Projet Nextflow

Claire Delamare Deboutteville

Octobre 2023 M2-Bioinformatique

1 Introduction

Dans le cadre de ce projet, nous allons utiliser Nextflow sur un cluster de calcul afin de réaliser des analyses de données de séquençage ARN. Nextflow permet l'automatisation des workflows bioinformatiques, ce qui simplifie grandement la réalisation d'analyses complexes. De plus, nous utiliserons le cluster Genotoul pour exécuter nos workflows, ce qui permettra de tirer parti de la puissance de calcul disponible pour traiter nos données de manière efficace.

Au cours de ce document, nous passerons en revue les différentes étapes du projet, depuis la préparation de l'environnement de travail, en passant par la configuration et l'implémentation des différents document d'entrée et l'exécution de workflows Nextflow, jusqu'à l'interprétation des résultats obtenus avec MultiQC. Nous utiliserons également des données provenant de la base de données NCBI pour appliquer notre workflow a un autre exemple que celui de la tomate (vu en TP).

L'objectif de ce projet est de se familiariser avec l'utilisation de Nextflow et du cluster Genotoul pour l'analyse de données de séquençage en bioinformatique.

2 Exercice 1 : Préparation de l'environnement de travail

2.1 Connexion à Genologin et préparation de l'espace de travail

- Connexion :

  Afin de se connecter au cluster Genologin, il faut un identifiant, ici 'geranium' :

  ```
  > ssh -XY geranium@genologin.toulouse.inrae.fr
  > mdp: f1o2r3!
  ```

- Préparation de son environnement de travail:

  On télécharge les fichiers fastq, annotation et génome de référence et on les range dans des sous-dossiers du même nom:

```
> wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/
> mkdir nom_nouveau_dossier
> mv source destination
```

3 Exercice 2 : Configuration et exécution de Nextflow

- Utilisation du pipeline Nextflow nf-core/rnaseq pour le traitement RNAseq des données de tomates.

- Paramétrage du run en expliquant les choix de paramètres.

```
#!/bin/bash //Permet de dire que c'est un executable shell bash
#SBATCH -J nfcornaseq // J pour "job" permet de donner le nom du job
#SBATCH -p unlimitq // Permet de notifier la file d'attente dans le cluster
#SBATCH --mem=6G  //Permet de limiter la mémoire necessaire pour le job
// Ces 4 lignes sont spécifiques a slurm qui est un gestionnaire de tâches

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

input=/work/geranium/projet/inputs.csv
gtf=/work/geranium/projet/annotation/ITAG2.3_genomic_Ch6.gtf
fasta=/work/geranium/projet/genome/ITAG2.3_genomic_Ch6.fasta
config=/work/geranium/projet/cd_config.cfg

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

  On definit nos chemins pour les options input, fasta et gtf et -c qui correspond a la configuration. Input contient le chemin vers notre plan d'experimentation qui contient les chemins vers les fastq, fasta contient le chemin vers le génome de référence et gtf contient le chemin vers le fichier d'annotation. "nextflow run" permet de lancer le script en nextflow. "nf-core/rnaseq" permet d'aller chercher la pipeline rnaseq d'analyse de données de sequençage ARN et -r précise la version de cette pipeline. "-profile" permet de lancer les configurations du cluster choisi, ici genotoul. "–aligner" permet de choisir l'aligneur, ici star-rsem.

- Lancement du job sur le cluster avec sbatch.

```
> sbatch run_pipeline.sh
```

- Suivi du job avec seff et utilisation de l'option resume si nécessaire.

- Explication de la sortie de seff et de l'intérêt de resume:

  L'outil -resume permet de reprendre la pipeline là ou on s'était arrêté (par exemple s'il y a eu une interruption due à une erreur) au lieu de tout recommencer depuis le début. C'est très utile, surtout si notre pipeline est très longue niveau temps.

  ```
  > seff 50726218
  ```

  L'outil seff permet de suivre l'exécution du job.

  ```
  > geranium@genologin2 ~/work/projet $ seff 50726218
  Job ID: 50726218
  Cluster: genobull
  User/Group: geranium/formation
  State: COMPLETED (exit code 0)
  Cores: 1
  CPU Utilized: 00:02:00
  CPU Efficiency: 19.77% of 00:10:07 core-walltime
  Job Wall-clock time: 00:10:07
  Memory Utilized: 1.83 GB
  Memory Efficiency: 30.47% of 6.00 GB
  ```

  Job ID est le numero du job qui permet de le suivre sur le cluster. Cluster est le nom du cluster , ici "genobull". User/Group donne le nom du groupe et de l'utilisateur. State affiche l'état du job, "Running" pour en cours d'éxecution, "Fail" pour échoué et "Completed" pour job terminé. Cores indique le nombre de coeur utilise. "CPU utilized" indique le temps total pendant lequel les CPU ont été utilisé, ici 2 minutes. "CPU efficiency" indique l'efficacite d'utilisation des CPU, ici 19.77 pourcent ce qui indique que seulement 19.77 pourcent du temps CPU alloué a été effectivement utilisé. "Job wall-clock time" indique le temps utilisé par la tâche, ici 10 minutes et 7 secondes. "Memory utilized" indique que 1.83 GB de RAM ont été utilisés par la tâche pendant son execution. "Memory efficiency" indique l'efficacite de l'utilisation de la memoire ce qui signifie ici que seulement 30.47 pourcent de la memoire allouée a la tâche a été utilisée.

- Une fois la tâche complétée, on doit copier coller les fichiers résultats du serveur à son pc local pour les analyser. On ouvre un terminal en local :

