Rapport Projet Tutoré Nextflow

Leclercq Etienne

07/10/23

1 Introduction

Ce projet à pour but l'utilisation de Nextflow et plus précisement du pipeline nfcore. Ce pipeline est concu pour faciliter l'analyse de données de séquencages accessibles au grand public au travers de plateformes et banques de données comme le NCBI. Nfcore permet notamment d'automatiser différentes analyses avec une excellente répétabilité et une documentation complète. Nous verrons au long de ce rapport comment utiliser ce genre d'outil pour différents exercices traitant petits et grands jeux de données.

2 Exercices

Exercice 1

Ce premier exercice représente simplement le chargement des données et leur affectation dans des dossiers dédiées crées au préalable. Il est nécessaire de se placer dans le répertoire /work puis /nextflow après avoir créé ce dernier. La figure suivante montre l'organisation complète du répertoire /nextflow après ces manipulations :

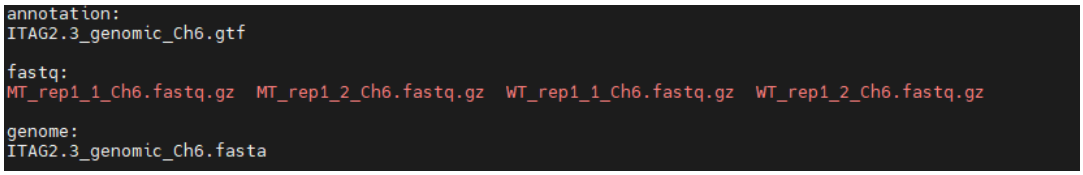

Figure 1: Organisation du répertoire Nextflow

Exercice 2

Dans ce deuxième exercice le but est de créer et paramètrer notre fichier bash de lancement Nextflow. Il nécessaire de créer au préalable de fichier inputs.csv appelé dans notre fichier bash. Ce fichiers comporte simplement des informations sur les échantillons biologiques avec une ligne pour chacun. Chaque colonne est séparée par une virgule on y retrouve le groupe, le nombre de replicats de l'échantillon puis les chemins vers les deux fastq et enfin l'orientation des brins lors de la séquence (on a ici aucune information). Pour finir ce fichier est sous le format .csv car on y stocke des données tabulaires.

```
hortensia@genologin1 ~/work/nextflow $ more inputs.csv
group,replicate,fastq_1,fastq_2,strandedness
mutant,1,/home/hortensia/work/nextflow/fastq/MT_rep1_1_Ch6.fastq.gz,/home/hortensia/work/nextflow/fastq/MT_rep1_2_Ch6.fastq.gz,un
wild,1,/home/hortensia/work/nextflow/fastq/WT_rep1_1_Ch6.fastq.gz,/home/hortensia/work/nextflow/fastq/WT_rep1_2_Ch6.fastq.gz,unst
```

Figure 2: Organisation du ficher inputs.csv

La figure suivante représente le script bash (.sh), c'est à dire le fichier pour lancer le pipeline d'analyse de séquencage RNA-Seq avec le framework nf-core/rnaseq :

```
hortensia@genologin1 ~/work/nextflow $ more run_pipeline.sh
#!/bin/bash
#SBATCH -J EtienneLeclercq
#SBATCH -p unlimitq
#SBATCH --mem=6G

module load bioinfo/nfcore-Nextflow-v21.04.1

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

nextflow run nf-core/rnaseq -r 3.0 -profile genotoul --input $input --fasta $fasta --gtf $gtf --aligner star_rsem -c $config --ma
6G --max-time 1d --name EtienneLeclercq
```

Figure 3: Organisation du ficher run pipeline.sh

Pour répondre à l'exercice plusieurs paramètres ont été modifié :

-Changement du nom du job sur le cluster : SBATCH -j EtienneLeclercq,–name EtienneLeclercq

-Révision et profil demandé : -r 3.0 -profile genotoul

-Mémoire max et temps max : SBATCH –mem=6G,config –maxMemory 6G –max-time 1d

Pour surveiller l'avancement du job on peut utiliser la commande seff couplée à son ID. La figure 3 suivante montre le résultat de cette requête :

