TP Sciences Omiques: données NGS

Thomas Dugé de Bernonville

November 26, 2018

1 Introduction

Les séances de TP prévues ont pour but de vous initier aux outils bioinformatiques pouvant être déployés pour le traitement de données **NGS** (Next Generation Sequencing) vues en cours. Tous ces outils s'utilisent normalement en ligne de commande dans un terminal sous Linux. Afin de simplifier l'accès de ces outils à la communauté scientifique, une interface graphique appelée **Galaxy** (`https://usegalaxy.org` ou `https://usegalaxy.eu`). Cette interface utilise le langage de programmation **Python** pour créer des raccourcis vers différents outils, sous forme d'interface graphique.

Pour lancer Galaxy, connectez-vous à une session Linux dans une salle info science, ouvrez un terminal (ctrl+alt+t) et tapez *galaxy* dans la console. Ouvrez alors un navigateur à l'adresse suivante: `http://127.0.0.1:8080`

TP2 Cette séance vous montrera comment traiter des données RNA-seq (Figure 1) obtenues par la technologie **Illumina**. L'exemple porte sur une plante médicinale, *Catharanthus roseus* et deux runs de RNA-seq obtenus sur des feuilles contrôle (échantillon1) et des feuilles consommées par une chenille (échantillon2). Deux stratégies seront envisagées pour illustrer le processus de reconstruction des transcrits à partir d'une référence génomique ou en absence de référence (Figure 2). Le TP finira sur l'étude de l'expression des transcrits.

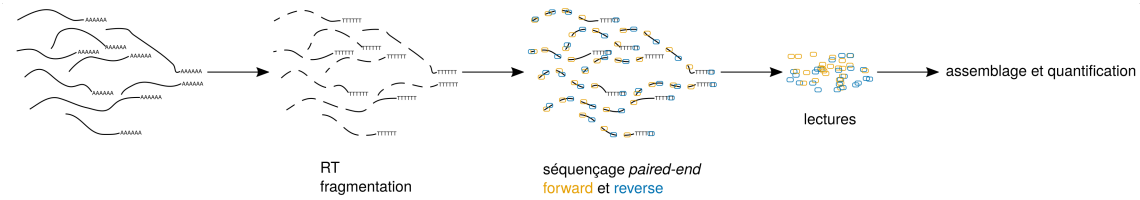

Figure 1: Principales étapes du RNA-seq.

Déroulement (Figure 2) Nous débuterons avec l'alignement des lectures d'un échantillon donné sur une portion du génome de référence. Des transcrits seront reconstruits et leur expresssion quantifiée à partir de ces alignements. Nous illustrerons ensuite comment un transcriptome peut être assemblé *de novo*, c'est-à-dire sans référence. A partir de l'assemblage *de novo*, une quantification sera également réalisée.

- **Assemblage avec référence**
  1. alignement des lectures avec **HiSAT2**
  2. prédiction des gènes avec **StringTie**
  3. quantification avec **StringTie**
- **Assemblage** *de novo*
  1. assemblage *de novo* avec **Trinity**
  2. quantification avec **Salmon**
  3. annotation par Blastx