```
> scp geranium@genologin.toulouse.inrae.fr:/home/geranium/work/projet/results/
multiqc/star_rsem/multiqc_report.html /home/cdelamare/Documents/resultat_nextflow/
projet1
```

4 Exercice 3 : Interpretation des resultats

4.1 Interprétation des fichiers de sortie

- Expliquation des principaux repertoires et fichiers de sortie :

Basic Statistics

Measure	Value
Filename	mutant_R1_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1624613
Sequences flagged as poor quality	0
Sequence length	101
%GC	41

Figure 1: Stat mut R1 2

A la fin de la pipeline, on obtient un dossier "results" contenant des sous dossiers "fastqc", "genome", "multiqc", "$pipeline_info$", "$star_rsem$" et "trimgalore". "fastqc" contient les html qui evaluent le contrôle qualite pour chaque échantillon. "genome" contients les index et différents formats du génome. "multiqc" contient le rapport multi contrôle qualité, c'est à dire pour tous nos échantillons. "$star_rsem$" contient les fichiers qui ont été nécessaire pour l'aligneur, tel que le compte des gènes, les fichiers bam et les fichiers bigwig. "trimgalore" contient les rapport de trimming.

- Interprétation des résultats d'un fichier fastqc : Nous allons tous d'abord regarder les fastqc pour analyser la qualité d'un échantillon. Nous allons détailler l'échantillon mutant R1 numero 2. Nous pouvons voir ici que le total des séquences est de 1 624 613 (fig1). Il y a 0 sequences de mauvaise qualité, ce qui est positif. La longueur des sequences est de 101pb et le pourcent de GC est de 41.

  Figure 2, nous avons la qualité des sequences par base. Cela commence a chuter vers 79 donc on peut dire que globalement nos sequences sont de bonne qualité. Figure 3, nous avons les scores de qualité phred des séquences. Il est gobalement de 38, ce qui est un bon score donc nos séquences sont de bonne qualité.

  En figure 4, on a le GC content qui est autour de 41 globalement. Ce qui est cohérent avec la figure 1.

  Nous n'avons pas de séquences surrépresentées ni de contenu des adaptateurs. Donc on peut en déduire que tout s'est bien passé.

4.2 Analyse du rapport MultiQC

- Explication des informations fournies par le rapport MultiQC.

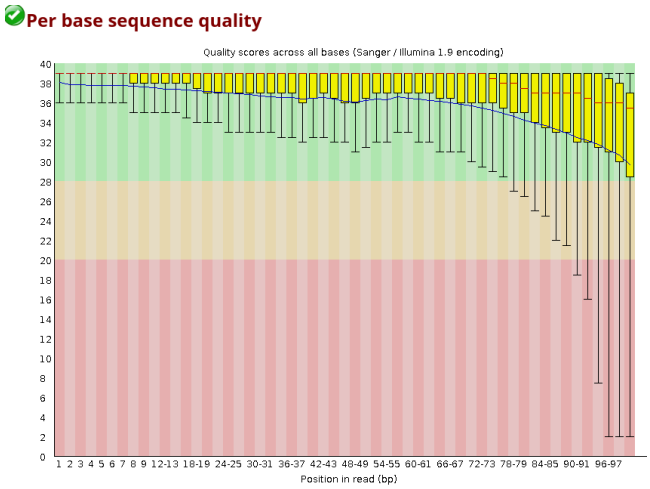


Figure 2: pbseqquality mut R1 2

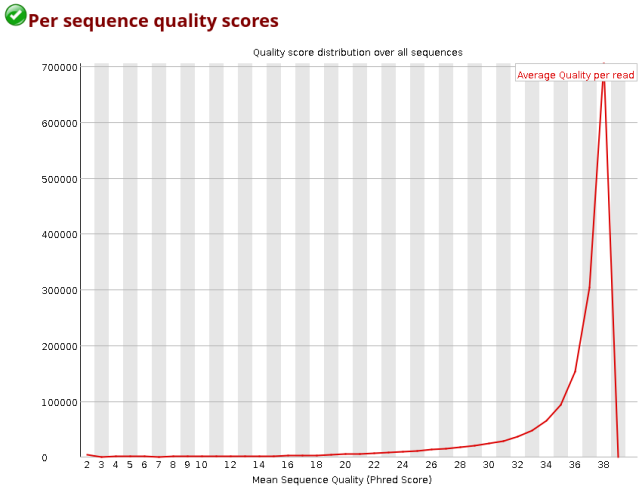


Figure 3: seqqualityscores mut R1 2

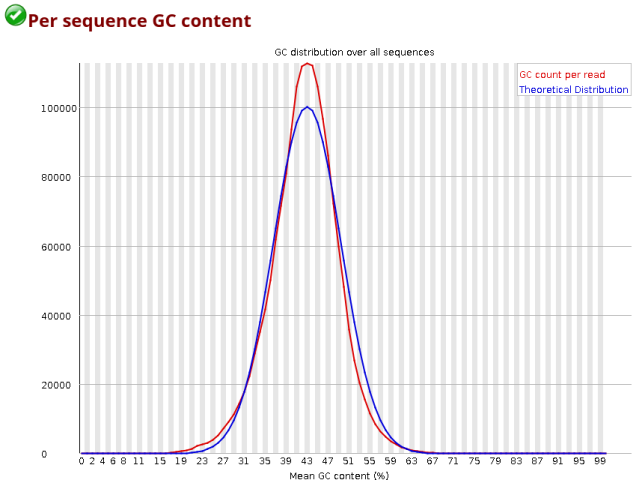


Figure 4: GC content mut R1 2

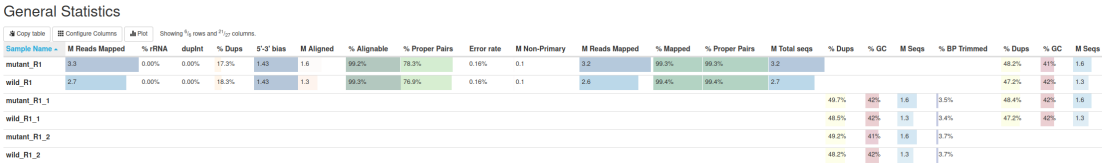


Figure 5: multiQC: General statistics

MultiQC permet de scanner tous les résultats obtenu en contrôle qualité pour générer un rapport unique reprenant tous ces résultats pour tous les échantillons en même temps.

En figure 5, nous avons le résumé des statistiques général du contrôle qualité. Etant donnée la largeur du tableau il est difficilement lisible. Voici les chiffres clés :

- Millions de lectures cartographiées : 3.3 (mutant) et 2.7 (wild)

