TRAITEMENT BIOINFORMATIQUE DE DONNÉES RNA-Seq

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 ?

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

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

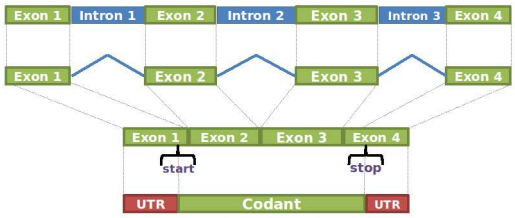
Rappels biologiques

Qu'est-ce qu'un transcrit?

Rappels biologiques

Qu'est-ce qu'un transcrit ?

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

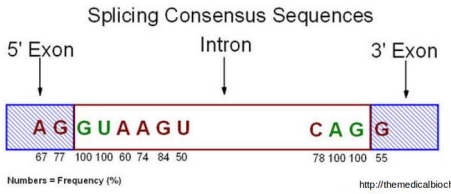


- **Transcrit :** portion d'ADN transcrite en molécule d'ARN
- **UTR :** région transcrite mais pas traduite

Rappels biologiques

Qu'est-ce qu'un site d'épissage?

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

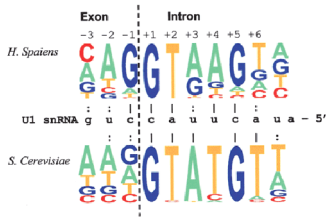


http://themedicalbiochemistrypage.org/rna.php

Rappels biologiques

Qu'est-ce qu'un site d'épissage?

- **Site** d'**épissage non-canonique :**
  - **GC**-**AG** ou **AT-AC** comme sites **donneurs** et **accepteurs**
- **Mammifère :**
  - 0.69% GC-AG
  - 0.05% AT-AC
- **Autre exemple :**



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

Rappels biologiques

Epissage alternatif et isoformes

- Excision d'exon

- Rétention d'intron

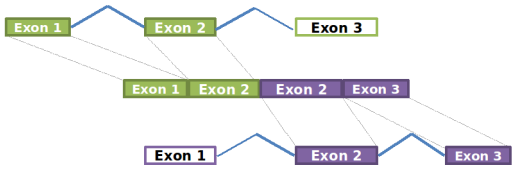
- TSS alternatif

- Exons exclusifs

Rappels biologiques

Et plus encore ?

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

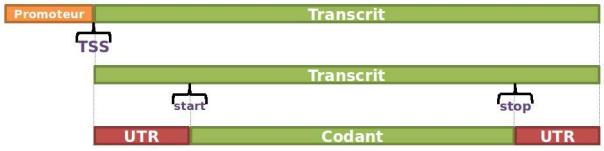


- **Chimère biologique**

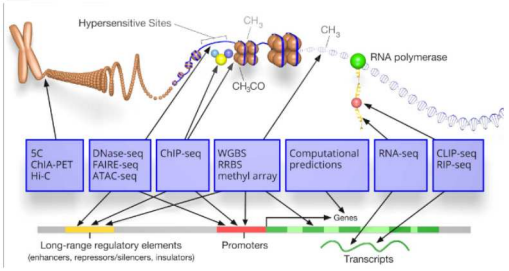
Rappels biologiques

Gène procaryote / gène eucaryote

- Pas d'intron chez les procaryotes



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

Les séquenceurs :

tinyurl.com/ngsspecs

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

http://tinyurl.com/ngsspecs

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 OPEN

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

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

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

Les données publiques

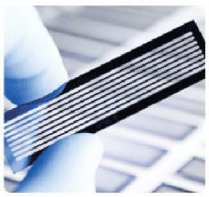
- Archive de short-read :
  - Données brutes de séquences
  - SRA https://www.ncbi.nlm.nih.gov/sra
  - ENA https://www.ebi.ac.uk/ena
- Gene Expression Omnibus
  - Données analysées (bed, peak, bigwig …),
  - Lien vers SRA
  - https://www.ncbi.nlm.nih.gov/geo/
- Expression Atlas
  - Interface d'exploration 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

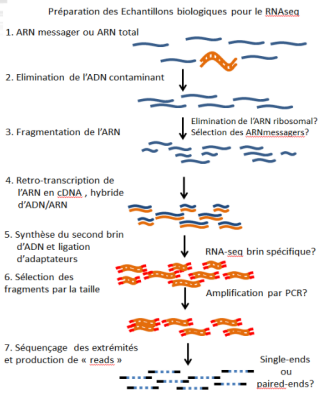
Flowcell : 1 plaque
(en général 1 run)



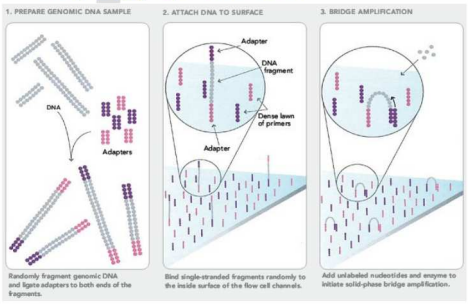
Lane : ligne de séquençage

- 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.

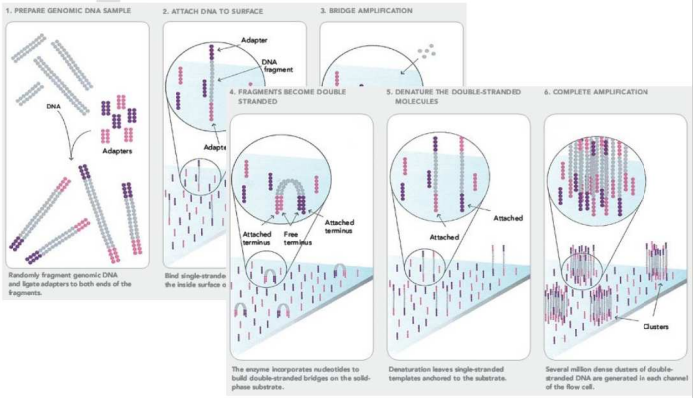
Le protocole RNAseq



Séquençage illumina



Séquençage illumina



Quels choix quand on fait du RNA-Seq ?

