

Nextflow Project

Youcef Ben Mohammed

October 10, 2023

Contents

1	Introduction	3
2	Mise en place d'un pipeline RNA-seq avec Nextflow	3
2.1	Préparation des données	3
2.2	Préparation des fichiers d'exécution	4
3	Analyse des résultats	6
3.1	Interprétation des principaux résultats	6
3.2	Interprétation du rapport MultiQC	8
3.2.1	General Statistics	8
3.2.2	Biotype Counts	8
3.2.3	Picard	9
3.2.4	QualiMap	9
3.2.5	RSeQC	10
3.2.6	FastQC	10
4	Lancement du pipeline sur des données NCBI	13
4.1	Récupération des données	13
4.1.1	Récupération des numéros d'accèsion des fastq	13
4.1.2	Téléchargement des fastq avec sratoolkit	14
4.1.3	Récupération du génome de référence et du fichier d'annotation	14
4.1.4	Lancement du pipeline	14
4.2	Analyse des résultats MultiQC	15
4.2.1	General Statistics	15
4.2.2	Biotype Counts	16
4.2.3	QualiMap	16
4.2.4	FastQC	17
5	Conclusion et Perspectives	18
6	Références	18

1 Introduction

L'un des défis majeurs auxquels un bioinformaticien peut être confronté aujourd'hui c'est d'être capable de reproduire les résultats exactes d'un workflow donné sans problèmes de dépendance et avec moins de debugage. En effet, reconstruire *de novo* ce dernier demande souvent un temps considérable. Un autre défi important est de garantir que ce workflow soit portable, c'est-à-dire qu'il puisse s'exécuter et produire les mêmes résultats dans différents environnements de travail. Dans ce cadre vient Nextflow, qui est un système de gestion de workflows qui assure la reproductibilité des workflows, permettant également le lancement simultané de plusieurs pipelines (principe de parallélisation), et garantit la portabilité de ce workflow d'un environnement à un autre sans dépendances système.

Ce projet a pour but de se familiariser avec l'utilisation de Nextflow, de ce fait nous allons construire un pipeline RNA-seq avec Nextflow ensuite nous allons le tester sur un jeu de données, puis nous analyserons les résultats de sortie, principalement le rapport MultiQC, enfin nous allons essayer de lancer le pipeline sur des données de *Gadus morhua* collectées à partir de site NCBI.

2 Mise en place d'un pipeline RNA-seq avec Nextflow

Après avoir connecter au serveur genologin, cette étape consiste à télécharger les données disponibles sur http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES, puis à préparer un fichier bash Nextflow qui sera lancé sur le cluster de calcul de genologin.

2.1 Préparation des données

Les données sont téléchargé avec les commandes suivantes;

- Genome de référence :

```
1 wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.fasta
```

- Fichier d'annotaion :

```
1 wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.gtf
```

- Fichiers fastq :

```
1 wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep1_1_Ch6.fastq.gz
```

```
2
```

```
3 wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep1_2_Ch6.fastq.gz
```

```
4
```

```
5 wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep2_1_Ch6.fastq.gz
```

```
6
```

```
7 wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep2_2_Ch6.fastq.gz
```

2.2 Préparation des fichiers d'exécution

Avant de lancer le pipeline RNA-Seq, il faudra au préalable préparer quatre fichiers.

- Un fichier CSV contenant les informations et les chemins vers les fichiers FASTQ, nommé "inputs.csv".

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ more inputs.csv
group,replicate,fastq_1,fastq_2,strandedness
mutant,1,/work/cyclamen/nextflow_project/NEXTFLOW/FASTQ/MT_rep1_1_Ch6.fastq.gz,/work/cyclamen/nextflow_project/NEXTFLOW/FASTQ/MT_rep1_2_Ch6.fastq.gz,unstranded
wild,1,/work/cyclamen/nextflow_project/NEXTFLOW/FASTQ/WT_rep1_1_Ch6.fastq.gz,/work/cyclamen/nextflow_project/NEXTFLOW/FASTQ/WT_rep1_2_Ch6.fastq.gz,unstranded
```

- Un fichier trace qui permet de récupérer les lignes de commande complètes lancées à chaque étape du pipeline, nommé "pipeline_trace.txt".

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ more pipeline_trace.txt
task_id name status exit realtime %cpu rss script
2 RNASEQ:PREPARE_GENOME:GTF2BED (ITAG2.3_genomic_Ch6.gtf) COMPLETED 0 0ms 84.2% 9.9 MB
gtf2bed ITAG2.3_genomic_Ch6.gtf > ITAG2.3_genomic_Ch6.bed
5 RNASEQ:PREPARE_GENOME:GET_CHROM_SIZES (ITAG2.3_genomic_Ch6.fasta) COMPLETED 0 1s 75.3% 2.1 MB
samtools faidx ITAG2.3_genomic_Ch6.fasta
cut -f 1,2 ITAG2.3_genomic_Ch6.fasta.fai > ITAG2.3_genomic_Ch6.fasta.sizes
echo $(samtools --version 2>&1) | sed 's/^.*samtools //; s/Using.*$//' > samtools.version.txt
```

- Un fichier de configuration nommé "sm_config.cfg".

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

- Un fichier de lancement nommé "run_pipeline.sh".

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ more run_pipeline.sh
#!/bin/bash
#SBATCH -J YoucefBENMOHAMMED
#SBATCH -p workq
#SBATCH --time=1-00:00:00
#SBATCH --mem=6G

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

input=/work/cyclamen/nextflow_project/NEXTFLOW/inputs.csv
gtf=/work/cyclamen/nextflow_project/NEXTFLOW/annotation/ITAG2.3_genomic_Ch6.gtf
fasta=/work/cyclamen/nextflow_project/NEXTFLOW/genome/ITAG2.3_genomic_Ch6.fasta
config=/work/cyclamen/nextflow_project/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
```

- Pour les options SBATCH -J indique le nom du job.
- p la chaîne de traitement à utilisée
- time la durée maximale du job
- mem la mémoire à utiliser

Pour l'option --resume dans le pipeline RNAseq indique à Nextflow, une fois le pipeline est relancé, de se redémarrer à partir d'un fichier caché.

La commande seff permet de suivre l'état du job.

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ seff 50706900
Job ID: 50706900
Cluster: genobull
User/Group: cyclamen/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:03:08
CPU Efficiency: 31.13% of 00:10:04 core-walltime
Job Wall-clock time: 00:10:04
Memory Utilized: 1.85 GB
Memory Efficiency: 30.78% of 6.00 GB
```

Job ID: indique l'identifiant du job. Cluster: le nom du cluster. User/Group: le nom de l'utilisateur qui a lancé le job. State: état actuelle du job, COMPLETED; la tâche est terminée. Cores: le nombre de cœurs. CPU Utilized: le temps pendant lequel le processeur est actif. CPU Efficiency: le pourcentage de temps pendant lequel le processeur est actif par rapport à au temps total. Job Wall-clock time: mesure la durée totale de l'exécution du job. Memory Utilized: la quantité de mémoire (RAM) utilisé par le job. Mémoire Efficiency: le pourcentage de mémoire utilisé par rapport à la mémoire disponible.

