Master Bioinformatique et Biologie des Systems

Projet Initiation sur genologin

Nextflow

Étudiante : Tuong Bao Han PHAN
Professeur : Sarah MAMAN

Toulouse, October 10, 2023

Contents

1 Exercice 1 : Connexion à *Genologin*, création d'un répertoire de travail, téléchargement de fichiers à traiter

1.1 Connexion à Genologin

Pour la connexion à Genologin, il faut taper la commande ci-dessous pour accéder à mon compte de formation qui correspond au nom de fleur iris et tapez les mots de passe correspondant :

```
$ ssh -XY iris@genologin.toulouse.inrae.fr
iris@genologin.toulouse.inrae.fr's password: f1o2r3!
```

1.2 Préparer son espace de travail

Après réussir à accéder au serveur Genologin, j'ai créé un répertoire de travail `projet` dans mon `/work/iris/` en utilisant `mkdir` et télécharger les fichiers d'entrée du TP `ttp://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/` en utilisant `wget` et enregistrer dans le répertoire `exo12` qui est un sous-répertoire du répertoire `projet`.

Les fichiers de génome (`.fasta`) et d'annotation (`.gtf`) sont enregistrées dans les répertoires génome et annotation qui sont les sous-répertoires du répertoire `exo12` respectivement. Les fichiers de `fastq.gz` sont enregistrées dans le répertoire `fastq` qui est le sous-répertoire du répertoire `exo12`.

Voici ce sont les commandes:

```
$ cd /work/iris
$ mkdir projet
$ cd projet
$ mkdir exo12
$ cd exo12
$ mkdir genome
$ mkdir annotation
$ mkdir fastq
$ cd genome
$ wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.fasta
$ cd ../annotation
$ wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.gtf
$ cd ../fastq
$ wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep1_1_Ch6.fastq.gz
```

Dans le fichier fastq, faire le meme chose pour les autres fichiers `fastq.gz` donc à la fin, on a 4 fichiers `fastq` différents qui correspondent MT_rep1_1, MT_rep1_2, WT_rep1_1 et WT_rep1_2.

2 Exercice 2 : Préparer son fichier `bash` de lancement Nextflow

2.1 Préparer le fichier de lancement et lancer le job sur les données tomates

Afin de préparer le fichier de lancement, il faut créer deux fichiers `inputs.csv` et `sm_config.cfg` :

- `inputs.csv` permet décrire des échantillons de séquençage d'ARN avec les données sur les groupes et les réplicats des échantillons. Ici, on a 2 groupes : *mutant* et *wild* (sauvage et mutant) avec les réplicats qui spécifiquent le numéro de réplicat au sein de son groupe, les `fastq_1` et `fastq_2` contient le chemin vers le fichier `FASTQ` de lecture 1 et lecture 2 associé à l'échantillon. `strandedness` indique l'orientation de la bibliothèque de séquençage. Dans ce cas-là, les échantillons sont `unstranded` indique que l'orientation n'est pas spécifiée (non orientée) pour ces échantillons

```
group,replicate,fastq_1,fastq_2,strandedness
mutant,1,/work/iris/projet/exo12/fastq/MT_rep1_1_Ch6.fastq.gz,/work/iris/projet/exo12/fastq/
MT_rep1_2_Ch6.fastq.gz,unstranded
wild,1,/work/iris/projet/exo12/fastq/WT_rep1_1_Ch6.fastq.gz,/work/iris/projet/exo12/fastq/
WT_rep1_2_Ch6.fastq.gz,unstranded
```

- `sm_config.cfg` est une configuration spécifique pour un pipeline de séquençage ARN. On active la fonction de traçage, ce qui signifie que le pipeline enregistrera des informations de traçage sur son exécution dans un fichier nommé `pipeline_trace.txt`. Les champs d'informations du traçage sont spécifiés pour chaque tache du pipeline incluant l'identifiant de la tache, nom de la tache, le statut, le code de sortie, le temps réel, le pourcentage d'utilisation du CPU, la mémoire RSS et le script exécuté.

```
trace {
enabled = true
file = 'pipeline_trace.txt'
fields = 'task_id,name,status,exit,realtime,%cpu,rss,script'
```

Enfin, mon fichier batch `run_pipeline.sh` permet de lancer le job (nextflow) sur les données tomates:

```
#!/bin/bash
#SBATCH -J HanPHAN
#SBATCH -p workq
#SBATCH --mem=6G
#SBATCH --time=24:00:00

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

input=/work/iris/projet/exo12/inputs.csv
gtf=/work/iris/projet/exo12/annotation/ITAG2.3_genomic_Ch6.gtf
fasta=/work/iris/projet/exo12/genome/ITAG2.3_genomic_Ch6.fasta
config=/work/iris/projet/exo12/sm_config.cfg

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

Les paramètres de `nextflow run nf-core/rnaseq`

- `-r 3.0` : spécifie la version du pipeline Nextflow à exécuter (version 3.0 ou 3.4)
- `-profile genotoul` : utilise le profil spécifique appelé *genotoul* pour configurer l'exécution du pipeline.
- `-input $input -fasta $fasta -gtf $gtf` : spécifie les fichiers d'entrée nécessaires pour le pipeline, en utilisant les variables définies précédemment. Le pipeline va marcher en utilisant le chemin d'accès aux fichiers inputs (ce sont les fichiers `fastq` fournis dans le fichier `inputs.csv`) et aux fichiers `gtf` et `fasta` (ce sont les fichiers téléchargés dans le répertoire annotation et génome sur les données de tomate)
- `-aligner star_rsem` : spécifie l'aligneur à utiliser pour analyse (dans ce cas-là on utilise `star_rsem`) qui permet d'alignement et quantification des niveaux d'expressions de gènes et d'isoformes à partir de données de séquençage d'ARN. C'est un des outils de quantification les plus précis pour l'analyse ARN-seq. Il permet d'effectuer à la fois l'alignement et la quantification dans un seul package et sa capacité à utiliser efficacement les lectures mappage ambigu.
- `-c $config` : spécifie le fichier de configuration à utiliser pour le pipeline, en utilisant la variable "config"

Enfin, lancer le job sur le cluster avec *sbatch* en utilisant de la commande:

```
$ sbatch run_pipeline.sh
```

2.2 Suivre le job avec seff, utiliser resume si nécessaire

2.2.1 Expliquer la sortie du seff

Au début, la sortie du fichier à l'état `RUNNING` qui indique le travail est actuellement en cours d'exécution. Cela signifie que les calculs ou les taches spécificiées dans le travail sont en train d'être traités par le cluster. Enfin, si tout a bien marché, la sortie du seff sera affiché :

```
Job ID: 50745055 # identifiant unique attribue a ce travail sur le cluster pour le suivi et
# la gestion.
Cluster: genobull # nom de cluster informatique sur lequel le travail a ete execute
User/Group: iris/formation # utilisateur ou le groupe d'utilisateurs qui a soumis ce travail
State: COMPLETED (exit code 0) # le travail a ete execute avec succes et termine normalement.
# Code 0 : pas d'erreur majeure pendant l'execution du travail
Cores: 1  # le nombre de coeurs de processeur utilises pour ce travail
CPU Utilized: 00:01:51  # la temps CPU reellement utilisee par le travail dans 1min 51sec
CPU Efficiency: 18.41% of 00:10:03 core—walltime  # l'efficacite d'utilisation du CPU par rapport
# du temps total de calcul.
Job Wall—clock time: 00:10:03 # la dur e totale de l'execution du travail dans 10 mins et 3 sec
Memory Utilized: 1.89 GB # la quantite de memoire RAM utilisee par le travail pendant execution
Memory Efficiency: 31.57% of 6.00 GB # l'efficatite d utilisation de la memoire par rapport a
# la memoire totale disponible
```

2.2.2 Expliquer l'intérêt du resume (qu'il soit utilisé ou pas)

L'option `-resume` dans Nextflow permet de reprendre l'exécution d'un pipeline à partir d'un point précédemment interrompu sans avoir à tout recommencer depuis le début donc qui permet de gestion des erreurs, d'économie de temps, de suivi et gestion. De plus, en termes de l'alignement, avec cette option, les résultats ne seront pas écrasés lors de la reprise du pipeline et pourront être utilisés pour une analyse comparative entre les algorithmes d'alignement si nécessaire.

3 Exercice 3 : Interpréter le report MultiQC ainsi que les principaux fichiers résultats obtenus

3.1 Interprétation des principaux résultats

```
total 3
drwxr-xr-x  2 iris formation 4096  5 oct.  23:40 fastqc
drwxr-xr-x  4 iris formation 4096  5 oct.  23:40 genome
drwxr-xr-x  3 iris formation 4096  5 oct.  23:47 multiqc
drwxr-xr-x  2 iris formation 4096  5 oct.  23:47 pipeline_info
drwxr-xr-x 14 iris formation 4096  5 oct.  23:47 star_rsem
drwxr-xr-x  3 iris formation 4096  5 oct.  23:41 trimgalore
```

Figure 1: *Les sous-répertoires dans le répertoire* `resultat`

3.1.1 fastqc

```
-rw-r--r-- 1 iris formation 658481  5 oct.  23:40 mutant_R1_1_fastqc.html
-rw-r--r-- 1 iris formation 416878  5 oct.  23:40 mutant_R1_1_fastqc.zip
-rw-r--r-- 1 iris formation 654797  5 oct.  23:40 mutant_R1_2_fastqc.html
-rw-r--r-- 1 iris formation 412887  5 oct.  23:40 mutant_R1_2_fastqc.zip
-rw-r--r-- 1 iris formation 658647  5 oct.  23:40 wild_R1_1_fastqc.html
-rw-r--r-- 1 iris formation 415938  5 oct.  23:40 wild_R1_1_fastqc.zip
-rw-r--r-- 1 iris formation 654999  5 oct.  23:40 wild_R1_2_fastqc.html
-rw-r--r-- 1 iris formation 413133  5 oct.  23:40 wild_R1_2_fastqc.zip
```

Figure 2: *Les fichiers dans le répertoire* `resultat/fastqc`

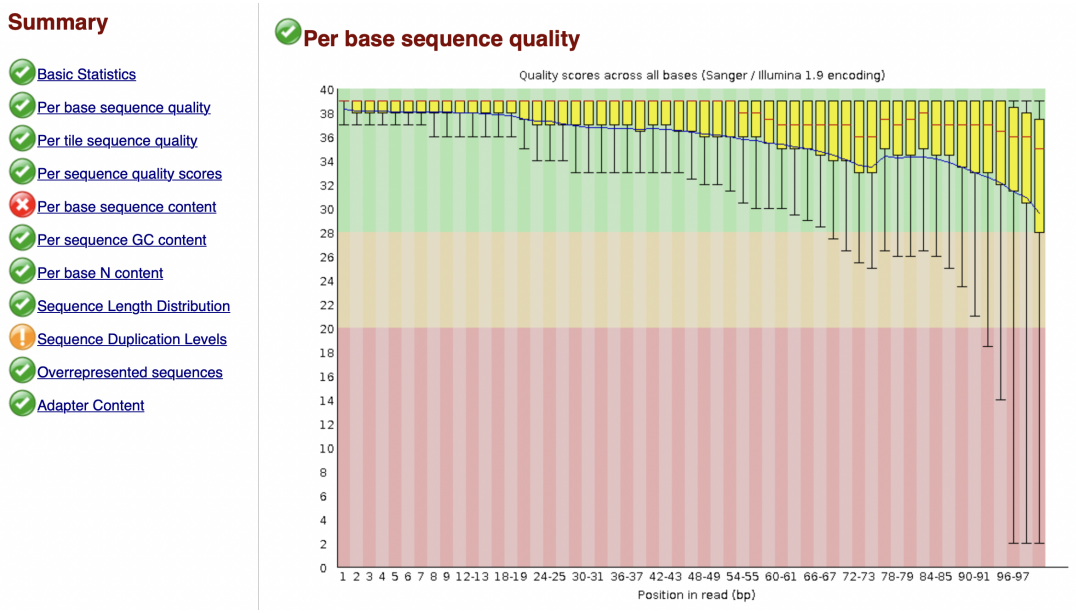


Figure 3: *FastQC de mutant_R1_1. Logiciel FastQC est un outil en bioinformatique pour évaluer la qualité des données de séquençage, en particulier les données issues du séquençage d'ADN ou d'ARN.*

La ligne rouge centrale est la valeur médiane. La case jaune représente l'écart interquartile (25-75%). Les moustaches supérieures et inférieures représentent les points 10% et 90%. La ligne bleue représente la qualité moyenne. L'axe des y sur le graphique montre les scores de qualité. Plus le score est élevé, meilleur est cette base. L'arrière-plan du graphique divise l'axe y en appels de très bonne qualité (vert), appels de qualité raisonnable (orange) et appels de mauvaise qualité (rouge). La qualité des appels sur la plupart des plateformes se dégradera à mesure que l'exécution progresse, il est donc courant de voir des appels de base tomber dans la zone orange vers la fin d'une lecture.

- Basic Statistiques [PASS] : cette section fournit des statistiques de base sur les données de séquençage, telles que la quantité totale de séquences, la longueur, moyenne des séquences, etc. Un résultat **PASS** signifie généralement que les données de base sont conformes aux normes attendues
- Per base sequence quality [PASS] : Cette section montre la qualité des bases à chaque position de séquence. Un résultat **PASS** indique que la qualité des bases est généralement bonne sur toute la longueur des séquences.
- Per tile sequence quality [PASS] : Cette section évalue la qualité des séquences en fonction de leur emplacement sur la puce de séquençage (carreaux ou "tiles"). Un résultat **PASS** signifie que la qualité est généralement homogène sur toute la puce.
- Per sequence quality scores [PASS] : Cela évalue la qualité globale des scores de qualité attribués à chaque séquence. Un résultat **PASS** signifie que la majorité des séquences ont des scores de qualité acceptables.
- Per base sequence content [FAIL] : Cette section évalue la répartition des bases à chaque position de séquence. Un résultat **FAIL** peut indiquer une distribution de bases inhabituelle qui peut nécessiter une attention particulière.
- Per sequence GC content [PASS] : Cette section évalue la répartition du contenu en GC (guanine-cytosine) sur l'ensemble des séquences. Un résultat **PASS** signifie que la distribution est conforme aux attentes.
- Per base N content [PASS]: Cette section évalue la répartition des bases ambiguës (représentées par "N") à chaque position de séquence. Un résultat **PASS** signifie que la proportion de bases ambiguës est généralement faible.
- Sequence Length Distribution [PASS] : Cette section montre la distribution des longueurs de séquence. Un résultat **PASS** signifie que les longueurs de séquence sont cohérentes avec

ce qui est attendu.

- Sequence Duplication Levels [WARNING] : Cette section évalue le niveau de duplication des séquences. Un avertissement **WARNING** peut indiquer une duplication élevée qui peut nécessiter une investigation plus approfondie.
- Overrepresented sequences [PASS] : Cette section identifie les séquences qui sont sur-représentées dans l'échantillon. Un résultat **PASS** signifie que les séquences sur-représentées sont sous contrôle.
- Adapter Content [PASS] : Cette section évalue la présence d'adaptateurs d'ADN dans les données de séquençage. Un résultat **PASS** signifie que la plupart des adaptateurs ont été correctement retirés.

3.1.2 genome

```
drwxr-xr-x 3 iris formation    4096  5 oct.  23:40 index
-rw-r--r-- 1 iris formation  320910  5 oct.  23:39 ITAG2.3_genomic_Ch6.bed
-rw-r--r-- 1 iris formation      29  5 oct.  23:39 ITAG2.3_genomic_Ch6.fasta.fai
-rw-r--r-- 1 iris formation      20  5 oct.  23:39 ITAG2.3_genomic_Ch6.fasta.sizes
-rw-r--r-- 1 iris formation 2034585  5 oct.  23:39 ITAG2.3_genomic_Ch6_genes.gtf
drwxr-xr-x 2 iris formation    4096  5 oct.  23:39 rsem
```

Figure 4: *Les fichiers dans le répertoire* `resultat/genome`

Ce répertoire contient le génome de référence avec l'annotation et d'autres fichiers en format différents. Le répertoire index contient les fichiers créés pour divers types de données, notamment les fichiers FASTA, BAM, etc. utilisé pour accélérer la recherche et l'accès aux données.

- Le fichier `bed` stocke les régions génomiques, les emplacements des gènes, les sites de liaison des facteurs de transcription, etc.
- Le fichier `fai` est un fichier d'index associé à un fichier `fasta` générée par le logiciel `samtools faidx` qui contient des informations sur les positions des séquences dans le fichier `FASTA`, ce qui permet un accès rapide et efficace à ses séquences.
- Le fichier `fasta.sizes` contient des informations sur les tailles des séquences ou des régions génomiques
- Le dossier `rsem` est associé au logiciel RSEM, qui est utilisé pour l'analyse de données de séquençages d'ARN pour quantifier l'expression génique et effectuer les analyses de séquençage ARN différentielles.

3.1.3 pipeline_info

