2018

- Galaxy -

Formation à l'analyse de données RNA-seq

- EXERCICES -

Liens utiles

Données publiques :

ENA European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
http://www.ebi.ac.uk/ena/

e!Ensembl

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
http://www.ensembl.org/index.html

Logiciels utilisés :

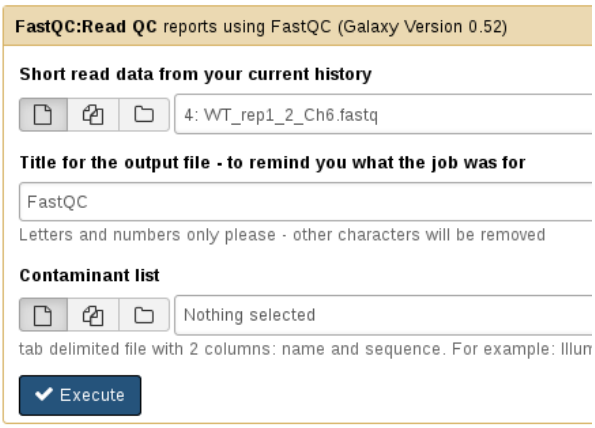
FastQC Report	**FastQC** aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis. *http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/*
cutadapt	**Cutadapt** Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads. Cutadapt helps to trim reads by finding the adapter or primer sequences in an error-tolerant way. It can also modify and filter reads in various ways. Adapter sequences can contain IUPAC wildcard characters. Also, paired-end reads and even colorspace data is supported. If you want, you can also just demultiplex your input data, without removing adapter sequences at all.
STAR	**STAR** is a Spliced Transcripts Alignment to a Reference. https://github.com/alexdobin/STAR
Cufflinks	**Cufflinks** assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols. http://cufflinks.cbcb.umd.edu/
SAMtools	**SAM** (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. http://samtools.sourceforge.net/

RSEM	RSEM: accurate quantification of gene and isoform expression from RNA-Seq data
Integrative Genomics Viewer	The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated datasets. It supports a wide variety of data types including sequence alignments, microarrays, and genomic annotations. http://www.broadinstitute.org/igv/
Bioconductor	**Bioconductor** provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. http://bioconductor.org/
R	**R** is a free software environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. http://www.r-project.org/

Objectifs:

Cette formation a pour objectif de vous aider à traiter les séquences issues des SGS (Second Generation Sequencing), en particulier les plates-formes Illumina HiSeq. Vous y découvrirez les nouveaux formats de séquences, les biais connus et mettrez en œuvre des logiciels d'alignement épissé sur génome de référence, la recherche de nouveaux gènes, de nouveaux transcrits et la quantification de l'expression de ces gènes et transcrits.

Exercice n°1: Analyse de la qualité des données avec FastQC

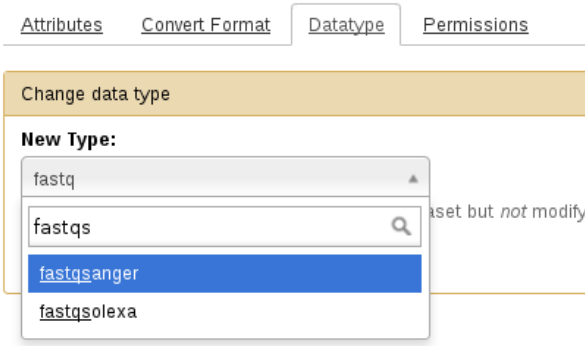
Recherchez l'outil « FastQC:Read QC reports using FastQC" à l'aide du champs "search" de Galaxy:	
Puis lancez cet outil, 4 fois, pour chacun des fichiers fastq. Renommez les fichiers obtenus : FastQC MT_rep1_1_Ch6.fastq FastQC MT_rep1_2_Ch6.fastq FastQC WT_rep1_1_Ch6.fastq FastQC WT_rep1_2_Ch6.fastq 	

1) Quelle est la longueur des lectures ? Est-ce la même que celle que vous avez obtenue à l'exercice précédent ?
2) La qualité du séquençage vous paraît-elle correcte ?
3) Regardez les résultats concernant les biais décrits lors du cours : lesquels retrouve-t-on ?