3 Analyse des résultats

Cette étape consiste à interpréter les principaux résultats ainsi que le rapport MultiQC généré par le pipeline Nextflow.

3.1 Interprétation des principaux résultats

Le pipeline renvoie un dossier results qui contient 5 dossier, Un dossier fastqc contenant les résultats du controle quality des fastq.

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/  
total 3  
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 00:57 fastqc  
drwxr-xr-x 4 cyclamen formation 4096 5 oct. 00:57 genome  
drwxr-xr-x 3 cyclamen formation 4096 5 oct. 01:04 multiqc  
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:04 pipeline_info  
drwxr-xr-x 14 cyclamen formation 4096 5 oct. 01:03 star_rsem  
drwxr-xr-x 3 cyclamen formation 4096 5 oct. 00:58 trimgalore  
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/fastqc/  
total 4208  
-rw-r--r-- 1 cyclamen formation 658481 5 oct. 00:57 mutant_R1_1_fastqc.html  
-rw-r--r-- 1 cyclamen formation 416877 5 oct. 00:57 mutant_R1_1_fastqc.zip  
-rw-r--r-- 1 cyclamen formation 654797 5 oct. 00:57 mutant_R1_2_fastqc.html  
-rw-r--r-- 1 cyclamen formation 412887 5 oct. 00:57 mutant_R1_2_fastqc.zip  
-rw-r--r-- 1 cyclamen formation 658647 5 oct. 00:57 wild_R1_1_fastqc.html  
-rw-r--r-- 1 cyclamen formation 415941 5 oct. 00:57 wild_R1_1_fastqc.zip  
-rw-r--r-- 1 cyclamen formation 654999 5 oct. 00:57 wild_R1_2_fastqc.html  
-rw-r--r-- 1 cyclamen formation 413133 5 oct. 00:57 wild_R1_2_fastqc.zip
```

Un dossier genome qui contient les informations relatives à l'indexation du génome.

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/genome/  
total 2314  
drwxr-xr-x 3 cyclamen formation 4096 5 oct. 00:57 index  
-rw-r--r-- 1 cyclamen formation 320910 5 oct. 00:56 ITAG2.3_genomic_Ch6.bed  
-rw-r--r-- 1 cyclamen formation 29 5 oct. 00:57 ITAG2.3_genomic_Ch6.fasta.fai  
-rw-r--r-- 1 cyclamen formation 20 5 oct. 00:57 ITAG2.3_genomic_Ch6.fasta.sizes  
-rw-r--r-- 1 cyclamen formation 2034585 5 oct. 00:57 ITAG2.3_genomic_Ch6_genes.gtf  
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 00:57 rsem  
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/genome/rsem/  
total 14905  
-rw-r--r-- 1 cyclamen formation 20 5 oct. 00:57 genome.chrlist  
-rw-r--r-- 1 cyclamen formation 12963 5 oct. 00:57 genome.grp  
-rw-r--r-- 1 cyclamen formation 3573367 5 oct. 00:57 genome.idx.fa  
-rw-r--r-- 1 cyclamen formation 3573367 5 oct. 00:57 genome.n2g.idx.fa  
-rw-r--r-- 1 cyclamen formation 3818268 5 oct. 00:57 genome.seq  
-rw-r--r-- 1 cyclamen formation 676102 5 oct. 00:57 genome.ti  
-rw-r--r-- 1 cyclamen formation 3573367 5 oct. 00:57 genome.transcripts.fa  
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $
```

Un dossier multiqc, qui rassemble l'ensemble des resultats du pipeline RNA-seq.

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/multiqc/star_rsem/  
total 1353  
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:04 multiqc_data  
-rw-r--r-- 1 cyclamen formation 1383618 5 oct. 01:04 multiqc_report.html  
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/multiqc/star_rsem/multiqc_data/  
total 991  
-rw-r--r-- 1 cyclamen formation 422 5 oct. 01:04 multiqc_cutadapt.txt  
-rw-r--r-- 1 cyclamen formation 972778 5 oct. 01:04 multiqc_data.json  
-rw-r--r-- 1 cyclamen formation 1204 5 oct. 01:04 multiqc_fastqc_1.txt  
-rw-r--r-- 1 cyclamen formation 1145 5 oct. 01:04 multiqc_fastqc.txt  
-rw-r--r-- 1 cyclamen formation 2622 5 oct. 01:04 multiqc_general_stats.txt  
-rw-r--r-- 1 cyclamen formation 21093 5 oct. 01:04 multiqc.log  
-rw-r--r-- 1 cyclamen formation 417 5 oct. 01:04 multiqc_picard_dups.txt  
-rw-r--r-- 1 cyclamen formation 222 5 oct. 01:04 multiqc_rsem.txt  
-rw-r--r-- 1 cyclamen formation 621 5 oct. 01:04 multiqc_rseqc_bam_stat.txt  
-rw-r--r-- 1 cyclamen formation 95 5 oct. 01:04 multiqc_rseqc_infer_experiment.txt  
-rw-r--r-- 1 cyclamen formation 726 5 oct. 01:04 multiqc_rseqc_junction_annotation.txt  
-rw-r--r-- 1 cyclamen formation 1727 5 oct. 01:04 multiqc_rseqc_read_distribution.txt
```

Un dossier pipeline_info qui contient les informations détaillées du pipeline, dans lequel nous trouvons également les versions des logiciels utilisés.

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/pipeline_info/
total 3602
-rw-r--r-- 1 cyclamen formation 3144070 5 oct. 01:04 execution_report.html
-rw-r--r-- 1 cyclamen formation 271620 5 oct. 01:04 execution_timeline.html
-rw----- 1 cyclamen formation 242205 5 oct. 01:04 pipeline_dag.svg
-rw-r--r-- 1 cyclamen formation 12994 5 oct. 01:04 pipeline_report.html
-rw-r--r-- 1 cyclamen formation 2543 5 oct. 01:04 pipeline_report.txt
-rw-r--r-- 1 cyclamen formation 377 5 oct. 00:57 samplesheet_valid.csv
-rw-r--r-- 1 cyclamen formation 235 5 oct. 01:03 software_versions.csv
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $
```

Un dossier star_rsem qui contient les résultats statistiques et des estimations quantitatives sur l'expression génique générés par le pipeline.

