Mini-projet Nextflow

Exploration du pipeline Nextflow RNAvar sur des échantillons RNA-seq du NCBI

Bouchra Boufenouche
Master 2 Bioinformatique et Biologie des Systèmes

3 octobre 2025

Résumé

L'objectif de ce mini-projet est de se familiariser avec le pipeline nf-core/rnavar pour l'analyse de données RNA-seq. Nous allons sélectionner deux échantillons de Gadus morhua sur le NCBI, lancer le pipeline Nextflow pour réaliser la qualité des reads, l'alignement sur le génome de référence, la détection et l'annotation des variants, et interpréter les résultats obtenus

1 Introduction

Le RNA-seq est une méthode de séquençage permettant d'étudier l'expression des gènes et de détecter des variants génétiques au niveau transcriptomique. Cette approche est particulièrement pertinente pour analyser des espèces modèles ou d'intérêt écologique et économique.Parmi elles,L'espèce Gadus morhua, également connue sous le nom de morue de l'Atlantique (morue), est particulièrement étudiée pour ses caractéristiques génomiques. Son génome, d'une taille estimée à environ 830 millions de paires de bases, comporte près de 22 154 gènes identifiés, traduisant un répertoire génétique riche [1]. Environ 25 % de ce génome est constitué de séquences répétitives, qui jouent un rôle majeur dans la dynamique et l'évolution génomique [2].

L'analyse de données RNA-seq issues de cette espèce nécessite une série d'étapes complexes : contrôle qualité des lectures, alignement sur le génome de référence, appel et annotation de variants. Pour automatiser et standardiser ces analyses, des pipelines tels que nf-core/rnavar, basés sur Nextflow, offrent un cadre reproductible et efficace pour le traitement de grands volumes de données. L'objectif de ce mini-projet est donc de se familiariser avec cet outil, d'exécuter l'analyse sur deux échantillons RNA-seq de Gadus morhua et d'interpréter les résultats obtenus en termes de qualité des reads, d'efficacité de l'alignement et de détection de variants.

2 Matériel et Méthodes

2.1 Récupération des données

Deux échantillons de Gadus morhua ont été choisis sur le NCBI : SRR2045415 et SRR2045416. Les fichiers SRA ont été téléchargés avec SRA Toolkit, qui permet d'accéder aux séquences brutes depuis la base NCBI. Pour chaque échantillon, la commande suivante a été utilisée :

Code Bash

```
prefetch SRR2045415
prefetch SRR2045416
```

Pour obtenir des fichiers FASTQ paired-end compressés, prêts à être alignés sur le génome de référence, les fichiers SRA des échantillons SRR2045415 et SRR2045416 ont été convertis et compressés à l'aide de la commande :

Code Bash

```
fastq-dump –split-files –gzip SRR2045415
fastq-dump –split-files –gzip SRR2045416
```

Le génome de référence de Gadus morhua (GCF_902167405.1_gadMor3.0_genomic.fna) et le fichier d'annotation (genomic.gtf) ont été téléchargés depuis NCBI Datasets. Ces fichiers sont nécessaires pour : aligner les reads sur le génome de référence (STAR) et annoter les variants détectés (GTF)

2.2 Préparation du fichier samplesheet (input.csv)

Avant de lancer le pipeline nf-core/rnavar, il est nécessaire de créer un fichier samplesheet décrivant les échantillons à analyser. Ce fichier est un tableau CSV avec au minimum une colonne sample (identifiant de l'échantillon) et une colonne fastq_1 pour le read 1 et fastq_2 pour le read 2 dans le cas de données paired-end.Le pipeline nf-core/rnavar est capable de détecter automatiquement si les échantillons sont single-end ou paired-end grâce à ces informations. Pour indiquer l'orientation des reads, une colonne supplémentaire strandedness a été ajoutée avec la valeur unstranded (non orienté) :

Path1 :/work/user/fmt116/TP_nextflow/projet_nextflow/sra_tools/fastq/SRR2045415_1.fastq.gz
Path2 :/work/user/fmt116/TP_nextflow/projet_nextflow/sra_tools/fastq/SRR2045415_2.fastq.gz
Path3 :/work/user/fmt116/TP_nextflow/projet_nextflow/sra_tools/fastq/SRR2045416_1.fastq.gz
Path4 :/work/user/fmt116/TP_nextflow/projet_nextflow/sra_tools/fastq/SRR2045416_2.fastq.gz

