projet Nextflow

Catherine Bonjour

6 octobre 2023

Contents

1 Les paramètres du fichier de lancement **2**

2 Les principaux répertoires **3**
2.1 fastqc . . . 3
2.2 genome . . . 5
2.3 multiqc . . . 6
2.3.1 General Statistics . . . 6
2.3.2 Biotype Counts . . . 7
2.3.3 DupRadar . . . 8
2.3.4 Picard Mark Duplicates . . . 9
2.3.5 Preseq . . . 9
2.3.6 QualiMap . . . 10
2.3.7 Rsem . . . 11
2.3.8 RSeqQC . . . 12
2.3.9 Samtools . . . 14
2.3.10 FastQC . . . 15
2.3.11 Cutadapt . . . 23
2.3.12 nfcore/rnaseq Software Versions . . . 24
2.3.13 nfcore/rnaseq Workflow Summary . . . 24
2.4 trimgalore . . . 25
2.5 pipeline info . . . 25
2.6 star rsem . . . 26
2.6.1 bigwig . . . 27
2.6.2 dupradar . . . 27

2.6.3 rseqc 28
2.6.4 featurecounts 29
2.6.5 stringtie 29
2.6.6 picard metrics 29
2.6.7 preseq 29
2.6.8 qualimap 29

3 RNA-seq avec des nouvelles données 29

4 Références 30

1 Les paramètres du fichier de lancement

```
dahlia@genologin1 ~/work/projet $ more run_pipeline.sh
#!/bin/bash
#SBATCH -J CatherineBonjour
#SBATCH -p workq
#SBATCH --mem=6G
#SBATCH --time=24:00:00

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

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

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

Figure 1: le fichier de lancement

Avec SBATCH -J on peut changer et personnaliser le nom du job. Avec SBATCH -p workq j'ai changé la partition. Il existe d'autres partitions comme unlimitq accessible pour tous et d'autres qui sont accessibles sur demande.Chacune à une priorité différente en fonction du temps exécution. Pour configurer le temps maximum à un jour j'ai fait SBATCH –time=24:00:00 pour déclarer que c'est 24 heures maximum. Le module utilisé est bioinfo/nfcore-Nextflow-v21.04.1

Pour savoir si le job est fini on utilise la **commande seff** avec l'identifiant de job. Il donne des informations comme l'identifiant du job,la durée du pipeline

```
dahlia@genologin2 ~/work/TP $ seff 50634553
Job ID: 50634553
Cluster: genobull
User/Group: dahlia/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:03:04
CPU Efficiency: 30.56% of 00:10:02 core-walltime
Job Wall-clock time: 00:10:02
Memory Utilized: 1.85 GB
Memory Efficiency: 30.90% of 6.00 GB
```

Figure 2: la sortie du seff

(Job Wall-clock time), la quantité et le pourcentage de mémoire utilisée (memory),le processeur (CPU), le nombre de nœuds (cores), le nom du cluster (genobull) et le statut : completed si le job a terminé sans erreurs.

Une option de Nextflow est la commande **resume** qui permet d'exécuter un script en utilisant des résultats en cache. C'est donc utile pour continuer un job qui était arrêté par une erreur.

2 Les principaux répertoires

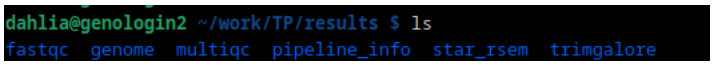

Figure 3: les principaux répertoires

2.1 fastqc

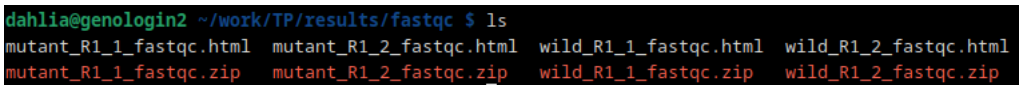

Figure 4: fichiers du répertoire fastqc

L'outil **fastqc** permet de contrôler la qualité de nos séquences avant de faire d'autres analyses. Les résultats du fastqc sont visibles dans les fichiers html générés.

- **Per base sequence quality**

  Ici on peut voir que la qualité des séquences est relativement bonne. Plus les boîtes jaunes représentent une grande partie de l'image et plus la qualité des séquences est mauvaise.

- **Per tile sequence quality**

  Elle est spécifique aux librairies Illumina. Chaque carré représente une partie de la flowcell.Un graphique convenable devrait être bleu sur toute sa surface.

- **Per sequence quality scores**

  Permet de voir si un sous-ensemble de nos séquences a des scores faibles. Il faut que ces séquences représentent un faible pourcentage parmi toutes les séquences. La qualité est mesurée avec le score phred.

- **Per base sequence content**

  Les 4 lignes représentent les 4 bases et doivent être le plus parallèles possible. Cela permet de contrôler que la distribution des 4 bases est équilibrée

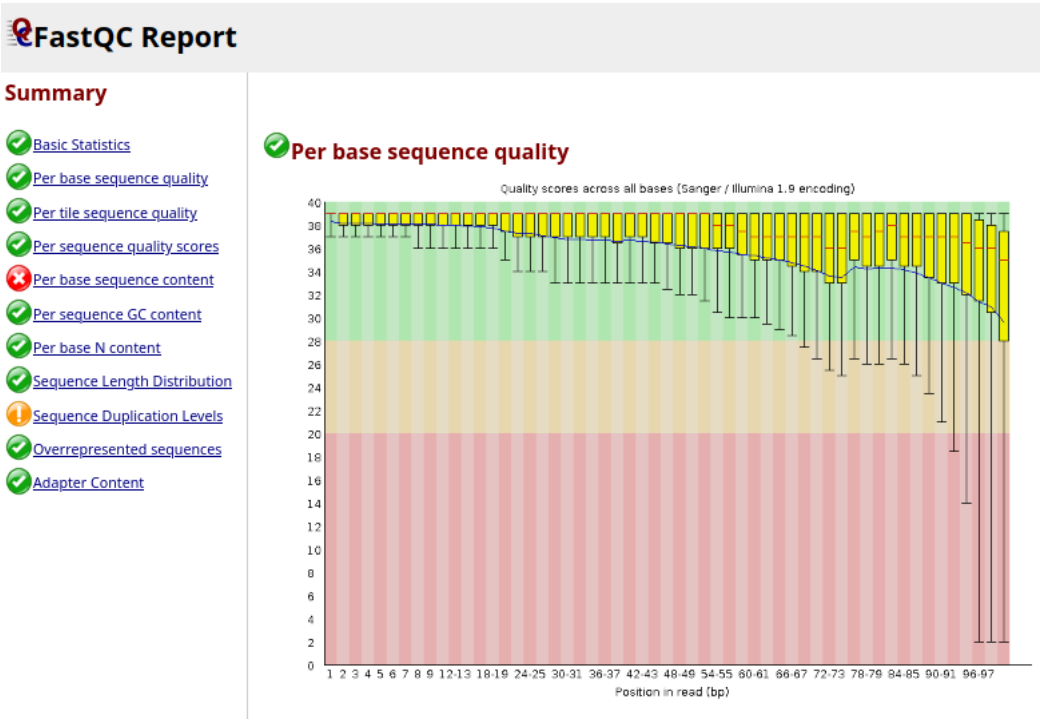

Figure 5: la qualité des séquences pour un réplicat

parmi les séquences. Pourtant avec certaines librairies on voit une composition déséquilibrée surtout au début des séquences.

- **Per sequence GC content**

  Compare la distribution en bases GC sur toute la longueur de chaque séquence (ligne rouge), avec la distribution théorique (ligne bleu)

- **Per base N content**

  Indique le pourcentage de bases non identifiées (N) par le séquenceur à chaque position.

- **Sequence length distribution**

  Le graphique montre la distribution de la taille des séquences. IL est normal que certains séquenceurs génèrent des fragments de taille différente.

- **Sequence Duplication Levels**

  Le plot montre le nombre relatif de séquences avec des degrès de duplication différents. Il faut que le niveau de duplication ne soit pas elevé.

- **Overrepresented sequences**

Parmi toutes les séquences il faut pas que une séquence soit surreprésentée car cela peut indiquer une contamination.

- **Adapter content**

  Pour vérifier si les adaptateurs utilisés lors des séquençages sont présents dans nos séquences.

2.2 genome

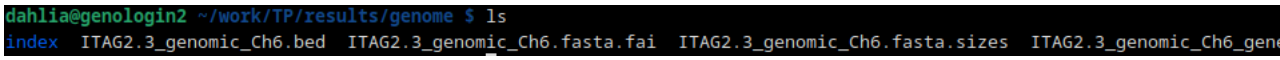

Figure 6: le répertoire genome

Ce répértoire contient le génome de référence et d'autres fichiers qui le décrivent.

Le fichier bed stocke les régions génomiques sous forme de coordonnées ainsi que les annotations associées.

le fichier .fai est générée par le logiciel samtools faidx. C'est l'index du fichier fasta correspondant.