```
-rw-r--r-- 1 iris formation 3137802  5 oct.  23:47 execution_report.html
-rw-r--r-- 1 iris formation  271331  5 oct.  23:47 execution_timeline.html
-rw------- 1 iris formation  242205  5 oct.  23:47 pipeline_dag.svg
-rw-r--r-- 1 iris formation   12804  5 oct.  23:47 pipeline_report.html
-rw-r--r-- 1 iris formation    2353  5 oct.  23:47 pipeline_report.txt
-rw-r--r-- 1 iris formation     309  5 oct.  23:39 samplesheet.valid.csv
-rw-r--r-- 1 iris formation     235  5 oct.  23:47 software_versions.csv
```

Figure 5: *Les fichiers dans le répertoire* `resultat/genome`

Ce dossier contient des générés dans le contexte de l'exécution d'un pipeline.

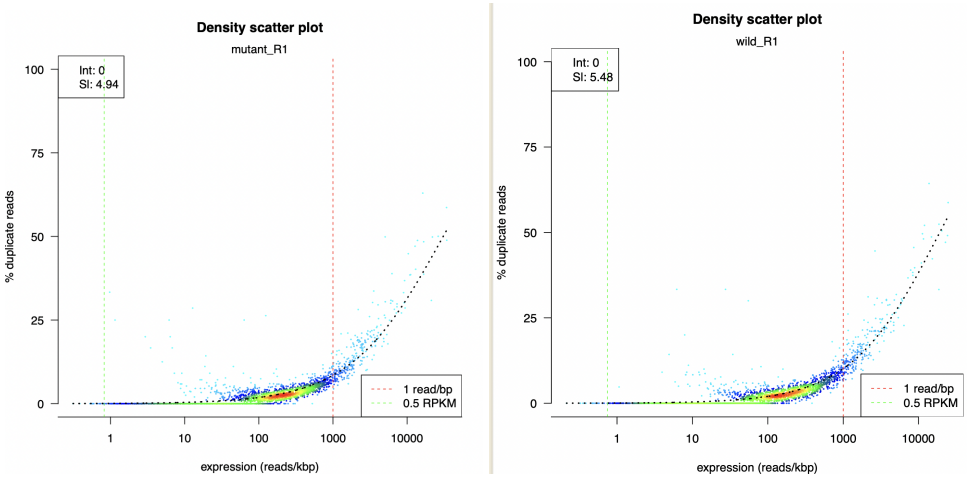
- `execution_report` contient des graphiques sur les mémoires d'utilisation, le temps d'exécution et le déroulement de chaque tache dans le déroulement de pipeline avec son état, son process, le pourcentage de CPU utilisé, etc.

- `execution_timeline` est une représentation graphique de la chronologie de l'exécution d'un pipeline ou d'une série de taches informatique. Il peut montrer quand chaque tache a été démarrée et terminée pour évaluer les performances et la répartition temporelle des calculs.
- `pipeline_dag.svg` représentant un graphe acyclique direct qui représente des dépendances entre les taches dans un pipeline.
- Les fichiers `csv` avec `software_version` avec les versions des logiciels utilisés et samplesheet pour les informations sur les échantillons utilisés.
- Le `pipeline_report` est une version de text de rapport de pipeline

3.2 star_rsem

Dans ce répertoire, on trouve les fichiers bam compressés, les fichiers `sam` et les fichiers générés à partir différents outils utilisés pendant l'analyse de pipeline.

- `bigwig` stocke les données séquençages sous forme de graphiques de densité qui permet voir la couverture de séquençage à travers le génome
- `dupradar` évalue la qualité des données de séquençage en identifiant les duplications



a b

Figure 6: *Ces plots montrent que plus le niveau de duplication est élevé, plus les gènes sont exprimés sur* ***a****) les génomes de mutant et* ***b****) génomes de sauvages. Nos génomes de mutant et de sauvages sont en bonne qualité. Les histogrammes semblent que les reads sont normaux.*

- `featuresCounts` : quantifie l'expression génique en comptant le nombre de lectures de séquençages attribuées à des caractéristiques génomiques spécifiques
- `picard_metrics` analyse des données de séquençage, notamment les fichiers BAM. Ces fichiers contiennent divers métriques et statistiques relatives à des données. Ils sont essentiels pour évaluer la qualité des données, les étapes de traitement
- `preseq`: estime la saturation de la bibliothèque de séquençage
- `qualimap`: décrit les résultats de contrôle de qualité de l'alignement des séquences pour détecter des biais due au séquençage.
- `rseqc` : fournit des métriques de qualité et des analyses de contrôle de qualité pour les données de séquençage. Les fichiers sont utilisés par MultiQC. Il contient les sous-répertoires.
  - `bam_stat` : générer des statistiques à partir de fichiers BAM (les données alignements des séquences sur un génome de référence)
  - `infer_experiment` : déduire l'orientation des données RNA-seq (déterminer si les lectures proviennent du brin sens ou antisens ou les deux)
  - `inner_distance` : estimer la distance moyenne entre les paires de lectures appariées dans les données. Utile pour la fragmentation d'ARN

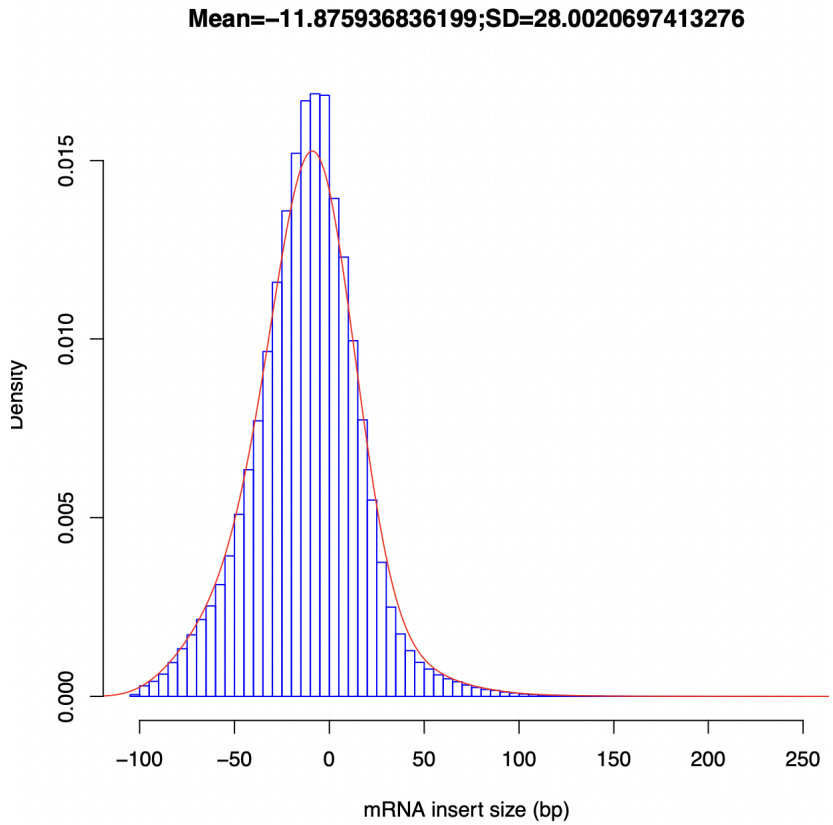


Figure 7: *La distance moyenne entre les lectures appariées pour le sauvage est 11pb*

- `junction_annotation`: annoter des jonctions d'épissage identifiées dans les données en les associant à des gènes ou des transcrits spécifiques
- `junction_saturation`: évaluer la saturation des jonctions d'épissage dans les données de séquençages qui indique si suffisamment de données ont été générées pour détecter toutes les jonctions d'épissage potentielles
- `read_distribution`: répartir les lectures de séquençages sur différentes catégories ou régions génomiques telles que les exons, les introns, les régions intergéniques, etc.

```
Total Reads                     2646455
Total Tags                      3550957
Total Assigned Tags             3402933
=======================================================================
Group               Total_bases         Tag_count           Tags/Kb
CDS_Exons           3502224             3221655             919.89
5'UTR_Exons         0                   0                   0.00
3'UTR_Exons         0                   0                   0.00
Introns             5503508             112410              20.43
TSS_up_1kb          2572830             14764               5.74
TSS_up_5kb          9200819             25032               2.72
TSS_up_10kb         13501456            30778               2.28
TES_down_1kb        2219057             25682               11.57
TES_down_5kb        7691998             34484               4.48
TES_down_10kb       12367516            38090               3.08
=======================================================================
```

- `read_duplication`: détecter et quantifier les lectures dupliquées dans les données de séquençage pour éviter des biais des analyses.

3.2.1 trimegalome

