TP Sciences Omiques: données NGS

Thomas Dugé de Bernonville

November 22, 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 console. Ouvrez alors un navigateur à l'adresse suivante: `http://127.0.0.1:8080`

Galaxy se manipule dans un navigateur web et se compose de trois fenêtres: les outils à gauche, la fenêtre principale au centre et les données et résultats à droite (Figure 1). L'utilisation de Galaxy se fait généralement de la manière suivante:

1. Sélection d'un outil dans la fenêtre de gauche
2. Paramétrage de l'analyse dans la fenêtre principale
3. Visualisation des résultats dans la fenêtre de droite

TP1 Cette séance vous montrera comment utiliser des données WGS pour assembler et analyser une séquence génomique. L'exemple porte sur une plante médicinale, *Catharanthus roseus*.

1. contrôle qualité séquences
2. trimming
3. contrôle qualité séquences
4. assemblage
5. prédiction des gènes
6. annotation (blastp XML, convert to gff)
7. visualisation
8. prédiction de cluster génomique

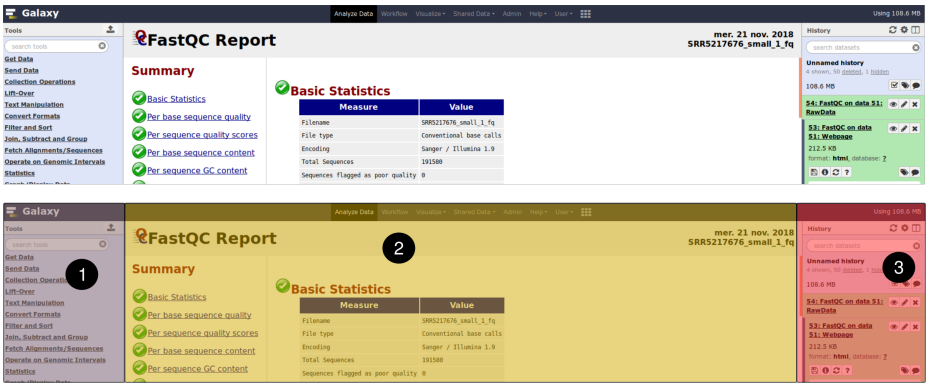

Figure 1: Organisation d'une instance Galaxy

2 Chargement des données et pré-traitement

La première étape est de charger des données sur lesquelles différentes opérations pourront être effectuées.

$Tools \rightarrow GetData \rightarrow UploadData \rightarrow ChooselocalFile \rightarrow Start \rightarrow Close$

Vous pouvez ensuite examiner les données en cliquant sur le symbole oeil. Les séquences sont issues d'un WGS sur *Catharanthus roseus*. **Récupérez des informations sur ce séquençage dans les bases SRA ou ENA.**

Contrôle qualité Lancer le programme **FastQC** pour contrôler la qualité des lectures sur chaque fichier fastq. Plusieurs indices sont utilisés dans ce but: la qualité des bases, la distribution de motifs aberrants, ou encore le % en GC. Lorsque la qualité n'est pas suffisante, comme parfois observé sur la fin des lectures, il peut être intéressant de *trimmer* les lectures pour éliminer les doutes sur les bases ayant un mauvais score.

Trimming des séquences Lancer le programme **Trimmomatic** sur la paire de lectures de l'échantillon. Ce programme va pouvoir scanner les lectures sur une fenêtre donnée et éventuellement supprimer l'extrémité 3' sur un nombre de nucléotides donnés si ceux-ci n'ont pas une qualité suffisante. Il peut aussi détecter la présence de séquences correspondant aux adaptateurs utilisés pour la construction de la banque.

Vous pouvez lancer à nouveau **FastQC** sur les séquences trimmées pour évaluer l'impact de la procédure. Les données sont à présents utilisables pour l'assemblage.