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

Déplétion / Enrichissement ?

❖Résultats semblables d'après :

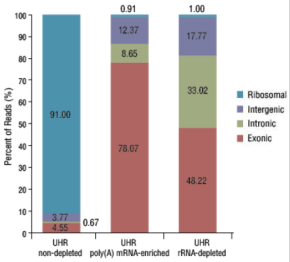
Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling, BMC Genomics , 2014

❖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



https://content.neb.com/products/e6310-nebnext-rrna-depletion-kit-human-mouse-rat

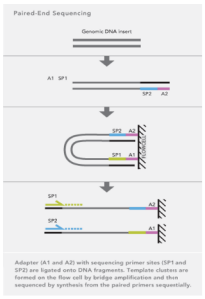
Paired-end

Protocole différent (Adaptateurs spécifiques)

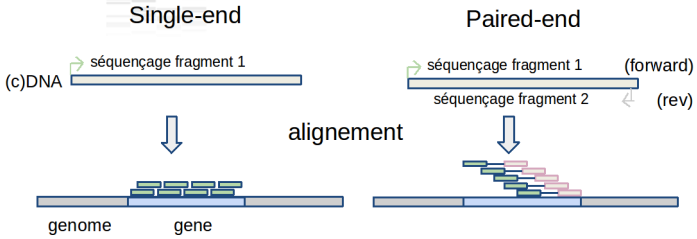
❖Améliore le mapping

❖Aide a la détection de variant alternatif

❖Plus généralement aide à la détection de : variation structurale de génome (insertion/délétion), CNV, réarrangement génomique



Single-end *vs* Paired-end



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

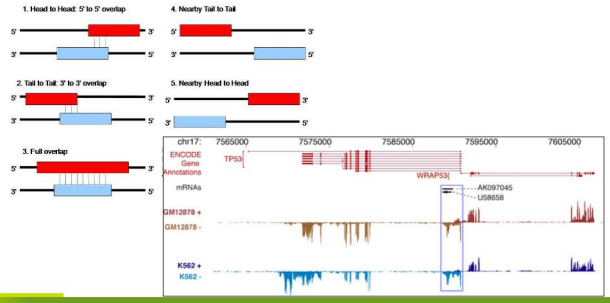
L'intérêt des librairies brin spécifique

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 ?

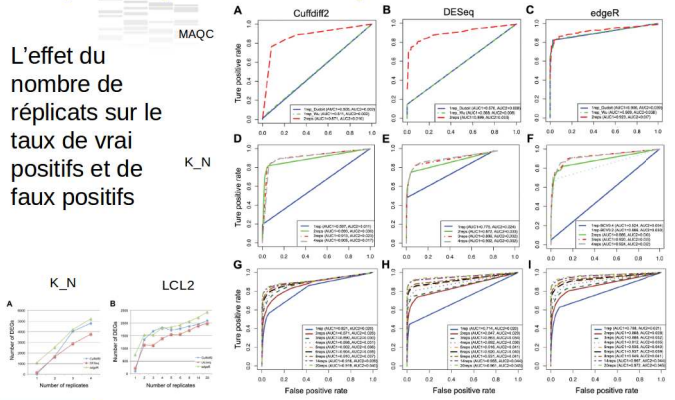
Pourquoi augmenter le nombre de répétitions biologiques ?

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)*

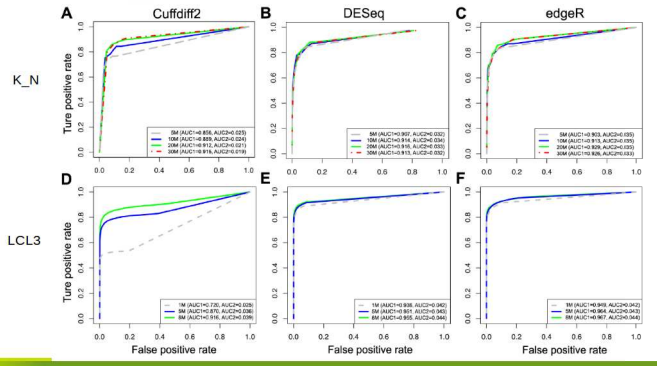
Equilibre profondeur / répétitions ?

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



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

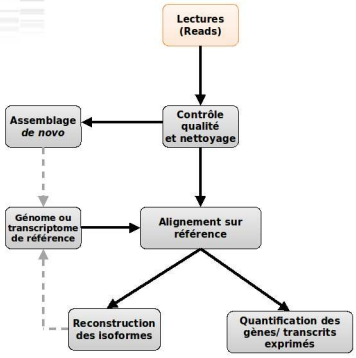
- **De novo :**
  - Pas de génome/transcriptome de référence
  - Outils en évolution permanente
  - Ressources (cpu/disque) +++
- **Transcriptome de reference**
  - Dépendant de la qualité de l'annotation structurale
  - Peu couteux
- **Génome de référence**
  - Permet une approche combinée :
    - sur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

Pipeline d'analyse RNA-Seq : avec référence

- **Contrôle qualité**
- **Pre-nettoyage** des lectures
  - **suppression** des **adaptateurs** de **séquençage**
  - (**suppression** des **adaptateurs** de **multiplexage**)
- **Nettoyage** des lectures
  - **tronquer** les **extrémités** de **mauvaise qualité** des lectures
- **Alignement** des **lectures sur** la **référence**
  - gènes ou génome complet
- **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
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- 1 ère ligne = identifiant de la séquence