```
drwxr-xr-x 2 iris formation 4096  5 oct.  23:41 fastqc
-rw-r--r-- 1 iris formation 3253  5 oct.  23:41 mutant_R1_1.fastq.gz_trimming_report.txt
-rw-r--r-- 1 iris formation 3471  5 oct.  23:41 mutant_R1_2.fastq.gz_trimming_report.txt
-rw-r--r-- 1 iris formation 3116  5 oct.  23:41 wild_R1_1.fastq.gz_trimming_report.txt
-rw-r--r-- 1 iris formation 3356  5 oct.  23:41 wild_R1_2.fastq.gz_trimming_report.txt
```

Le fichier `trimegalome` est un script enveloppe *wrapper*, ce qui signifie qu'il rassemble plusieurs logiciels (`cutadapt` et `fastq`) pour effectuer une évaluation de la qualité des données. Trimegalome permet de supprimer les lectures de mauvaise qualité (basées sur le score Phred), les adaptateurs, ainsi que les lectures dont la taille est inférieure à un seuil donné. Ce script génère des fichiers au format `fastq` qui sont ensuite soumis à une réanalyse avec *FastQC*. En conséquence, un sous-répertoire nommé `fastqc` est créé pour stocker les résultats de cette réanalyse.

Le rapport de la sortie de `trimegalome` a des paramètres de l'exécution, l'information sur l'exécution de `cutadapt`, résumé des statistiques, les informations spécifiques à l'adaptateur, les séquences supprimées et les statistiques de l'exécution.

3.2.2 multiqc

MultiQC est un outil en bioinformatique pour simplifier et améliorer le processus d'analyse et de visualisation des résultats de contrôle de qualité (QC) provenant des différentes étapes d'analyses de données. Ce outil prend en compte de nombreux outils couramment utilisés en bioinformatique, tels que *FastQC, Picard, Trimmomatic, STAR, HISAT2, etc.* Cet outil permet de consolider les informations de qualité en provenant de multiples logiciels et analyses en un seul rapport convivial et informatif.

Il contient un répertoire `multiqc_data` qui contient les données brutes générées par MultiQC lorsqu'il compile et agrège les rapports de qualité de différentes analyses et un `multiQC_report` en format `html` qui présente les informations de qualité extraites de différentes analyses

3.3 Interprétation du report MultiQC

Le report MultiQC est un rapport de contrôle de qualité consolidé généré par MultiQC au format HTML. Ce rapport présente de manière claire et visuelle les informations de qualité extraites de différentes analyses, telles que des informations sur la qualité des bases, la distribution de la taille des lectures, les métriques d'alignement, etc. Il est souvent utilisé pour évaluer rapidement la qualité globale des données génomiques ou transcriptomiques et pour prendre des décisions sur les étapes suivantes de l'analyse.

3.3.1 General Statistics

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
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%
mutant_R1_1												
mutant_R1_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%
wild_R1_1												
wild_R1_2												

Pour les différents échantillons (`mutant_R1` et `wild_R1`) générées à partir d'une analyse de données de séquençage.

- **M Reads Mapped** : Le nombre de millions de lectures mappées (alignées) à un génome de référence. Ici, on a 3,3 millions de lectures pour mutant et 2,7 millions de lectures pour wild.
- **% rRNA** : Le pourcentage de lectures qui sont identifiées comme rRNA (ARN ribosomique). Dans ce cas, il est de 0,00%, ce qui signifie que les lectures ne sont pas principalement des rRNA.
- **dupInt** : Duplication interne, peut être interprété comme le pourcentage de lectures en double.
- **% Dups** : Le pourcentage de lectures en double (dupliquées). Dans ce cas, il est de 17,3% pour "mutant_R1" et 18,3% pour "wild_R1".
- **5'-3' bias** : Le biais de séquence entre les brins 5' et 3' des lectures.
- **M Aligned** : Le nombre de millions de lectures alignées avec succès.
- **% Alignable** : Le pourcentage de lectures alignées par rapport au total des lectures.
- **% Proper Pairs** : Le pourcentage de paires de lectures alignées de manière appropriée.
- **Error rate** : Le taux d'erreurs dans les lectures.
- **M Non-Primary** : Le nombre de millions de lectures non primaires, c'est-à-dire les lectures qui ne sont pas la lecture principale pour une paire.
- **% Mapped** : Le pourcentage de lectures mappées par rapport au total des lectures. Dans ce cas, il est de 99,3% pour "mutant_R1" et 99,4% pour "wild_R1".
- **% Proper Pairs (suite)** : Le pourcentage de paires de lectures alignées de manière appropriée.
- **M Total seqs** : Le nombre de millions de séquences totales.
- **% Dups (suite)** : Le pourcentage de séquences dupliquées.
- **% GC** : Le pourcentage de bases G-C dans les séquences. Dans ce cas, il est de 48,2% pour `mutant_R1` et 47,2% pour `wild_R1`.
- **% BP Trimmed** : Le pourcentage de bases coupées (trimmées) dans les séquences.
- **% Dups (suite)** : Le pourcentage de séquences dupliquées.
- **% GC (suite)** : Le pourcentage de bases G-C dans les séquences.

En outre, on permet de choisir les paramètres différents :

General Statistics: Columns

Jncheck the tick box to hide columns. Click and drag the handle on the left to change order.

Show All | Show None

Sort	Visible	Group	Column	Description	ID	Scale
‖	☐	Samtools	M Reads	Total reads in the bam file (millions)	flagstat_total	read_count
‖	☑	Samtools	M Reads Mapped	Reads Mapped in the bam file (millions)	mapped_passed	read_count
‖	☑	Biotype Counts	% rRNA	% reads overlapping rRNA features	percent_rRNA	None
‖	☑	DupRadar	dupInt	Intercept value from DupRadar	dupRadar_intercept	None
‖	☑	Picard	% Dups	Mark Duplicates - Percent Duplication	PERCENT_DUPLICATION	None
‖	☑	QualiMap	5'-3' bias	5'-3' bias	5_3_bias	None
‖	☑	QualiMap	M Aligned	Reads Aligned (millions)	reads_aligned	read_count
‖	☑	Rsem	% Alignable	% Alignable reads	alignable_percent	None
‖	☑	RSeQC	% Proper Pairs	% Reads mapped in proper pairs	proper_pairs_percent	None
‖	☑	Samtools	Error rate	Error rate: mismatches (NM) / bases mapped (CIGAR)	error_rate	None
‖	☑	Samtools	M Non-Primary	Non-primary alignments (millions)	non-primary_alignments	read_count
‖	☑	Samtools	M Reads Mapped	Reads Mapped in the bam file (millions)	reads_mapped	read_count
‖	☑	Samtools	% Mapped	% Mapped Reads	reads_mapped_percent	None
‖	☑	Samtools	% Proper	% Properly Paired Reads	reads_properly_paired_percent	None

3.3.2 Biotype Counts

Comptage des biotypes est une métrique qui indique combien de lectures de séquençage chevauchent sur le génome. Ces données sont générées à l'aide de l'outil « featureCounts » ou un outil similaire.

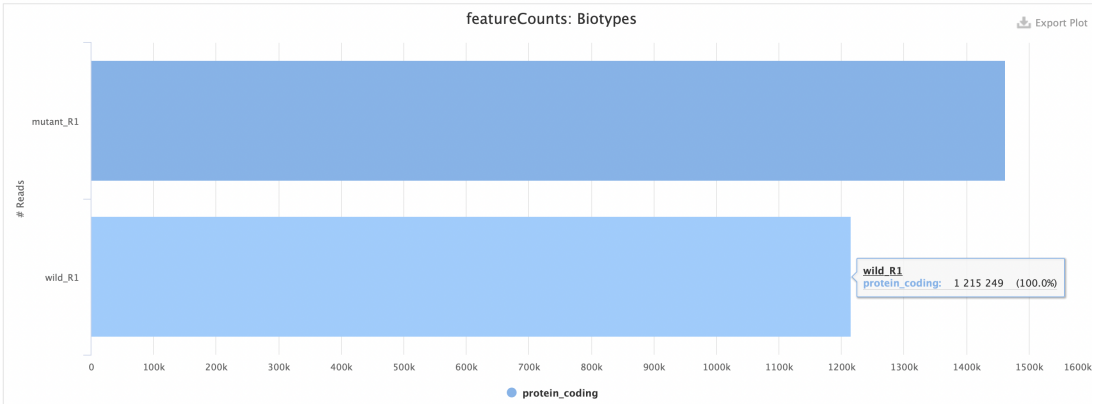


Figure 8: *Dans ce cas, toutes nos séquences s'alignent. On a 1 462 722 protéine_coding pour le mutant (mutant_R1) et 1 215 249 protéine_coding pour le sauvage (wild_R1)*

3.3.3 DupRadar

DupRadar est un outil qui permet à effectuer un contrôle de qualité concernant le taux de duplication dans l'ensemble de données de séquençage d'ARN. Objectif est d'évaluer la qualité des données en identifiant les lectures dupliquées, c'est-à-dire les lectures qui sont des copies identiques les unes des autres.

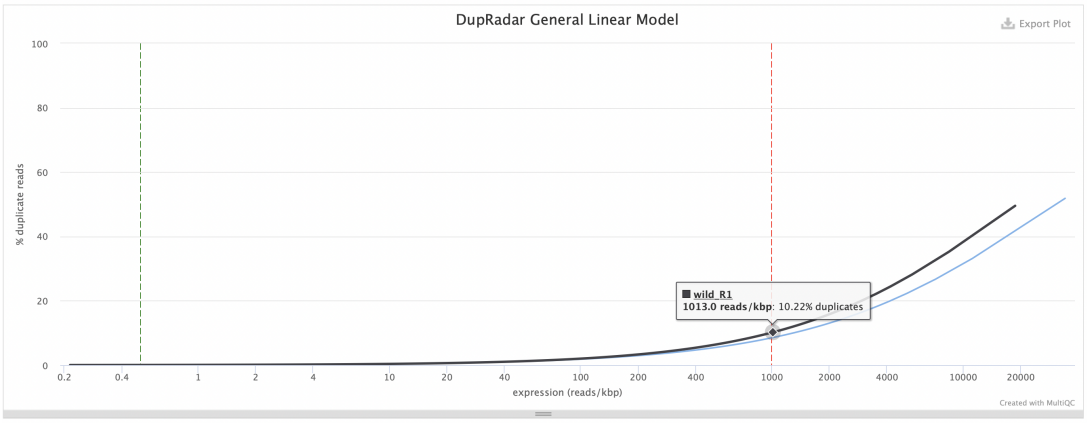


Figure 9: *Ce qu'on attend est plus la lecture du gène est grande, plus le nombre de duplications attendu est grand. C'est logique pour notre cas. Pour les 2 échantillons mutants et sauvages, le niveau de duplication est faible pour les gènes faiblement exprimés et à l'inverse.*

3.3.4 Picard - Mark Duplicates

Picard est un outil en utilisant Java pour manipulation de données de séquençage. Il effectue diverses tâches comme la détection et le marquage des lectures dupliquées, etc. liées à la gestion de à l'analyse des données de séquençages. L'outil Mark Duplicates de Picard identifie les lectures qui sont des duplicatas exacts les unes des autres, marque ces lectures et fournit des informations sur le nombre de lectures marquées comme dupliquées pour évaluer la qualité des données et les performances des étapes de séquençage.

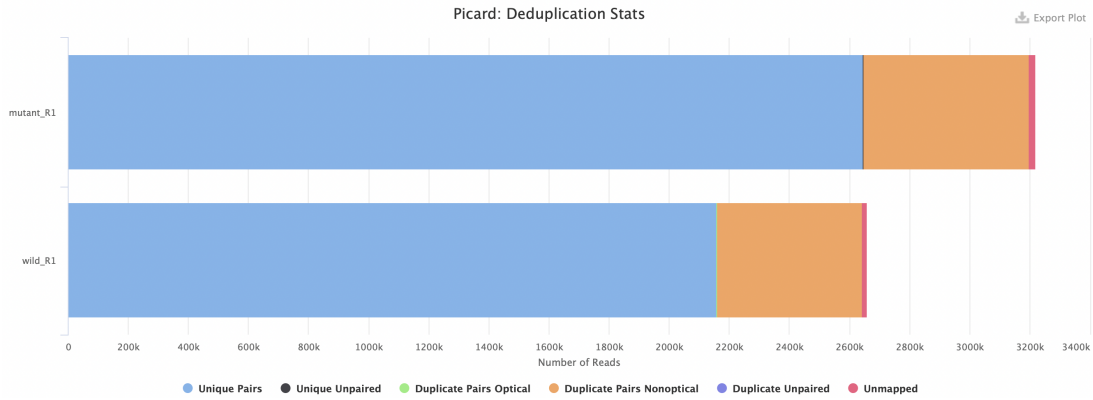


Figure 10: *Pour mutant_R1 et wild_R1, on a un grand pourcentage (>80%, partie bleu) des paires de lectures dans l'échantillon sont uniques (elles ne sont pas des duplicatas exacts d'autres paires de lectures dans l'échantillon), 0% de lectures uniques non appariées (c'est-à-dire toutes les lectures sont soit en paires, soit dupliquées), 0% de paires dupliquées optiques, plus de 10% des paires dupliquées non optiques (partie orange, ces duplicatas sont des copies exactes d'autres paires de lectures dans l'échantillon), 0% de lectures dupliquées non appariées et très peu de pourcentage de lectures non mappées (vers 0,6-0,7%,partie rose : ces lectures dans l'échantillon ne sont pas mappées sur le génome de référence).*

3.3.5 Preseq – Complexity curve

`Preseq` évalue combien de nouvelles lectures uniques sont séquencées à mesure que le nombre total de lectures augmente.

- **Axe des x (abscisse)** : représente le nombre total de lectures séquencées
- **Axe des y (ordonnée)** : représente le nombre de lectures uniques détectées à mesure que le nombre total de lectures augmente.

- **Ligne en pointillés** : représente l'idéal d'une bibliothèque parfaitement complexe, où le nombre total de lectures est égal au nombre de lectures uniques.
- **Courbe de complexité** : montre comment le nombre de lectures uniques évolue à mesure que le nombre total de lectures augmente.
- **Trimming de l'axe des x** : Il est important de noter que l'axe des x est parfois tronqué (coupé) lorsque toutes les courbes atteignent 80% de leur valeur maximale sur l'axe des y.

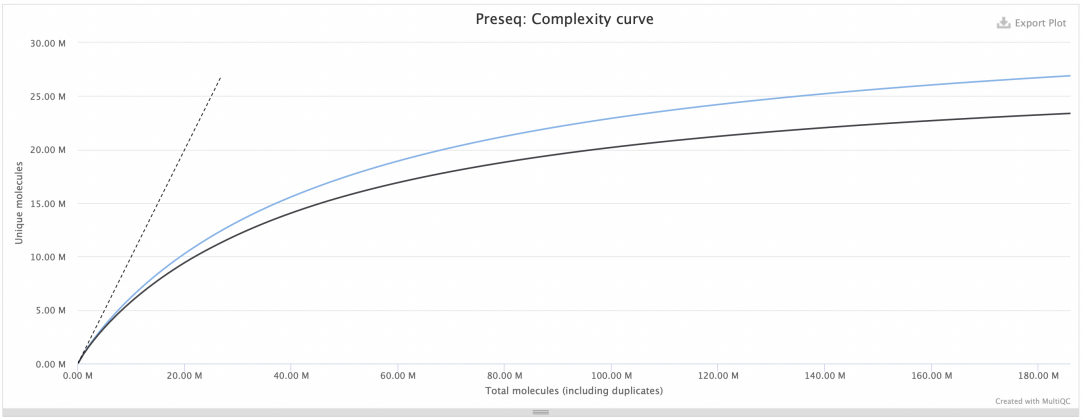


Figure 11: *Au début, lorsque le nombre total de lectures est faible, la courbe augmente rapidement, indiquant que de nombreuses nouvelles lectures uniques sont découvertes. Cependant, à mesure que le nombre total de lectures augmente, la courbe commence à s'aplatir, ce qui signifie que de moins en moins de nouvelles lectures uniques sont découvertes à mesure que la bibliothèque devient plus saturée en lectures dupliquées.*

