Traitement bioinformatique de données RNA-Seq sous Galaxy

23 et 24 mai 2022
Plateforme Migale
INRAE - Jouy-en-Josas

cyprien.guerin @inrae.fr

valentin.loux @inrae.fr

Tour de table

- **Vous ? Votre labo ? Vos données ?**
- Avez-vous **déjà utilisé Galaxy ?**
- Avez-vous **déjà traité** des données de type **NGS ?**
- **Vos attentes ?**

Qu’allez-vous apprendre ?

- A l’issue de cette journée de **formation vous saurez :**
  - Utiliser un environnement **Galaxy** pour **quantifier** et **découvrir** de **nouveaux transcrits**.

- **Vous ne saurez pas :**
  - Etudier un transcriptome **sans génome de référence**,
  - Traiter bioinformatiquement de **nombreuses librairies**.

Programme de la formation

- **Lundi**
  - Introduction au RNA-Seq (Biologie et Protocole)
  - Vérification de la qualité
  - Algorithmes d'alignement
  - Visualisation
  - Assemblage de transcrits

- **Mardi matin**
  - Quantification
  - Création et utilisation d'un *workflow* d'analyse
  - Discussion

Un peu de vocabulaire

- **Transcriptome :** Ensemble des transcrits d'un organisme
- **RNA-Seq de novo :** Etude du transcriptome sans génome de référence
- **Read :** Lecture

Rappels biologiques

Qu'est-ce qu'un gène ?

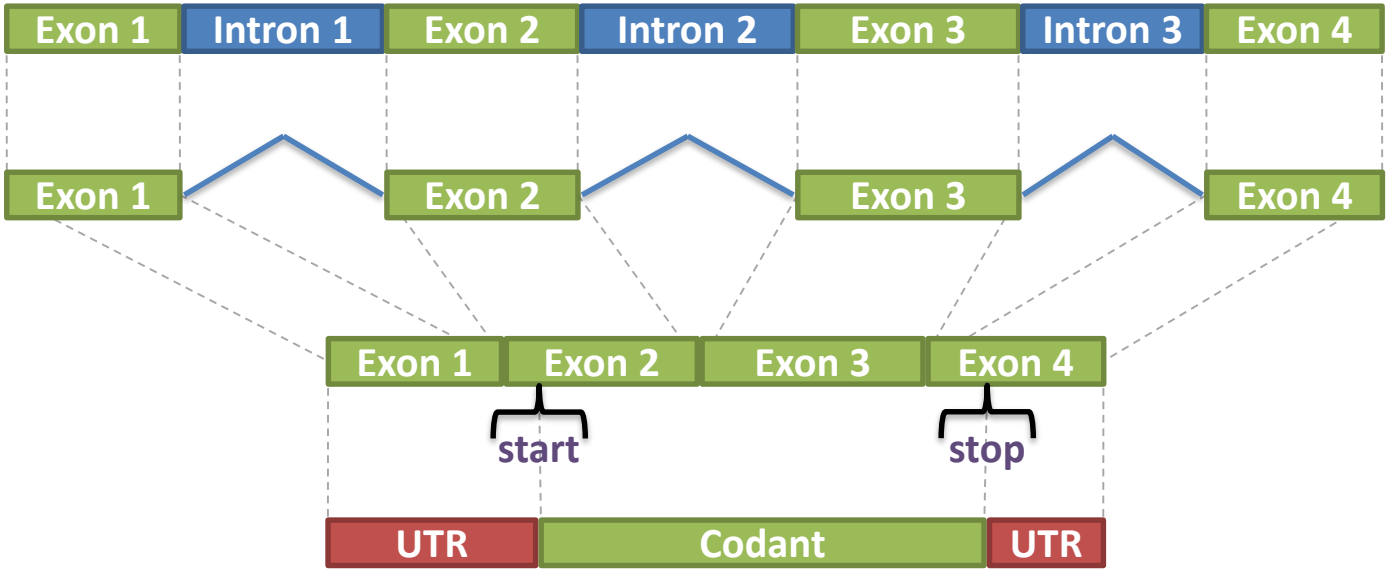
- **Gène :** unité fonctionnelle de l'ADN qui contient les instructions nécessaires à la création d'un produit fonctionnel

- **Promoteur :** zone de fixation de l'ARN-polymérase
- **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

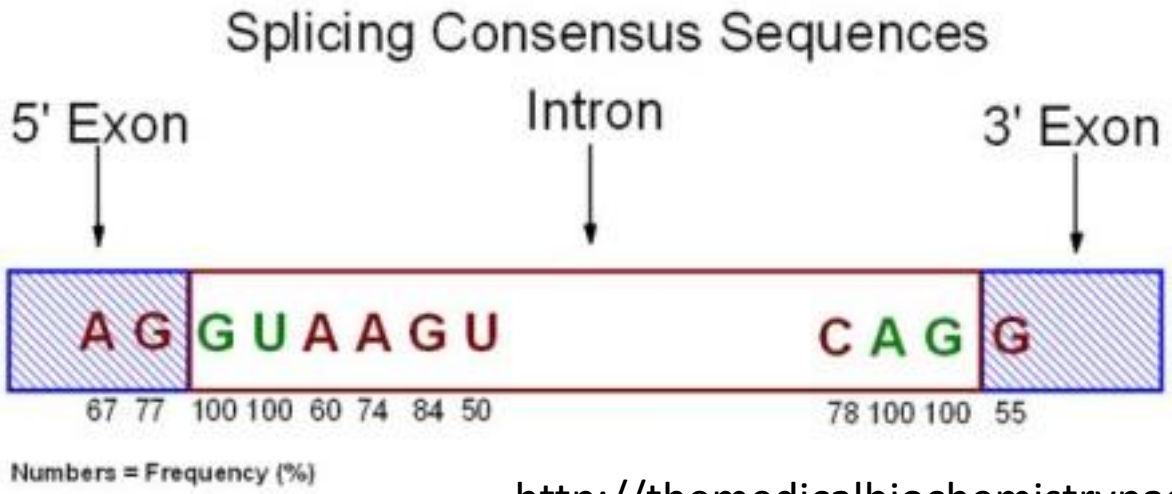


- **Transcrit :** portion d'ADN transcrite en molécule d'ARN
- **UTR :** région transcrite mais pas traduite

Rappels biologiques

Qu'est-ce qu'un site d'épissage?

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



http://themedicalbiochemistrypage.org/rna.php

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

Rappels biologiques

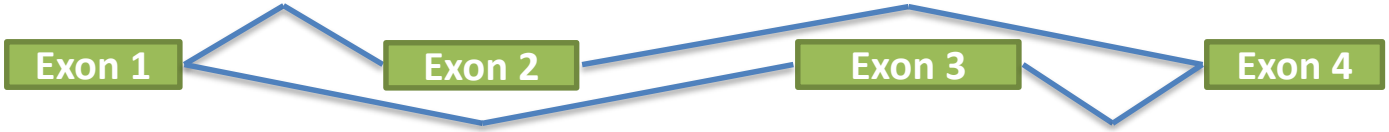
Epissage alternatif et isoformes

- Excision d'exon

- Rétention d'intron

- TSS alternatif

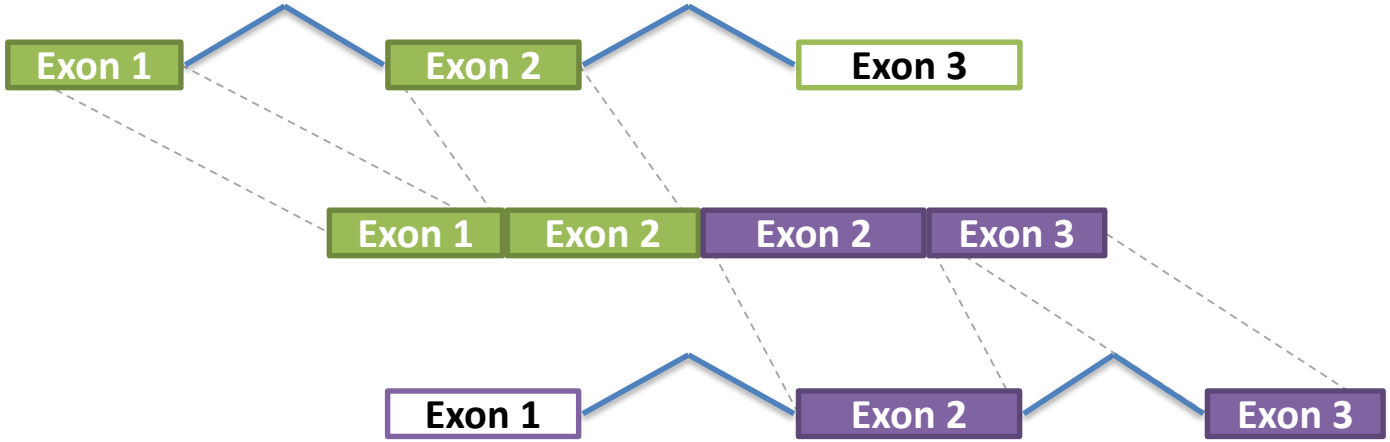
- Exons exclusifs



Rappels biologiques

Et plus encore ?

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

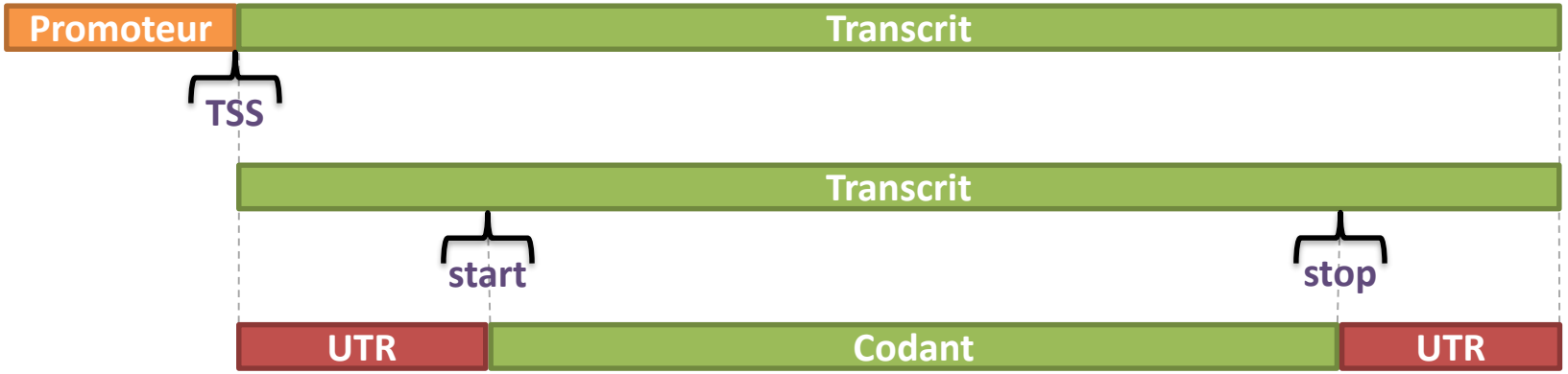


- **Chimère biologique**

Rappels biologiques

Gène procaryote / gène eucaryote

- Pas d'intron chez les procaryotes



- Opérons

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 : séquençage de l'ensemble des transcrits</u>**
  - (+) séquence du génome pas forcément nécessaire, **séquençage** de l'ensemble des transcrits **sans *apriori***
  - (-) **quantitatification moins fine**, encore en développement

Une expérience de RNA-seq de A à Z

Bio

Préparation des Echantillons biologiques pour le RNAseq

1. ARN messager ou ARN total

2. Elimination de l'ADN contaminant

3. Fragmentation de l'ARN

Elimination de l'ARN ribosomal?
Sélection des ARNmessagers?

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

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

RNA-seq brin spécifique?

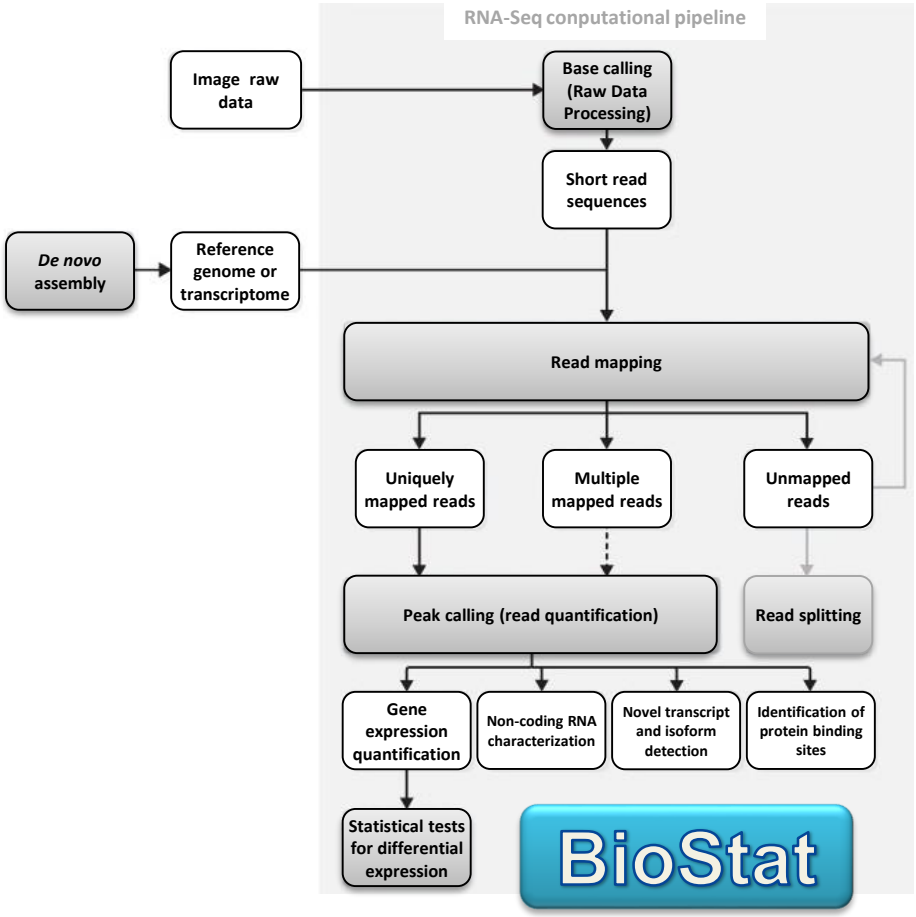
6. Sélection des fragments par la taille

Amplification par PCR?

7. Séquençage des extrémités et production de « reads »

Single-ends ou paired-ends?

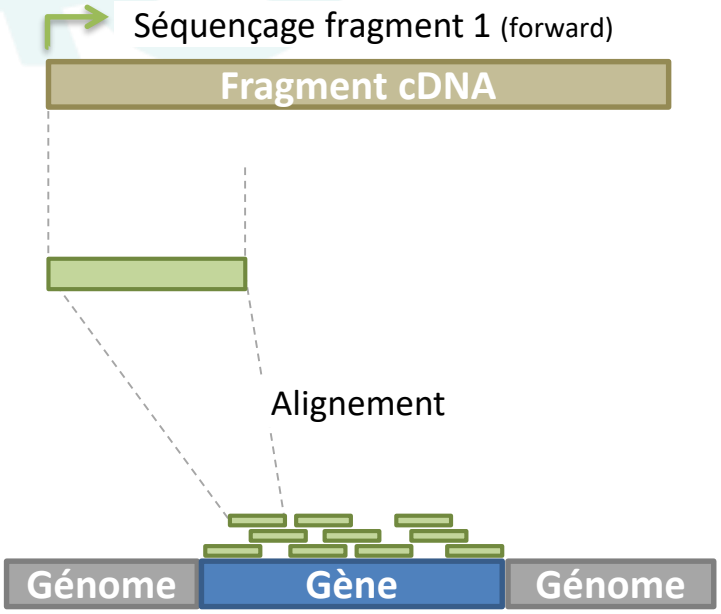
BioInfo



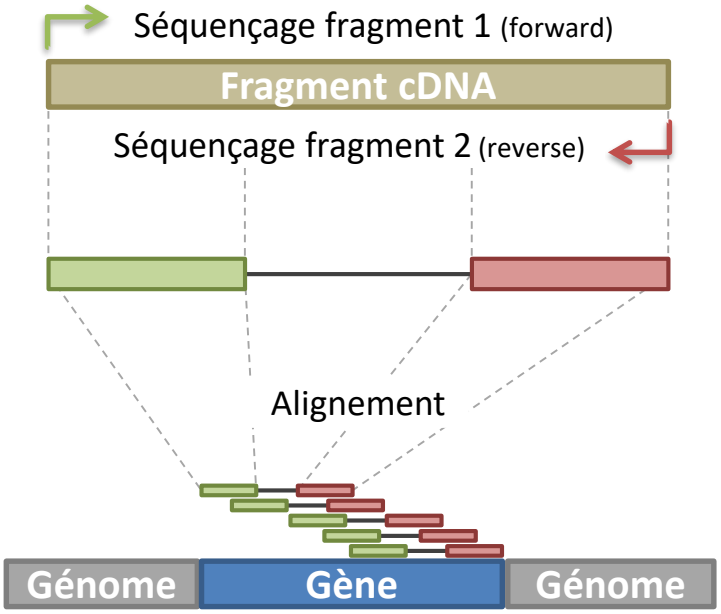
Mutz 2013, Current Opinion in Biotechnology

Lectures simples ou appariées

Single-end

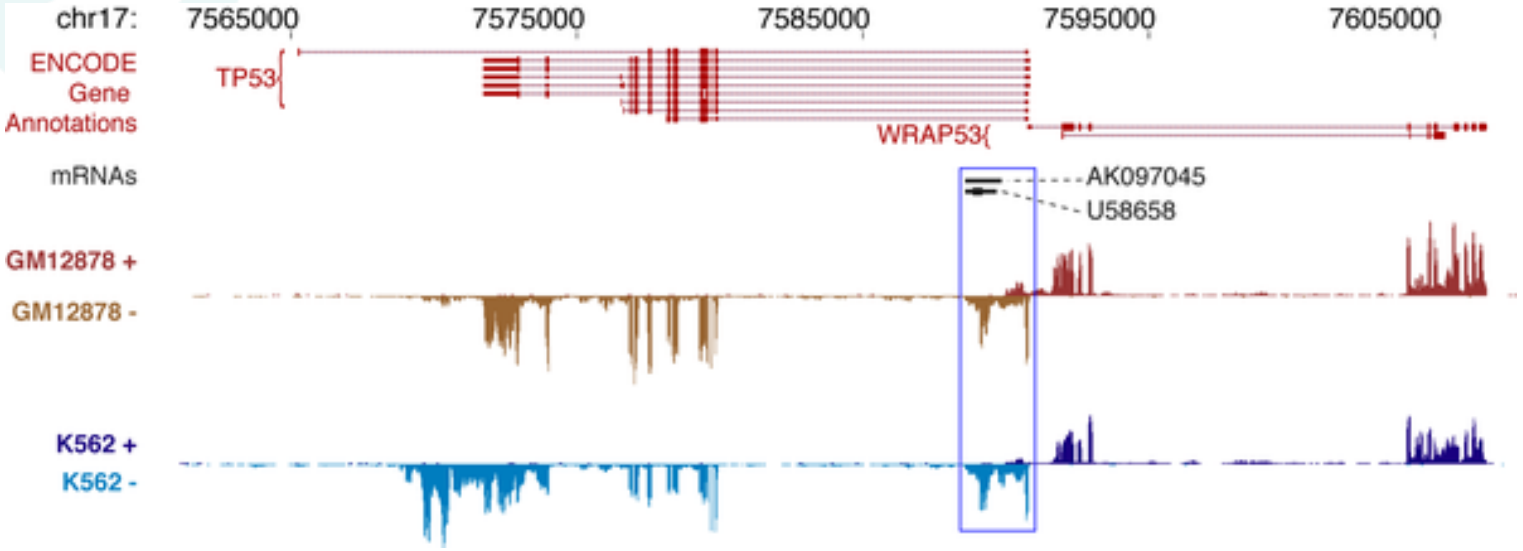


Paired-end



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

Intérêt des librairies brin-spécifique



Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]

NATURE METHODS | VOL.7 NO.9 | SEPTEMBER 2010 | 709

A quelles questions biologiques PEUT répondre le RNA-seq ?

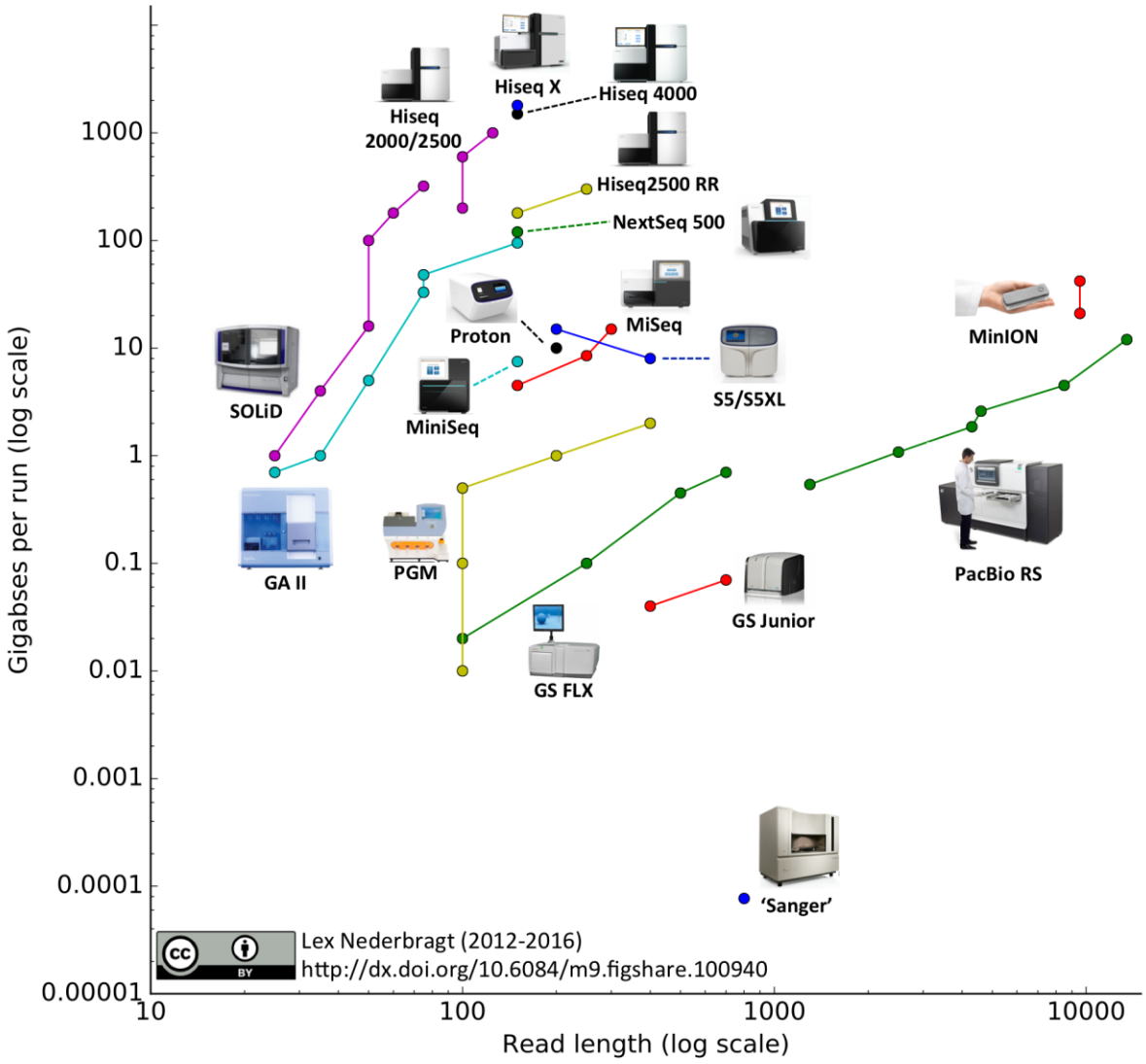
- L'**analyse d'expression différentielle** (différence d'expression) au niveau du transcriptome
- L'étude de l'**épissage alternatif** (isoformes) et recherche de **nouveaux transcrits**
  - amélioration des annotations structurales existantes
- La recherche d'**allèles spécifiques** et la **quantification**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

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

- Etude 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
  - découverte des petits ARN et leurs niveaux d'expression
  - étude des gènes faiblement exprimés

Les séquenceurs de 2ème et 3ème génération



Les séquenceurs 3ème génération

- **PacBio**, **Oxford Nanopore**, …
- ***Single molecule*** (quantification)
- **Lectures** (très) **longues** (pas d'assemblage)
- **Faible débit** (pour l'instant)
- **Grand nombre d'erreurs** (pour l'instant)

A single-molecule long-read survey of the human transcriptome

Donald Sharon[1–3], Hagen Tilgner[1,3], Fabian Grubert[1] & Michael Snyder[1]

nature biotechnology

Highly parallel direct RNA sequencing on an array of nanopores

Daniel R Garalde[1], Elizabeth A Snell[1], Daniel Jachimowicz[1], Botond Sipos[1], Joseph H Lloyd[1], Mark Bruce[1], Nadia Pantic[1], Tigist Admassu[1], Phillip James[1], Anthony Warland[1], Michael Jordan[1], Jonah Ciccone[1], Sabrina Serra[1], Jemma Keenan[1], Samuel Martin[1], Luke McNeill[1], E Jayne Wallace[1], Lakmal Jayasinghe[1], Chris Wright[1], Javier Blasco[1], Stephen Young[1], Denise Brocklebank[1], Sissel Juul[2], James Clarke[1], Andrew J Heron[1] & Daniel J Turner[1]

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 ?