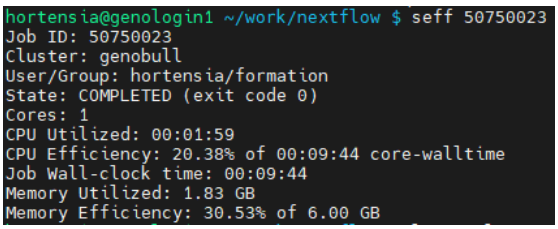

Figure 4: Sortie du seff

Cet affichage nous permet d'obtenir différentes informations importantes comme état du job FAILED/RUNNING/COMPLETED, le nombre de coeurs et cpu utilisés ainsi que leur efficacité. Ces données couplées à la mémoire utilisée permettent d'adapter les paramètres dans le script bash si jamais. Si jamais l'état est FAILED il possible de récupérer les messages d'erreurs dans le slurm avec la commande grep.
Pour finir, il est possible d'inclure dans le script bash l'option -resume. Cette option indique à Nextflow de reprendre l'exécution du workflow à partir du point où il s'était arrêté en cas d'interruption ou d'échec précédent. Il est utilisé pour la gestion des travaux interrompus ou pour reprendre l'exécution d'un workflow existant.

Exercice 3

Cet exercice consiste en l'analyse des différents dossiers compris dans le dossier results obtenu à la fin du job. On a un dossier multiqc que nous aborderons plus tard mais également les dossiers suivants:

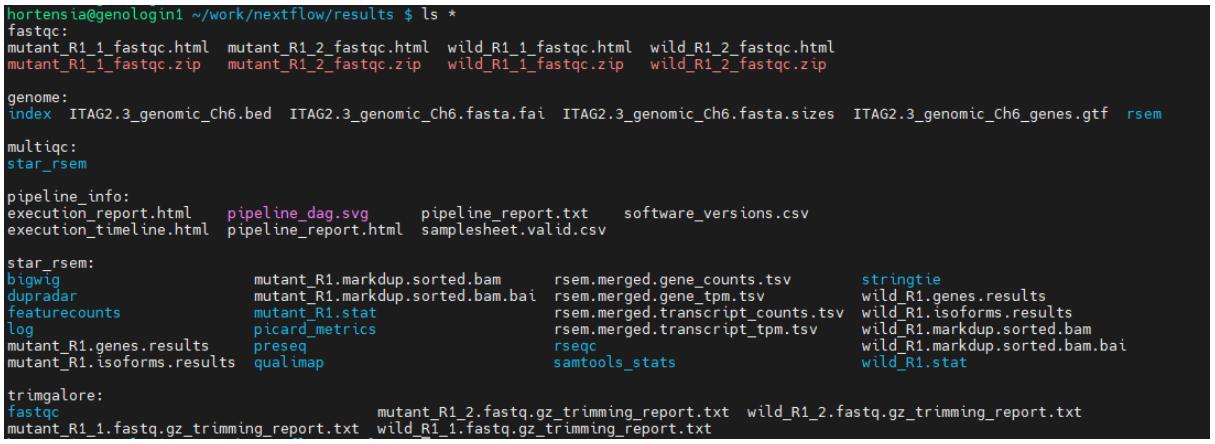

Figure 5: Répertoire results

On retrouve donc les répertoires:

-fastqc: contient les résultats de l'analyse de qualité des données brutes à l'aide de FastQC.

-genome: contient le genome de référence en format .bed ainsi que d'autres informations et descriptions.

-pipeline info:contient des informations spécifiques au pipeline d'analyse, notamment des rapports d'exécution, des rapports de pipeline, un fichier de suivi temporel de l'exécution, un fichier DAG (Directed Acyclic Graph) et des informations sur les versions des logiciels utilisés. Le fichier DAG est un graphe qui représente les dépendances entre les différentes tâches et processus.

-star rsem:contient des résultats liés à l'alignement des données à l'aide de STAR et à l'estimation de l'expression génique à l'aide de RSEM. On peut y trouver des fichiers BAM, des fichiers de comptage d'expression (gene counts.tsv et transcript counts.tsv), des fichiers de TPM (transcript per million), des fichiers de statistiques et d'autres résultats associés à ces étapes.

-trimgalore: contient des résultats liés à l'outil trimgalore. Cet outil est argement utilisé en bioinformatique pour la qualité et la découpe des données de séquençage génétique. On a donc des ces fichiers les étapes de prétraitement des données.