Exercice n°2: Nettoyage des données avec cutadapt et Sickle

1) Supprimez les adaptateurs à l'aide de cutadapt (outil "Remove adaptators with cutadapt").

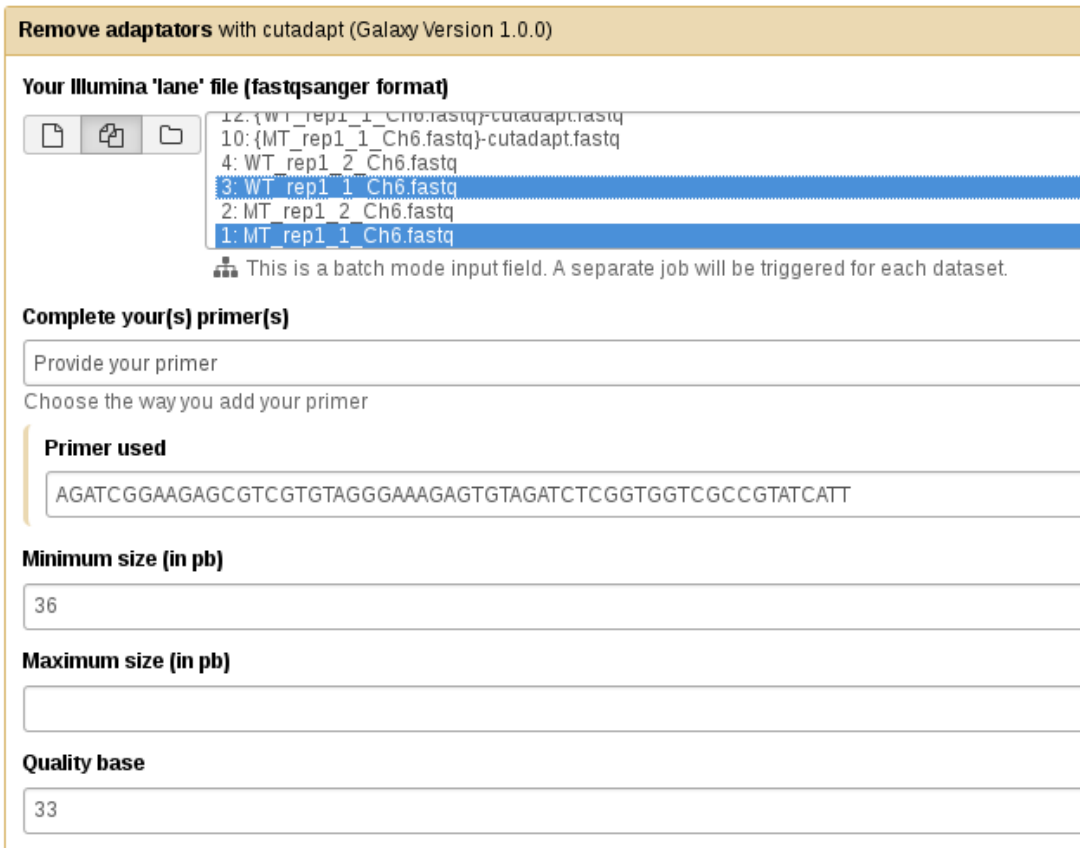
L'outil cutadapt ne reconnaît en entrée les fichiers fastq que si ces derniers sont au format "fastqsanger". Pour modifier le datatype de fastq à fastqsanger, veuillez cliquer sur l'icône  pour accéder au menu "datatype".



Spécifiez que la taille minimale des reads trimmées doit être de 36 bp.
Adaptateurs : -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

Il est possible de traiter plusieurs fichiers fastq en entrée de cutadapt afin de ne pas avoir à relancer l'outil 4 fois. Il s'agit du mode "batch". Pour cela, il suffit de cliquer sur l'icône "Multiple datasets"  .

L'outil cutadapt sera lancé une première fois pour les reads R1 pour enlever l'adaptateur AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC puis cutadpat sera lancé une seconde fois en R2 pour supprimer l'adapteur AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
De plus, spécifiez que la taille minimale des reads trimmées doit être de 36 bp.

