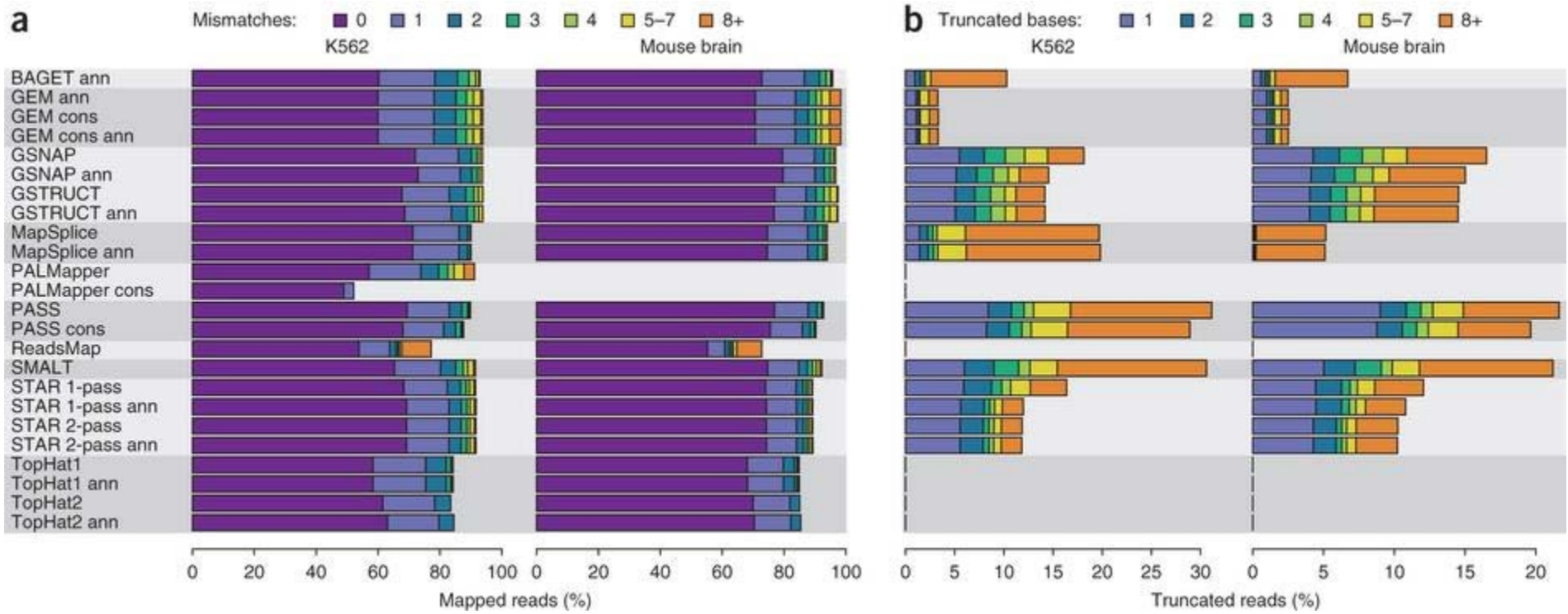
RGASP 3

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



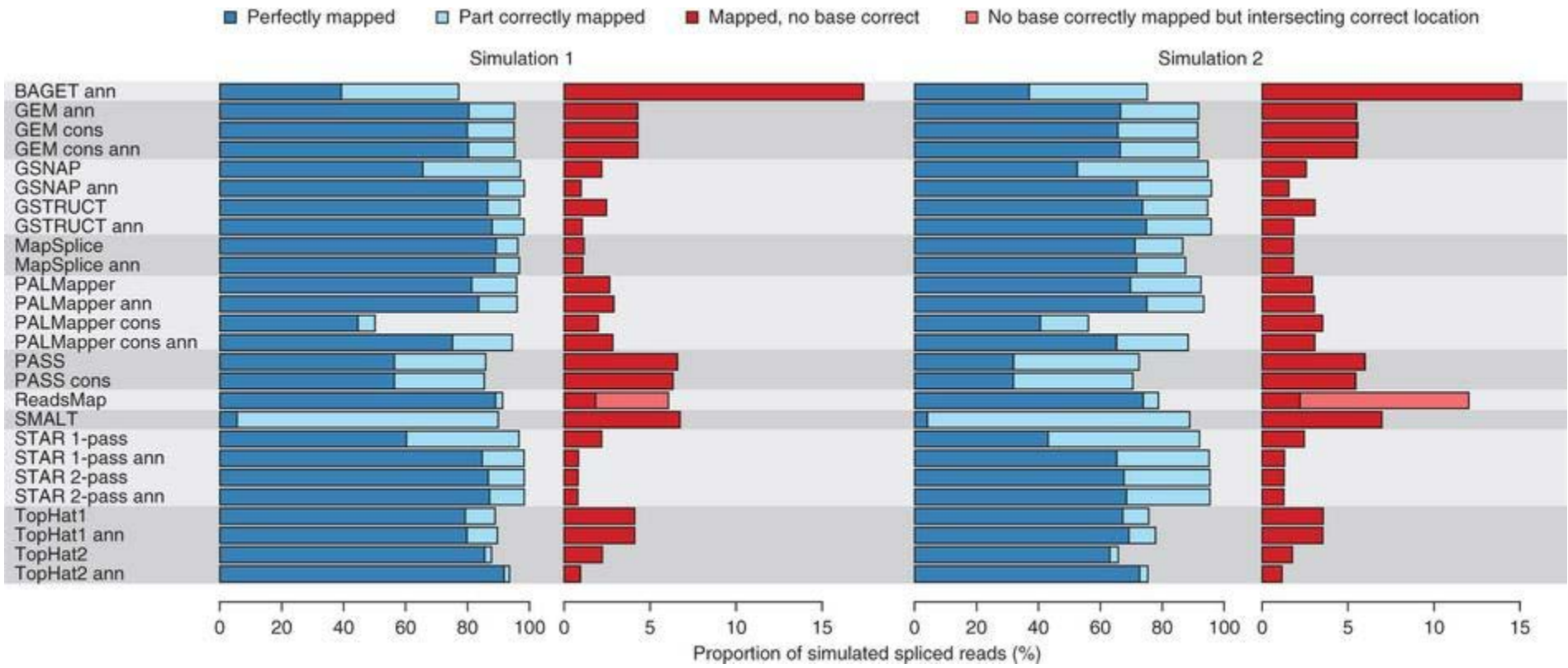
RGASP 3

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	**+**

L’alignement en pratique

Les données :

- **Lectures (brutes / nettoyées ?)**

- **Génome de référence éventuellement annoté :**
  - Séquence nucléique (fasta)
  - Annotation structurale (GTF)

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

L'alignement en pratique avec Tophat

- En **entrée** :
  - **lectures** (**.fastq**)
  - **index bowtie2** de la **référence** (**.bt2**)
  - **annotation structurale** du **génome** (**.gtf**) [optionnel]
  - **jonction** (**.bed**) [optionnel]
  - **insertions** / **délétions** (**.bed**) [optionnel]

- Toutes les options de tophat :

```
-i/--min-intron-length      <int>        [ default: 50     ]
-I/--max-intron-length      <int>        [ default: 500000 ]
-segment-length             <int>        [ default: 25     ]

-p/--num-threads            <int>        [ default: 1      ]

-G/--GTF                    <filename>   (GTF/GFF with known transcripts)
-T/--transcriptome-only                  (map only to the transcriptome)
-j/--raw-juncs              <filename>
   --insertions             <filename>
   --deletions              <filename>
 --no-novel-juncs
 --no-novel-indels
```

L'alignement en pratique avec Tophat

Attention aux paramètres par défaut !

- **Dans le manuel :**

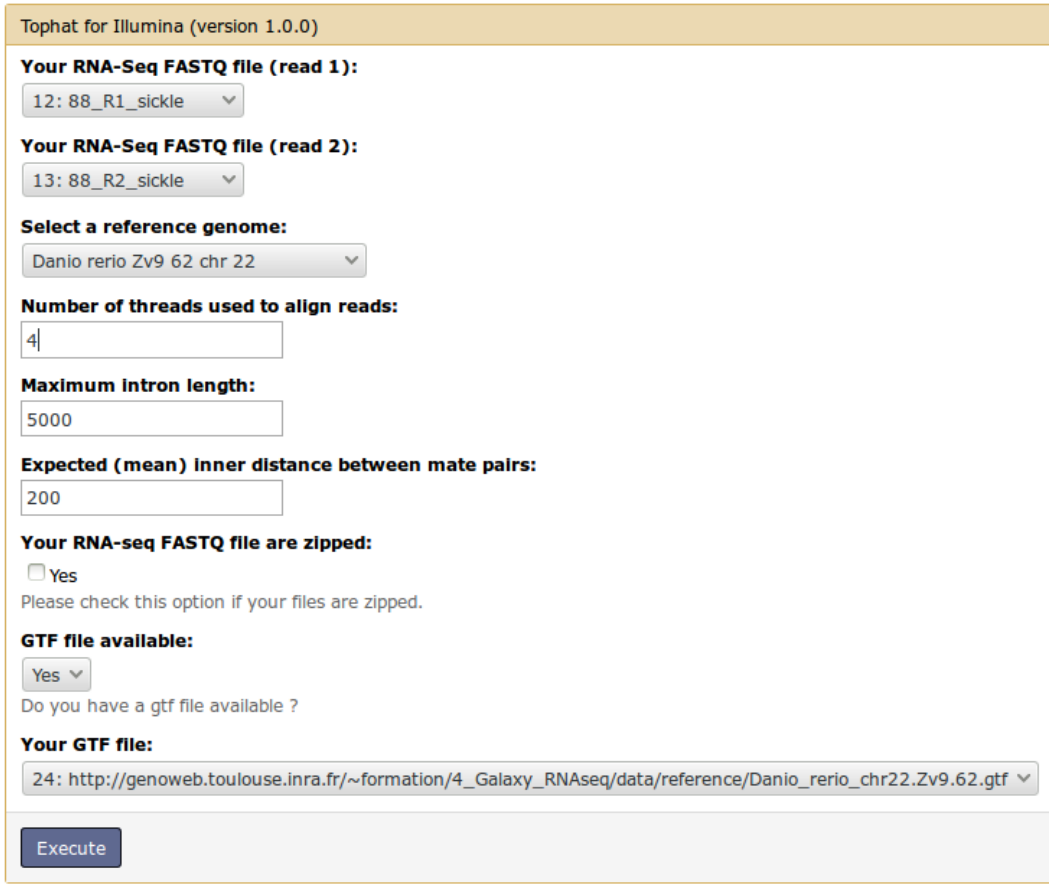
> **Please Note TopHat has a number of parameters and options, and their default values are tuned for processing mammalian RNA-Seq reads.**
>
> **If you would like to use TopHat for another class of organism, we recommend setting some of the parameters with more strict, conservative values than their 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