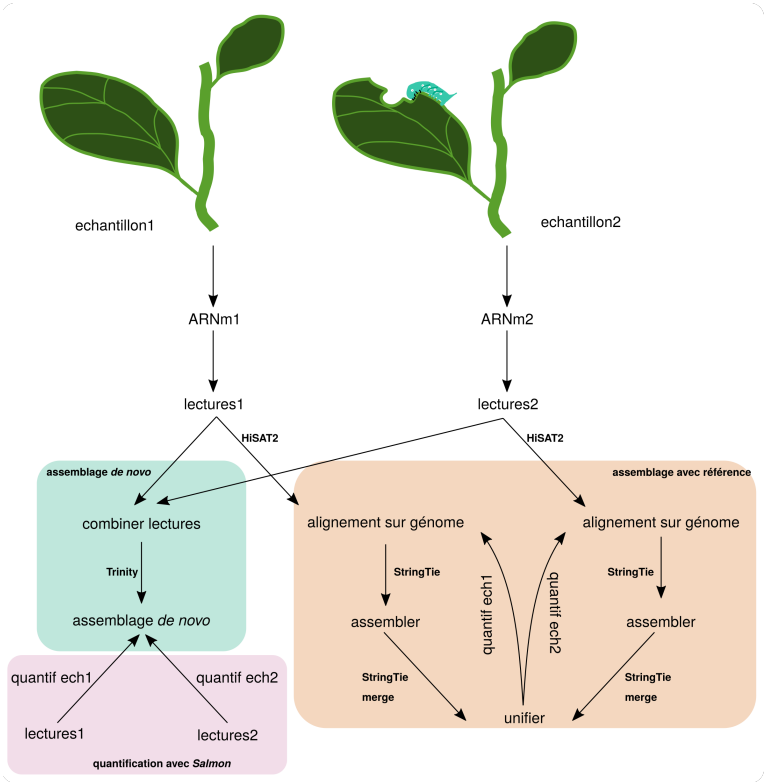

Figure 2: Stratégies pour le traitement des données RNA-seq.

Visualisation Dans la mesure du possible, nous utiliserons l'Integrative Genomics Viewer du Broad Institute pour visualiser les résultats des différentes analyses (`http://software.broadinstitute.org/software/igv/igv2.3`, charger une version online avec 750Mo de mémoire.) Les fichiers **.bam** (format d'alignement binarisé) et **.gff** (annotation au format tabulaire) seront téléchargés depuis Galaxy et ouvert dans IGV.

2 Alignement sur une séquence de référence

La suite **HiSAT2** et **StringTie** va être consécutivement utilisée pour créer des alignements sur la séquence de référence, reconstruire des transcrits et quantifier leur expression (Figure 4).

Lorsqu'une séquence génomique est disponible au préalable, les lectures issues du séquençage peuvent cartographiées directement sur cette séquence. Alors que les lectures obtenues par RNA-seq devraient être dépourvues d'introns, ces-derniers sont présents dans la séquence génomique. Cela implique que certaines lectures ne pourront pas être correctement cartographiées car elles seront interrompues sur la séquence génomique par les introns (Figure 3). Pour résoudre ce genre de problème, il faut utiliser des programmes dédiés qui analyseront les sites de multiples cartographies pour les lectures non alignées dans un premier temps et mettre en lumière des sites de jonctions intron-exon. Une fois les lectures alignées, un assembleur est utilisé pour retrouver un chemin de séquence correspondant à la séquence transcrite. La suite la plus utilisée jusqu'à présent était **TopHat** et **Cufflinks** mais ces deux derniers sont en cours de remplacement par des variantes plus précises et plus performantes, **HiSAT2** et **StringTie**. **TopHat** et **HiSAT2** reposent tous les deux sur des transformations de **Burrows-Wheeler** compressées en index **FM**, utilisées par l'aligneur **Bowtie2**. Cette conversion lexicographique est reconnue pour être plus économe en taille et plus performante en termes de recherche.

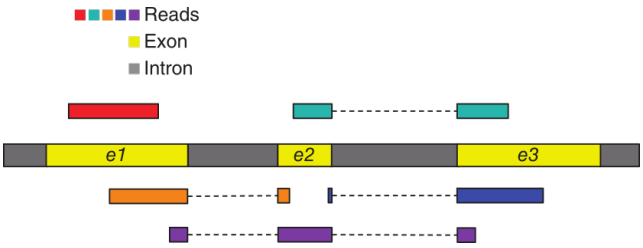

Figure 3: Alignement des lectures sur les séquences exoniques.