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

- Le **RNA-Seq** n'est **pas** complètement **mature !**

- Quelques chiffres :
  - **100M de lectures** sont suffisantes pour détecter **90 % des transcrits** de **81 % des gènes** du **transcriptome humain**. *(Toung et al. 2011)*
  - **20M de lectures** (75bp) permettent de détecter des **transcrits exprimés à un niveau moyen ou faible** chez le **poulet**. *(Wang et al. 2011)*
  - **10 M de lectures** permettent que **90% des transcrits** (**humain, zebrafish**) soient **couverts par 10** lectures en moyenne. *(Hart et al. 2013)*

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

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

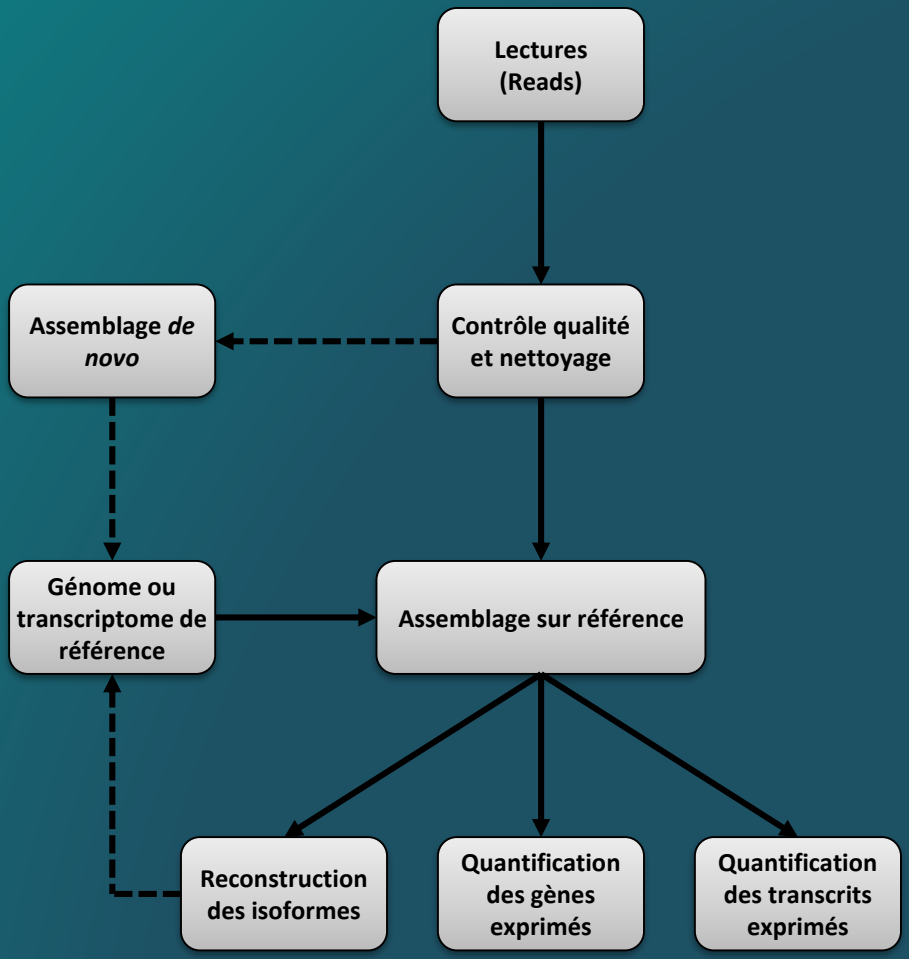
Stratégie d'analyse en fonction des données disponibles

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

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 la **Danio Rerio** (*zebra fish*)
  - extraits des runs **ERR022486** et **ERR022488**
  - explorer les données sur l'*European Nucleotide Archive* :
    - http://www.ebi.ac.uk/ena/
  - **Pour le TP** :
    - utilisation de données réduites (temps de calculs raccourcis)

Travaux pratiques

Quelques détails pratiques

- **Identifications de formation**
  - **Ordinateurs portable de formation**
  - **Galaxy :** https://galaxy.migale.inrae.fr/

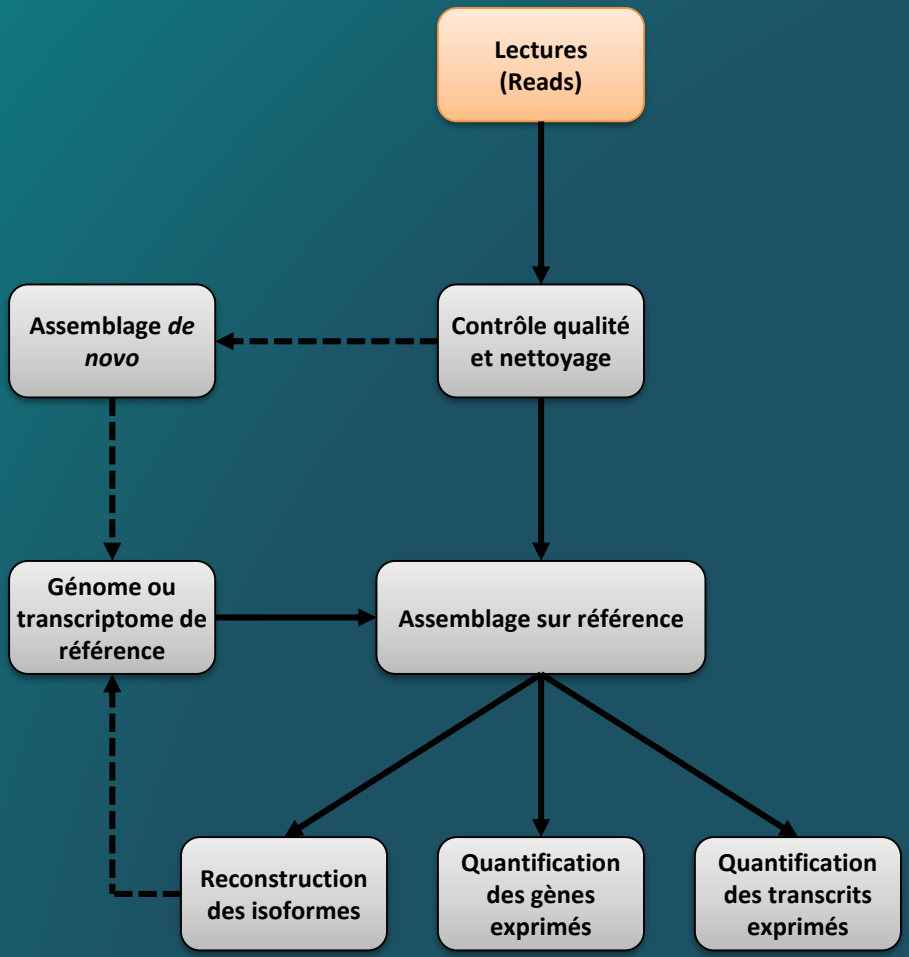
Login: **stage01**
Password: ***donné en début de formation***

Login: **stage02**
Password: ***donné en début de formation***

Login: **stage03**
Password: ***donné en début de formation***

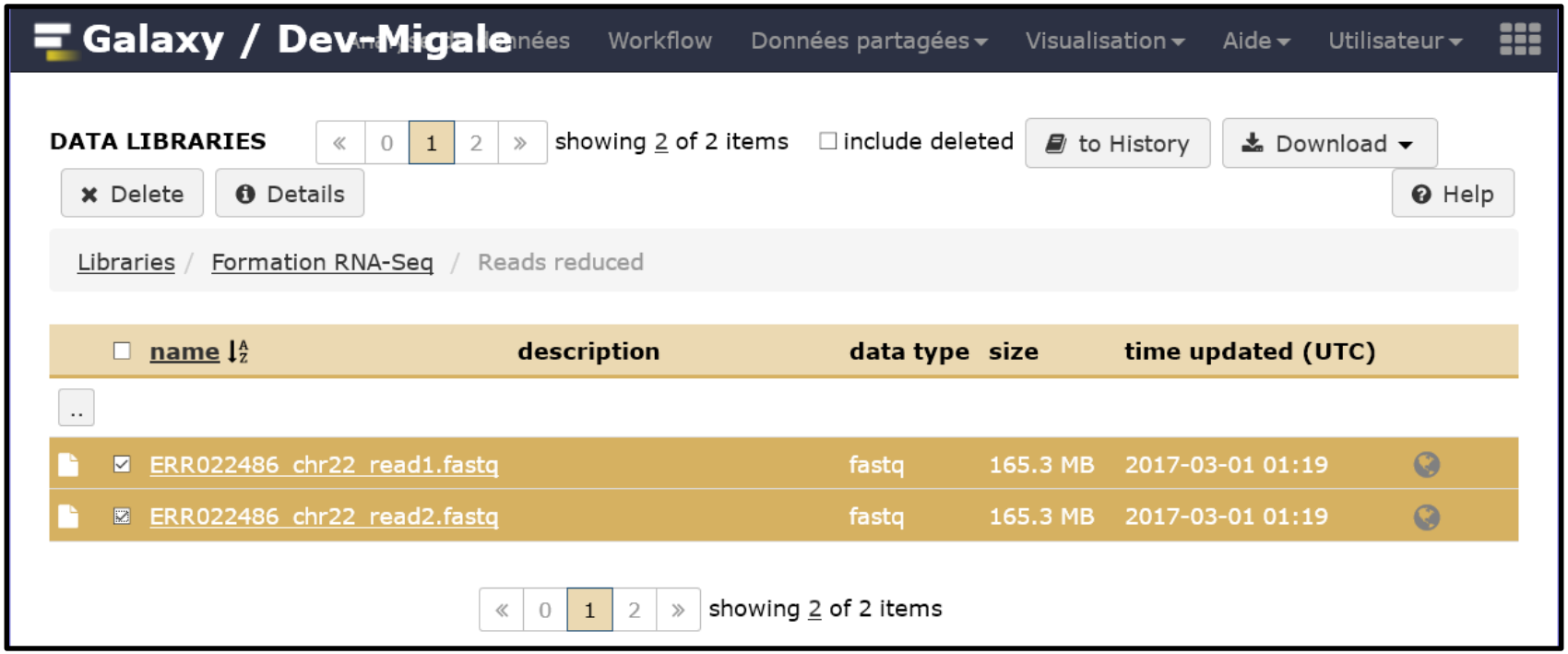
…

Workflow d'analyse RNA-Seq



Lectures brutes

Chargement des données par *Data Library*



Lectures brutes

Format Fastq

- 1 lecture = 4 lignes dans le fichier

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

- 1ère 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 clluster within the tile
197393	'y'-coordinate of the clluster 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

Lectures brutes

Format Fastq

- 2ème ligne = **séquence nucléotidique**

```
GCACACGTGTGGGTACTATTGCAGGGCTACTATGGATCGGCTAGTACCAGTGAGTCAGCTA
```

- 3ème ligne = répétition de l'identifiant de séquence (**ou rien**)

```
+
```

- 4ème ligne = **qualité**
  - **Phred quality score (format Sanger)**
  - $Q_{sanger} = -10 \log_{10} p$