TP : Alignement avec Tophat

- **Lancez tophat**
  - 4 CPU
  - en paired-end
  - avec une taille d'insert de 200bp
  - taille max d'intron de 5000bp
  - contre la référence nommée « Danio rerio Zv9 62 chr 22 »
  - avec le trancriptome (gtf)

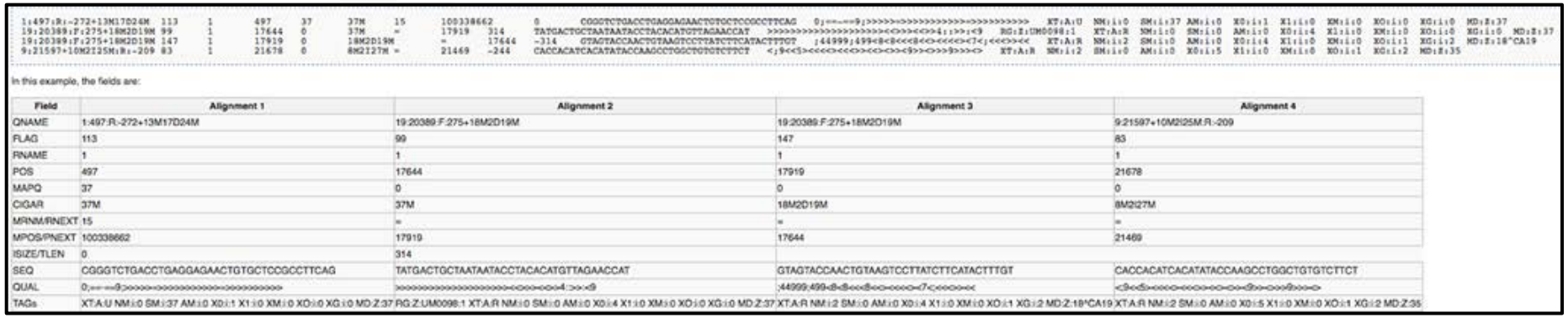
- **Vous obtenez 3 fichiers résultats :**
  - Fichier de junction (bed)
  - Fichier d'alignement (bam)
  - Fichier des reads non alignées (unmapped.bam)



- ***Avant de continuer la présentation, lancer Tophat.***

Alignement épissé : Format SAM / BAM

Sequence / Binary Alignment Map format



In this example, the fields are:

Field	Alignment 1	Alignment 2	Alignment 3	Alignment 4
QNAME	1:497:R:-272+13M17D24M	19:20389:F:275+18M2D19M	19:20389:F:275+18M2D19M	9:21597+10M2I25M:R:-209
FLAG	113	99	147	83
RNAME	1	1	1	1
POS	497	17644	17919	21678
MAPQ	37	0	0	0
CIGAR	37M	37M	18M2D19M	8M2I27M
MRNM/RNEXT	15	=	=	=
MPOS/PNEXT	100338662	17919	17644	21469
ISIZE/TLEN	0	314		
SEQ	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG	TATGACTGCTAATAATACCTACACATGTTAGAACCAT	GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT	CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT
QUAL	0;==-=9;>>>>>>>>>>>>>>>>>>>>>>>>>>	>>>>>>>>>>>>>>>>>>>>><<<<<<<<<4;>>>;<9	;44999;499<8<8<<<8<<<<<<<7<;<<<<><<	<9><5><<<<<<<<<<<<<<<<9>><<<<9>>><>
TAGs	XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	RG:Z:UM0098:1 XT:A:R NM:i:0 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37	XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:4 X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:18^CA19	XT:A:R NM:i:2 SM:i:0 AM:i:0 X0:i:5 X1:i:0 XM:i:0 XO:i:1 XG:i:2 MD:Z:35

- Spécifications du format : *http://samtools.github.io/hts-specs/SAMv1.pdf*
- SAM Flags to Plain English : *http://picard.sourceforge.net/explain-flags.html*
- **Samtools** : **bam to sam**

Alignement épissé : Chaîne CIGAR

Lettre supplémentaire : N

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

SRR031714.5132309 97 3R 8629 50 21M**789N**16M

Alignement épissé : Format BED

Browser Extensible Data format

- **Format tabulé pour la représentation d'objets**
  - **1 ligne d'en tête**
  - **3 champs obligatoires** : <chr> <start> <end>
  - **9 champs optionnels** (informations sur la façon de représenter l'objet)

```
track name=junctions description="TopHat junctions"
3R      8628    9471    JUNC00000001    4       +       8628    9471    255,0,0 2       21,33   0,810
3R      20575   20687   JUNC00000002    2       +       20575   20687   255,0,0 2       20,17   0,95
3R      23263   23382   JUNC00000003    3       +       23263   23382   255,0,0 2       21,32   0,87
3R      23248   23491   JUNC00000004    17      +       23248   23491   255,0,0 2       36,33   0,210
3R      23558   24016   JUNC00000005    4       +       23558   24016   255,0,0 2       35,36   0,422
3R      24458   24586   JUNC00000006    13      +       24458   24586   255,0,0 2       36,35   0,93
3R      24613   24762   JUNC00000007    37      +       24613   24762   255,0,0 2       35,36   0,113
3R      24991   27600   JUNC00000008    6       +       24991   27600   255,0,0 2       24,36   0,257
3R      28008   28130   JUNC00000009    69      +       28008   28130   255,0,0 2       36,33   0,89
3R      28486   29539   JUNC00000010    25      +       28486   29539   255,0,0 2       33,36   0,101
3R      29848   29970   JUNC00000011    57      +       29848   29970   255,0,0 2       36,36   0,86
3R      29866   29959   JUNC00000012    1       +       29866   29959   255,0,0 2       18,19   0,74
3R      29857   30023   JUNC00000013    18      +       29857   30023   255,0,0 2       27,36   0,130
3R      47005   47508   JUNC00000014    14      +       47005   47508   255,0,0 2       34,31   0,472
3R      47672   47810   JUNC00000015    8       +       47672   47810   255,0,0 2       36,31   0,107
3R      47895   48006   JUNC00000016    16      +       47895   48006   255,0,0 2       24,27   0,84
3R      48876   50352   JUNC00000017    13      +       48876   50352   255,0,0 2       36,32   0,1444
```

http://genome.ucsc.edu/FAQ/FAQformat.html#format1

Alignement épissé

Pour aller plus loin : Tophat-fusion

- Recherche de **points de fusion** :
  - **identification** des points de fusion dus à des **réarrangements inter-** ou **intra-chromosomiques**
  - lectures d'**au moins 50** nucléotides PE ou SE (coupées en 2)

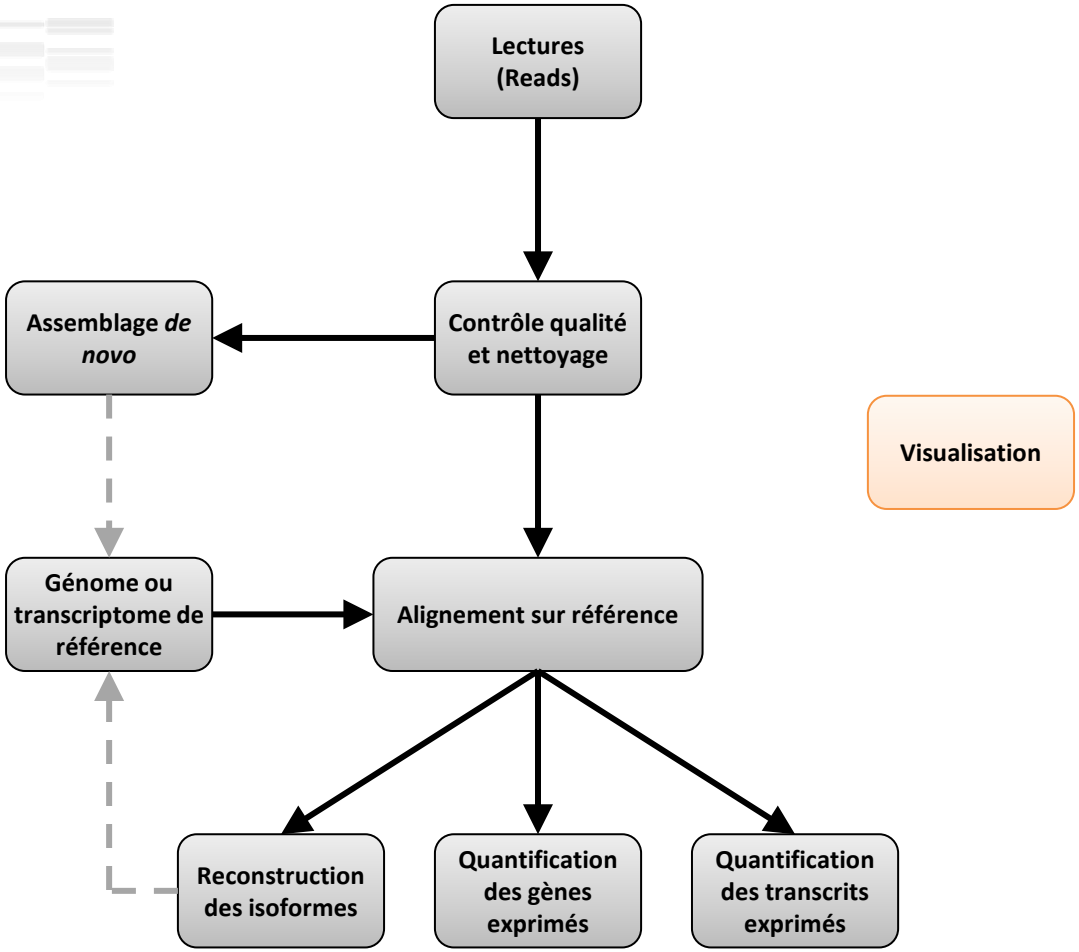
Daehwan et al. Genome Biology 2011

Alignement épissé

Conclusion

- **Alignement épissé est plus coûteux** et **compliqué que** l'**alignement** de lectures courtes **classique**
- Il peut néanmoins s'**appuyer** sur des **outils d'alignement généralistes**
- Cette **étape** est **coûteuse** en **espace disque** et **temps de calcul**, mais **facilement parallélisable par les données**
- Dans le cas de **génomes procaryotes**, on utilisera un outil d'**alignement généraliste**