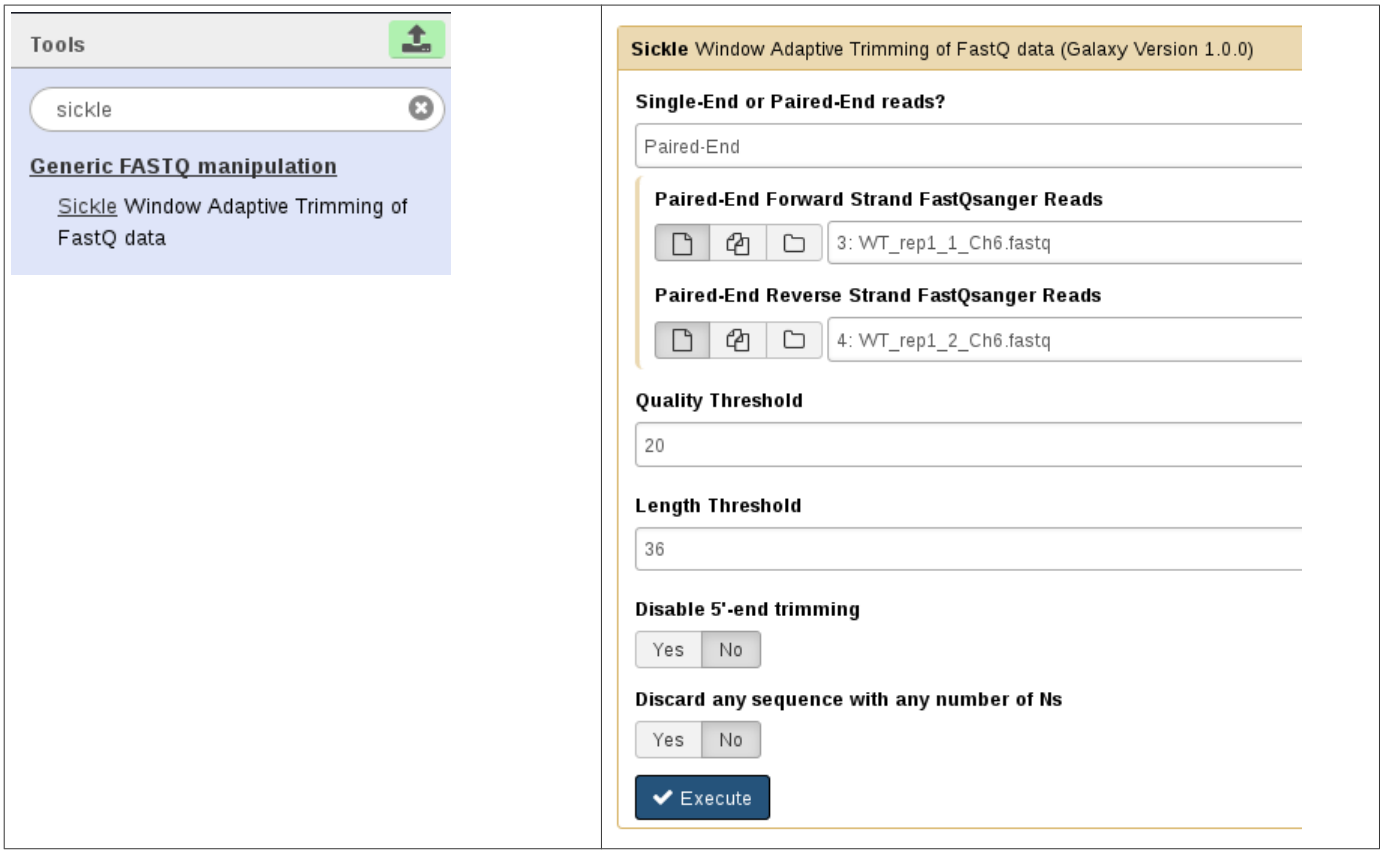


Chaque job génère deux fichiers sortants pour un fichier fastq entrant : un fichier fastq nettoyé {MT_rep1_1_Ch6.fastq}-cutadapt.fastq, et un fichier de log {MT_rep1_1_Ch6.fastq}-cutadapt.log

Pensez aussi à aller voir les log du logiciel.

2) Testez un autre type de nettoyage avec l'outil “Sickle Window Adaptive Trimming of FastQ data”, enlevez les reads trimmées lorsqu’elles sont plus courtes que 36 pbs, pour WT et pour MT.

