Mini Rapport sur un pipeline de nextflow LaTeX

Moussa Sow

October 11, 2023

Master 2 Bioinformatique biologie des systèmes
Université Tououse III Paul Sabatier

Contents

1 Introduction

Les travaux en génomique impliquent souvent l'analyse de grandes quantités de données génomiques. Ces données complexes nécessitent dans la plupart du temps une étape de prétraitement et de netoyage afin d'être mieux exploitable. La conception de pipeline est souvent de rigueur pour effctuer les différentes étapes de l'anlyse selon les objetifs finaux recherchés. De ce fait, nexflow est aujourd'hui un des outils les plus utilisés pour réaliser des tâches successives de façon automatique. Sa popularité vient du fait qu'il permet de mieux gérer les ressources (calcul, mémoire), d'accéder à diverses programmes sans forcémment les installer un par un et donne la possiblité de suivre le pipeline qui est exécuté. Dans ce présent rapport nous avons utilisé un pipeline de nexflow pour faire du RNAseq dans un premier lieu avec un génome de tomate et en second lieu avec un génome bactérien. Nous avons détailé sur la manière de mettre en place ce pipeline et avons choisi une partie des résultats qu'on a essayé de détaillé au mieux. L'éxécution du pipeline a fait appel aux programmes suivant : nextflow nf-core/rnaseq , fastqc,rsem,samtools, trimgalore, deseq2.

- nf-core est un programme géré par une communauté considérable visant à collecter un ensemble organisé de pipelines d'analyse construits à l'aide de Nextflow.
- FastQC est un outil de contrôle qualité pour les données de séquence à haut débit, écrit par Simon Andrews du Babraham Institute de Cambridge.
- RSEM est un progiciel permettant d'estimer les niveaux d'expression de gènes et d'isoformes à partir de données RNA-Seq.
- SAMtools est un ensemble d'utilitaires pour interagir avec et post-traiter les alignements de lecture de courtes séquences d'ADN dans les formats SAM, BAM et CRAM, écrits par Heng Li.
- TrimGalore est un wrapper autour de Cutadapt et FastQC pour appliquer de manière cohérente l'adaptateur et le découpage de qualité aux fichiers FastQ, avec des fonctionnalités supplémentaires pour les données RRBS.
- DESeq2 est un package de r utilisé pour l'analyse différentielle d'expression génique à partir de données de séquençage d'ARN (RNA-Seq).

2 Matériel et Méthode

2.1 Création d'un répertoire de travail

Une première étape a été de télécharger (à l'aide de la commande wget) les données avec le lien qui a été fourni. En vue d'une bonne gestion des fichiers d'entrées et de sorties qui seront manipulés, nous avons créé un répertoire nextflow/ dans laquelle il y a 3 sous répertoires : un répertoire genome/ qui contient le génome de référence sous format FASTA un répertoire annotation/

qui contient le fichier d'annotation du génome de tomate sous format GTF un répertoire fastq/ qui contient les séquences des reads sous format FASTQ Pour lancer un pipeline sur le cluster de calcul BioInfo Genotoul il est plus efficace de mettre dans un fichier de lancement sbatch avec l'extension .sh toutes les lignes de commandes nécessaires. En outre, il faut également deux autres fichiers à savoir un fichier de configuration qui surcharge le fichier en local et un fichier de description des échantillons.

2.2 Fichier de configuration

Ce fichier permet de suivre le déroulement du pipeline qu'on a lancé avec trace qui a un attribut "enabled = true" et enregistre un fichier de sortie dans le répertoire par lequel on a lancé le fichier du script de lancement (file = pipeline_trace.txt) et dans fields on a le type d'informations qu'on veut suivre comme identifiant de la tâche (task_id) ou encore le nom du script lancé (script).

```
jacinthe@genologin2 ~/work/nextflow $ more sm_config.cfg
trace { enabled = true
                file = 'pipeline_trace.txt'
                fields = 'task_id,name,status,exit,realtime,%cpu,rss,script'
                }
```

2.3 Fichier de description des échantillons tests

Dans ce fichier csv, on trouve les échantillons de séquences fastq de tomates ainsi que leur replica.

```
jacinthe@genologin2 ~/work/nextflow $ more inputs.csv
group,replicate,fastq_1,fastq_2,strandedness
mutant,1,/home/jacinthe/work/nextflow/fastq/MT_rep1_1_Ch6.fastq.gz,/home/jacinthe/work/nextflow/fastq/MT_rep1_2_Ch6.fastq.gz,unstranded
wild,1,/home/jacinthe/work/nextflow/fastq/WT_rep1_1_Ch6.fastq.gz,/home/jacinthe/work/nextflow/fastq/WT_rep1_2_Ch6.fastq.gz,unstranded
```