- Pourcent d'ARN ribosomique : 0 pour les 2

- Pourcentage de duplication : 17.3 (mutant) et 18.3 (wild)

- Millions de lectures alignées : 1.6 (mutant) et 1.3 (wild)

- Pourcentage de lectures alignables : 99.2 (mutant) et 99.3 (wild)

- Pourcentage de GC : Entre 41 et 42

- Pourcentage de pair de base trimmés avec cutadapt: entre 3.4 et 3.7

- Pourcentage de reads dupliqués avec fastQC avant "trimmage" : entre 48.2 et 49.7

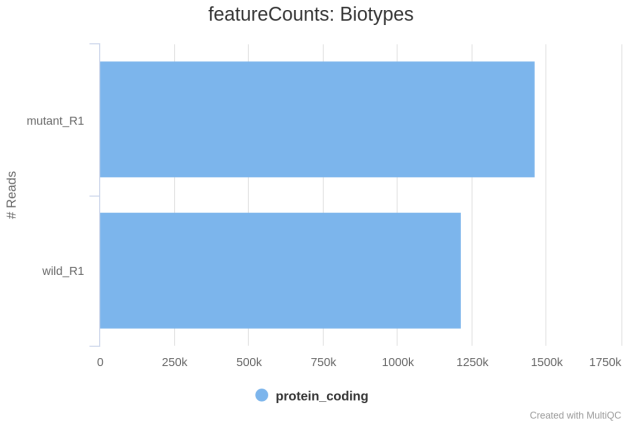


Figure 6: multiQC: feature counts mut/wild

Les statistiques avant et aprés trimmage sont très semblables. Globalement, on peut dire que nos statistiques sont bonnes.

Figure 6, on a le comptage des features. Ici, on a uniquement des reads codants pour des protéines pour nos 2 échantillons.

La figure 7 nous indique qu'il y a une majorité (autour de 80 pourcent) de duplicats unique, ce qui est une bonne chose. Moins on a de duplicatat non optical mieux c'est. On a également des lectures non cartographiées en faible quantité.

L'origine des séquences est majoritairement exonique (environ 83 pourcent) et on a environ 8 pourcent d'introns et de séquences intergénique pour nos deux échantillons (figure 8).

Le graphe suivant (9) montre la répartition de l'alignement des lectures par échantillon. Pour nos deux échantillons, on a une immense majorité (inf à 98 pourcent) des lectures qui s'alignent de manière unique à un gène, tandis que 0.6 (wild) et 0.5 pourcent (mutant) s'alignent à de multiples gènes. Aucun gène n'a été filtré due à de trop nombreux alignements. En revanche, environ 0.7 pourcent des gènes n'ont pas pu être alignés pour nos deux échantillons.

Sur le figure 10, on peut voir le nombre d'occurence d'une lecture. Ainsi, on a presque 100 000 lectures qui ont une seule occurence, et environ 100 lectures qui ont une 50aine d'occurences.

La figure 11 présente une inférence du sens des lectures. Ici, pour les deux échantillons on a 50 pourcent de sens et d'antisens.

- Conclusion: L'analyse des résultats a révélé que la qualité des échantillons de tomate était globalement bonne, avec des statistiques conformes aux attentes pour la plupart des mesures.

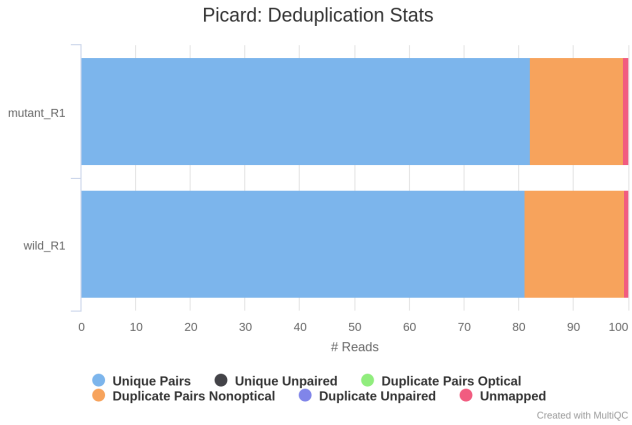


Figure 7: multiQC: Picard deduplication rate

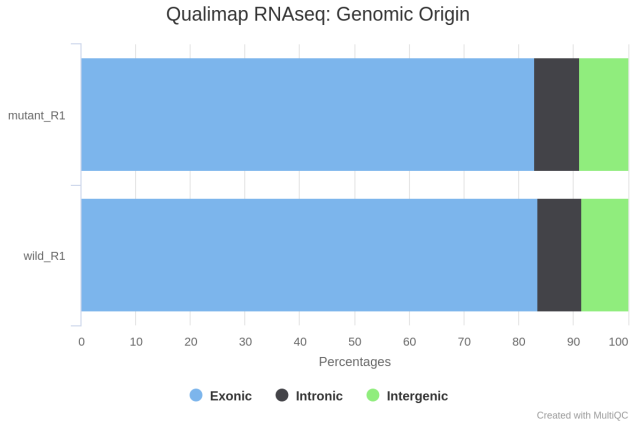


Figure 8: multiQC: Qualimap genomic origin

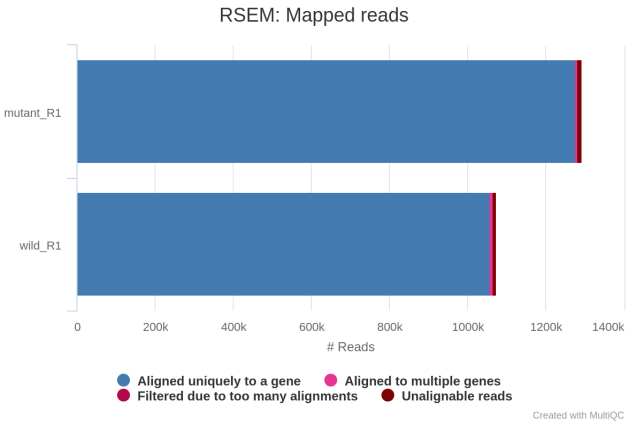


Figure 9: multiQC: Assignment plot

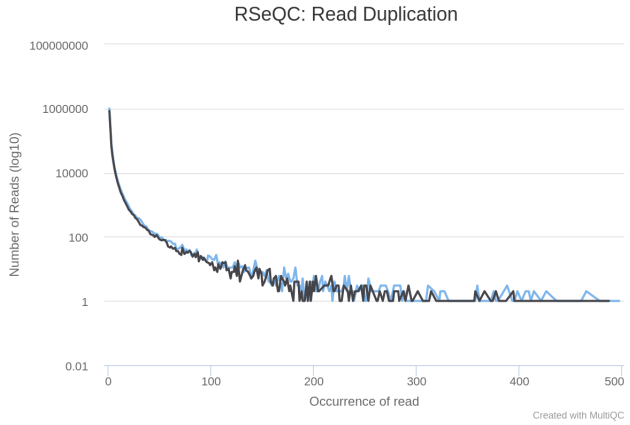


Figure 10: multiQC: Reads duplicate plot

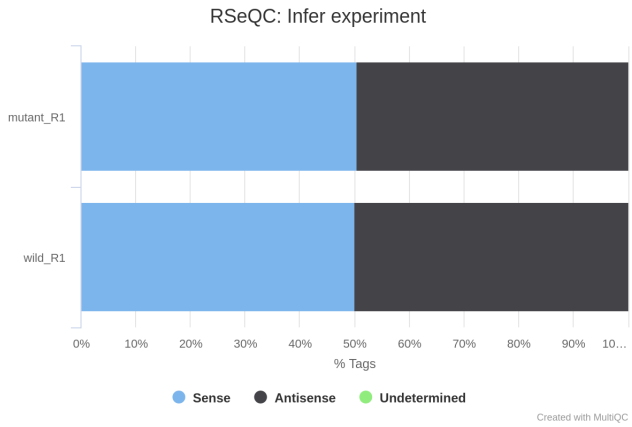


Figure 11: multiQC: Reads sens inference

5 Exercice 4 : Utilisation de données NCBI

- Sélection de trois échantillons à partir du site NCBI avec des références FASTA associées. Les fichiers fastq se trouvent sur le NCBI:

  https://www.ncbi.nlm.nih.gov/sra/?term=SRR2045415

  Pour télécharger les données fastq au format gunzip, on utilise fasterq-dump de sratoolkit sur nos 3 échantillons:

  ```
  > fasterq-dump -e 3 --gzip SRR2045415 SRR2045416 SRR2045417
  ```

- Description de nos échantillons:

  Ces 3 échantillons ont été prélevés sur la morue dans un projet d'assemblage de génome de la morue atlantique ($https://figshare.com/collections/Transcript_assembly_and_peptide_se$ SRR2045415 est un prélévement des ovaires, SRR2045416 du cerveau et SRR2045417 des branchies. Nous les appellerons respectivement 15, 16 et 17 pour la suite.

- On télécharge également le génome de référence et le fichier d'annotation de la morue "gadus morhua" sur le site Ensembl:

  $https://ftp.ensembl.org/pub/release-110/fasta/gadus_morhua/dna/$

  $https://ftp.ensembl.org/pub/release-110/gtf/gadus_morhua/$

  On importe le tout sur le cluster.

  ```
  > scp source destination
  ```

  Chaque fichier est rangé dans un sous-dossier fastq, annotation ou genome.

```
geranium@genologin2 ~/work/projet2 $ cat inputs.csv
group,replicate,fastq_1,fastq_2,strandedness
SRR2045415,1,/work/geranium/projet2/fastq/SRR2045415_1.fastq.gz,/work/geranium/projet2/fastq/SRR2045415_2.fastq.gz,forward
SRR2045416,1,/work/geranium/projet2/fastq/SRR2045416_1.fastq.gz,/work/geranium/projet2/fastq/SRR2045416_2.fastq.gz,forward
SRR2045417,1,/work/geranium/projet2/fastq/SRR2045417_1.fastq.gz,/work/geranium/projet2/fastq/SRR2045417_2.fastq.gz,forward
```

Figure 12: Cat inputs

```
geranium@genologin2 ~/work/projet2 $ cat run_pipeline.sh
#!/bin/bash
#SBATCH -J clairedelamarernaseq2
#SBATCH -p unlimitq
#SBATCH --mem=6G

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

input=/work/geranium/projet2/inputs.csv
gtf=/work/geranium/projet2/annotation/Gadus_morhua.gadMor3.0.110.gtf.gz
fasta=/work/geranium/projet2/genome/Gadus_morhua.gadMor3.0.dna.toplevel.fa.gz
config=/work/geranium/projet2/cd_config.cfg

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

Figure 13: Cat Run pipeline

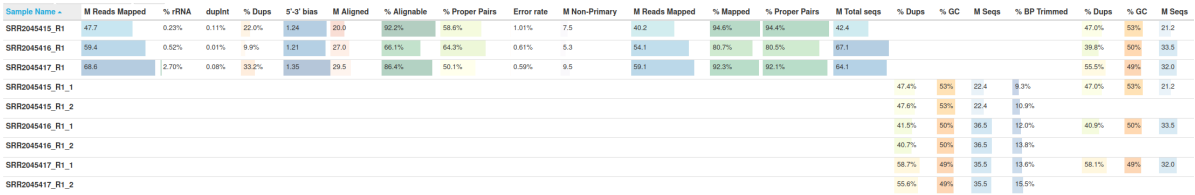
- On modifie tous les chemins dans les fichiers inputs (Fig 12) et runpipeline (Fig 13).

  Et on lance le pipeline sur nos trois échantillons.

  ```
  > sbatch run_pipeline.sh
  > seff 50754051
  Job ID: 50754051
  Cluster: genobull
  User/Group: geranium/formation
  State: COMPLETED (exit code 0)
  Cores: 1
  CPU Utilized: 00:03:04
  CPU Efficiency: 1.05% of 04:53:09 core-walltime
  Job Wall-clock time: 04:53:09
  Memory Utilized: 2.01 GB
  Memory Efficiency: 33.49% of 6.00 GB
  ```

- Pour importer les résultats:

  ```
  scp geranium@genologin.toulouse.inrae.fr:/home/geranium/work/projet2/results/multiqc/
  star_rsem/multiqc_report.html /run/media/cdelamare/ESD-USB/M2/Galaxy_Nextflow/Projet/
  ```



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	% GC	M Seqs	% BP Trimmed	% Dups	% GC	M Seqs
SRR2045415_R1	47.7	0.23%	0.11%	22.0%	1.24	20.0	92.2%	58.6%	1.01%	7.5	40.2	94.6%	94.4%	42.4					47.0%	53%	21.2
SRR2045416_R1	59.4	0.52%	0.01%	9.9%	1.21	27.0	66.1%	64.3%	0.61%	5.3	54.1	80.7%	80.5%	67.1					39.8%	50%	33.5
SRR2045417_R1	68.6	2.70%	0.08%	33.2%	1.35	29.5	86.4%	50.1%	0.59%	9.5	59.1	92.3%	92.1%	64.1					55.5%	49%	32.0
SRR2045415_R1_1															47.4%	53%	22.4	9.3%	47.0%	53%	21.2
SRR2045415_R1_2															47.6%	53%	22.4	10.9%			
SRR2045416_R1_1															41.5%	50%	36.5	12.0%	40.9%	50%	33.5
SRR2045416_R1_2															40.7%	50%	36.5	13.8%			
SRR2045417_R1_1															58.7%	49%	35.5	13.6%	58.1%	49%	32.0
SRR2045417_R1_2															55.6%	49%	35.5	15.5%			

Figure 14: multiQC: general statistics

```
resultats_nextflow/projet2/

scp geranium@genologin.toulouse.inrae.fr:/home/geranium/work/projet2/results/fastqc/
*.html /run/media/cdelamare/ESD-USB/M2/Galaxy_Nextflow/Projet/resultats_nextflow/projet
```