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

Rappels sur les variants de Fastq

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |       |                              |                     |
33                          59   64      73                            104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.......................26...31........41
```

S - Sanger Phred+33, scores des séquences brutes compris entre 0 et 40
X - Solexa Solexa+64, scores des séquences brutes compris entre -5 et 40
I - Illumina 1.3+ Phred+64, scores des séquences brutes compris entre 0 et 40
J - Illumina 1.5+ Phred+64, scores des séquences brutes compris entre 3 et 40 avec 0=inutilisé, 1=inutilisé, 2=Indicateur de contrôle qualité de segment de séquence (en gras)
L - Illumina 1.8+ Phred+33, scores des séquences brutes compris entre 0 et 41

Biais spécifiques au RNA-Seq

Biais spécifiques:

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

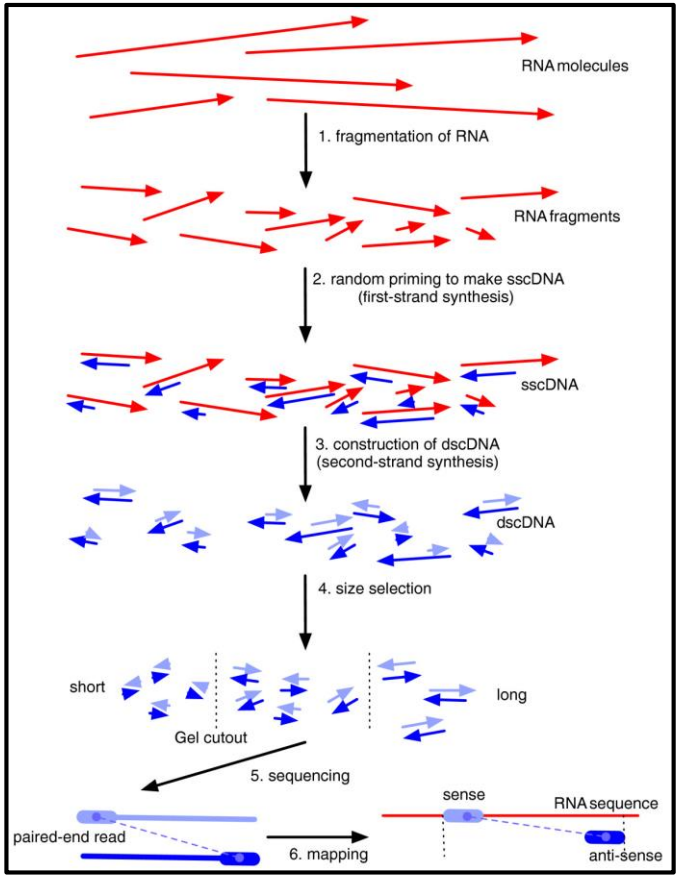
- Influence du séquençage
  - biais de position, de composition en séquence (contenu en GC)
  - influence de la longueur des transcrits

- « Mapabilité » du génome/transcriptome

Préparation de la banque

Etapes 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

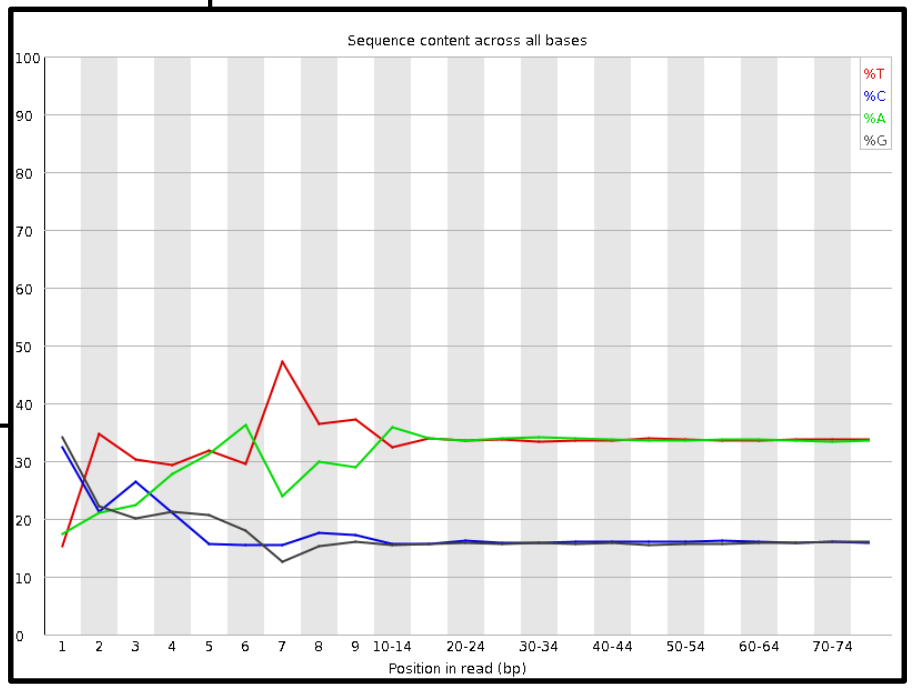
Published online 14 April 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 12 e131*
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

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

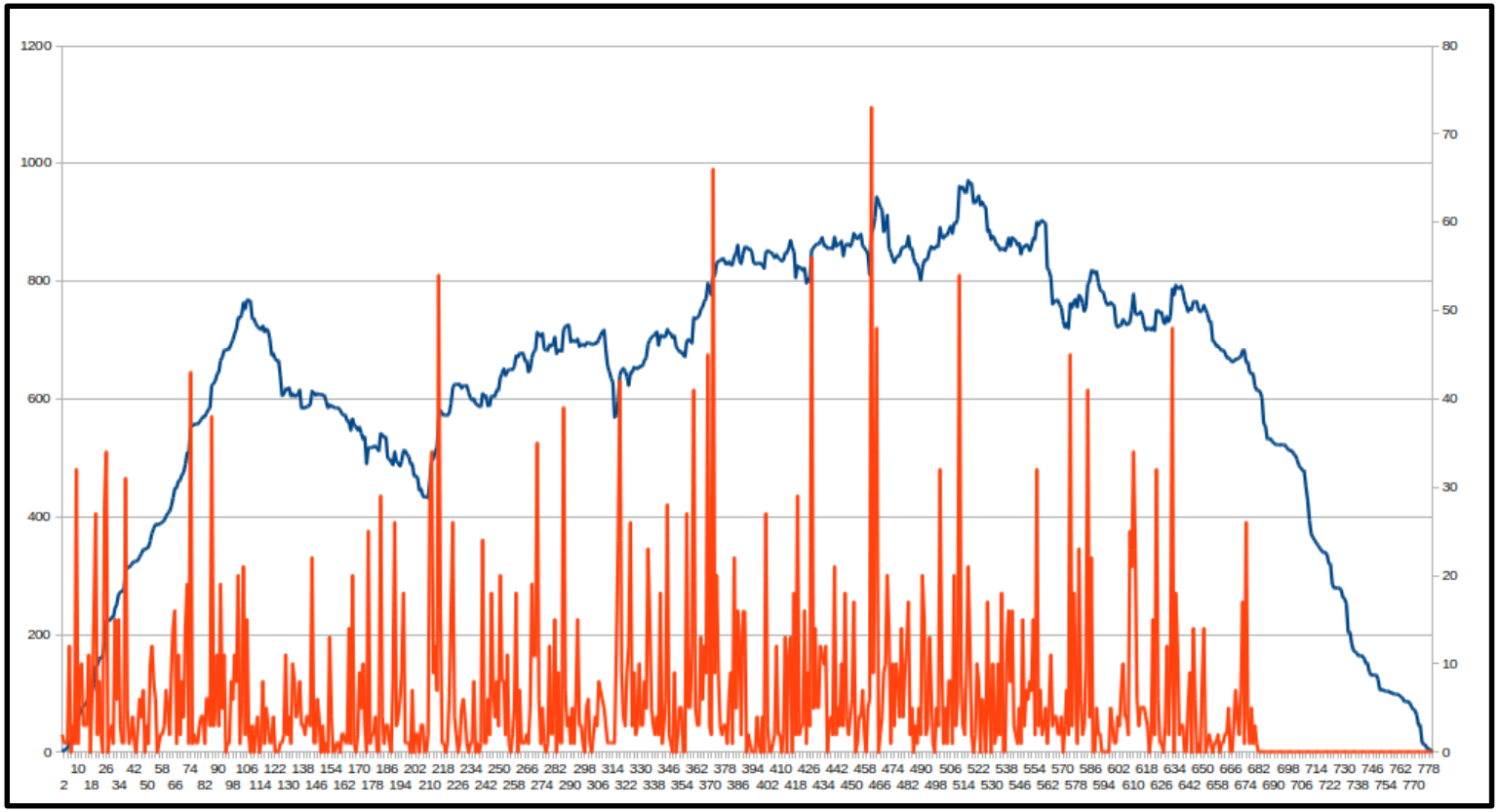
ABSTRACT

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.



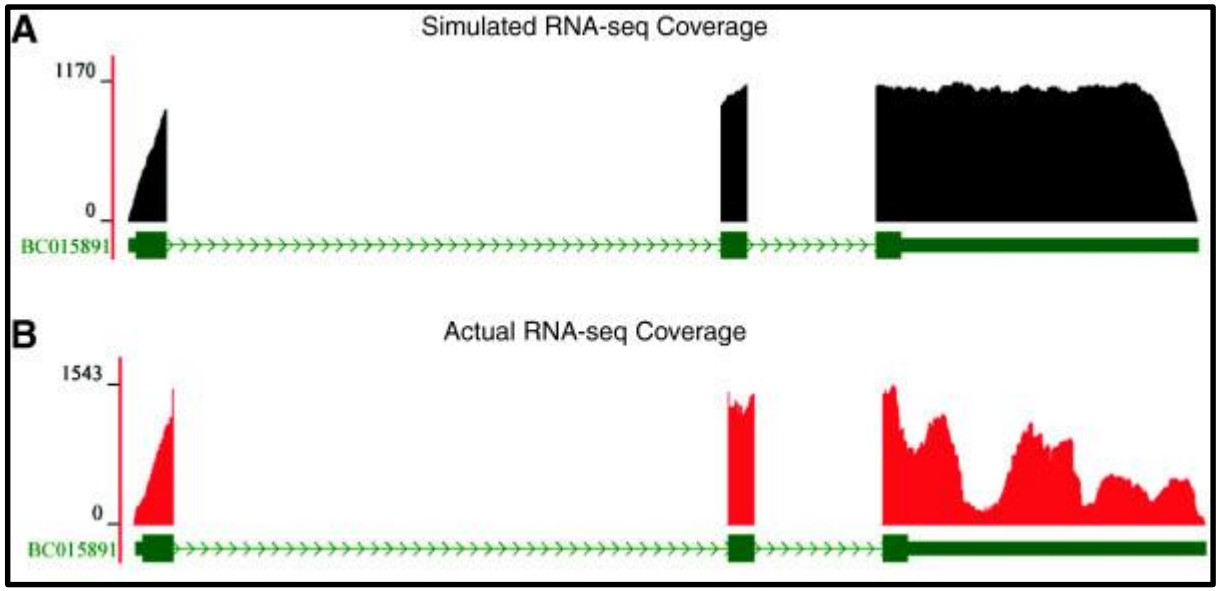
Biais : *random hexamer priming*

- **Influence** sur la **couverture**
  - Orange : première position des lectures
  - Bleu : couverture



Biais : *random hexamer priming*

- **Problèmes** dans l'**étude des nouveaux transcrits**

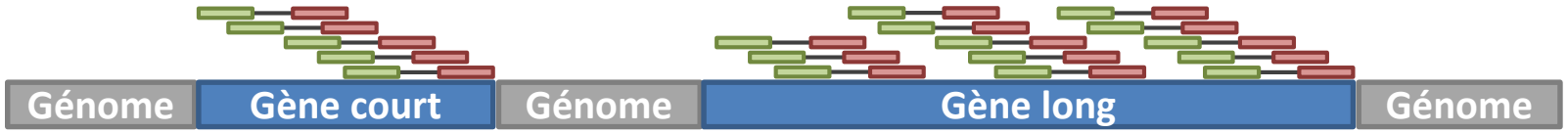


IVT-seq reveals extreme bias in RNA sequencing, Lahens et al. 2014

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**.

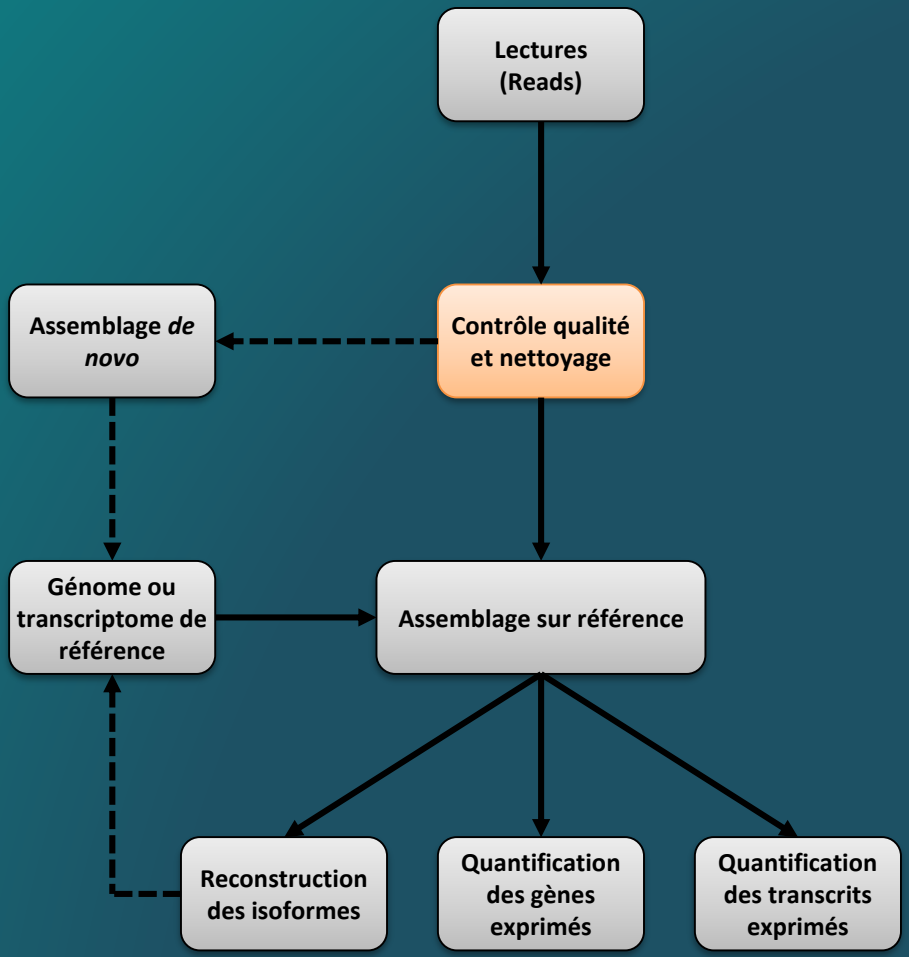
- Pour un **même gène** ayant **deux isoformes**, l'une faisant la moitié de l'autre, exprimé en **même abondance dans deux conditions différentes** :
  - L'isoforme la plus courte sera deux fois moins « comptée » que la plus longue



Biais : « mappabilité »

- Les étapes bioinformatiques peuvent être **influencées** par :
  - La **qualité** de la **référence**
    - **Assemblage**
    - **finition**
  - La **différence** (**distance**) entre la **référence** et **l'organisme étudié**
  - 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,…)**

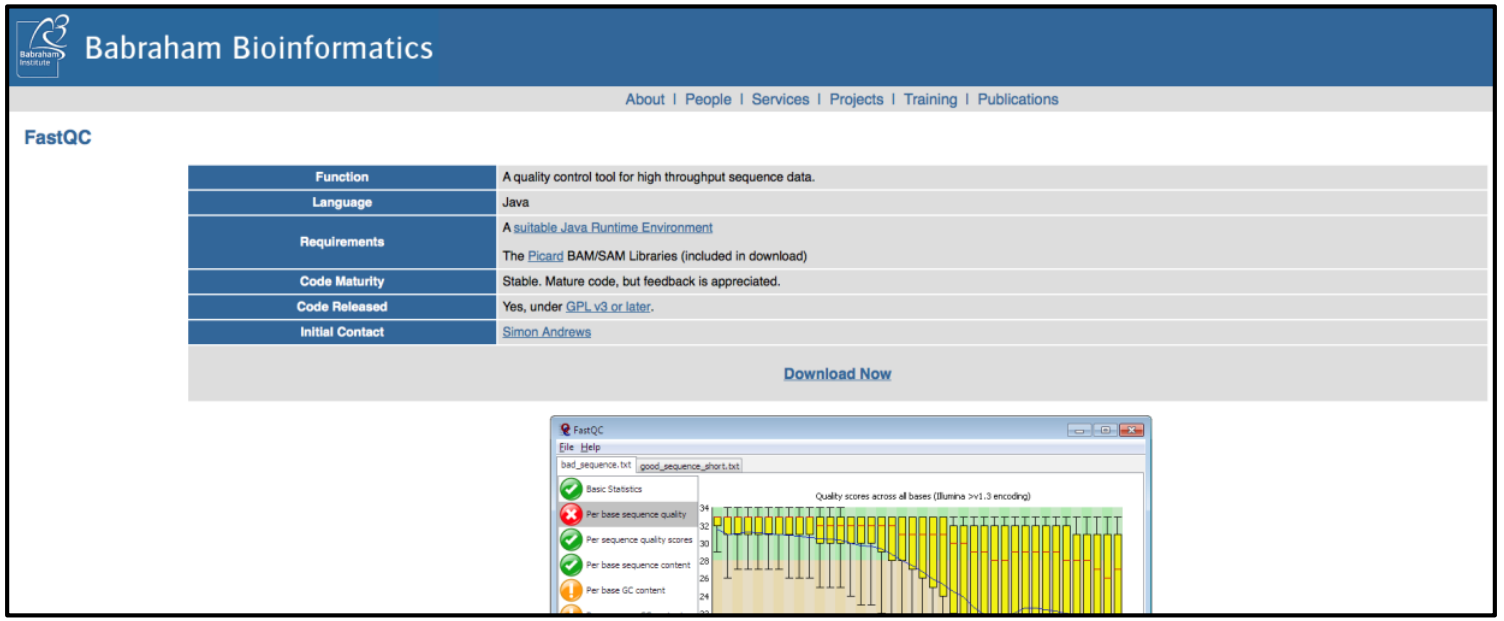
- Vérifier que les séquences peuvent **répondre au questions biologiques** posées :
  - **Biais techniques**
  - **Biais biologiques**

- Aider à paramètrer pour le **nettoyage** des données