2.4 Fichier de lancement du pipeline

Dans ce script on a utilisé nf-core/rnaseq qui est un pipeline de nextflow adapté pour analyser les données de séquençage d'ARN obtenues à partir d'organismes avec un génome de référence (fasta) et une annotation (gtf). Il prend une feuille d'échantillons et des fichiers FASTQ en entrée (input), effectue un contrôle qualité, un découpage et un (pseudo-) alignement, et produit une matrice d'expression génique et un rapport de contrôle qualité détaillé. Après avoir spécifié le langage qui est utilisé (bash), on a indiqué au cluster le nom du job (nfcornernaseq) et la mémoire allouée est fixée à 6 Go. Ensuite, on élimine tous les modules chargés auparavant avant de charger le module de nexflow. nf-core/rnaseq a été utilisé avec les paramètres ou arguments suivants :

```
jacinthe@genologin2 ~/work/nextflow $ more run_pipeline.sh
#!/bin/bash
#SBATCH -J nfcornernaseq
#SBATCH -p unlimitq
#SBATCH --mem=6G


module purge
module load bioinfo/nfcore-Nextflow-v21.04.1

input=/home/jacinthe/work/nextflow/inputs.csv
gtf=/home/jacinthe/work/nextflow/annotation/ITAG2.3_genomic_Ch6.gtf
fasta=/home/jacinthe/work/nextflow/genome/ITAG2.3_genomic_Ch6.fasta
config=/home/jacinthe/work/nextflow/sm_config.cfg

nextflow run nf-core/rnaseq -r 3.0 -profile genotoul --input $input --fasta $fasta --gtf $gtf --aligner star_rsem -c $config -resume
```

```
--fasta $fasta --gtf $gtf --aligner star_rsem -c $config -resume
```

profile : (genotoul) correspond au profil de configuration du cluster de genotoul. - input, - - fasta, - - gtf : sont les entrées qu'on a évoqué ci-dessus. - - aligner : on définit le programme pour aligner nos reads fastq sur le génome de référence, ici on a utilisé star_ sem. resume : permet de gagner du temps surtout si le pipeline doit durer. Lorsqu'il est utilisé à partir du deuxième lancement du script, Nextflow utilisera les résultats mis en cache de toutes les étapes du pipeline où les entrées sont les mêmes, en continuant là où elles sont arrivées précédemment. config : pour retrouver le fichier de configuration Par la suite, on peut maintenant lancer le script sbatch, et avec la commande seff on peut suivre l'avancement du processus, si ça marche ou s'il se plante pour une quelconque raison. L'un des avantages de nextflow est qu'il permet de faire quasiment toutes les étapes de l'analyse de rnaseq à savoir l'alignement des reads, contrôle de qualité, indexation, comptages des transcrits et même des analyses statistiques. On peut retrouver les étapes de prétraitement des données, telles que la qualité des lectures, l'adaptateur trimming, le contrôle de qualité dans cette figure ci-dessous :

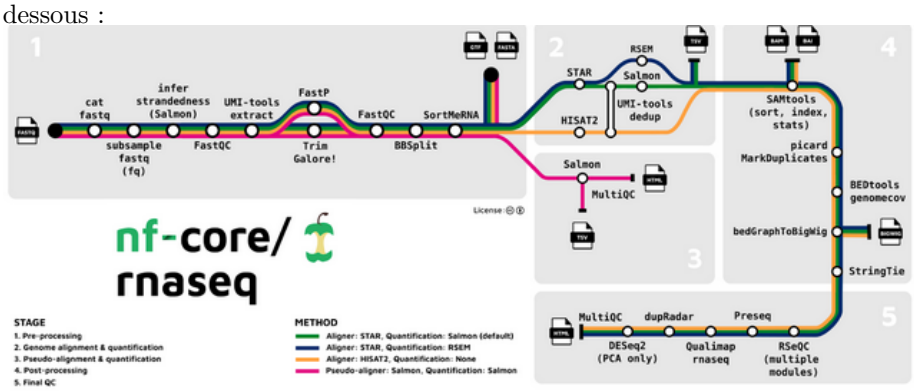


image (https://github.com/nf-core/rnaseq/blob/master/docs/images/nf-core-rnaseq_metro_map_grey.png)

3 Résultats obtenus

A la fin de l'exécution de tous les tâches du script de lancement un répertoire results est contenant un ensemble de résultats des différentes étapes dispatcher sur cinq sous répertoires tels que : fastqc/, genome/, multiqc/, pipeline_info/, star_rsem/ et trimgalore/.

- Dans fastq/ on retrouve les séquences des reads ainsi que leur équivalent en fastqc qui est permet d'analyser la qualité de séquençage.
- Dans genome/ on a les génomes de référence indexés.
- Dans pipeline_info/ se trouve des informations relatives au temps d'éxécution du pipeline ou encore la versions des programmes utilisés. On peut visualiser avec firefox par exemple ces résultats.