Code Bash

```
sample,fastq\_1,fastq\_2,strandedness

SRR2045415,/Path1,/Path2,unstranded

SRR2045416,/Path3,/Path4,unstranded
```

2.3 Réalisation du script de lancement du pipeline

Après avoir récupéré les données RNA-seq (FASTQ) et le génome de référence avec son annotation, et préparé le fichier samplesheet (input.csv), l'étape suivante consiste à préparer et exécuter le script Bash pipeline_rnavar.sh pour lancer le pipeline nf-core/rnavar. Cette étape consiste à créer un script Bash compatible avec Slurm afin de soumettre le pipeline nf-core/rnavar sur le cluster. L'aligner choisi, STAR, est très rapide mais gourmand en mémoire, notamment lors du tri des reads alignés, et une mémoire insuffisante peut provoquer des erreurs ; pour cette raison, 100 Go ont été alloués afin de garantir la stabilité de l'exécution. Il est toutefois possible d'optimiser la consommation mémoire sans allouer autant, en ajustant des paramètres STAR tels que –star_max_memory_bamsort, –star_bins_bamsort ou –star_max_collapsed_junc [3]. Pour accélérer l'alignement, STAR peut paralléliser ses calculs sur plusieurs threads ; dans ce pipeline, 12 cœurs ont été attribués via –cpus-per-task=12, ce qui permet de réduire le temps d'exécution des étapes parallélisables. Le job est identifié sur le cluster par -J nfcorernavar pour faciliter son suivi et est soumis sur la partition workq du cluster.

Le script charge également les modules nécessaires, notamment Nextflow, qui permet d'exécuter les pipelines nf-core dans un environnement contrôlé, en gérant la parallélisation, la reproductibilité et la gestion des dépendances. STAR a été installé localement afin de générer les fichiers d'index du génome à partir du fichier .fna, lesquels sont ensuite référencés via l'option –star_index pour permettre un alignement rapide des reads.

L'exécution du pipeline se fait avec le profil Singularity (-profile singularity), garantissant la reproductibilité grâce à l'utilisation de conteneurs, tout en configurant certains paramètres par défaut adaptés à l'environnement du cluster. L'option –aligner star spécifie STAR comme aligner principal . Le nombre de threads dédiés au contrôle qualité avec FastQC est fixé à 4 (–fastqc_threads 4), tandis que STAR utilise 12 threads (–star_threads 12) correspondant aux cœurs SLURM alloués. La longueur des reads (–read_length 150) est indiquée pour que le pipeline effectue correctement certains calculs internes et corrections. L'étape de recalibrage des bases (–skip_baserecalibration) est désactivée, car elle peut être longue et n'est pas toujours nécessaire lorsque la qualité des reads est déjà correcte. Enfin, l'option -resume permet de reprendre l'exécution du pipeline là où elle s'était arrêtée en cas d'interruption, évitant ainsi de relancer toutes les étapes précédemment complétées. pat : /work/user/fmt116/TP_nextflow/projet_nextflow/sra_tools

Code Bash

```
#!/bin/bash
#SBATCH -J nfcorernavar
#SBATCH -p workq
#SBATCH --mem=100G
#SBATCH --cpus-per-task=12


module purge
module load bioinfo/NextflowWorkflows/nfcore-Nextflow-v25.04.0

export PATH=/pat/star_index/STAR-2.7.11b/source:$PATH

nextflow run nf-core/rnavar -r 1.2.1 \
  -profile singularity \
  --input /pat/input.csv \
  --fasta /pat/GCF_902167405.1_gadMor3.0_genomic.fna \
  --gtf /pat/genomic.gtf \
  --aligner star \
  --star_index /pat/star_index/ \
  --fastqc_threads 4 \
  --star_threads 12 \
  --read_length 150 \
  --outdir /pat/results_rnavar \
  --skip_baserecalibration \
  -resume
```

2.4 Soumission du pipeline sur le cluster

Le pipeline ne peut pas être exécuté directement sur le nœud frontal du cluster , car il requiert des ressources importantes (mémoire, processeurs). L'exécution doit donc être faite sur un nœud de calcul .