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW $ ll results/star_rsem/
total 401942
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:03 bigwig
drwxr-xr-x 7 cyclamen formation 4096 5 oct. 01:02 dupradar
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:02 featurecounts
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:01 log
-rw-r--r-- 1 cyclamen formation 201057 5 oct. 01:01 mutant_R1.genes.results
-rw-r--r-- 1 cyclamen formation 211121 5 oct. 01:01 mutant_R1.isoforms.results
-rw-r--r-- 1 cyclamen formation 224454717 5 oct. 01:02 mutant_R1.markdup.sorted.bam
-rw-r--r-- 1 cyclamen formation 65224 5 oct. 01:02 mutant_R1.markdup.sorted.bam.bai
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:01 mutant_R1.stat
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:02 picard_metrics
drwxr-xr-x 3 cyclamen formation 4096 5 oct. 01:02 preseq
drwxr-xr-x 4 cyclamen formation 4096 5 oct. 01:04 qualimap
-rw-r--r-- 1 cyclamen formation 141712 5 oct. 01:01 rsem.merged_gene_counts.tsv
-rw-r--r-- 1 cyclamen formation 141550 5 oct. 01:01 rsem.merged_gene_tpm.tsv
-rw-r--r-- 1 cyclamen formation 141709 5 oct. 01:01 rsem.merged_transcript_counts.tsv
-rw-r--r-- 1 cyclamen formation 141547 5 oct. 01:01 rsem.merged_transcript_tpm.tsv
drwxr-xr-x 9 cyclamen formation 4096 5 oct. 01:03 rseqc
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:03 samtools_stats
drwxr-xr-x 4 cyclamen formation 4096 5 oct. 01:03 stringtie
-rw-r--r-- 1 cyclamen formation 200315 5 oct. 01:00 wild_R1.genes.results
-rw-r--r-- 1 cyclamen formation 210125 5 oct. 01:00 wild_R1.isoforms.results
-rw-r--r-- 1 cyclamen formation 185538085 5 oct. 01:02 wild_R1.markdup.sorted.bam
-rw-r--r-- 1 cyclamen formation 59920 5 oct. 01:02 wild_R1.markdup.sorted.bam.bai
drwxr-xr-x 2 cyclamen formation 4096 5 oct. 01:00 wild_R1.stat
```

3.2 Interprétation du rapport MultiQC