Aidez-vous du manuel pour bien paramétrer Sickle: https://github.com/najoshi/sickle



Pensez à renommer vos datasets sortantes.

3) A quelle valeur de qualité sickle trimme-t-il par défaut ?
4) Relancez fastQC sur les résultats.
5) Quel est le nombre de séquences après nettoyage ?

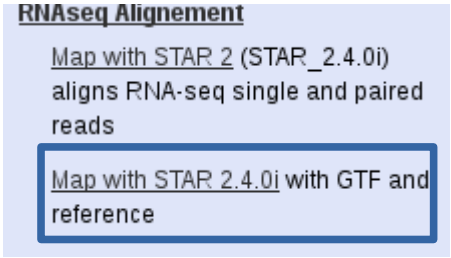
Exercice n°3: alignement/visualisation

1) Alignement avec STAR

L'outil “Map with STAR 2.4.0i with GTF and reference” indexe la référence, ou bien utilise les index déjà disponibles sur le serveur Genotoul, dans /bank/STARdb/, puis réalise les alignements épissés.

a) Recherchez le manuel de « rnaSTAR » sur internet.

b) Quelle version de STAR est utilisée par défaut ?



c) Chargez le fichier fasta de référence à l’aide de l'outil “Upload File from your computer” dans la section “Get data” de Galaxy. Voici l'URL d'accès au fichier :

“http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2018/data/reference/ITAG2.3_genomic_Ch6.fasta”

d) Renommez la dataset sortante en “ITAG2.3_genomic_Ch6.fasta”.

e) Lancez “Map with STAR 2.4.0i with GTF and reference” en indiquant les paramètres suivants:

Map with STAR 2.4.0i with GTF and reference (Galaxy Version 1.0.0) — Options

Paired or single reads

Paired reads

First input fastq gzipped file (read1.fastq.gz)

5: http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2018/data/reads/WT_rep1_1_Ch6.fastq

Second input fastq gzipped file (read2.fastq.gz)

6: http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2018/data/reads/WT_rep1_2_Ch6.fastq

compressed fastq file

Not compressed

fastq files are compressed or not

Threads number

8

alignIntronMin

20

alignIntronMax

1000000

outFilterMismatchNmax

10

Genotoul reference genome or your own fasta file

Your own fasta file

Your own reference genome

2: http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2015/data/reference/ITAG2.3_genomic_Ch6....

Your own GTF file

1: http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/2015/data/reference/ITAG_pre2.3_gene_mo...

Do you want to perform Cufflinks quantification after ?

Yes

Cufflinks needs strand information. This option add STAR options: --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonical --outFilterType BySJout

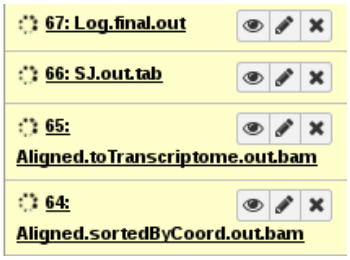
Transcriptome alignment

Yes

STAR will outputs alignments translated into transcript coordinates (for RSEM, eXpress, etc.), works if indexation has been performed with gtf.

Execute

f) Quels sont les fichiers de sortie ? Les renommer (WT_…)
g) Combien de reads sont alignés de façon unique et de façon multiple ? (voir Log.final.out)

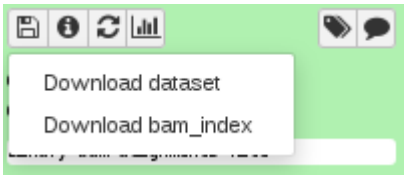


2) Alignement avec HISAT2

Lancez "**HISAT2** A fast and sensitive alignment program (Galaxy Version 2.0.5.2)" sur une paire de fastq. **HISAT2 remplace Tophat.**

3) Visualisation des résultats de STAR

Téléchargez sur votre ordinateur les fichiers de résultats de STAR (genome bam et SJ.out.tab) et le fichier d'indexation (bai)