2.1 Alignement des lectures

Mapping avec HiSAT2 Charger les données suivantes (télécharger à cette adresse:

- la séquence de référence: cro_v2_scaffold_18.fa
- l'annotation *in silico* de cette séquence: cro_v2_scaffold_18.gff3
- fichiers fastq Forward (_1) et Reverse (_2) des échantillons 1 et 2

Executer HiSAT2 en choisissant la séquence de référence chargée et en utilisant les lectures de l'**échantillon 1** (Figure 5). Nous utiliserons les autres paramètres par défaut. Un clic sur *Advanced Options* vous montrera les différents paramètres qu'il est possible d'optimiser pour affiner l'alignement. Dans l'item *view details* de l'objet **.bam** vous pourrez trouver les statistiques du mapping.

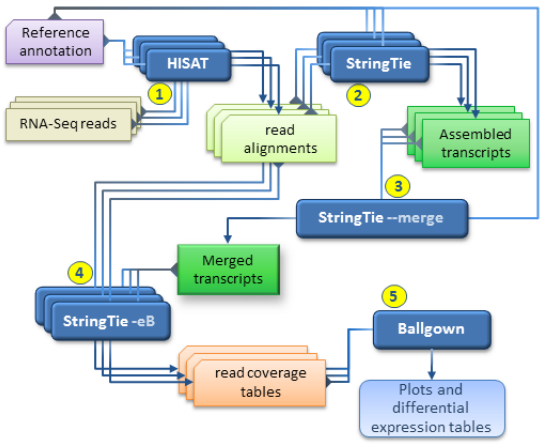

Figure 4: Assemblage et quantification avec la suite HiSAT2 et StringTie.

Visualisation dans IGV Commencer par charger la séquence de référence dans IGV avec le menu *load Genome from file...*. Télécharger le fichier **.bam** depuis Galaxy. Pour le visualiser, il faut d'abord créer un index qui sera utilisé par IGV pour se répérer sur cette séquence. Dans IGV, utiliser le menu *run IGVtools* avec le mode *index* sur le fichier **.bam**. Cette opération va créer un fichier **.bai** dans le dossier de travail. Vous pouvez à présent charger l'alignement avec le menu *load from file* (Figure 6). Charger à présent le fichier **.gff3** correspondant à l'annotation *in silico* de la référence.

2.2 Reconstruction des transcrits

Assemblage avec StringTie Dans la stratégie utilisée, un assemblage est reconstruit pour chaque run RNA-seq (Figures 2 et 4). Il faut donc disposer des fichiers **.bam** pour chaque run. Au final, StringTie sera appelé par trois vagues successives:

- Assemblage: **StringTie** pour chaque échantillon, sans fichier **.gff** de référence (Figure 7)
- Assemblage: **StringTie merge** pour créer un assemblage commun à tous les échantillons (Figure 8)
- Quantification: **StringTie** pour chaque échantillon en utilisant l'assemblage ci-dessus comme référence (Figure 9)

Visualisation et fichiers de sortie StringTie produit un fichier d'annotation **.gff** que vous pouvez visualiser dans IGV après l'avoir téléchargé depuis Galaxy et en utilisant *load from file*. Vous pouvez comparer les zones de couvertures des lectures, les transcrits prédits *in silico* et la prédiction basée sur le RNA-seq par StringTie.