3.3.6 Qualimap

`Qualimap` est une application qui facilite le contrôle de qualité des données de séquençage d'alignement et de ses dérivés.

Genomic origin of reads
L'une des fonctionnalités de QualiMap est d'identifier l'origine génomique des lectures mappées dans les données d'alignement. Cette fonction classe les lectures qui ont été correctement alignées sur le génome en fonction de leur emplacement par rapport aux régions génomiques.

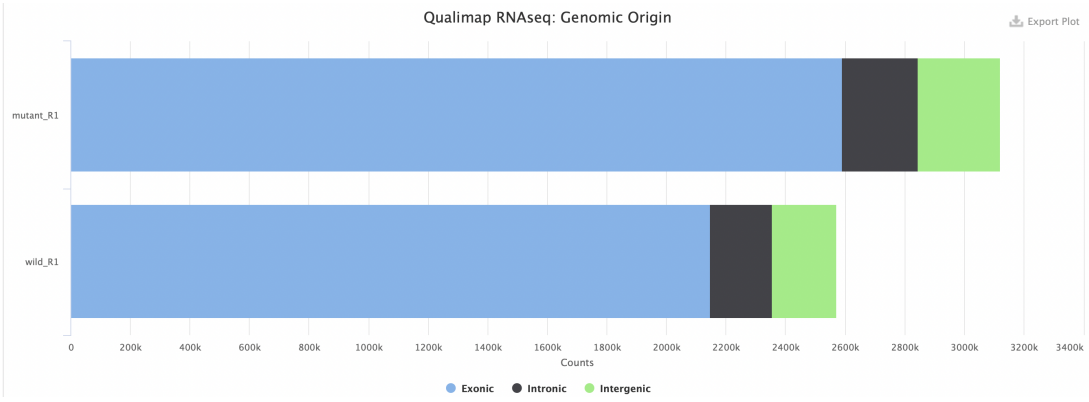


Figure 12: *La plupart des lectures de mutant_R1 et wild_R1 ont correctement aligné sur des régions exoniques (exonic) du génome. Il y a peu de lectures qui ont été alignées sur des régions introniques et sur des régions intergéniques (ni exons ni introns) du génome.*

Gene coverage profile
Ce profil est principalement utilisé pour évaluer la profondeur de séquençage et la couverture des gènes. Lorsque vous n'observez pas de biais dans le profil de couverture des gènes, vous vous attendez généralement à voir une couverture élevée

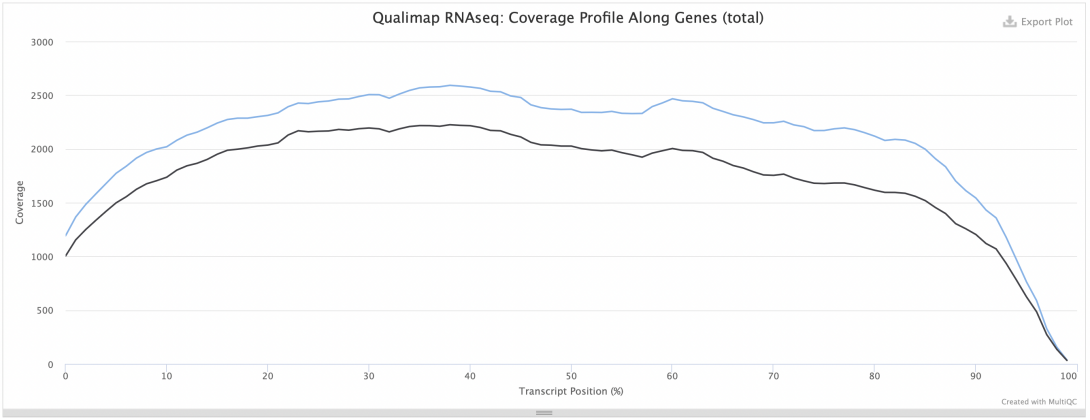


Figure 13: *La couverture élevée au milieu des séquences géniques indique que de nombreuses lectures sont alignées sur les parties centrales des gènes. Cela suggère une grande quantité de séquences géniques bien couvertes. La couverture plus faible aux extrémités 5′ (début) et 3′ (fin) des séquences géniques peut être le résultat de plusieurs facteurs.*

3.3.7 Rsem

`Rsem` (RNA-Seq by Expectation-Maximization"), est un ensemble de logiciels utilisé pour estimer les niveaux d'expression des gènes et des isoformes à partir de données de séquençage d'ARN (RNA-Seq). Il est largement utilisé en bioinformatique pour quantifier l'expression génique et isoformique à partir de données de séquençage.

Mapped Reads : Une répartition détaillée de la manière dont toutes les lectures (ou reads) ont été alignées pour chaque échantillon. Elle peut fournir des informations importantes sur la qualité de l'alignement des lectures sur le génome de référence ou le transcriptome.

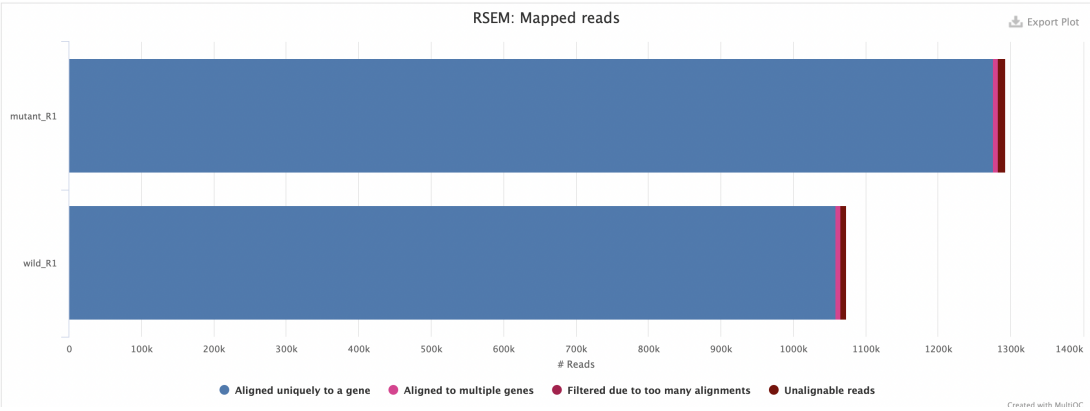


Figure 14: *Pour l'échantillon mutant_R1 et wild_R1, la plupart des lectures de ces échantillons ont été alignées de manière unique sur un seul gène (>98%). Cela signifie que ces lectures correspondent clairement et exclusivement à un gène spécifique du génome de référence. Il y a très peu de lectures ont été alignées sur plusieurs gènes différents (<1%) ou n'ont pas pu être alignées (<1%) pour les deux échantillons et aucun de lectures ont été filtrés en raison du nombre excessif d'alignements.*

Multimapping rates : Une mesure qui indique à quelle fréquence les lectures de séquençage ont été alignées sur plusieurs régions de référence. Elle est importante pour évaluer la spécificité de l'alignement des lectures et peut fournir des informations sur la complexité des échantillons de séquençage.

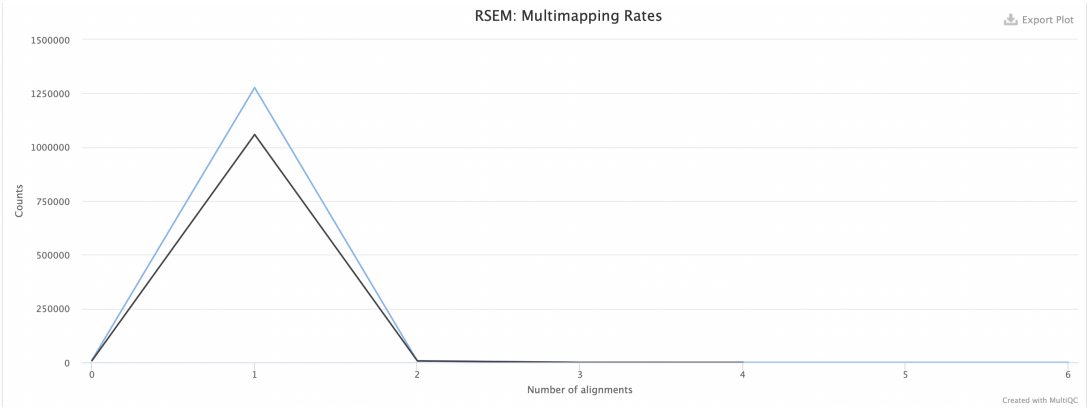


Figure 15: *La majorité des séquences s'alignent une fois sur un génome de référence. Une spécificité élevée de l'alignement des lectures.*

3.3.8 RSeQC

`RSeQC` est un ensemble d'outils qui permet d'évaluer de manière approfondie les données de séquençage à haut débit de type RNA-seq.

Read Distribution : Ce module calcule comment les lectures mappées sont distribuées sur les éléments du génome ou les régions spécifiques du génome. Il permet d'évaluer la répartition des lectures le long des caractéristiques génomiques telles que les gènes, les exons, les introns, les régions intergéniques,

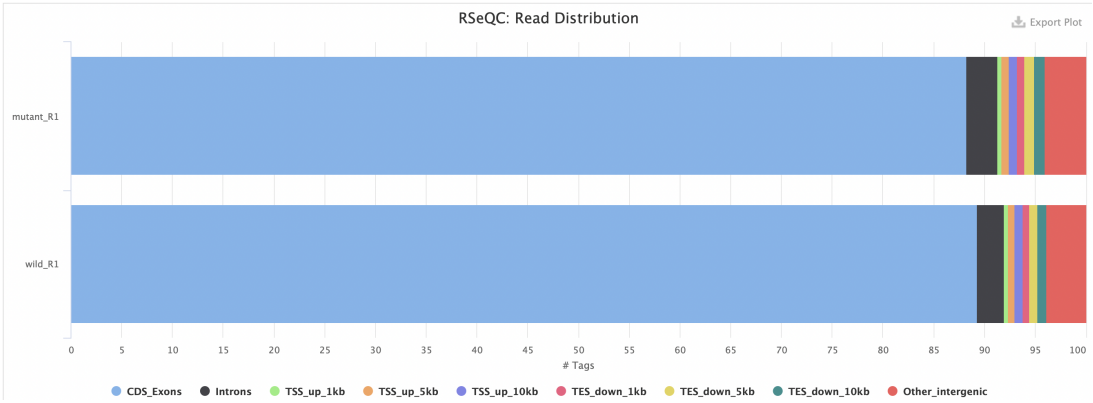


Figure 16: *La distribution des lectures mappées du mutant_R1 et wild_R1 semblent identique avec la majorité des régions qui correspondent à exons. Dans RNA-seq, ça semble un bon résultat.*

Inner Distance : permet de calculer la distance intérieure (ou taille d'insertion) entre deux lectures appariées (paired-end) provenant de séquençage RNA.

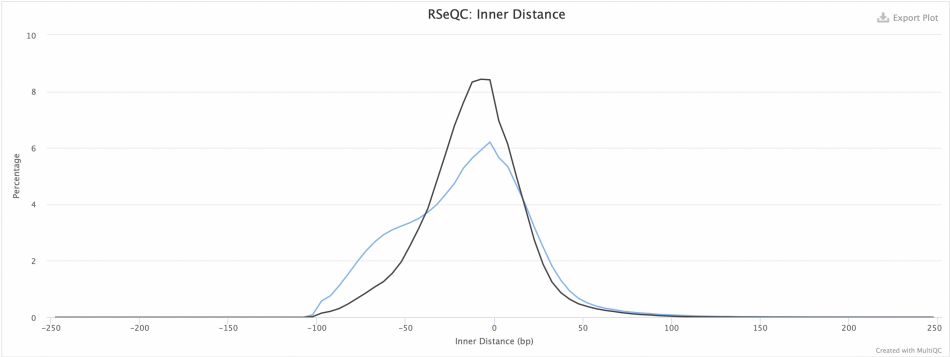


Figure 17: *La figure montre que la distance entre les lectures contiguës est faible, cela signifie généralement que la taille d'insertion (distance intérieure) entre les paires de lectures est courte.En d'autres termes, les lectures appariées sont proches les unes des autres sur le brin d'ARN ou d'ADN, et il y a peu d'espace entre elles.*

Read Duplication : calcule combien de positions d'alignement ont un certain nombre de duplicatas exacts dans les données de séquençage. Cela signifie qu'il évalue la fréquence à laquelle des positions spécifiques dans le génome ou le transcriptome ont des lectures identiques qui sont alignées à cet endroit.

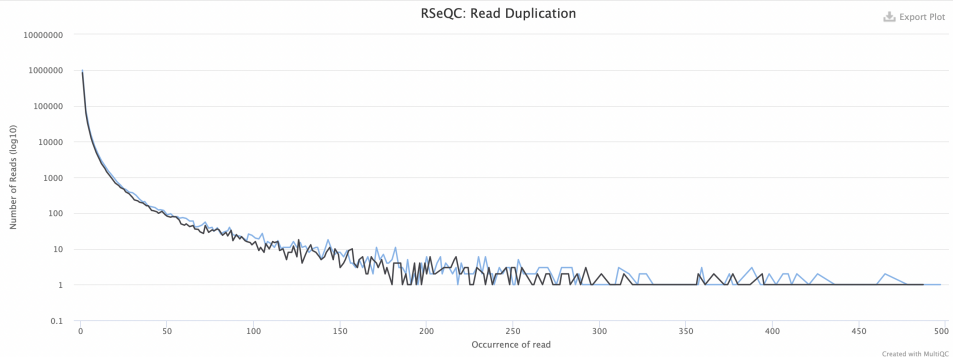


Figure 18: *L'analyse de duplication de lectures illustre la quantité de lectures (axe des ordonnées) en relation avec leur fréquence d'apparition respective. On observe que nos échantillons présentent quelques duplications, mais elles demeurent dans des limites acceptables. Une importante surface sous la courbe pourrait suggérer que les échantillons contiennent un grand nombre de lectures avec de fréquentes duplications.*

Junction Annotation : permet de comparer les jonctions d'épissage détectées dans les données de séquençage à un modèle de gènes de référence.

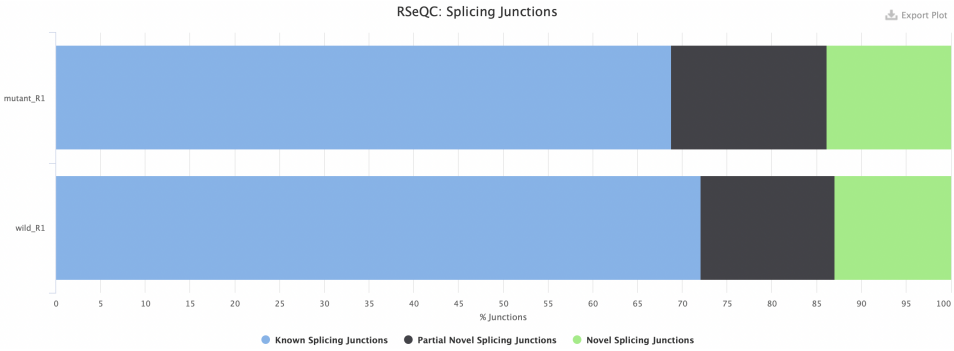


Figure 19: *La plupart des jonctions d'épissage détectées correspondent à des épissages qui sont déjà répertoriés et connus dans un modèle de gènes de référence. Il y a une partie de jonctions d'épissage sont partiellement nouvelles et sont complètement nouvelles sur les deux échantillons sauvages et mutants. Cependant, l'échantillon de mutant contient plus de nouveaux sites d'épissages que l'échantillon sauvage.*

Junction Saturation : évalue le nombre de jonctions d'épissage connues qui sont observées dans chaque ensemble de données de séquençage RNA-seq.

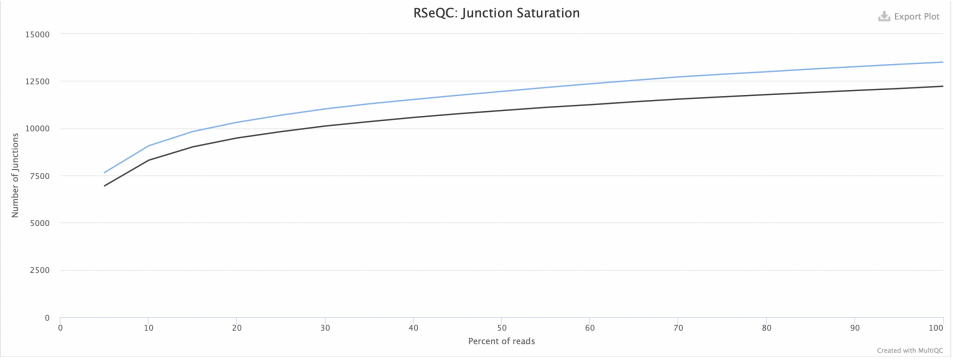


Figure 20: *Plus la courbe est plate, plus c'est mieux le résultat. L'idée est que si la profondeur de séquençage est suffisante, la courbe atteindra un plateau où toutes les jonctions d'épissage connues auront été redécouvertes. C'est logique avec nos échantillons.*

Infer experiment : déterminer la proportion de lectures et de paires de lectures qui correspondent à la stratégie d'orientation (strandeness) des transcrits qui se chevauchent.

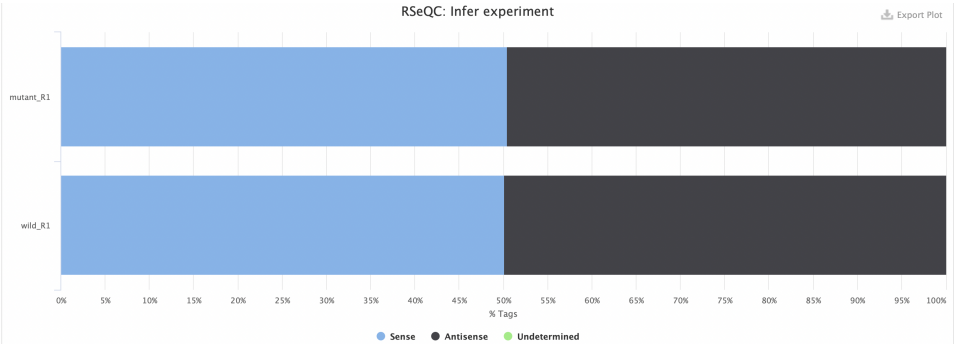


Figure 21: *Pour toutes les 2 échantillons, la stratégie d'orientation des lectures est répartie de manière égale entre le sens et l'antisens des transcrits, avec aucune lecture dont l'orientation n'est pas déterminée. Cela suggère que les données RNA-seq de cet échantillon sont orientées (stranded) et que l'information sur l'orientation des transcrits est préservé.*

Bam Stat : est un script qui fournit les statistiques d'alignement des séquences (reads) à partir du fichier BAM.

3.3.9 Samtools

`Samtools` est une suite d'outils puissants conçus pour interagir avec les données résultant du séquençage RNA-seq, offrant une gamme de fonctionnalités essentielles pour l'analyse de ces données.

Percent Mapped : révèle que la grande majorité des séquences générées au cours du séquençage se sont correctement alignées sur le génome de référence, ce qui est un indicateur crucial de la qualité des données et de la précision de l'alignement.

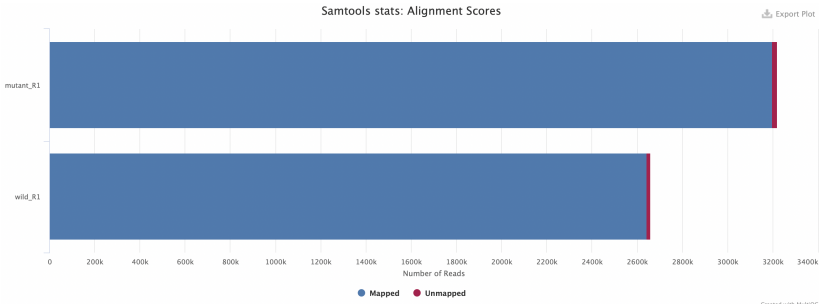


Figure 22: *Sur les 2 échantillons, la majorité des séquences s'alignent sur le génome de référence* (>99%).

Alignement Metrics : permet d'évaluer divers aspects de l'alignement des lectures.

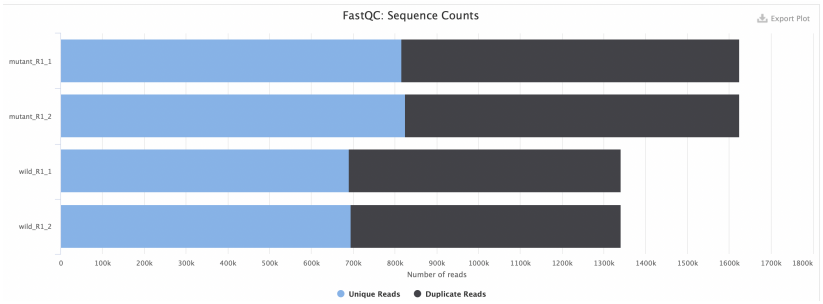
Samtools Flagstat : compte le nombre d'alignements pour chaque catégorie de drapeau (flag). Cette métrique permet de vérifier que l'alignement des lectures est plausible et de détecter d'éventuelles incohérences ou problèmes.

Mapped Reads per Contig : utilise l'outil "Idxstats" pour compter le nombre de lectures qui s'alignent sur chaque chromosome et pour chaque contig. Cette métrique fournit des informations précises sur la distribution des lectures alignées dans le génome, ce qui peut être essentiel pour comprendre la répartition spatiale des informations génétiques.

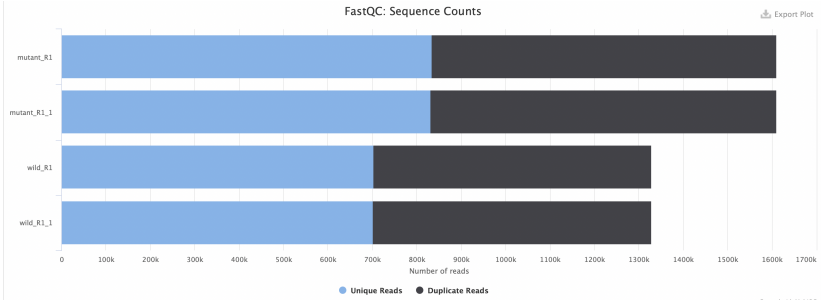
3.3.10 FastQC

- FastQC (**Avant** `cutadapt`) : un rapport généré par FastQC qui présente les résultats de FastQC avant tout processus de suppression d'adaptateurs ou de traitement des données brutes. Cette section fournit des informations de base sur les comptages de séquences pour chaque échantillon dans leur état brut, avant tout nettoyage ou manipulation des données.
- FastQC (**Après** `cutadapt`) qui présente les résultats de FastQC après le processus de suppression d'adaptateurs ou d'autres étapes de prétraitement des données

Sequence Counts : fournit le nombre de séquences (ou lectures) qui ont été générées pour chaque échantillon. Il s'agit du décompte brut des lectures sans tenir compte des lectures en double



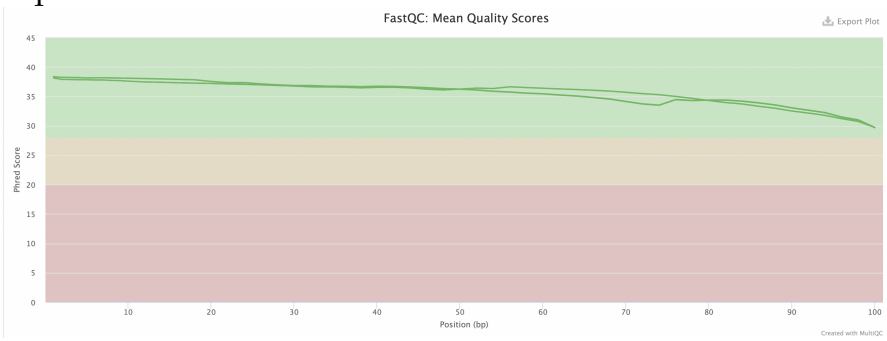
Avant `cutadapt`



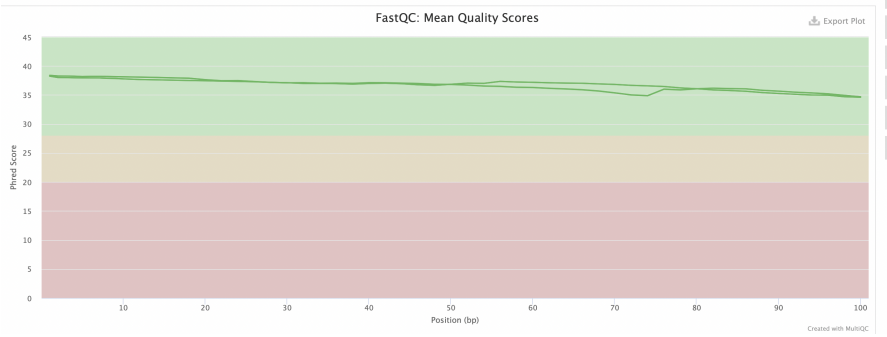
Après `cutadapt`

Figure 23: *Les résultats avant et après suppression des adaptateurs semblent identiques. Dans tous les échantillons mutants ou sauvages avant ou après traitement cutadapt, la majorité des lectures sont uniques (>50%) tandis que <50% sont des lectures dupliquées. Il y a presque autant de lectures uniques que de doublons donc la profondeur de séquençage ne semble pas très grande*

Sequence Quality Histograms : Ces histogrammes fournissent une visualisation de la qualité des bases de séquences dans un échantillon de données de séquençage avec le score phred de qualité parmi toutes les séquences



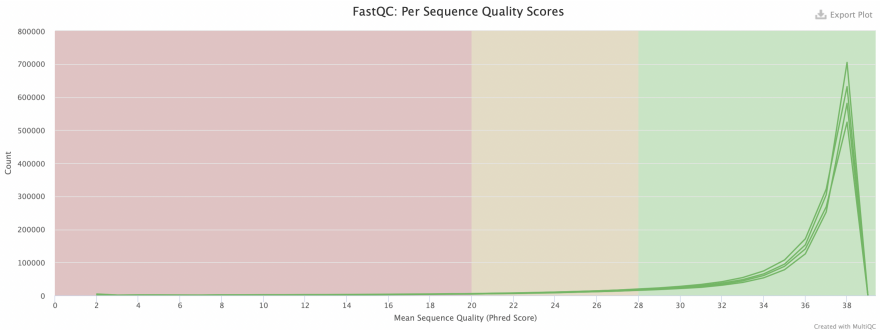
Avant `cutadapt`



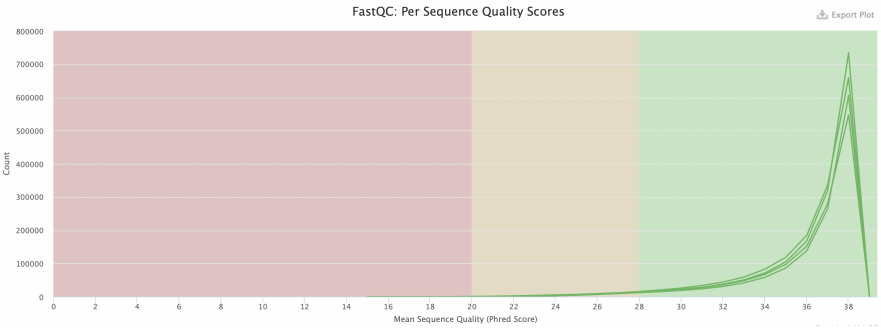
Après `cutadapt`