3.2.1 General Statistics

General Statistics

Copy table | Configure Columns | Plot | Showing 10 rows and 19 columns.

Sample Name	M Reads Mapped	% Dups	5'-3' bias	M Aligned	% Alignable	% Proper Pairs	Error rate	M Non-Primary	M Reads Mapped	% Mapped	% Proper Pairs	M Total seqs	% Dups	% GC	M Seqs	% BP Tri
mutant_R1	3.3	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%	42%	1.6	3.5%
mutant_R1_2													49.2%	41%	1.6	3.7%
wild_R1	2.7	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%	42%	1.3	3.4%
wild_R1_2													48.2%	42%	1.3	3.7%

Figure 1: General Statistics

Nous avons obtenu un taux d'alignement autour de 99 % pour les deux conditions, ce qui est très bon.

Un pourcentage des duplicats 17,3% pour mutant et 18,8% pour Wild-type, les duplicats sont intéressant dans le cadre de RNA-seq, puisque on ignore leurs origine (PCR) ou taux d'expression génique.

Un pourcentage des reads mapped autour des 99% ce qui est rassurant.

Un pourcentage de GC 42% qui est un peu élevé par rapport au taux de GC chez la tomate (38%), cela est dû probablement au taux de duplicats élevé.

3.2.2 Biotype Counts

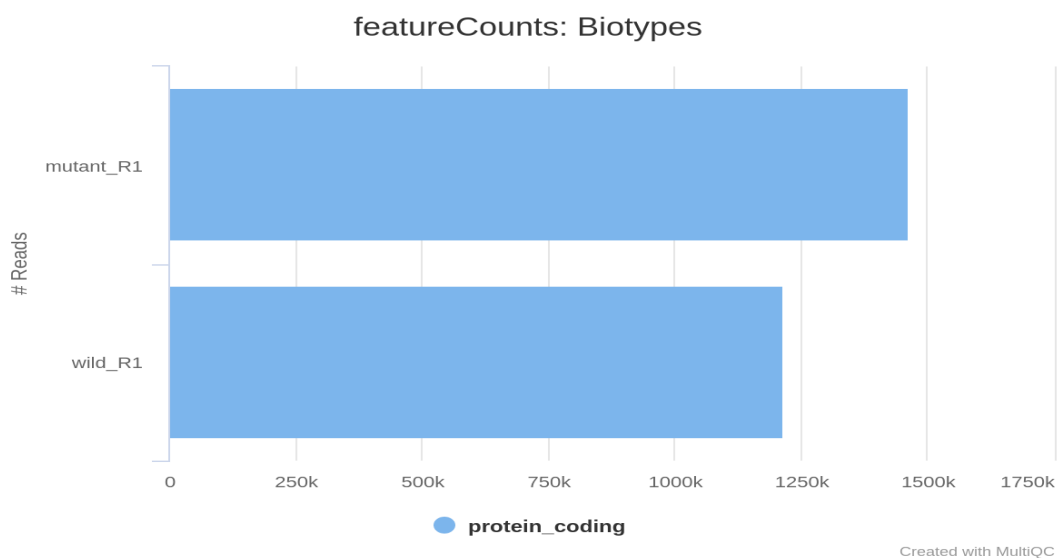


Figure 2: Biotype Counts

Nous pouvons constaté que pour les deux conditions mutantes et sauvages, 100% des reads sont associés à des protéines codantes.

3.2.3 Picard

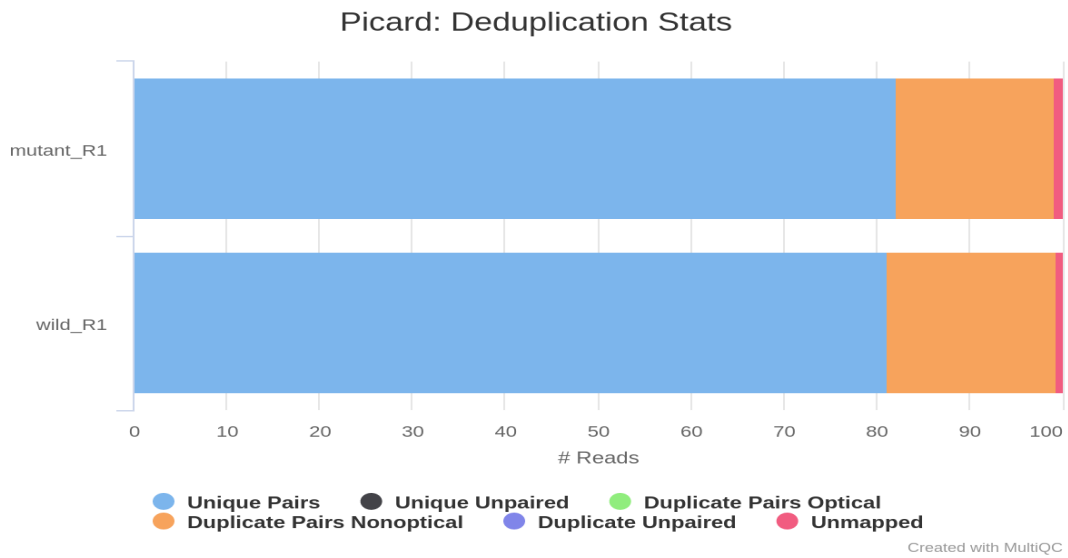


Figure 3: Mark Duplicates

Cela renvoie le nombre de des reads classé par l'état de duplication,

3.2.4 QualiMap

Nous pouvons également identifier les problèmes liés à notre library ou à la contamination de nos samples en examinant le pourcentage de lectures qui sont exoniques, introniques ou intergénomiques.

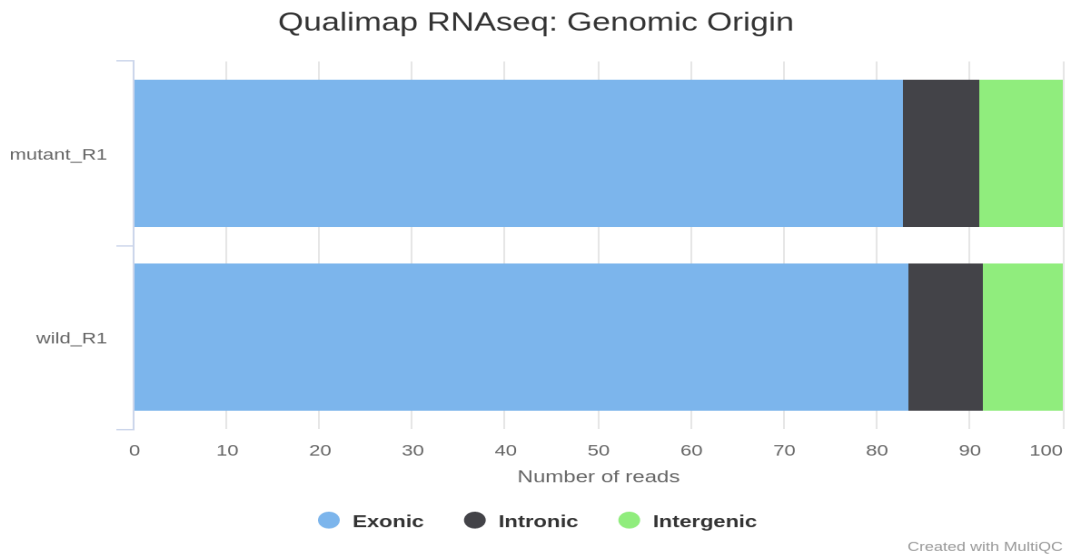


Figure 4: Genomic origin of reads

Un taux élevé de reads intergénomiques, indique la présence de contamination, ici avons un pourcentage de 8% pour les deux samples, donc les contaminations sont faibles.