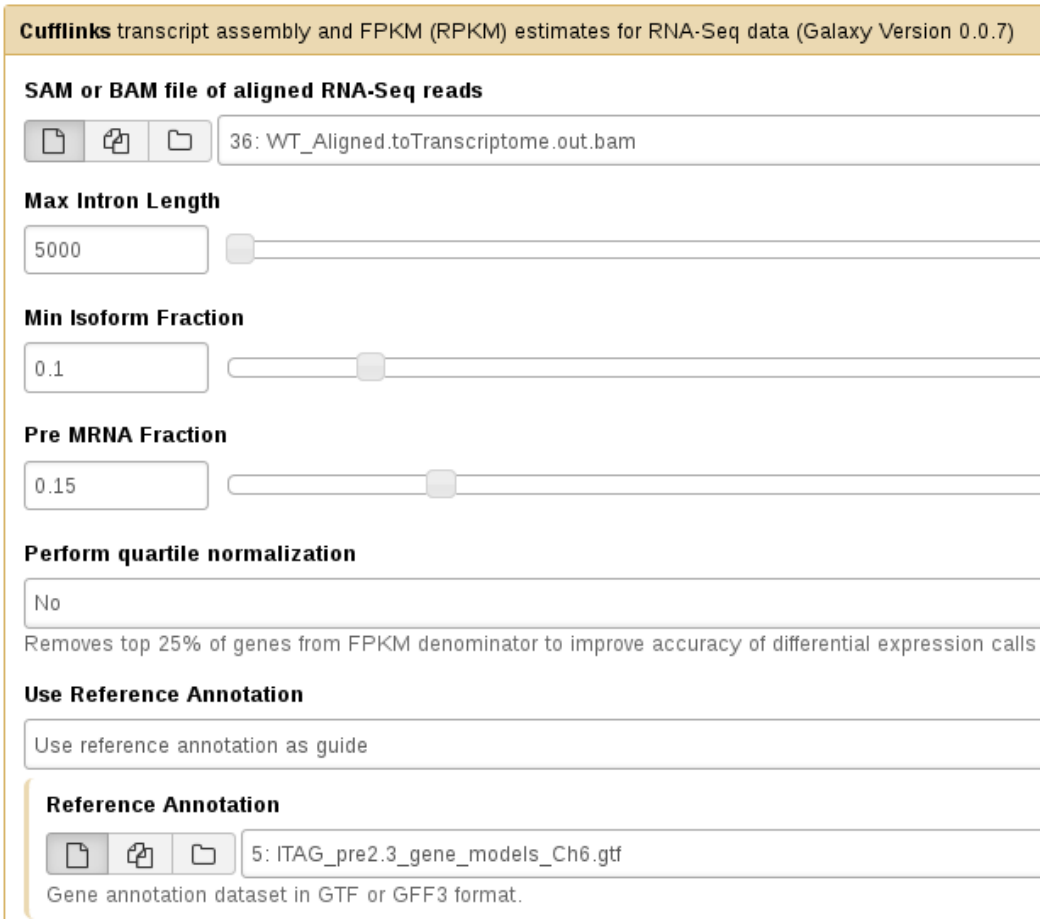


a) Utilisez IGV pour visualiser les résultats sur votre poste de travail.
b) Lancez IGV depuis « download » du site web de la formation (en bas de la page): http://www.broadinstitute.org/software/igv/download
a) Chargez les annotations (fichier gtf mis à disposition dans http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/2017/data/reference/)
b) Chargez les .bam et les .bai
c) Explorez l'interface, en utilisant le clic-droit (pour visualiser toutes les isoformes, les paires de lectures associées....)
d) Regardez les régions suivantes :

- SL2.40ch06:34,298,666-34,306,292
- SL2.40ch06:34,209,900-34,260,000
- SL2.40ch06:2,786,806-2,807,064
- SL2.40ch06:38,479,173-38,483,269
- SL2.40ch06:10,694,176-10,704,838
- Solyc06g009140.2.1
- SL2.40ch06:7,973,823-7,977,708

NB. Pour des jeux de données de tailles plus conséquentes, pensez à trier et indexer vos fichiers GTF avec les outils igvtools sort et igvtools index.