- Enfin, dans multiqc/, star_rsem/ et trimgalore/ il y a un condensé de l'ensemble des résultats obtenus avec ces programmes du même nom.

TrimGalore est un wrapper autour de Cutadapt et FastQC pour appliquer de manière cohérente l'adaptateur et le trimming de qualité aux fichiers FastQ. Le programme star_rsem combine les étapes de mapping et d'estimation du niveau d'expression des gènes. Et multiqc permet de regrouper l'ensemble des résultats obtenus à la fin de l'exécution des différentes étapes du pipeline pour fournir une visualisation d'un rapport sous format html. Ce rapport peut être ouvert avec firefox, et pour la suite on se focalise sur les résultats de ce rapport.

3.1 Interprétation des principaux résultats

General Statistics

Copy table | Configure Columns | Plot | Showing 6/6 rows and 21/22 columns.

Sample Name	M Reads Mapped	% rRNA	dupInt	% Dups	5'-3' bias	M Aligned	% Alignable	% Proper Pairs	Error rate	M Non-Primary	M Reads Mapped	% Mapped	% Proper Pairs	M Total seqs	% Dups
mutant_R1	3.3	0.00%	0.00%	17.3%	1.43	1.6	99.2%	78.3%	0.16%	0.1	3.2	99.3%	99.3%	3.2	
mutant_R1_1															49.7%
mutant_R1_2															49.2%
wild_R1	2.7	0.00%	0.00%	18.3%	1.43	1.3	99.3%	76.9%	0.16%	0.1	2.6	99.4%	99.4%	2.7	
wild_R1_1															48.5%
wild_R1_2															48.2%

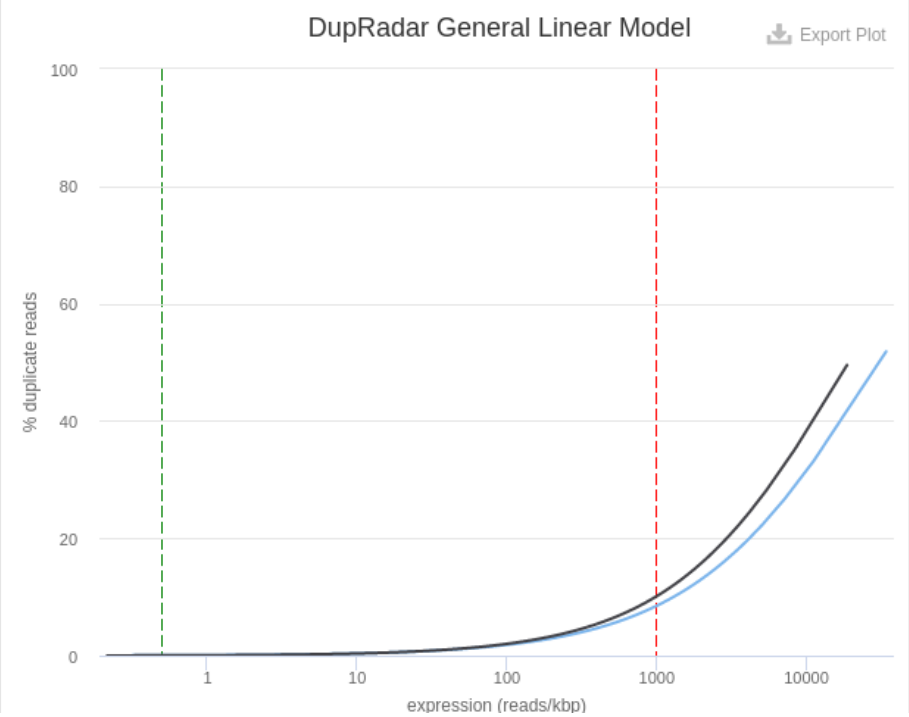
Dans ce tableau (general statistics), on a le résumé général sur nos alignements :

- On a un plus de 99% de reads bien alignés que cela soit chez le mutant comme chez le sauvage, même constat aussi pour les reads mappés.
- Le taux de GC est entre 41 et 42% pour les mutants comme les sauvages, ce qui est un bon signe sachant que pour le génome de la tomate (Solanum lycopersicum) ce taux varie entre 40 et 45
- Enfin, si le taux d'erreur (0.16%) est faible, le taux de reads duplicatas est assez élevé pour le mutant (17,3%) comme pour le sauvage (18,3%) donc il sera nécessaire de filtrer.

Dans la suite, on a un diagramme qui montre le pourcentage de reads alignés, on a un taux à 100% ce qui est cohérent par rapport aux résultats du tableau ci-dessus.