Contrôle qualité

Outils :

- **FastQC**
  - plutôt orienté DNA-Seq



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

Contrôle qualité

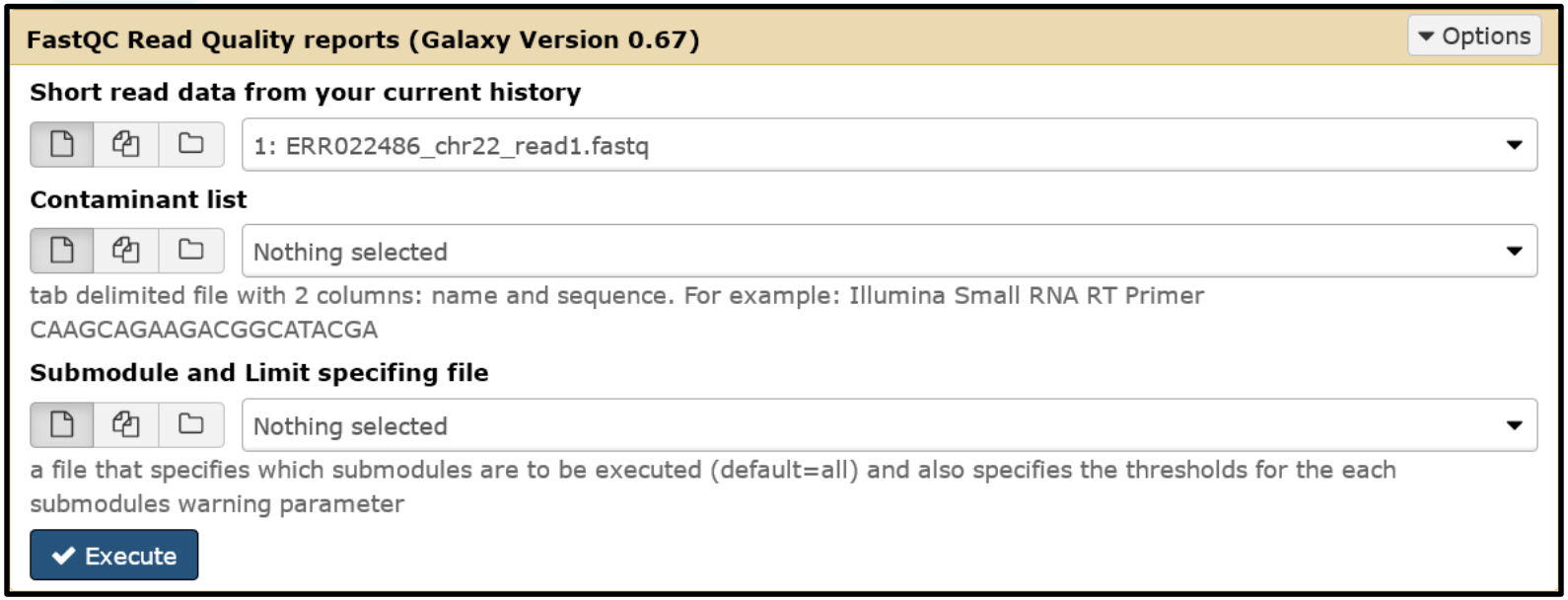
TP : **Métriques de qualité** avec **FastQC**

- Objectif :
  - vérifier la **qualité** des ***reads*** **bruts**
  - quels sont les **biais** que vous pouvez **identifier ?**

- En **entrée** :
  - **lectures** (**.fastq/.fastq.gz**)

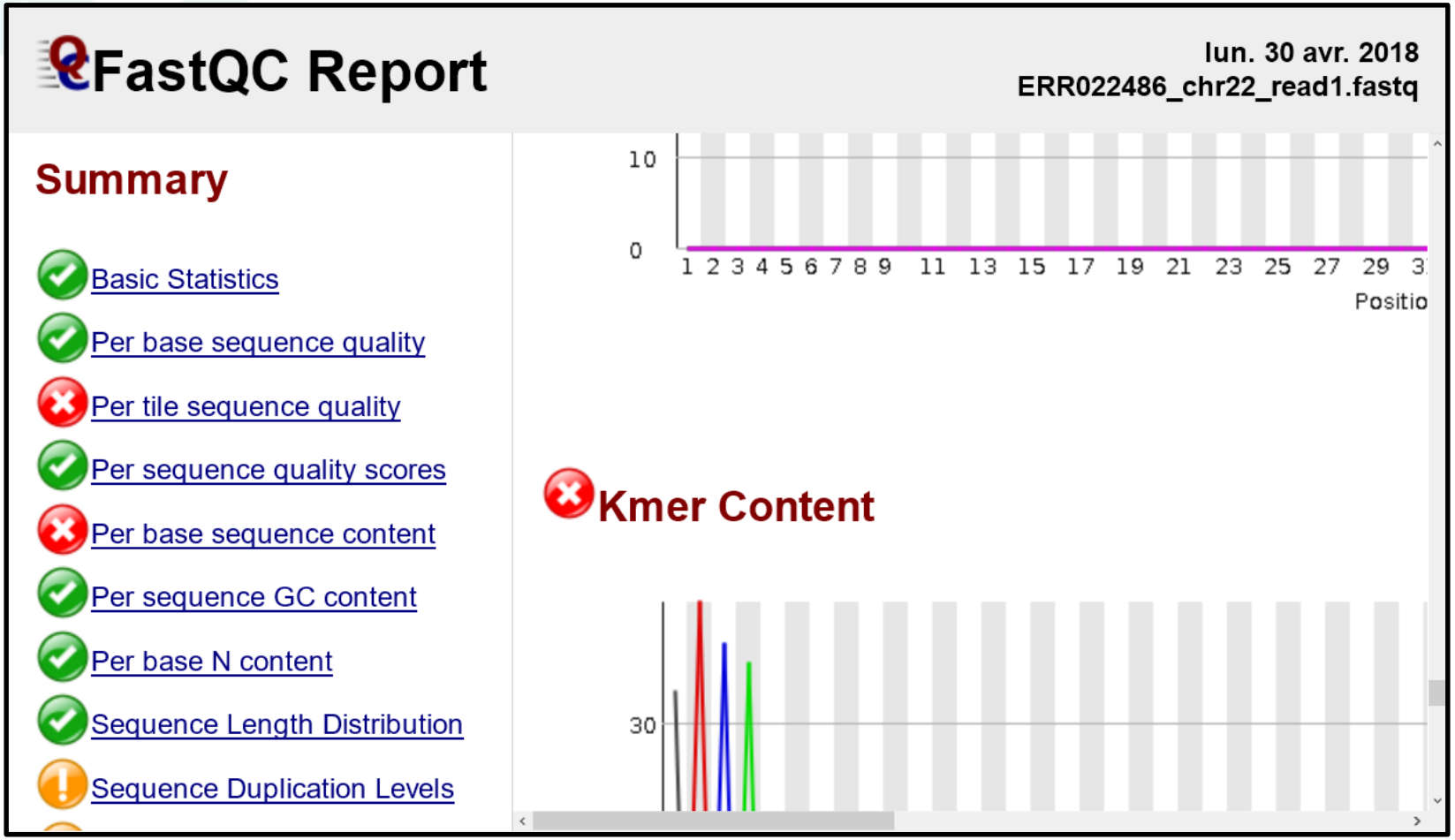
Contrôle qualité

TP : **Métriques de qualité** avec **FastQC**



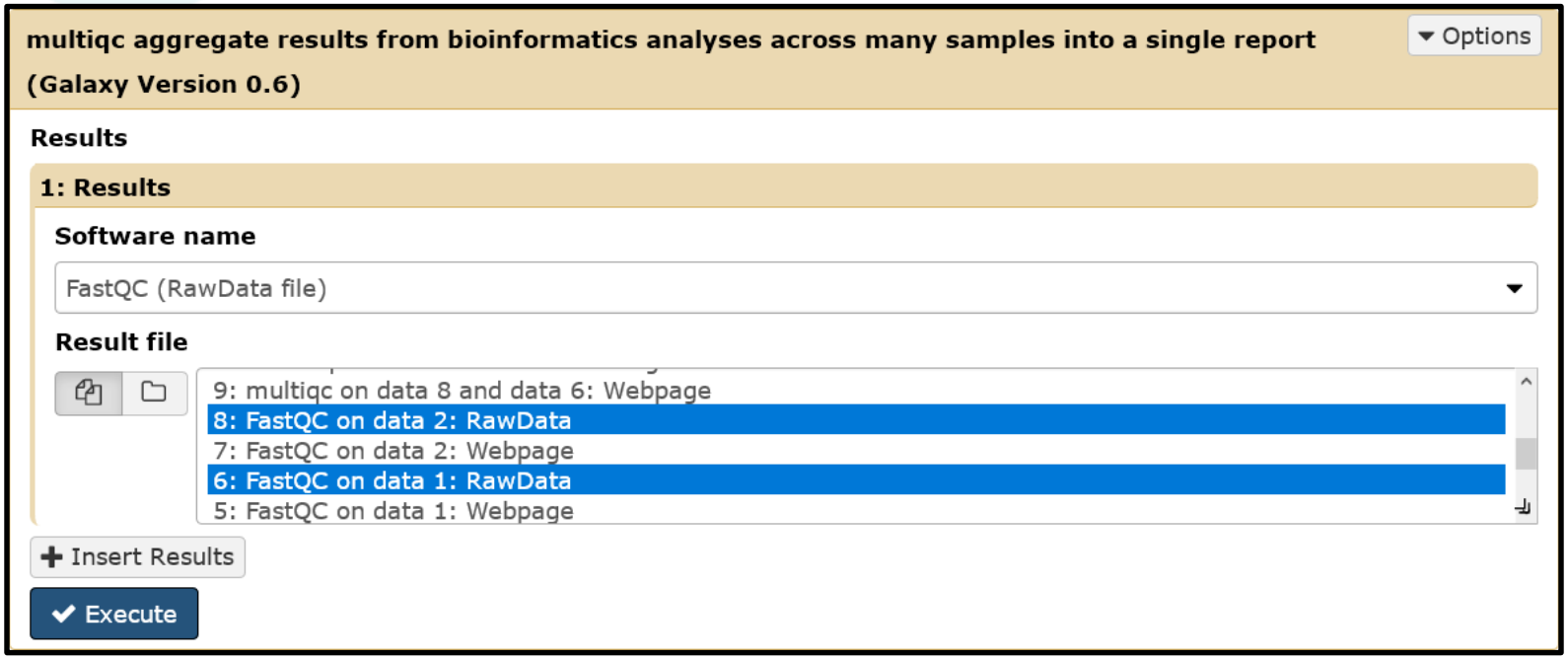
Contrôle qualité

TP : **Métriques de qualité** avec **FastQC**



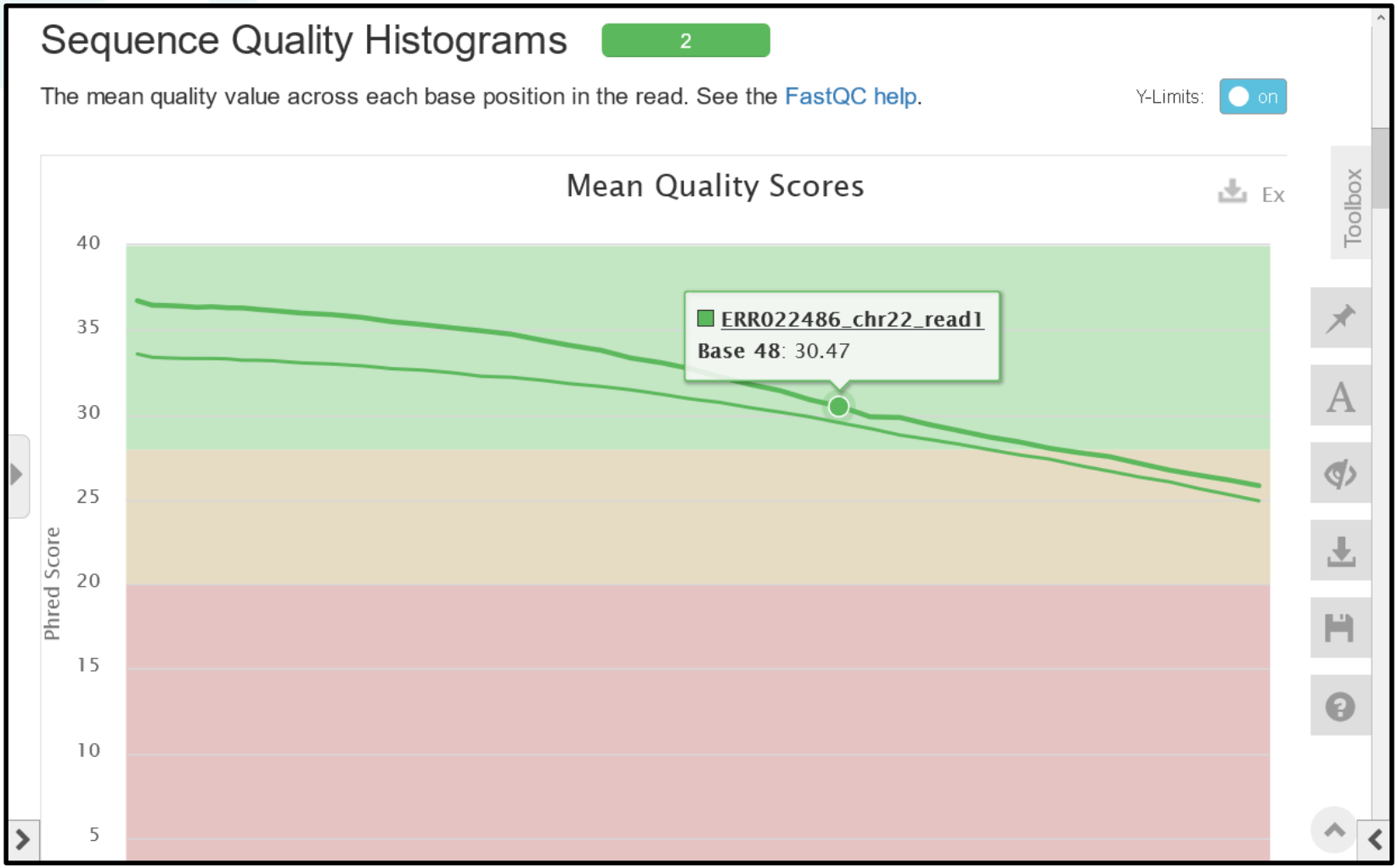
Contrôle qualité

TP : Regrouper les sorties FastQC avec MultiQC



Contrôle qualité

TP : Regrouper les sorties FastQC avec MultiQC

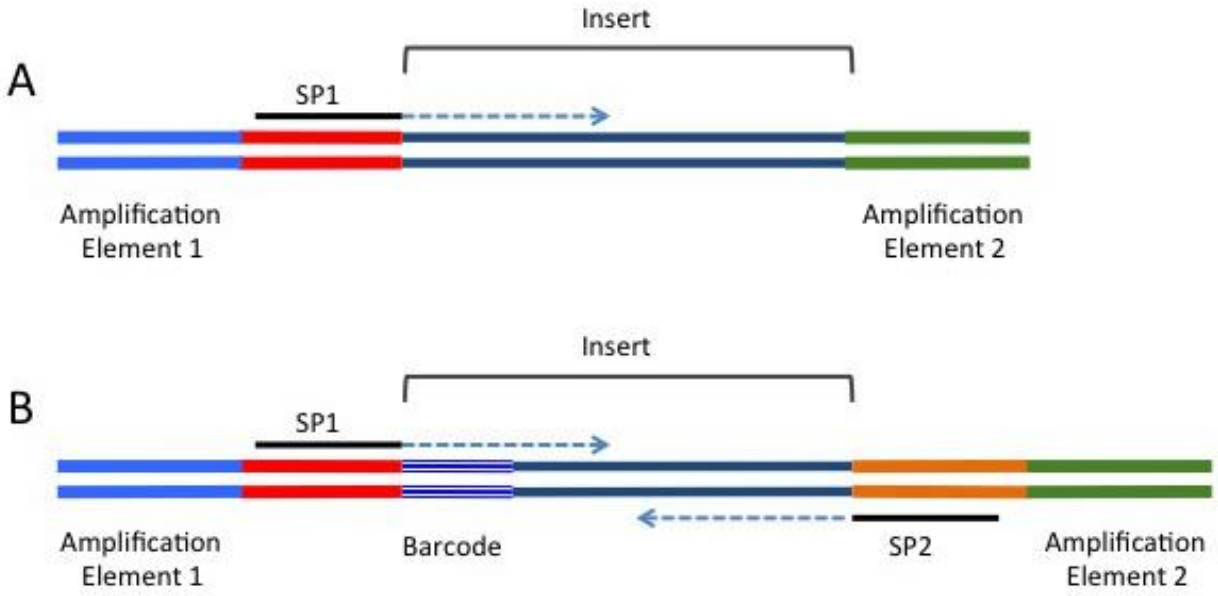


Nettoyage des données

- **Adaptateurs** et **Tags**
  - Cutadapt

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

Pourquoi nettoyer des données



http://rnaseq.uoregon.edu/

Nettoyage des données

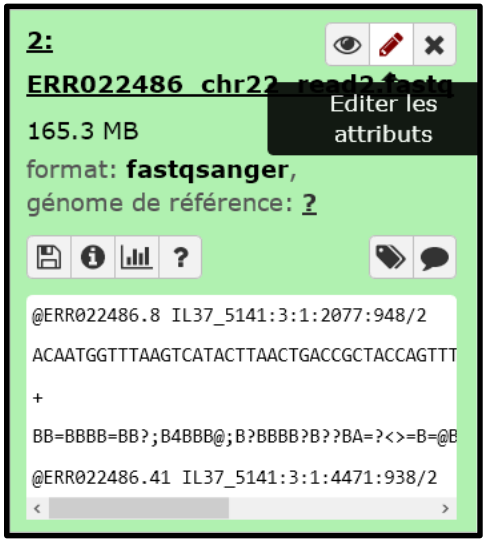
TP : **Nettoyage des lectures** avec **Sickle**

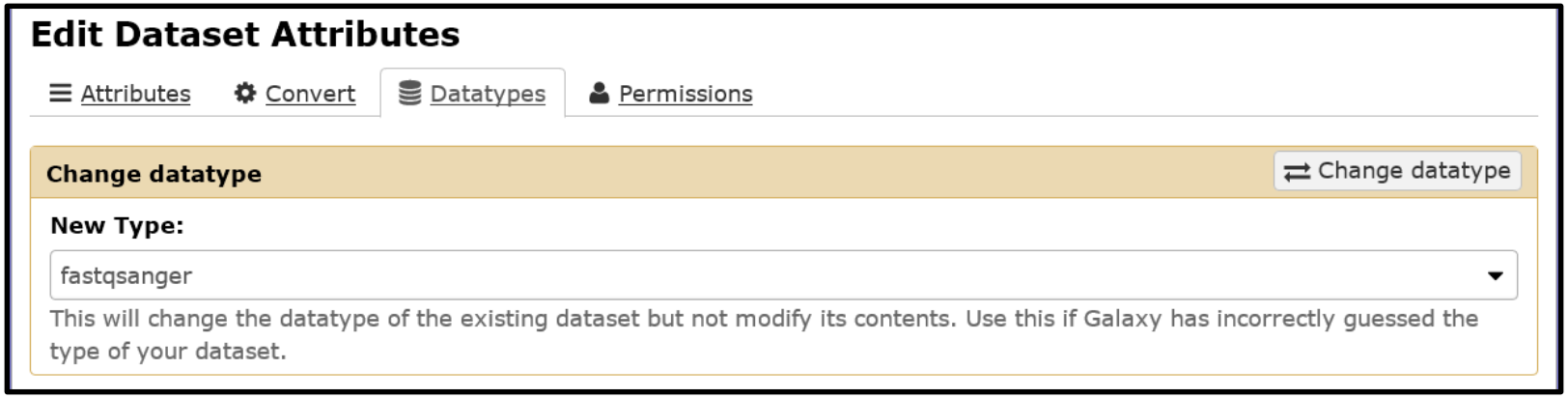
- **Nettoyer** les **lectures** avec les **paramètres** suivants :
  - Minimal length of 20
  - Minimal mean quality of 20
  - No N in seq
  - No 5’ trimming

- En **entrée** :
  - **lectures** (**.fastq/.fastq.gz**) **qualité** de **type Sanger**

Nettoyage des données

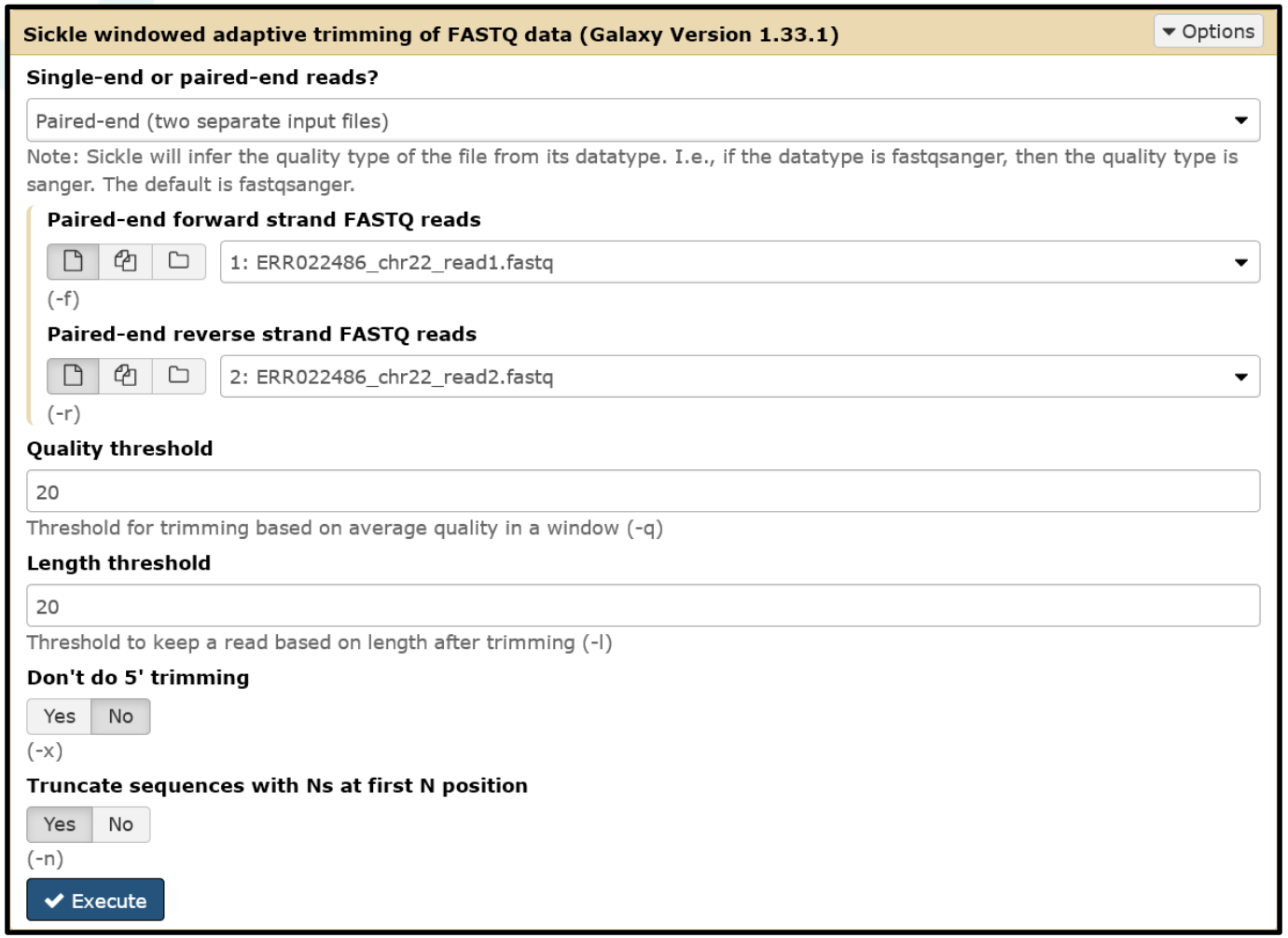
TP : Conversion de *datatype* fastq





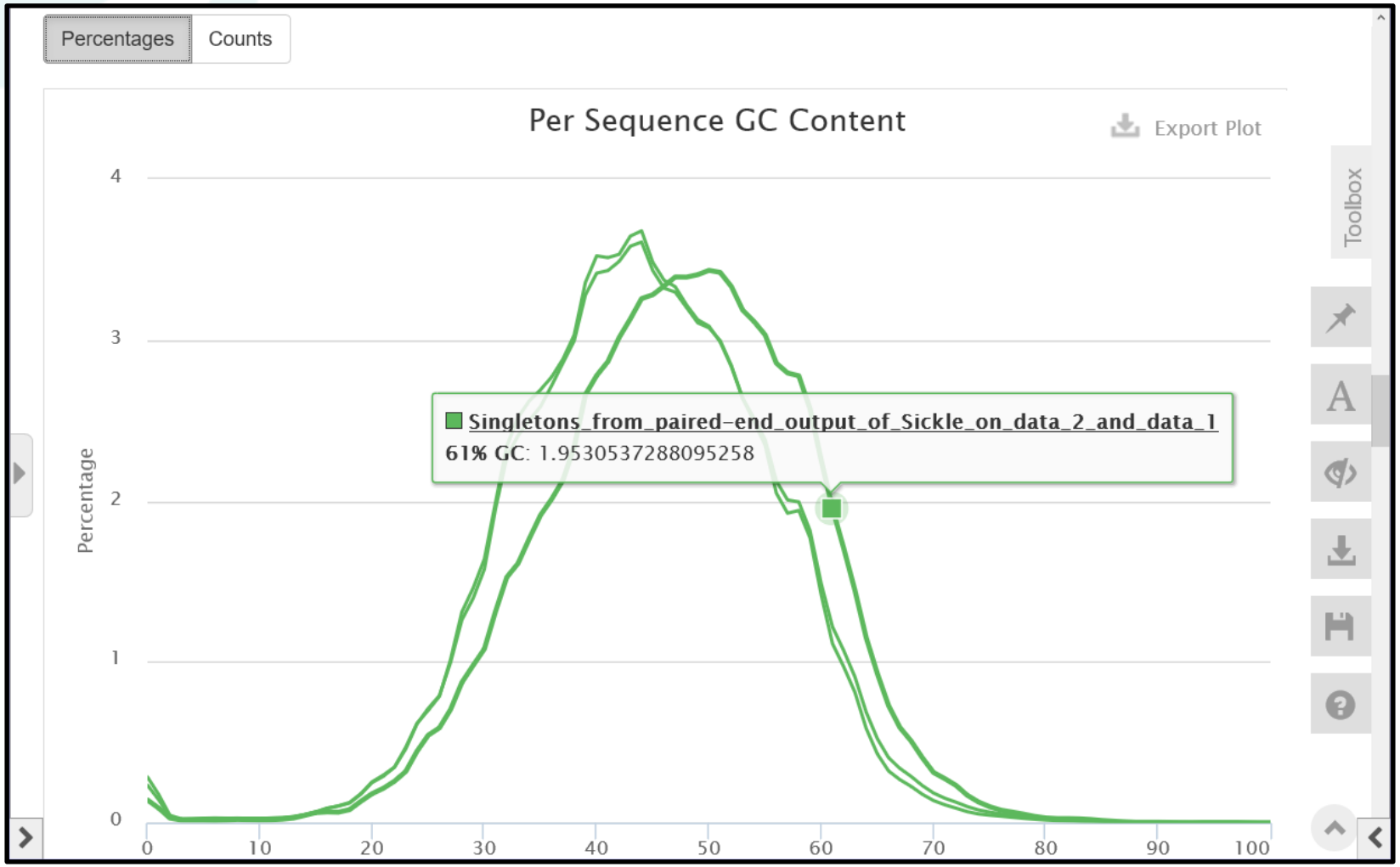
Nettoyage des données

TP : Nettoyage des lectures avec Sickle

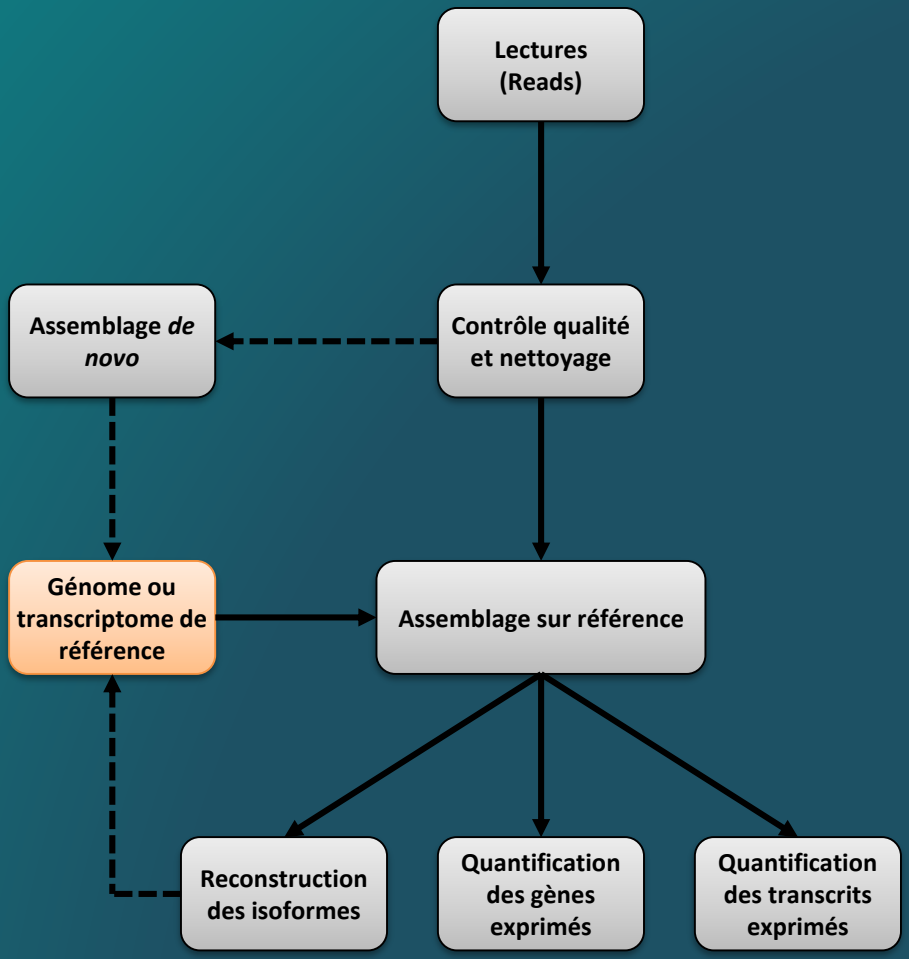


Nettoyage des données

TP : Nettoyage des lectures avec Sickle



Workflow d'analyse RNA-Seq



TP : chargement de la référence

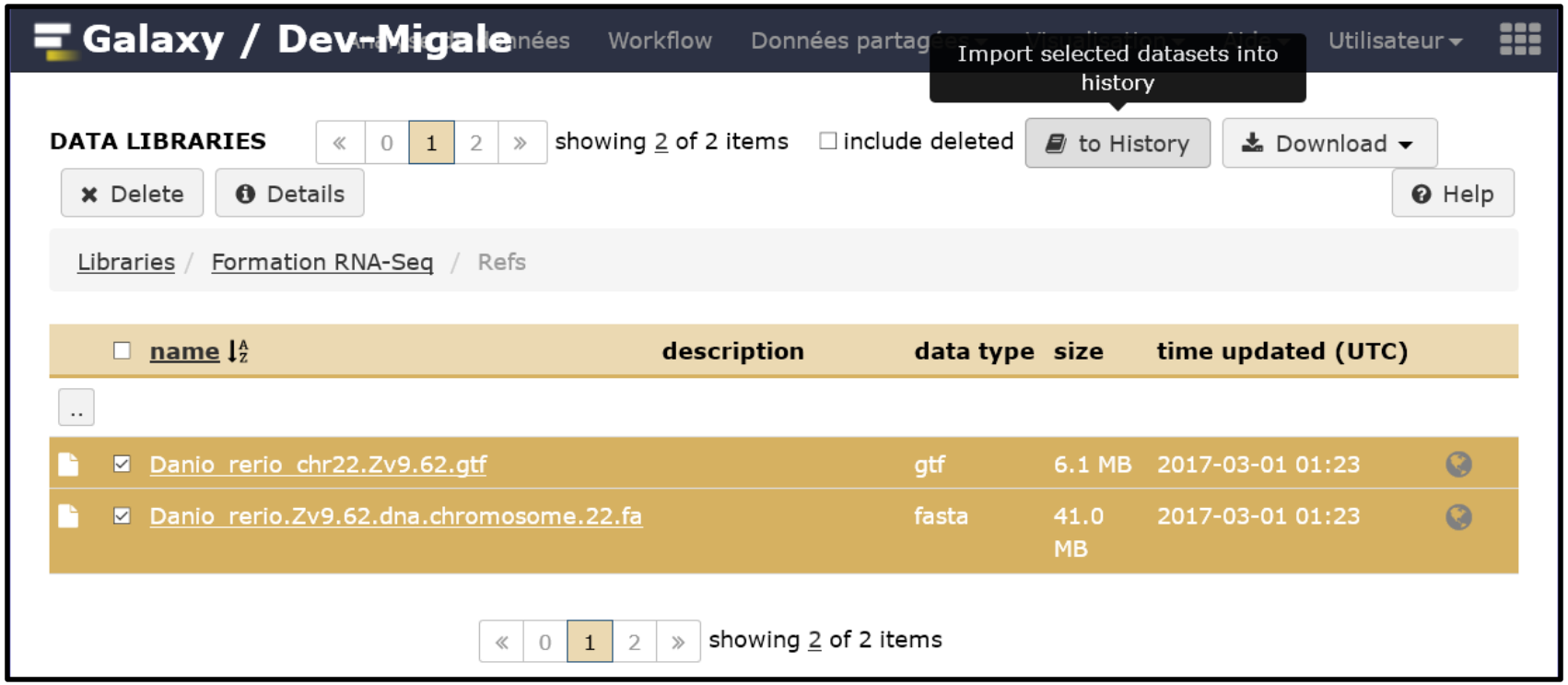
Objectif

- **Charger** la **séquence** et l'**annotation** de ***Danio rerio***

- La **référence** utilisée pour le **TP** ne contient que **le chromosome 22** (séquence et annotations)

Lectures brutes

Chargement des données du génome de référence



Format GTF : Gene Transfert Format

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

1.Seqname	2.Source	3.Feature	4.Start	5.End	6.Score	7.Strand	8.Frame	9.Attributes
22	protein_coding	exon	4683	4766	.	-	.	gene_id "ENSDARG00000089609"; transcript_id "ENSDART00000131195"; exon

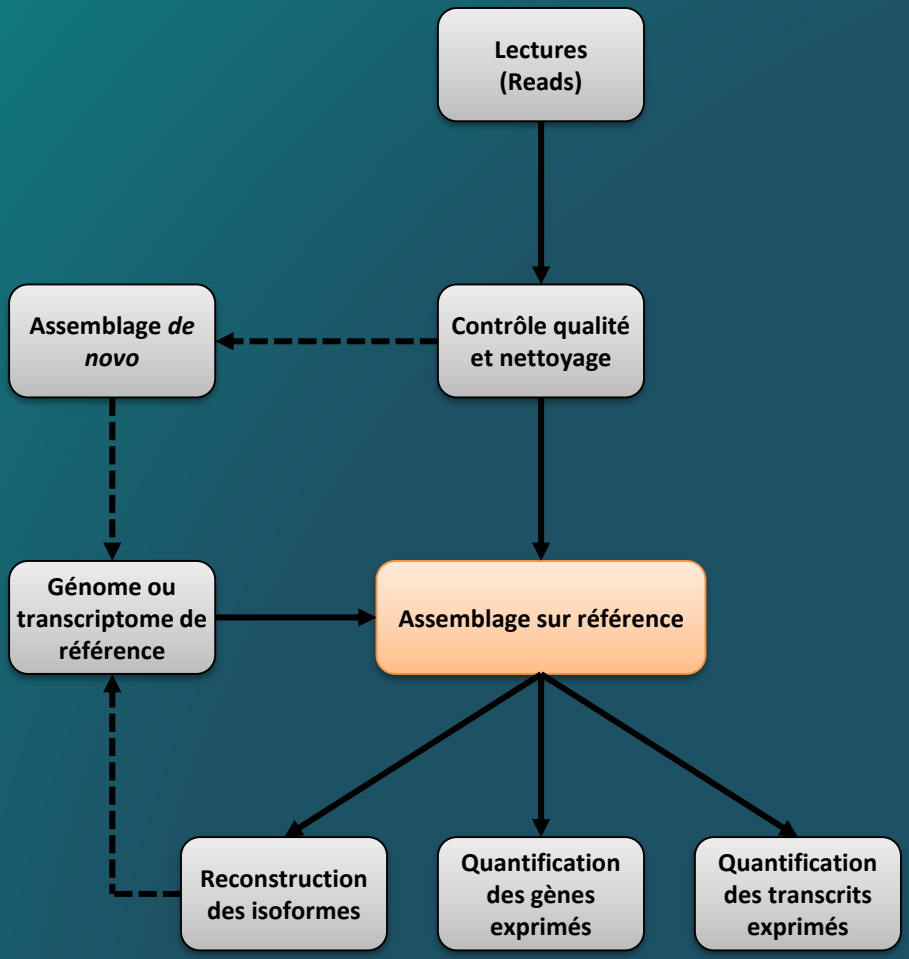
ttributes
ne_id "ENSDARG00000089609"; transcript_id "ENSDART00000131195"; exon_number "1"; gene_name "CABZ01094378.1"; transcript_name "CABZ01094378.1-201";

- **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

Workflow d'analyse RNA-Seq



Alignement épissé

Objectifs :

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

Alignement épissé

Données initiales

- **Lectures (brutes / nettoyées ?)**

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