3.2.5 RSeQC

Inner Distance calcule la distance interne entre deux reads pair-end.

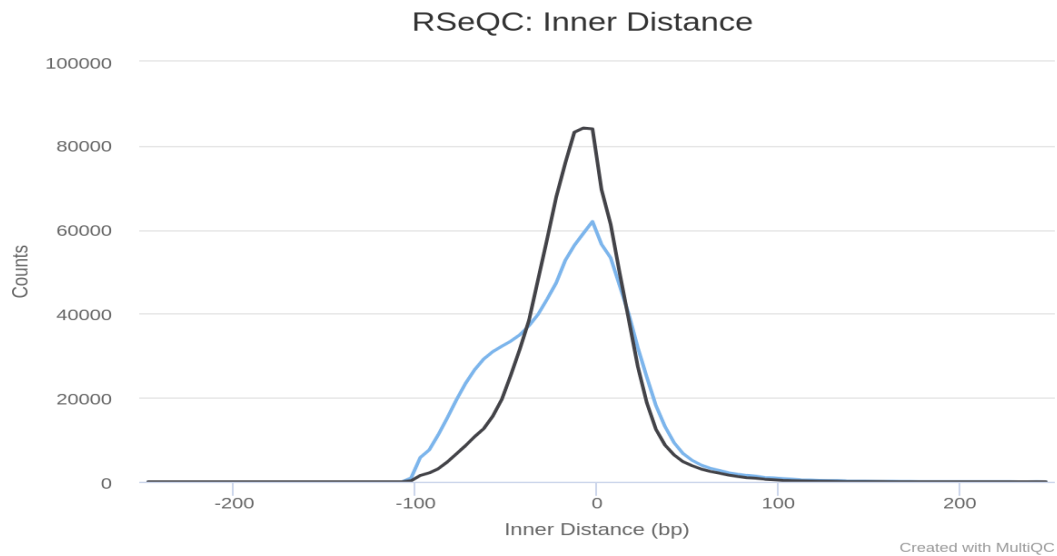


Figure 5: RSeQC: Inner Distance

La distribution des distances pour WT suit une distribution normale, les mutants tend vers une distribution plus ou moins normale, ce qui est un bon indicateur.

3.2.6 FastQC

- Sequence Quality Histograms



Figure 6: Sequence Quality Histograms

Nous pouvons constater que la distribution des Phread-scores pour les deux échantillons tout le long des positions des reads est presque supérieur à 30, ce qui rassurant d'une très bonne qualité des reads.

- Per Sequence Quality Scores

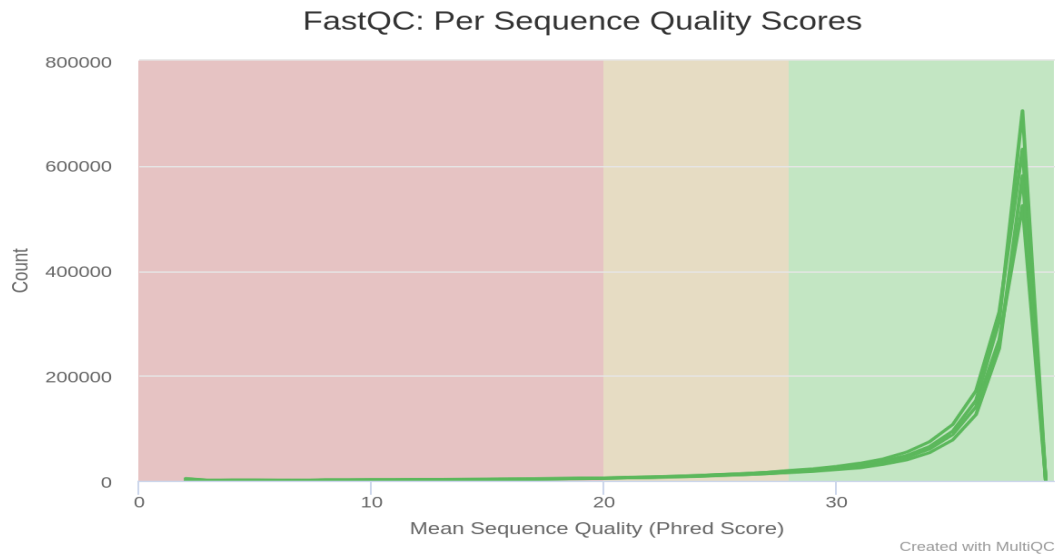


Figure 7: Per Sequence Quality Scores

Nous pouvons constater que les Phread-Scores sont au-dessus de 28 et qu'ils semblent être majoritairement unimodaux.

- Per Sequence GC Content

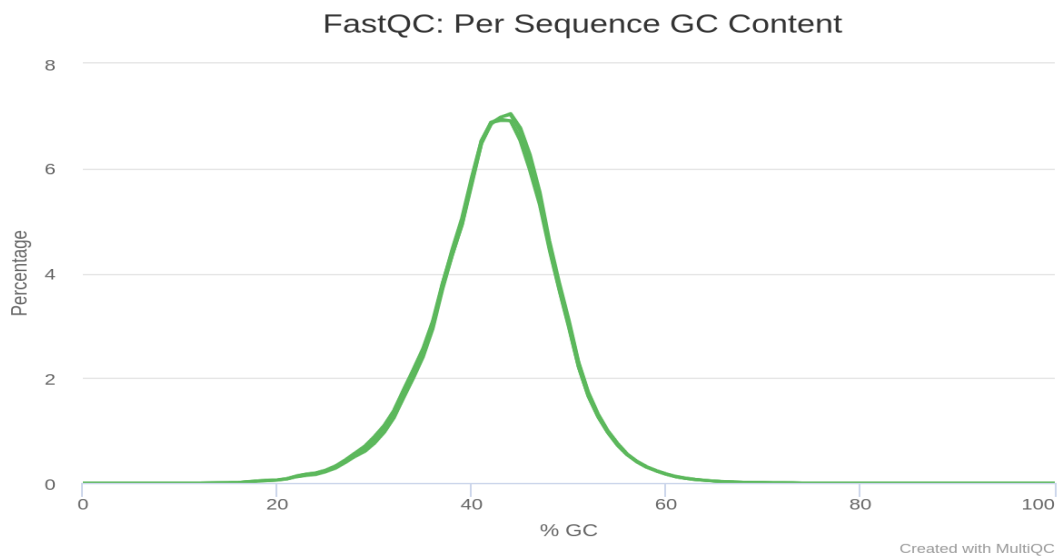


Figure 8: GC Content Distribution

Nous pouvons voir que la distribution des pourcentages de GC sont unimodale et tend vers une loi normale, ce qui est un bon indicateur.

- Sequence Duplication Levels

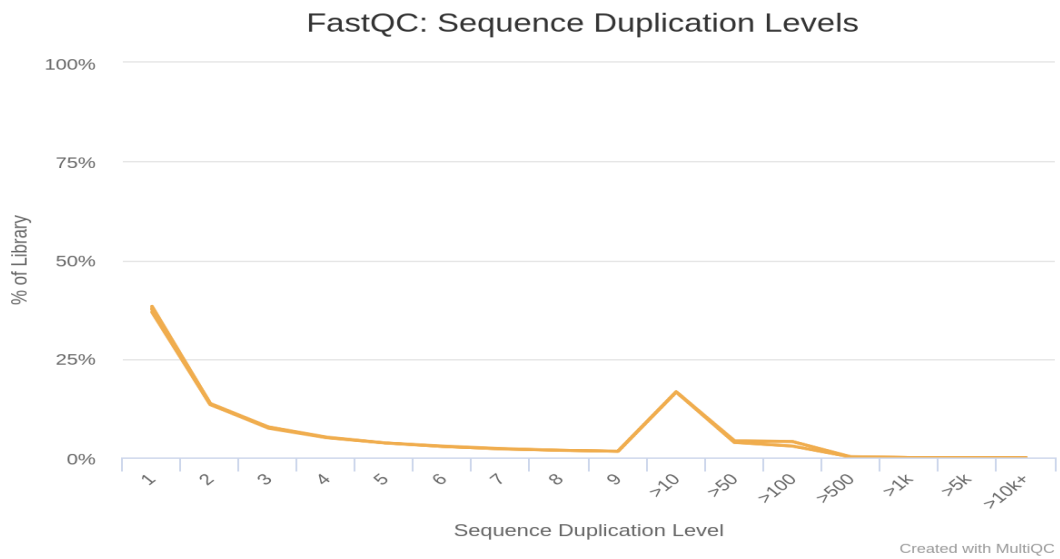


Figure 9: Sequence Duplication Levels

Nous pouvons voir la taux de sequences dupliques dans les reads, la figure montre un pic dans l'ensemble pas de présence un taux élevé de duplication, à l'exception pour 10 sequences un taux de duplication d'environ 16% pour les deux conditions.

4 Lancement du pipeline sur des données NCBI

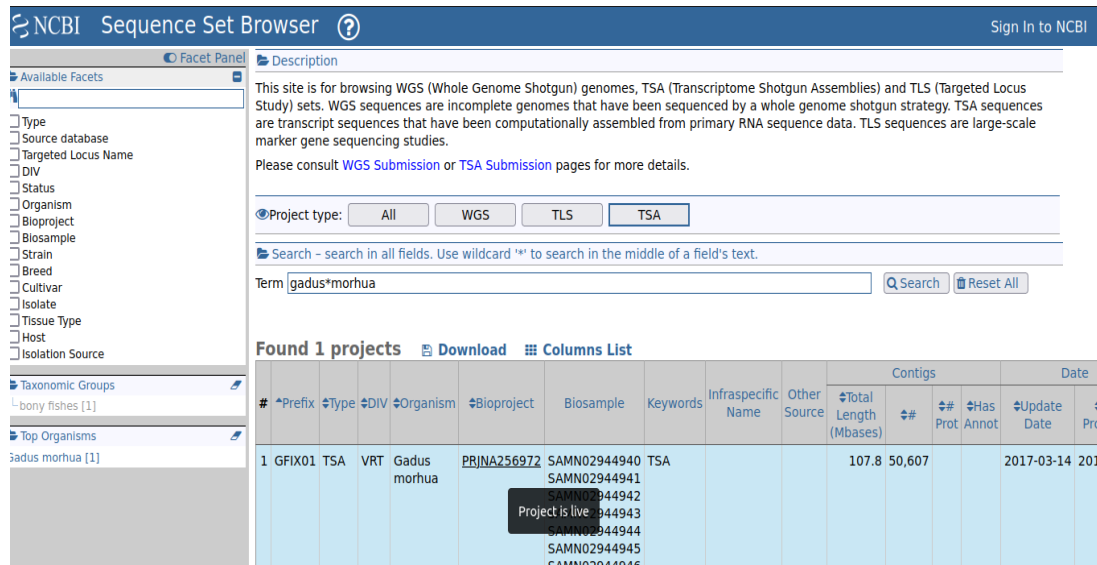
Cette étape consiste à lancer le pipeline sur deux échantillons sur le site NCBI.

Commande pour compresser

4.1 Récupération des données

4.1.1 Récupération des numéros d'accèsion des fastq

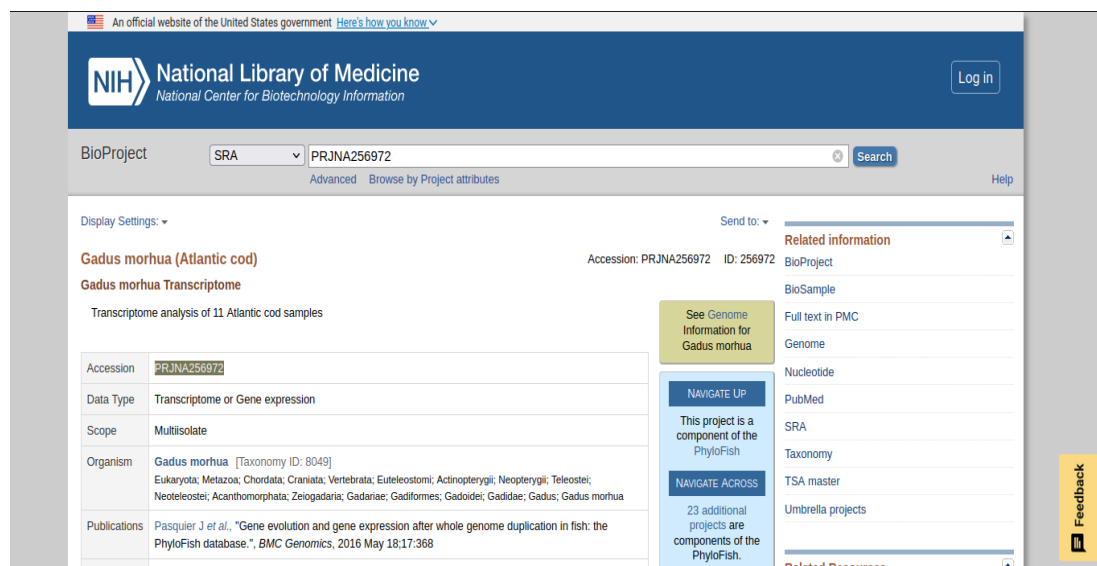