Le fichier fasta.sizes contient la taille du génome de référence (içi le chromosome 6).

le fichier gtf contient les annotations du génome de référence. C'est comme un fichier gff il contient les exons.

On y retrouve aussi les fichiers générés par le logiciel **rsem**. Ce logiciel estime les niveaux d'expression des gènes et des isoformes à partir des données du rna-seq

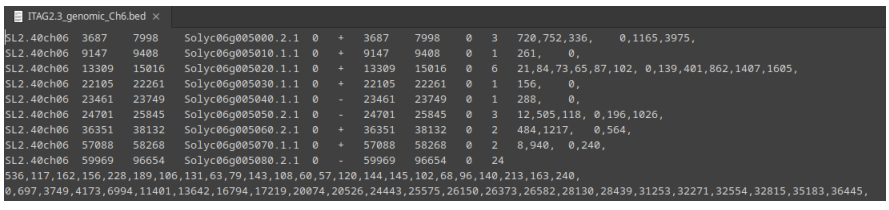

Figure 7: fichier format bed

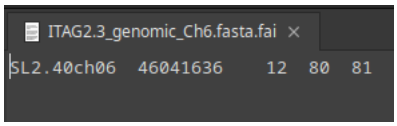

Figure 8: fichier format fai

2.3 multiqc

MultiQC résume tous les résultats du pipeline en un seul fichier html. Pour cela il utilise les résultats générés par des logiciels différents,comme il est possible de voir dans la figure 10. Dans le fichier html il a un bouton help au dessus de chaque graphique qui permet de comprendre chaque analyse. Cela aide à l'interprétation des graphiques. Au tout début du fichier html on a un résumé

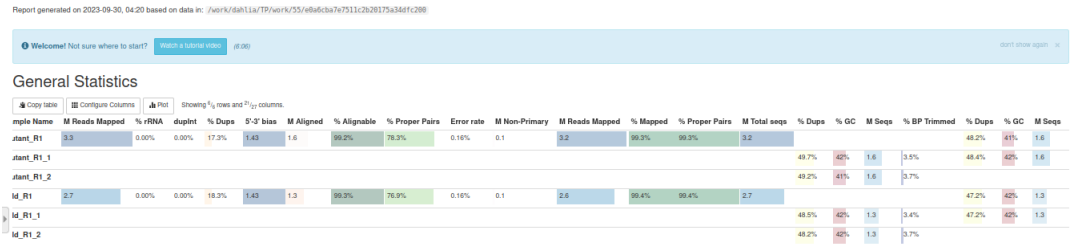

Figure 9: les statistiques pour chaque échantillon

des statistiques.

2.3.1 General Statistics

Il regroupe toutes les statistiques pour chaque échantillon. Le paramètre calculé le plus important est le nombre total de reads bruts (M seqs). Il donne aussi le pourcentage de reads dupliqués. IL faut que ce pourcentage ne soit pas grand et aussi que la différence de pourcentages entre les 2 échantillons ne soit pas grande. Le 5' 3' bias est pour savoir si nos données ont des biais 5' ou 3' qui pourraient être dus à de la dégradation de l'ARN survenue durant les étapes

différentes de la préparation de nos échantillons. Dans notre cas la valeur de 5' 3' bias et le pourcentage de duplication est faible.

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

Figure 10: les outils de multiqc

2.3.2 Biotype Counts

Il montre le nombre et le pourcentage de reads qui se chevauchent sur le génome. Dans notre cas toutes les séquences s'alignent.

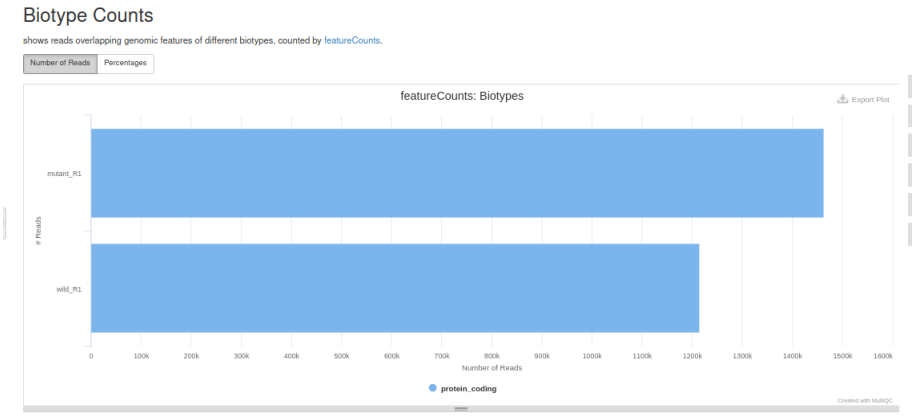

Figure 11: biotype counts

2.3.3 DupRadar

Nous informe sur le niveau de duplication de nos données.Pour les gènes très exprimés (nombre de reads grand, il est attendu d'avoir un grand nombre de duplication aussi.Donc il ne doit pas y avoir un grand pourcentage de reads dupliqués pour les gènes faiblement exprimés. Dans notre cas pour l'échantillon mutant et l'échantillon sauvage, le niveau de duplication reste faible pour les gènes faiblement exprimés (nombre de reads faible).

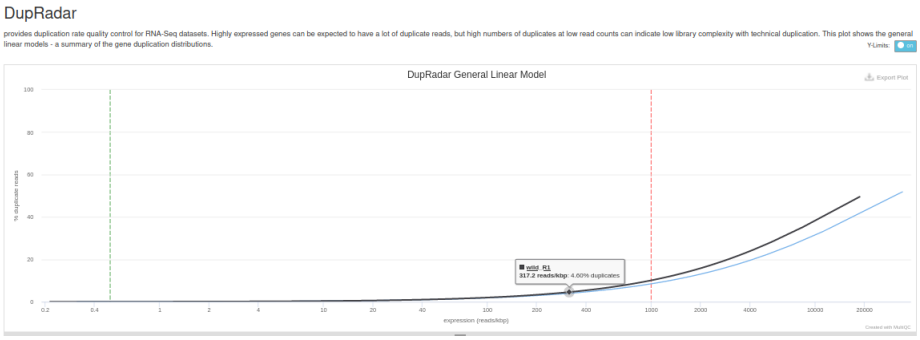

Figure 12: plot qui résume le niveau de duplication en fonction du nombre de reads

2.3.4 Picard Mark Duplicates

.

On a un graphique qui résume le nombre de reads triés par statut de duplication .Les duplicate pairs optical sont des artéfacts qui proviennent d'un cluster d'amplification qui est reconnu comme des clusters multiples par le séquenceur. Ici on a pas cet artéfact ou alors très peu (couleur verte)

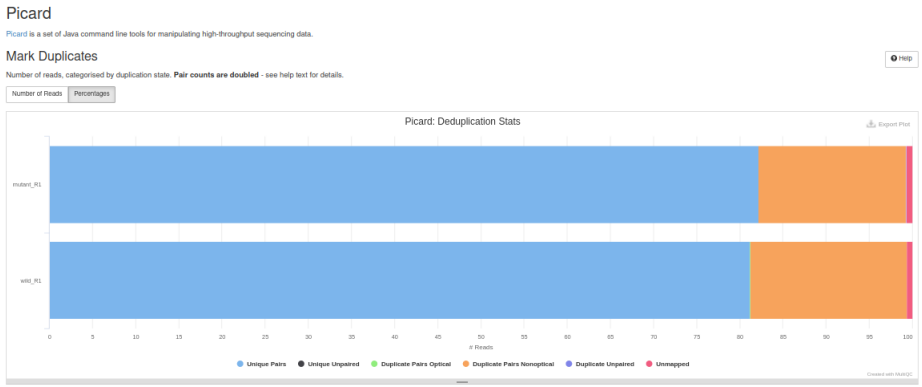

Figure 13: Mark Duplicates

2.3.5 Preseq

Il estime la variété des séquences issues d'une librairie de séquençage. PLus les courbes se rapprochent de la ligne en pointillé et plus la complexité de la librairie est grande. Dans la figure 14 on voit alors que la complexité n'est pas très grande.

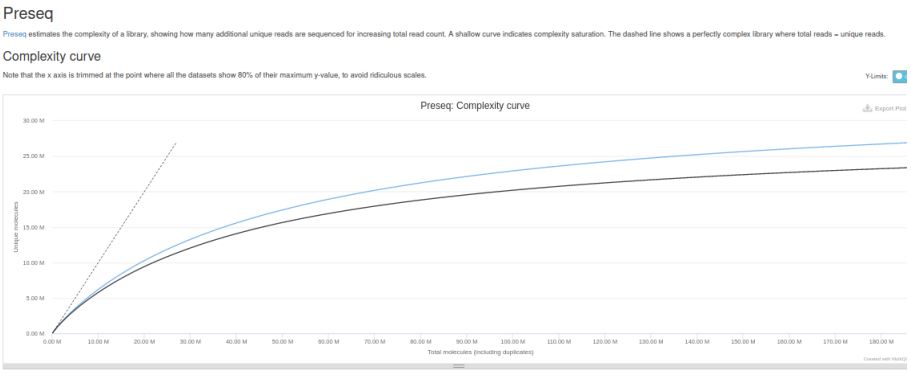

Figure 14: Complexity curve

2.3.6 QualiMap

Avec QualiMap on voit que la plus grande partie de nos séquences représentent des régions où se trouvent des exons(couleur bleu). Les régions où se trouvent des introns en vert et en noir les régions intergéniques.