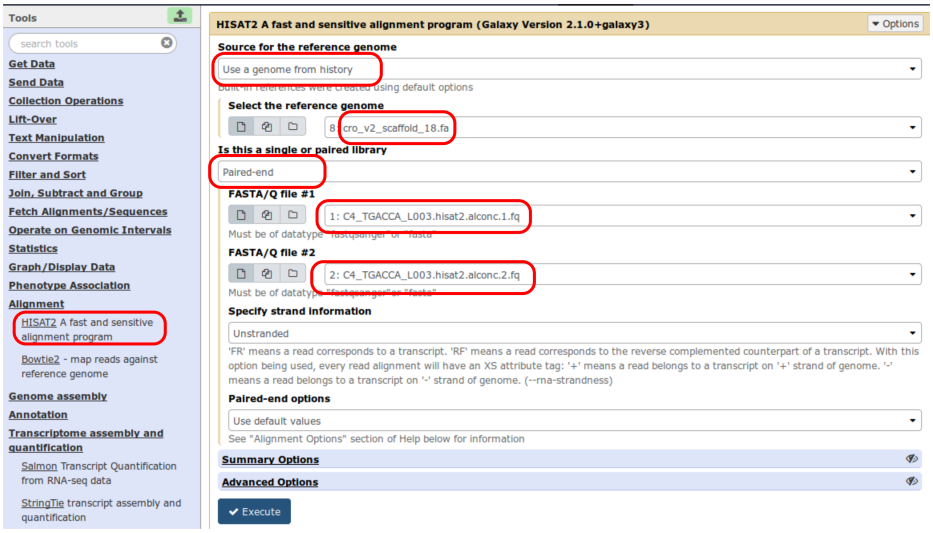

Figure 5: Utilisation de HiSAT2 sous Galaxy.

3 Assemblage *de novo*

Préparation du jeu de données L'assemblage *de novo* sera réalisé sur les données de séquences combinées. Dans un terminal, placez vous dans votre dossier contenant les données. Utilisez les commandes suivantes pour concatener les fichiers **Forward** d'une part et **Reverse** d'autre part.

```
/$ cd votredossier
/votredossier$ cat *_1.fq > all_1.fq
/votredossier$ cat *_2.fq > all_2.fq
```

Charger ces données combinées dans Galaxy.

Assemblage avec Trinity et Annotation Vous utiliserez le programme **Trinity** pour assembler les lectures courtes sous forme de contigs plus longs (Figure 10). Il se déploie en 3 grandes étapes, (i) pour créer un catalogue de mots de longueurs k, (ii) chercher des chevauchements de longueur k-1 et (iii) établir les chemins préférentiels les plus probables. Cette dernière étape utilise les lectures brutes comme soutien à la résolution du graphe de **De Bruijn**. Les premières étapes sont les plus consommatrices en terme de mémoire, alors que la dernière est généralement lancée en parallèle sur de nombreux processeurs.

Exécuter le programme sur les données combinées, en sélectionnant la normalisation *in silico* pour plus d'efficacité et un nombre minimum de k-mer égal à 2 pour plus de robustesse (Figure 11).

Vous pouvez annoter les transcrits ainsi reconstruits avec un **Blastx** contre la banque Uniprot, comme vu au TD précédent. Il existe aussi des programmes tels que **Transdecoder** capables de détecter les ORF dans ces séquences de transcrits.

Visualisation Il est possible de cartographier les transcrits reconstruits *de novo* de manière à obtenir un fichier **.gff** qui pourra être visualisé dans IGV.

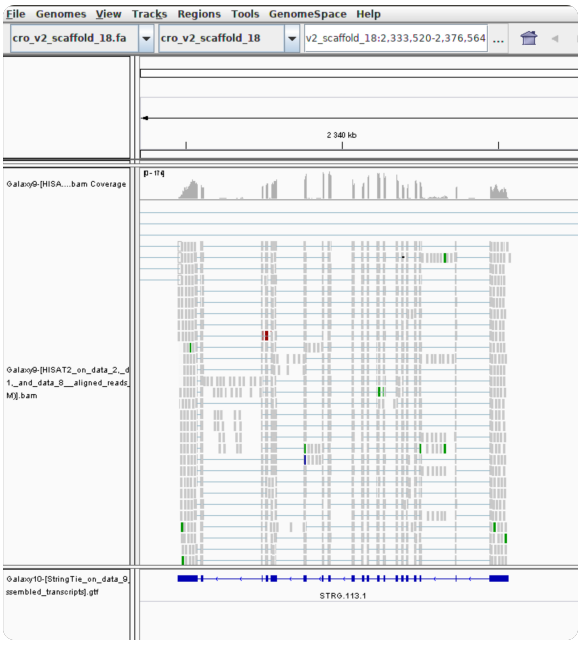

Figure 6: Visualisation d'un fichier .bam et d'un assemblage par StringTie dans IGV.