Figure 24: *Dans toutes les 2 cas, la courbe se trouve sur la partie verte signifie que les scores sont de bonnes qualités. Il n'y a pas trop de différents avant et après le traitement des adaptateurs. Il semble avoir une augmentation légère après suppressions de l'adaptateur*

Per Sequence Quality Scores : Cette métrique compte le nombre de reads en function du score moyen de qualité.



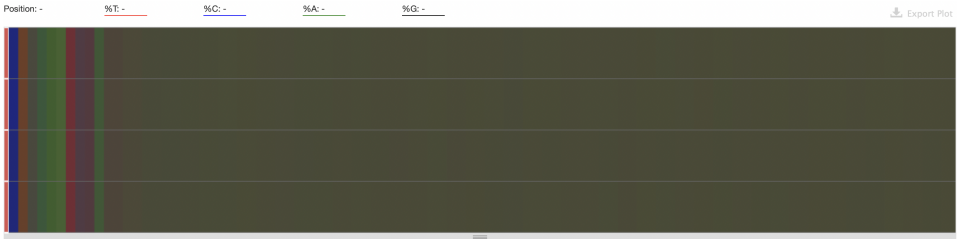
Avant cutadapt



Après cutadapt

Figure 25: *Les courbes avant et après suppression d'adaptateur sont trouvés sur le fond vert, cela signifie que la majorité de séquences ont un score élevé. Pas de changement significatif sur les données brutes et les données après le traitement cutadapt*

Per Base Sequence Content : Cette métrique sert à évaluer la distribution des bases nucléotidiques (A, C, G, T) à chaque position le long de toutes les séquences d'un échantillon de données de séquençage. Cette métrique permet de détecter d'éventuels problèmes ou biais dans les données de séquençage.



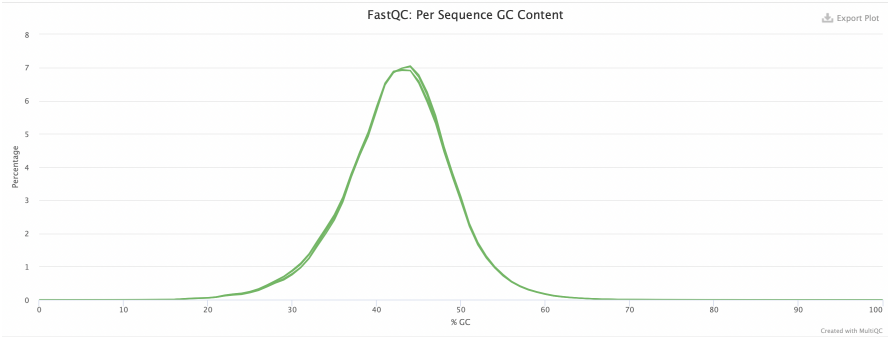
Avant cutadapt

Après cutadapt

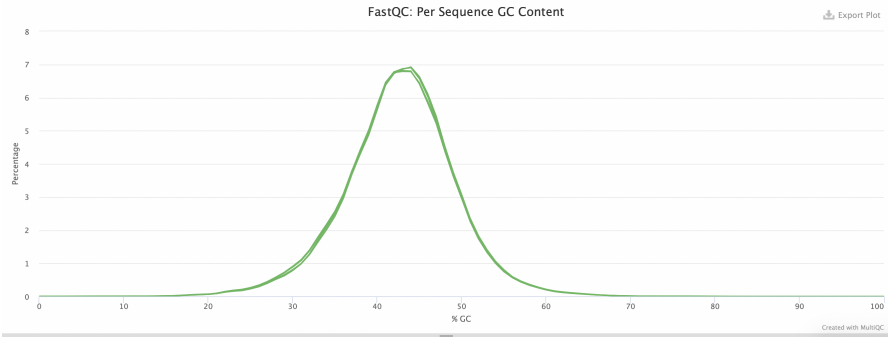
Figure 26: *Les résultats montrés sont des heatmaps. Dans toutes les 2 cas, il y a un déséquilibre au début de la séquence peut être causé à cause du biais dans la partie 5' et pas une contamination, car on ne trouve pas la contamination dans notre séquence. De plus, après la suppression des adaptateurs, on trouve un biais au niveau de la partie 3' en fin de la séquence. C'est lié au Trimegalore sur l'endroit ou on a coupé les adaptateurs et créé un déséquilibre sur les bases*

Per Sequence GC Content : évalue la répartition du contenu en GC (cytosine-guanine) dans chaque séquence individuelle d'un échantillon de données de séquençage. Cette métrique permet

d'analyser la variabilité du contenu en GC des séquences et de détecter d'éventuels problèmes ou biais.



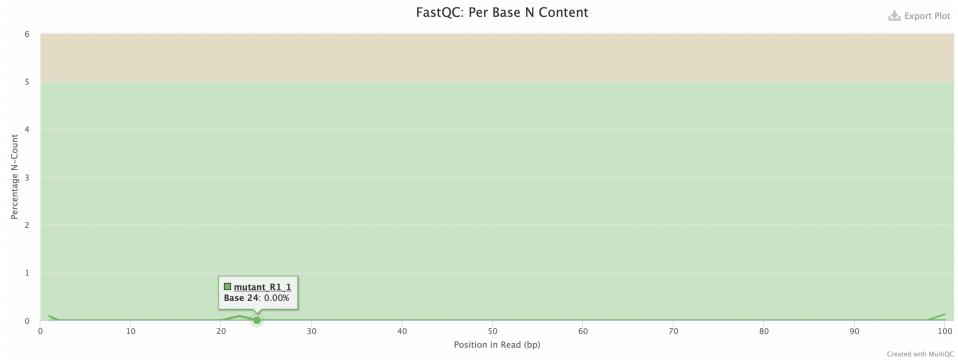
Avant `cutadapt`



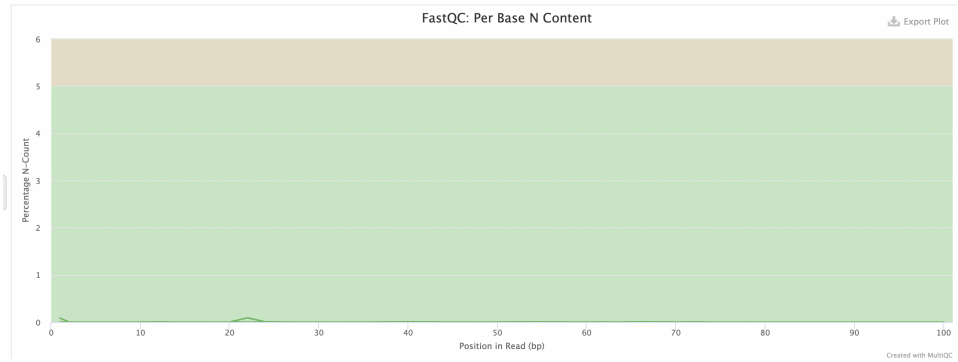
Après `cutadapt`

Figure 27: *La répartition du contenu en GC semble normale avant et après de traitement d'adaptateurs. On ne remarque pas une différence significative entre les deux cas.*

Per Base N Content : évalue la distribution des bases N (qui représentent des positions inconnues ou ambiguës dans la séquence) à chaque position le long de toutes les séquences d'un échantillon de données de séquençage. Cette métrique permet de détecter d'éventuels problèmes liés aux bases ambiguës ou manquantes dans les données.



Avant `cutadapt`



Après `cutadapt`

Figure 28: *Sur nos échantillons dans les 2 cas, le pourcentage de bases N est très faible. Cela indique que les données de séquençage sont de bonne qualité en ce qui concerne la présence de bases ambiguës ou manquantes (bases N). C'est généralement un signe positif qui suggère que la majorité des séquences ont été séquencées avec succès et que les positions ambiguës sont rares ou négligeables.*

Sequence Length Distribution : Indique la répartition des longueurs des séquences dans un échantillon de données de séquençage.

Avant `cutadapt` : Le résultat indique que toutes les séquences de votre échantillon ont une longueur de 101 bases (101bp).

Cela signifie que chaque séquence de votre échantillon est de la même longueur, ce qui est un résultat attendu dans de nombreuses expériences de séquençage, en particulier lorsqu'une stratégie de séquençage spécifique est utilisée pour générer des lectures de longueur uniforme.

Après `cutadapt` :

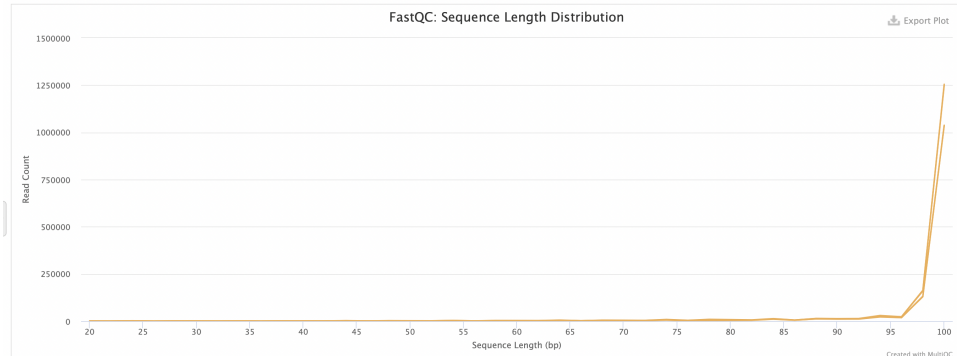
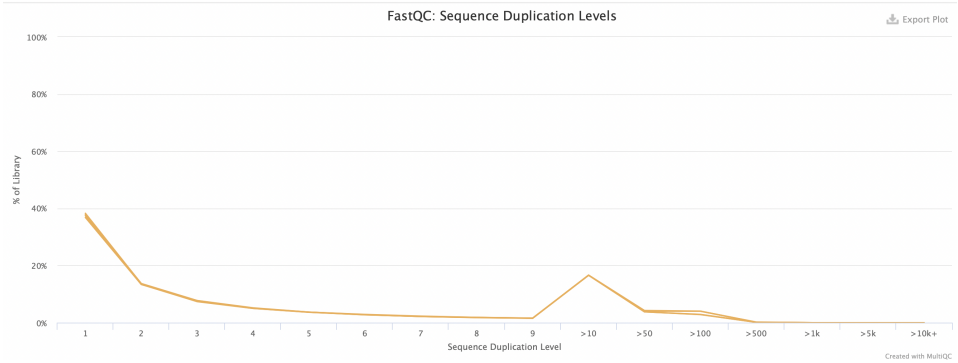
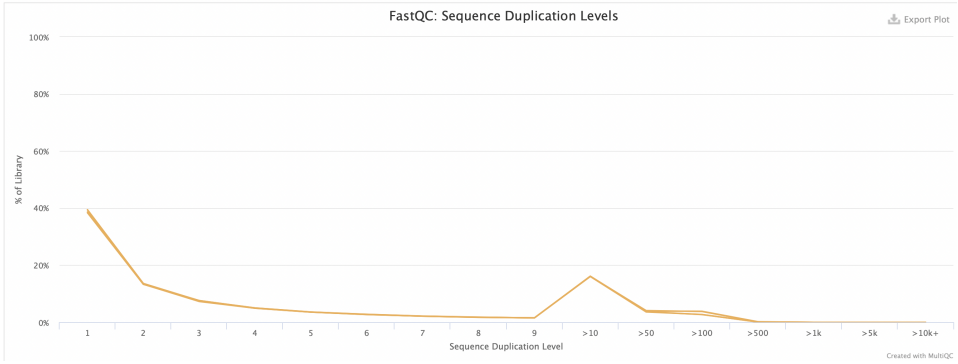


Figure 29: *Après suppression de l'adaptation, des séquences ont des tailles différents.*

Sequence Duplication Levels : évalue le degré de duplication des séquences dans un échantillon de données de séquençage. Cette métrique permet de détecter la présence de séquences dupliquées ou répétées et d'évaluer leur fréquence dans les données.



Avant `cutadapt`



Après `cutadapt`

Figure 30: *Il n'y a pas de différence significative entre les données avant et après réduction des adaptateurs. La majorité des séquences a été dupliquée 1 fois ou 2 fois, cela signifie qu'une distribution idéale avec peu ou pas de duplications. Le pourcentage de duplication globale est faible. Il y a un pic de duplication dû à duplication optique, mais il reste faible.*

Overrepresented sequences : évalue la présence de séquences qui sont fortement représentées dans l'échantillon de données de séquençage. Ces séquences surexprimées peuvent indiquer des contaminants ou des artefacts dans les données de séquençage.

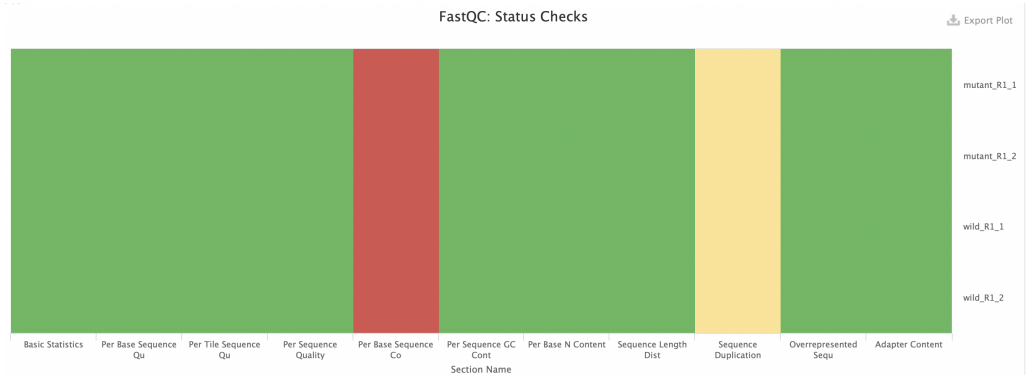
Avant `cutadapt` + *Après* `cutadapt` : 4 échantillons ont moins de 1% de leurs lectures composées de séquences surexprimées). Cela signifie que, pour ces échantillons, la proportion de séquences surexprimées par rapport à l'ensemble des séquences est inférieure à 1%. Cela indique que la majorité des séquences dans l'échantillon ne sont pas dominées par des séquences surexprimées. Cela suggère que les échantillons sont bien préparés et ne contiennent pas de contaminants.

Adapter Content : évalue la présence d'adaptateurs d'ADN ou d'ARN dans les données de séquençage. Les adaptateurs sont des séquences courtes ajoutées aux extrémités des fragments d'ADN ou d'ARN lors de la préparation de la bibliothèque pour le séquençage. La présence d'adaptateurs indique généralement que les fragments n'ont pas été correctement purifiés avant le séquençage.

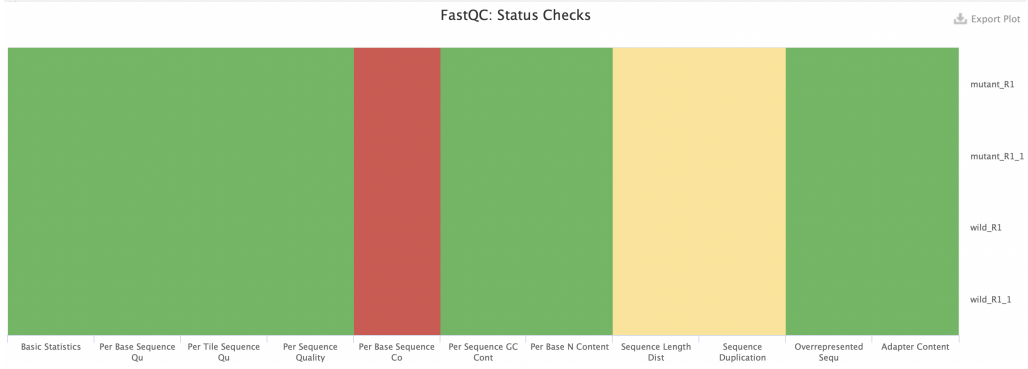
Avant `cutadapt` + *Après* `cutadapt`

Aucun échantillon n'a montré une contamination par des adaptateurs supérieure à 0,1%. Cela ne signifie qu'aucun de vos échantillons n'a montré de contamination significative par des adaptateurs. Un seuil de 0,1% est généralement considéré comme faible et acceptable. Cela suggère que les données de séquençage ont été bien préparées, avec une élimination efficace des adaptateurs ce qui garantit la qualité des données et des résultats analysées

Status Checks : Ce sont les résultats des vérifications de l'état global des données de séquençage. Ces vérifications incluent généralement plusieurs indicateurs et métriques qui permettent d'évaluer la qualité globale des données.



Avant `cutadapt`



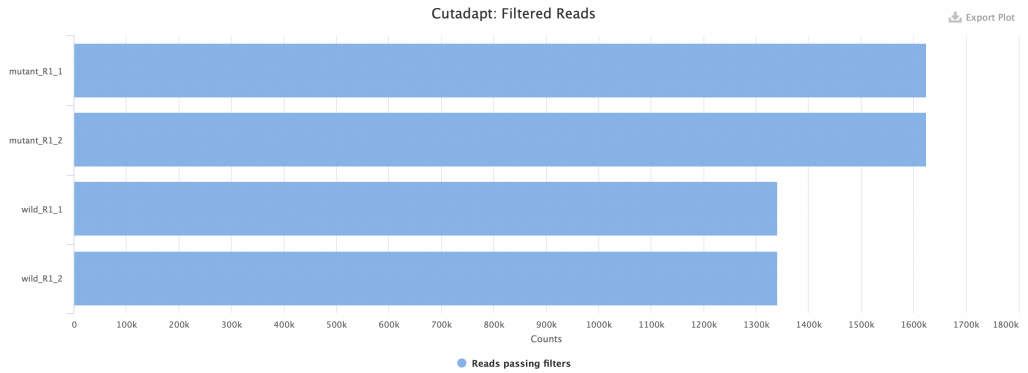
Après `cutadapt`

Figure 31: *La plupart des parties d'analyses sont en verts. Cela signifie que les données sont conformes aux normes de qualités attendues. Cependant, la distribution en bases (Per Base Sequence Content) est marquée en rouge. Cela peut être dû à des variations inattendues dans la composition des bases le long des séquences. Le niveau de duplication (Sequence Duplication) est marqué en jaune, ce qui signifie qu'il est moyen mais acceptable. Cela peut indiquer que certaines séquences sont dupliquées, mais que cela reste dans une plage acceptable.*
Après le trimming des adaptateurs, vous mentionnez qu'il y a aussi une distribution de la taille des lectures qui n'est pas très homogène

3.3.11 Cutadapt

détecte et supprimer des séquences d'adaptateurs, des amorces, des queues poly-A et d'autres types de séquences indésirables de vos lectures de séquençage à haut débit.

Filtered Reads : Il montre le nombre de lectures (ou paires de lectures) qui ont été supprimées de vos données par le biais de l'application de Cutadapt sur ces échantillons. Cette métrique vous indique combien de lectures ont été retirées de vos données en raison de la détection d'adaptateurs, d'amorces, de queues poly-A ou d'autres séquences indésirables par Cutadapt.



Trimmed Sequence Lengths : Il montre le nombre de reads en fonction de la taille des adaptateurs enlevés.

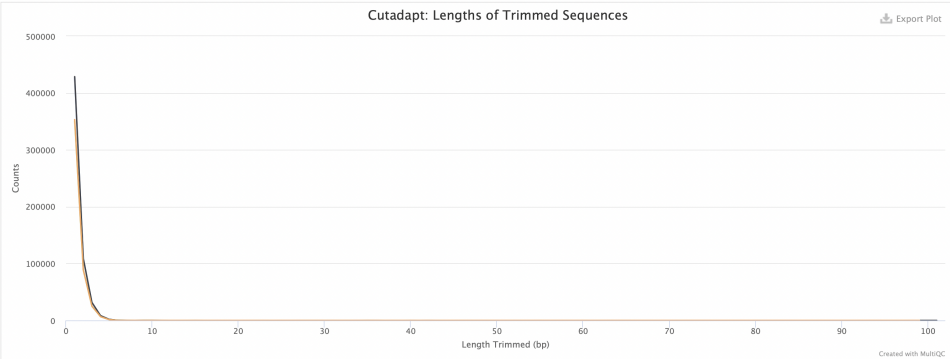


Figure 32: *Le graphique montre que la plupart des adaptateurs (300000 reads pour le mutant et 40000 reads pour le sauvage) enlevées sont très petites (vers 1bp).*

4 Exercice 3 : Lancer ce pipeline sur des données NCBI

4.1 Création du répertoire de travail et téléchargement des fichiers analysés

Pour cette partie, on a choisi les 3 échantillons de Gadus morhua (Morue franche) qui est une espèce de poissons à nageoires rayonnées de la famille des morues. Ils sont originaires de l'océan Atlantique. Ce sont des carnivores.

Le numéro accession sur ce projet d'analyse transcriptomique est PRJNA256972 avec les 3 échantillons :

- SRR2045415 : Cod ovary (1 ILLUMINA (Illumina HiSeq 2000) run: 22.4M spots, 4.5G bases, 2.8Go downloads)
- SRR2045416 : Cod brain (1 ILLUMINA (Illumina HiSeq 2000) run: 36.5M spots, 7.3G bases, 4.7Go downloads)
- SRR2045417 : Cod gills (1 ILLUMINA (Illumina HiSeq 2000) run: 35.5M spots, 7.1G bases, 4.6Go downloads)

Ces échantillons sont téléchargés sur les données du site NCBI. Comme ce sont les grands échantillons (>5Go), il faut télécharger les fichiers `fastq` par **SraToolKit**:

1. D'abord, il faut installer `sra-toolkit` sur le système de mon ordinateur local avec la commande :

   ```
   $ sudo apt install sra-toolkit
   ```

2. Ensuite, on utilise fastq-dump pour télécharger le fichier FASTQ

   ```
   $ fastq-dump --outdir /chemin/sortie --split-files SRR2045415 SRR2045416 SRR2045417
   ```

3. Après cette commande, il télécharge les 6 fichiers différents pour les 3 numéros accession SRA qui correspondent au SRR2045415_1, SRR2045415_2 , SRR2045416_1, SRR2045416_2, SRR2045417_1, SRR2045417_2
4. Après réussir à télécharger toutes les 6 fichiers, il faut compresser toutes les 6 fichiers avant transférer à mon serveur `genologin` au répertoire souhaité (répertoire `cd /work/iris/projet/exo31/fastq`). On utilise `gzip` sur chaque fichier téléchargé :

   ```
   $ gzip SRR2045415_1
   $ gzip SRR2045415_2
   ...
   ```

   Sur mon ordinateur,

   ```
   scp -r <chemin_source> iris@genobioinfo.toulouse.inrae.fr:/home/iris/work/projet/exo3/fastq
   ```

Ensuite, il faut télécharger le génome de référence avec son annotation correspondante associé par `wget` sur serveur local

- Le génome (fichier fasta):