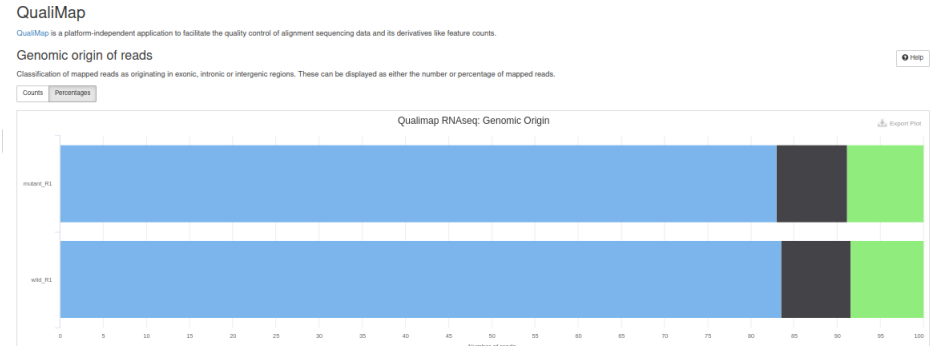

Figure 15: QualiMap: Genomic origin of reads

Le deuxième graphique montre la distribution de la profondeur de séquençage parmi tous les reads. Quand il y a pas de biais on s'attend à avoir une couverture élevée au milieu de nos séquences avec une couverture faible aux extrémités 5' et 3'. C'est le cas içi.(figure 16)

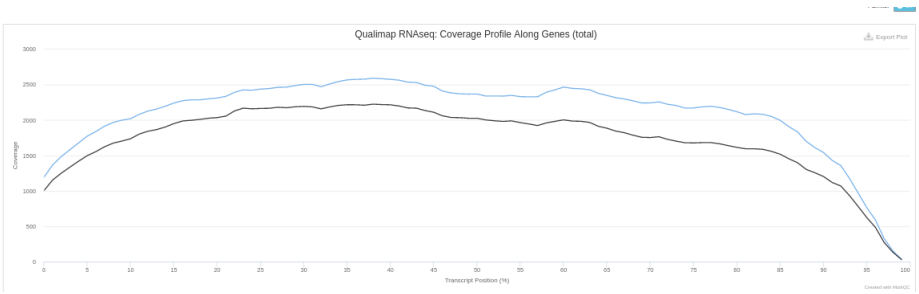

Figure 16: QualiMap: Gene coverage profile

2.3.7 Rsem

Il donne deux graphiques mapped reads et multimapping rates. Dans le premier on voit que la majorité des séquences s'alignent sur des gènes uniques (en bleu). Dans le deuxième graphique on voit que la majorité des séquences s'alignent une fois sur le génome de référence. Cela indique que nos échantillons sont de bonne qualité.

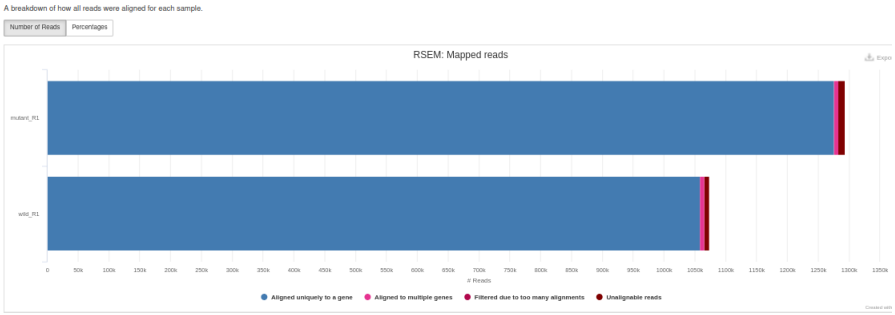

Figure 17: rsem: mapped reads

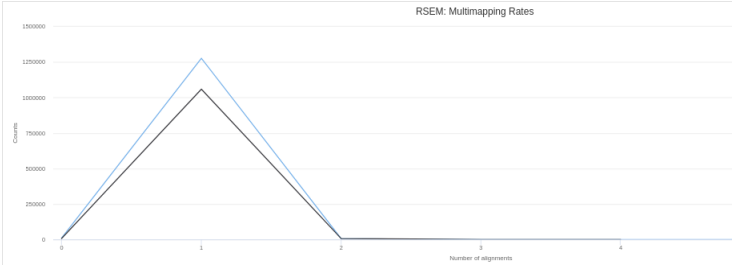

Figure 18: rsem: multimapping rates

2.3.8 RSeqQC

C'est une librairie qui contient des scripts qui évaluent la qualité des données rna-seq. Dans le graphique **read distribution** on voit que nos séquences couvrent en majorité des régions qui correspondent à des exons. Ceci est attendu dans une expérience rna-seq avec des bons résultats.

Avec **Inner Distance** on peut voir que la distance entre deux reads contigus est faible. Ceci est souvent observé dans des séquences vieilles ou dégradées.

Read duplication montre le nombre de lectures (axe y) en fonction du nombre d'occurrences correspondant. On voit que nos échantillons ont quelques duplications mais cela reste acceptable. Une grande région sous la courbe pourrait indiquer que les échantillons on beaucoup de lectures avec beaucoup de duplications.

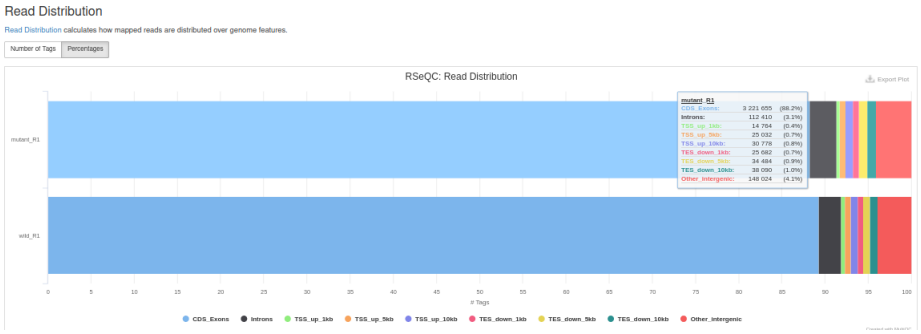

Figure 19: **RSeQC**:Read distribution

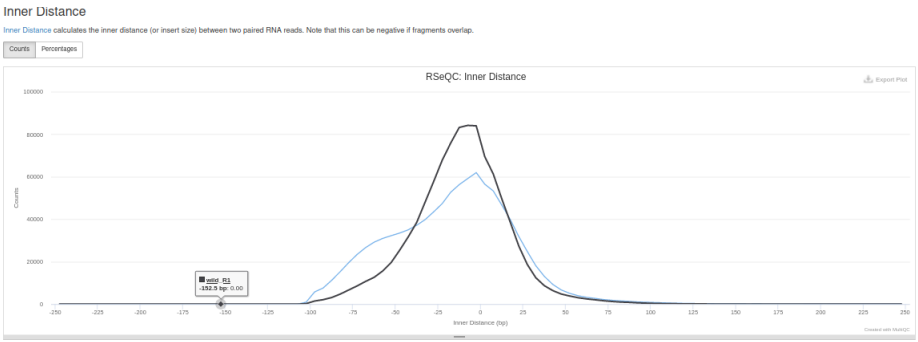

Figure 20: **RSeQC**:Inner distance

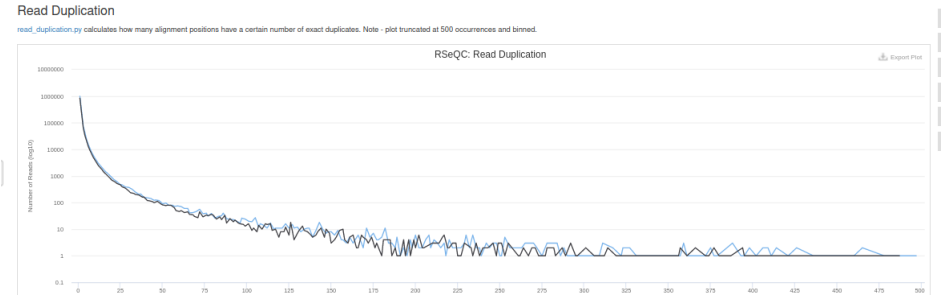

Figure 21: **RSeQC**:Read duplication

L'outil **Junction Annotation** compare les jonctions d'épissage détectées à celles d'un génome de référence. Dans figure 22 on voit que le mutant comporte plus de nouveaux sites d'épissage (en vert)

Figure 22: **RSeQC**:Splicing junctions

Junction saturation mesure le nombre de jonctions d'épissage connues retrouvées dans nos séquences. La courbe doit être le plus plate possible signe que la profondeur de séquençage est suffisante. C'est le cas pour nos échantillons.