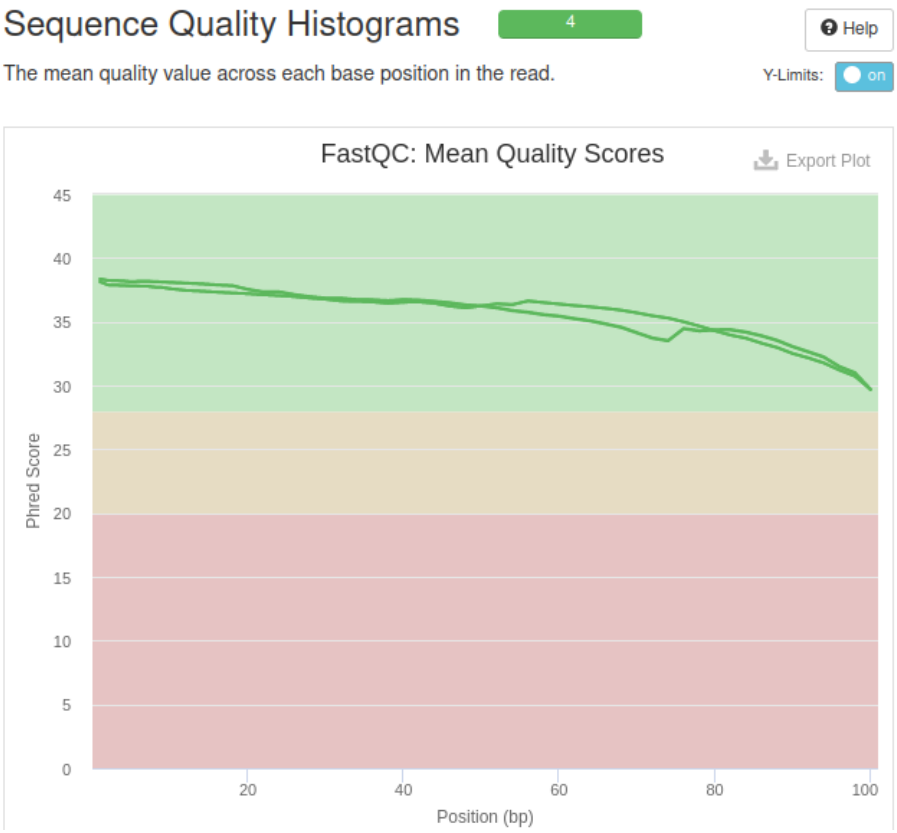
Relation entre taux d'expression génique et le nombre de reads dupliqués.

On peut voir que le taux élevé de duplicatas s'accorde avec un taux élevé de reads, ce qui est bien normal du fait que les gènes fortement exprimés ont plusieurs reads dupliqués.

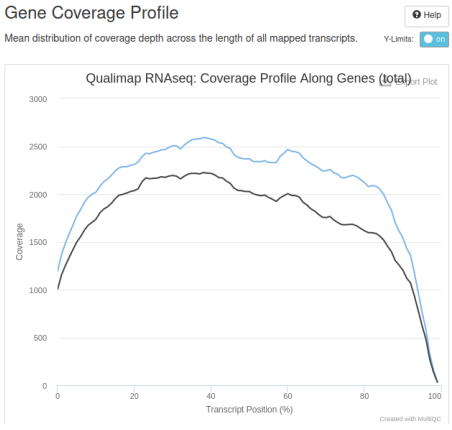


Histogramme de qualités des séquences

Ce graphique représente la qualité (score Phred, en ordonnée) de chaque base (en abscisse) pour tous les reads des 4 échantillons du jeu de données. Le code couleur indique les scores de très bonne qualité en vert, bonne qualité en orange et mauvaise en rouge. Généralement, la qualité baisse en fin de reads. Ce qui fait qu'on a des scores élevés qui indique des reads de très bonne qualité. Ce constat est confirmé par la figure sur les scores de qualité par séquences (voir figure ci-dessous).



Profondeur de séquençage :



On a globalement un taux de couverture moyen des transcrits assez élevé, une baisse à la fin est bien normal et souvent le cas.

Distribution de pourcentage de GC

Semblable aux résultats précédents sur le contenu des séquences, à nous assurer qu'il existe une distribution de forme raisonnablement normale autour du contenu GC attendu pour un génome/exome de référence de la tomate. De fortes asymétries, des formes multimodales ou des pointes abruptes pourraient indiquer des erreurs de séquençage ou de contamination. Ce qui n'est pas le cas ici.

3.2 Conclusion partielle

Pour conclure, on peut voir que l'exécution des différentes tâches du pipeline s'est bien déroulée. Il n'y a pas de métriques dans les différentes sorties qui indiquent des erreurs majeurs de mapping ou de paramètres qui sont éloignés des valeurs attendues.