3 Assemblage d'une séquence génomique

Construire du long à partir du court Comme vous pouvez le constater, les séquences utilisées ici sont asssez courtes. Provenant de fragments chevauchant,

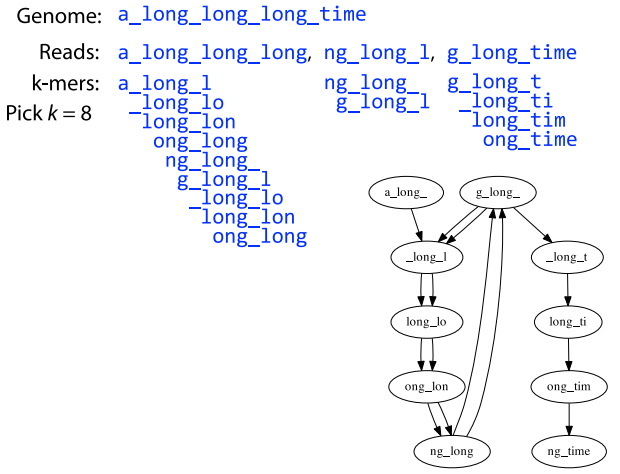

Figure 2: principe d'un graphe de De Bruijn

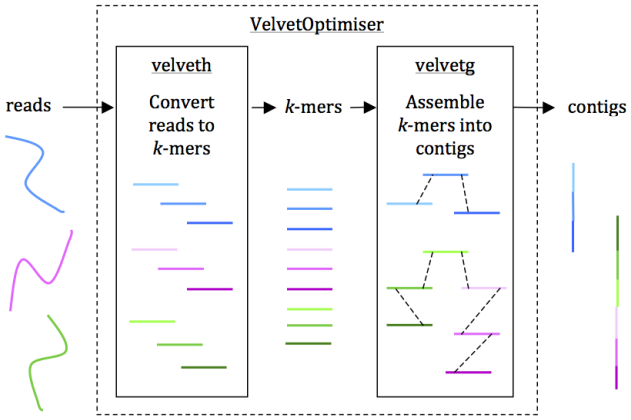

Figure 3: assemblage avec Velvet

il est probable que certaines lectures se chevauchent également et cette ressemblance peut être utilisée pour reconstruire une séquence plus longue. Les algorithmes permettant ce genre de calcul sont généralement basés sur de la reconstruction de graphes de De Bruijn (basée sur les k-mer, Figure 2) mais certains utilisent également le Overlap-Layout-Consensus (basé sur la comparaison des lectures directes). Quelques noms fréquemment retrouvés de programmes résolvant des graphes de De Bruijn: **Velvet**, **Abyss** et **ALLPATHS-LG**.

Assemblage avec Velvet Vous allez utiliser **Velvet** pour reconstruire des contigs à partir des lectures. Ce programme travaille en deux étapes (Figure 3) impliquant (i) la conversion des lectures en mots de longueur k puis l'assemblage de ces mots par chevauchements de longueur k-1. Avec Velvet, comme pour la majorité des autres assembleurs, il faut choisir une valeur pour k. Il existe aussi une manière d'optimiser l'assemblage en testant une gamme large de valeurs de k.

Tester deux valeurs contrastées de k, puis lancer l'optimisateur.

Caractériser l'assemblage Vous pouvez calculer un certain nombre de paramètres qui permettront d'estimer la qualité de l'assemblage. Parmi ces valeurs, le **N50** donne la longueur à laquelle 50% des bases assemblées se trouvent dans des contigs de cette longueur. Vous trouverez également des valeurs présentant la dispersion des longueurs de contigs et le %G+C.