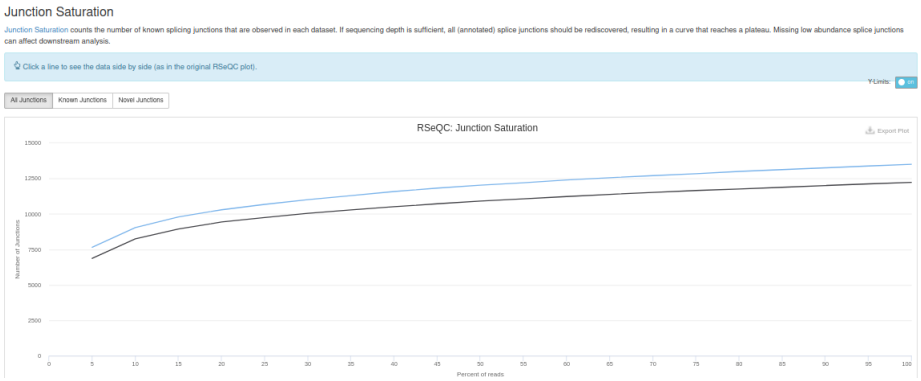

Figure 23: **RSeQC**:Junction saturation

Infer experiment prédit le sens des brins du protocole. Ce sens est utilisé pour préparer l'échantillon à séquencer en examinant le sens avec laquelle les reads vont s'orienter par rapport aux 'features/annotations' des gènes dans le génome de référence. Ici la quantité de brins sens et antisens est équilibrée.

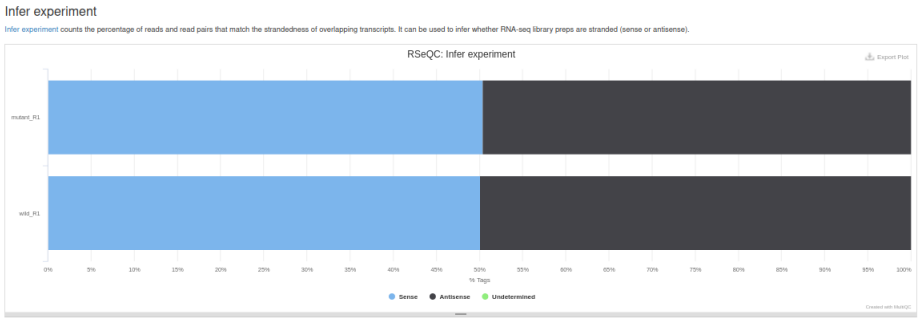

Figure 24: **RSeQC**:Infer experiment

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

2.3.9 Samtools

Samtools est un ensemble de programmes permettant d'interagir avec les données issues du séquençage RNA-seq. Dans le graphique **percent mapped** on voit que la majorité des séquences s'alignent sur le génome de référence.

Alignement metrics est un ensemble de mesures sur l'alignement.

Samtools Flagstat compte le nombre d'alignements pour chaque type flag(catégorie). Permet de contrôler que l'alignement est vraisemblable

Mapped reads per contig le logiciel idxstats tool compte le nombre de lectures qui s'alignent par chromosome et pour chaque contig.

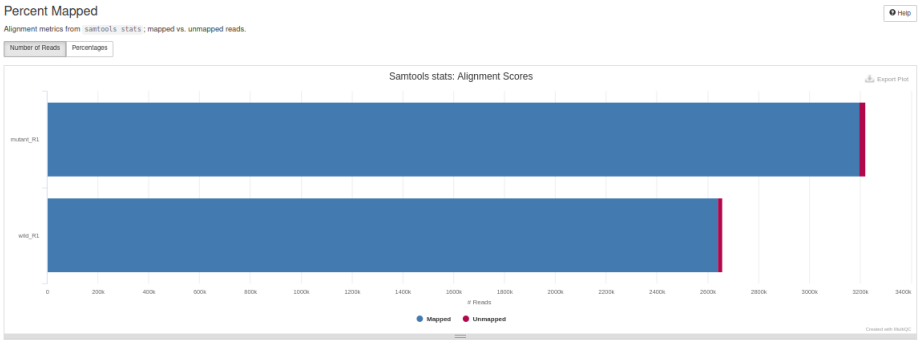

Figure 25: **Samtools**:percent mapped

2.3.10 FastQC

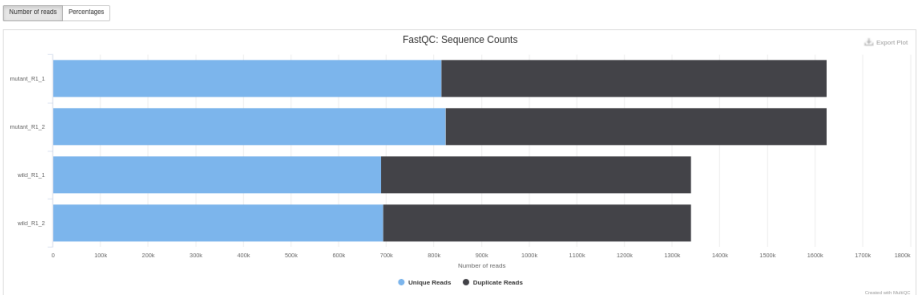

Figure 26: **FastQC**:Sequence Counts

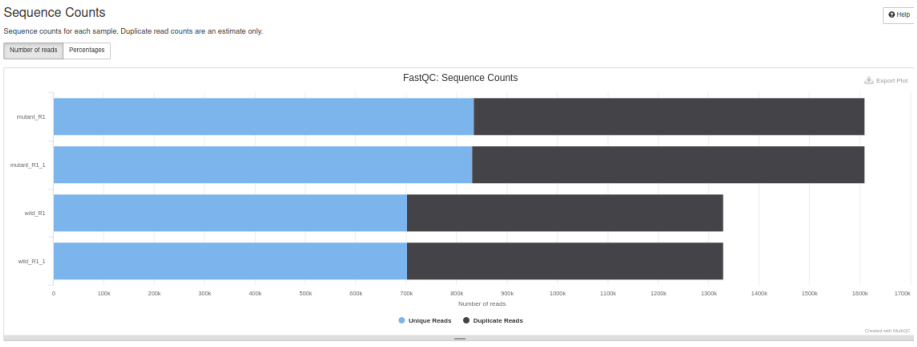

Figure 27: **FastQC**:Sequence Counts after trimming

Le graphique **Sequence counts** montre le nombre total de reads uniques (en bleu) et le nombre total de reads dupliqués (en noir) avant et après retrait des adaptateurs. On ne remarque pas de changement après retrait des adaptateurs.

Il y a presque autant de lectures uniques que de doublons ce qui suggère que la profondeur de séquençage n'est pas très grande sinon on aurait eu beaucoup plus de séquences en double.

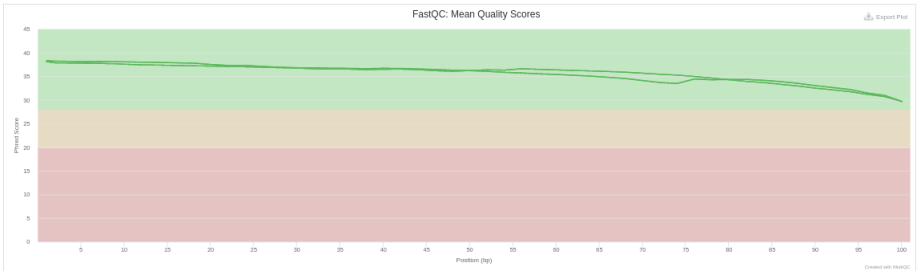

Figure 28: **FastQC**:Mean Quality Scores

Figure 29: **FastQC**:Mean Quality Scores after trimming

Mean Quality Scores montre le score phred de qualité parmi toutes les séquences. La courbe se retrouve sur un font vert signe que les scores sont de très bonne qualité. On remarque une légère augmentation du score après réduction des adaptateurs.

Per sequence quality score Montre le nombre de reads en fonction du score moyen de qualité. On retrouve un pic à la fin signe que la majorité des séquences ont un score elevé.