- Résultats obtenus :

  Nous allons décrire la figure 14 qui reprends les statistiques générales de nos échantillons 15, 16 et 17 dans cet ordre:

  - Millions de lectures cartographiées : 47.7, 59.4 et 68.6

  - Pourcent d'ARN ribosomique : 0.23, 0.52 et 2.7

  - Pourcentage de duplication : 22, 9.9 et 33.2

  - Millions de lectures alignées : 20, 27 et 29.5

  - Pourcentage de lectures alignables : 92.2, 66.1 et 86.4

  - Taux d'erreurs en pourcent : 1.01, 0.61 et 0.59

  - Pourcentage de GC : 53, 50 et 49

  - Pourcentage de pair de base trimmés avec cutadapt: 9.3 et 10.9, 12 et 13.8, 13.6 et 15.5

  - Pourcentage de lectures dupliquées avec fastQC avant "trimmage" : 47.4 et 47.6, 41.5 et 40.7, 58.7 et 55.6

  Globalement, les statistiques sont bonnes. Sauf pour l'échantillon 16 qui n'a que 66.1 pourcent de ses lectures alignable.

  Figure 15, on a la répartition des reads par "feature". On peut voir que la majorité des reads sont des sequences codant des protéines (en bleu). Il y a des long ARN non codant a moins de 3 pourcent pour tous les échantillons (en noir) et en jaune les arn ribosomiaux également en faible quantité (inf à 3 pourcent). Pour l'échantillon 16, on a la présence faible de pseudogènes en vert (inf à 0.3 pourcent).

  La figure 16 montre un résumé de la distribution de la duplication des gènes. On peut s'attendre, pour les gènes fortement exprimés, à avoir beaucoup de lectures en double, mais un grand nombre de duplicatas avec de faibles comptages de lectures peuvent indiquer une faible complexité de

