Octobre 2022

UNIVERSITÉ TOULOUSE III PAUL SABATIER Université de Toulouse

Nextflow

Initiation sur genologin

Date de rendu: 12 oct. 2022
Cursus : M2 Bioinformatiques et Biologie des systèmes

Kory Jason

Table des matières

1. Connexion à Genologin

Identifiant : dahlia

1.1 Organisation de l'espace de travail

Vous trouverez sur le server Genologin, l'ensemble des fichiers rapporter dans ce rapport dans le sous-répertoire : `~/work/dahlia/tomatoes*` . Les fichiers de données sont dans le sous-répertoire `~/work/dahlia/tomatoes/data` ; les résultats générés se trouvent dans le sous-répertoire `~/work/dahlia/tomatoes/resultats`.

2. Préparation fichier bash de lancement Nextflow

2.1 Pipeline nf-core/rnaseq

La pipeline nf-core/rnaseq de Nextflow est un outil bioinformatique pour l'analyse de données RNA seq, provenant d'organisme avec un génome de référence et des annotations.

La pipeline procède au pré-traitement des données, à l'alignement et la quantification des reads, ainsi qu'au mapping des reads sur le génome de référence. Elle contient de plus, des outils pour un contrôle qualité des résultats RNAseq. L'ensemble des sorties sont brièvement commentées dans la suite du rapport.

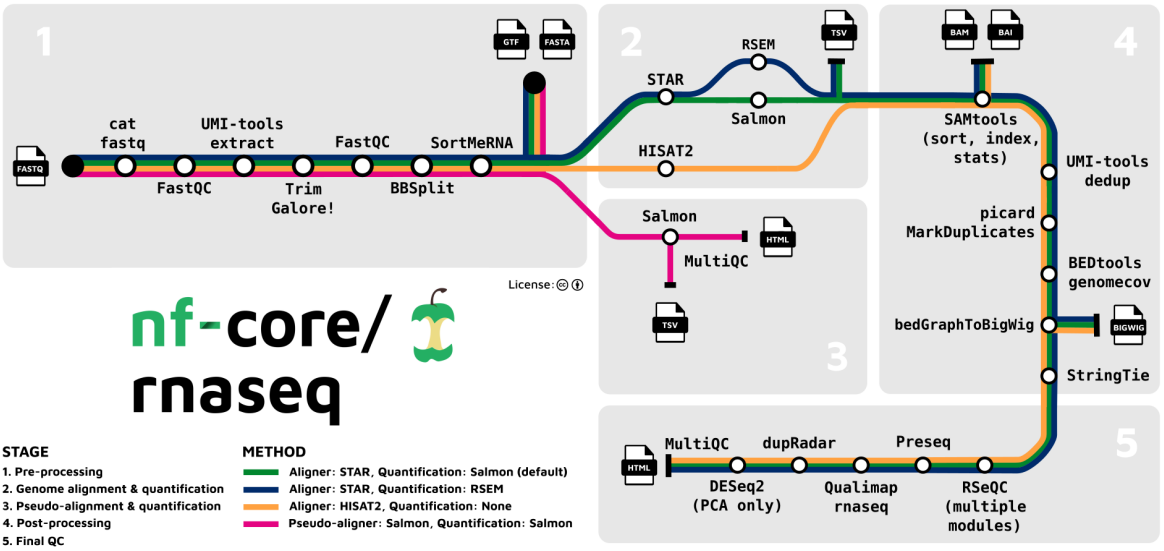

Figure 1 : Etape de la pipeline nf-core/rnaseq.

2.2 Préparation fichier bash

2.2.1 Dossier data

L'organisme étudié est la tomate. Plus particulièrement la séquence nucléotidique de référence est le chromosome 6 de *Solanum lycopersicum* (SL2.40ch06). Il s'agit de la tomate "cookie" devenue un élément incontournable de la gastronomie européenne.