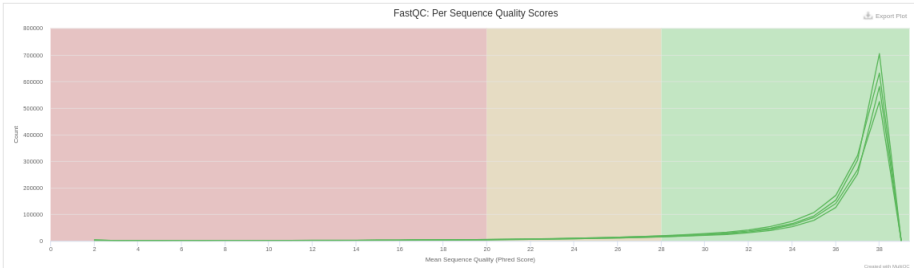

Figure 30: **FastQC**:Per sequence quality scores

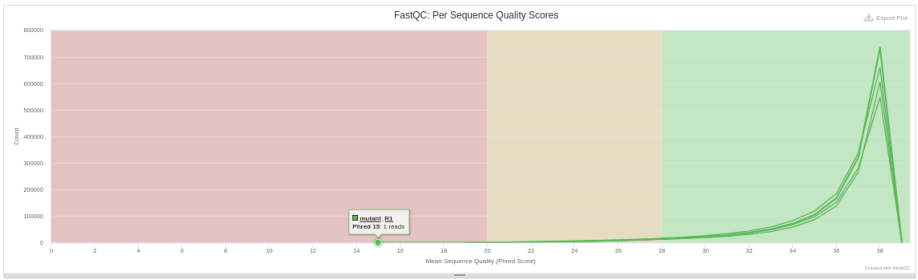

Figure 31: **FastQC**:Per sequence quality scores after trimming

Per base sequence content

On a des heatmap qui montrent la proportion en bases de nos données.On remarque dans tous les cas un déséquilibre dans la proportion en bases sur le début de la séquence. Dans les analyses précédentes nous n'avions pas trouvé de contamination mais on avait remarqué un biais dans la partie 5' de la séquence qui laissait suggérer un problème sur les séquences sur cette partie(échantillon abimé). On remarque qu'après retrait des adaptateurs un déséquilibre apparait en fin de séquence. C'est surement lié au fait qu'on a retiré les adaptateurs avec Trim galore qui en fonction de l'endroit où il a coupé il va créer un déséquilibre sur les bases.

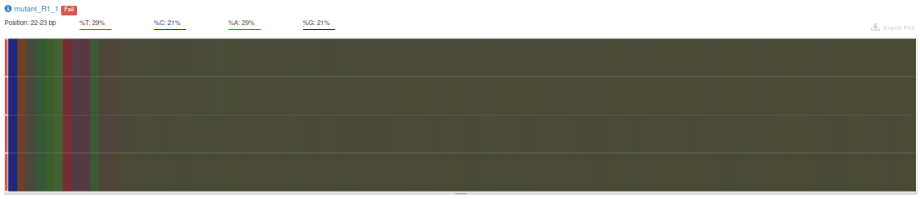

Figure 32: **FastQC**:Per base sequence content

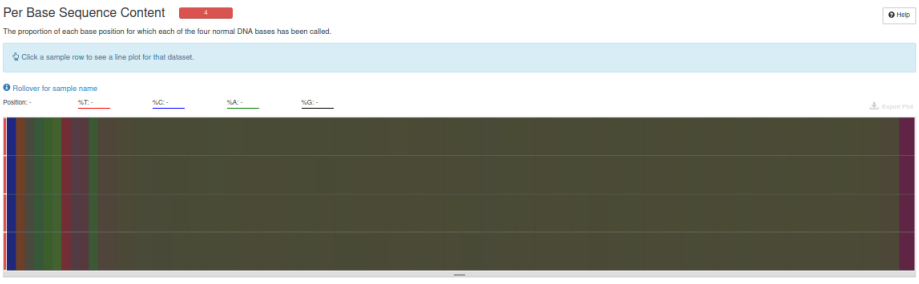

Figure 33: **FastQC**:Per base sequence content after trimming

sequence GC content

montre le contenu moyen en bases GC. En remarque pas de différence avant et après retrait des adaptateurs et que la distribution en bases GC est normale.

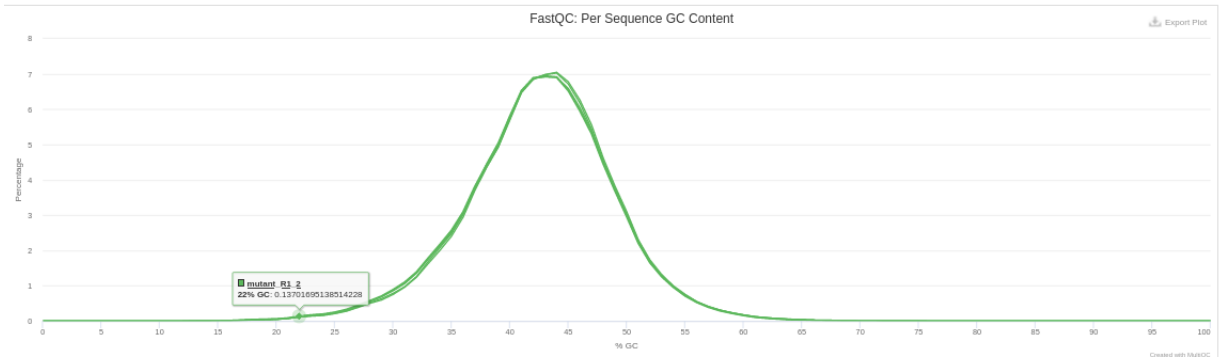

Figure 34: **FastQC**:Per sequence GC content

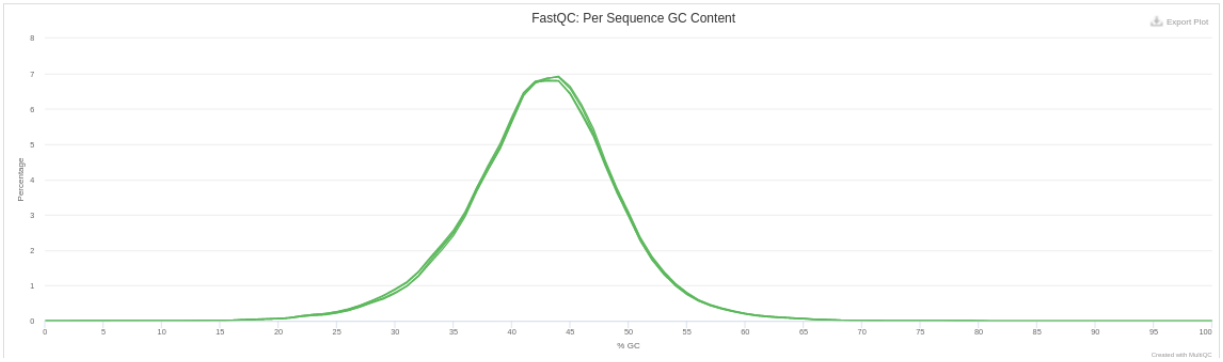

Figure 35: **FastQC**:Per sequence GC content after trimming

Per base N content Cela aide à contrôler que le pourcentage de bases non reconnues (N) n'est pas élevée. Pour nos échantillons le pourcentage de bases N reste assez faible. Cela est satisfaisant.