Exercice 4 : Recherche de nouveaux transcrits et comparaison des gtf

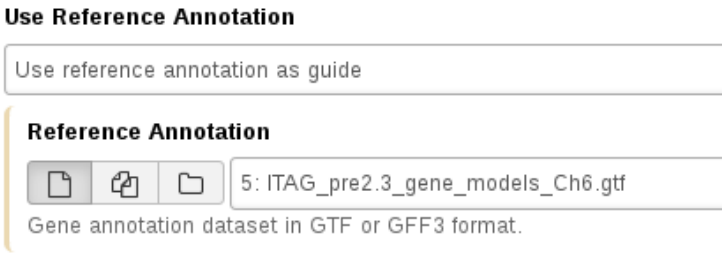


1) Dans Cufflinks, que signifie RABT ?

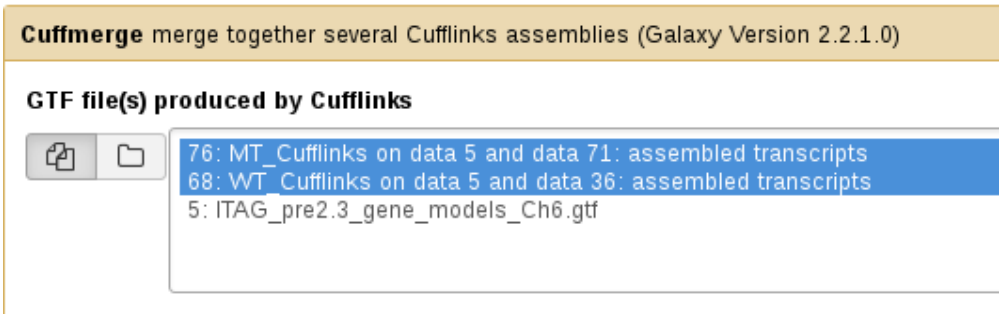
2) Quelle version de cufflinks est disponible sur genotoul ? Et sur internet ?

3) Lancez “Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data “ sur chaque condition avec les options suivantes :

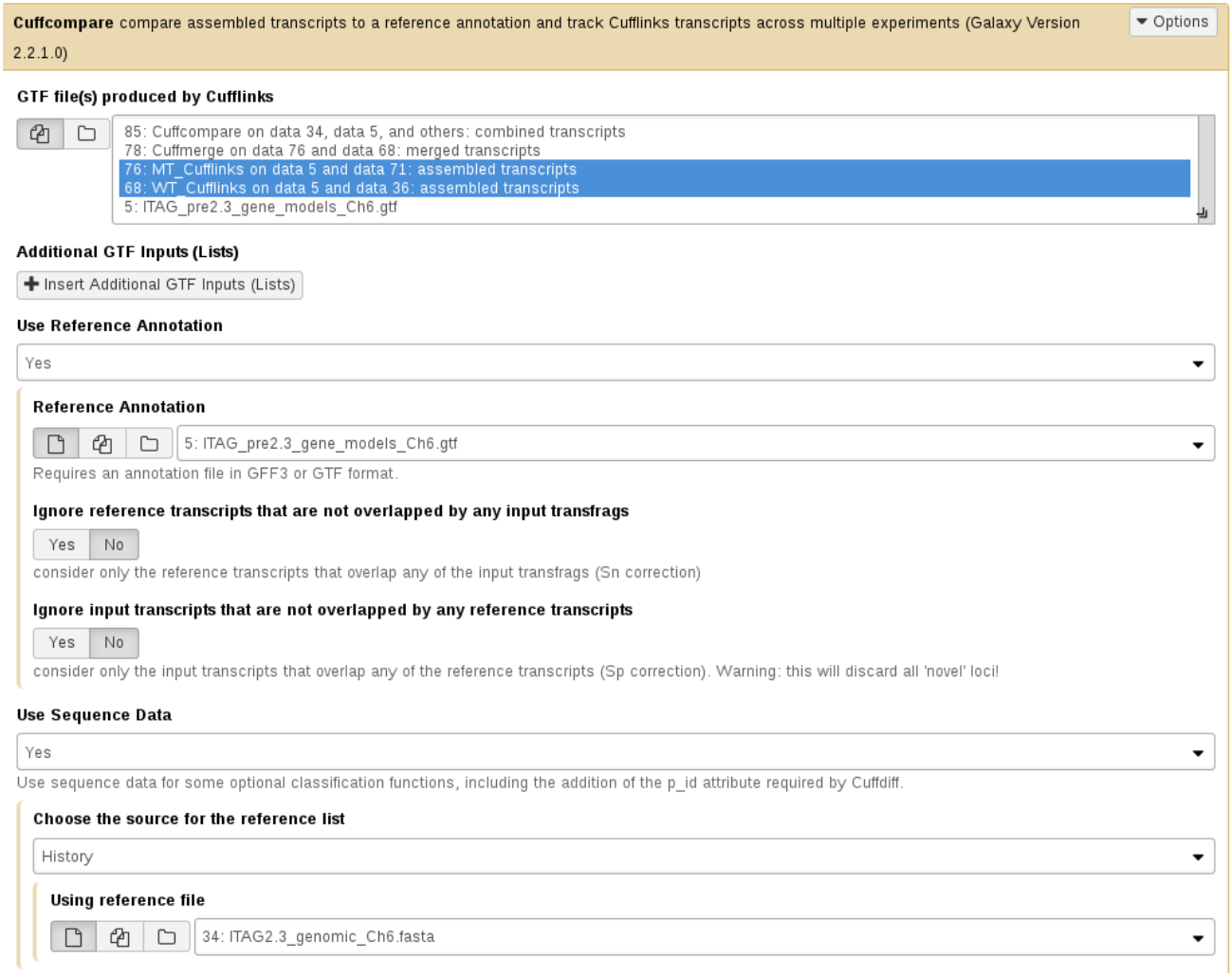
   - -g pour faire un assemblage RABT
   - max-intron-length : 5000