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

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

Format Fastq

- 4ème ligne = Qualité

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

- Appelée aussi Phred quality score (Sanger format)

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

Probabilité qu'une base soit incorrecte

Format Fastq

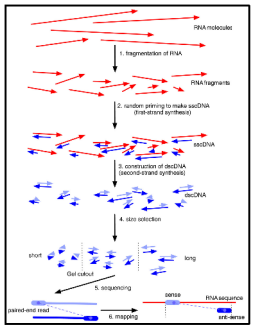
- Qualité Encodée en ASCII

Biais spécifiques au RNA-Seq

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

Préparation de la banque

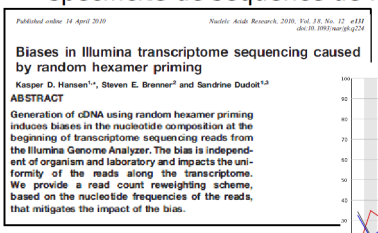
- **Extraction ARN total**
- **Déplétion** (queue polyA)
- **Fragmentation**, **reverse transcription** avec des **hexamères aléatoires** -> **dscDNA**
- **Séquençage**



Roberts et al. Genome Biology 2011, 12:R22

Biais : *random hexamer priming*

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

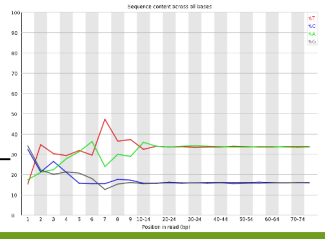

Published online 14 April 2010 Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

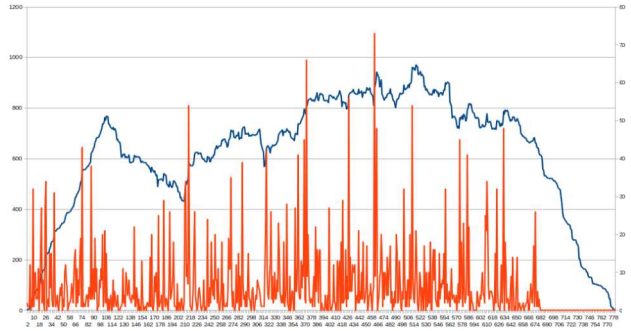
Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

ABSTRACT

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



Biais : *random hexamer priming*



Orange = reads start sites
Blue = coverage

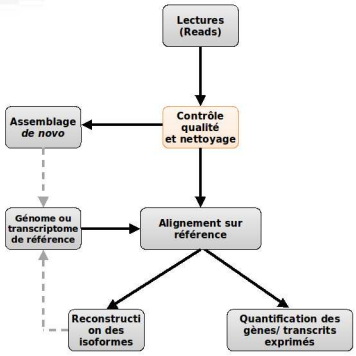
Biais : longueur des transcrits

- La capacité, en utilisant des **comptages** obtenus par **RNA-Seq**, a observer un transcrit comme étant **différentiellement exprimé** est **directement reliée** à sa **longueur**.
- Pour un **même gène** ayant **deux isoformes**, l'une faisant la moitié de l'autre, exprimé en **même abondance dans deux conditions différentes** :
  - L'isoforme la plus courte sera deux fois moins « comptée » que la plus longue

Biais : « mappabilité »

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

Workflow d'analyse RNA-Seq



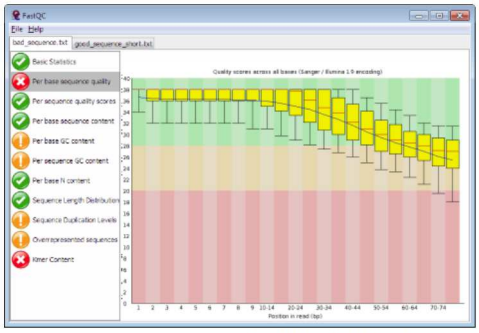
Contrôle qualité

Objectifs :

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

Contrôle qualité avec FastQC

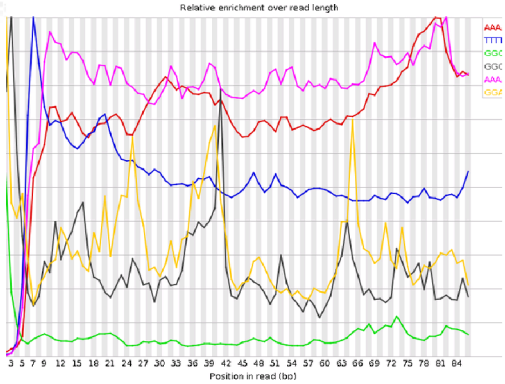
- orienté DNA-Seq



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

Contrôle qualité

Kmer content



Contrôle qualité

Over-expressed sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	64433	0.20451959970239228	No Hit

→ Détecter les adaptateurs

Nettoyage des données

Nettoyage « optionnel »

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

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

Nettoyage des données

Principe de Sickle :

- Traite les paires ensemble
  - Fenêtre glissante : 10% de la taille des reads
  - Calcul de la qualité moyenne des lectures

exemple : Longueur = 23

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

Travaux pratiques

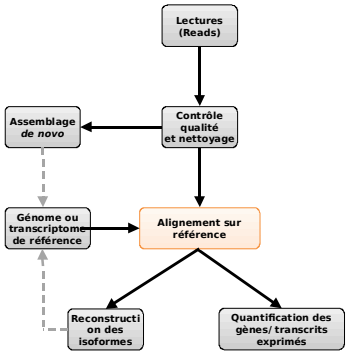
Présentation des objectifs

- **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

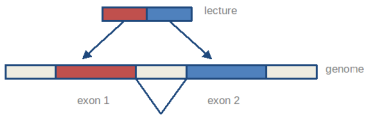


Alignement épissé

Objectifs :

- **Aligner** les **lectures** issues du séquençage de **dscDNA** (transcrits) sur le **génome**, en tenant compte de l'**épissage alternatif**
- Être capable d'**exploiter** les listes des **jonctions exons-exons connues**, mais également d'en **détecter** de **nouvelles**
- Tout cela dans un **temps raisonnable**...