Alignement épissé

Outils :

- **HISAT2**
  - **alignement** des **lectures RNA-Seq** sur une **référence** avec ***Bowtie2***
  - identifier les **jonctions** d'**épissage entre** les **exons**
  - capable de **travailler avec** ou sans liste de jonctions connues (transcriptome de référence)

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim[1,2], **Ben Langmead**[1,3], and **Steven L Salzberg**[1,3]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

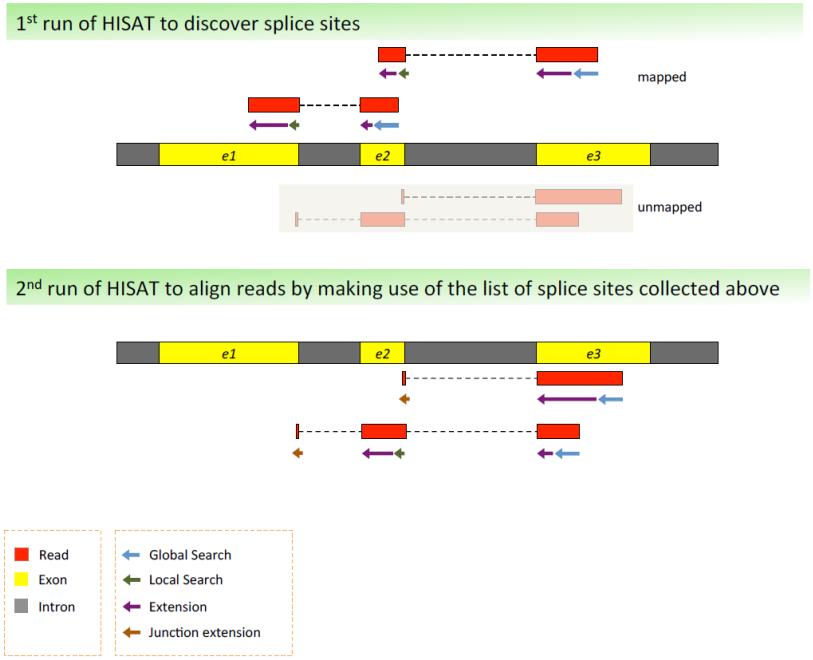
[2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA

[3]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA

https://ccb.jhu.edu/software/hisat2/index.shtml

Alignement épissé

- **HISAT2**



HISAT: a fast spliced aligner with low memory requirements
Supplementary Figures 8 & 9

HISAT2

Alignement épissé

- En **entrée** :
  - **lectures** (**.fastq**)
  - **séquence génomique (.fasta)**
  - **annotation structurale** du **génome** (**.gtf**) [optionnel]

Alignement épissé

Attention aux paramètres par défaut !

- **Tailles des introns par défaut :**
  - **min : 20**
  - **max : 500000**

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

Alignement épissé

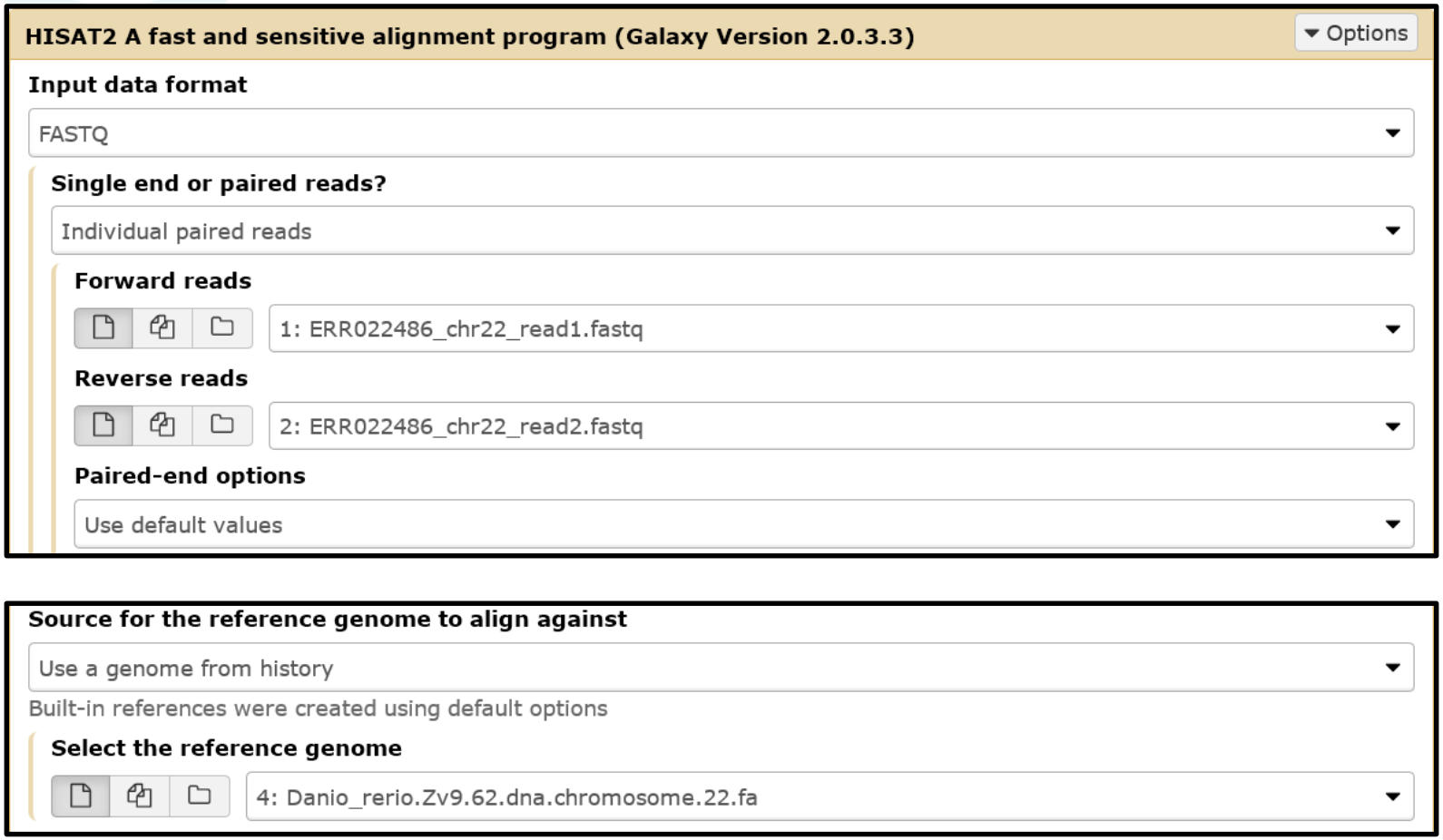
TP : lancement de *HISAT2*

- **Utiliser HISAT2** :
  - **avec** un transcriptome de référence
  - ne **pas** chercher des **introns** de taille **supérieur à 5000**
  - avec une **taille d'insert** de **200**

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

Alignement épissé

TP : lancement de *HISAT2*



Alignement épissé

TP : lancement de *HISAT2*

Spliced alignment parameters

Specify spliced alignment parameters

Maximum intron length

5000

Specify strand-specific information

FR Unstranded

'F' means a read corresponds to a transcript. 'R' means a read corresponds to the reverse complemented counterpart of a transcript. (--rna-strandness)

Disable spliced alignment

Yes | No

(--no-spliced-alignment)

GTF file with known splice sites

3: Danio_rerio_chr22.Zv9.62.gtf

Alignement épissé

HISAT2

- En **sortie** :
  - liste des **lectures alignées** (.bam)

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS
@HD VN:1.0 SO:coordinate							
@SQ SN:22 LN:42261000							
@PG ID:hisat2 PN:hisat2 VN:2.0.3-beta CL:"/projet/galaxydev/galaxy/database/dependencies/_conda/envs/mulled-v1-999b52958							
ERR022486.20746436	137	22	77	255	76M	=	77
ERR022486.20746436	69	22	77	0	*	=	77
ERR022486.27659380	99	22	178	255	76M	=	1670
ERR022486.360792	145	22	196	255	76M	=	78334
ERR022486.13842468	133	22	238	0	*	=	238
ERR022486.13842468	89	22	238	255	1S71M1077N4M	=	238
ERR022486.20148694	69	22	242	0	*	=	242
ERR022486.20148694	153	22	242	255	67M1077N9M	=	242
ERR022486.27365215	99	22	246	255	63M1077N13M	=	1707
ERR022486.500352	81	22	253	255	56M1077N20M	=	77561

Alignement épissé : Format SAM / BAM

Sequence / Binary Alignment Map format