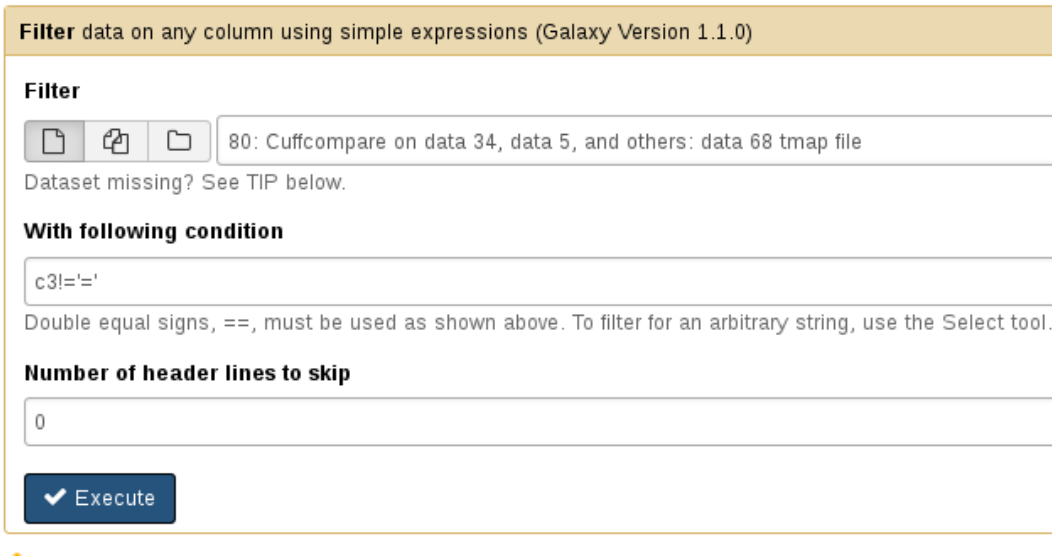
4) Utilisez cuffmerge pour fusionner les 2 nouveaux gtf.



- Combien de transcrits obtenez-vous ? Comparer ce résultat au comptage de l'exercice 6.
- L'outil “Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments “ permet d'obtenir une comparaison entre deux fichiers d'annotation.



- Extrayez du fichier tmap les lignes dont la troisième colonne n'est pas **'='** et allez voir pour chaque type de transfrag un exemple issu du nouveau gtf (merged.gtf) dans IGV, puis retournez voir les zones citées dans l'exercice 5.