Workflow d'analyse RNA-Seq



Visualisation

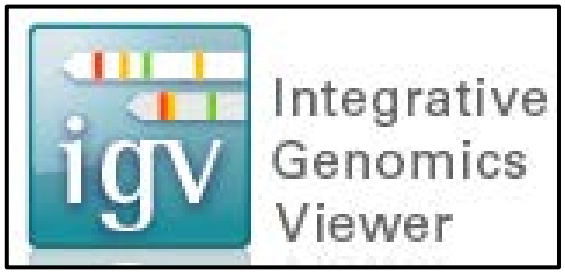
Une bonne dizaine d'outils

- **tview** (samtools)
- **IGV**
- **Tablet**
- **GenomeViewer**
- **Savant**
- **Artemis**
- **Trackster** (Galaxy)
- …

Visualisation avec IGV

Integrative Genomics Viewer

- Outil **open-source** développé au **Broad Institute**
- **Performant**, capable de gérer une **grande quantité de données**
- **Multiple formats** d'entrée
  - sam, bam, wig, biwgwig, gff, gtf, bed, custom, …
- **Documenté** et **maintenu**

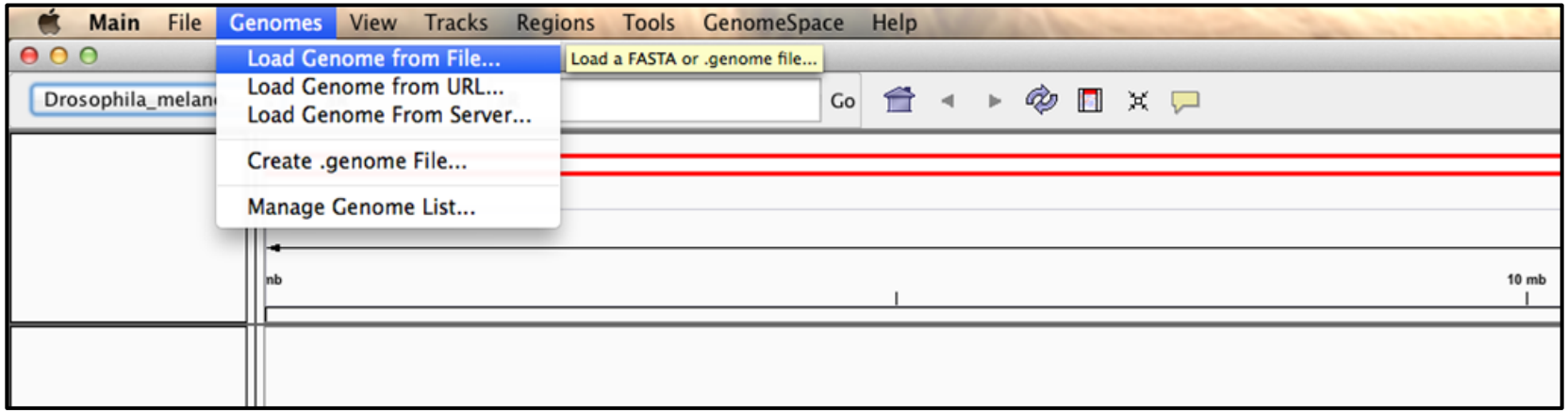


http://www.broadinstitute.org/igv/home

Visualisation avec IGV

Chargement du génome

- Charger la séquence du génome
  - fichier fasta

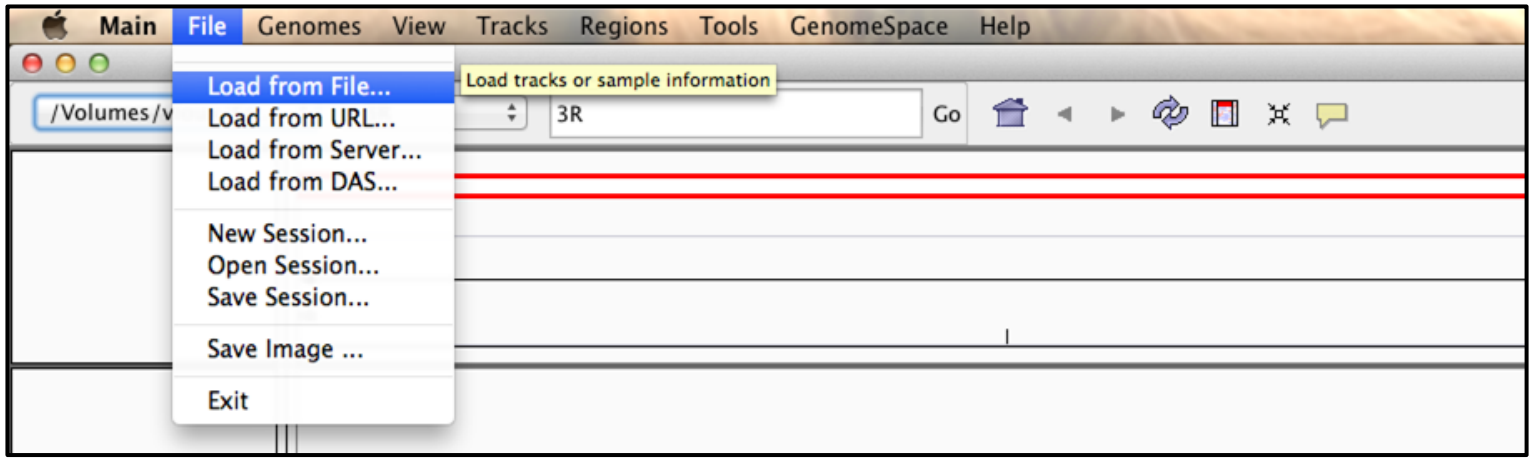


Visualisation avec IGV

Chargement des résultats

- Charger l'**annotation**
  - fichier gtf

- Charger les **résultats**
  - accepted_hits.bam
  - junctions.bed

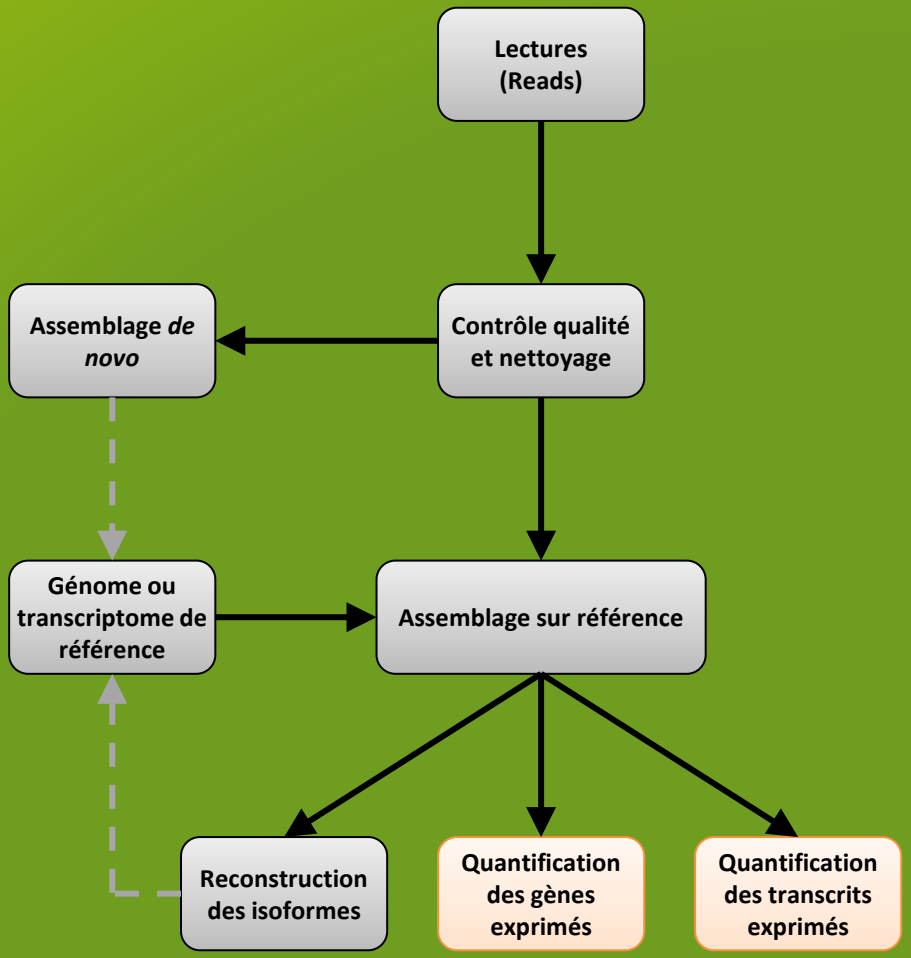


Visualisation avec IGV

TP : Visualiser les deux alignements précédents

- **Indexer les alignements** (samtools index)
- Sur votre poste de travail :
  - **Télécharger les fichiers bam, bai, bed de galaxy.**
  - **Télécharger les fichiers de séquence** (fasta) et d'**annotation** (gtf) sur www.genoweb.toulouse.inra.fr/~formation
- **Dans IGV ouvrir importer toutes ces informations**
- S'intéresser plus particulièrement aux régions :
  - chr22:586,901-596,104 : nouvel exon?
  - chr22:669,413-678,616: frontières d'exons
  - chr22:20,729,869-20,739,072 : UTR
  - chr22:25,962,028-25,970,244 : nouveau transcrit ?