Pour cela, on utilise la commande :

Code Bash

```
sbatch pipeline_rnavar.sh
```

Cette commande soumet le script pipeline_rnavar.sh à la file d'attente du cluster. Slurm se charge ensuite d'allouer les ressources demandées dans le script (–mem, –cpus-per-task, partition, etc.) et de lancer le pipeline sur un nœud de calcul adapté

3 Résultats

3.1 l'organisation des résultats

À l'issue de l'exécution du pipeline nf-core/rnavar, le dossier de résultats généré est structuré en plusieurs sous-répertoires correspondant aux étapes clés de l'analyse. Le dossier pipeline_info/ contient les journaux d'exécution de Nextflow ainsi que les informations de configuration (profil, paramètres, version du pipeline). Il constitue une trace technique utile pour vérifier le bon déroulement de l'analyse et assurer sa reproductibilité, sans nécessiter d'interprétation biologique approfondie. Le répertoire preprocessing/ regroupe les résultats liés à la préparation des données, notamment les rapports FastQC sur la qualité des lectures brutes et éventuellement des lectures nettoyées. Ces rapports permettent une première évaluation de la qualité des données d'entrée. Le dossier reports/ rassemble les rapports globaux générés par MultiQC, dont le fichier central multiqc_report.html qui offre une vue d'ensemble intégrée de la qualité, de l'alignement, de la couverture et des variants détectés ; il constitue le point d'entrée idéal pour explorer rapidement les résultats. Le répertoire samtools/ fournit des statistiques d'alignement calculées avec Samtools, comme le taux de mapping des reads et la profondeur de séquençage, ce qui permet de vérifier la qualité et l'efficacité de l'alignement réalisé avec STAR. Enfin, le dossier variant_calling/ contient les fichiers VCF produits par GATK HaplotypeCaller, correspondant aux variants détectés ainsi que leurs index. C'est la partie centrale de l'analyse, car elle permet d'examiner le nombre et la qualité des variants, d'exploiter les scores de qualité (QUAL) et de profondeur (DP), et de relier ces variants aux gènes concernés grâce à l'annotation fournie.

3.2 Qualité et alignement des reads

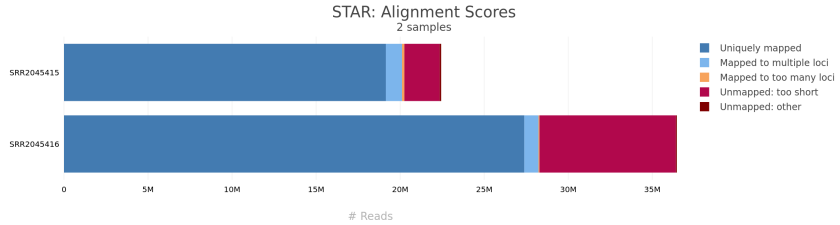

FIGURE 1 – Schéma du Score d'alignement STAR

Sample Name	Total reads	Aligned	Uniq aligned	Avg. mapped len	Annotated splices	Mismatch rate	Del rate	Del len	Ins rate	Ins len
SRR2045415	22.4 M	89.6%	85.3%	197.3 bp	16.5 M	0.9%	0.0%	2.3 bp	0.0%	1.9 bp
SRR2045416	36.5 M	77.4%	75.1%	196.9 bp	15.4 M	1.0%	0.0%	2.4 bp	0.0%	1.9 bp

FIGURE 2 – Schéma des données statistiques de l'alignement STAR

3.3 Résultats du variant calling

La capture ci-dessous illustre le contenu du fichier VCF généré par GATK HaplotypeCaller pour l'échantillon SRR2045415. On y retrouve les informations essentielles pour chaque variant détecté : la position chromosomique, l'allèle de référence, l'allèle alternatif,

la qualité du variant et les métriques de filtrage. Cet extrait permet de visualiser directement le format des données et la nature des variants identifiés (SNPs et INDELs).