Les fichiers requis en entrée pour la pipeline nf-score/rnaseq, sont les suivants :

- fichier **fastq** (MT_rep{1,2}_{1,2}_Ch6.fastq.gz ; ***figure 1)*** décrivant les lectures (reads) de séquençage à haut débit générées et leur qualité par base.
- fichier **fasta/gtf** (ITAG2.3_genomic_Ch6.fasta/.gtf ; ***figure 2***) contenant la séquence du génome de référence, et une description de la structure des gènes respectivement.

- tableau (sample.csv ; ***figure 3***) contenant les informations sur les fichiers utilisés en entrée (requis pour le run)

2.2.1 Fichier tomatoes.sh

Le fichier bash, intitulé `tomatoes.sh` contient les commandes pour l'exécution de la pipeline.

```
#!/bin/bash

#SBATCH -J JASONKory
#SBATCH --time=1-0
#SBATCH -p workq
#SBATCH -e error.out
#SBATCH --mem=6G
#SBATCH --cpus-per-task=32
#SBATCH --mail-type=BEGIN,END,FAIL

#Purge previous module

module purge

#Load the application
module load bioinfo/Nextflow-v21.04.1
module load system/singularity-3.5.3

#Command to run on the cluster

nextflow run nf-core/rnaseq -r 3.4 -profile genotoul \
--input /work/dahlia/tomatoes/data/sample.csv \
--fasta /work/dahlia/tomatoes/data/ITAG2.3_genomic_Ch6.fasta \
--gtf /work/dahlia/tomatoes/data/ITAG2.3_genomic_Ch6.gtf \

tomatoes.sh (END)
```

Figure 2 : Contenu du fichier bash "tomatoes.sh"

Le job a été soumis avec la commande suivante :

```
sbatch tomatoes.sh
```

2.2.3 Suivi du job avec seff

La commande seff est utilisée pour trouver le rapport d'efficience du job qui a été complété, en plus de suivre sa progression.

Syntax

```
seff <job-id>
```

Output

```
Job ID: 37825179
Cluster: genobull
User/Group: dahlia/formation
State: COMPLETED (exit code 0)
Nodes: 1
```

```
Cores per node: 32
CPU Utilized: 00:03:32
CPU Efficiency: 1.13% of 05:12:32 core-walltime
Job Wall-clock time: 00:09:46
Memory Utilized: 1.51 GB
Memory Efficiency: 25.19% of 6.00 GB
```

On peut voir l'état du job (COMPLETED), la mémoire utilisé (1.51 GB), la quantité de mémoir alloué à été utilisé en pourcentage (25.19% of 6.00 GB), le temps qu'à pris le job pour être réalisé (00:09:46), des information sur les CPU, etc.

PS : la commande squeue à aussi été utilisé pour le suivi, car il est indiqué que l'utilisation de la commande seff lorsque le job est en train de tourner peux envoyer de fausses informations.

Syntax

```
squeue -u dahlia
```

2.2.4 Commande -resume

La commande permet de reprendre le précédent job après le fail de celui-ci après avoir effectué les modifications nécessaires.
Ce qui permet de "resume", donc de re-soumettre le job en cours à l'endroit du "fail" au lieu de relancer un nouveau job.

3 : Interprétation des résultats

3.1 Interprétation des principaux résultats

En sortie de la pipeline d'analyse se trouvant dans le dossier `results`, on retrouve six fichiers :

- fastqc
- genome
- multiqc
- pipeline_info
- star_salmon
- trimgalor

3.2.1 Sortie Fastqc

Fastqc est un outil pour le contrôle qualité des données de séquençage.

NB : le principe d'interprétation et les résultats étant les mêmes pour chaque sample, dans les paragraphes qui suivent une interprétation détaillée est faite des résultats pour un sample pour exemple.

Per based sequence quality : Graphique de la distribution de score de qualité pour chaque position dans un read, pour chaque reads (encode avec Illumina v1.9).

Le score de qualité est élevé pour chaque base des reads (> 28). La qualité baisse au fur et à mesure d'un run, ce qui est attendu.

On peut conclure que le séquençage et les reads sélectionnés sont de bonne qualité.