```
$ cd /work/iris/projet/exo3
$ mkdir genome
$ cd genome
$ wget https://ftp.ensembl.org/pub/release-110/fasta/
gadus_morhua/dna/Gadus_morhua.gadMor3.0.dna.toplevel.fa.gz
```

- L'annotation (fichier gtf):

```
$ cd /work/iris/projet/exo3
$ mkdir annotation
$ cd anootation
$ wgethttps://ftp.ensembl.org/pub/release-110/gtf/gadus_morhua/Gadus_morhua.gadMor3.0.110.gtf.gz
```

4.2 Préparation de fichier bash et lancement du `nextflow`

Comme la partie précédente, ici il faut créer les fichiers inputs.csv, sm_config.cfg et run_pipeline.sh pour lancer le `nextflow`. Le fichier sm_config.cfg rassemble à ceux qui a présenté dans l'exercice 2. Pour le script run_pipeline.sh, on ne change que le chemin d'accès correspondant aux nouveaux fichiers de données correspondantes, on a laissé tous les paramètres qui sont rassemblés, ceux qu'on a présentés dans l'exercice 2 sauf ajouter l'option `-resume`. Par contre, il faut changer totalement le fichier inputs.csv qui correspond à 6 nouveaux fichiers SRA qu'on a téléchargé.