Pour cela, plusieurs étapes sont nécessaires:

1. créer une database BLAST sur l'assemblage Trinity avec **blastdbcmd**
2. blaster la séquence de référence (**query**) contre cette database avec l'algorithme **megablast** avec les options suivantes:
   - output format= BLAST XML
   - dans les options avancées, une valeur de **1** pour *Maximum hits to consider/show*
3. convertir l'objet BLAST XML grâce à l'outil **BlastXML to gapped GFF3**

Le fichier **.gff3** peut maintenant être visualisé dans IGV.

Quantification La quantification des transcrits assemblés *de novo* est réalisée avec **Salmon**. Ce programme est ultra-rapide et très performant. Il se base sur une notion de **quasi-mapping** beaucoup plus efficace que le mapping avec *Bowtie2* par exemple. La difficulté de cette quantification est d'attribuer correctement les lectures sur les transcrits qui peuvent parfois être très similaires. **Salmon** traite le problème de manière inverse en émettant une hypothèse sur les lectures potentiellement générables à partir de l'assemblage traité. Il faut à présent exécuter **Salmon** sur les deux échantillons séparément (Figure 12).

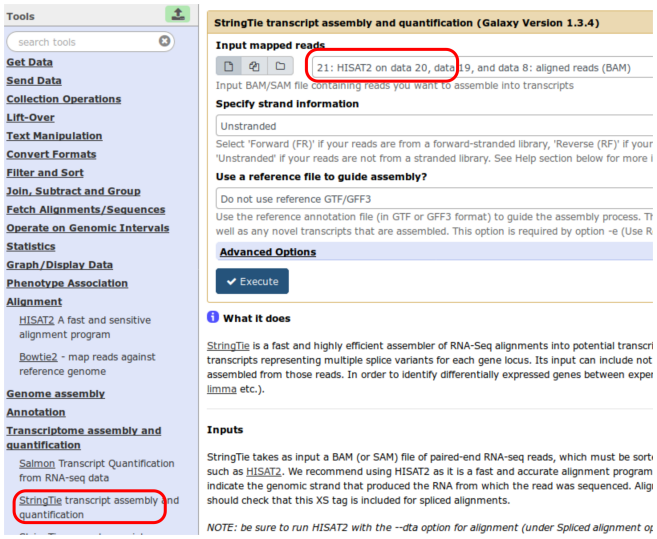

Figure 7: Utilisation de StringTie sous Galaxy.