Introduction



Définition

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

- **Prédiction :** chaque outil propose un/plusieurs locus.
- **Locus :** le résultat est un ensemble de positions génomiques (ex.: chr1:100..150)

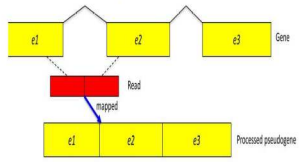
- Mapping ARN ≠ Mapping ADN

- Mapping ≠ Alignement

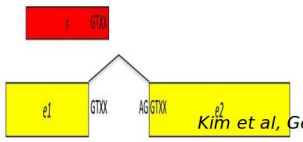
Les outils de mapping font de mauvais alignements (sauf aux jonctions).

Cas difficiles

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



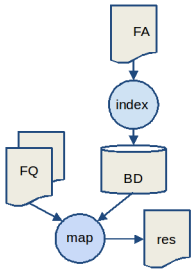
- Fin de la lecture sur un exon propre
- Lecture sur une jonction non-connue d'un gène peu exprimé



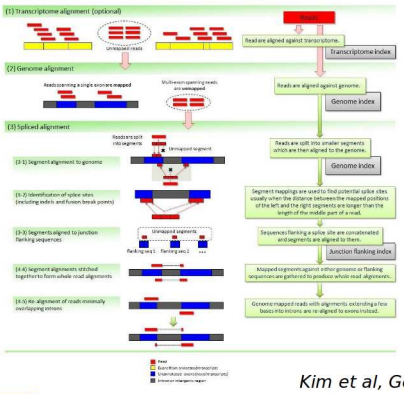
Kim et al, Genome Biology, 2013

Étapes de mapping

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



Tophat2

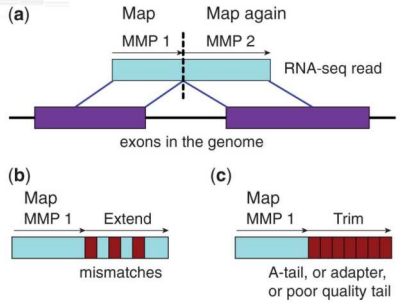


Tophat2 est constitué de beaucoup d'étape pour résoudre chaque cas difficile.

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

Kim et al, Genome Biology, 2013

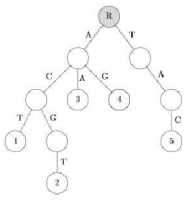
STAR is an ultrafast universal RNA-seq aligner



Dobin et al, Bioinformatics, 2011

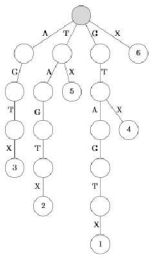
STAR is an ultrafast universal RNA-seq aligner

Aligneurs index BWT (BWA, Bowtie, SOAP)



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

STAR



GTAGT.

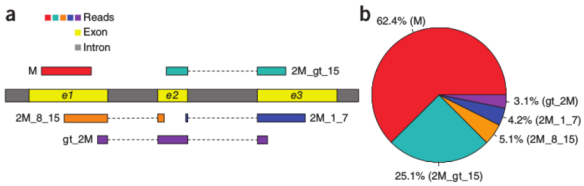
STAR is an ultrafast universal RNA-seq aligner

- 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

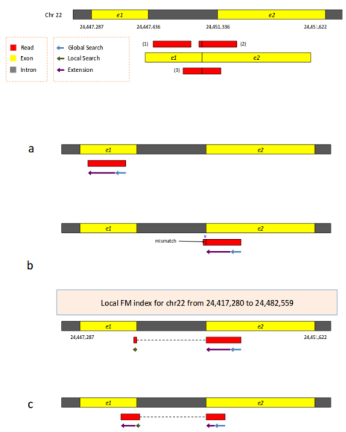
- **"We recommend that the HISAT and TopHat2 users switch to HISAT2."**
- **2 FM index : tout genome + regions de 64kb**



Kim et al, Nature, 2015

HiSAT2

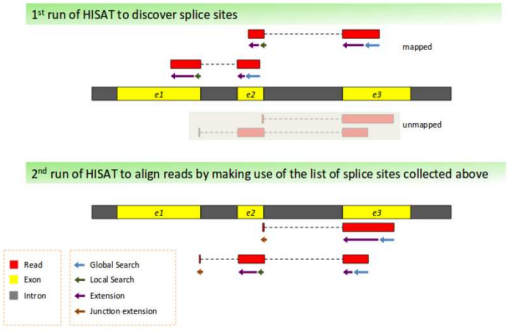
- 3 types reads
- a) en plein
- b) épissé avec petite region
- c) épissé 2 grandes régions