```
group,replicate,fastq_1,fastq_2,strandedness
SRR2045415,1,/home/iris/work/projet/exo3/fastq/SRR2045415_1.fastq.gz,/home/iris/work/projet/exo3/fastq/SRR2045415_2.fastq.gz,unstranded
SRR2045416,1,/home/iris/work/projet/exo3/fastq/SRR2045416_1.fastq.gz,/home/iris/work/projet/exo3/fastq/SRR2045416_2.fastq.gz,unstranded
SRR2045417,1,/home/iris/work/projet/exo3/fastq/SRR2045417_1.fastq.gz,/home/iris/work/projet/exo3/fastq/SRR2045417_2.fastq.gz,unstranded
inputs.csv (END)

group,replicate,fastq_1,fastq_2,strandedness
SRR2045415,1,/home/iris/work/projet/exo31/fastq/SRR2045415_1.fastq.gz,/home/iris/work/projet/exo31/fastq/SRR2045415_2.fastq.gz,foward
SRR2045416,1,/home/iris/work/projet/exo31/fastq/SRR2045416_1.fastq.gz,/home/iris/work/projet/exo31/fastq/SRR2045416_2.fastq.gz,foward
SRR2045417,1,/home/iris/work/projet/exo31/fastq/SRR2045417_1.fastq.gz,/home/iris/work/projet/exo31/fastq/SRR2045417_2.fastq.gz,foward
```

Comme sur la description des échantillons SRA sur NCBI, les échantillons sont orientés mais il ne remarque pas quel type de l'orientation (`forward` : orientation dans le sens de lecture 1; `reverse` : orientation dans le sens de lecture 2). C'est pourquoi, pour le paramètre `strandedness` dans le fichier inputs.csv, on a essayé mettre `unstranded` dans le répertoire `exo3` ou `forward` dans le répertoire `exo3` trouvé dans mon espace de travail, les résultats du `nextflow` semblent identiques dans les deux cas. Les raisons pour expliquer dans ce cas est peut-être :

- Ces données ne dépendent pas de l'orientation des données
- Le pipeline qu'on a utilisé peut détecter automatiquement l'orientation
- Les données ne représentent pas significatives entre `forward` et `unstranded`

Enfin, après avoir toutes les 3 fichiers `inputs.csv`, `sm_config.cfg` et `run_pipeline.sh`, on peut lancer le `nextflow` par la commande `sbatch` :

```
$ sbatch run_pipeline.sh
```

La sortie de seff est :