```
1:497:R:-272+13M17D24M  113     1       497     37      37M     15      100338662       0       CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG    0;==-==9;>>>>>>=>>>>>>>>>>>>>=>>>>>>>>>>    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
19:20389:F:275+18M2D19M 99      1       17644   0       37M     =       17919   314     TATGACTGCTAATAATACCTACACATGTTAGAACCAT    >>>>>>>>>>>>>>>>>>>>>><>>><<>>4:>>:<9   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
19:20389:F:275+18M2D19M 147     1       17919   0       18M2D19M        =       17644   -314    GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT    ;44999;499<8<8<<<8<<><<<<<>7<;<<<>><<   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
9:21597+10M2I25M:R:-209 83      1       21678   0       8M2I27M =       21469   -244    CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT    <;9<<5><<<<><<<><<><<><9>><>>>9>>><>    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
```

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*
- **Conversion avec Samtools**

NGS: SAM Tools

SAM-to-BAM converts SAM format to BAM format

BAM-to-SAM converts BAM format to SAM format

Alignement épissé : Chaine 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

Autres outils d'alignement épissé

Comparaison et unification d'outil de mapping

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 18 2011, pages 2518–2528
doi:10.1093/bioinformatics/btr427

Sequence analysis

Advance Access publication July 19, 2011

Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)

Gregory R. Grant[1,2,3,*], Michael H. Farkas[4], Angel D. Pizarro[2], Nicholas F. Lahens[5], Jonathan Schug[3], Brian P. Brunk[1], Christian J. Stoeckert[1,3], John B. Hogenesch[1,2,5] and Eric A. Pierce[4,*]

[1]Penn Center for Bioinformatics, [2]Institute for Translational Medicine and Therapeutics, [3]Department of Genetics, [4]F.M. Kirby Center for Molecular Ophthalmology and [5]Department of Pharmacology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Associate editor: Ivo Hofacker

RESEARCH ARTICLE

RNA-Seq differential expression analysis: An extended review and a software tool

Juliana Costa-Silva[1], Douglas Domingues[1,2], Fabricio Martins Lopes[1]*

Autres outils d'alignement épissé

STAR

Table 1. Mapping speed and RAM benchmarks on the experimental RNA-seq dataset

Aligner	Mapping speed: million read pairs/hour		Peak physical RAM, GB	
	6 threads	12 threads	6 threads	12 threads
STAR	309.2	549.9	27.0	28.4
STAR sparse	227.6	423.1	15.6	16.0
TopHat2	8.0	10.1	4.1	11.3
RUM	5.1	7.6	26.9	53.8
MapSplice	3.0	3.1	3.3	3.3
GSNAP	1.8	2.8	25.9	27.0

Dobin et al. Bioinformatics 2013

CRAC

Table 2 Comparative evaluation of splice junction prediction tools

	75 bp		200 bp	
Tool	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
CRAC	79.43	**99.5**	*86.02*	**99.18**
GSNAP	*84.17*	97.03	72.94	97.09
MapSplice	79.89	*97.68*	84.72	*98.82*
TopHat	**84.96**	89.59	54.07	94.69
TopHat2	82.25	92.71	**88.65**	91.35

We compared the sensitivity and precision of different tools on the human simulated RNA-seq (42M, 75 nt and 48M, 200 nt) against the human genome for splice junction prediction. The sensitivity is the percentage of correctly reported cases over all sequenced cases, while the precision is the percentage of correct cases among all reported cases. Values in bold in the three tables indicate the maximum of a column, and those in italics the second highest values. For all tasks with the current read length, CRAC combines good sensitivity and very good precision. Importantly, CRAC always improves sensitivity with longer reads, and yields the best precision (that is the fewer false positives) over all solutions, even against specialized tools like TopHat.

Philippe et al. Genome Biol. 2013

Autres outils d'alignement épissé

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

- «We have identified that the **impact of the mapping tool on the**
- **final results is minimal** »

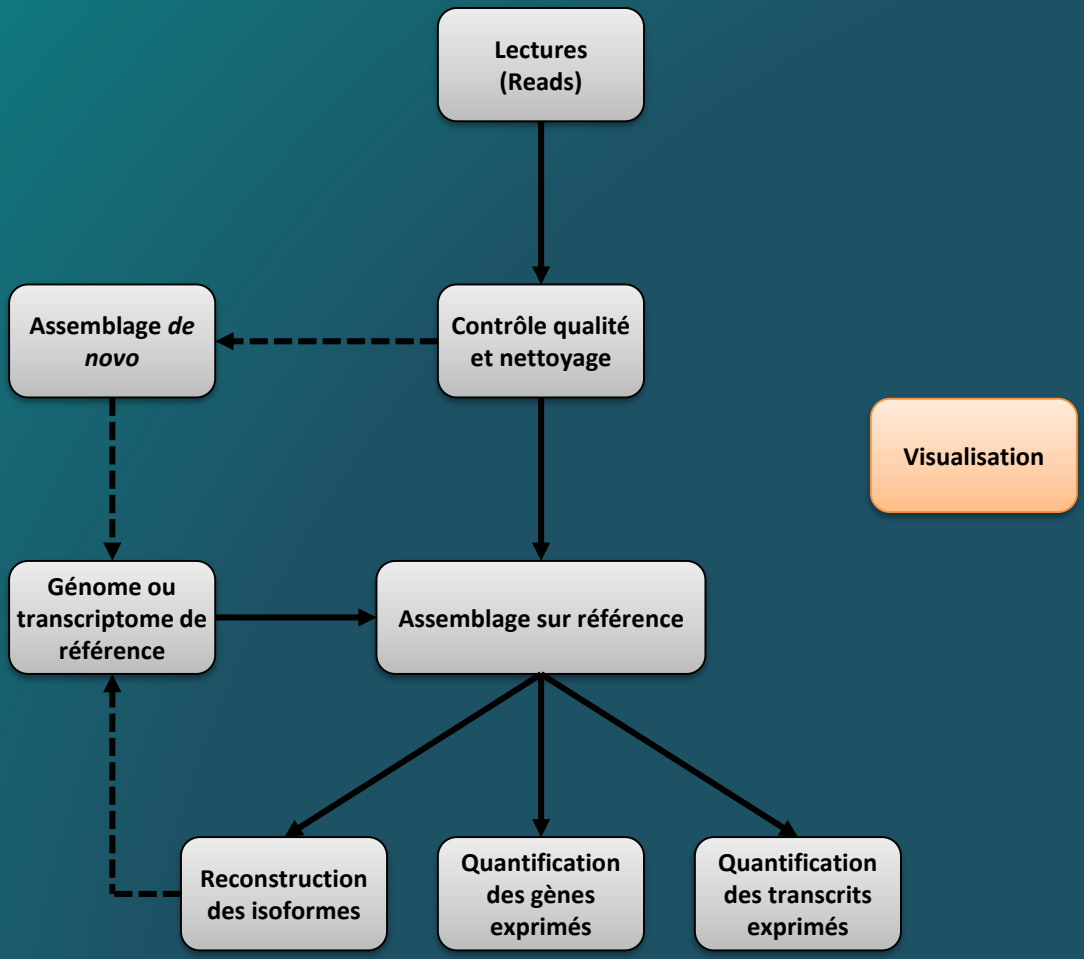
Engström et al., Nature Methods, 2013
Cosat-Silva et al., PLOS One, 2017

Alignement épissé

Conclusion

- **Alignement épissé est plus couteux** et **compliqué que** l'**alignement** de lectures courtes **classique**
- Il peut néanmoins s'**appuyer** sur des **outils d'alignement généralistes** (**Bowtie2**)
- Cette **étape** est **couteuse** 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** (**Bowtie2 ou BWA**)

Workflow d'analyse RNA-Seq



Visualisation

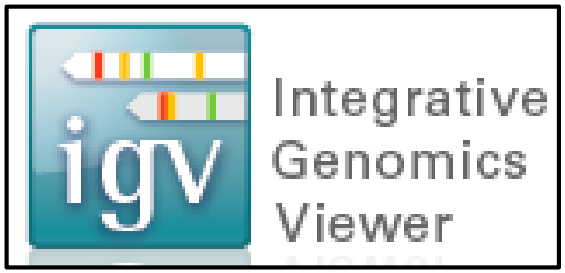
Une bonne dizaine d'outils

- **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, bigwig, gff, gtf, bed, custom, …
- **Documenté** et **maintenu**

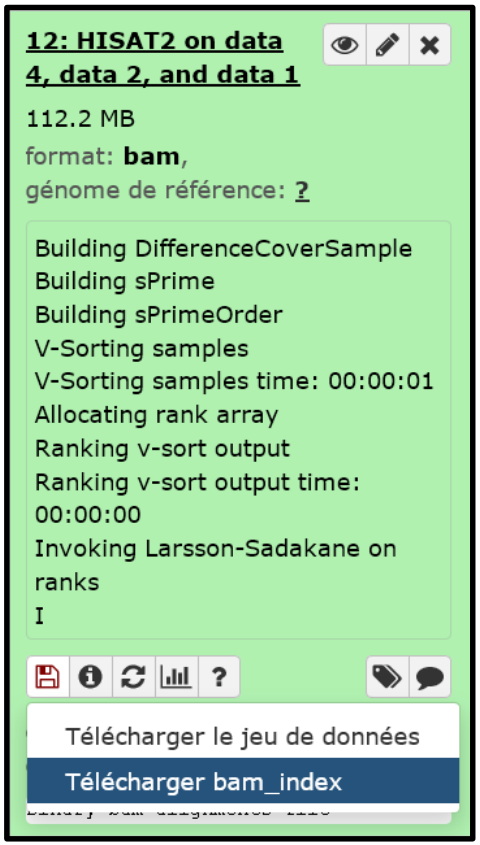


http://www.broadinstitute.org/igv/home

TP : Récupération des BAM et références

Objectif

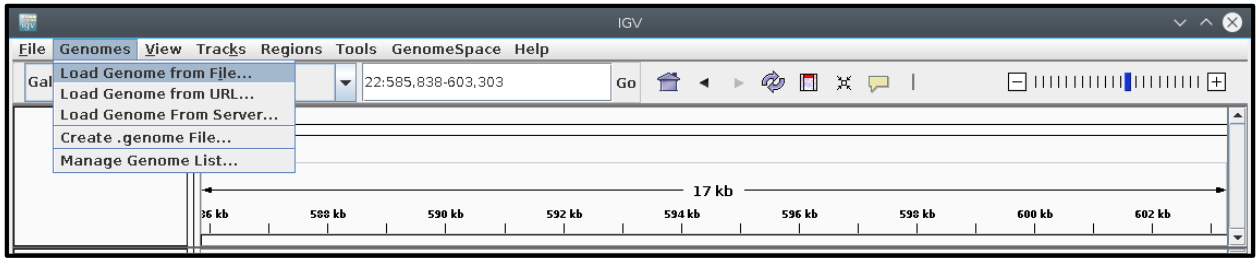
- **Téléchargement** des **alignements épissés** (**bam**)
- **Téléchargement** des **index** (**bai**)
- **Lancer IGV sous linux :**
  - touches « **Alt** » + « **F2** »
  - taper « **igv** »
  - touche « **Entrée** »



Visualisation avec IGV

Chargement du génome

- Charger la séquence du génome
  - fichier fasta

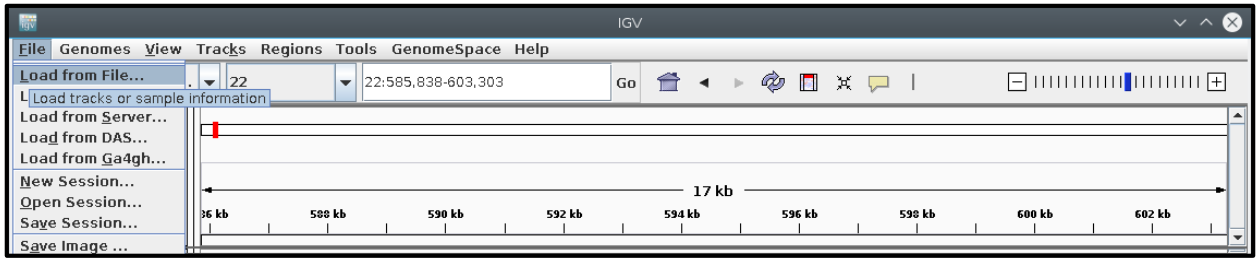


Visualisation avec IGV

Chargement des résultats

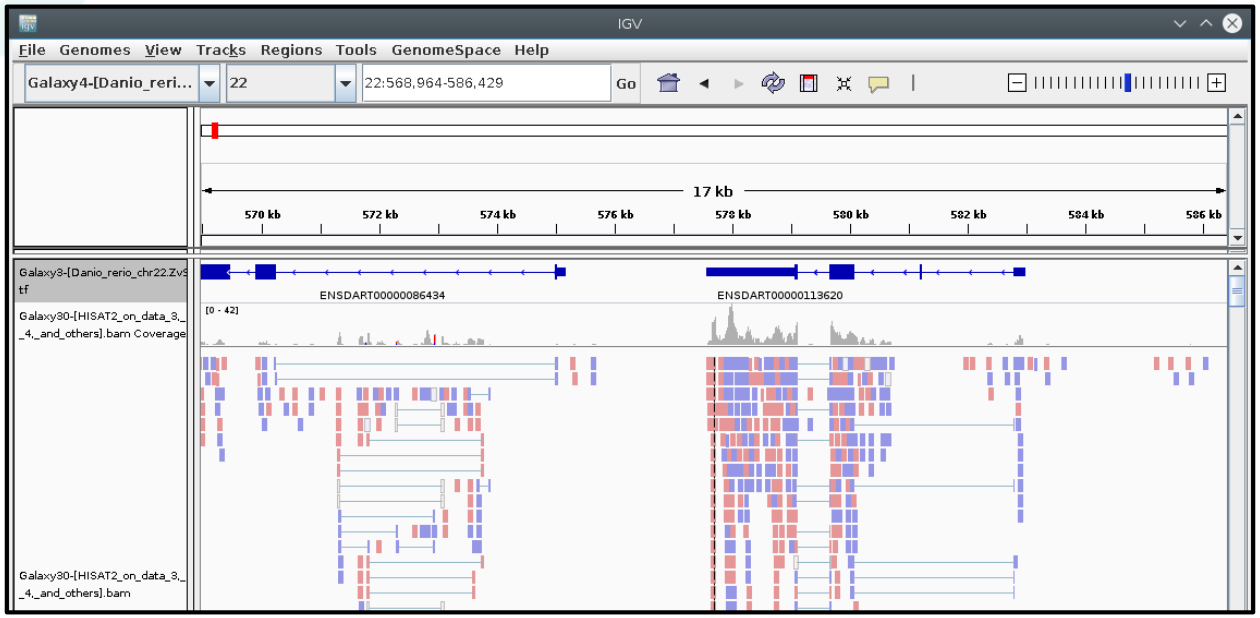
- Charger l'**annotation**
  - fichier gtf

- Charger les **résultats**
  - fichier .bam



Visualisation avec IGV

Navigation à la souris



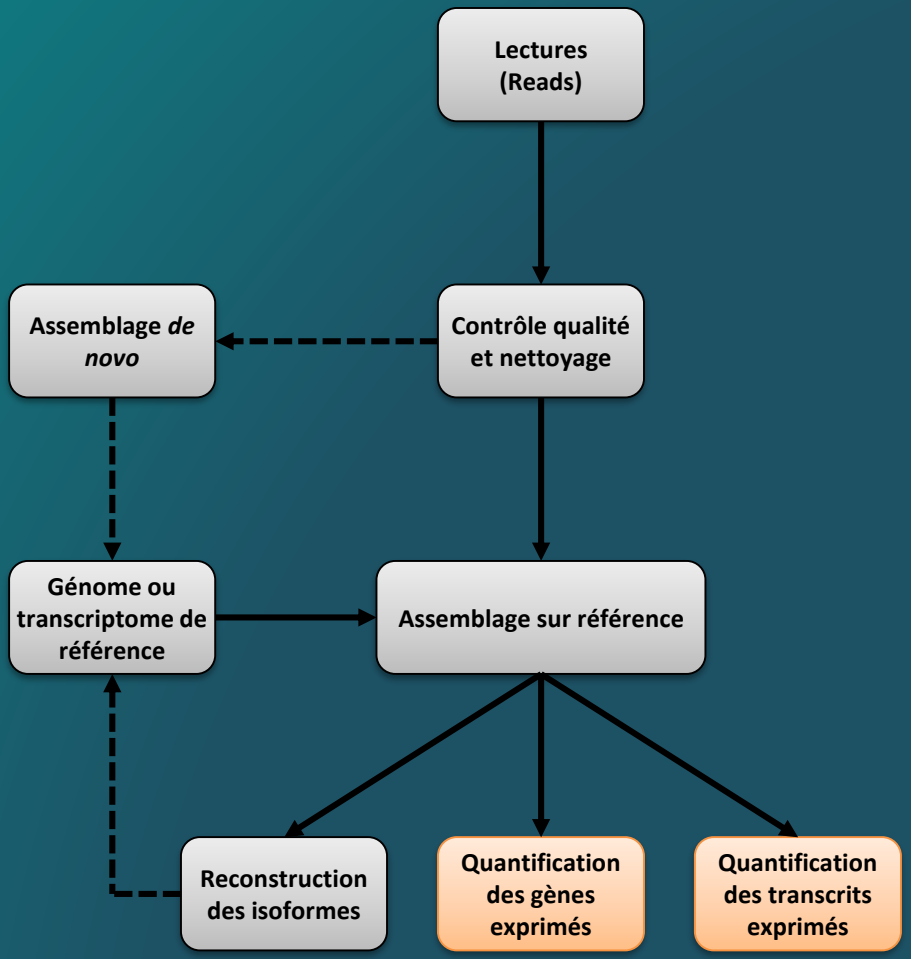
- Penser à **sauvegarder la session**

Visualisation avec IGV

TP : Visualiser les deux alignements précédents

- **Séquence** (fasta) et **annotation** (gtf)
- **Alignements** (bam), l'**index** (bai) doit être au même endroit.
- **S'intéresser** plus particulièrement **aux régions** :
  - **22:585,838-603,303** : rétention d'intron ?
  - **22:669,413-678,616** : frontières d'exons ? UTR ?
  - **22:2,754,99-2,772,496** : nouveau transcrit ? intérêt des reads orientées ?

Workflow d'analyse RNA-Seq



Quantification

Que cherche t'on à compter ?

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

Seqname	Source	Feature	Start	End	Score	Strand
22	protein_coding	exon	4683	4766	.	-
22	protein_coding	CDS	4683	4766	.	-
22	protein_coding	start_codon	4764	4766	.	-
22	protein_coding	exon	4414	4607	.	-
22	protein_coding	CDS	4414	4607	.	-
22	protein_coding	exon	4213	4340	.	-

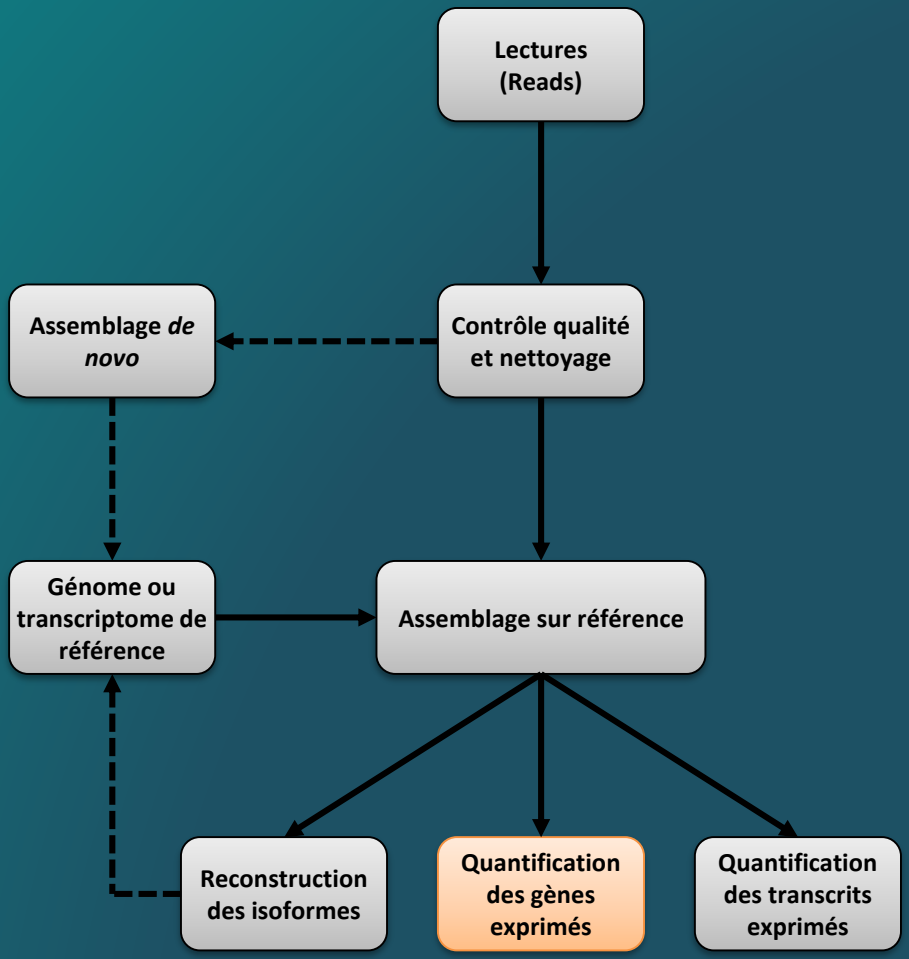
- **Comptage brut** sur les **gènes** ou les **exons** (pas les isoformes) :
  - packages DE (**htseq-count**)

- **Estimation** de l'**abondance** des **transcrits reconstruits** (isoformes) :
  - **StringTie**

- Dépend des données disponibles

Seqname	Source	Feature	Start	End
22	StringTie	transcript	77	4900
22	StringTie	exon	77	308
22	StringTie	exon	1386	1565
22	StringTie	transcript	5714	9036
22	StringTie	exon	5714	5869
22	StringTie	exon	6005	6204
22	StringTie	transcript	3511	3780
22	StringTie	exon	3511	3780

Workflow d'analyse RNA-Seq

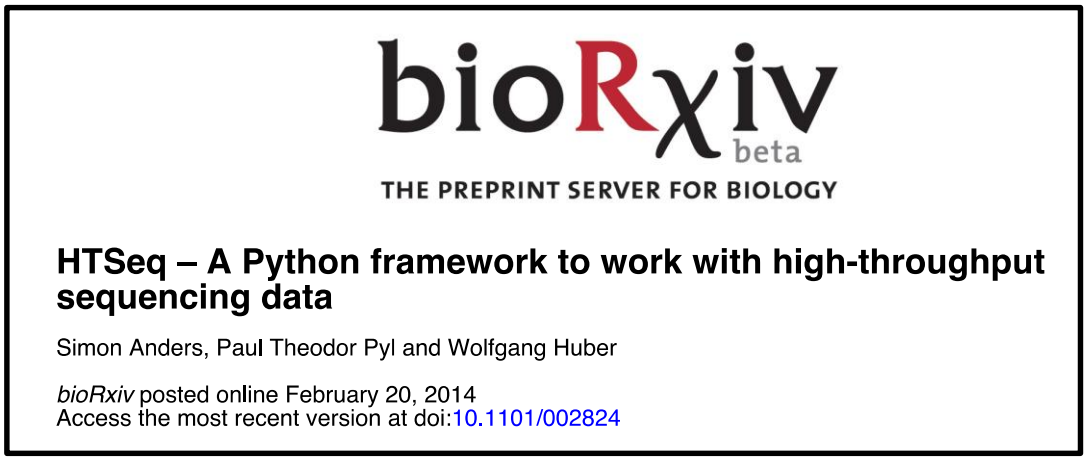


Comptage des gènes et exons

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

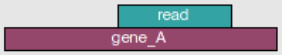
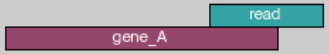
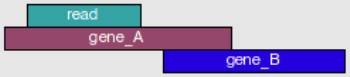
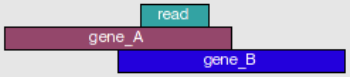
HTSeq – A Python framework to work with high-throughput sequencing data