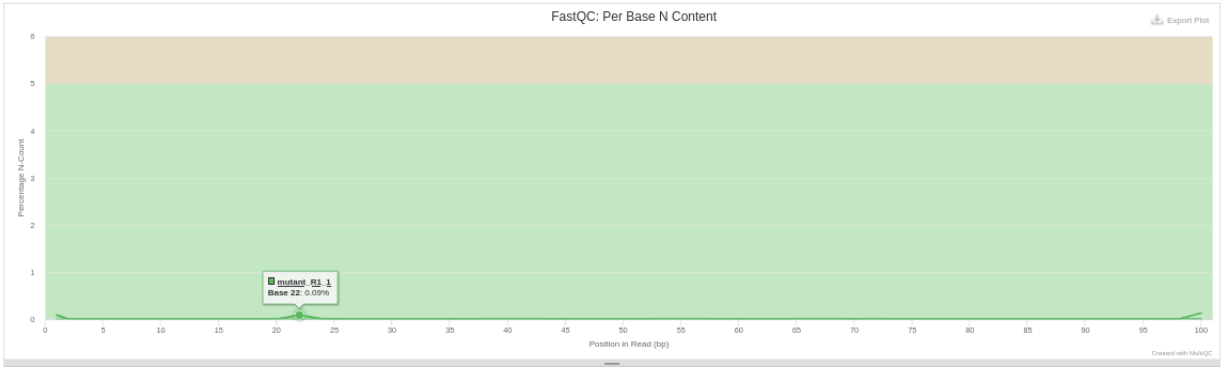

Figure 36: **FastQC**:Per base N content

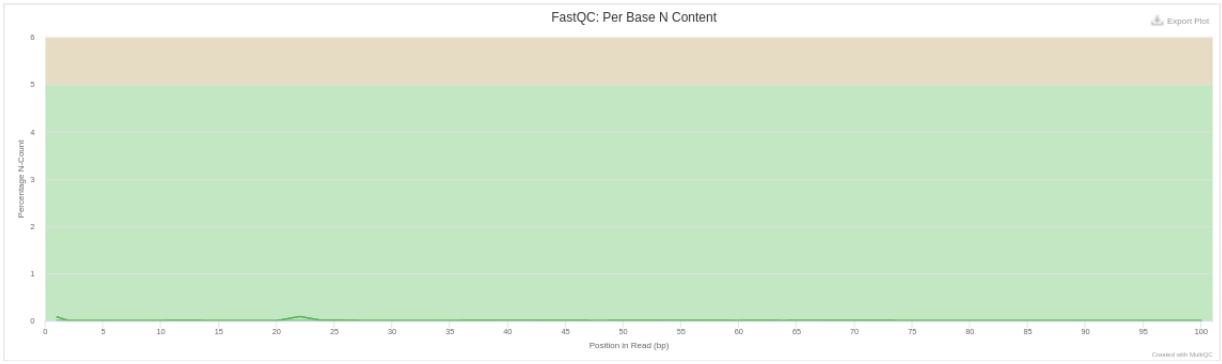

Figure 37: **FastQC**:Per base N content after trimming

Sequence Length Distribution Toutes les séquences ont la même taille de 101 paires de bases avant d'enlever les adaptateurs. Après réduction on remarque qu'on a des séquences de tailles différentes.(Figure 38)

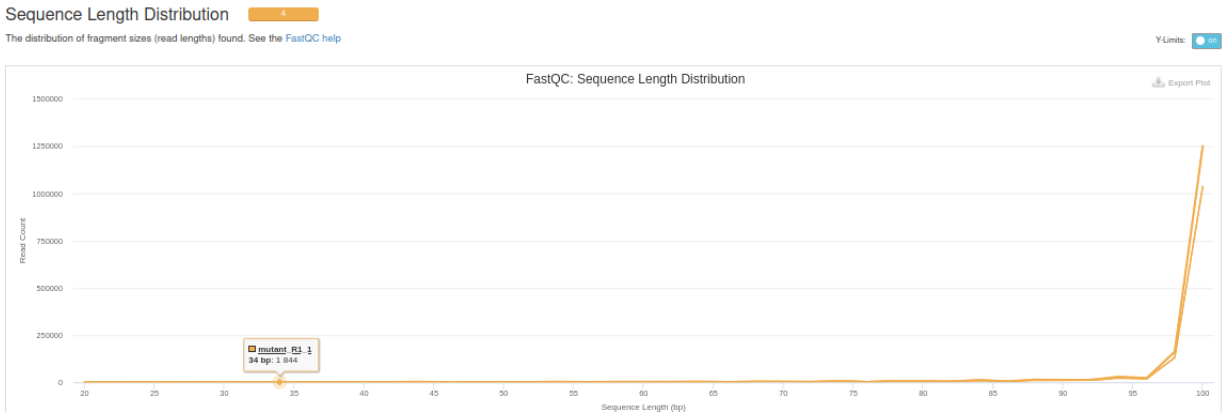

Figure 38: **FastQC**:Sequence Length Distribution after trimming

Sequence duplication levels

Les graphiques montrent le degrés de duplication pour chaque séquence dans une librairie. Il y a pas de différence avant et après retrait des adaptateurs. Dans un cas optimal la majorité des duplications doivent se retrouver dans la partie gauche du graphique. Cette situation est retrouvée avec nos échantillons. De plus le pourcentage de duplication est globalement faible. Il y a seulement un pic de duplication pour l'échantillon sauvage mais il reste faible. Il est peut-être dû à des duplications optiques.

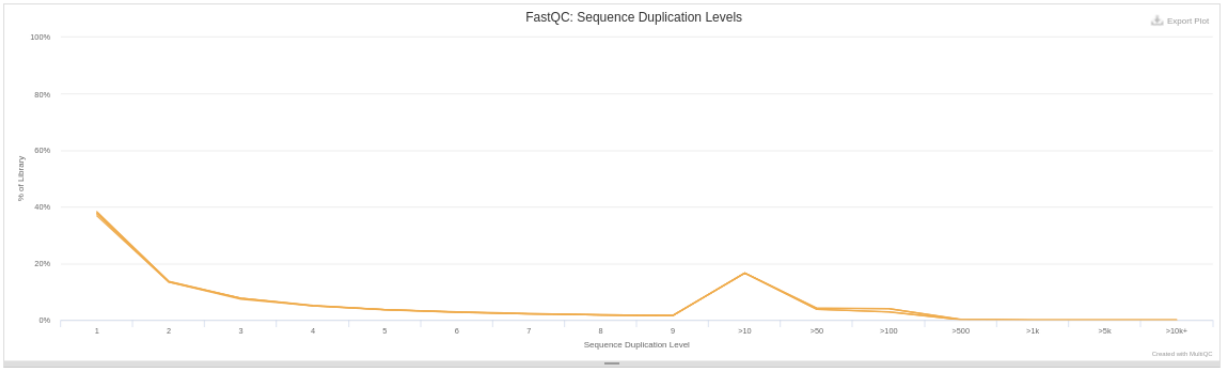

Figure 39: **FastQC**:Sequence Dulication Levels

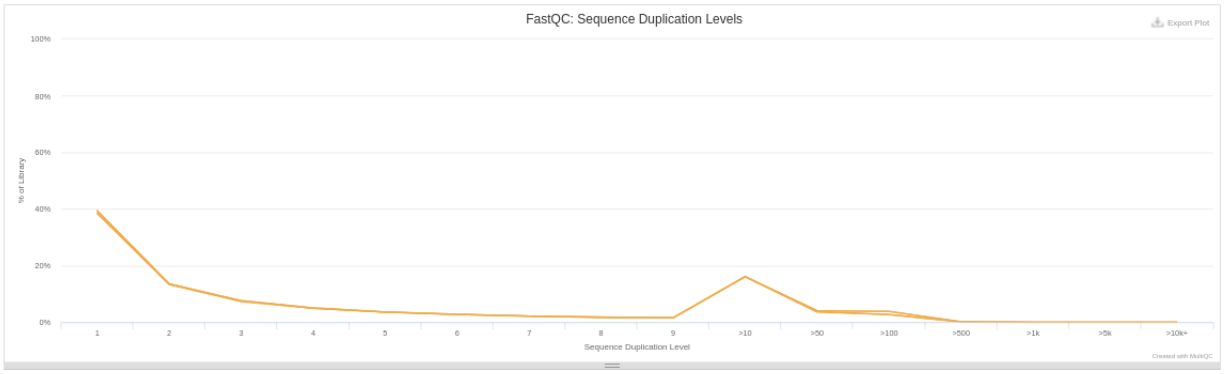

Figure 40: **FastQC**:Sequence Duplication Levels after trimming

Overrepresented sequences Le nombre total de séquences surreprésentées pour une librairie n'est pas significatif dans notre cas donc la distribution en séquences différentes est homogène. Aucun changement observé après retrait des adaptateurs.

Adapter content

Sert à contrôler qu'il n'y a pas eu de contamination sur les adaptateurs utilisés. Ici le pourcentage de contamination n'est pas significatif.

Status checks

FastQC évalue chaque partie des analyses. Quand les résultats sont ceux attendus alors c'est vert sinon c'est jaune ou rouge. Pour nos expériences, la zone rouge indique que la distribution en bases n'est pas satisfaisante (per base sequence content). Le niveau de duplication(Sequence duplication)(zone de couleur jaune) est moyen mais acceptable. Après trimming des adaptateurs, il y a aussi une distribution de la taille des reads qui n'est pas très homogène.

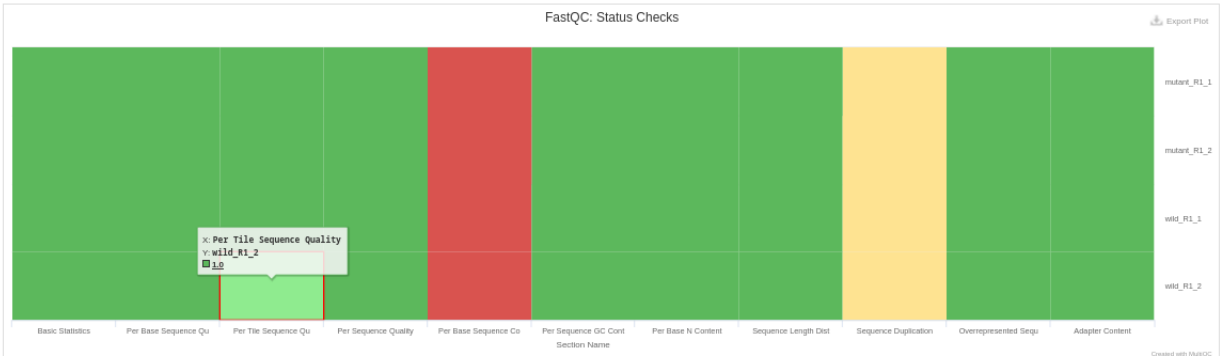

Figure 41: **FastQC**:Status Checks

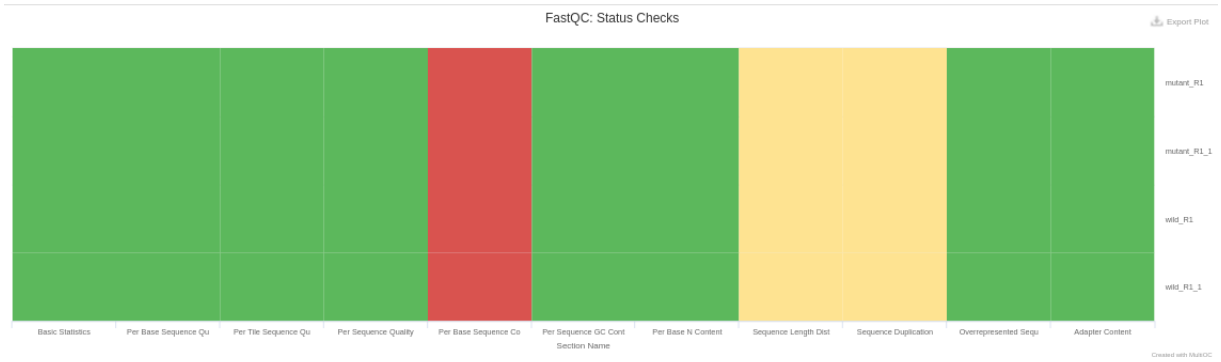

Figure 42: **FastQC**:Status Checks after trimming

2.3.11 Cutadapt

Cutadapt est un outil qui permet de nettoyer les séquences comme les queues polyA et les linkers/adaptateurs qui ne sont pas nécessaires pour la suite des analyses bio-informatiques et statistiques. Le barplot filtered reads montre le nombre de reads single-end/nombre de reads pair-end qui ont été enlevés par cutadapt. Il indique avec les barres bleus que toutes les séquences ont subi un filtrage.