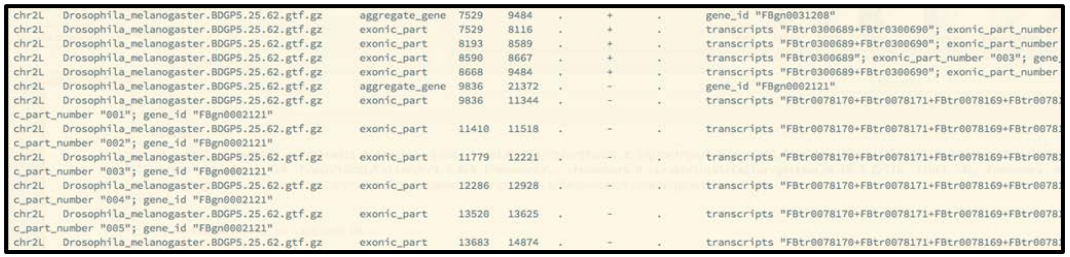
04 Quantification



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 "FBgn0031208"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     7529    8116    .       +       .       transcripts "FBtr0300689+FBtr0300690"; exonic_part_number
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     8193    8589    .       +       .       transcripts "FBtr0300689+FBtr0300690"; exonic_part_number
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     8590    8667    .       +       .       transcripts "FBtr0300689"; exonic_part_number "003"; gene
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     8668    9484    .       +       .       transcripts "FBtr0300689+FBtr0300690"; exonic_part_number
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      aggregate_gene  9836    21372   .       -       .       gene_id "FBgn0002121"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     9836    11344   .       -       .       transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBtr0078
c_part_number "001"; gene_id "FBgn0002121"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     11410   11518   .       -       .       transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBtr0078
c_part_number "002"; gene_id "FBgn0002121"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     11779   12221   .       -       .       transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBtr0078
c_part_number "003"; gene_id "FBgn0002121"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     12286   12928   .       -       .       transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBtr0078
c_part_number "004"; gene_id "FBgn0002121"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     13520   13625   .       -       .       transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBtr0078
c_part_number "005"; gene_id "FBgn0002121"
chr2L   Drosophila_melanogaster.BDGP5.25.62.gtf.gz      exonic_part     13683   14874   .       -       .       transcripts "FBtr0078170+FBtr0078171+FBtr0078169+FBtr0078
```

- **Obtenir une matrice de comptage :**

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

Vers une Analyse Différentielle …

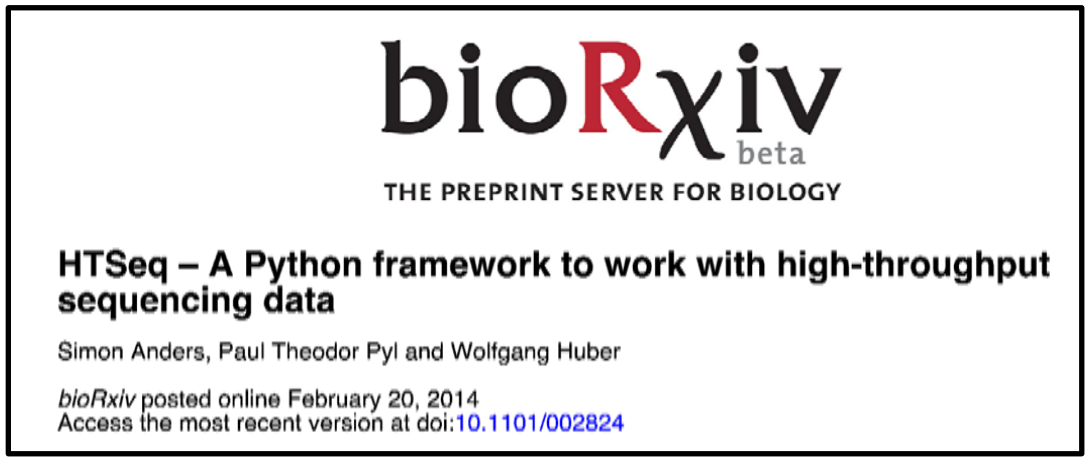
- Au niveau Gènes :
  - quantification : Htseq-count
  - analyse différentielle : DEseq ou EdgeR (Package R)
- Au niveau Transcrits (suite cufflinks) :
  - abondance : cufflinks
  - analyse différentielle : cuffdiff
- Au niveau Transcrits (à venir) :
  - quantification : featureCounts
  - analyse différentielle : DEseq ou EdgeR (Package R)

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

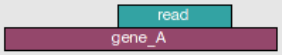
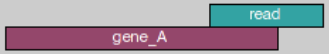
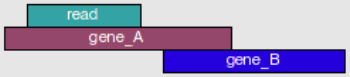
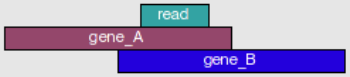
HTSeq – A Python framework to work with high-throughput sequencing data

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
```

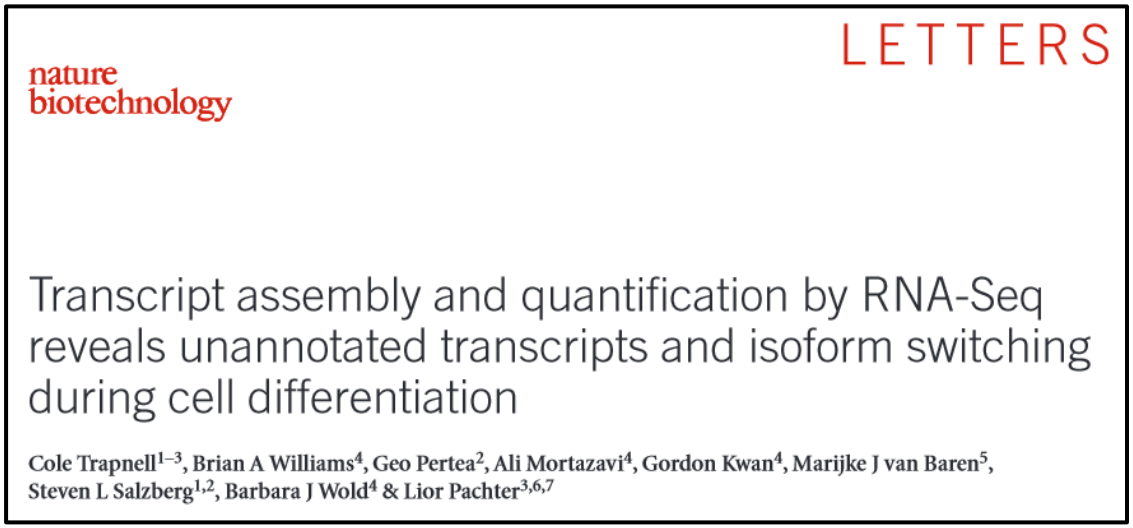
HTSeq-count

TP : Quantification des gènes

- Galaxy (Toulouse) : les bams doivent être trié par readname.
- Produire la **table de comptage** des **gènes**, en **mode intersection non-empty**, à partir des alignements de chaque échantillon.
- **Créer la matrice de comptage à l'aide de « Merge tabulated files »**

Cufflinks

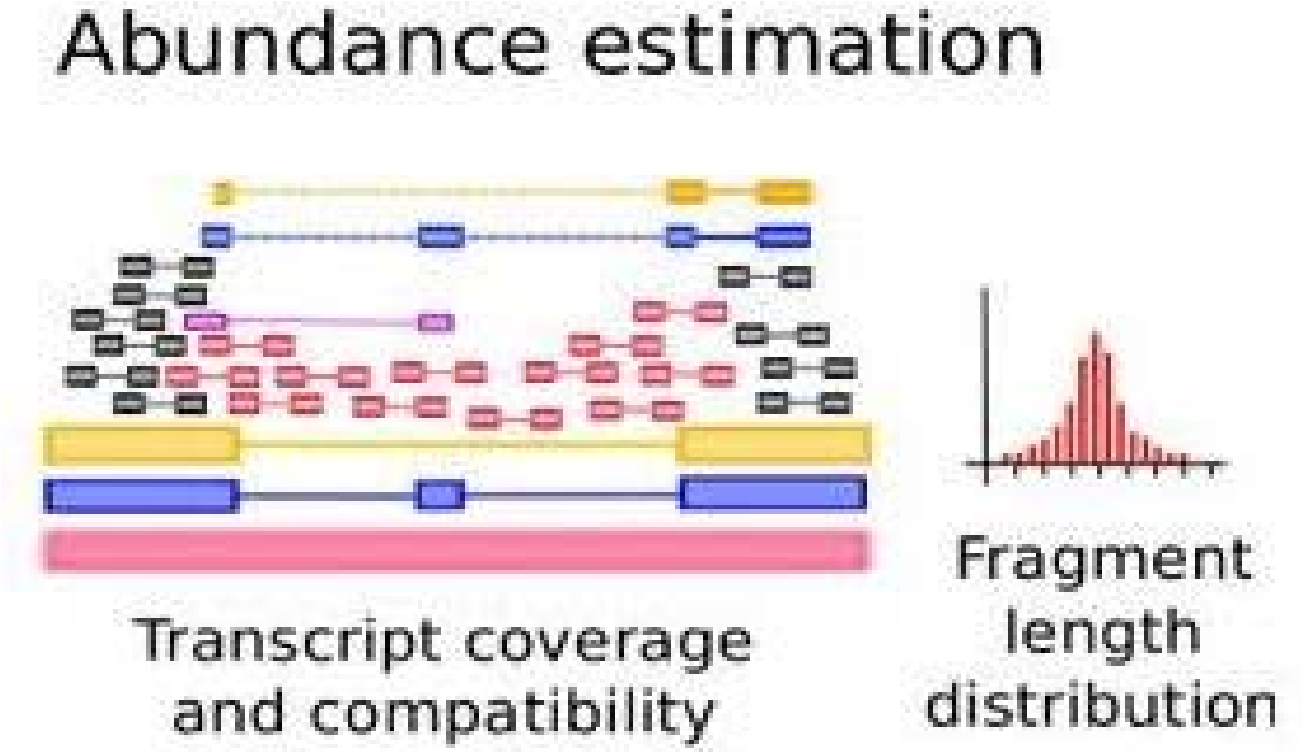
- **Pipeline / suite logiciel** de **traitement RNA-Seq** :
  - **assemble** les **transcrits**
  - **quantifie** l'**abondance** des transcrits
  - **compare** les **annotations** des transcrits
  - **analyse** l'**expression différentielle** des transcrits



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

Cufflinks

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



Trapnell et al. Nat Biotechnol. 2010

Cufflinks

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

  C = Nombre de read mappés

  N = Nombre total de read de la librairie

  L = taille des exons du gène en bp

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

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

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

Mortazavi et al. Nature Methods 2008

Comptage brut de cufflinks ?

sigcufflink

- Modification du code de cufflinks (membre de Sigenae) sur genotoul et galaxy de toulouse
- Exactement les même options que cufflinks

gene_id	transcript_id	pairs	forward		reverse	
CUFF.6	CUFF.6.1	4873	4873		3431	
CUFF.6	CUFF.6.2	5222	5222		3769	
CUFF.6	ENSDART00000067635		4819	4819		3580

FeatureCounts

featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

Yang Liao[1,2], Gordon K. Smyth[1,3] and Wei Shi[1,2,*]