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 ficihier d'entrée
  - **__alignement_not_unique** : lectures avec alignement multiple (BAM)

ENSDARG00000000183	514
ENSDARG00000000212	4074
ENSDARG00000000229	652
ENSDARG00000000568	609
ENSDARG00000000853	1641
ENSDARG00000001057	2496
ENSDARG00000001734	289
ENSDARG00000001818	231
ENSDARG00000002002	53
ENSDARG00000002192	14
ENSDARG00000002215	4093
ENSDARG00000002293	0

__no_feature	188766
__ambiguous	13165
__too_low_aQual	0
__not_aligned	4880
__alignment_not_unique	102545

HTSeq-count

TP : Quantification des gènes

- Produire la **table de comptage** des **gènes**, en **mode union**, à partir de **l'alignement épissé** en tenant compte du **sens** des gènes

HTSeq-count

TP : Quantification des gènes

htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version 0.9.1) Options

Aligned SAM/BAM File

30: HISAT2 on data 3, data 4, and others

GFF File

3: Danio_rerio_chr22.Zv9.62.gtf

Mode

Union

Mode to handle reads overlapping more than one feature. (--mode)

Stranded

No

Feature type

exon

Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for RNA-Seq and Ensembl GTF files, is exon. (--type)

ID Attribute

gene_id

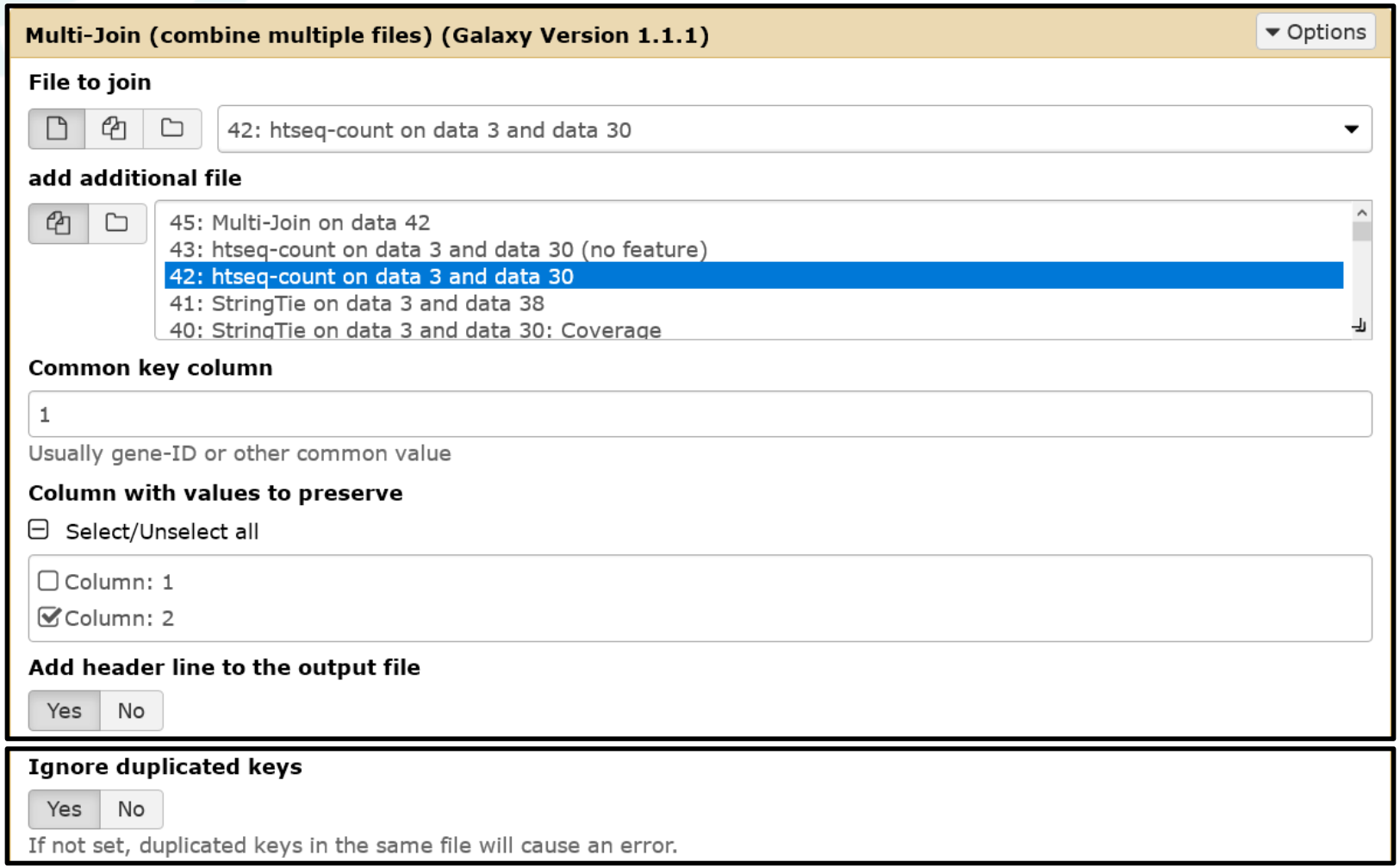
HTSeq-count

TP : Grouper les comptages de plusieurs analyses

- Produire la **table de comptage** des **gènes** avec **une colonne par expérience**

HTSeq-count

TP : Grouper les comptages de plusieurs analyses

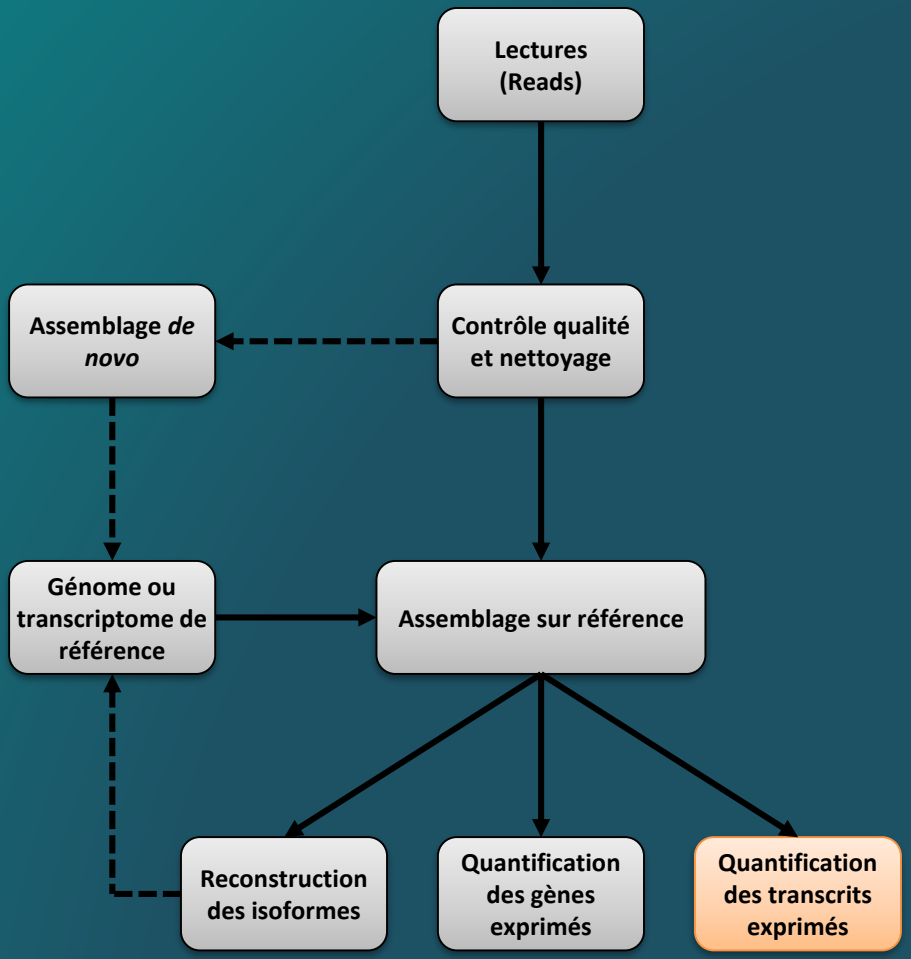


HTSeq-count

TP : Grouper les quantifications dans un fichier

1	2	3
key	dataset_37255_V2	dataset_37255_V2
ENSDARG00000000183	514	514
ENSDARG00000000212	4074	4074
ENSDARG00000000229	652	652
ENSDARG00000000568	609	609
ENSDARG00000000853	1641	1641
ENSDARG00000001057	2496	2496
ENSDARG00000001734	289	289
ENSDARG00000001818	231	231
ENSDARG00000002002	53	53
ENSDARG00000002192	14	14
ENSDARG00000002215	4093	4093
ENSDARG00000002293	0	0

Workflow d'analyse RNA-Seq



StringTie

- **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

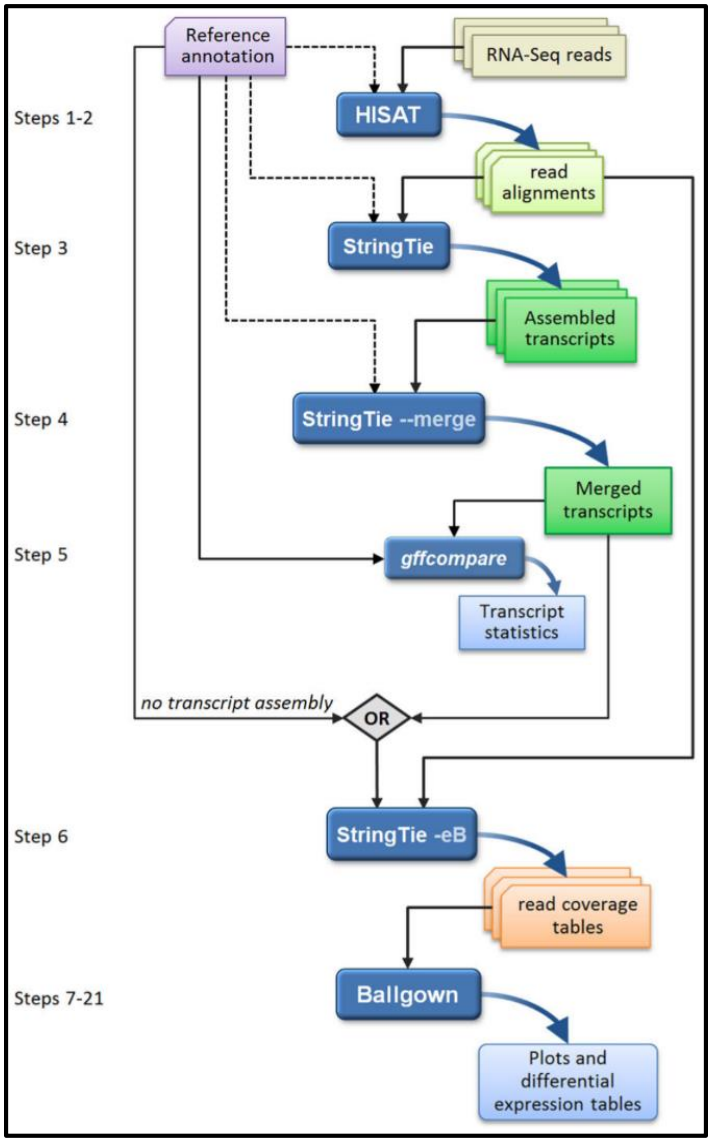
Nat Biotechnol. 2015 March ; 33(3): 290–295. doi:10.1038/nbt.3122.

StringTie enables improved reconstruction of a transcriptome from RNA-seq reads

Mihaela Pertea[1,2], **Geo M Pertea**[1,2], **Corina M Antonescu**[1,2], **Tsung-Cheng Chang**[3,4], **Joshua T Mendell**[3,4,5], and **Steven L Salzberg**[1,2,6,7]

https://ccb.jhu.edu/software/stringtie/

HISAT, StringTie & Ballgown : Pipeline

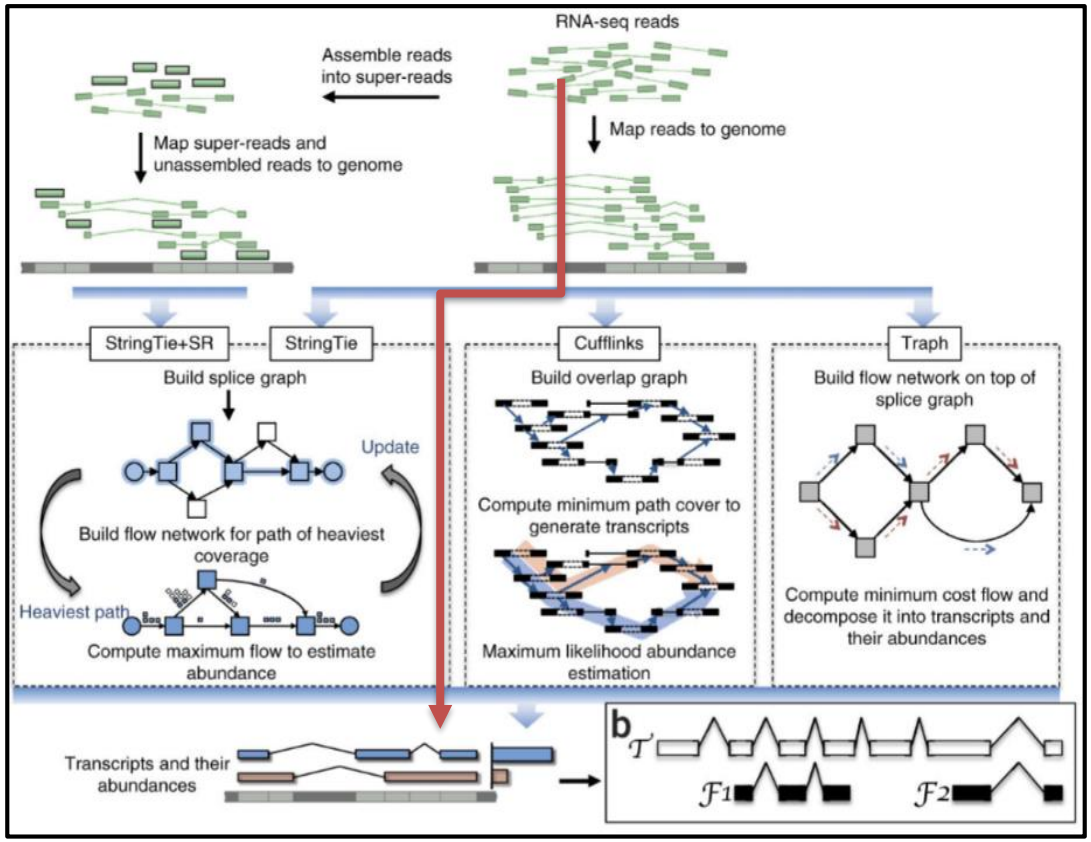


Pertea et al. 2016 Figure 1.

StringTie avec référence

Principes

- **Assignation des lectures** à un transcrit



Pertea et al. 2015 Figure 1.a

StringTie avec référence

RPKM / FPKM

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

- **RPKM** :
  - **R**eads **P**er **K**ilobase of exon per **M**illion fragments mapped
    - nombre de lectures alignées
    - nombre total de lectures de la librairie
    - taille des exons du gène en paire de bases

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

Mortazavi et al. Nature Methods 2008

StringTie avec référence

Quantification des transcrits de référence

- En **entrée** :
  - **lectures** (**sam/bam**)
  - **annotations** (**gtf**)
  - **utilisation** de l'annotation de **référence uniquement**

StringTie avec référence

TP : Quantifier les transcrits **annotés**

StringTie transcript assembly and quantification (Galaxy Version 1.2.3) ▾ Options

Mapped reads to assemble transcripts from

30: HISAT2 on data 3, data 4, and others

Use GFF file to guide assembly

Use GFF

Reference annotation to use for guiding the assembly process

3: Danio_rerio_chr22.Zv9.62.gtf

(-G)

Perform abundance estimation only of input transcripts

Yes | No

(-e)

Options

Specify advanced options

Additional gene abundance estimation output file

Yes | No

(-A)

StringTie avec référence

TP : Quantifier les transcrits **annotés**

- En **sortie** :
  - **assembled transcripts** (**gtf**), annotation et abondance des transcrits et exons
  - **coverage** (**gff3**), annotation des transcrits retrouvés couverts
  - **gene abundance estimates** (**tab**), abundance des transcrits

StringTie avec référence

Description de la sortie **assembled transcripts**

- **Format GTF** avec coordonnées et informations par :
  - **transcript** (les **isoforms**)
  - **Exon**
- **Expression :**
  - **cov** : The average per-base coverage for the transcript or exon
  - **FPKM** : Fragments per kilobase of transcript per million read pairs
  - **TPM** : Transcripts per million

Seqname	Source	Feature	Start	End	Score	Strand	Frame
22	StringTie	transcript	78	4766	1000	-	.
22	StringTie	exon	78	308	1000	-	.
22	StringTie	exon	1386	1565	1000	-	.

Attributes
gene_id "ENSDARG00000089609"; transcript_id "ENSDART00000131195"; ref_gene_name "CABZ01094378.1"; cov "3.858156"; FPKM "43.400478"; TPM "88.519356";
gene_id "ENSDARG00000089609"; transcript_id "ENSDART00000131195"; exon_number "1"; ref_gene_name "CABZ01094378.1"; cov "2.783550";
gene_id "ENSDARG00000089609"; transcript_id "ENSDART00000131195"; exon_number "2"; ref_gene_name "CABZ01094378.1"; cov "4.088889";

http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

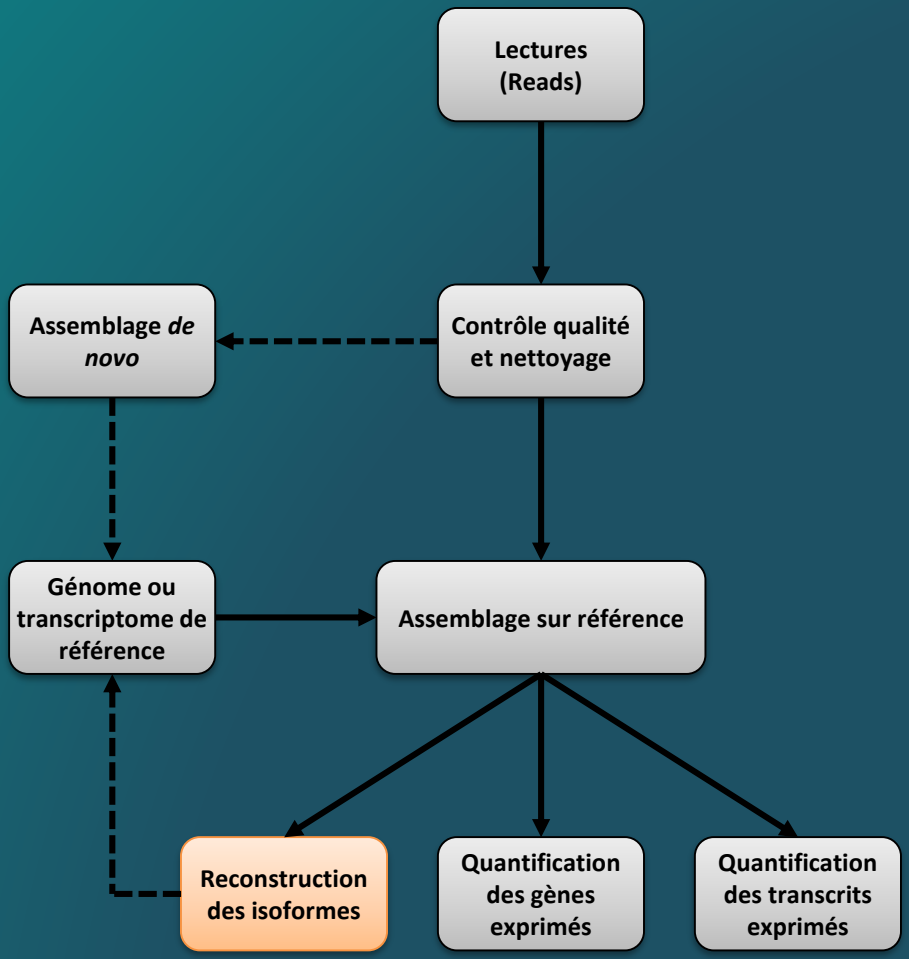
StringTie avec référence

Description du format tabulé d'abondance

- Informations similaires au GTF

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
Gene ID	Gene Name	Reference	Strand	Start	End	Coverage	FPKM	TPM
ENSDARG00000089609	CABZ01094378.1	22	-	78	4766	3.858156	43.400478	88.519356
ENSDARG00000088231	CELSR2	22	-	11787	24372	46.957535	528.226318	1077.367310
ENSDARG00000046004	capzb	22	-	26943	41610	406.710175	4575.091309	9331.330078
ENSDARG00000086848	atad3b	22	-	47436	49330	56.332581	633.686340	1292.462891
ENSDARG00000090462	MRPL20	22	-	51445	53106	82.068466	923.189880	1882.932617
ENSDARG00000075768	sdhb	22	+	53663	58397	238.769730	2685.925781	5478.198242
ENSDARG00000005879	zgc:112334	22	-	53931	237175	3.530009	41.883244	85.424820
ENSDARG00000059367	mfap2	22	+	62540	66555	489.302704	5504.176270	11226.285156
ENSDARG00000090150	CABZ01072698.1	22	-	66607	72822	43.454659	488.822357	996.999268
ENSDARG00000088201	CABZ01072699.2	22	-	72952	75837	7.594595	85.431755	174.246109
ENSDARG00000086607	CABZ01072699.1	22	-	77426	78210	5.729544	64.451767	131.455429
ENSDARG00000031898	NBL1	22	-	99197	107460	6.723083	75.628105	154.250626
ENSDARG00000087545	fbxo42	22	+	120113	121585	40.852818	459.554199	937.304016
ENSDARG00000087401	SLC25A34	22	+	125533	133222	1.351430	46.514229	94.870140

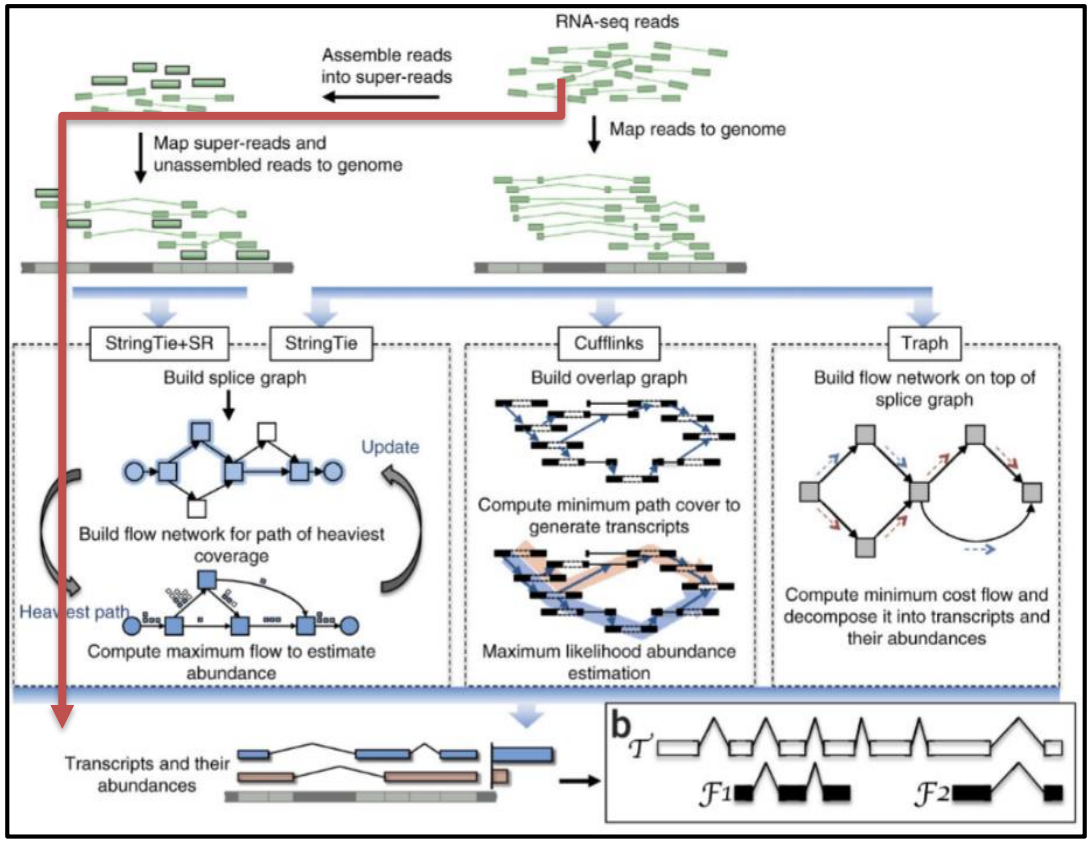
Workflow d'analyse RNA-Seq



StringTie pour une nouvelle annotation

Principes

- Ajout de la **reconstruction** de **nouveaux transcrits**

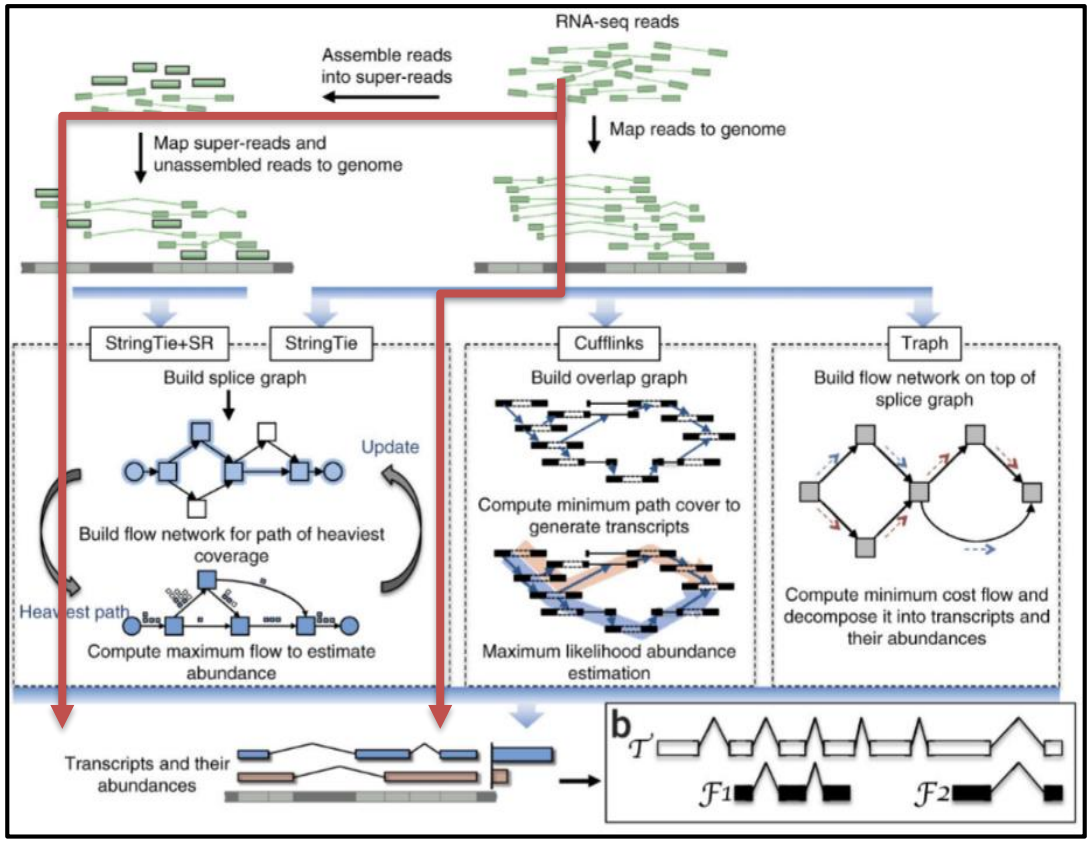


Pertea et al. 2015 Figure 1.a

StringTie pour enrichir l'annotation

Principes

- Ajout de la **reconstruction** de **nouveaux transcrits**



Pertea et al. 2015 Figure 1.a

StringTie pour enrichir l'annotation

TP : Recherche et quantification des nouveaux **transcrits**

- En **entrée** :
  - **lectures** (**.sam/.bam**)
  - **annotations** (**.gtf**)
  - **enrichir** l'annotation par l'**assemblage de transcrits**

StringTie pour enrichir l'annotation

TP : **Recherche** et quantification des nouveaux **transcrits**

StringTie transcript assembly and quantification (Galaxy Version 1.2.3) Options

Mapped reads to assemble transcripts from