Le deuxième graphique montre le nombre de reads en fonction de la taille des adaptateurs enlevés. Dans notre cas la majorité des adaptateurs enlevés sont de taille inférieure à 5pb.

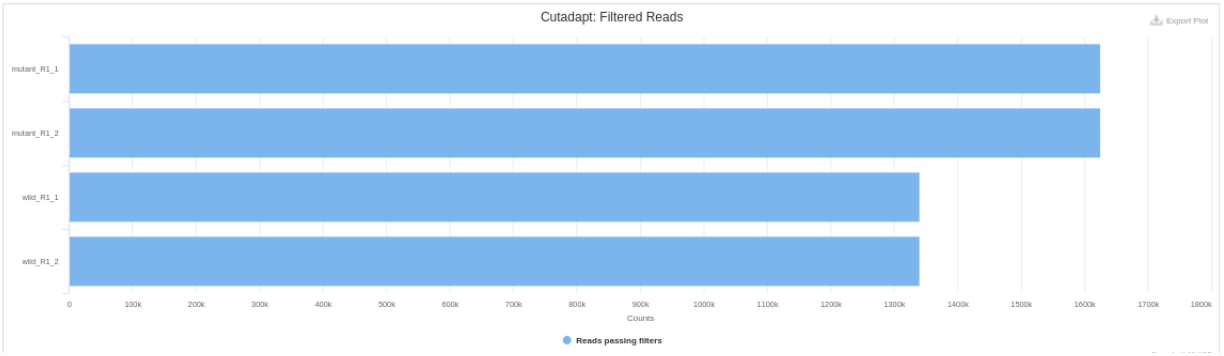

Figure 43: Filtered Reads

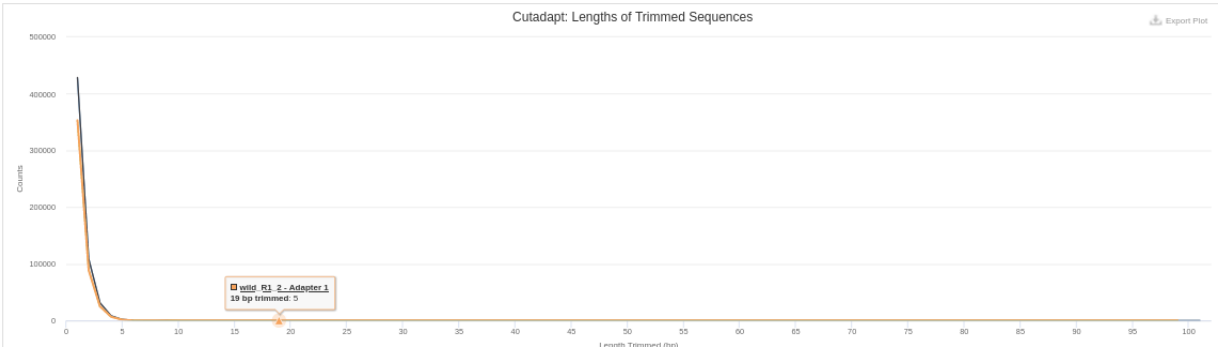

Figure 44: Lengths of Trimmed Sequences

2.3.12 nfcore/rnaseq Software Versions

Sont indiqués ici tous les logiciels utilisés pour le pipeline nfcore-renaseq ainsi que leurs versions.

bedtools	2.29.2
deseq2	1.28.0
dupradar	1.18.0
fastqc	0.11.9
nextflow	21.04.1
nf-core/rnaseq	3.0
picard	2.23.9
preseq	2.0.3
qualimap	2.2.2-dev
rsem	1.3.1
rseqc	3.0.1
samtools	1.10
stringtie	2.1.4
subread	2.0.1
trimgalore	0.6.6
ucsc	377

2.3.13 nfcore/rnaseq Workflow Summary

Une partie du summary:

Core Nextflow options

revision	3.0
runName	kickass_carlsson
containerEngine	singularity
launchDir	/work/dahlia/TP
workDir	/work/dahlia/TP/work
projectDir	/home/dahlia/.nextflow/assets/nf-core/rnaseq
userName	dahlia
profile	genotoul
configFiles	/home/dahlia/.nextflow/assets/nf-core/rnaseq/nextflow.co

Input/output options

input	/home/dahlia/work/TP/inputs.csv

Reference genome options

fasta	/home/dahlia/work/TP/genome/ITAG2.3_genomic_Ch6.fasta
gtf	/home/dahlia/work/TP/annotation/ITAG2.3_genomic_Ch6.gtf
save_reference	true
igenomes_ignore	true

Alignment options

aligner	star_rsem

Institutional config options

config_profile_descri...	The Genotoul cluster profile
config_profile_contact	support.bioinfo.genotoul@inra.fr

2.4 trimgalore

Ce répertoire contient des fichiers issus du traitement par le script Trimgalore. Il peut être définit comme un script wrapper c'est à dire qu'il regroupe plusieurs logiciels (cutadapt et fastq) pour effectuer un contrôle qualité des données. Trimgalore permet de retirer les reads de mauvaise qualité(phred score),les adaptateurs, ainsi que les reads avec une taille inférieure à un seuil. Ce script produit des fichiers fastq qui sont réanalysés par FastQC. Ainsi on retrouve un sous-répertoire fastqc.

2.5 pipeline info

Il contient des fichiers avec des informations sur le déroulement du pipeline. Le fichier **execution report** contient des graphiques qui nous informent sur la mémoire utilisée par chaque outil du pipeline, la durée pour chaque outil, la CPU ainsi que des informations pour chaque tâche séparée du pipeline.

Le fichier **execution timeline** liste les différentes tâches par ordre chronologique.

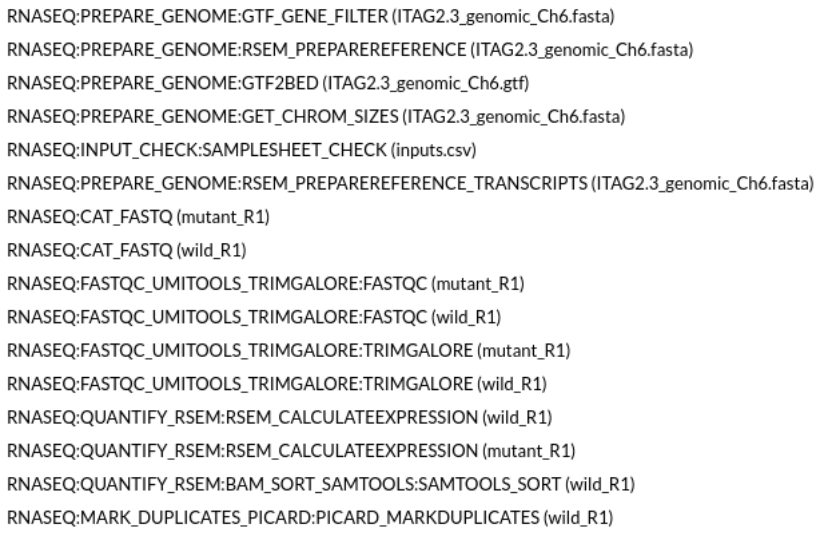

Figure 45: partie du fichier execution timeline

On retrouve aussi un fichier format csv qui liste les logiciels et leurs versions, le fichier des entrées (inputs) et enfin le graphe orienté acyclique (dag) du pipeline nextflow rnaseq.

2.6 star rsem

Dans le répertoire rsem on retrouve des fichiers bam qui sont des fichiers sam compressés. Les fichiers sam nous montrent les reads alignés sur le génome de référence.