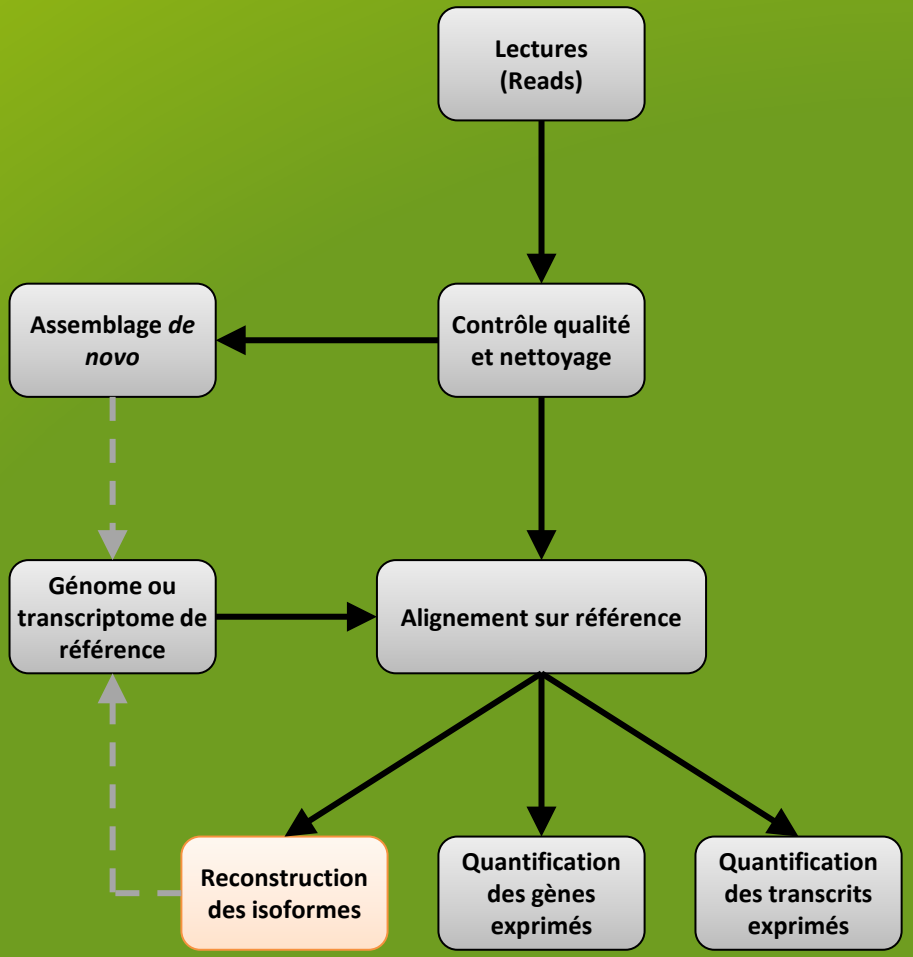
- **htseq-count ++**
- **Niveau exon, gène, transcrit**
- **1 read peut être attribué à plusieurs Feature,**
- **Reads avec alignement multiples peuvent être pris en compte**
- **Brin-spécifique très bien géré.**

TP : Cufflinks

Quantification des transcrits

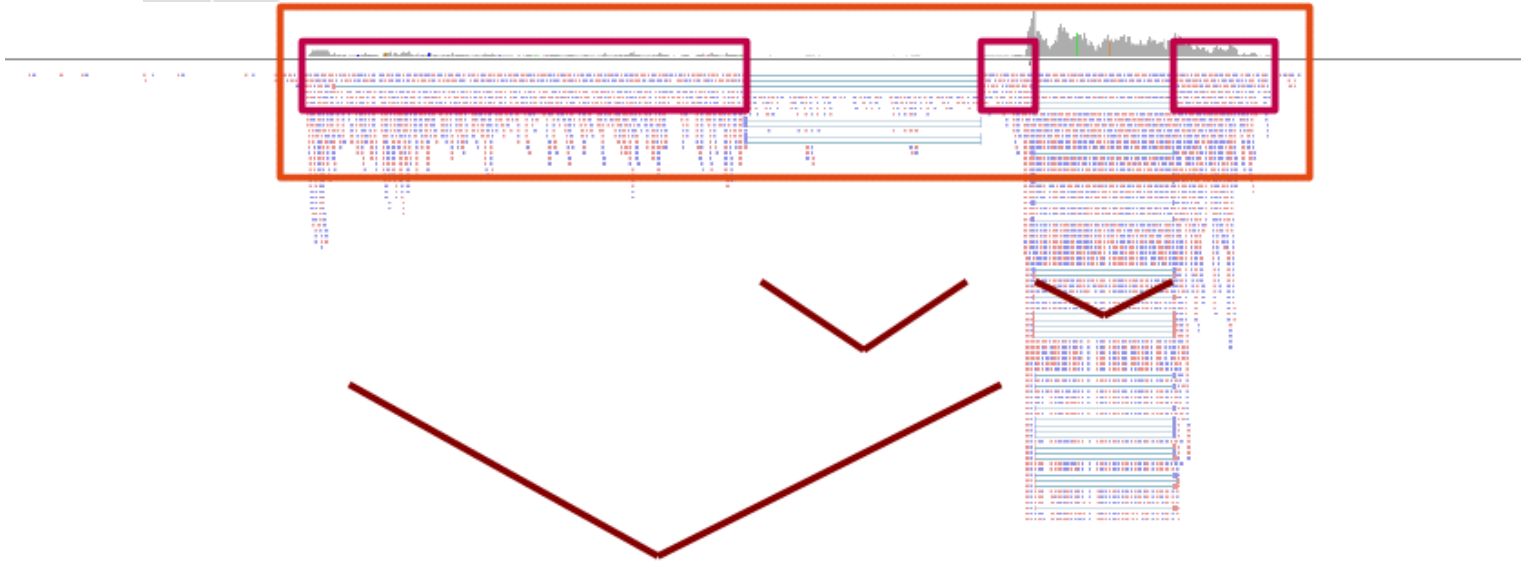
- Avec Cufflinks pour les deux échantillons
- Avec Sigcufflinks pour les deux échantillons
- En **entrée** :
  - **lectures** (**.sam/.bam**)
  - **annotations** (**.gtf**) : option « Use Reference Annotation » à « estimate isoform expression »
  - **Taille d'intron** 5000

- Créer la matrice de comptage à partir des données Sigcufflinks a l'aide de l'outil « Merge sigcufflinks »



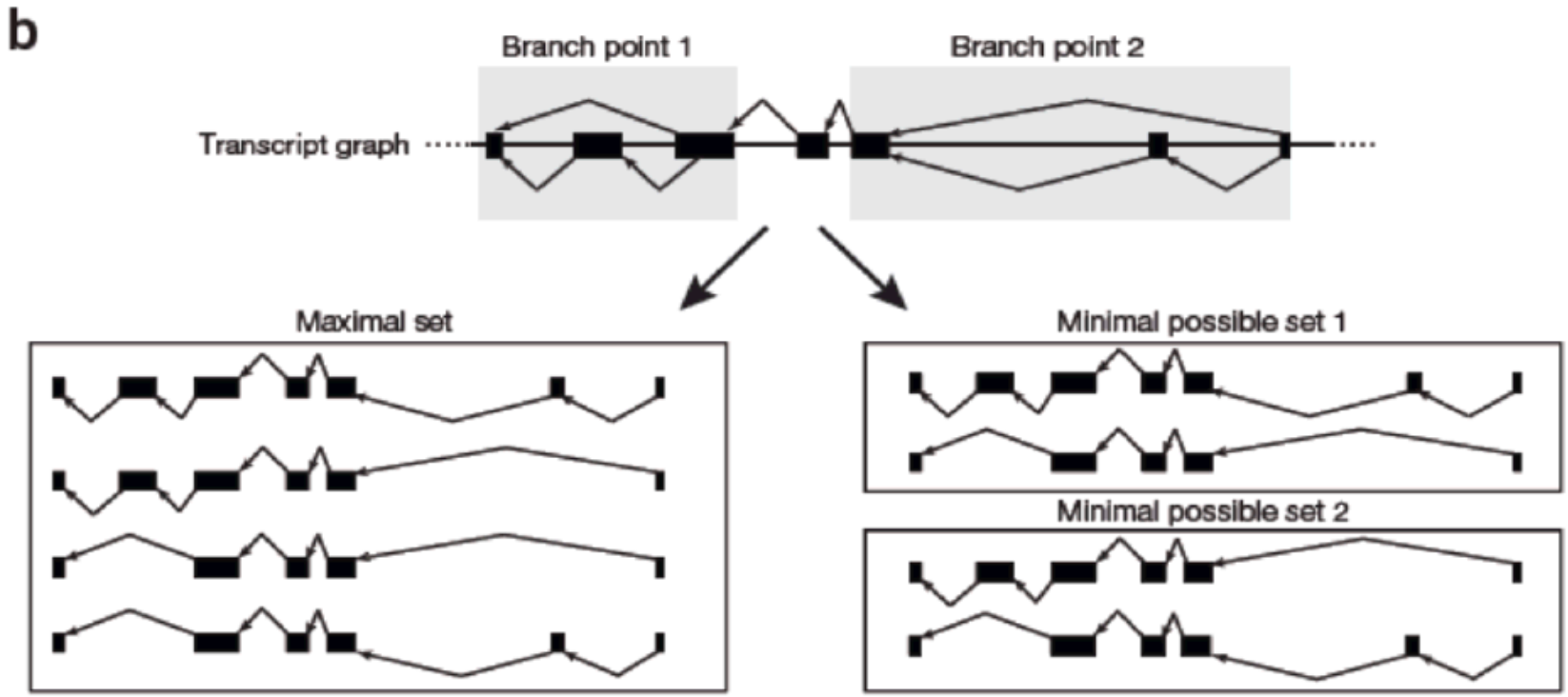
_05 Reconstruction de transcrits

Reconstruction de transcrits



gene location
Exon location
Junctions :
- Between read pair junction
- Within read junction

Stratégie de reconstruction d'un modèle



REVIEW

Computational methods for transcriptome annotation and quantification using RNA-seq

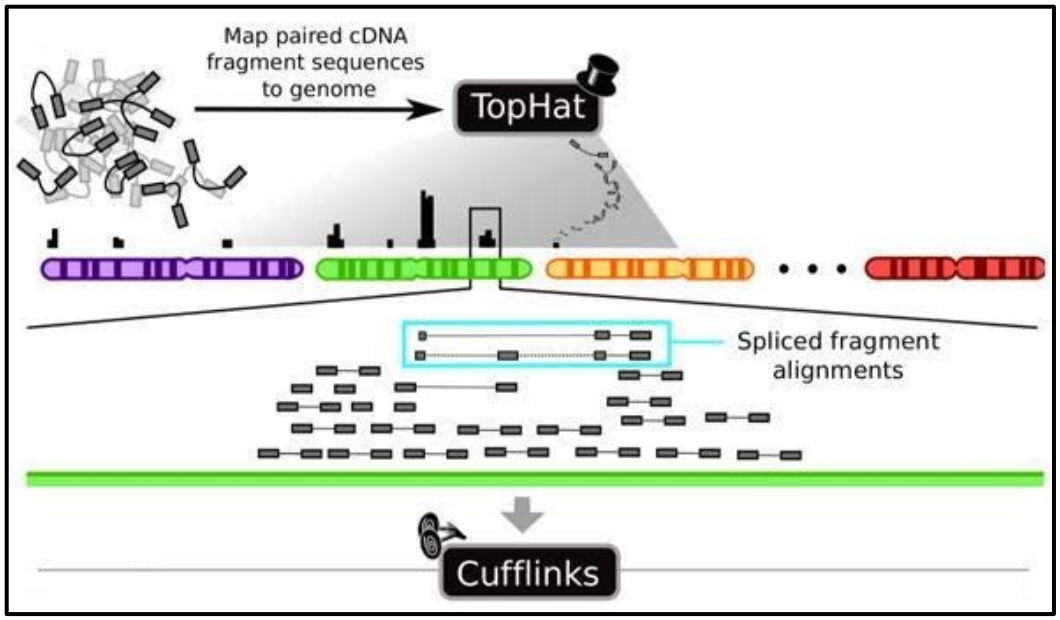
Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