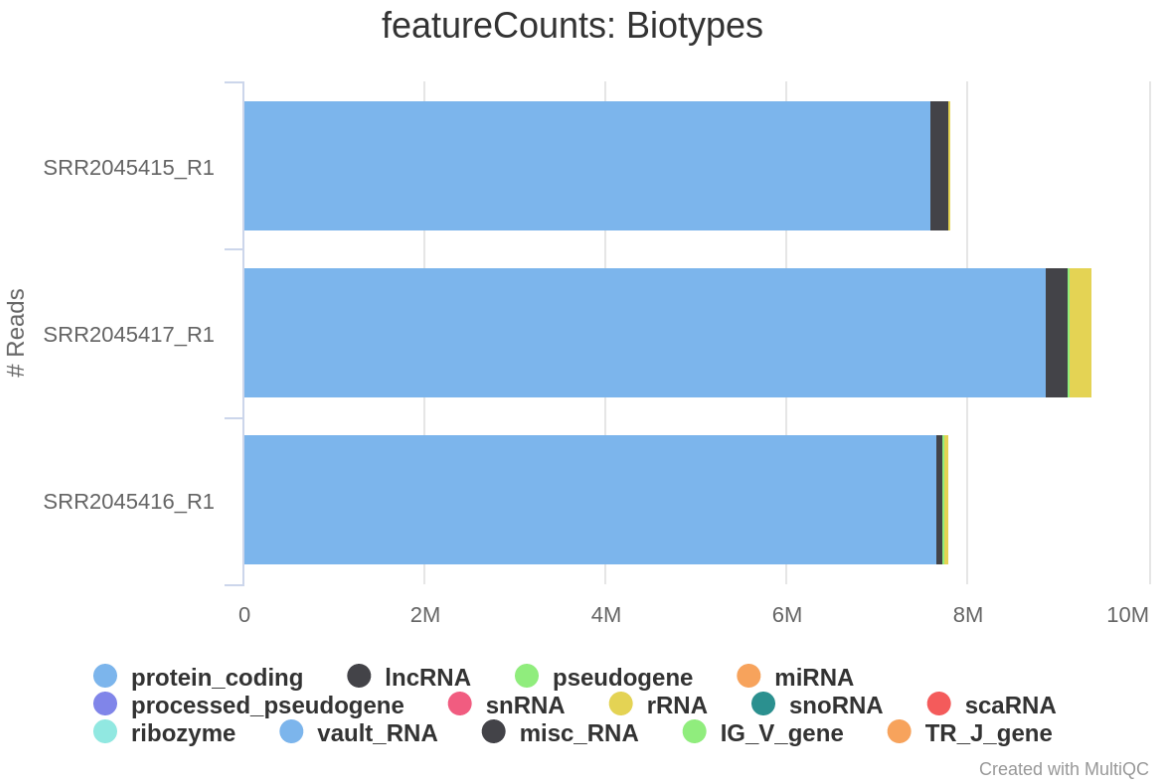


Figure 15: multiQC: feature counts

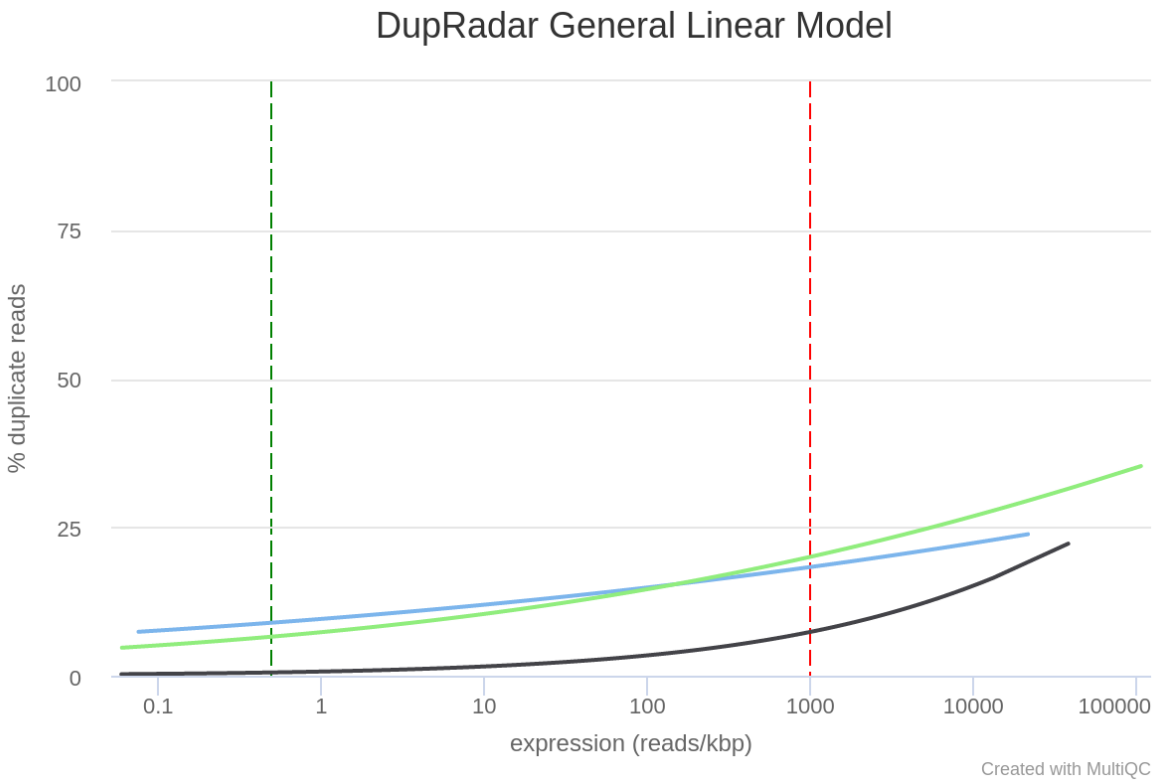


Figure 16: multiQC: Dupradar general linear model

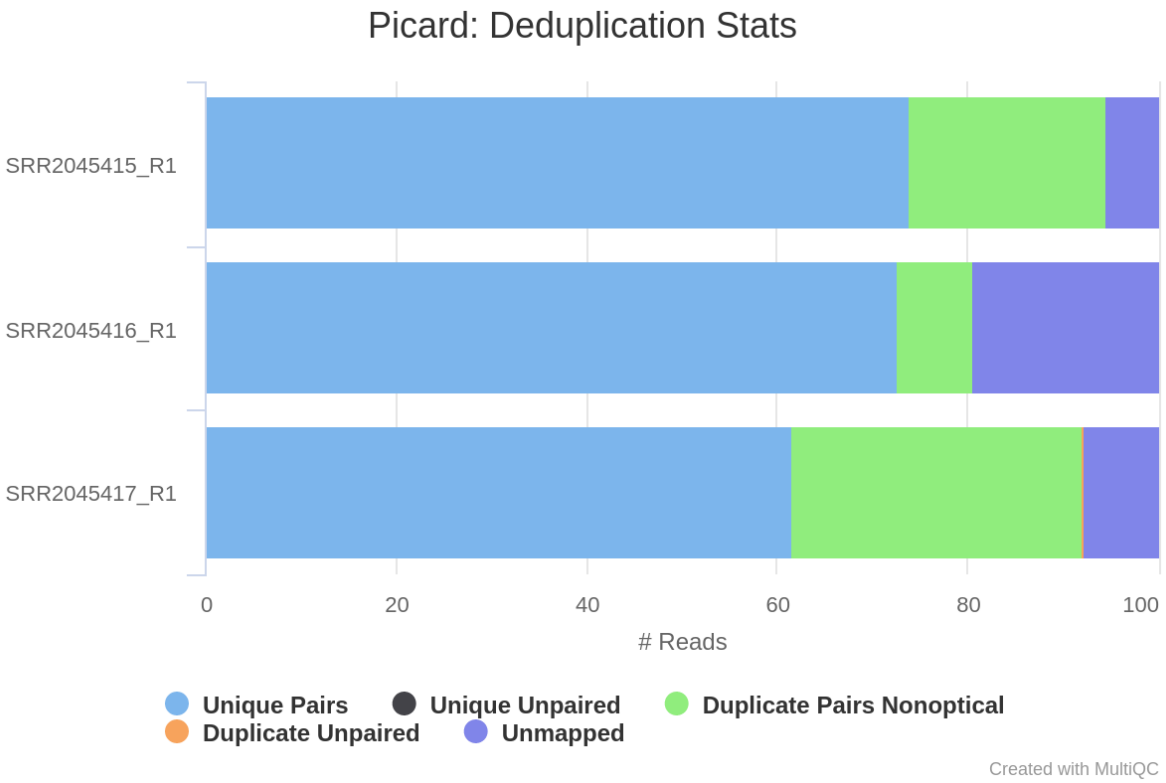


Figure 17: multiQC: Picard deduplication

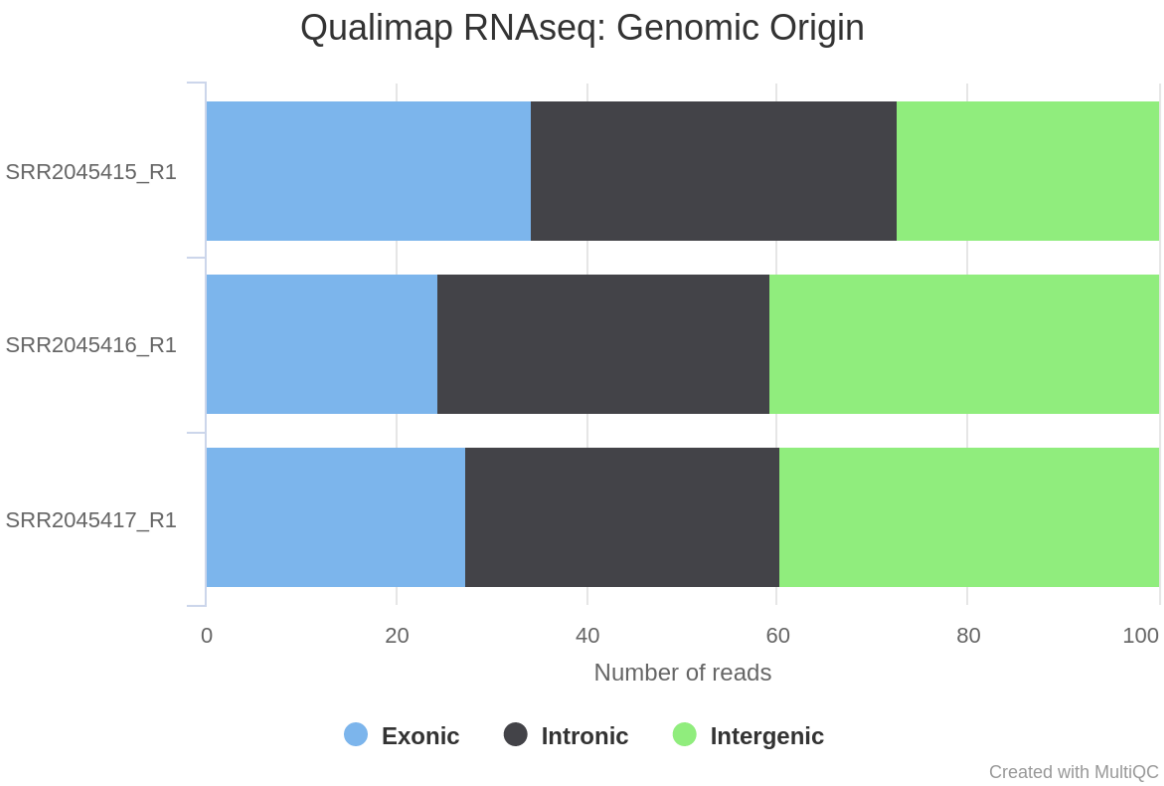


Figure 18: multiQC: Qualimap genomic origin

la bibliothèque due à une duplication technique. On voit ici que ce n'est pas le cas. Notre distribution de la duplication des gènes est cohérente.