4 Annotation

Prédiction des gènes avec Augustus La prédiction *in silico* des gènes est possible et devient de plus en plus pertinente avec le développement de nouveaux outils (Figure 4). Si une prédiction *ab initio* est possible (uniquement basée sur la séquence donnée), ces derniers sont de plus en plus basés sur un principe d'apprentissage utilisant un jeu de données provenant d'une espèce proche de référence lorsque celle-ci est disponible. Actuellement la manière la plus correcte d'établir des modèles de gènes (TSS, UTRs, jonctions introns-exons,...) reste le RNA-seq (voir **TP2**) et la cartographie des lectures sur la séquence génomique. Il existe plusieurs programmes permettant de réaliser une prédiction *in silico*, tels que **Maker**, **Augustus** et **SNAP**. Vous utiliserez **Augustus** avec les paramètres par défaut.

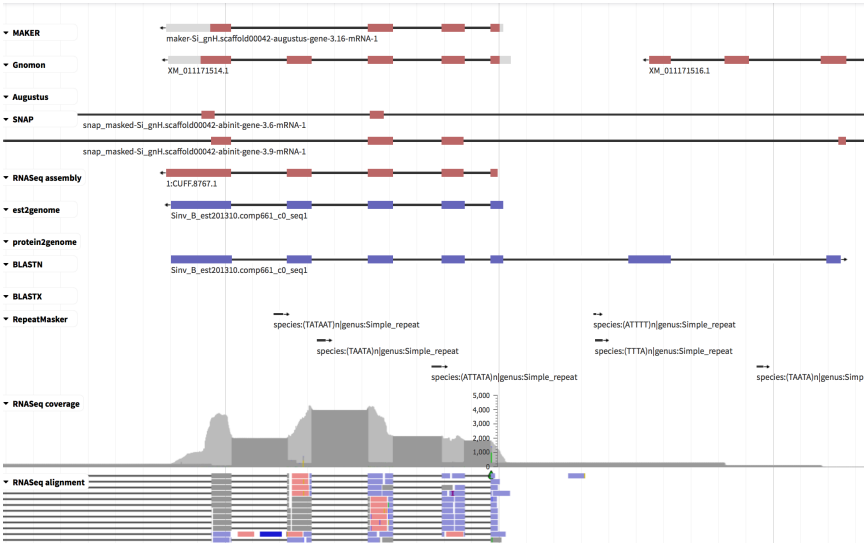

Figure 4: Détermination d'un modèle de gène par plusieurs approches.

Annotation des gènes Une fois la présence de gène suspectée, il est intéressant de pouvoir les annotant en recherchant des homologies avec d'autres gènes connus par l'utilisation de bases de données. Les deux principales manières d'attribuer des homologies sont:

1. Par BLAST contre une base de séquences de référence, par exemple **Uniprot**

2. Par scan de domaines fonctionnels, par exemple **Pfam** avec Hmmer

Vous utiliserez la suite NCBI BLAST+ pour trouver des homologies contre Uniprot pour votre assemblage.

Une fois la recherche d'homologie terminée, convertissez la sortie XML en fichier GFF.

Les fichiers **GFF** sont des fichiers txt reprennant des annotations sur une ou plusieurs séquences nucléotidiques. Voir ce lien pour plus d'informations: `https://www.ensembl.org/info/website/upload/gff.html`.

Visualisation Vous disposez à présent d'un assemblage annoté. Nous allons visualiser cette assemblage grâce au **Integrated Genomics Viewers** développé par le Broad Institute. Il faudra charger l'assemblage puis le fichier d'annotation Augustus pour visualiser les loci où ont été prédits les gènes. Un raccourci est disponible dans Galaxy.

5 Prédiction de clusters génomiques

Les gènes de voies de biosynthèse peuvent être physiquement proches sur un génome Des algorithmes ont été récemmment développés pour mettre en évidence des clusters de gènes chez les bactéries, les champignons et les plantes à partir de séquences génomiques annotés. De tels clusters regroupent des gènes codant des réductases, des oxidases, des transférases... et leur proximité génomique pourrait indiquer qu'ils interviennent dans une même voie.

Consultez la page `http://plantismash.secondarymetabolites.org/` pour plus d'infos et tester vos résultats.