Cufflinks

Reconstruction de **transcrits**

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

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

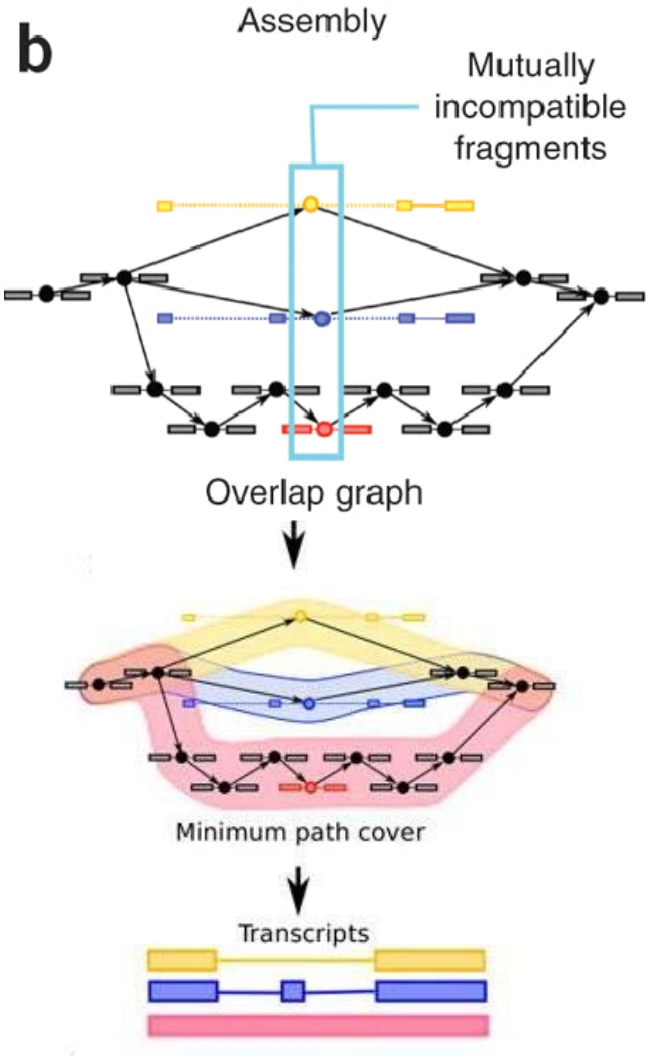


Trapnell et al. Nat Biotechnol. 2010

Cufflinks : Assembly

Reconstruction de **transcrits**

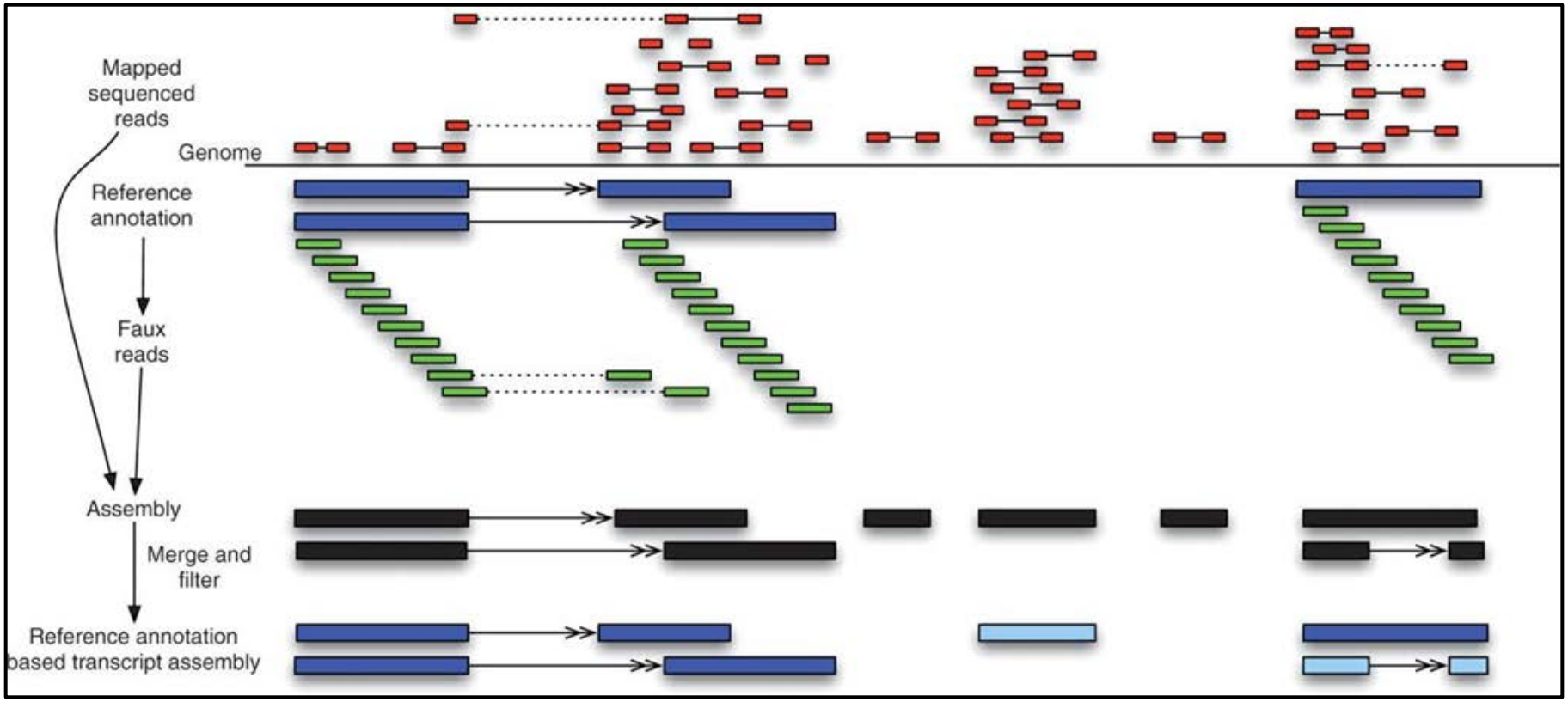
- **Stratégie** de **construction** du **modèle**
  - trouver le **nombre minimum de modèles qui expliquent les lectures** :
    - **minimum de chemins**, théorème de Dilworth
    - **nb** de **lectures incompatibles** = **nb minimum** de **transcrits** nécessaires
    - **1 chemin = 1 isoforme**



Trapnell et al. Nat Biotechnol. 2010

Cufflinks : RABT

Reference Annotation Based Transcripts Assembly



Roberts et al. Bioinformatics 2011

Cufflinks

Reconstruction des **transcrits**

- En **entrée** :
  - **lectures** (**.sam/.bam**)
  - **annotations** (**.gtf**)

```
-g/--GTF-guide <reference_annotation.(gtf/gff)> :guide RABT (Reference Annotation Based
Transcript) assembly
```

- 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

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**

22	Cufflinks	transcript	9743035	9747366	349	-	.	
22	Cufflinks	exon	9743035	9745254	349	-	.	

Si RABT, indique si tous les introns et les exons sont couvert par les reads

gene_id "CUFF.560"; transcript_id "CUFF.560.1"; FPKM "23.7787563790"; frac "0.143485"; conf_lo "8.754478"; conf_hi "38.803035"; cov "2.840328"; full_read_support "yes";
gene_id "CUFF.560"; transcript_id "CUFF.560.1"; exon_number "1"; FPKM "23.7787563790"; frac "0.143485"; conf_lo "8.754478"; conf_hi "38.803035"; cov "2.840328";

Estimation de la profondeur

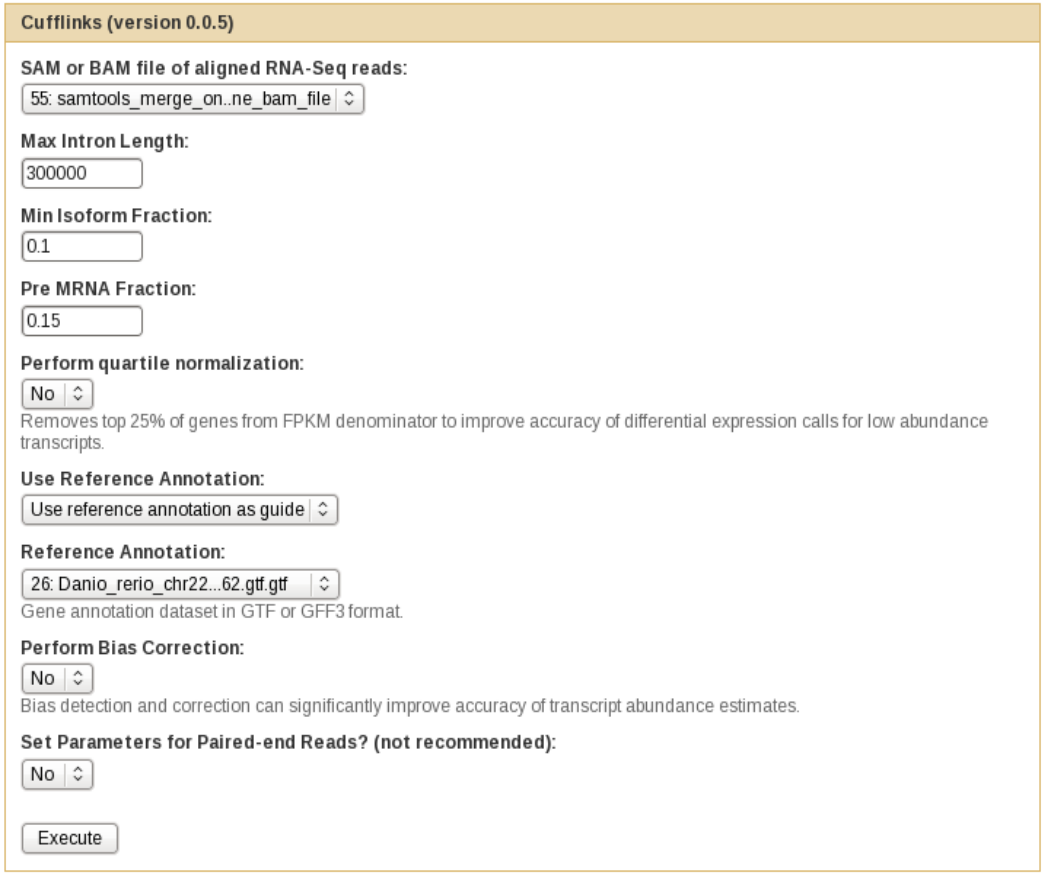
Workflow reconstruction de transcrits ?

