Coraline SIPRA
Master BBS

Rapport Nextflow

2022-2023

Sommaire:

Commandes :

Seff :

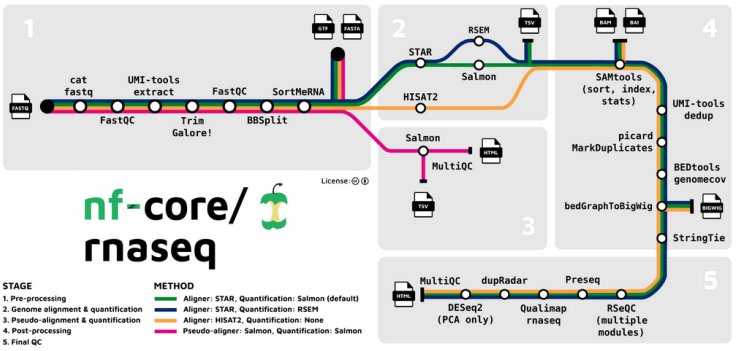
seff <job id>
La commande seff fournit des statistiques relatives à l'efficacité de l'utilisation des ressources du job terminé (Job id,Etat du job, Cluster, Utilisateur, CPU, Mémoire)

Resume :

<commande> -resume
Permet lorsque l'on relance un Job de reprendre là où il s'est arrêté précédemment car les étapes précédentes ont été mises en cache.

Répertoires:

Ces répertoires contiennent des fichiers issus des différents logiciels utilisés tout au long du pipeline.



Répertoire Fastqc:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/fastqc $ ls
CONTROL_REP1_1_fastqc.html  CONTROL_REP1_1_fastqc.zip  CONTROL_REP1_2_fastqc.html  CONTROL_REP1_2_fastqc.zip
```

Ce dossier contient des fichiers relatifs à l'étape de prétraitement du pipeline.

FastQC permet de faire des contrôles de qualité sur les données de séquence provenant de séquençage à haut débit. Cela permet de montrer si les données (ici nos 4 échantillons) comportent des problèmes avant de faire d'autres analyses.

```
Archive:  CONTROL_REP1_1_fastqc.zip
   creating: CONTROL_REP1_1_fastqc/
   creating: CONTROL_REP1_1_fastqc/Icons/
   creating: CONTROL_REP1_1_fastqc/Images/
  inflating: CONTROL_REP1_1_fastqc/Icons/fastqc_icon.png
  inflating: CONTROL_REP1_1_fastqc/Icons/warning.png
  inflating: CONTROL_REP1_1_fastqc/Icons/error.png
  inflating: CONTROL_REP1_1_fastqc/Icons/tick.png
  inflating: CONTROL_REP1_1_fastqc/summary.txt
  inflating: CONTROL_REP1_1_fastqc/Images/per_base_quality.png
  inflating: CONTROL_REP1_1_fastqc/Images/per_tile_quality.png
  inflating: CONTROL_REP1_1_fastqc/Images/per_sequence_quality.png
  inflating: CONTROL_REP1_1_fastqc/Images/per_base_sequence_content.png
  inflating: CONTROL_REP1_1_fastqc/Images/per_sequence_gc_content.png
  inflating: CONTROL_REP1_1_fastqc/Images/per_base_n_content.png
  inflating: CONTROL_REP1_1_fastqc/Images/sequence_length_distribution.png
  inflating: CONTROL_REP1_1_fastqc/Images/duplication_levels.png
  inflating: CONTROL_REP1_1_fastqc/Images/adapter_content.png
  inflating: CONTROL_REP1_1_fastqc/fastqc_report.html
  inflating: CONTROL_REP1_1_fastqc/fastqc_data.txt
  inflating: CONTROL_REP1_1_fastqc/fastqc.fo
```

On retrouve deux types de fichiers ici. Un fichier de rapport en zip qui contient un ensemble de fichiers qui sont les résultats des analyses effectuées et qui présente notamment un fichier **fastqc_report.html** qui renvoie aux fichiers d'analyse du rapport.

L'autre fichier disponible est le rapport HTML des analyses effectuées.

Le rapport HTML:

Différents paramètres sont calculés:

Per Base Sequence Quality: Montre des valeurs de qualité pour chaque position des bases dans le fichier. L'axe des Y sur le graphique montre les scores de qualité. Plus le score est élevé, meilleur est le séquençage de la base. La qualité de séquençage diminue au fur et à mesure que l'on progresse, il est donc courant de voir des bases dans la zone orange vers la fin.

Per Sequence Quality Scores: Permet de voir si un sous-ensemble de nos séquences ont des valeurs de qualité faibles (en général un petit pourcentage du total de la séquence).

Per Base Sequence Content: Proportion de chaque base de l'ADN dans le fichier. La quantité relative de chaque base doit refléter la quantité globale des bases dans le génome. Si c'est déséquilibré les unes par rapport aux autres cela peut montrer un problème de séquençage ou une contamination.

Per Base GC Content: Contenu de GC à chaque position de base. La quantité relative doit refléter la quantité globale des bases dans le génome. Si c'est déséquilibré cela peut montrer un problème de séquençage ou une contamination.

Per sequence GC content: Mesure le contenu GC sur toute la longueur de la séquence et la compare à une distribution normale modélisée de la teneur en GC. Une distribution de forme inhabituelle (pas normale) pourrait indiquer une contamination ou des sous-ensembles biaisés.

Per Base N content: Trace le pourcentage de base à chaque position pour laquelle un N était placé. Ne doit pas dépasser quelques pourcentages sinon cela suggère qu'il n'a pas était possible d'interpréter les données suffisamment bien pour être valides.

Sequence length distribution: Montre la distribution des tailles de fragment dans le fichier.

Duplicate Sequence: Montre le nombre relatif de séquences avec différents degrés de duplication. Un haut niveau de duplication peut indiquer une sorte de biais d'enrichissement (par exemple PCR sur amplification).

Overrepresented Sequences: Répertorie toutes les séquences qui représentent plus de 0,1 % du total.

Adapteur Content: Analyse de tous les Kmers pour trouver ceux qui n'ont pas une couverture uniforme sur toute la durée des lectures. La présence de séquences sur-représentées (telles que des dimères adaptateurs) entraînera la domination du tracé par les Kmers que contiennent ces séquences.

Répertoire Trim galore:

Ce dossier contient des fichiers relatifs à l'étape de prétraitement du pipeline.

Ce répertoire contient les fichiers de séquence obtenus après être passé par Trim Galore.