30: HISAT2 on data 3, data 4, and others

Use GFF file to guide assembly

Use GFF

Reference annotation to use for guiding the assembly process

3: Danio_rerio_chr22.Zv9.62.gtf

(-G)

Perform abundance estimation only of input transcripts

Yes | No

(-e)

Options

Specify advanced options

Additional gene abundance estimation output file

Yes | No

(-A)

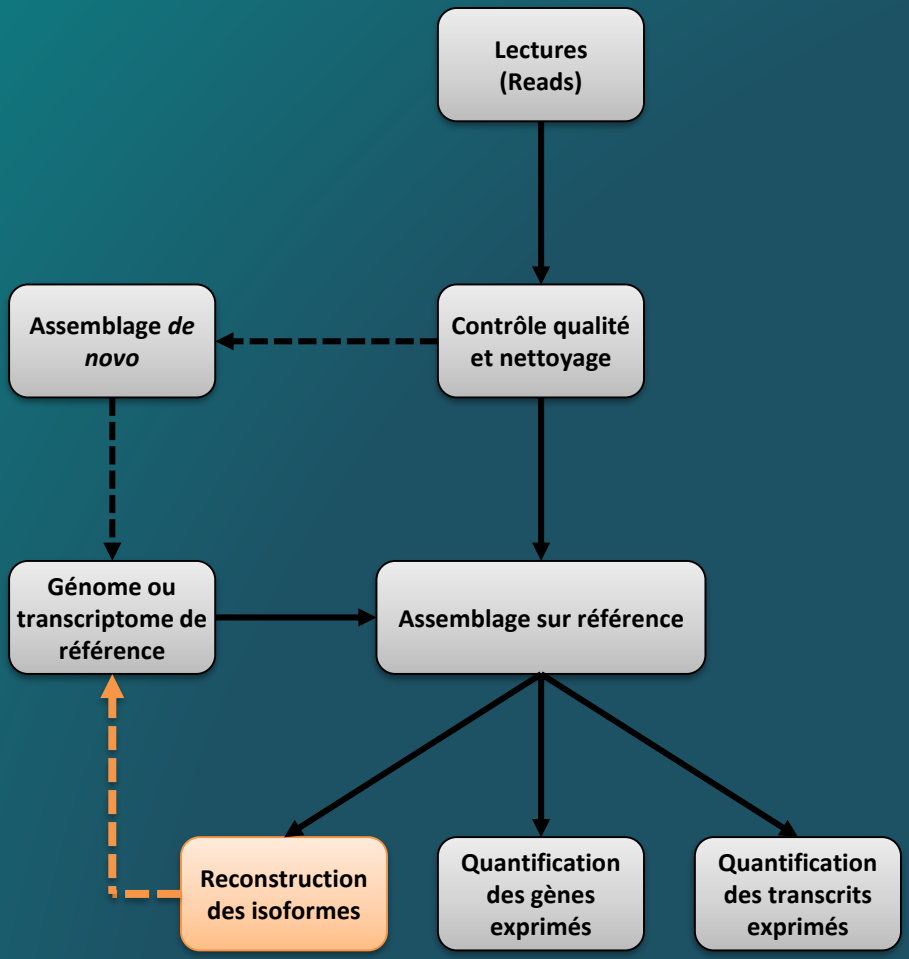
StringTie pour enrichir l'annotation

TP : **Recherche** et quantification des nouveaux **transcrits**

- En **sortie** :
  - **assembled transcripts** (**gtf**), annotation et abondance des transcrits et exons
  - **coverage** (**gff3**), annotation des transcrits retrouvés couverts
  - **gene abundance estimates** (**tab**), abundance des transcrits

  - une impression de déjà vu ?

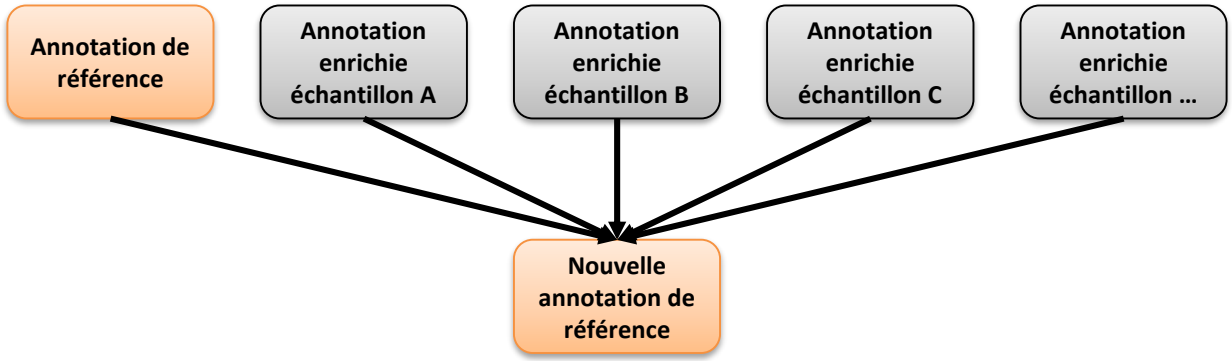
Workflow d'analyse RNA-Seq

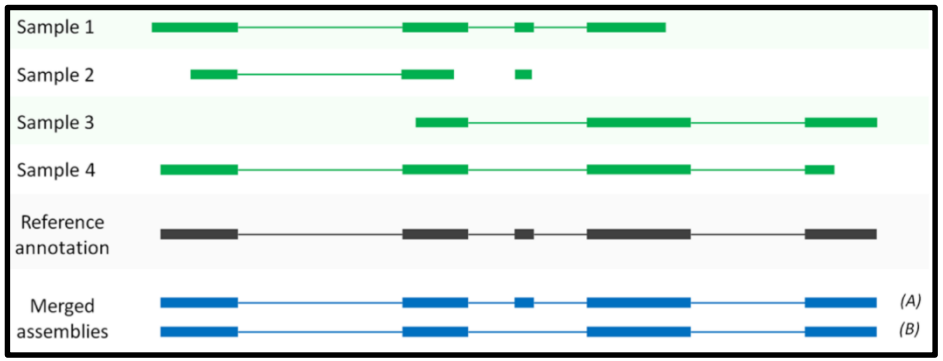


StringTie merge

Principes

- **Regrouper** l'**annotation initiale** et celle des **nouveaux transcrits**
  - traiter les transcrits conditions spécifiques
  - obtenir **une seule annotation** de **référence**





Pertea et al. 2016 Figure 2.

StringTie merge

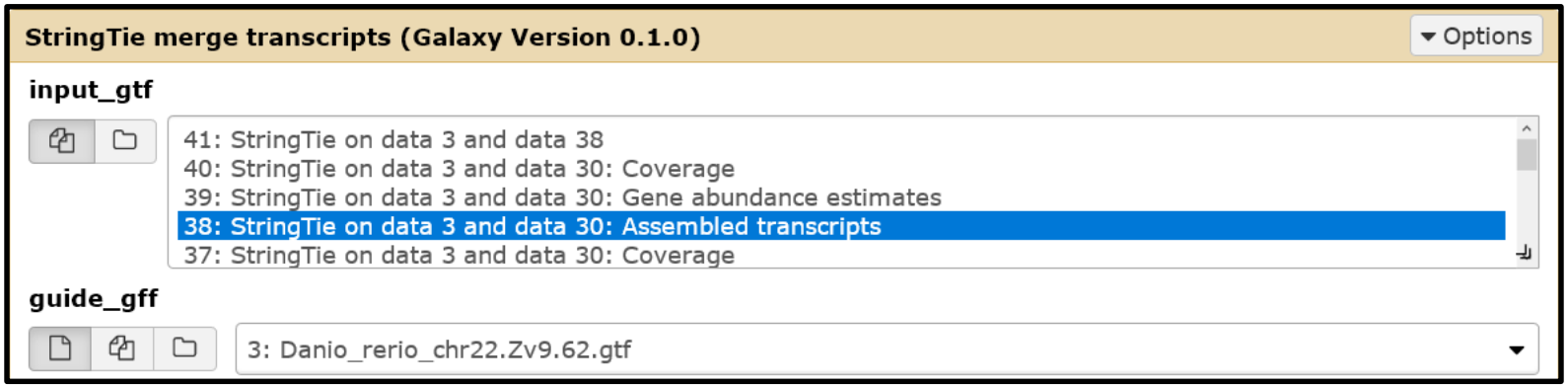
TP : **Fusionner** la **nouvelle annotation avec** l'**initiale**

- Objectif :
  - **fusionner** l'**annotation initiale** avec celle **reconstruite** par **StringTie**

- En **entrée** :
  - **annotations initiale** (**.gtf**)
  - [annotation initiale +] **nouveaux transcrits** (**.gtf**)

StringTie merge

TP : Fusionner la nouvelle annotation avec l'initiale



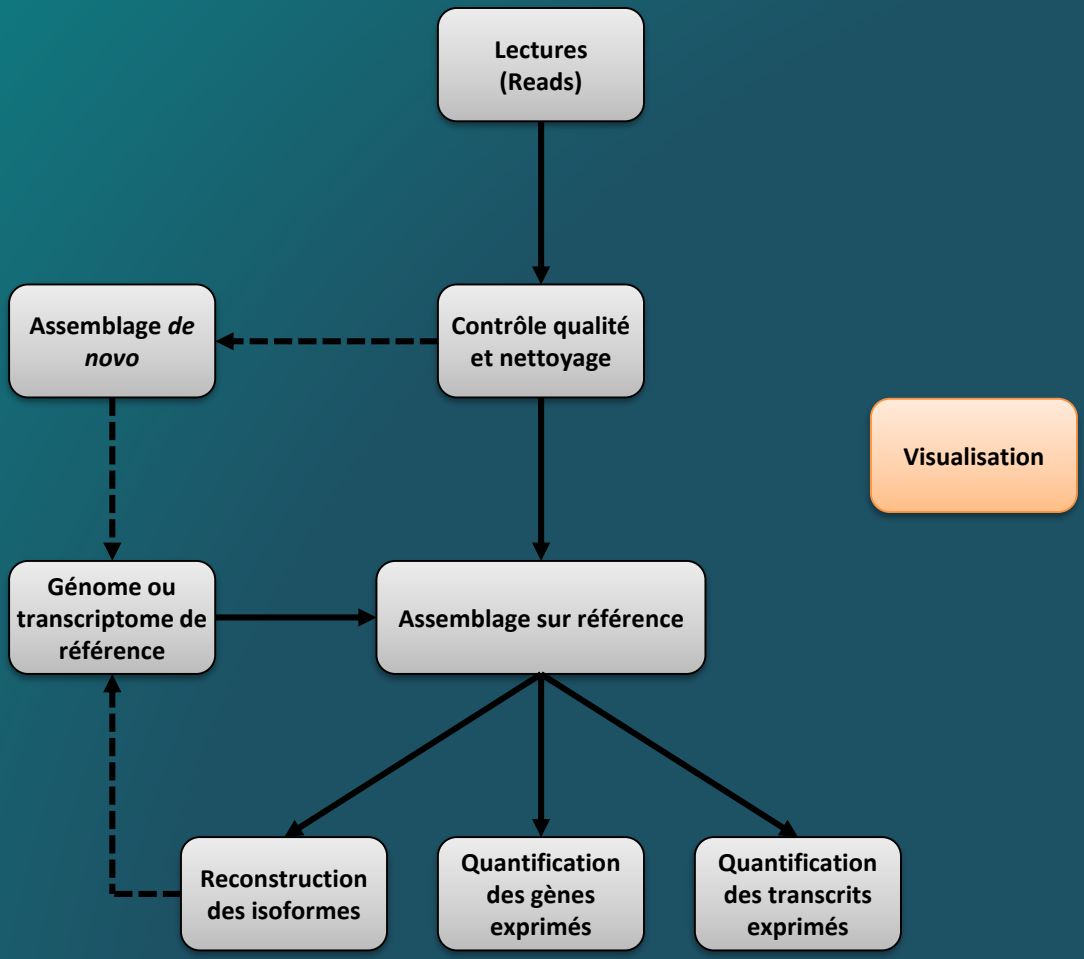
StringTie merge

TP : **Fusionner** la **nouvelle annotation avec** l'**initiale**

- En **sortie**:
  - **nouvelle annotation** (**.gtf**)
    - union des annotations des transcrits

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
22	StringTie	transcript	78	4766	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	exon	78	308	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; exon_number "1"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	exon	1386	1565	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; exon_number "2"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	exon	1644	1813	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; exon_number "3"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	exon	4213	4340	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; exon_number "4"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	exon	4414	4607	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; exon_number "5"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	exon	4683	4766	1000	-	.	gene_id "MSTRG.1"; transcript_id "ENSDART00000131195"; exon_number "6"; gene_name "CABZ01094378.1"; ref_gene_id "ENSDARG00000089609";
22	StringTie	transcript	3511	3780	1000	.	.	gene_id "MSTRG.2"; transcript_id "MSTRG.2.1";
22	StringTie	exon	3511	3780	1000	.	.	gene_id "MSTRG.2"; transcript_id "MSTRG.2.1"; exon_number "1";
22	StringTie	transcript	5714	9036	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1";
22	StringTie	exon	5714	5869	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1"; exon_number "1";
22	StringTie	exon	6005	6204	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1"; exon_number "2";
22	StringTie	exon	6709	6867	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1"; exon_number "3";
22	StringTie	exon	7023	7109	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1"; exon_number "4";
22	StringTie	exon	7194	7284	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1"; exon_number "5";
22	StringTie	exon	8816	9036	1000	-	.	gene_id "MSTRG.3"; transcript_id "MSTRG.3.1"; exon_number "6";

Workflow d'analyse RNA-Seq



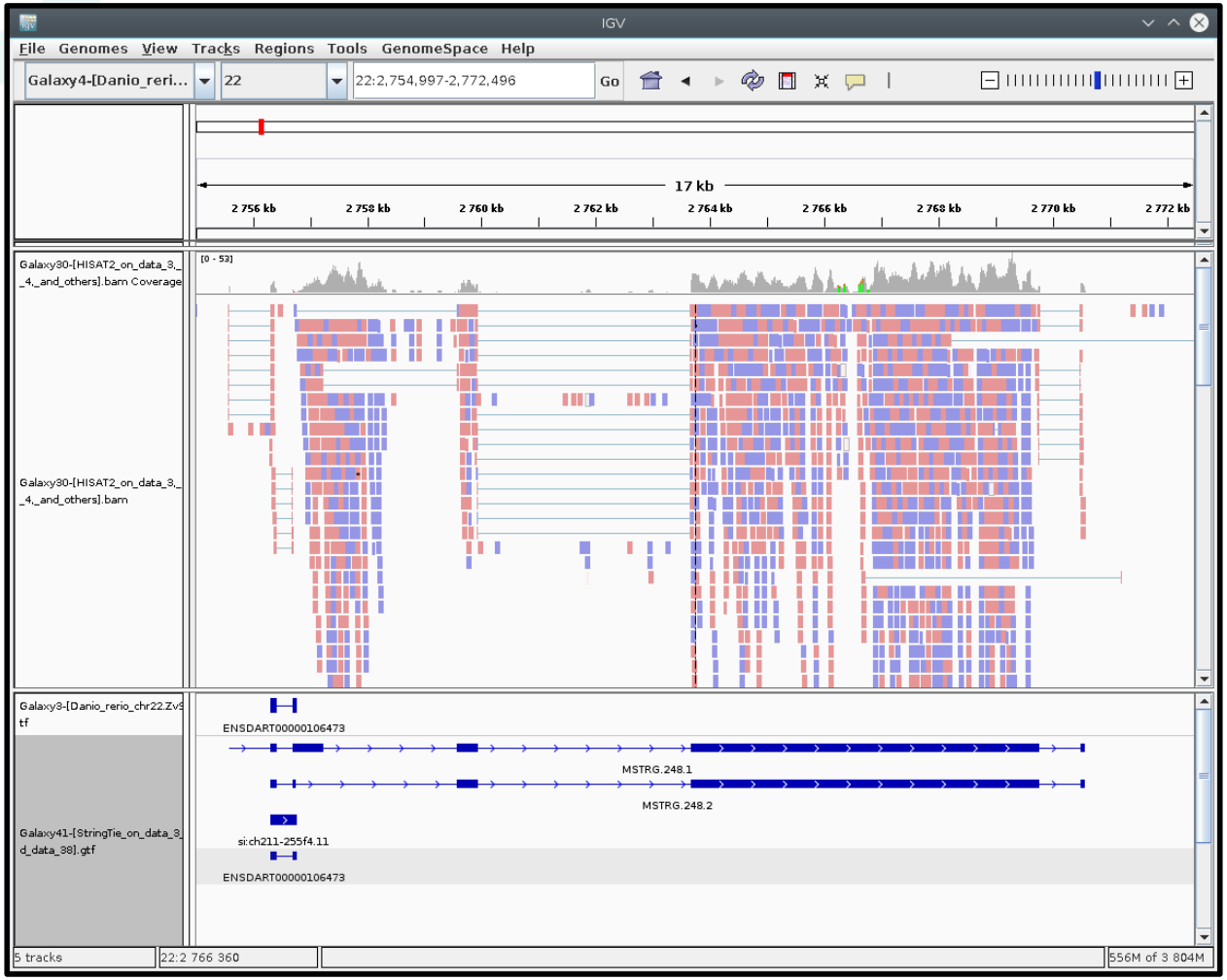
StringTie merge - IGV

TP : Visualiser les différences avec l'annotation initiale

- **Séquence** (fasta) et **annotation** (gtf)
- **Alignements** (bam), l'**index** (bai) doit être au même endroit.
- **transcripts** (.gtf)
- **Regarder** de nouveau les **régions** :
  - **22:585,838-603,303** : rétention d'intron ?
  - **22:669,413-678,616** : frontières d'exons ? UTR ?
  - **22:2,754,99-2,772,496** : nouveau transcrit ? intérêt des reads orientées ?

StringTie merge - IGV

TP : Visualiser les différences avec l'annotation initiale



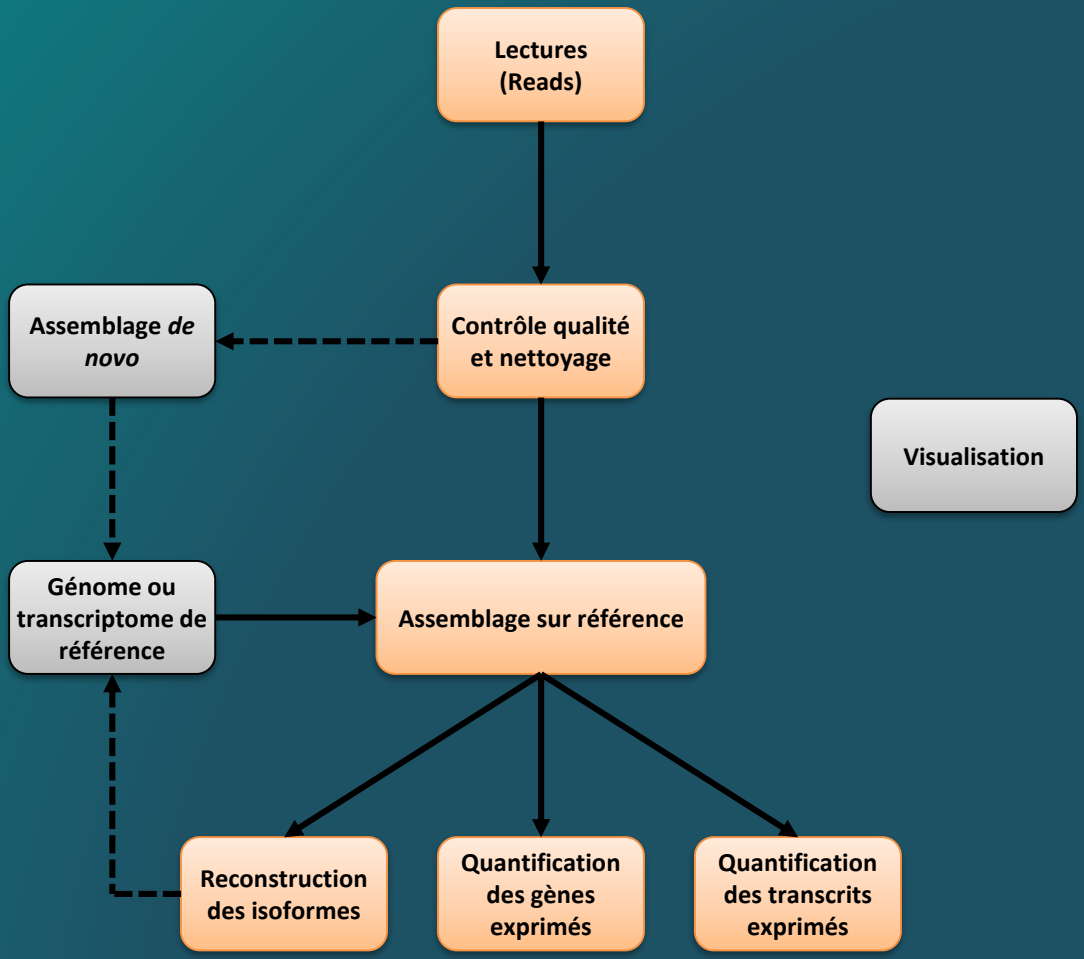
Quantification

Conclusion

- Plusieurs façons de faire : comptage gène, comptage transcrits

- Deux gamme outils :
  - **Htseq-count** : analyse avec DESeq, edgeR, …
  - **StringTie/Ballgown** : tout intégré

Workflow d'analyse RNA-Seq



Galaxy – création d'un *workflow*

TP : Extraire et exécuter un *worflow*

- **Créer un *worflow*** à partir d'un **historique de test**
- **Visualiser, modifier et partager un *worflow***
- **Exécuter** un *workflow* sur un nouveau **jeu de données**

Conclusion générale

- **Workflow**
- **Choix** des **outils dépendent** des **données disponibles** et de la **question biologique**
- Tous les **outils** sont **dispo** sur **Migale et Galaxy**

Liens utiles

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

Remerciements

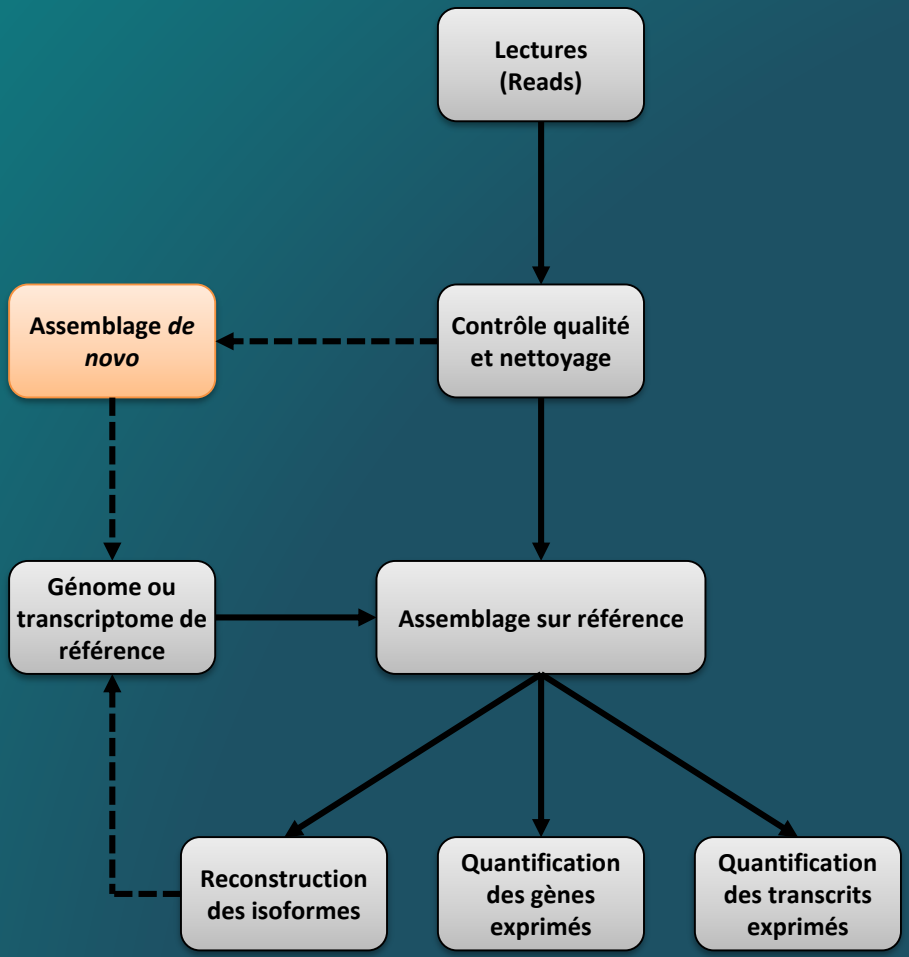
- Le groupe de travail **« Planification d'experiences et RNA-seq »** du **PEPI IBIS**

Pour aller plus loin…

cyprien.guerin @inrae.fr
valentin.loux @inrae.fr

INRAE
la science pour la vie, l'humain, la terre

Workflow d'analyse RNA-Seq



Assemblage *de novo*

Le problème général

- **Semblable** à un **puzzle** :
  - **millions** de **pièces**
  - **sans** l'**image d'origine**
  - **avec** des **pièces** dans les **deux sens**
  - **erreurs** de **séquençage** : les **pièces** ne s'assemblent pas
  - **couverture + biais de séquençage** : des **parties** du **puzzle** sont **absentes**

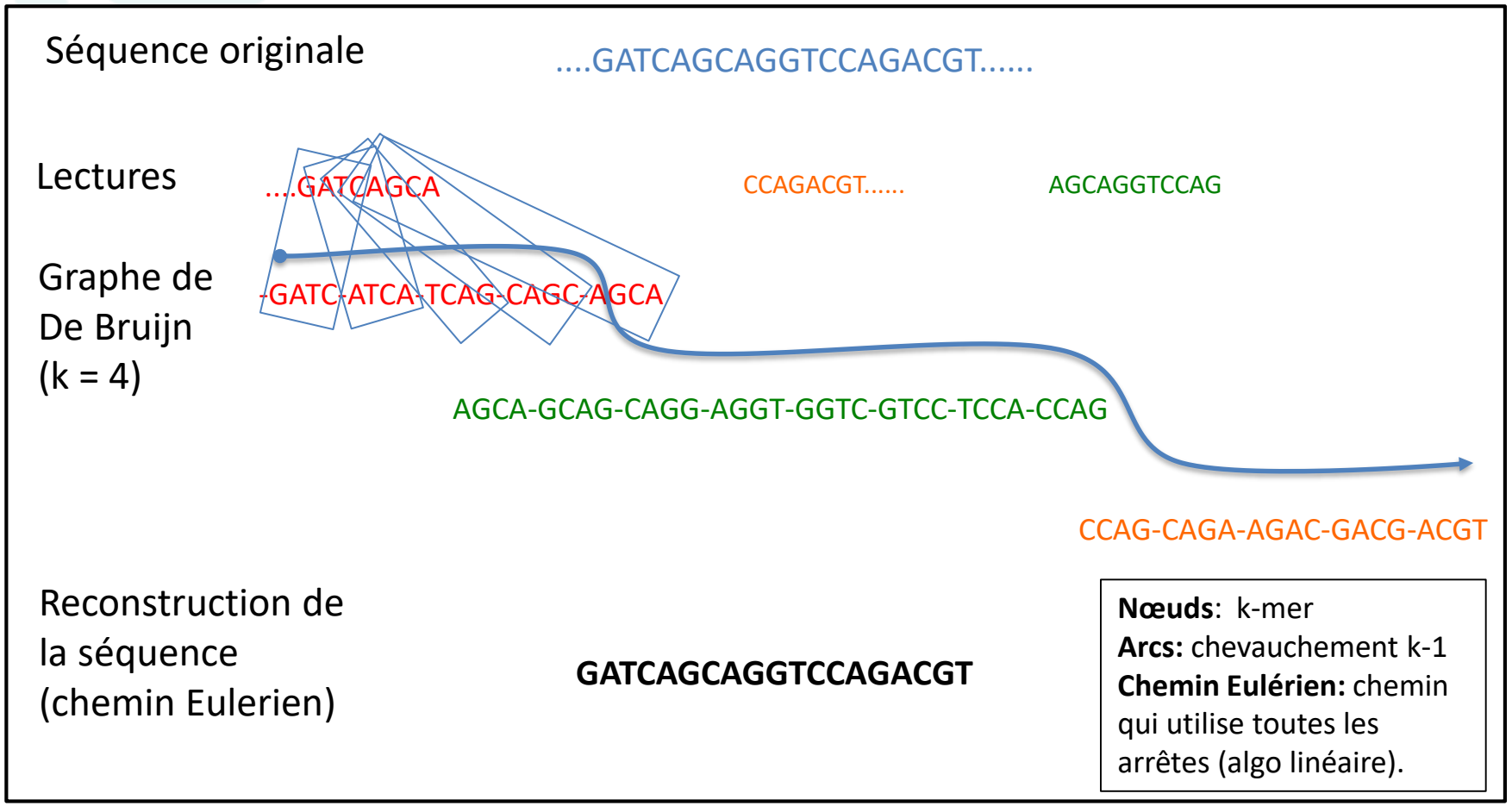
Assemblage *de novo*

Algorithmes d'assemblage

- **Tous** les **algorithmes** sont **basés** sur le **chevauchement** des **lectures :**
  - **Overlap Layout Consensus** (Chemin hamiltonien dans un graphe)
  - **Graphes de De Bruijn** (Chemin Eulerien dans un graphe)

Assemblage *de novo*

Chemin Eulerien dans un Graphe de De Bruijn



Assemblage *de novo*

En pratique

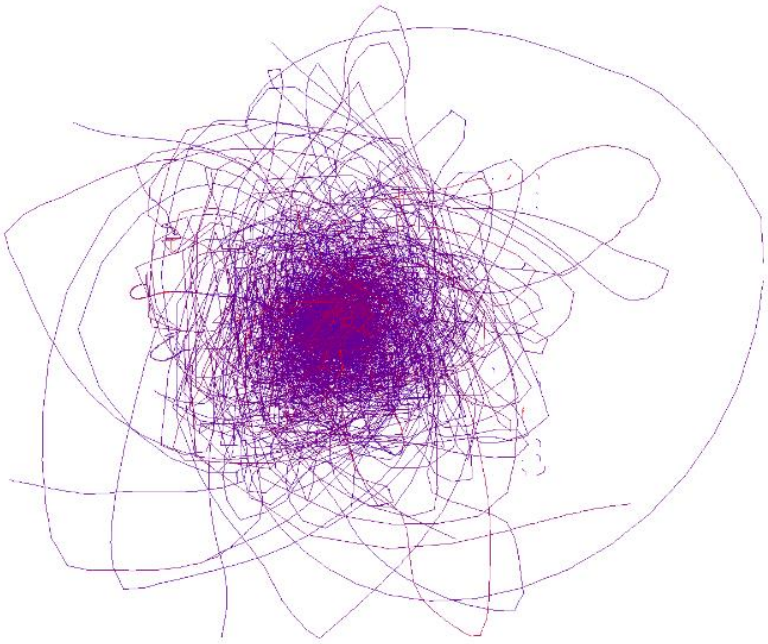


Figure 2.9: Graphical representation of the de Bruijn of the *Streptococcus suis* genome

In this representation, node sequences are represented as curves, which connect at their tips.

D. Zerbino, Genome assembly and comparison using de Bruijn graphs. 2009. Phd Thesis

Assemblage *de novo*

Limites de l'assemblage

- **Tous** les **outils** utilisent des **heuristique** pour résoudre les principaux problèmes :
  - **Gestion** de **répétitions**
  - **Nettoyage** des **données**
  - **Volume** de **données**

- Choix des **paramètres sensibles** (taille de k pour DBG)

Assemblage *de novo*

Assemblage transcriptome ≠ Assemblage génome

- **Assemblage** de **génome**
  - **couverture uniforme**
  - **double brin**
  - **un *contig*** (idéalement)

- **Assemblage** de **transcriptome**
  - **couverture hétérogène** sur une **échelle exponentiels**
  - **brin spécifique**
  - **autant** de ***contig*** que d'**isoformes** de gènes (idéalement)

Assemblage *de novo*

Conclusion

- L'**assemblage** de **transcriptome** reste un **problème ouvert** de **recherche** et d'**ingénierie**

- **Problèmes** et **heuristiques spécifiques** à ce type d'assemblage