```
Job ID: 50755859
Cluster: genobull
User/Group: iris/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:04:18
CPU Efficiency: 1.38% of 05:11:24 core-walltime
Job Wall-clock time: 05:11:24
Memory Utilized: 1.85 GB
Memory Efficiency: 30.82% of 6.00 GB
```

Figure 33: *Comme les 6 échantillons `fastq` analysés sont assez grand (2Go – 4Go), le temps de lancement de ce `nextflow` est 5h11mins avec 1,85Go mémoire utilisés.*

4.3 Interprétation du résultat

Le fichier de résultat contient 6 fichiers comme on a présenté dans l'exercice 3. Les principaux résultats obtenus sont trouvés dans le report de multiQC

4.3.1 General statistics

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
SRR2045415_R1	47.7	0.23%	0.09%	22.0%	1.24	20.0	92.2%	58.6%	1.01%	7.5	40.2	94.6%
SRR2045415_R1_1												
SRR2045415_R1_2												
SRR2045416_R1	59.4	0.53%	0.01%	9.9%	1.21	27.0	66.1%	64.3%	0.61%	5.3	54.1	80.7%
SRR2045416_R1_1												
SRR2045416_R1_2												
SRR2045417_R1	68.6	2.80%	0.07%	33.2%	1.35	29.5	86.4%	50.1%	0.59%	9.5	59.1	92.3%
SRR2045417_R1_1												
SRR2045417_R1_2												

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
SRR2045415_R1	47.7	0.23%	0.09%	22.0%	1.24	20.0	92.2%	58.6%	1.01%	7.5	40.2	94.6%
SRR2045415_R1_1												
SRR2045415_R1_2												
SRR2045416_R1	59.4	0.53%	0.01%	9.9%	1.21	27.0	66.1%	64.3%	0.61%	5.3	54.1	80.7%
SRR2045416_R1_1												
SRR2045416_R1_2												
SRR2045417_R1	68.6	2.80%	0.07%	33.2%	1.35	29.5	86.4%	50.1%	0.59%	9.5	59.1	92.3%
SRR2045417_R1_1												
SRR2045417_R1_2												

Figure 34: *Le pourcentage de duplication est plus grand chez SRR2045415(20%) et SRR2045417 (32%) et plus petit chez SRR2045416 (9,9%).*
Le pourcentage de lectures qui sont alignables par rapport au nombre total de lectures est un peu différent selon les échantillons : SRR2045415 avec 92,2% et SRR2045417 avec 86,4% et SRR2045416 avec 66,1%.
Le pourcentage de lecture mappées par rapport au nombre total de lecture : SRR2045415 a 94,4%, SRR2045416 a 80,7% et SRR2045417 a 92,3%.

4.3.2 Biotypes

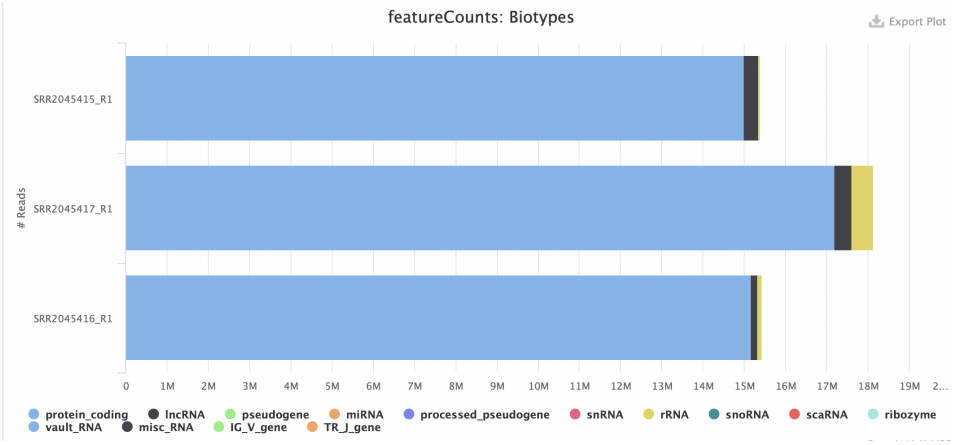


Figure 35: *La majorité de lectures de séquençage chevauchent sur les 3 échantillons sont des protéine_coding. Il y a une petite partie de lectures qui est lncARN (ARN long non codant, partie noir) et une très petite partie de rARN (partie jaune)*

4.3.3 DupRadar

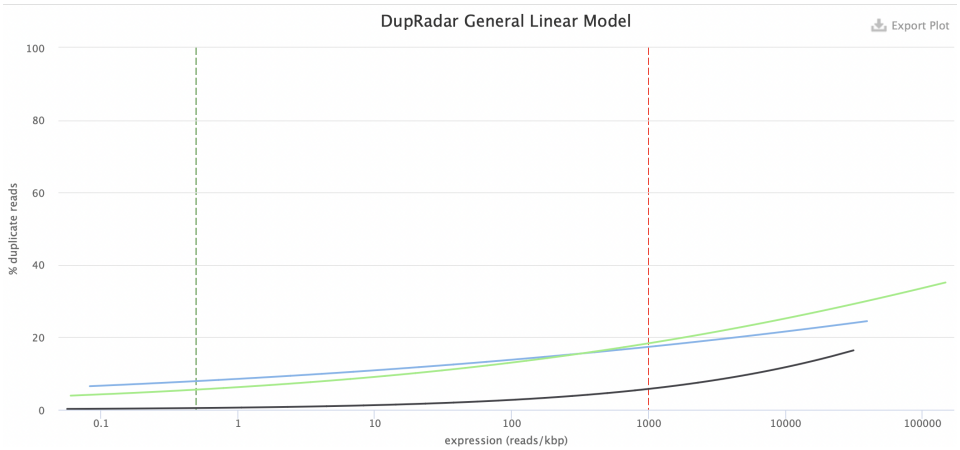


Figure 36: *Le pourcentage de duplication par rapport le nombre de lecture de SRR2045416 (courbe noir) est plus faible que les deux autres échantillons. SRR2045417 a le niveau de duplication plus grand que les deux autres sur les gènes fortement exprimés.*

4.3.4 Picard- Mark Duplicates

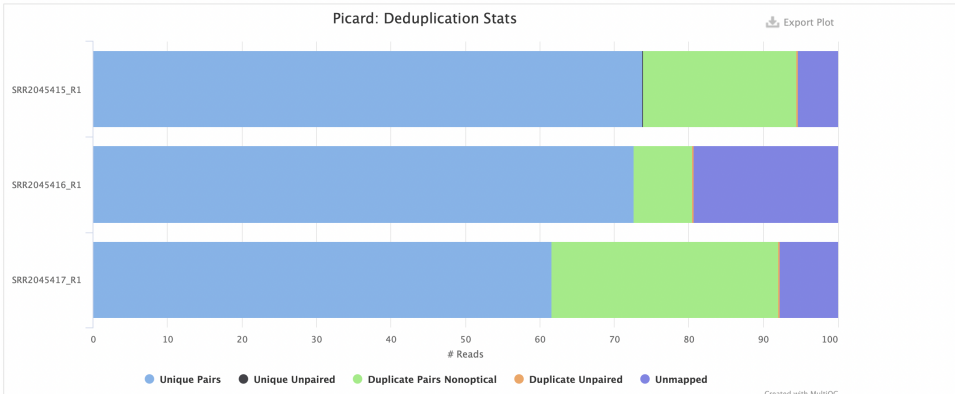


Figure 37: *SRR2045416 : >70%, SRR2045417 : >60% (partie bleu). Il y a une partie de lectures qui sont des paires de duplication non optiques (partie vert). Cela signifie que ces paires de lectures sont des copies identiques d'une paire de brins d'ADN, mais qu'elles ne sont pas le résultat d'une duplication artificielle due à des artefacts optiques ou des erreurs de séquençage. (SRR2045415 : 20%, SRR2045416 : 8%, SRR2045417 : 30%). L'autre partie (partie violet) est des lectures non mappées celles qui n'ont pas été alignées avec succès sur un génome de référence (SRR2045415 : 5%, SRR2045416 : 19%, SRR2045417 : 7%)*

4.3.5 Preseq- Complexity curve

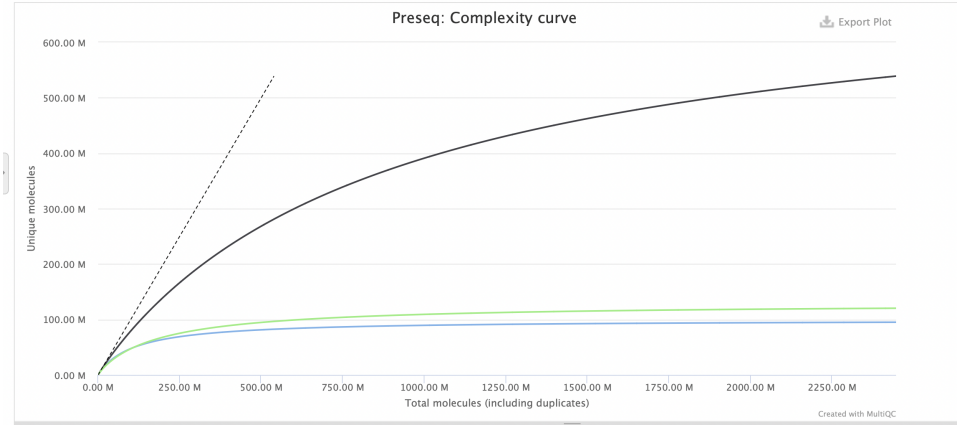


Figure 38: *SRR2045416 a la courbe plus élevée que celles des échantillons SRR2045415 et SRR2045417. Cela veut dire que cet échantillon SRR2045416 a une complexité plus élevée, ça signifie qu'il contient une plus grande variété de molécules d'ARN ou d'ADN provenant de différentes régions du génome, il a les séquences moins biaisées ou dupliquées que les autres. Cela semble conforme au résultat de picard sur mark duplicates.*

4.3.6 Qualimap

Genomic origin of reads

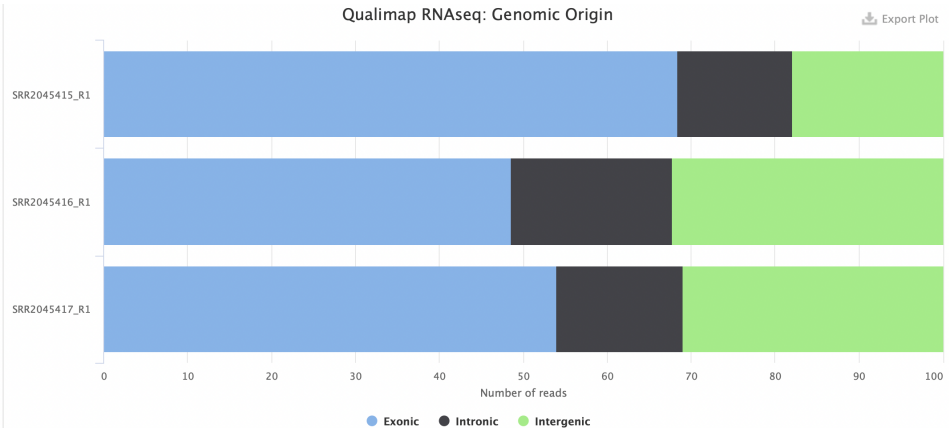


Figure 39: *La plupart de lectures des 3 échantillons ont aligné sur des régions exoniques du génome. SRR2045415 a de mêmes lectures alignées sur des régions introniques et intergéniques . Les deux échantillons SRR2045416 et SRR2045417 ont plus de 30% de lectures sur des régions intergéniques et plus de 15% de lectures sur des régions introniques.*

Gene coverage profile

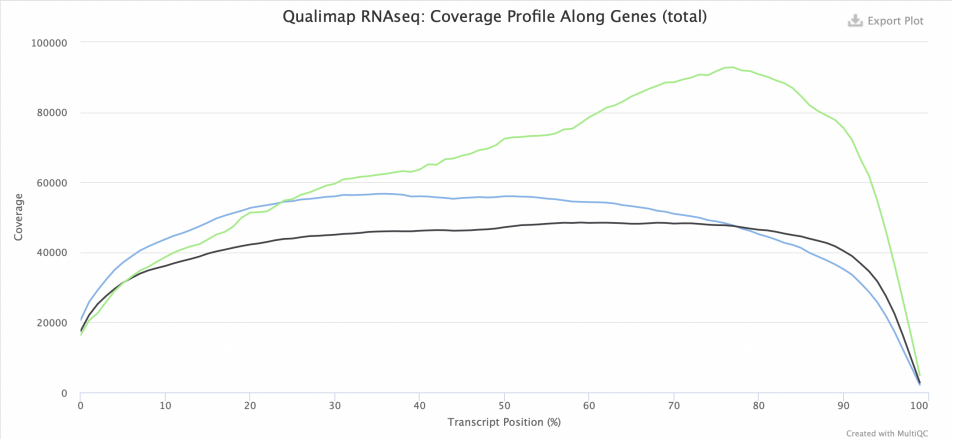


Figure 40: *SRR2045417 a une couverture génique plus élevée ou plus profonde que SRR2045415 et SRR2045416. Cela signifie qu'il a réussi à séquencer un plus grand nombre de transcrits ou de régions de gènes différents par rapport aux autres échantillons, il contient une plus grande diversité de transcrits, les régions spécifiques qui ne sont pas aussi bien couvertes aux autres échantillons.Les lectures séquencées à partir de cet échantillon ont pu être alignées plus efficacement.*

4.3.7 Rsem

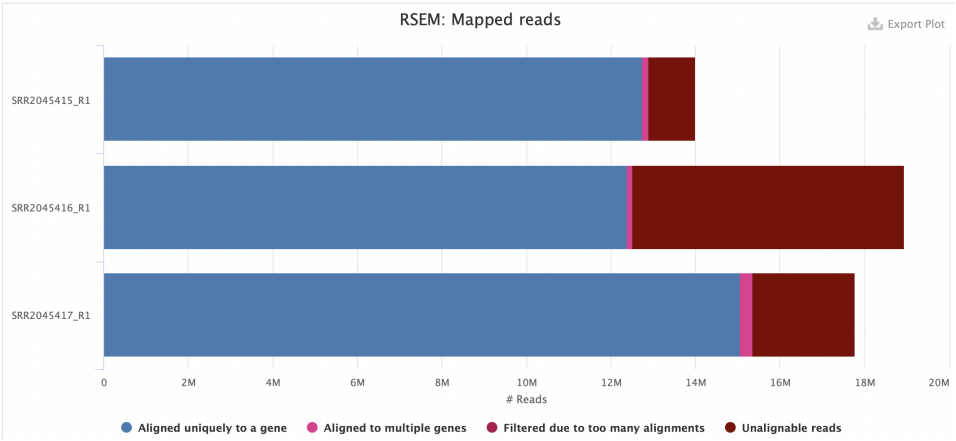


Figure 41: *En général, la majorité de lectures de ces 3 échantillons a été aligné de manière unique sur un seul gène et un très peu de lectures ont été alignées sur plusieurs gènes différents. Pour l'échantillon SRR2045416, il a une partie assez grande pour les lectures qui n'ont pas pu etre alignées. Cela correspond au résultat %Alignable dans la partie 4.3.1.*

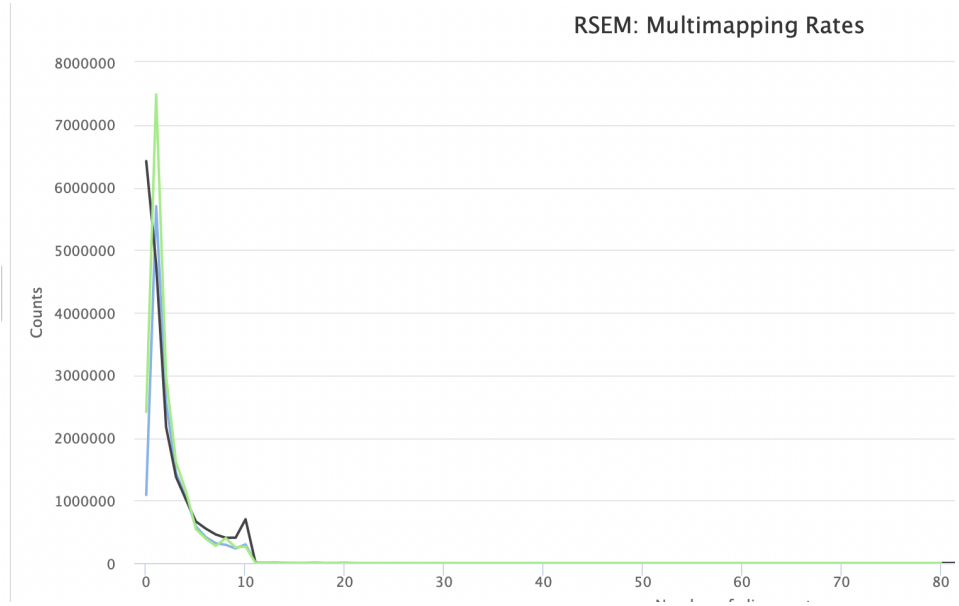


Figure 42: *Pour l'échantillon SRR2045415 et SRR2045417, la majorité des séquençages s'aligne une seule fois sur un génome de référence. Tandis que SRR2045416 a la plupart des séquençages qui ne s'aligne pas sur un génome de référence. Cependant, SRR2045416 a une grande partie de séquençages qui a été aligné une seule fois sur le génome de référence. Cela indique une spécificité élevée de l'alignement des lectures pour SRR2045415 et SRR2045417 (surtout SRR2045415) et aussi SRR2045416 mais peu que les deux autres.*

4.3.8 RSeQC

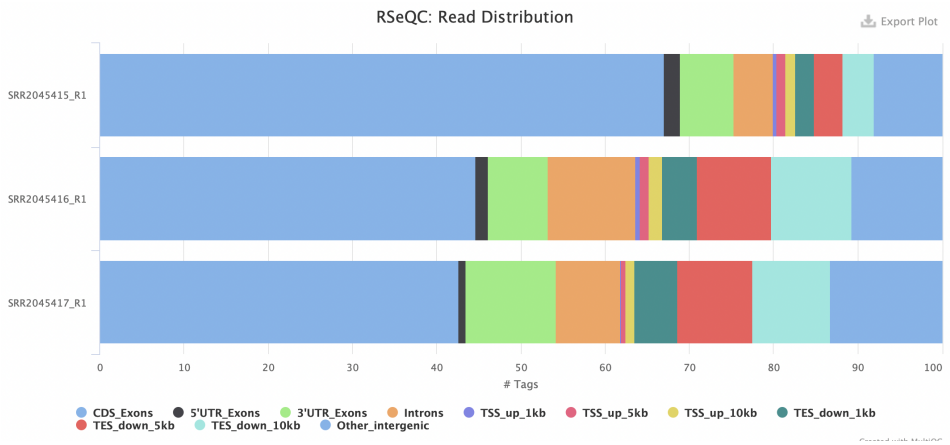


Figure 43: *La distribution de 3 échantillons sont très variées surtout pour les deux échantillons SRR2045416 et SRR2045417*

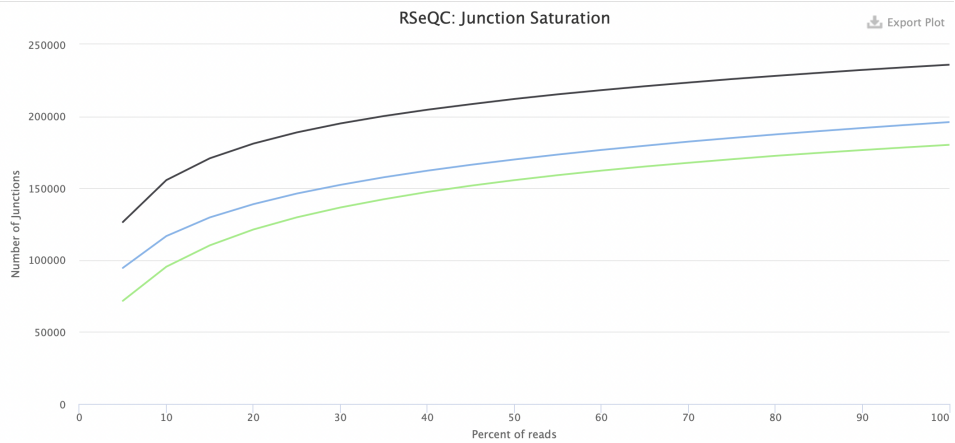


Figure 44: *L'échantillon SRR2045416 contient une plus grande diversité de jonctions d'épissage par rapport aux autres échantillons. cet échantillon a une meilleure couverture des jonctions d'épissage, ce qui peut être le résultat d'une meilleure qualité des données de séquençage ou d'une meilleure capacité à séquencer des régions de jonction.*

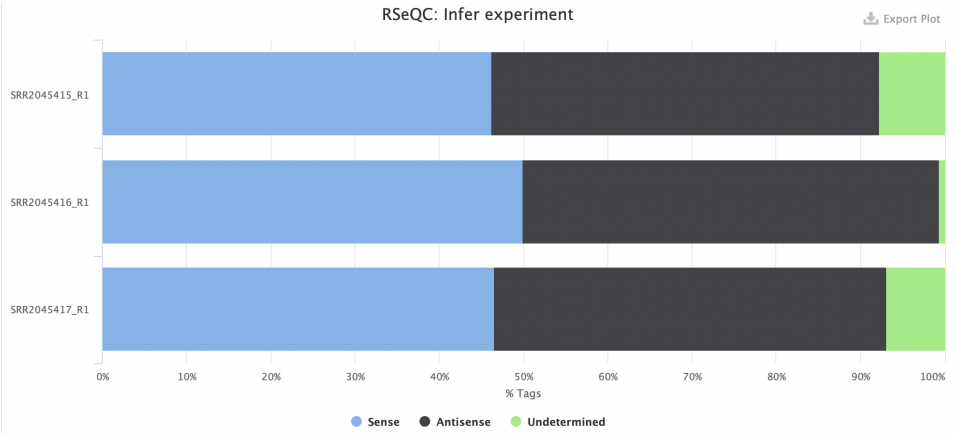


Figure 45: *Pour toutes les 3 échantillons, l'orientation des lectures est répartie de manière quasi-égale entre le sens et l'antisens des transcrits, avec très peu de lectures dont l'orientation n'est pas déterminée.*

4.3.9 Samtools

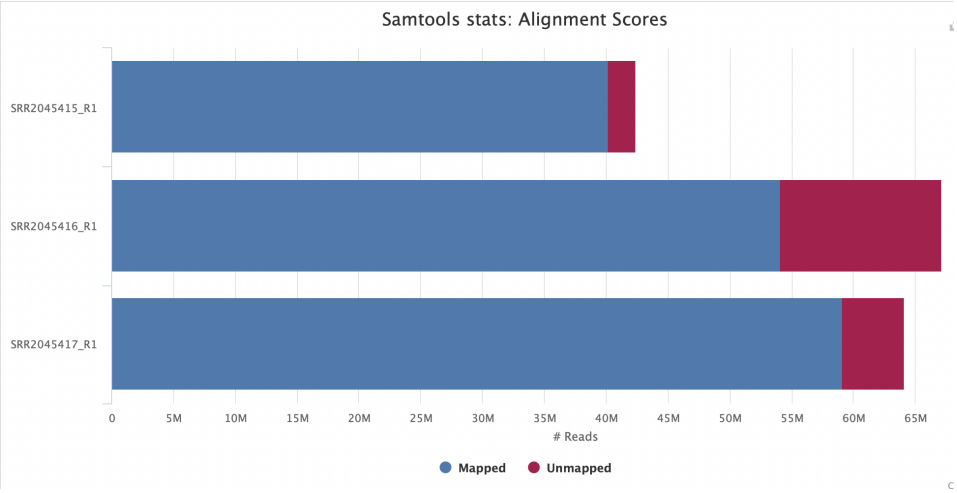
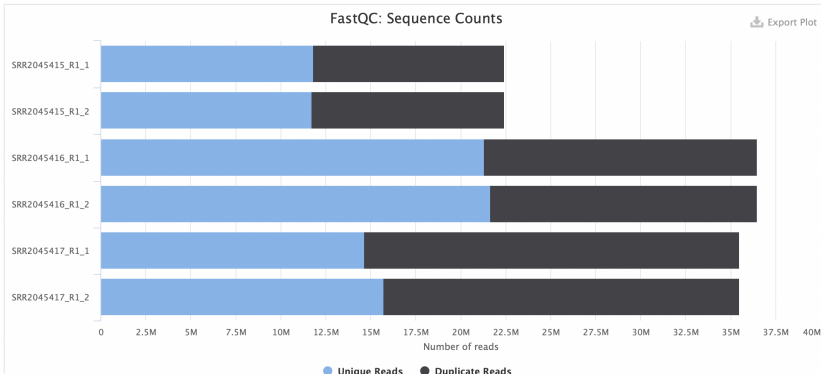
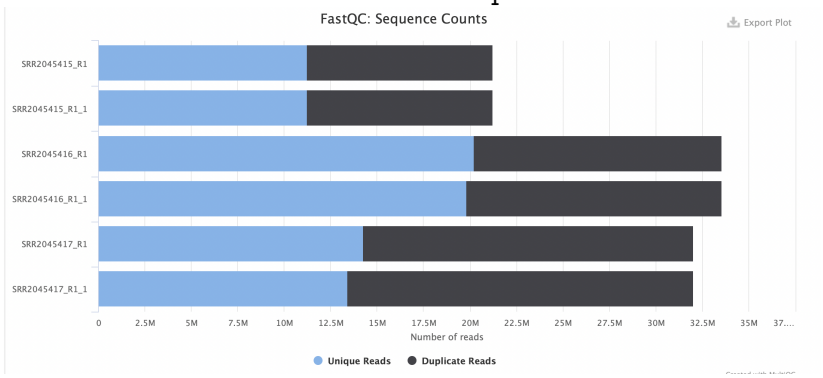


Figure 46: *Sur les 2 échantillons SRR2045415 et SRR2045417,le score d'alignement est grand (>90%) tandis que SRR2045416 est un peu plus petit que les 2 autres (>80%). Ce résultat correspond au résultat de* `rsem` *(mapped reads)*

4.3.10 FastQC



Avant cutadapt

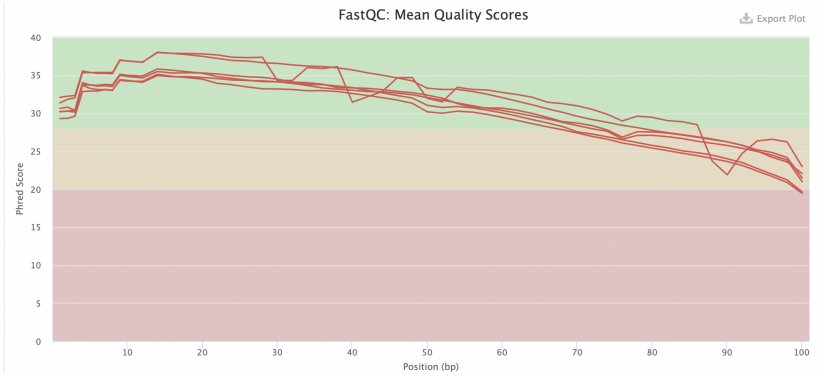


Après cutadapt

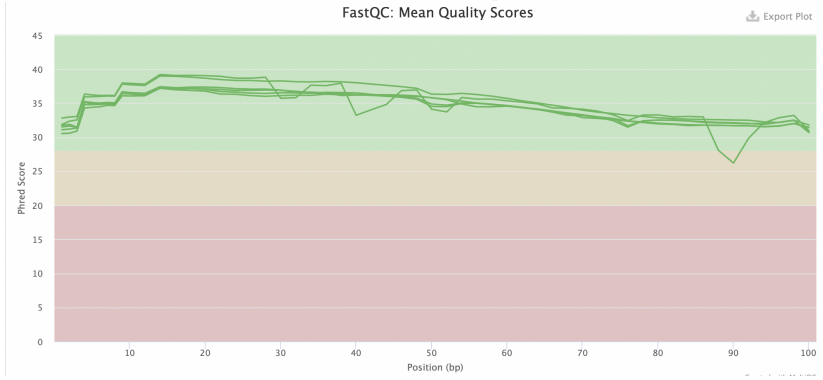
Figure 47: *FastQC: Sequence Counts*

Il n'y a pas de différence significative avant et après le traitement des adaptateurs

Mean Quality Scores



Avant cutadapt



Après cutadapt

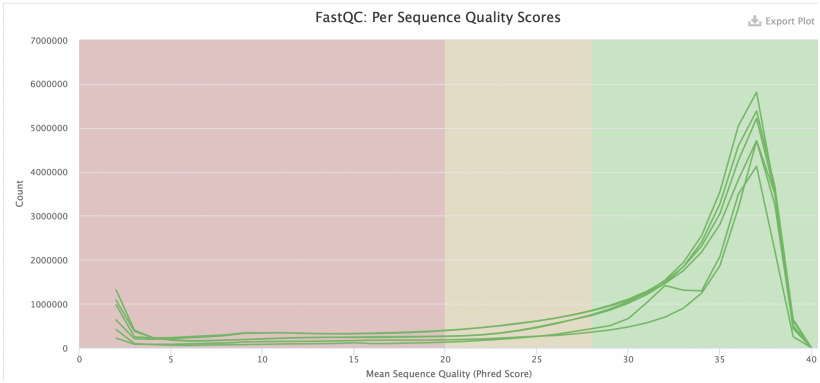
Figure 48: *FastQC: Mean Quality Scores*

On observe une amélioration des résultats du *Sequence Quality Histograms* après la suppression des adaptateurs dans le processus de prétraitement des données de séquençage.

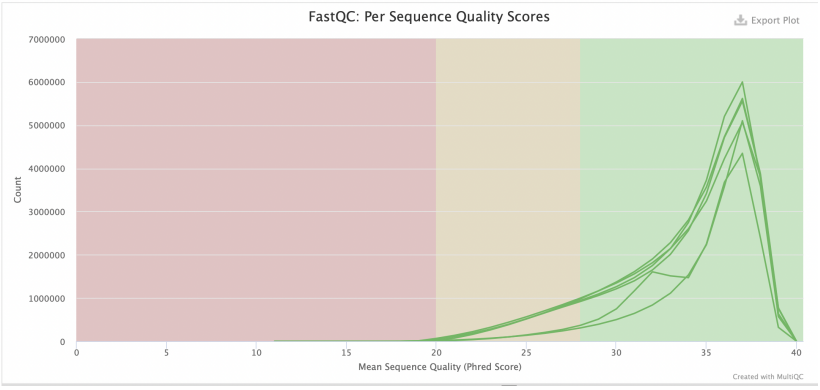
Avant la suppression des adaptateurs, les échantillons ont échoué dans la "zone verte", il y a peut-être un artéfact qui entraine des scores de qualité de séquence plus bas.

L'élimination des adaptateurs peut effectuer les scores de qualité des bases de séquence ont probablement augmenté, car les bases de séquence étaient maintenant plus fiables et ne contenaient plus d'artefacts ou d'adaptateurs

Per Sequence Quality Scores



Avant `cutadapt`

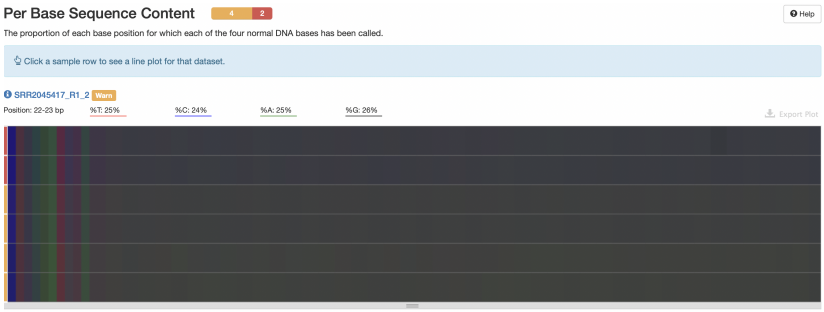


Après `cutadapt`

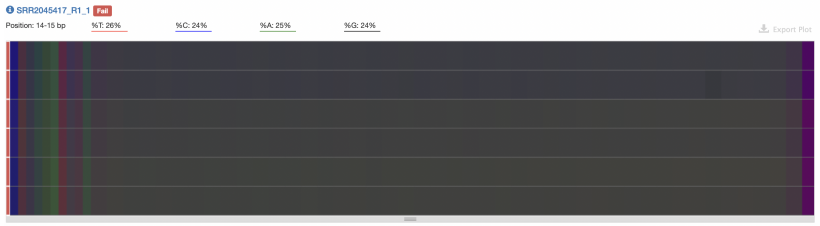
Figure 49: *FastQC: Per Sequence Quality Scores*

On observe une amélioration de la qualité des données à la suite de la suppression des adaptateurs qui supprime les counts avec les scores de 2-20 phred score. Le traitement de `Cutadapt` peut aider à éliminer les séquences qui étaient mal alignées en raison de la présence d'adaptateurs, d'artéfacts, ce qui peut entraîner des scores de qualité (phred score) bas pour ces séquences.

Per Base Sequence Content



Avant `cutadapt`

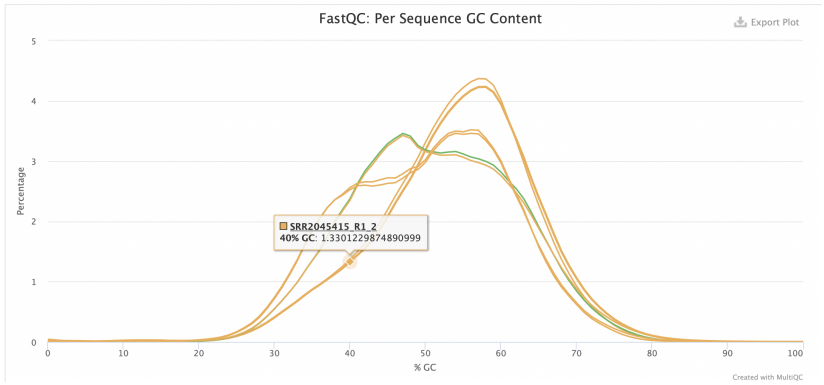


Après `cutadapt`

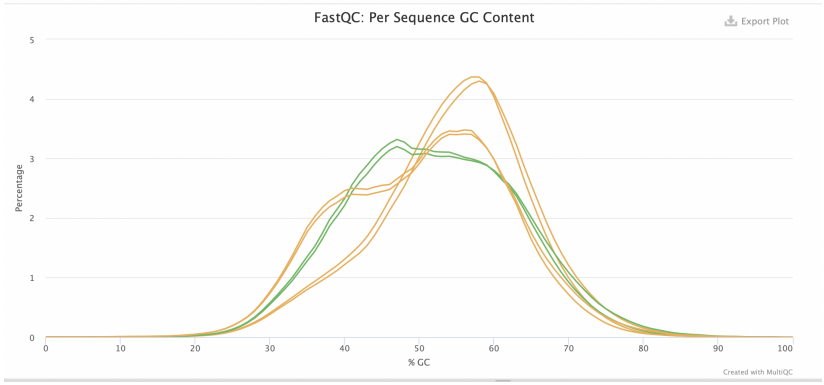
Figure 50: *FastQC: Per Base Sequence Content*

Dans toutes les 2 cas, il y a un déséquilibre au début de la séquence peut être causé à cause du biais dans la partie 5′. Cutadapt supprime les adaptateurs et d'autres séquences indésirables à l'extrémité 3′ ou initialement présenté à 3′. Cette suppression peut modifier la composition en bases des séquences, ce qui peut se manifester sous forme d'un nouveau biais à 3′.

Per Sequence GC Content



Avant `cutadapt`



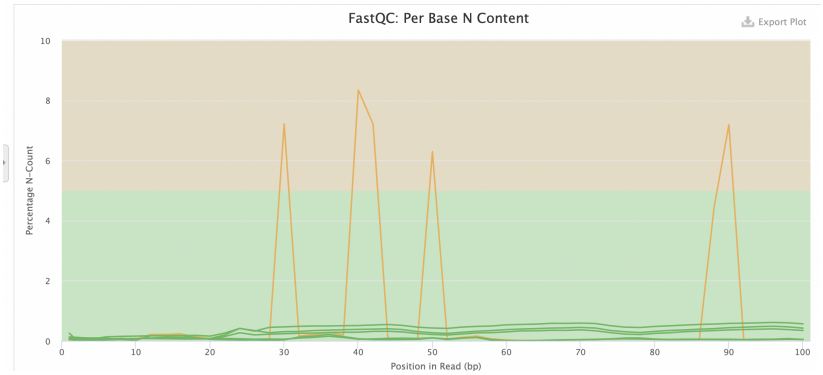
Après `cutadapt`

Figure 51: *FastQC: Per Sequence GC Content*

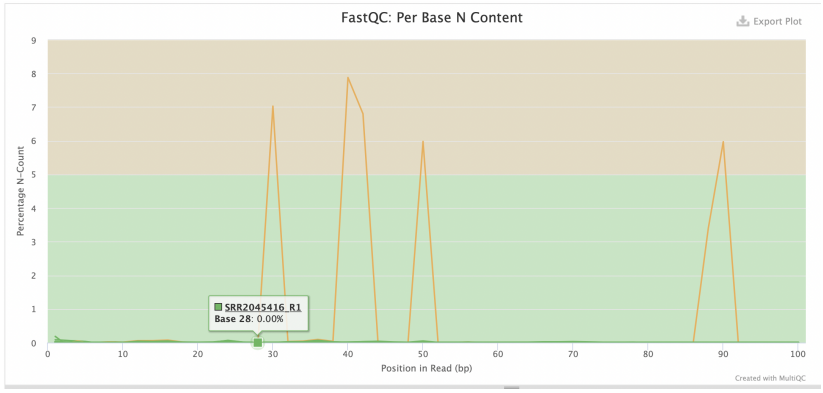
On observe une amélioration légère de la qualité des données, en particulier en ce qui concerne

le contenu en GC des séquences qui fait un échantillon de l'état warning (avertissement : jaune) à passed (réussi :vert). Si les adaptateurs contiennent des régions avec des compositions en GC variables ou de mauvaise qualité, leur élimination peut réduire la variabilité du contenu en GC dans les données, conduisant ainsi à une courbe de contenu en GC plus homogène.

Per Base N Content



Avant `cutadapt`



Après `cutadapt`

Figure 52: *FastQC: Per Base N Content*

Le fait que les courbes des échantillons "passed" (réussi, en vert) présentent un pourcentage de N bas et restent relativement plates après le traitement avec `Cutadapt` est généralement un signe de nettoyage efficace des données. L'échantillon SRR2045415_R1_1(courbe jaune) avait initialement un pourcentage de N très élevé et que ce pourcentage n'a pas changé après le traitement avec `Cutadapt`, cela peut indiquer que le problème de qualité des données était sévère dans cet échantillon. Cet échantillon est peut-être de mauvaise qualité, être contaminé, etc.

Avec les données brutes, chaque séquence de votre échantillon est de la même longueur.

Sequence Length Distribution

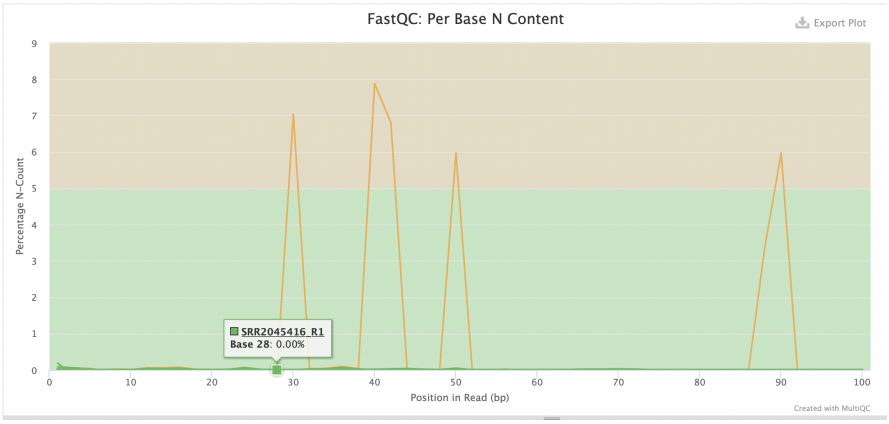
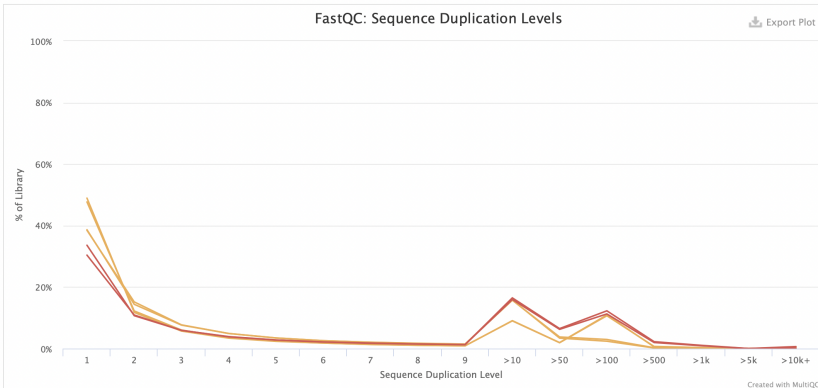


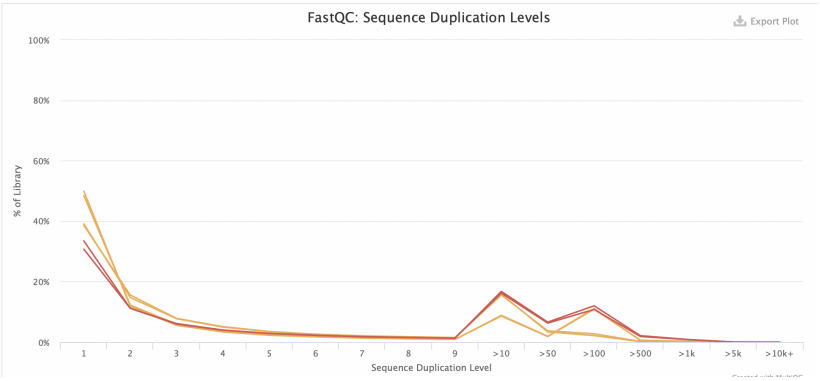
Figure 53: *FastQC: Sequence Length Distribution après* `cutadapt`

`Cutadapt` peut éliminer des séquences d'adaptateurs ou de mauvaise qualité, des séquences spécifiques qui étaient initialement présentes en grand nombre dans les données, des artéfacts avec des longueurs inhabituelles ce qui peut entraîner une modification de la distribution des longueurs des séquences.

Sequence Duplication Levels



Avant `cutadapt`



Après `cutadapt`

Figure 54: *FastQC: Sequence Duplication Levels*

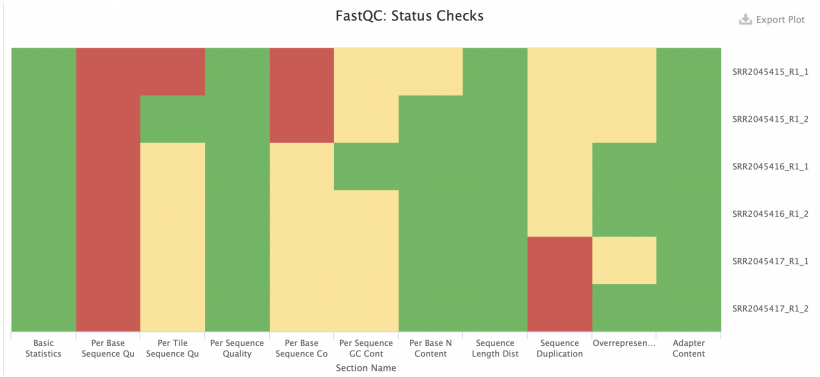
On observe une absence de différence significative dans les niveaux de duplication avant et après `Cutadapt` peut être due à plusieurs facteurs, notamment la nature de ces données de séquençage (faible de niveau de duplication initiale donc l'utilisation `Cutadapt` n'a pas eu un impact significatif sur les niveaux de duplications), l'efficacité de `Cutadapt` pour éliminer la duplication, etc.

Overrepresented sequences

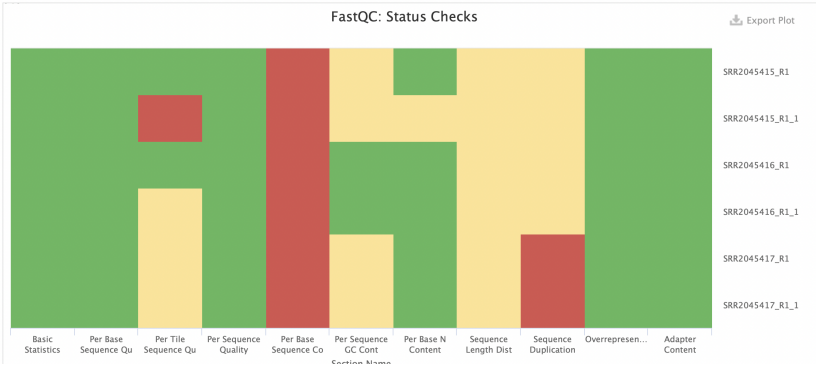
- Avant `cutadapt` : les 3 échantillons sont passed(réussi) et les 3 échantillons sont warning (avertissements)
- Après `cutadapt` : les 6 échantillons sont passed(réussi)

Trois de nos échantillons qui étaient initialement en "warning" pour les séquences sur-représentées sont maintenant passés en "réussi," cela signifie que Cutadapt a efficacement supprimé ces séquences sur-représentées de ces échantillons. (due à contamination, artéfacts de séquençages, préparation de la bibliothèque)

Status Checks



Avant `cutadapt`



Après `cutadapt`

Figure 55: *FastQC: Status Checks*

Les résultats après utilisation Cutadapt améliore les états de certaines parties : per Base sequence Quality, Per Sequence GC Content, Per Base N Content , Overpresented sequences.

Cependant, `Cutadapt` dégénère l'état de Per Base Sequence Content et Sequence Length Distribution.

5 Références

- 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
- https://youtu.be/qPbIlO_KWNO
- https://multiqc.info/docs/#using-multiqc
- https://github.com/nf-core/rnaseq/tree/master/docs/images
- https://academic.oup.com/bioinformatics/article/32/19/3047/2196507?login=false
- https://nf-co.re/rnaseq/3.12.0/docs/usage
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/