Kim et al, Nature, 2015

HiSAT2

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



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

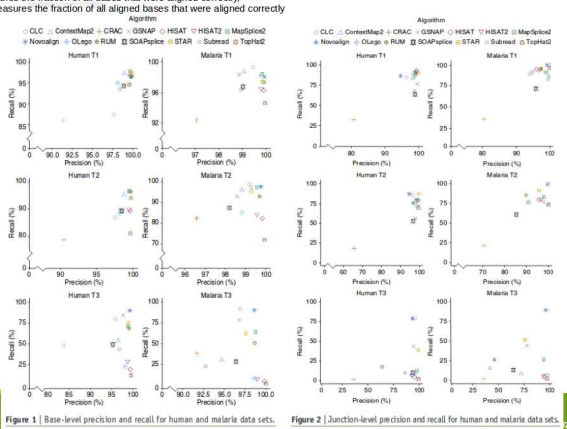
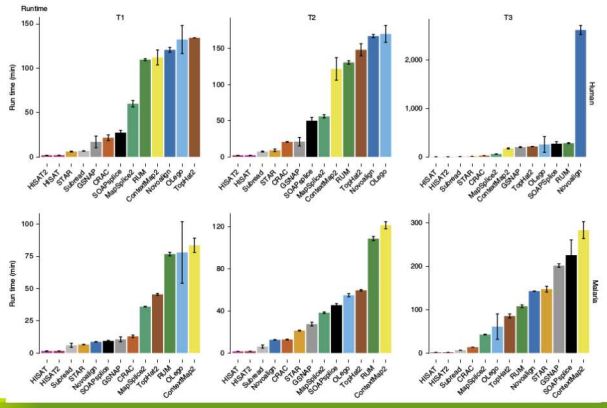


Figure 1 | Base-level precision and recall for human and malaria data sets.

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

Baruzzo et al. Nature Method.

Benchmark of RNAseq aligners



Baruzzo et al, Nature Method.

Alignement : données initiales

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

Format GTF : Gene Transfert Format

- **Dérivé** du format généraliste GFF (General Feature Format)
- Contient l'**annotation structurale** du **génome** (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 **1ère 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

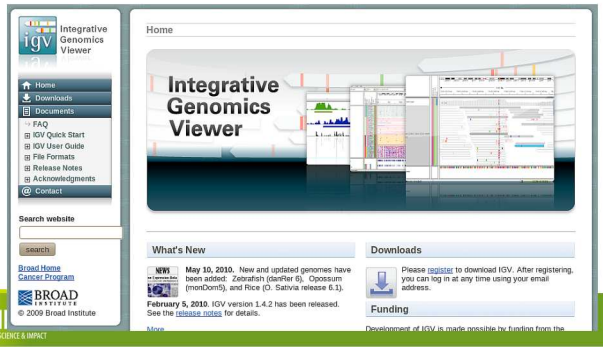
Quelles informations doivent être stockées dans un fichier d’alignement SAM ?

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

TP: lancer l’alignement

Visualisation des alignements avec IGV

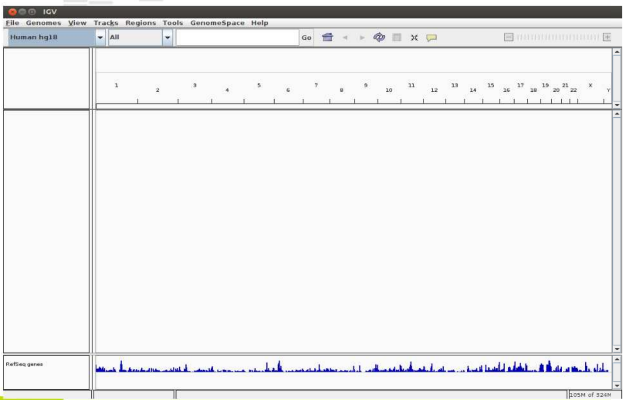
- IGV : Integrative Genomics Viewer
- Website : http://www.broadinstitute.org/igv



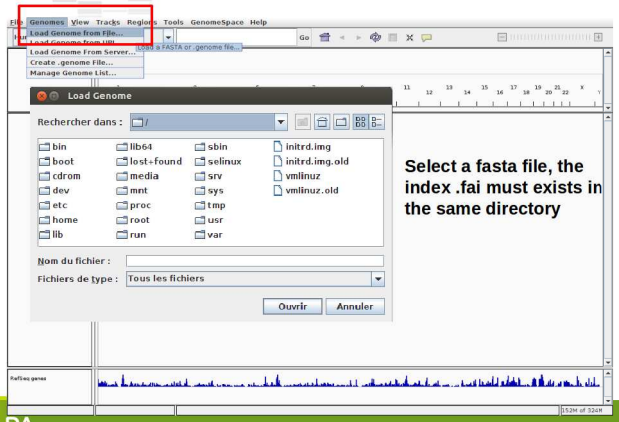
Visualisation des alignements avec IGV

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

Visualisation des alignements avec IGV

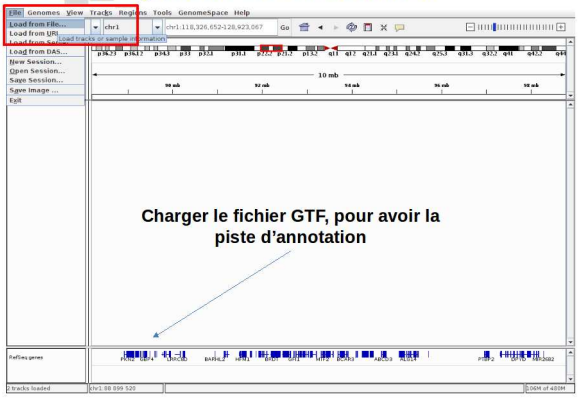


IGV : Chargement de la référence



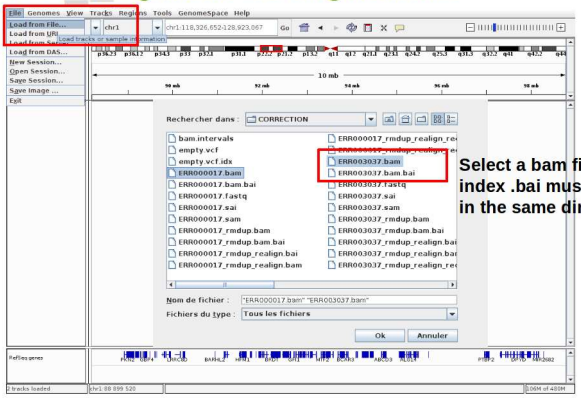
Select a fasta file, the index .fai must exists in the same directory

