R Notebook

Sandrine AUGER

2020-08-21

Description du projet

Loading the used R libraries

1. Traitement des données RNAseq

Les données fournies sont des données brutes issues du séquençage à haut débit par la technologie Illumina à partir d'une extrémité ("single-end sequencing"). Le séquençage a été effectué sur la plateforme I2BC (Paris-Saclay). Les 18 fichiers fastqc ont été transférés sur le cluster de la plateforme Migale.

Organisation de l'espage de travail sur la plateforme Migale

Personal space: save_home/saauger/ The space for configuration files (.ssh, .bashrc , ...) and personal saved data, not intended for computation. Backup is active on this space. **Personal work space: work_home/saauger/** Personal space for work, secured, but without backup, for computational data, results data.

Création des dossiers de travail

Transfert des scripts sur Migale depuis une fenêtre shell

Conversion des scripts Windows sous Unix

Les scripts bash sont crées sur l'ordinateur avec Notepad puis transférés sur le cluster Migale. Il convient de convertir les scripts provenant de Windows en mode Unix.

1.1. Contrôle qualité des données brutes

1.1.1. Utilisation de l'outil FastQC

Recherche des options disponibles pour un outil dans l'environnement conda :

La qualité de séquençage est vérifiée avec le programme FastQC (Babraham Institute, Cambridge) en lançant le script FASTQC.sh.

Le temps mis pour exécuter ce job est de 37 minutes.

1.1.2. Compilation des rapports avec l'outil MultiQC

L'outil MultiQC compile un rapport HTML. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller. Bioinformatics (2016).

Le temps mis pour exécuter ce job est de moins de 1 minute.

1.1.3. Résultats

Le **nombre de lecture varie entre 25.3 et 33.4 millions**. Le nombre de lecture unique est d'environ 20 % et le nombre de **lectures dupliquées** est d'environ 80 %. **Le niveau de duplication est trop élevé ???** Tous les échantillons ont des séquences d'une seule longueur (**75 nts**), avec un **GC% = 41**. La valeur moyenne de qualité pour chaque position de base dans la lecture est **Phred Scor e= 35**. Le pourcentage d'appels de base à chaque position pour lequel a été appelé N est très faible **(N < 0.01%)**. **Le contenu des séquences ayant un adaptateur est de moins de 1%** Conclusion : données de bonne qualité, avec une bonne profondeur, les adaptateurs ont déjà été enlevés.

1.2. Preprocess : nettoyage et filtrage des séquences avec fastp

L'outil de prétraitement fastp permet d'enlever les adaptateurs, de filtrer par la qualité et la longueur des reads. Chen S, Zhou Y, Chen Y, Gu J "fastp: an ultra-fast all-in-one FASTQ preprocessor." Bioinformatics. 2018 Sep 1;34(17):i884-i890.

1.2.1. Gestion du nom des fichiers de sortie

Le but est de renommer les fichiers de sortie en gardant uniquement la première partie du nom du fichier d'entrée fastqc. La commande cut -d"_" -f1-4 permet de sélectionner les 4 premiers champs dans les noms des fichiers. Par exemple, I213_C_t53_S18_all_R1_001.fastq.gz devient I213_C_t53_S18.fastq.gz. Je fais un test en lançant le script renommer.sh.

1.2.2. Trimming et filtrage avec l'outil fastp

Le trimming des adaptateurs et leur autodétection sont activés par defaut.

QUALITY FILTER fastp prend en charge la découpe de fenêtre coulissante par lecture en évaluant les scores de qualité moyens dans la fenêtre coulissante. **-W, –cut_window_size** range 1-1000, default=4. **-u, –unqualified_percent_limit**, how many percents of bases are allowed to be unqualified (0~100). Default 40 means 40%. **-q, –qualified_quality_phre**, the quality value that a base is qualified. Default 15 means phred quality >=Q15 is qualified. La limite maximale du nombre de **N** de 5 par read est laissée par défaut, elle peut être modifiée par l'option -n(-n_base_limit). L'outil fastp trimme les **polyG** (>= 10 bases) en fin de reads. In Illumina NextSeq/NovaSeq data, polyG can happen in read tails since G means no signal in the Illumina two-color systems. LENGTH FILTER **-l, –lenght_required**, seuil de longueur requis, par defaut = 15, reads shorter than lenght_limit will be discarded. La longueur de l'adaptateur fait 32 nts, la longueur des reads trimmés va diminuer fortement.

1.2.3. FastQC et MultiQC après le trimming et le filtrage

Les fichiers de sortie fastp.fastq.gz sont analysés avec FastQC. Puis, une compilation des rapports est effectuée avec MultiQC.

Globalement, plus de 99,5% des reads ont passé les filtres, 0.4% avait une mauvaise qualité et 0.04% une longueur trop courte.

1.3. Mapping

BWA (le Burrows-Wheeler Aligner) est un aligneur rapide à lecture courte. Il utilise la transformation burrows-wheeler pour effectuer l'alignement de manière efficace en termes de temps et de mémoire.

1.3.1. Construction de l'index

The complete genome sequence of *S. thermophilus* strain N4L has been deposited in GenBank under the accession no. LS974444 (BioProject no. PRJEB27286) (Proust *et al.*, 2018). La séquence du génome est récupérée sur NCBI puis transférée dans le dossier save_home/saauger/projet_tutore/MAPPING.

1.3.2. Mapping sur le génome

Les outputs sont des fichiers .sam. Ces fichiers contiennent toutes les informations sur la position de chaque read sur le génome de référence. L'outil SAMTools est ensuite utilisé pour convertir les fichiers .sam en .bam, les trier et les indexer.

1.3.3. Génération de la table de comptage

Une attention particulière doit être apportée pour décider comment traiter les lectures qui s'alignent ou se chevauchent avec plus d'un gène. Le script HTseq-count est utilisé dans ce projet.

The are one or more files containing the aligned reads in SAM format.