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

Cufflinks

TP : **Reconstruction** des **transcrits**

- Fusionner les bam (samtools merge) des deux échantillons
- Détecter les nouveaux transcrits avec Cufflinks
- En sortie vous obtenez:
  - **accepted_hits** (**.bam**)
  - **junctions** (**.bed**)
  - **transcripts** (**.gtf**)
- Indexer le fichier bam
- Télécharger ces fichiers pour les visualiser dans IGV



Quantification & Découverte de nouveau transcripts

Conclusion

- En général découverte puis quantification.
- Découverte :
  - Merge des échantillons
  - Suppression des dupplicats
  - Cufflinks

- Quantification : comptage gène, comptage transcrits
- Deux gammes outils :
  - **Htseq-count** : analyse avec DESeq, edgeR, …
  - **cufflinks/cuffdiff**: tout intégré

06

Conclusion

Conclusion générale

- Workflow à construire …

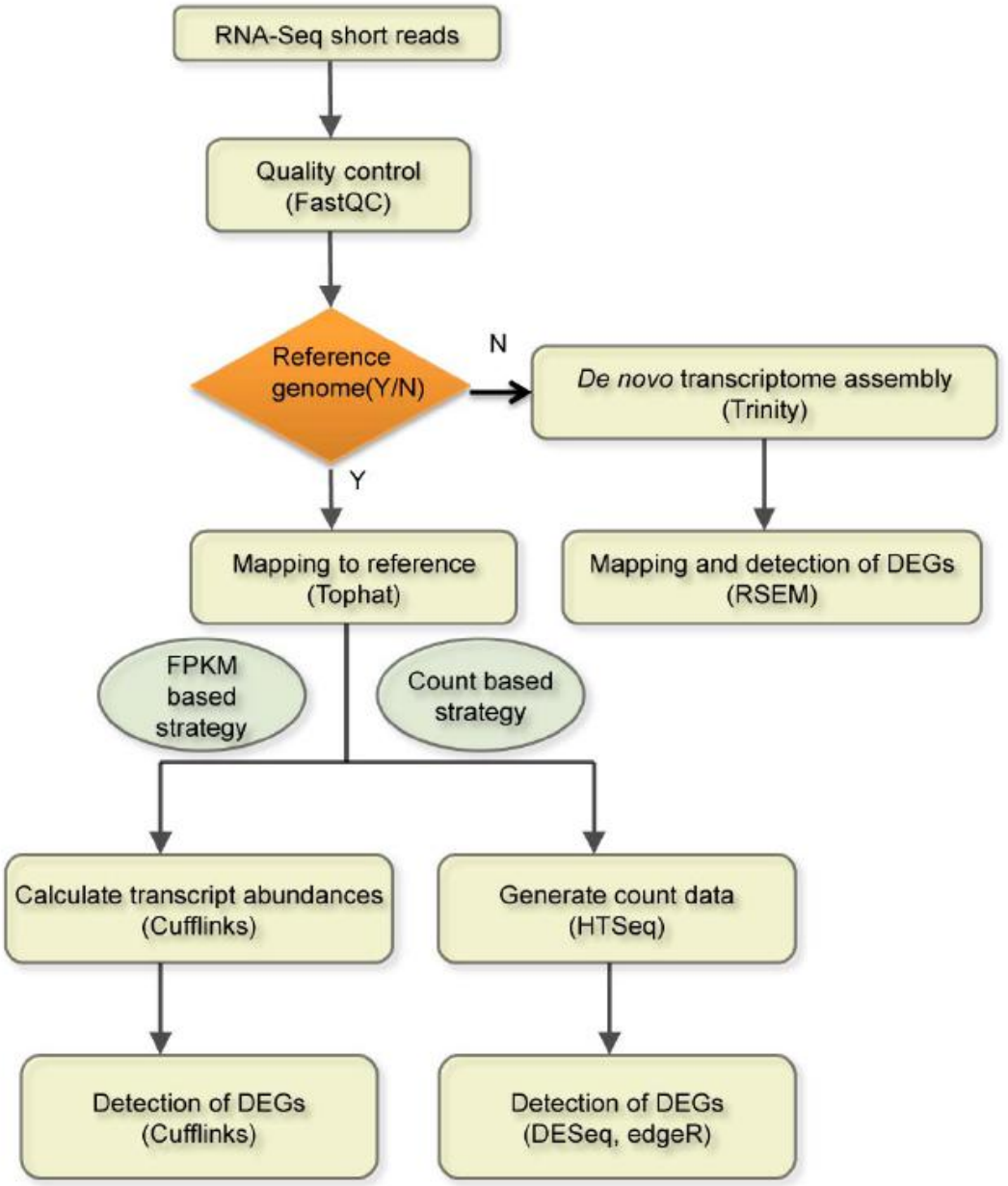
- Choix des outils dépendent des données disponibles et de la question biologique