IGV : Chargement de l'annotation



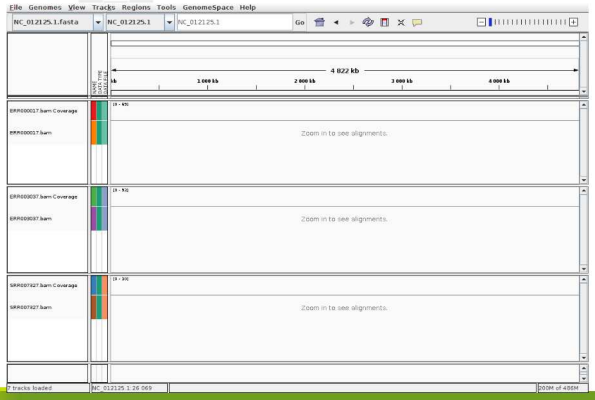
Charger le fichier GTF, pour avoir la piste d'annotation

IGV : Chargement des alignements

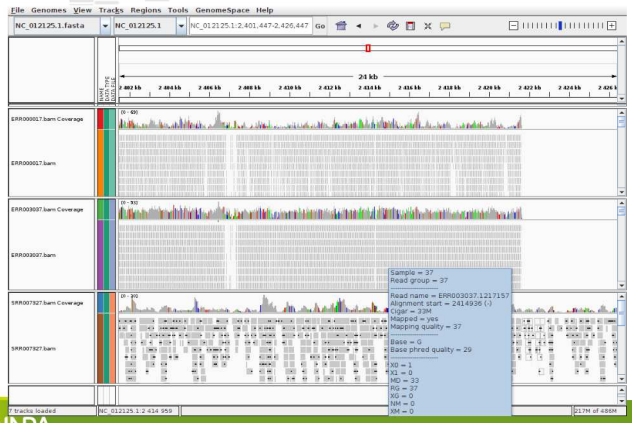


Select a bam file, the index .bai must exists in the same directory

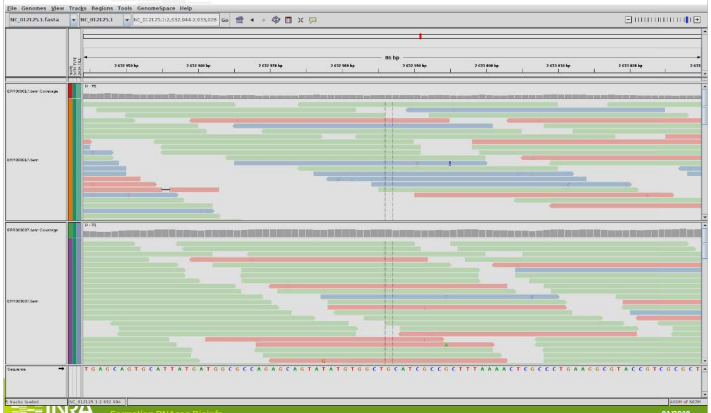
IGV : Chargement des alignements



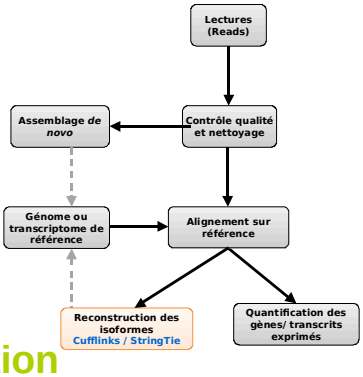
IGV : Chargement des alignements



IGV : Chargement des alignements

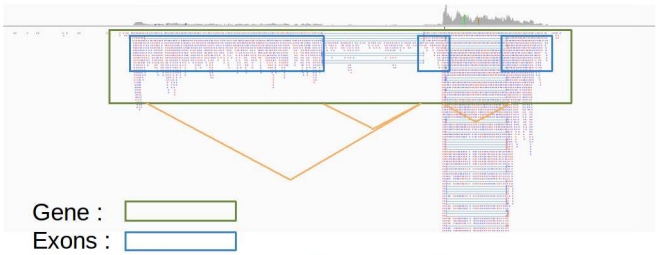


TP – Visualisation



_05 Reconstruction de transcrit

Modélisation

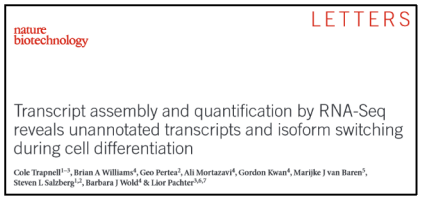


Gene :
Exons :
Jonctions (dans les paires & les Reads)

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)