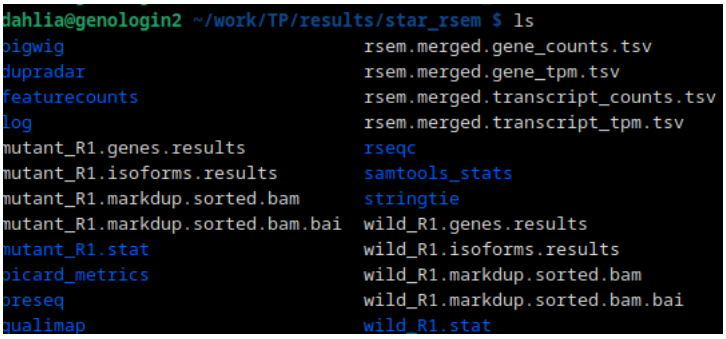

Figure 46: les fichiers et sous répertoires

2.6.1 bigwig

Ce répertoire contient des fichiers compressés, indexés et en format binaire pour des analyses de séquençage RNA-seq comme la profondeur des séquences.

2.6.2 dupradar

Il contient différents graphiques issues de l'outil dupradar qui évalue les niveaux de duplication dans nos données RNA-seq.

Les scatters plot et les boxplot montrent que plus les gènes sont exprimés et plus le niveau de duplication est élevé. Cela signifie que nos échantillons sont de bonne qualité. Les histogrammes montrent que la distribution des reads est normale. Les rpk sont obtenus en divisant le nombre de reads correspondant à un gène par la taille du gène exprimé en kilobases afin de pouvoir comparer deux librairies différentes dans notre cas celle issue du mutant et celle issue du sauvage.

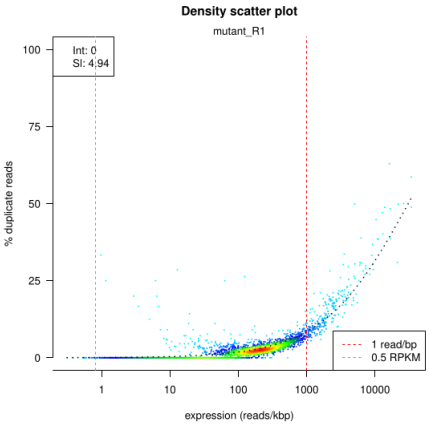

Figure 47: échantillon mutant

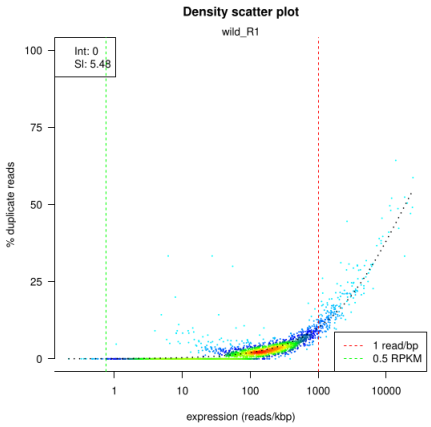

Figure 48: échantillon sauvage

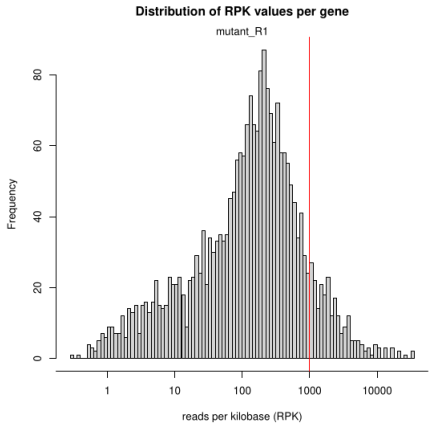

Figure 49: échantillon mutant

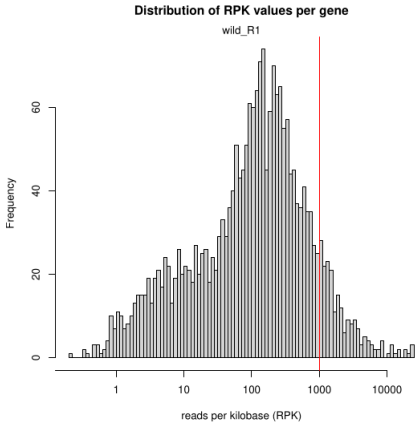

Figure 50: échantillon sauvage

2.6.3 rseqc

Les fichiers de sortie sont utilisés par MultiQC Contient les sous-répertoires :

- bam stat

  Résume les statistiques d'alignement des reads

- inner distance

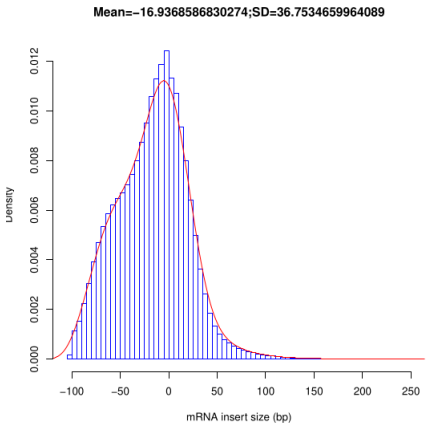

Figure 51: échantillon mutant

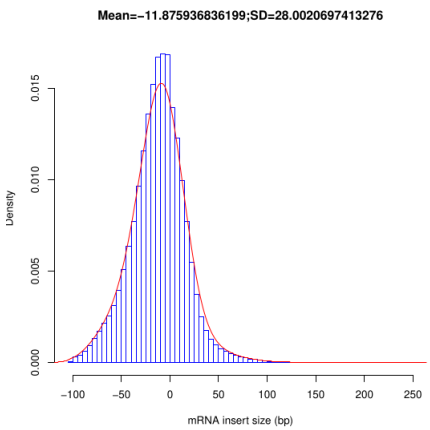

Figure 52: échantillon sauvage

  La distance moyenne entre reads contigus est de 16pb en moyenne pour le mutant et de 11pb pour le sauvage.

- read duplication

- infer experiment
- junction annotation
- junction saturation
- read distribution

2.6.4 featurecounts

Pour présenter les annotations des séquences du RNA-seq. Featurecounts est un outil qui effectue un contrôle qualité supplémentaire pour s'assurer qu'il y a pas d'annotations surreprésentés et donc il n y a pas eu de contamination.

2.6.5 stringtie

stringtie permet d'aligner efficacement les reads sur le génome. Les sorties de ce logiciel sont utilisées par DESeq2 qui estime l'expression différentielle des gènes. Ce répertoire contient des fichiers gtf pour chacun de nos 2 échantillons (mutant et sauvage).Ce sont les séquences annotées.

2.6.6 picard metrics

Dans les fichiers de ce répertoire il y a différentes informations concernant l'alignement (mapping) des séquences ainsi que le nombre de séquences non alignées.

2.6.7 preseq

Contient des fichiers de sortie de la librairie preseq. Cet outil permet de prédire et d'estimer la complexité d'une librairie issue du séquençage. Normalement il ne faut pas qu'il y ait des séquences surreprésentées dans une librairie.

2.6.8 qualimap

Contient un rapport html pour chaque échantillon qui décrit les résultats du contrôle qualité effectué pour l'alignement des séquences. Ce contrôle est nécessaire pour détecter des éventuels biais dues au séquençage.

3 RNA-seq avec des nouvelles données

Pour les paired end reads:

Il faut télécharger sra-toolkit:
sudo apt install sra-toolkit
Après sur le terminal depuis le répertoire qui stockera les données:
prefetch SRR2045415
Pour avoir les 2 réplicats dans des fichiers séparés:
fasterq-dump SRR2045415 –split-files –skip-technical

Ces 2 fichiers fastq il faut les compresser avec gzip
Ensuite il faut copier ces fichiers vers le compte genologin:
scp SRR2045415.fastq.gz dahlia@genologin.toulouse.inra.fr:/home/dahlia/work/projet

Pour les single-end reads:
vdb-config –prefetch-to-cwd
Ensuite télécharger:
prefetch SSR2045415
fasterq-dump SRR2045415

Le génome de référence et les annotations associées:
Genome de NCBI

4 Références

https://nf-co.re/docs/usage/troubleshooting
https://nf-co.re/rnaseq/2.0/docs/usage#running-the-pipeline
https://bioinfo.genotoul.fr/index.php/faq
Interprétation des sorties de fastqc: https://documents.migale.inrae.fr/posts/tutorials/illumina-qc/
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help
Le logiciel rsem: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323
le fichier bed: https://fr.wikipedia.org/wiki/BED_(format_de_fichier)
https://genotoul-bioinfo.pages.mia.inra.fr/training-rnaseq
https://nf-co.re/rnaseq/3.12.0/docs/output
https://hbctraining.github.io/Intro-to-rnaseq-fasrc-salmon-flipped/lessons/11_multiQC.html
https://github.com/nf-core/rnaseq/blob/master/docs/output.md#multiqc
https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
https://www.iame-research.center/wp-content/uploads/2022/05/club_bionfo_QC.pdf
https://gatk.broadinstitute.org/hc/en-us/articles/360036834611-MarkDuplicates-Picard-
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4029031
https://www.nature.com/articles/nbt.3122