La figure 17, nous montre les statistiques de déduplication. Par défaut, le pipeline utilise Picard MarkDuplicates pour marquer les lectures en double identifiées parmi les alignements, ce qui nous permet d'évaluer le niveau global de duplication dans nos échantillons. Une grande proportion de duplicatas non optiques suggère que la duplication n'est pas principalement due à une véritable duplication biologique, mais plutôt à des biais techniques introduits lors de la préparation de la bibliothèque ou de l'amplification par PCR. Cela peut être dû à des étapes de préparation de la bibliothèque qui favorisent la réplication artificielle de certaines régions, ce qui peut fausser les résultats de séquençage. Ici, on a en majorité des duplicatas unique (entre 61 et 73 pourcent) mais également une grande proportion de duplicatas non optique (20 et 30 pourcent pour l'échantillon 15 et 17). On en déduit une meilleure qualité de l'échantillon 16 de ce point de vue.

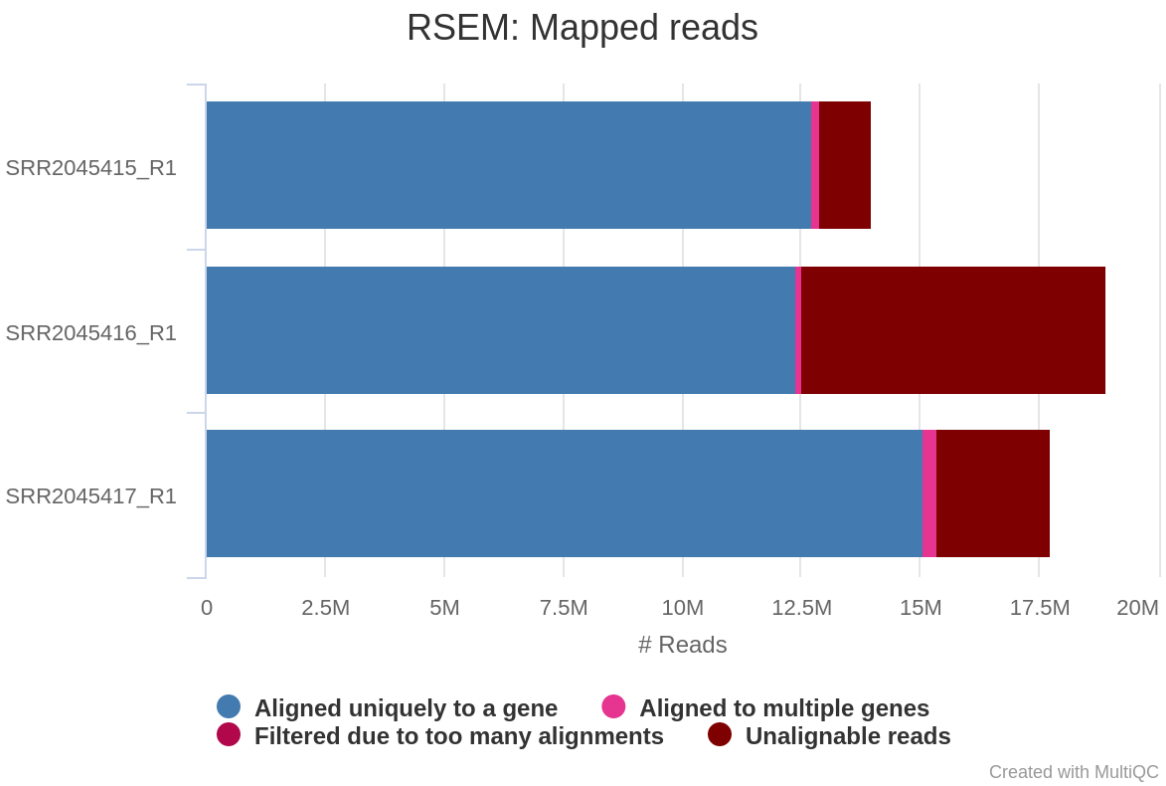


Figure 19: multiQC: Assignment plot

La répartition d'origine des lectures est d'environ 1/3 d'exons, 1/3 d'introns et 1/3 d'intergenic pour nos 3 échantillons (Fig 18). Une répartition équilibrée entre ces trois catégories peut indiquer que le séquençage RNA-seq a capturé de manière représentative l'ensemble du transcriptome.

La figure 19, comme vu précédemment (Fig 14), nous montre que l'échantillon 16 a le plus fort taux de lectures non alignables (34 pourcent) contre 7.8 et 13.6 pour les échantillons 15 et 17. La majorité des lectures s'alignent de manière unique à un gène mais on a 1.1 , 0.6 et 1.7 pourcent des gènes qui s'alignent de facçon multiple pour les échantillons 15, 16 et 17.

Sur la figure 20, on peut voir que la répartition des lectures ne se résume pas a des séquences codantes (exons) mais qu'il y a également des 5'UTRExons, 3'UTRExons, introns, TSS, TES et des séquences intergeniques.

En figure 21, on a la fraction de lectures alignées par chromosome. On observe une forte différence pour le chromosome 19, ou l'échantillon 16 (en bleu) contient beaucoup plus de lectures que les échantillon 15 et 17.