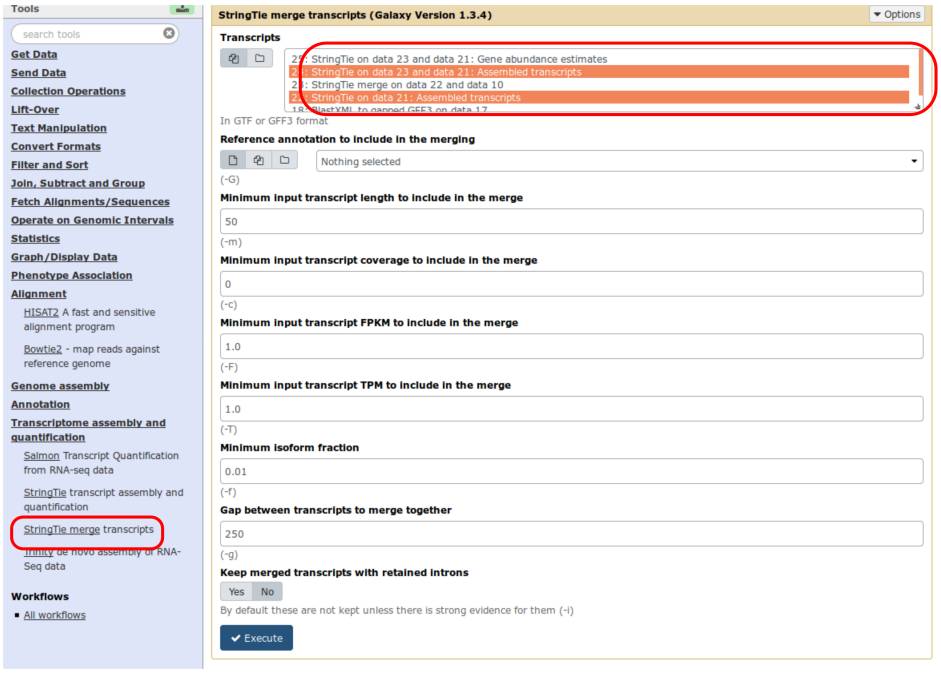

Figure 8: Utilisation de StringTie en mode *merge* sous Galaxy.

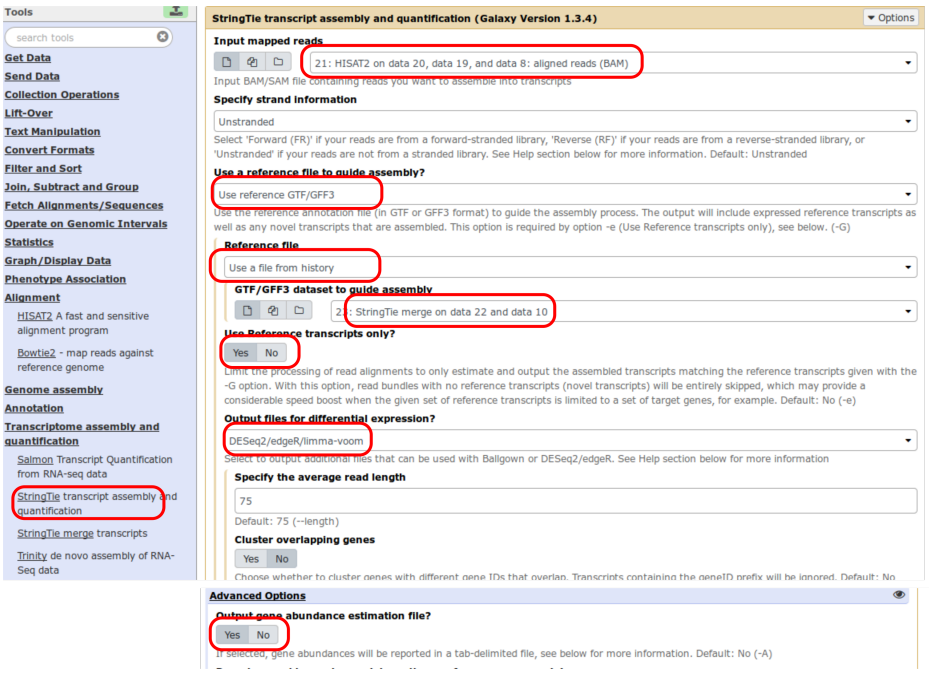

Figure 9: Utilisation de StringTie en mode quantification avec référence sous Galaxy.