Multiqc :

Multiqc est un outil bioinformatique qui permet d'agréger et de résumer les résultats des diverses analyses décrites précédemment. Il est placé à la fin du pipeline mais peut également être compilé manuellement. Il produit un fichier .html complet ainsi qu'un nouveau répertoire contenu dans le répertoire results contenant différents fichiers de données utilisables pour de nouvelles analyses plus complexes.

Pour cette première analyse on peut donc observer dans ce rapport un résumé du pipeline ainsi que les versions des outils bioinformatiques utilisés par faciliter la lecture et la compréhension de l'analyse par le lecteur. Les statistiques générales révèlent un excellent score d'alignement de plus de 99% avec un taux d'erreur très faible de 0.16%. On peut aussi observer le taux mapping (se rapprochant du taux d'alignement) c'est à dire la proportion de lectures qui ont pu être alignés et cartographiés avec succès sur le génome de référence. Ces premières statistiques démontre une excellent qualité des données de séquençage ce qui est essentiel pour des analyses ultérieures. La partie biotypes counts permet de se rendre compte de la proportion de chaque biotype dans les échantillons, ici on constate que les lectures codent uniquement pour des protéines.
L'outil DupRadar évalue et quantifie la présence de duplications d'ADN dans les échantillons. On peut noter une augmentation significative après le seuil de 1000 lectures/kpb ce qui pourrait indiqué une saturation du séquençage ou tout simplement un grand nombre réel de duplications. L'outil Picard précise ces résultats en indiquant 80% de lectures uniques ce qui semble valider l'hypothèse de saturation. La courbe de complexité de Preseq continue dans cette lignée et nous montre une bonne diversité de la bibliothèque et indique qu'un séquençage supplémentaire n'est pas nécessaire pour obtenir plus de nouvelles information. Pour compléter cette partie l'outil Rsem nous montre que 98% de ces lectures sont alignées de manière unique à un gène. On a donc en résumé une excellente qualité d'alignement.
Dans un autre registre l'outil Qualimap nous informe sur l'origine génomique des reads avec les pourcentages d'exons (84%),introns(8%) et intergènes(8%). L'outil permet aussi d'étudier le pourcentage de couverture des gènes par le biais de profils.
L'outil RSeQC permet également de vérifier la qualité des données de séquencage alignées. On y retrouve la distribution des reads sur les différentes régions génomiques avec 80% sur des exons codant pour des protéines comme vu précédemment. Puis d'autres métriques dont la saturation sur les duplications au niveau des jonctions exon-exon. Un bon score comme ici démontre que la grande majorité de ces dernières sont couvertes par les lectures. On retrouve pour finir diverses statistiques sur les fichiers BAM, qui sont des fichiers d'alignement d'ARN. L'outil Samstools nous redonne certaines informations déjà abordées comme le pourcentage de mapping ou d'alignement.
La dernière partie du rapport Multiqc aborde fastqc, génère un rapport qui donne un aperçu de la qualité des données brutes, telles qu'elles ont été pro-

duites par le séquenceur avant et après trimming (nettoyage) grâce à cutadapt. Dans cette première étude le trimming ne change que très peu les résultats ce qui signifie que les données brutes ne contenaient pas de séquences de mauvaise qualité (100% des lectures sont passées au travers du filtrage). On a donc pour chaque échantillon plus de 50% de lectures uniques de très bonne qualité avec en majorité 43% de GC et quasiment jamais de N (base nucléotidique indéterminée. On a également moins d'1% de séquences sur-représentées et pas d'adaptateur contaminant. Néanmoins on observe un problème au niveau du contenu de la séquence par base, malgré un bonne répartition de 25% pour chaque base en moyenne on observe le statut fail. Ceci est peut-être du à des seuils de qualité stricts, mais ceci ne semble par très inquiétant.
En résumé l'analyse initiale indique des données de séquençage de haute qualité, avec un excellent taux d'alignement et une faible duplication, préparant ainsi le terrain pour des analyses ultérieures.