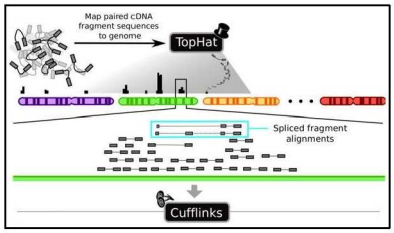
http://cufflinks.cbcb.umd.edu/

Cufflinks

Reconstruction de **transcrits**

❖ Fragments divisés en ***loci*** **non chevauchants**

❖ Chaque ***locus*** est **assemblé indépendamment**



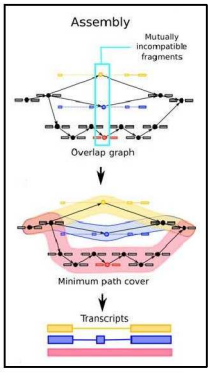
Trapnell et al. Nat Biotechnol. 2010

Cufflinks

Reconstruction de **transcrits**

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

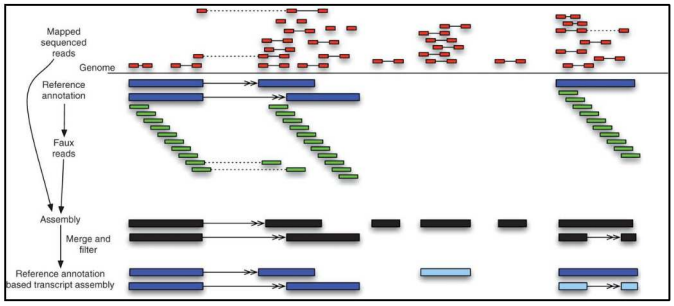
❖ **Stratégie** de **construction** du **modèle** :
- trouver le **nombre minimum de modèles qui expliquent les lectures** :
  - **minimum de chemins**
  - **Nb de lectures incompatibles** = **nb minimum** de **transcrits** nécessaires
  - **1 chemin = 1 isoforme**



Trapnell et al. Nat Biotechnol. 2010

Cufflinks

Reference Annotation Based Transcripts Assembly



Roberts et al. Bioinformatics 2011

Cufflinks

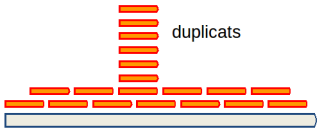
- Reference fasta (génome)
- Référence gtf (transcriptome)
- 1 bam par échantillon
- Quelles sont les stratégies possibles pour identifier le **maximum** de transcrits ?

Fusion d'aligements

- Samtools : suite logicielle permettant la manipulation de fichiers SAM/BAM/CRAM
- **Samtools view** : visualisation / conversion
- **Samtools merge** : fusion de fichiers d'aligement
- 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.

Cufflinks

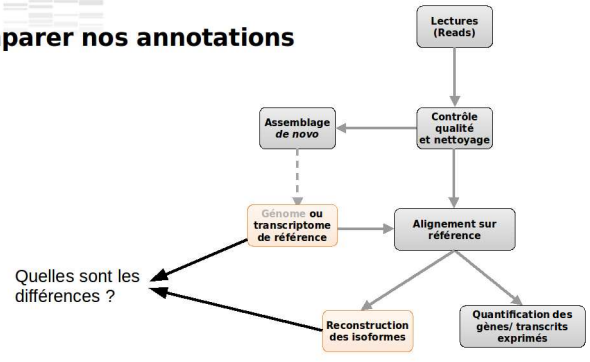
Reconstruction des **transcrits**

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

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

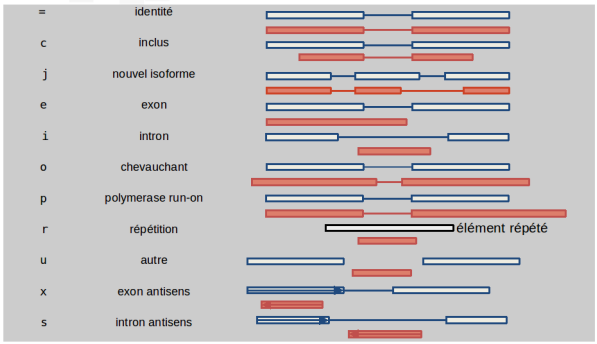
Cufflinks - Cuffcompare

Comparer nos annotations



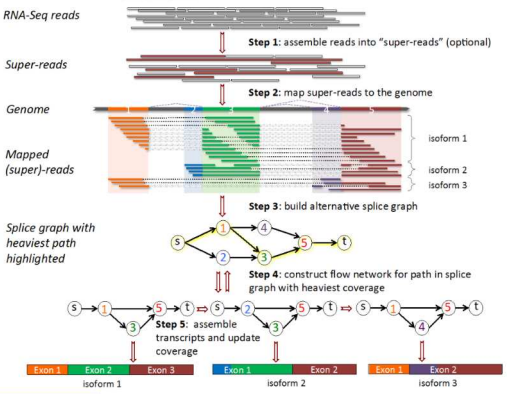
Cufflinks - Cuffcompare

Class code de cuffcompare



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

StringTie

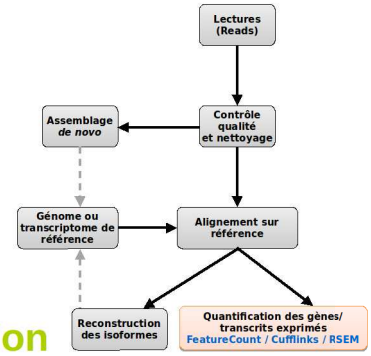


Pertea et al. Nature Biotechnology 2015

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

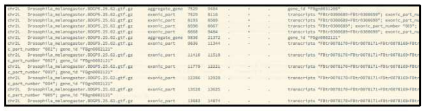


_06 Quantification

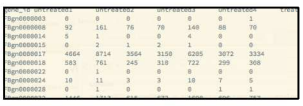
Quantification

Que cherche-t-on à compter ?