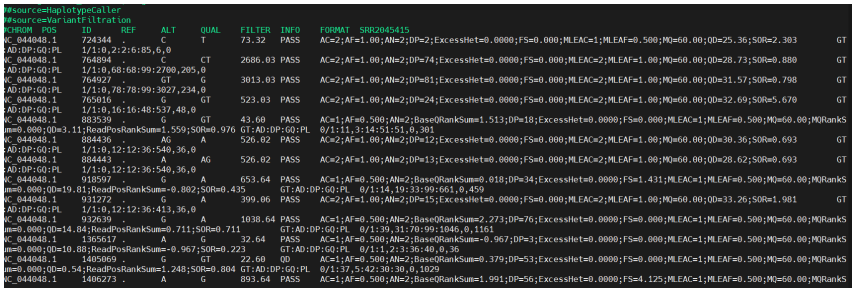

FIGURE 3 – Extrait du fichier VCF pour SRR2045415

4 Discussion

L'évaluation des statistiques d'alignement obtenues avec Samtools montre un alignement de très haute qualité. Sur l'échantillon SRR2045415, les 40 millions de reads générés se sont tous alignés sur le génome de référence, avec un appariement correct des lectures en mode paired-end. Le taux de duplication est d'environ 22 %, ce qui reste acceptable dans le cadre d'une analyse RNA-seq, où certaines régions sont naturellement très couvertes en raison de niveaux d'expression élevés. La longueur moyenne des lectures est de 100 pb, ce qui confirme les caractéristiques techniques des données de séquençage. Globalement, la qualité de l'alignement est jugée excellente et permet de poursuivre l'analyse des variants avec confiance. L'échantillon SRR2045416 contient environ 56 millions de lectures, dont 100 % se sont alignées correctement sur le génome de référence. La proportion de duplications est plus faible (11 %) que dans l'échantillon SRR2045415, indiquant une meilleure complexité de la librairie et réduisant les risques de biais liés à la surexpression de certaines régions. Le taux de reads avec une qualité de mapping nulle est inférieur à 0,5 %, confirmant la fiabilité de l'alignement. La longueur moyenne des lectures est de 100 pb, avec un insert size moyen d'environ 695 pb. La qualité moyenne des bases est élevée (Q32), traduisant une excellente précision de séquençage. Ces résultats confirment que les données de SRR2045416 sont de très bonne qualité et adaptées à la détection fiable de variants. L'étape de variant calling a permis d'identifier un nombre significatif de SNPs et d'INDELs dans les deux échantillons analysés (SRR2045415 et SRR2045416). Les statistiques extraites des fichiers .stats et des VCF indiquent la présence de variants à la fois hétérozygotes et homozygotes, reflétant la diversité génétique au sein des échantillons. Ces variants sont répartis sur plusieurs chromosomes et présentent des profondeurs de lecture suffisantes pour garantir leur fiabilité. Les fichiers .vcf.gz et leurs index .tbi contiennent l'ensemble des informations nécessaires pour une annotation fonctionnelle, incluant la position chromosomique, l'allèle de référence et l'allèle alternatif, ainsi que les scores de qualité et les métriques de filtrage.

L'analyse des résultats RNA-seq indique que les données brutes étaient de bonne qualité, avec un taux d'alignement élevé et une couverture satisfaisante sur le génome de référence de Gadus morhua. Le pipeline nf-core/rnavar a permis de détecter un nombre conséquent de variants, incluant à la fois des SNPs et des INDELs. Parmi ces variants,

certains sont présents à l'état homozygote tandis que d'autres sont hétérozygotes, reflétant une diversité génétique au sein des échantillons étudiés.

5 Références

[1] Derrien T., Johnson R., Bussotti G., et al. The GENCODE v7 catalog of human long noncoding RNAs : Analysis of their gene structure, evolution, and expression. *Nature*, 2011. Lien vers l'article

[2] Ma Y., Lou F., Yin X., Cong B., Liu S., Zhao L., Zheng L. Whole-genome survey and phylogenetic analysis of *Gadus macrocephalus*. *Bioscience Reports*, 2022. Lien vers l'article (PMC9289796)

[3] Pipeline d'appel de variants Lien vers la documentation)