- Bientôt : featureCount, cuffdiff

- Tous les outils sont dispo sur Migale et Genotoul

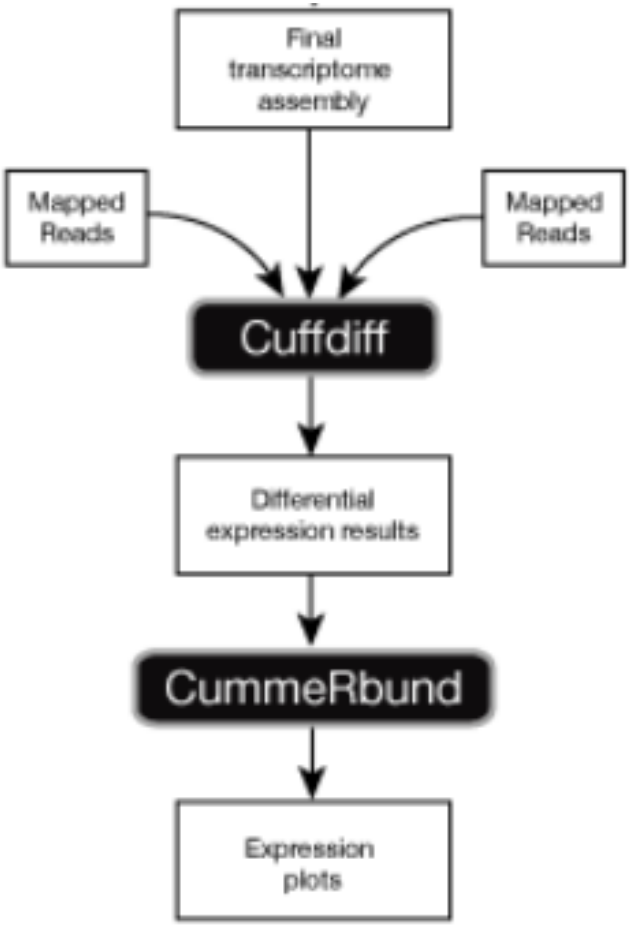
Conclusion générale

- Deux stratégies

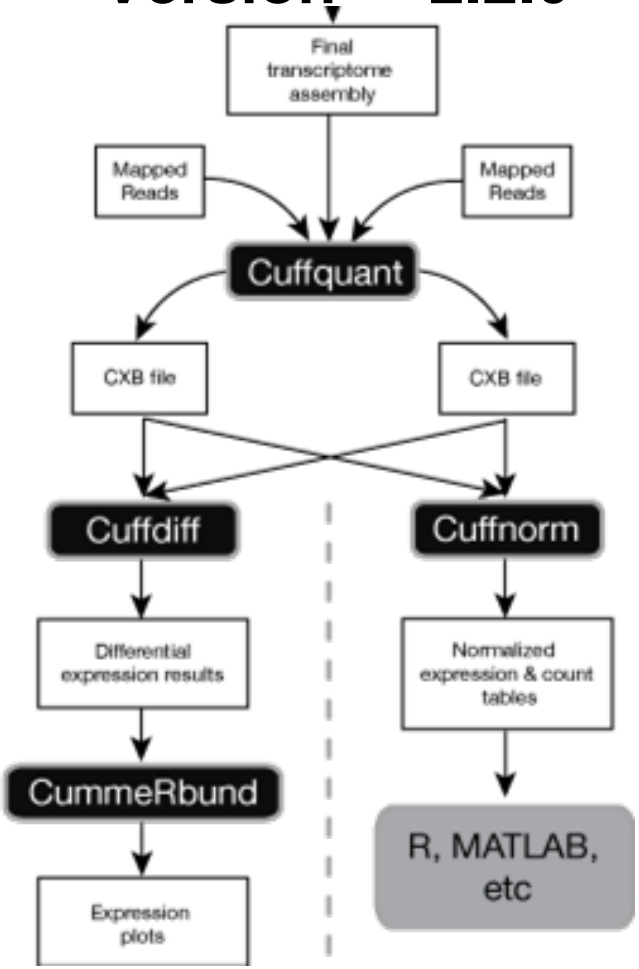


Stratégie Cufflinks

Version <2.2.0



Version >=2.2.0



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

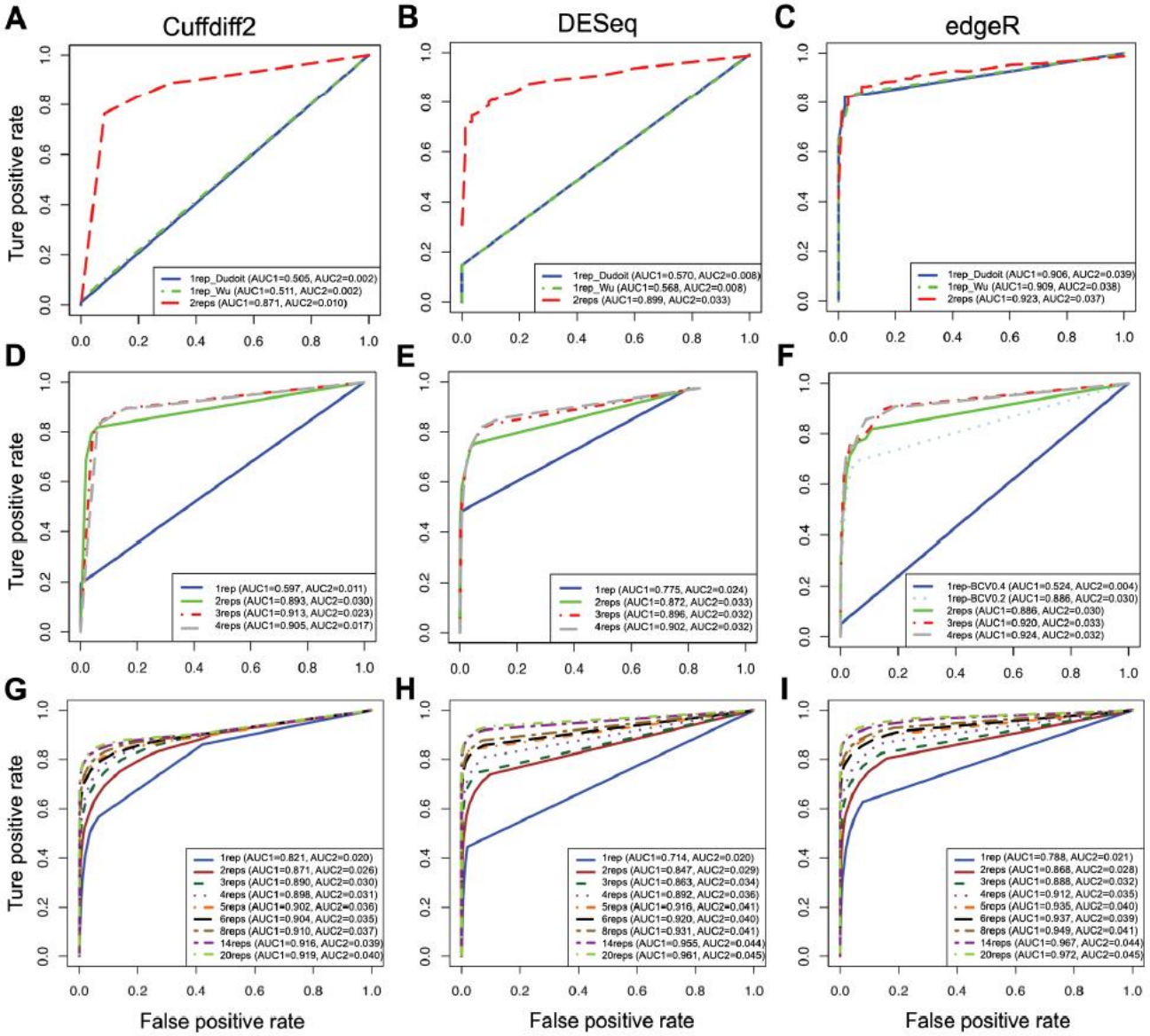
Stratégie comptage bruts

- Les étapes d'une étude stats:
  - Statistiques sur les données d'entrée (descriptives)
  - Normalisation
  - Suppression des données aberrantes
  - Analyse différentielle

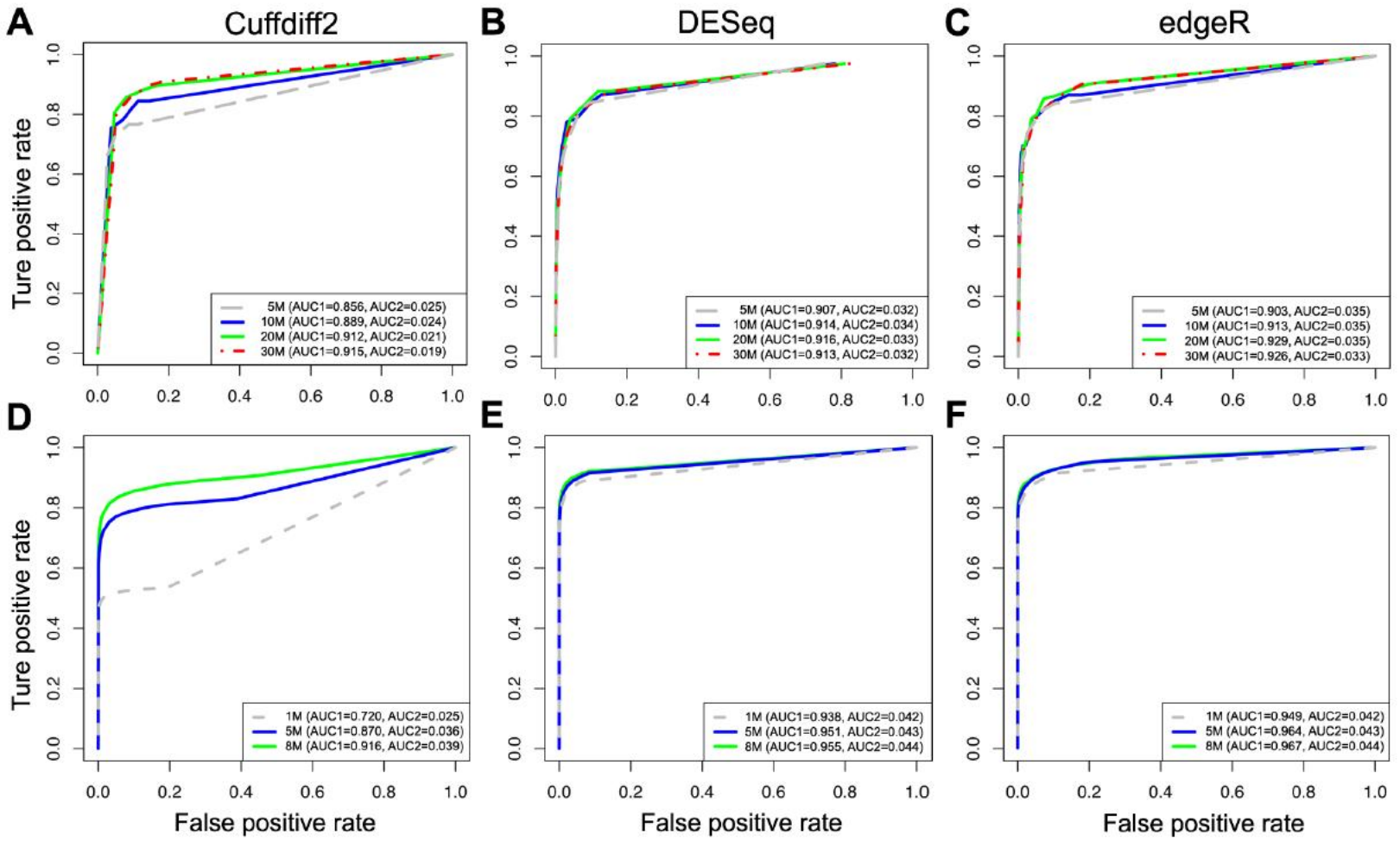
- Outils DEseq dans Galaxy , mais peu paramétrable

L'effet des réplicats

[Zhang et al. 2014]

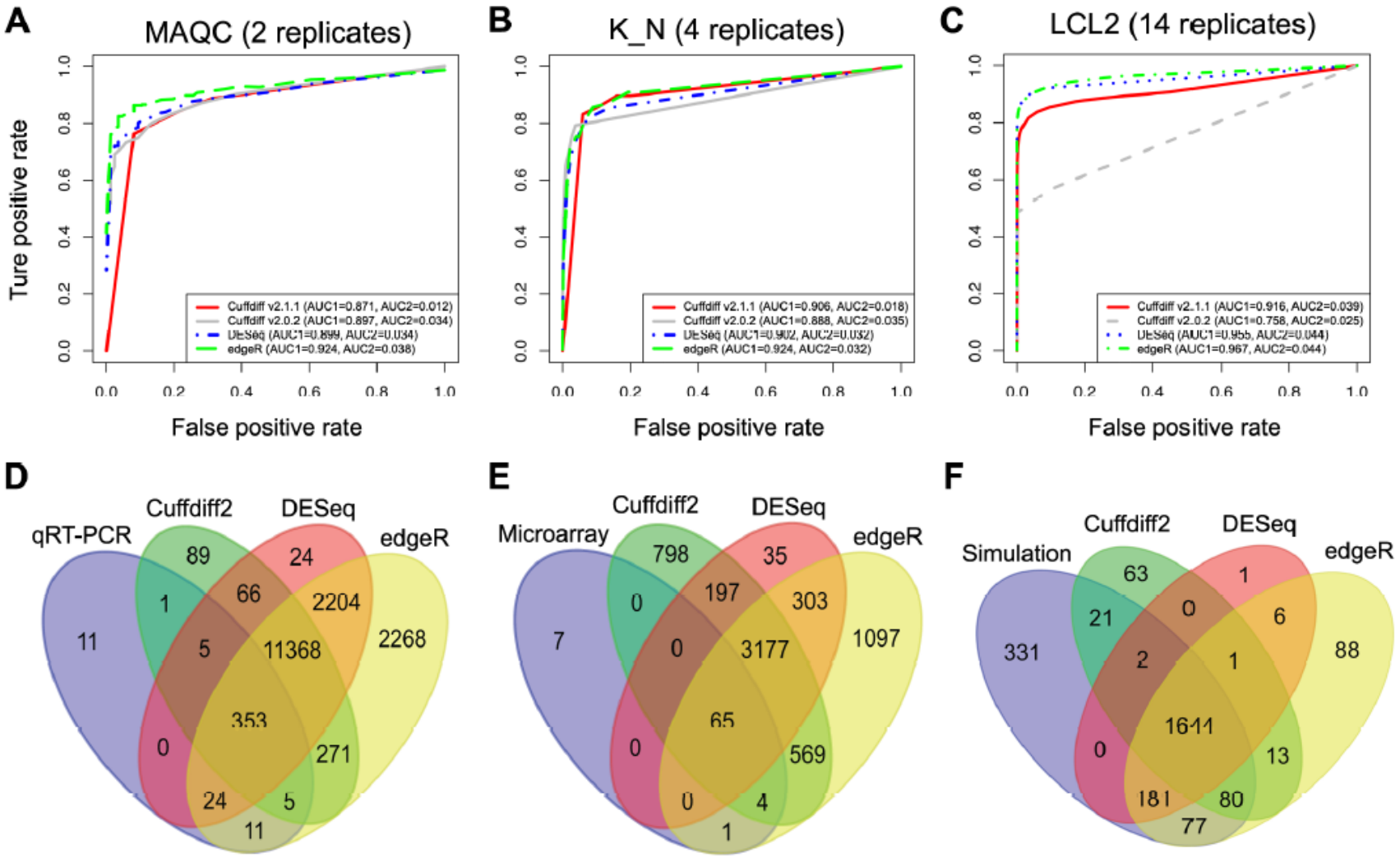


L'effet de la profondeur



Zhang et al. 2014

Recouvrement des méthodes



[Zhang et al. 2014]

Liens utiles

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

Discussion / Questions