Exercice 4

Ce dernier exercice consiste "simplement" en la réutilisation du pipeline mais cette fois sur des données du NCBI. Les échantillons choisis sont 3 SRR correspondant à des séquences liées aux ovaires, au cervau et aux branchies de l'espèce Gadeus Morhua, la morue de l'Atlantique.
La partie la plus dure à été la téléchargement des données avec des essais de plusieurs méthodes sur systèmes d'exploitation différents et même ordinateurs différents (problèmes de fichiers .dll étonnant). Au final, la méthode concluante consistait seulement en l'utilisation du module sratool.3.0.0 depuis le cluster avec un prefectch, fasterq-dump et pour finir un gzip. Le génome et l'annotation ont été récupérés depuis la banque de données Ensembl à cause d'un problème d'ID sur l'annotation du NCBI causant un arrêt impromptu du pipeline. Ce dernier peut-être retrouvé en annexe avec les fichiers utilisés et les sorties seff et tail du slurm pour une vérification de l'état du job. Malgré un statut failed après 5h de running le pipeline nf-core/rnaseq à été compilé et le répertoire results/ est complet.

L'analyse de ce pipeline sur les trois échantillons SRR permet de caractériser et de quantifier l'expression génique dans différents tissus ou conditions biologiques de l'espèce étudiée. Cela permet de comprendre quels gènes sont actifs, à quel niveau ils sont exprimés, et comment cette expression varie entre les échantillons. Cette information est cruciale pour étudier les mécanismes biologiques, les réponses aux stimuli, les régulations géniques, et bien d'autres aspects de la biologie de l'espèce en question. En résumé, ce pipeline aide à décrypter le profil transcriptomique de l'espèce, ouvrant ainsi la voie à une meilleure compréhension de ses processus biologiques.

Multiqc :

Analyses	cod brain R1	cod gills R1	cod ovary R1
General Statistics	Alignement : 66.1% Taux d'erreur : 0.61% Mapping : 80.7%	Alignement : 86.4% Taux d'erreur : 0.59% Mapping : 92.3%	Alignement : 92.2% Taux d'erreur : 1.01% Mapping : 94.6%
Biotype Counts	Seq. codante protéine : 98% incRNA : 2%	Seq. codante protéine : 98% incRNA : 2%	Seq. codante protéine : 98% incRNA : 2%
DupRadar (à 1000 reads/kpb)	5.5%	18%	18%
Picard	Paires uniques : 72% Paires de duplicats : 8% Non-mappé : 20%	Paires uniques : 62% Paires de duplicats : 30% Non-mappé : 8%	Paires uniques : 74% Paires de duplicats : 21% Non-mappé : 5%
Qualimap Genomic origin of reads	Exons : 49% Introns : 19% Intergènes : 32%	Exons : 54% Introns : 15% Intergènes : 31%	Exons : 68% Introns : 14% Intergènes : 18%
RSeQC Read Distribution	cf.annexe	...	...

Au niveau de l'analyse RSeQC on retrouve également une bonne saturation. Pour cette run de pipeline l'étape de nettoyage/trimming à été plus convaincante en améliorant globalement les status check et plus précisement en supprimant des adaptateurs contaminant. La distribution de longeurs des séquences à aussi été logiquement modifiée en retirant certaines régions problèmatiques. Pour les points positifs : on a donc plus d'adaptateurs contaminants et les séquences ne sont pas sur-représentées. Les scores de qualités sont aussi très bons. On remarque le même phénomène de fail malgré une distribution normale de chaque basse. Pour les points négatifs, on peut signaler un taux de GC légèrement trop élevé se rapprochant de 60% pour les séquences ovariennes et des branchies ainsi qu'un nombre non négligeable de bases N pour l'échantillon ovarien. De plus, sans surprise, on a un taux de duplications problèmatiques pour chacun des échantillons. Pour conclure, l'analyse initiale révèle des données de séquençage de bonne qualité malgré quelques zones d'ombres. Cependant, une préoccupation majeure réside dans le taux excessif de duplications observé dans les échantillons. Cette prévalence élevée de duplications pourrait indiquer une possible saturation du séquençage ou une amplification biaisée pendant la préparation des échantillons. Il serait judicieux d'envisager des étapes de correction et de normalisation des données, telles que la suppression des duplications ou l'utilisation de méthodes avancées de traitement des données, afin d'obtenir des résultats plus fiables et informatifs pour les analyses ultérieures.

