Protocol Utilisation de la Plateforme Galaxy pour alignement d'ARN

I. Introduction

La plateforme Galaxy regroupe plusieurs outils bio-informatiques avec une interface graphique pour faciliter l'accès.

Avec la plateforme Galaxy on peut manipuler les séquences, faire de l'alignement, ou convertir les fichiers en différents formats pour être utilisés ailleurs.

On peut se servir de deux plateformes Galaxy :

1. La plateforme publique maintenue par Galaxy Project : https://usegalaxy.org/

2. La plateforme locale maintenue par nous-mêmes : http://galaxyserver.riboclub.org:8080/

La plateforme publique propose plus d'outils et est plus performant mais le quota est limité (250G par utilisateur) et le temps d'attente peut être insupportable (quelques jours par exemple, dépendant de la charge de serveur).

Sur la plateforme locale se trouvent la plupart des outils courants mais il se peut que certains outils soient manquants. En outre, il est fortement conseillé de ne rouler qu'une analyse par fois. Si vous rencontrez des problèmes de logiciel ou ne trouvez pas l'outil dont vous avez besoin, n'hésitez pas à contacter Léandro Fequino via Leandro.Fequino@USherbrooke.ca. Leandro installera vos logiciels d'intérêt et nous bâtirons ainsi progressivement une plateforme locale qui répondra à tous nos besoins.

II. Utilisation

Avant de commencer

* **Il est important de s'enregistrer** et de **se loguer** pour pouvoir accéder aux analyses dans le futur. Pour ce faire, cliquer sur « **user** » puis « **register** » dans la partie noir en haut du page.

La liste d'outils se trouve à gauche. Les fichiers uploadés et les résultats d'analyse se trouvent à droite dans la colonne history. Vous pouvez créer autant d'history que vous voulez et changer les noms pour vous rappeler de quelle analyse il s'agit.

Entrer les données

* Cliquer sur « **Get Data** » puis « **Upload data** » pour uploader les fichiers fastq. Pour les fichiers plus gros, utiliser l'option ftp avec filezilla (cf. III Miscellaneous) par exemple. Choisir le génome de référence à utiliser.

Control de qualité

* **NGS : QC and manipulation – FASTQ Groomer** sur les fichiers fastq pour convertir ces fichiers en un format que Galaxy peut reconnaître. Faire attention au « input FASTQ quality score type ».

* **NGS : QC and manipulation – FastQC : Read QC** pour générer le rapport de qualité.

« Per base Sequence Quality » donne la distribution de score de qualité pour chaque base. La qualité en 3' est souvent plus basse qu'en 5'.

* **NGS : QC and manipulation – FASTQ Quality Trimmer** pour Trimer les séquences selon la qualité : souvent un score en dessous de 20 est considéré comme trop bas. Jouer sur les paramètres window size et step size pour améliorer le trim.

Alignement

* **NGS : Mapping Bowtie2** ou **NGS : RNA Analysis – Tophat for Illumina** ou **Tophat2** pour aligner les séquences contre un génome.

Sur Galaxy il existe plusieurs logiciels pour faire l'alignement. Tophat utilise Bowtie pour faire un alignement primaire ensuite génère une liste de jonction d'épissage possible pour aligner les lectures qui couvrent la jonction de deux exons donc est mieux adapté à l'alignement d'ARNm.

Séquençage single end et paired end : le séquençage paired end permet de lire les fragments plus longs mais « tronqués », donc Tophat demande la distance entre deux lectures venant du même fragment pour mieux aligner les séquences.

Le fournisseur de fastq est censé de connaître le « Mean Inner Distance between Mate Pairs » et « Std. Dev for Distance between Mate Pairs » du raw data, on peut également rouler Bowtie sur une partie des lectures pour avoir une estimation sur ces deux paramètres.

Il est important de garder en tête que ces deux paramètres vont probablement avoir changé après le trimming.

Choisir « Full parameter list » dans l'onglet « Tophat settings to use » pour voir la liste complète d'options.

Lire attentivement les descriptions avant de lancer une analyse.

Assemblage de transcrits ou Analyse d'ARN :

***NGS : RNA Analysis – Cufflinks** sur chaque fichier .bam produit par Tophat pour assembler les lectures en transcrits.

Le « mean inner distance » est important pour utiliser Cufflinks.

Pour plus d'info : http://cufflinks.cbcb.umd.edu/manual.html#cufflinks

*__NGS : RNA Analysis – Cuffmerge__ sur les données d'assemblage de transcrits et un génome de référence si besoin :

*__NGS : RNA Analysis – Cuffdiff__ sur a) sortie de Cuffmerge ; b) .bam de « accepted hits » produit par Tophat de chaque base de données pour comparer l'expression entre les données.

Lire ce document pour mieux comprendre la sortie de Cuffdiff :

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

III. Miscellaneous

Pour transférer des gros fichiers :

- Utilisation Filezilla : télécharger Filezilla https://filezilla-project.org/download.php et installer.

Connecter avec le nom d'utilisateur et mot de passe de login comme pour se loguer sur la plateforme. Mettre **usegalaxy.org** dans « **host** » pour se connecter sur le serveur public ou **galaxyserver.riboclub.org** pour le serveur local. La connexion est établie lorsque le message « Status: Directory listing successful » apparait.

Visualisation sur un Genome Browser :

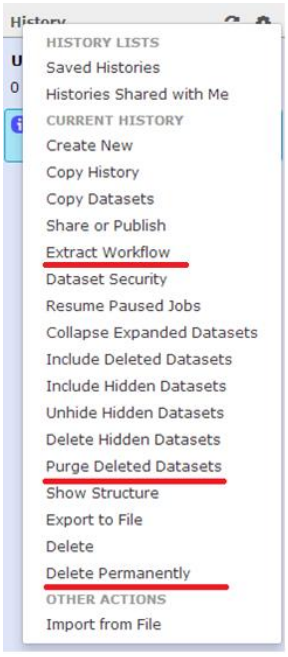
- Pour visualiser l'alignement ou la différence d'expression : faire **Visualisation – New visualisation** ou **exporter les résultats** vers le genome browser de UCSC.

Manipulation de fichiers :

- Editez les data sets en cliquant l'icône crayon. Ajoutez des annotations et changez les noms pour garder une trace de ce que vous faites.

- Supprimez les données dont vous n'avez plus besoin. Faites « **Purge Deleted Datasets** » pour supprimez immediatement les données ou « **Delete Permanently** » pour supprimez l'history en entier. **N'EST PAS REVERSIBLE.**

Sur le serveur local : ecrivez un mail à Léandro pour lui notifier – des fois il faut supprimer les fichiers manuellement via l'interface en ligne de commande.



- Si vous faites tout le temps le même type d'analyse avec les mêmes paramètres, vous pouvez créer un workflow (**Workflow – Create new workflow**) ou faire « **Extract Workflow** » pour enchainer automatiquement les outils avec les paramètres pré-définis.