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



- Comptage brut sur les gènes ou les exons ou les transcrits:
  - **featureCount**
- Estimation de l'abondance des transcrits reconstruits :
  - **Cufflinks, RSEM**
- Comptage brut des multimap
  - **mmquant**

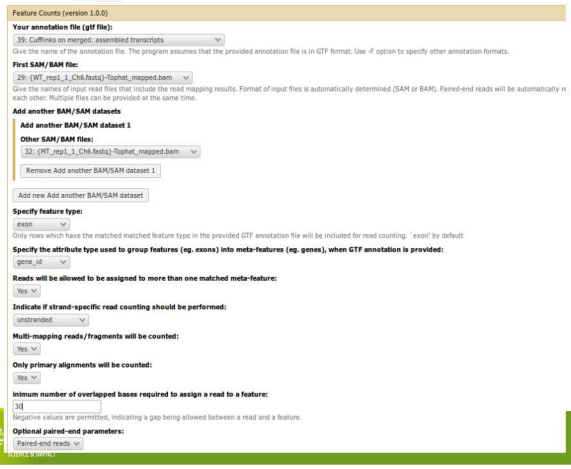


- Dépend des données disponibles

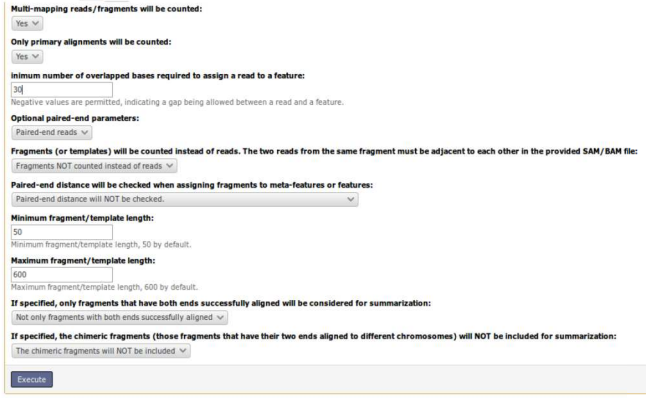
featureCounts

- 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



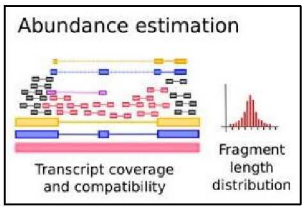
featureCounts : options



Cufflinks

Principes

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



Trapnell et al. Nat Biotechnol. 2010

Cufflinks

RPKM / FPKM

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

❖**RPKM** :
Reads **P**er **K**ilobase of exon per **M**illion fragments mapped :
R = Nombre de read mappés
N = Nombre total de read de la librairie
L = taille des exons du gène en bp

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

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

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

Mortazavi et al. Nature Methods 2008

RSEM

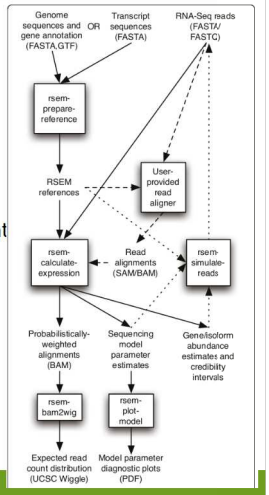
❖ Logiciel qui fait :
- l'alignement
- l'estimation des isoformes

❖ 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

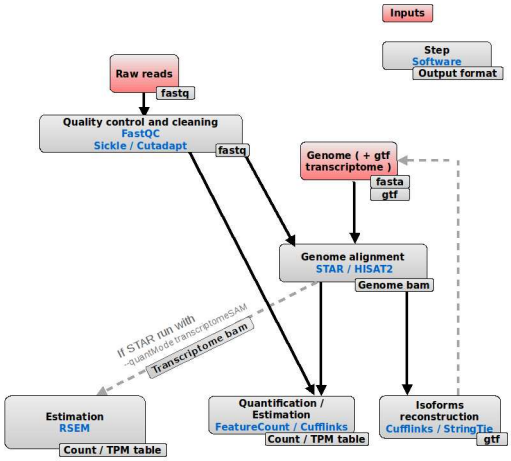
Conclusion générale

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

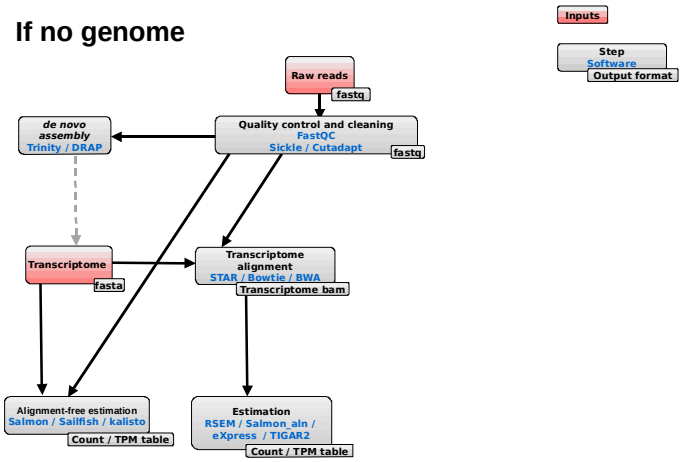
Tous les outils sont dispo sur Migale et Galaxy

- ❖ Et maintenant en avant pour les stats !

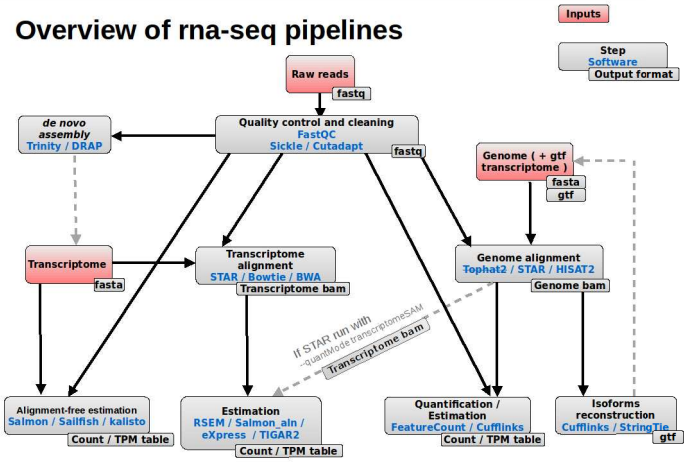
What we did



If no genome



Overview of rna-seq pipelines



Liens utiles

- **Seqanswer** : *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