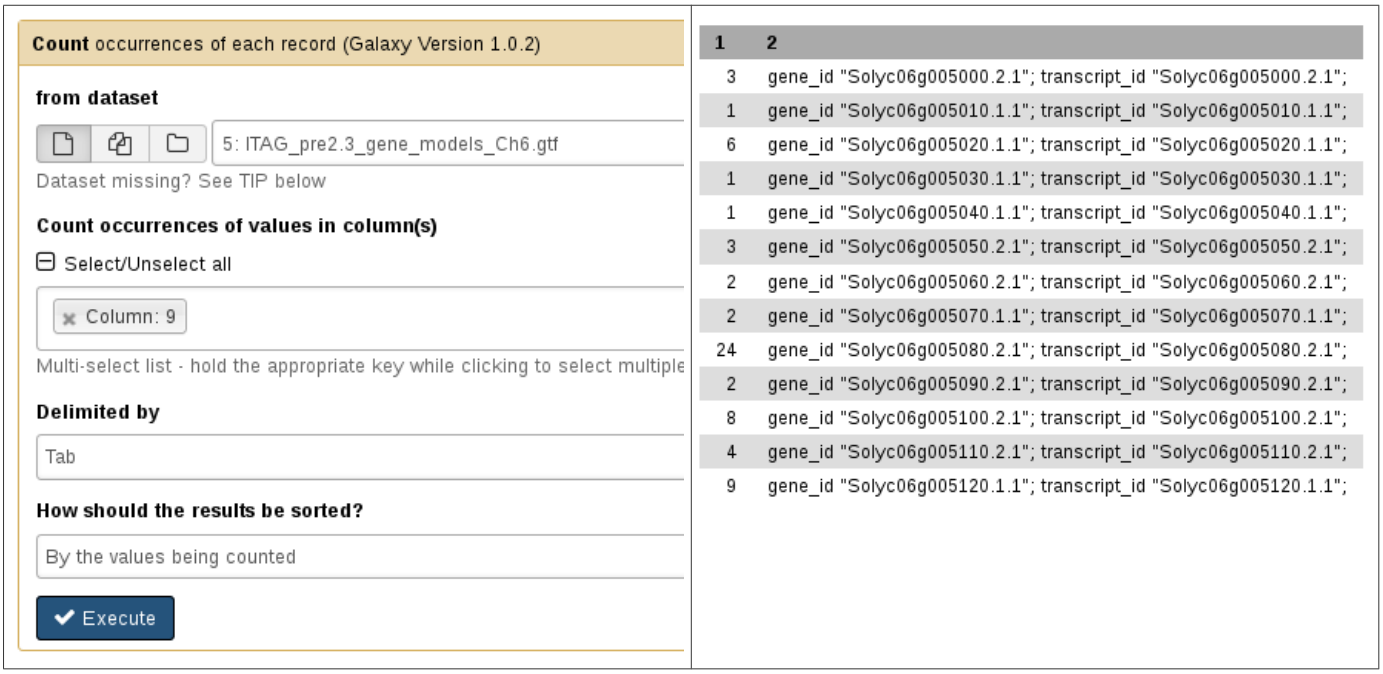


Dans IGV, pensez à étendre la piste du gtf (view « extended »).

Exercice n°5 : mesure d'expression au niveau gènes/transcripts

1) Manipulation du GTF, se familiariser avec sa référence :

À partir du fichier ITAG_pre2.3_gene_models_Ch6.gtf, comptez combien il y a de transcrits ?



2) Quantification brute avec FeatureCount

a) Lancez feature count avec le GTF de référence sur l'ensemble des échantillons (utilisez les paramètres définis dans le tutoriel)
b) Testez également le comptage au niveau des gènes avec cet utilitaire.

Exercice 6 : Sauvegarder votre chaîne de traitements en convertissant votre historique en workflow

Créez un workflow à partir des traitements bioinformatiques précédemment réalisés :

Etape 1 : Depuis le menu « Options » de votre historique en cours, choisir « Extract Workflow ».
Un workflow est ainsi généré automatiquement et disponible depuis le menu « Workflow ».

Etape 2 : Exportez ensuite ce workflow en cliquant sur la flèche noire à côté de l'intitulé du workflow et choisir « Download or Export ».

Etape 3 : Download to File : Veuillez cliquer sur « Download workflow to file so that it can be saved or imported into another
Galaxy server.»
Puis enregistrez ce fichier sur votre PC. Il vous sera ensuite possible de le ré-importer dans votre instance Galaxy.

Création de workflow :

- A partir de rien : Menu « Workflow » puis « Create a new workflow »
- A partir d'un historique : « History panel » Options → « Extract workflow »

Comme pour les historiques, il est possible de partager des workflows.