4 Partie sur l'exercice 4

4.1 Enoncé

Lancer ce pipeline sur des données NCBI. Choisir deux échantillons sur le site du NCBI avec la référence fasta associées ; puis lancer ce pipeline RNAseq sur ces deux samples. Expliquer vos démarches et les résultats obtenus.

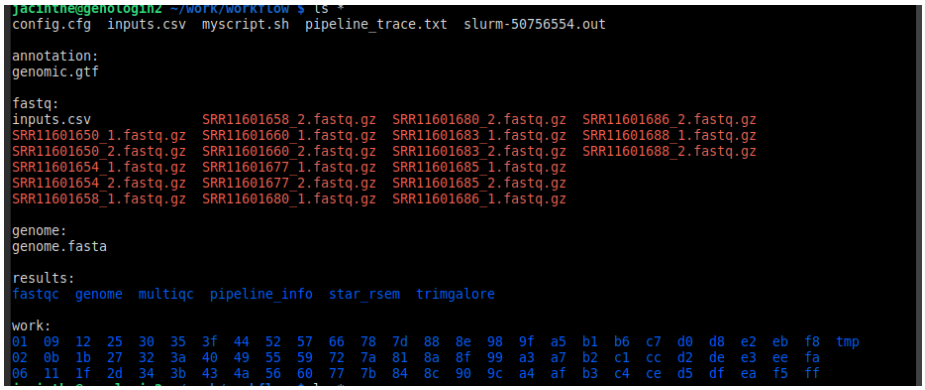
4.2 Introduction

Dans notre recherche de génome, on était confronté au fait que le pipeline prenait beaucoup de temps à tourner, et à la survenue d'erreurs lors de l'exécution, il nécessite beaucoup de temps d'attente pour voir si ça marche ou pas. De ce fait, on a choisi une espèce bactérienne avec un génome de petite taille (moins de 1000 pb) : la bactérie *Mycoplasma genitalium*. C'est un mycoplasme endoparasitaire des cellules épithéliales du tractus urogénital humain, agent infectieux pathogène pour l'être humain, responsable d'urétrites et d'autres infections sexuellement transmissibles, éventuellement en association avec un autre mycoplasme.

4.3 Données utilisés

Nous avons téléchargé le génome de référence et les annotations du mycoplasma dans NCBI genome, nous avons repris les données de séquençage (20 reads) utilisées lors d'un projet (numéro PRJNA627746) de Plummer et al. 2020 dans la banque de données ena (european nucleotide archive).

Nous avons conservés la même architecture de répertoires que celui de nextflow/ sauf que le nom du répertoire est workflow/ dans work/ Les répertoires et sous répertoires de travail.



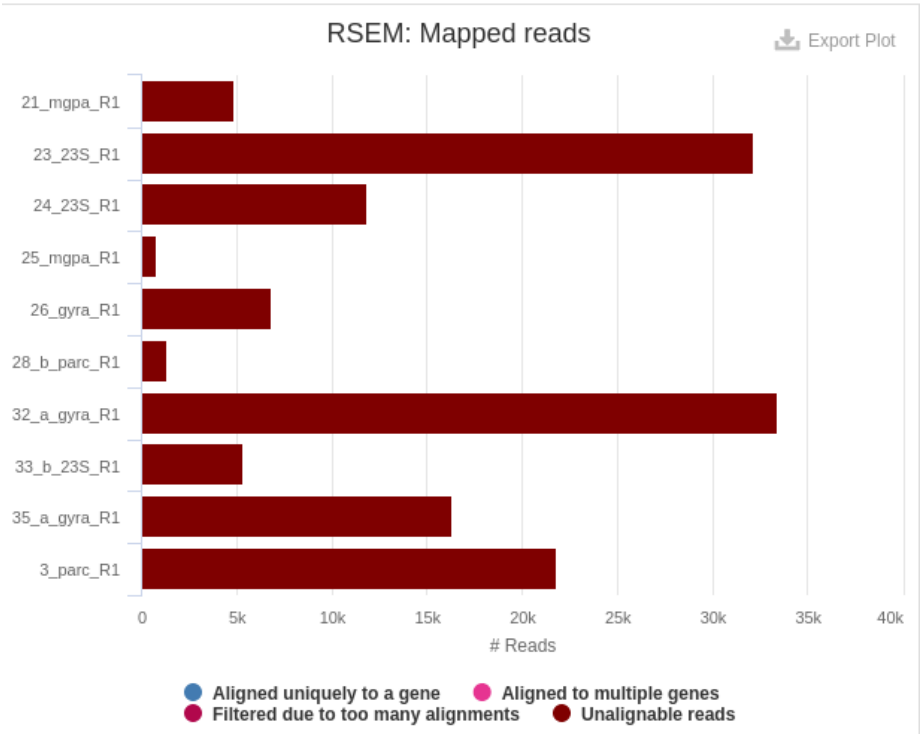
Après avoir réglé tous les paramètres convenablement, nous avons pu avoir des résultats avec un temps d'exécution assez court. Donc pour la suite, nous allons analyser la sortie du rapport html MultiQC; ce fichier peut être avec firefox.