La première étape consiste à chercher sur le site de NCBI le nom de l'espèce, dans notre cas il s'agit de *Gadus morhua*.



The screenshot shows the NCBI Sequence Set Browser interface. The search term "gadus*morhua" is entered in the search bar. The results table shows one project with 107.8 million contigs and 50,607 contigs. The project ID is PRJNA256972.

#	*Prefix	Type	DIV	Organism	BioProject	Biosample	Keywords	Intraspecific Name	Other Source	Contigs	Date
										Total Length (Mbases)	Update Date
1	GFIX01	TSA	VRT	Gadus morhua	PRJNA256972	SAMN02944940 SAMN02944941 SAMN02944942 SAMN02944943 SAMN02944944 SAMN02944945 SAMN02944946	TSA			107.8	2017-03-14

En choisissant un project avec minimum de 30000 contigs, nous allons cliquer sur le numéro de bioProject, qui va nous rediriger vers le site de NCBI, ensuite dans la barre de recherche nous tapons le numéro d'accèsion.



The screenshot shows the National Library of Medicine BioProject page for PRJNA256972. The page displays project details for *Gadus morhua* (Atlantic cod), including the accession number PRJNA256972 and the title "Gadus morhua Transcriptome". The page also shows related information and navigation options.

Accession: PRJNA256972

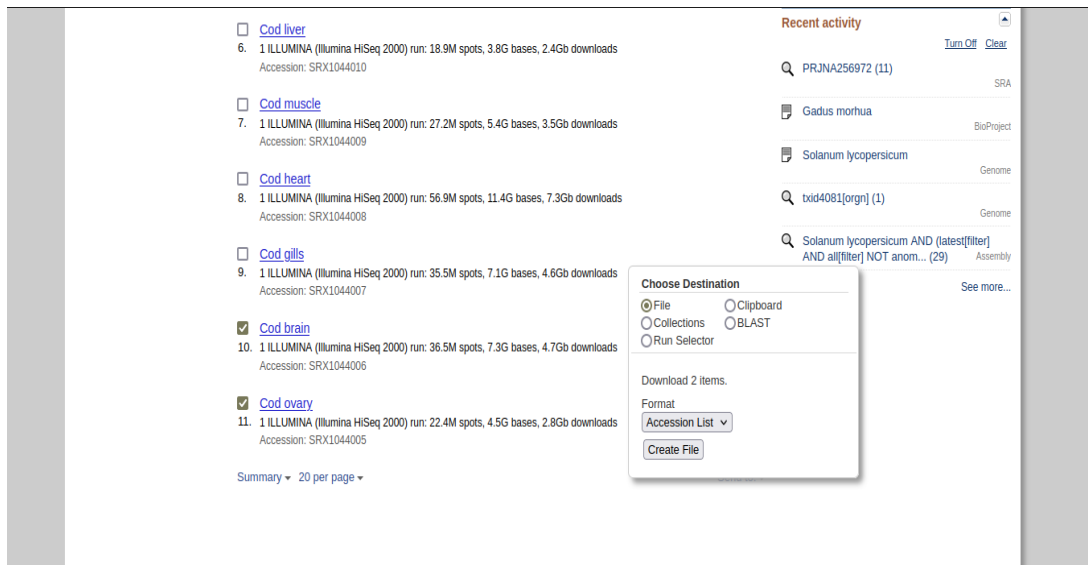
Data Type: Transcriptome or Gene expression

Scope: Multisolate

Organism: *Gadus morhua* [Taxonomy ID: 8049]
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Neoteleostei; Acanthomorphata; Zeiogadaria; Gadariae; Gadiformes; Gadoidei; Gadidae; Gadus; *Gadus morhua*

Publications: Pasquier J et al. "Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database.", *BMC Genomics*, 2016 May 18;17:368

Enfin nous allons choisir deux fastq, puis nous allons télécharger les numéros d'accèsion correspondant.



4.1.2 Téléchargement des fastq avec sratoolkit

Dans le répertoire de travail du cluster, nous allons préparer créer le fichier suivant:

```
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua $ more sratoolkit.sh
#!/bin/bash
#SBATCH -J sra_toolkit
#SBATCH -p unlimitq #queue de traitement
#SBATCH --mem=6G #memoire

module purge
module load bioinfo/sratoolkit.3.0.0

fastq-dump --split-files SRR2045415;
fastq-dump --split-files SRR2045416;
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua $
```

Après avoir récupéré les fastq, nous allons les compresser puisque le pipeline nf-rnaseq prend en entrée des fichiers en format fastq.gz.

```
1 for i in *.fastq; do gzip i;
2
```

4.1.3 Récupération du génome de référence et du fichier d'annotation

```
1 wget https://ftp.ensembl.org/pub/release-110/fasta/gadus_morhua/dna_index/Gadus_morhua.gadMor3.0.dna.toplevel.f
2
3 wget https://ftp.ensembl.org/pub/release-110/gtf/gadus_morhua/ Gadus_morhua.gadMor3.0.110.chr.gtf.gz
4
```

4.1.4 Lancement du pipeline

Après avoir tout les fichiers d'entrée, nous allons modifier les chemins d'accès sur les fichiers crée dans le pipeline précédent, ensuite nous allons lancer le script sur le cluster avec la commande sbatch, et suivre l'état du job avec la commande seff.

```

cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/pipeline $ more run_pipeline.sh
#!/bin/bash
#SBATCH -J YoucefBENMOHAMMED
#SBATCH -p workq
#SBATCH --time=1-00:00:00
#SBATCH --mem=6G