La qualité des lectures décroit et passe de bonne a moyenne a partir de

Figure 20: multiQC: Distribution plot

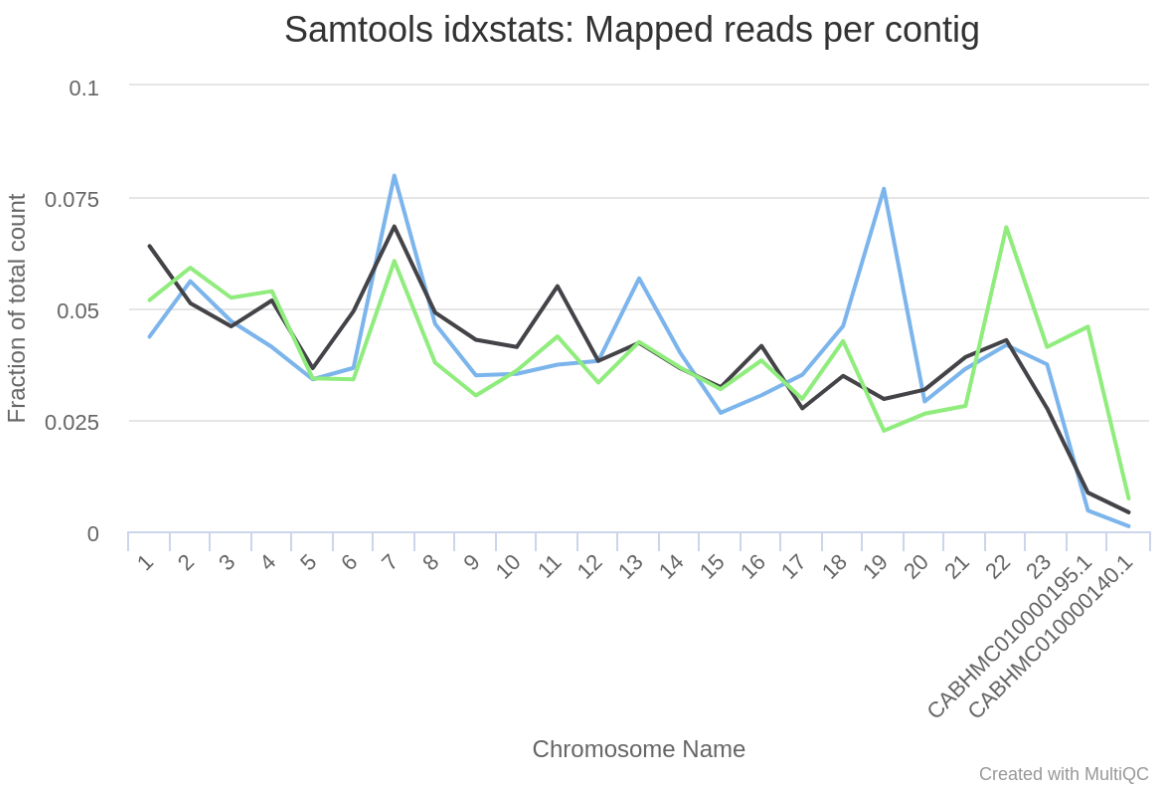


Figure 21: multiQC: mapped reads/chromosome

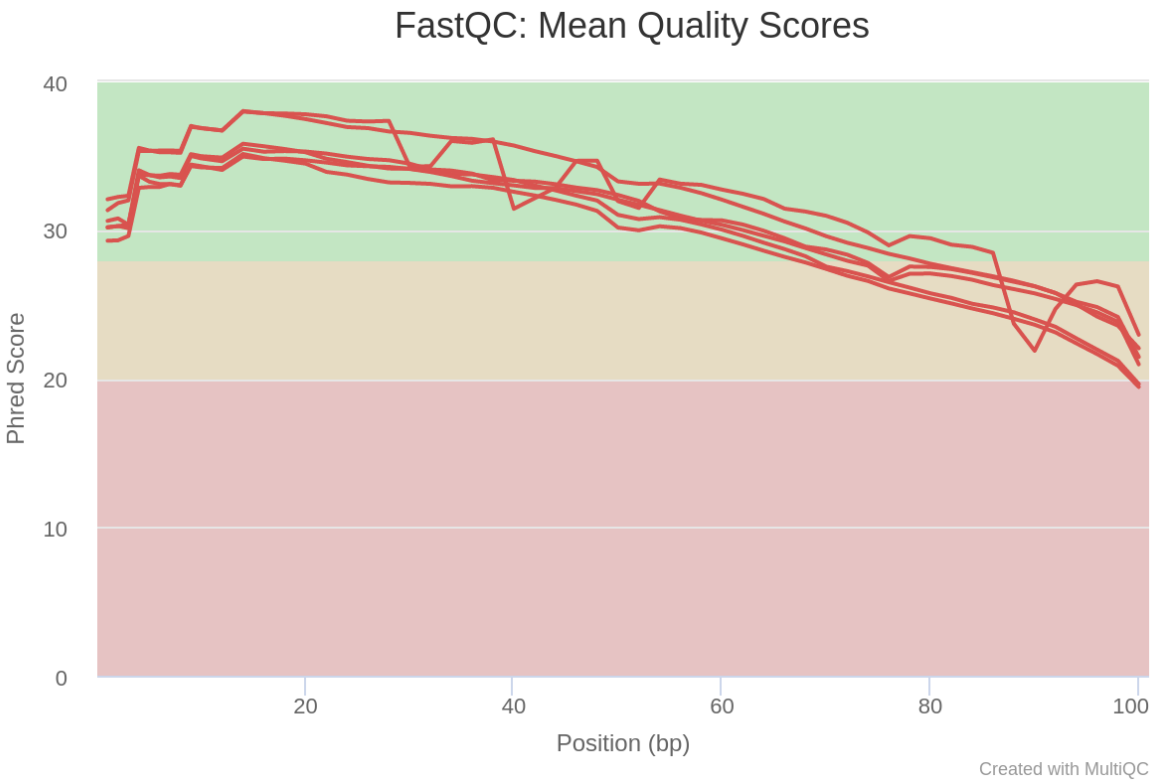


Figure 22: multiQC: Quality per base

la 70ieme base en moyenne pour tous les échantillons (Fig 22). On a tout de même un phred score de 37 pour tous nos échantillons, ce qui n'est pas mauvais.

- On peut en conclure globalement a une qualité relative de nos lectures et du séquençage. Les différences observées entre les échantillons de morue peuvent être liées à divers facteurs biologiques ou expérimentaux, et une analyse plus approfondie serait nécessaire pour en comprendre la signification biologique.

6 Conclusion

Nous avons étendu notre analyse en utilisant des données provenant de la morue atlantique, montrant ainsi notre capacité à adapter le pipeline à différentes données biologiques. Ce projet nous a permis de prendre en main l'outil Nextflow pour la gestion de pipelines bioinformatiques. Nous avons appris à configurer, exécuter et interpréter les résultats d'une pipeline RNAseq (notamment en analysant les rapport MultiQC) et a les lancer sur un cluster de calcul.