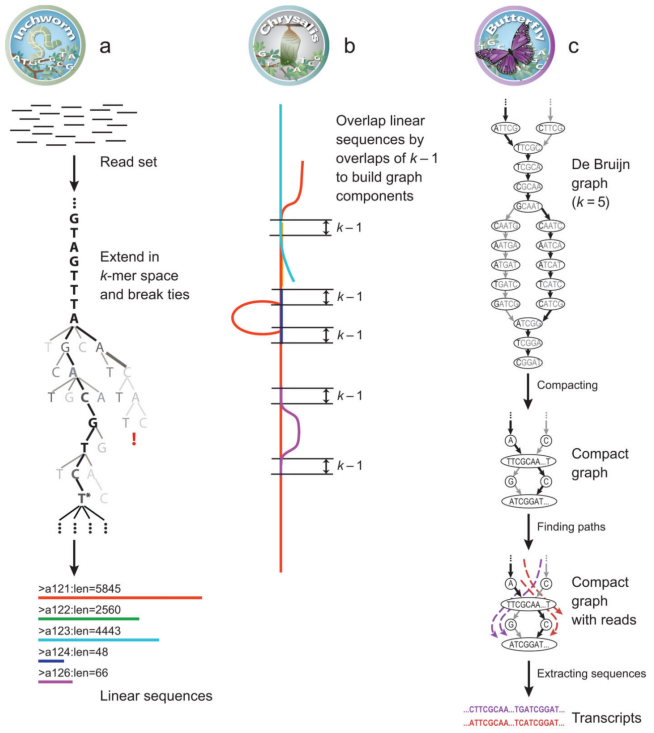

Figure 10: Fonctionnement de Trinity

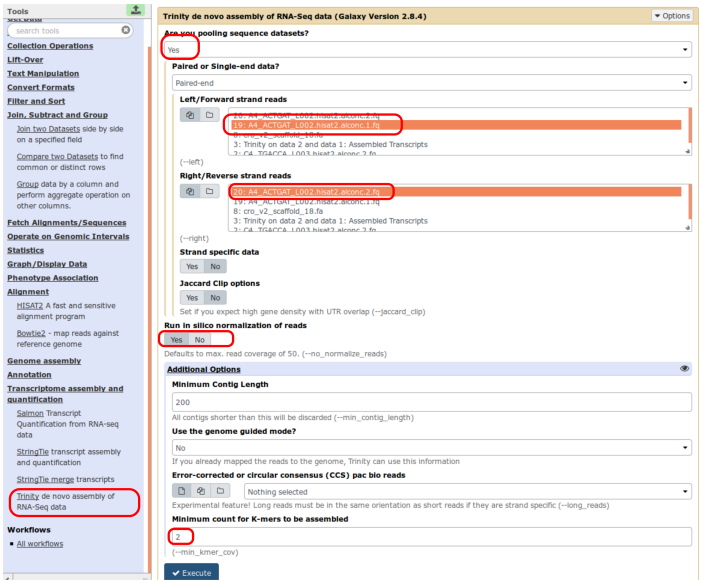

Figure 11: Utilisation de Trinity sous Galaxy.

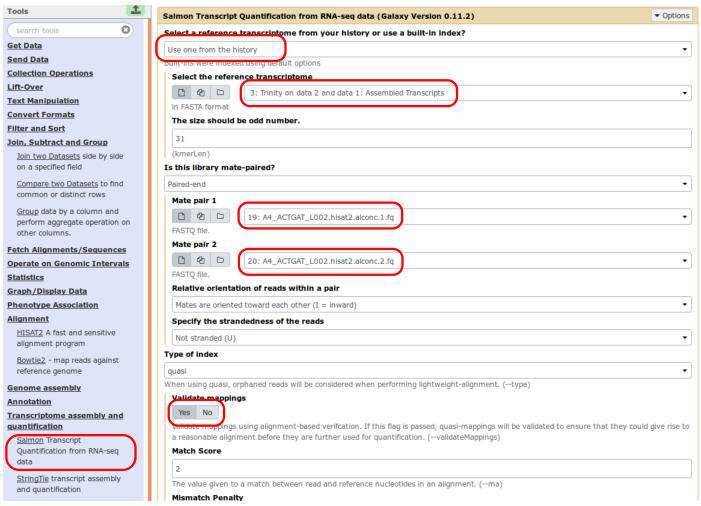

Figure 12: Utilisation de Salmon sous Galaxy.