Per tile sequence quality :
Il n'y a pas eu de perte de qualité associée à la flowcell.

Per sequence quality scores : Graphique de la distribution du score moyen par séquence (pour chaque base).

L'ensemble des séquences forme une distribution très serrée, avec un score de qualité moyen élevé. Quasiment pas de séquences de faible qualité sont observées.

Per base sequence content :
On devrait observer une distribution égale des 4 acides aminés qui ne changent pas avec la position des bases. Ce n'est pas le cas ici.

Cela est dû au random hexamer priming utilisé pour la préparation de la librairie RNA-seq. Ces primeurs enrichissent certaines bases. D'où le FAIL observé pour les samples.

Per base GC content : Variation du contenu en GC le long des reads

On observe un %GC important en milieu de séquence (position 23 - 59) qui est en accord avec le %GC théorique, selon une distribution normale (même moyenne et écart-type que la librairie).
Cependant on observe un second pic, plus haut que le pic théorique, ce qui peut être à l'origine d'une contamination de la librairie.

Per base N content : Graphique de la distribution de base non ordinaire (A, T, G, C) dans la librairie.
Il n'y a pas de base non ordinaire dans la librairie.

Sequence length distribution : La longueur des séquences est la même pour toute la librairie, 101 bases.

Sequence duplication levels : Graphique du pourcentage de duplication des séquences.
On observe environ 30% des séquences avec un pourcentage de duplication > 10 selon le graphique théoriquement. Cependant ces séquences ne sont pas détectées dans la librairie.

Il est pourtant indiqué que 50,25% de la librairie proviendrait de séquences dupliquées.

Overrepresented sequences : Aucune séquences n'est sur-représenté dans la librairie.

Adapter content : Analyse générique de tous les Kmers de votre bibliothèque pour trouver ceux qui n'ont pas une couverture uniforme sur toute la longueur des lectures.

Les séquences ont une couverture uniforme sur toute la longueur des lectures.

Interprétation de FastQC

Les résultats de l'analyse FastQC démontre des reads avec une bonne qualité pour l'analyse quantitative. Donc les phases de pré-traitement et avant cela l'analyse RNA-seq ont fonctionné correctement.

3.2.2 Sortie trimgalor

Trimalgor est un outil utilisé pour couper les adaptateurs (a.k.a trimming) des séquences dans un fichier fastq. Il détecte et ajuste automatiquement la séquence de l'adaptateur à couper.

En sortie de Trimalgor, un rapport sur l'adaptateur coupé dans chaque sample est généré, contenant un sommaire, sur le nombre de séquences traitées, le nombre qui contient l'adaptateur et le nombre de séquences qui ont passé le filtre (??).

L'adaptateur détecté et coupé des reads est : AGATCGGAAGAGC.
Il a été détecté et coupé de 578,951 séquences pour un brin et dans 473,901 séquences sur l'autre brin.

Automatiquement FastQC à été lancé sur les données une fois le “trimming” terminé. Il n’y a pas de différence significative dans les résultats avant et après Trimalgor. Les lectures restent de qualité.

3.2.3. Sortie genome

Le dossier `genome` contient les différents fichiers du génome, simple et indexé.

3.2.3 Sortie pipeline_info

Ce dossier contient des rapport générés par Nextflow sur l’exécution de la pipeline.

3.2.4 Sortie star_salmon

STAR et Salmon sont des outils pour l’alignement et la quantification des reads respectivement.

Leur résultats étant contenus dans le rapport MultiQC, ils seront explicités dans cette partie.

3.2 Interprétation du rapport MultiQC

Le rapport MultiQC est une agrégation des résultats d’analyse bioinformatique de la pipeline.

Read distribution : Biotype Counts , QualiMap, RSeqQC

La distribution des reads montre que ces derniers s’alignent majoritairement sur des exons (88%). Ce qui démontre un bon résultat pour ces données RNA-seq.

Les résultats de RSeQC : read duplication montre une duplication d’une duplication de certains gènes dans les 2 échantillons. Ce qui rejoint les résultats observés avec FastQC.