4.4 Rapport html de MultiQC

Le premier tableau (general statistics) apportent des informations sur la qualité des reads permettant de juger si les échantillons sont de bonne qualité et peuvent donc être utilisés pour les étapes suivantes ou non. Cependant, il n'existe pas de seuil consensuel basé sur les métriques FastQC pour classer les échantillons comme étant de bonne ou de mauvaise qualité.

Sample Name	% Alignable	% Dups	% GC	Length	M Seqs	% BP Trimmed	% Dups	% GC	Length	M Seqs
21_mgpa_R1	0.0%						94.9%	37%	185 bp	0.0
21_mgpa_R1_1		95.5%	36%	204 bp	0.0	1.4%	95.4%	36%	200 bp	0.0
21_mgpa_R1_2		95.0%	38%	189 bp	0.0	1.2%				
23_23S_R1	0.0%						84.3%	49%	113 bp	0.0
23_23S_R1_1		95.0%	50%	124 bp	0.0	1.3%	95.0%	51%	123 bp	0.0
23_23S_R1_2		84.5%	49%	119 bp	0.0	5.6%				
24_23S_R1	0.0%						88.1%	48%	107 bp	0.0
24_23S_R1_1		93.2%	51%	122 bp	0.0	1.2%	93.1%	51%	121 bp	0.0
24_23S_R1_2		88.3%	48%	110 bp	0.0	2.8%				
25_mgpa_R1	0.0%						90.6%	42%	94 bp	0.0
25_mgpa_R1_1		92.1%	39%	109 bp	0.0	1.1%	92.1%	39%	108 bp	0.0
25_mgpa_R1_2		91.0%	42%	94 bp	0.0	0.9%				
26_gyra_R1	0.0%						89.8%	42%	77 bp	0.0

Le premier constat qu'on a est que les pourcentage de read alignés alignés en seul emplacement du génome (%Alignable) issue du programme de star est quasiment nul, ce qui est un signal que les échantillons sont de mauvaise qualité. Ce résultat n'est pas cohérent avec le fait qu'il y a de fortes de duplicatas dans les reads qui varient entre 51 et 9,1%, ce qui veut dire qu'on a une bonne profondeur de séquençage et peu d'erreurs liés à la contamination. Il n'y a pas de reads qui s'aligne sur le génome de référence comme le montre la sortie du programme de



rsem.

Reads mappés

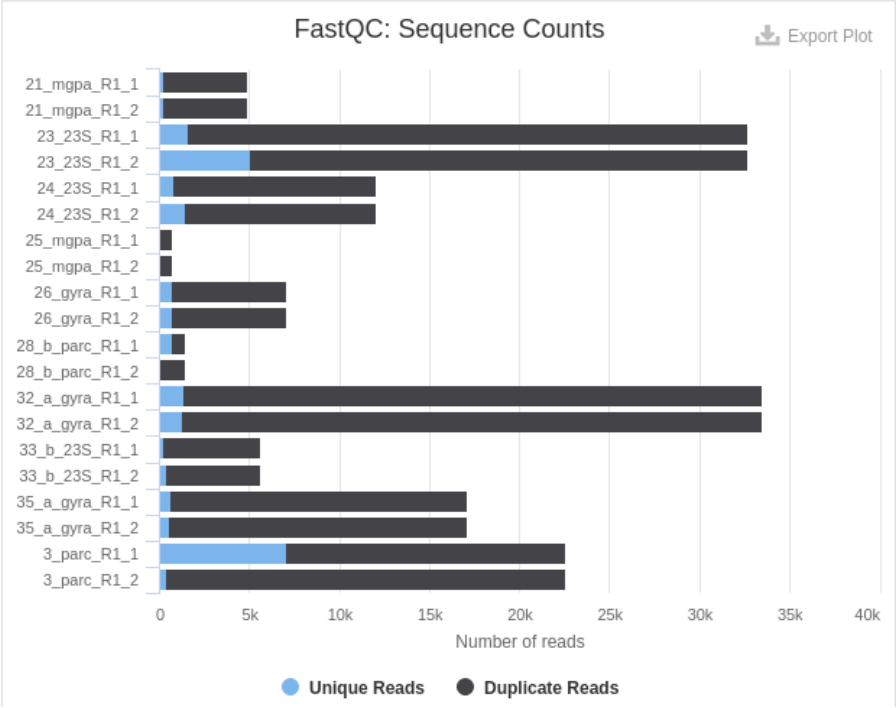
il y a en réalité trop peu de reads alignés sur un seul emplacement du génome.

Nombre de séquences

Le graphique montre le nombre de séquences pour chaque échantillons.