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

input=/work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/pipeline/inputs.csv
gtf=/work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/annotation/Gadus_morhua.gadMor3.0.110.chr.gtf.gz
fasta=/work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/genome/Gadus_morhua.gadMor3.0.dna.toplevel.fa.gz
config=/work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/pipeline/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
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/pipeline $ seff 50755746
Job ID: 50755746
Cluster: genobull
User/Group: cyclamen/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:04:58
CPU Efficiency: 1.66% of 04:59:43 core-walltime
Job Wall-clock time: 04:59:43
Memory Utilized: 1.86 GB
Memory Efficiency: 30.95% of 6.00 GB
cyclamen@genologin2 /work/cyclamen/nextflow_project/NEXTFLOW_Gadus_morhua/pipeline $

```

4.2 Analyse des résultats MultiQC

4.2.1 General Statistics

Sample Name	M Reads Mapped	% Dups	5'-3' bias	M Aligned	% Alignable	% Proper Pairs	Error rate	M Non-Primary	M Reads Mapped	% Mapped	% Proper Pairs	M
brain_R1	59.4	9.9%	1.21	27.0	65.6%	64.3%	0.61%	5.3	54.1	80.7%	80.5%	67
brain_R1_1												
brain_R1_2												
ovary_R1	47.7	22.0%	1.24	20.0	92.2%	58.7%	1.01%	7.5	40.2	94.6%	94.4%	42
ovary_R1_1												
ovary_R1_2												

Figure 10: General Statistics

Sur la figure nous pouvons voir un taux des reads alignées 65.6% pour brain, et 92.2 % pour ovary, ce qui semble pas souhaitable d'une première vue.

Un pourcentage des duplicats 9.9% pour brain et 22% pour ovary, mais cela ne peut être comme un inconvénient.

Un pourcentage des reads mapped (alignés avec succès) autour des 80.7% pour brain et 94.6% pour ovary ce qui est suggère potentiellement à faire une étape de filtrage.

Un pourcentage de GC 50% qui est un peu élevé par rapport au taux de GC chez cette espèce, ref ncbi(45.5%), cela est dû probablement au taux de duplicats élevé, comme suggéré dans l'analyse précédente.

4.2.2 Biotype Counts

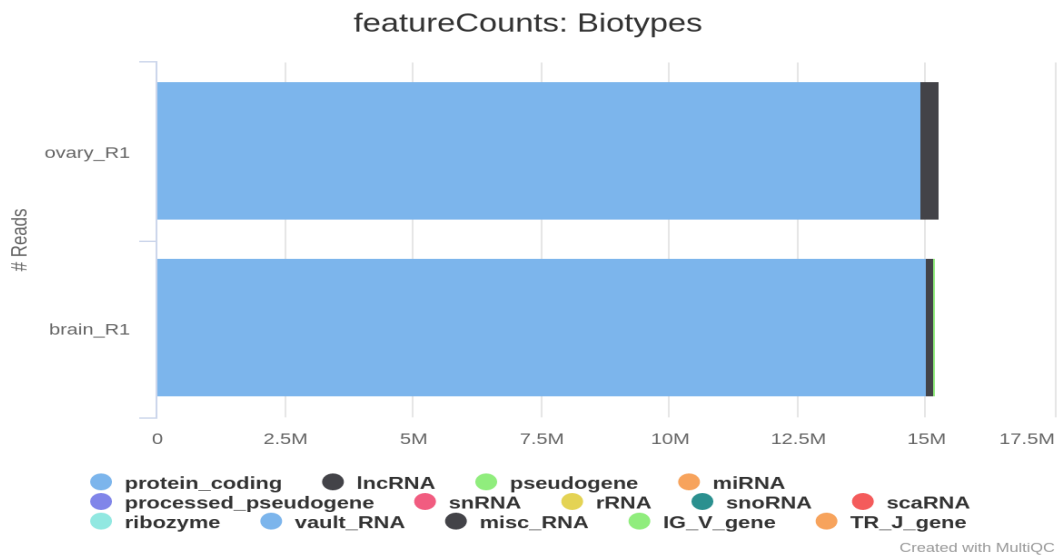


Figure 11: Biotype Counts

Nous pouvons voir sur cette figure que pour les deux conditions, environ 98% des reads sont associés à des protéines codantes, par contre nous observons la présence d'autres petits RNA, également de pseudogènes.

4.2.3 QualiMap

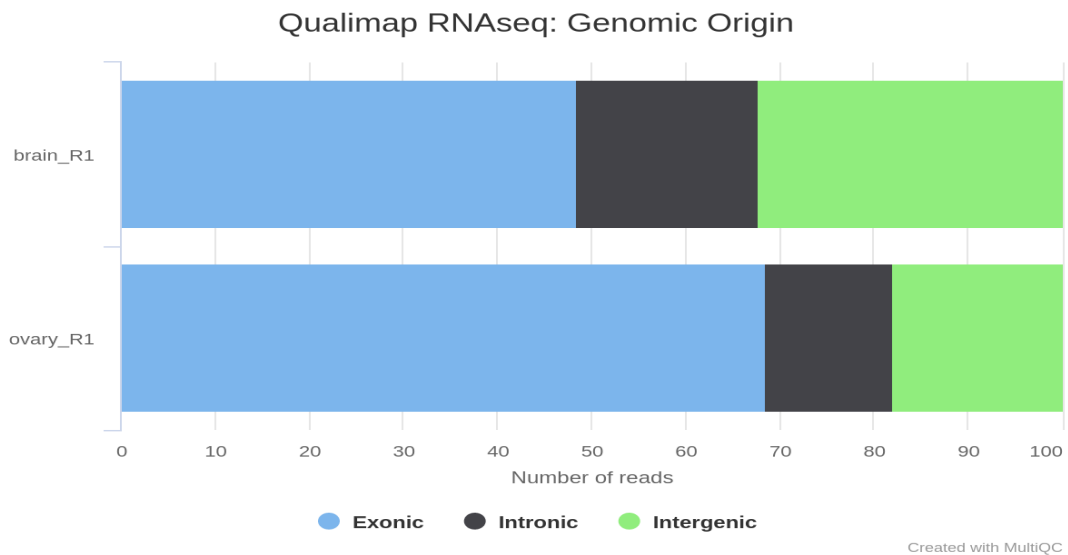


Figure 12: Genomic origin of reads

Sur cette figure nous pouvons également voir que le taux des reads exonique représente 48.4% pour brain et 68.5% pour ovary, et les taux intergénique et introniques sont élevés, ce qui indique la présence de contaminants.

4.2.4 FastQC

Le Phread-score des reads est bon au début, après se baisse un peu avant la fin des reads, cela et difficile de dire que tout nos reads sont de mauvaises qualité, puisque comme vu précédement il y a présence des petits ARN.



Figure 13: Sequence Quality Histograms before Trimming

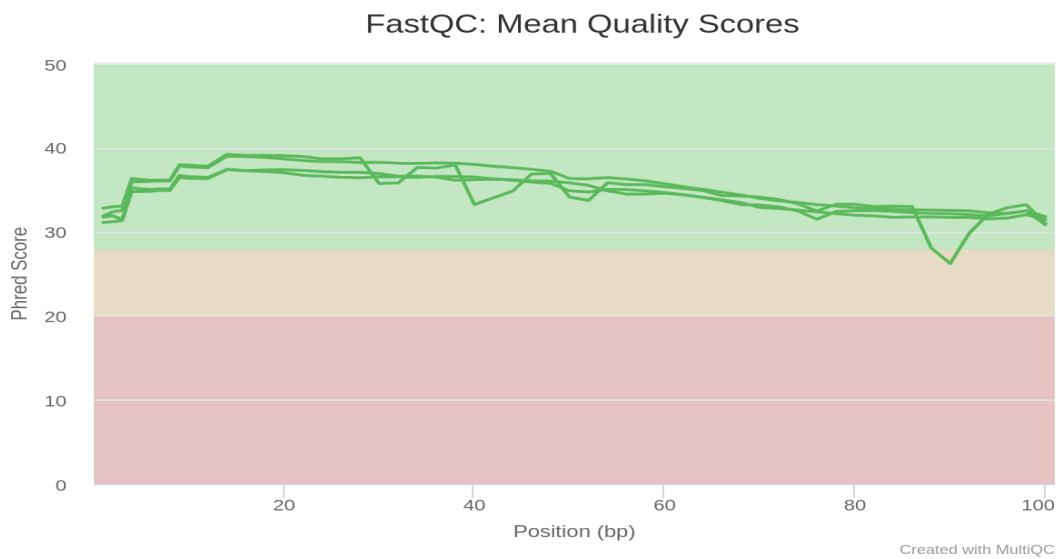


Figure 14: Sequence Quality Histograms After Trimming

5 Conclusion et Perspectives

Ce projet nous a permis de nous initier à la construction d'un pipeline RNA-seq avec Nextflow, de nous familiariser avec les paramètres et leur configuration, ainsi que de lancer ce pipeline sur des données en local ou présentes sur le site de NCBI. Dans un second temps, nous avons essayé d'interpréter les résultats du MultiQC, qui rassemble l'ensemble des résultats des outils de traitement en un seul fichier HTML. Néanmoins, il y aura des améliorations à apporter au niveau de la dernière partie, concernant les données NCBI. Il faudra revoir les fichiers inputs utilisés et essayer d'enlever les contaminants causés par les petits ARN, et également ajouter une partie sur DESeq2 au pipeline afin de quantifier les gènes qui sont différentiellement exprimés selon différentes conditions.

6 Références

https://hbctraining.github.io/variant_analysis/lessons/08_evaluate_QC.html

<https://github.com/nf-core/rnaseq/blob/master/docs/output.md#quality-control>

https://github.com/hbctraining/Intro-to-rnaseq-hpc-salmon/blob/master/lessons/qc_fastqc_assessment.md