Annexe

```
hortensia@genologin1 ~/work/ncbi $ more inputs2.csv
group,replicate,fastq_1,fastq_2,strandedness
cod_ovary,1,/home/hortensia/work/ncbi/fastq/SRR2045415_1.fastq.gz,/home/hortensia/work/ncbi/fastq/SRR2045415_2.fastq.gz,unstranded
cod_brain,1,/home/hortensia/work/ncbi/fastq/SRR2045416_1.fastq.gz,/home/hortensia/work/ncbi/fastq/SRR2045416_2.fastq.gz,unstranded
cod_gills,1,/home/hortensia/work/ncbi/fastq/SRR2045417_1.fastq.gz,/home/hortensia/work/ncbi/fastq/SRR2045417_2.fastq.gz,unstranded
hortensia@genologin1 ~/work/ncbi $ more run_pipeline2.sh
#!/bin/bash
#BATCH -J EtienneLeclercq
#SBATCH -p unlimitq
#SBATCH --mem=6G

module load bioinfo/nfcore-Nextflow-v21.04.1

input=/home/hortensia/work/ncbi/inputs2.csv
gtf=/home/hortensia/work/ncbi/Gadus_morhua.gadMor3.0.110.gtf
fasta=/home/hortensia/work/ncbi/Gadus_morhua.gadMor3.0.dna.toplevel.fa
config=/home/hortensia/work/ncbi/config.cfg

nextflow run nf-core/rnaseq -r 3.0 -profile genotoul --input $input --fasta $fasta --gtf $gtf --aligner star_rsem -c $config --maxMemory
6G --max-time 1d --name EtienneLeclercq -resume
```

Figure 6: Fichiers inputs et run pipeline

```
hortensia@genologin1 ~/work/ncbi $ seff 50757439
Job ID: 50757439
Cluster: genobull
User/Group: hortensia/formation
State: FAILED (exit code 127)
Cores: 1
CPU Utilized: 00:05:09
CPU Efficiency: 1.77% of 04:50:58 core-walltime
Job Wall-clock time: 04:50:58
Memory Utilized: 1.86 GB
Memory Efficiency: 30.93% of 6.00 GB
```

Figure 7: Rapport seff

```
-[nf-core/rnaseq] 3/3 samples passed STAR 5% mapped threshold:
    88.48%: cod_ovary_R1
    85.98%: cod_gills_R1
    77.3%: cod_brain_R1
-
-[nf-core/rnaseq] Pipeline completed successfully-
Completed at: 09-Oct-2023 22:43:42
Duration    : 4h 49m 55s
CPU hours   : 86.6
Succeeded   : 100
```

Figure 8: Document slurm

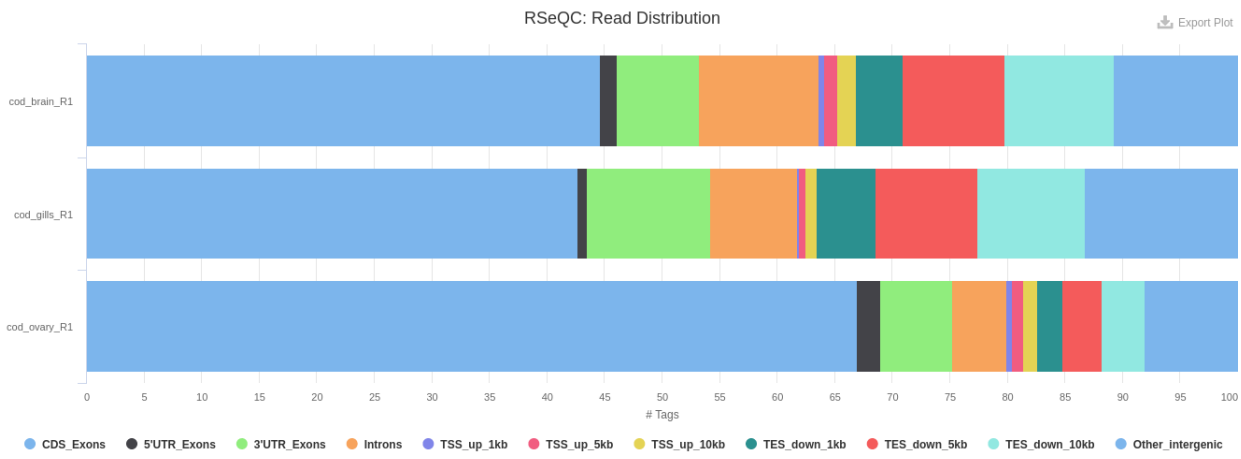

Figure 9: RSeQC