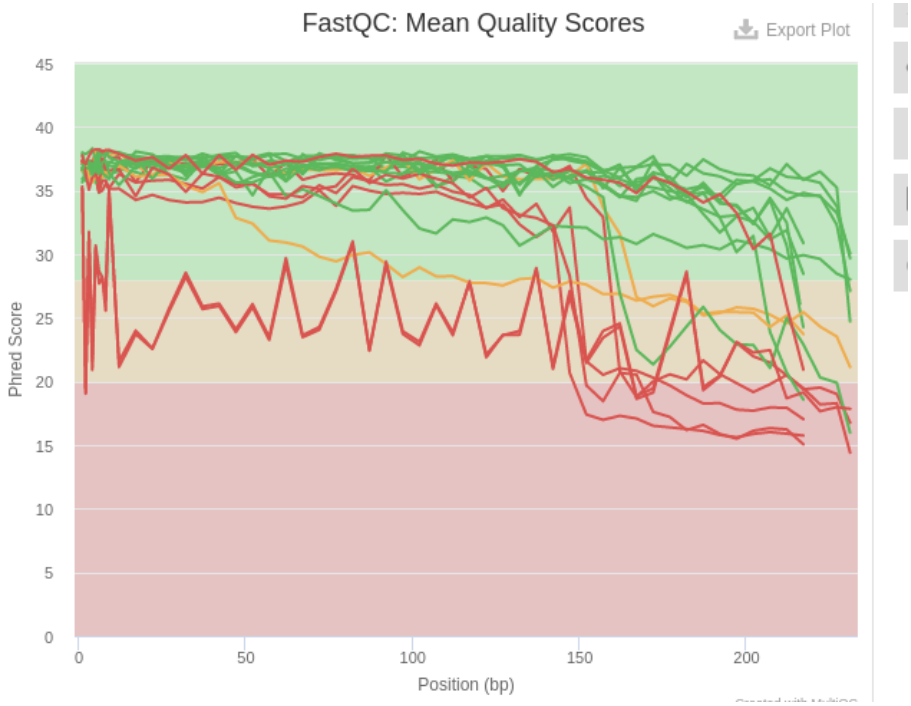
Le génome de ce parasite étant minimaliste c'est-à-dire il a dans son génome le strict minimum pour survivre et se reproduire, donc il n'est pas sujet à des duplications très importantes qui peuvent être coûteuses. De ce fait ces duplications s'expliquent:

- soit par des biais d'ordre technique comme lors de l'amplification par PCR ou bien par une contamination humaine lors de la manipulation.
- soit par une faible complexité des librairies dû à la présence de beaucoup de reads de courte taille.
- soit par une bonne profondeur de séquençage. Vue le peu de reads qui s'alignent sur le génome de référence, il est plus probable que l'on est dans le deuxième cas de figure d'une part et d'autre part dans la table 'Overrepresented sequences' cette tendance est bien représenté avec la tottalité des reads ayant une surepresentation. (voir Annexe 1)



Qualité des séquences

La mauvaise qualité des reads est plus explicite avec l'histogramme de qualité, pour 9 sur 20 échantillons la qualité est assez mauvaise même au début des reads, ils sont mis en évidence ici en couleur rouge. Ce état de fait se superpose sur le fait qu'il y a pour la majeure partie des reads des taux de GC assez faibles mais aussi des reads de longueurs courtes (Annexe 2). Il y a majoritairement de reads de petite taille que de reads de taille moyenne ou grandes. Ce qui peut expliquer aussi le pourcentage asséz élevés de reads dupliqués.



5 Conclusion

Actuellement nextflow est assez populaire dans le milieu de recherche en génomique et en bioinformatique vue son efficacité et sa tendance à faciliter l'automatisation de plusieurs pipelines. Il est nécessaire d'avoir une bonne configuration des répertoires de travail pour organiser les données en entrée, les scripts ainsi que les résultats. Nous avons utilisé le pipeline de nf-core/rnaseq qui fait appel de façon automatique à différents programmes. Nous avons pu constater que l'état des reads peut être de bonne ou de mauvaise qualité et avons proposé des explications sur cette différence par rapport à ce qui est attendu en inspectant par exemple les statistiques générales de nos alignements ou évaluer la qualité des reads et analysé la distribution du contenu en GC. Dans le cas du génome Mycoplasma, on a pu voir que les librairies de reads n'étaient pas de bonne qualité et qu'il y avait beaucoup de reads de quelques dizaines de bp (paires de basses) et peu de reads de taille moyenne ou grande. En résumé, la conception et l'exécution de ce pipeline rnaseq sur nextflow nous a permis de mieux comprendre le fonctionnement et l'utilité de cet outil bioinformatique dans le cadre des analyses en génomique qui nécessite souvent un flux de travail ou workflow.

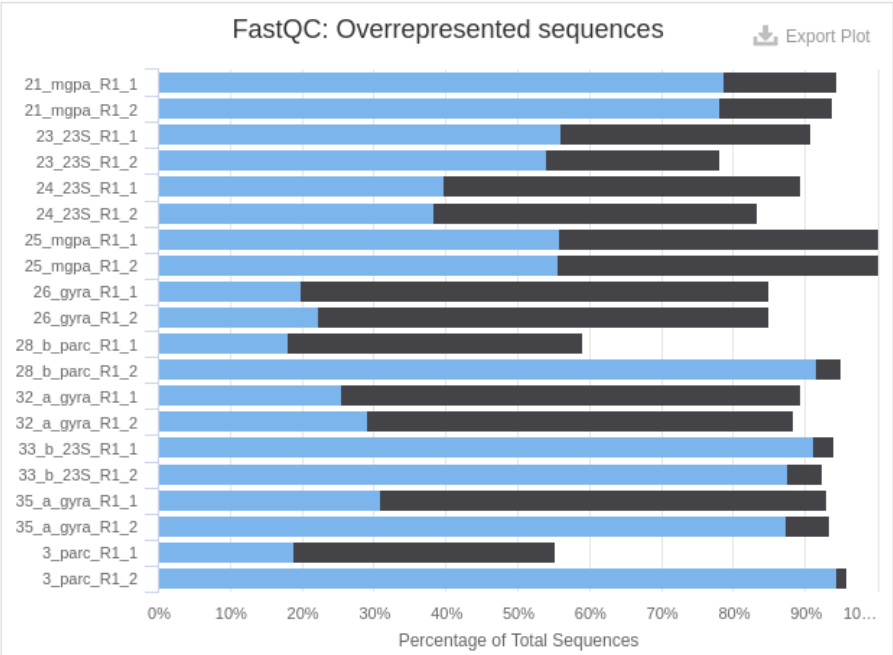
5.1 Référence

E.L. Plummer, G.L. Murray, K. Bodiyabadu, J. Su, S.M. Garland, C.S. Bradshaw, T.R.H. Read, S.N. Tabrizi, J.A. Danielewski,A custom amplicon sequencing approach to detect resistance associated mutations and sequence types in Mycoplasma genitalium,Journal of Microbiological Methods,Volume 179,2020,106089,ISSN 0167-7012, https://doi.org/10.1016/j.mimet.2020.106089.

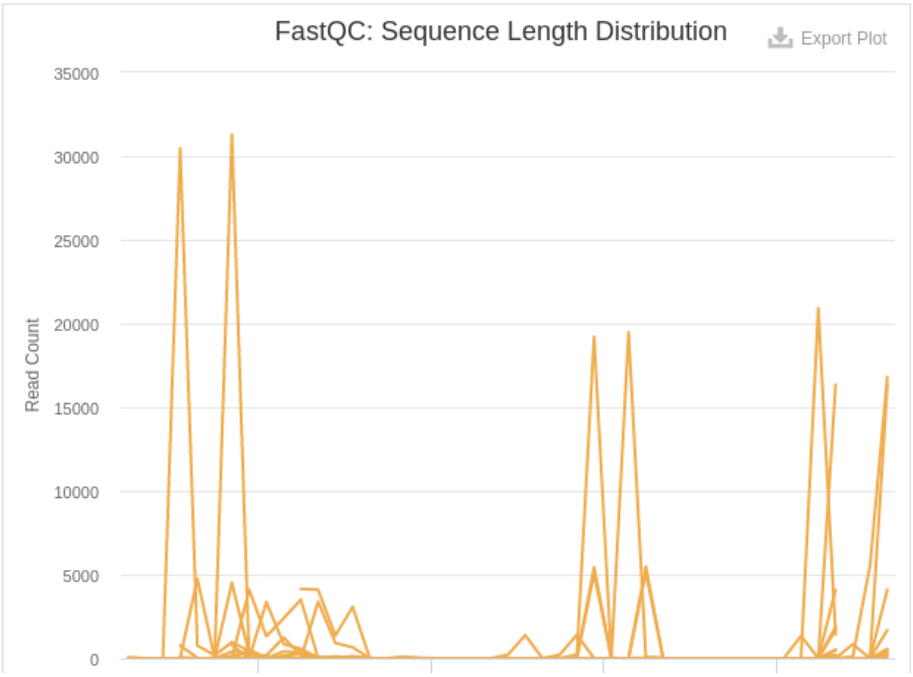
5.2 Annexes

1. Reads surreprésentées par librairie

The total amount of overrepresented sequences found in each library.



2. Distribution des longueurs de reads


FastQC: Sequence Length Distribution
Export Plot
Read Count
35000
30000
25000
20000
15000
10000
5000
0