```
This is cutadapt 3.4 with Python 3.9.6
Command line parameters: -j 4 -e 0.1 -q 20 -O 1 -a AGATCGGAAGAGC CONTROL_REP2_1.fastq.gz
Processing reads on 4 cores in single-end mode ...
Finished in 13.70 s (8 µs/read; 7.11 M reads/minute).

=== Summary ===

Total reads processed:               1,624,613
Reads with adapters:                   578,951 (35.6%)
Reads written (passing filters):     1,624,613 (100.0%)

Total basepairs processed:   164,085,913 bp
Quality-trimmed:               4,900,627 bp (3.0%)
Total written (filtered):    158,377,334 bp (96.5%)
```

Trim Galore est un script qui permet de découper les adaptateurs ou de supprimer les positions de méthylation biaisées pour les fichiers de séquence RRBS ainsi que faire du contrôle de la qualité. Il peut utiliser des adaptateurs standard ou que l'on a préalablement spécifiés. Ce script prend en entrée et produit des fichiers Fastq qui peuvent ensuite être repassés dans FastQC pour une nouvelle analyse: c'est le répertoire **fastqc** que l'on retrouve ici.

Répertoire Genome:

Ce dossier contient des fichiers relatifs à l'étape d'alignement et de quantification.

Ce répertoire comporte tous les fichiers de référence du génome (gtf,fai…etc) mais aussi plusieurs sous dossiers.

STAR:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/genome/index/star $ ls
chrLength.txt  chrNameLength.txt  chrName.txt  chrStart.txt  exonGeTrInfo.tab  exonInfo.tab  geneInfo.tab  Genome  genomeParameters.txt  SA  SAindex  sjdbInfo.txt  sjdbList.fromGTF.out.tab
```

STAR est un aligneur de lecture conçu pour la cartographie sensible à l'épissage des données de séquençage d'ARN.
L'alignement multiple et la quantification peuvent se faire dans le sens STAR -> Rsem ou STAR -> Salmon.

On retrouve plusieurs fichiers tels que des informations sur les jonctions annoté **sjd*,** des informations sur les exons et génome ainsi que des fichiers nécessaires au fonctionnement de STAR (index…etc).

Rsem:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/genome/rsem $ ls
genome.chrlist  genome.grp  genome.idx.fa  genome.n2g.idx.fa  genome.seq  genome.ti  genome.transcripts.fa  ITAG2.3_genomic_Ch6.fasta
```

Un logiciel permettant d'estimer les niveaux d'expression des gènes et des isoformes à partir des données RNA-Seq. Il fournit des estimations de la moyenne postérieure et de l'intervalle de confiance à 95 % pour les niveaux d'expression.

Répertoire star salmon:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/star_salmon $ ls
bigwig                          CONTROL_REP2.markdup.sorted.bam      log                 salmon.merged.gene_counts_length_scaled.rds  salmon.merged.gene_counts.tsv        salmon_tx2gene.tsv
CONTROL_REP1                    CONTROL_REP2.markdup.sorted.bam.bai  picard_metrics      salmon.merged.gene_counts_length_scaled.tsv  salmon.merged.gene_tpm.tsv           samtools_stats
CONTROL_REP1.markdup.sorted.bam deseq2_qc                            preseq              salmon.merged.gene_counts.rds                salmon.merged.transcript_counts.rds  stringtie
CONTROL_REP1.markdup.sorted.bam.bai dupradar                         qualimap            salmon.merged.gene_counts_scaled.rds         salmon.merged.transcript_counts.tsv
CONTROL_REP2                    featurecounts                        rseqc               salmon.merged.gene_counts_scaled.tsv         salmon.merged.transcript_tpm.tsv
```

Ce répertoire contient des fichiers issus de différents logiciels et qui sont utilisé pour l'étape de post traitement et celle d'analyse de qualité du pipeline (Rseq,Preseq,Qualimap, dupRADAR,Deseq,MultiQC)

Salmon:

```
salmon.merged.gene_counts_length_scaled.rds  salmon.merged.gene_counts.tsv
salmon.merged.gene_counts_length_scaled.tsv  salmon.merged.gene_tpm.tsv
salmon.merged.gene_counts.rds                salmon.merged.transcript_counts.rds
salmon.merged.gene_counts_scaled.rds         salmon.merged.transcript_counts.tsv
salmon.merged.gene_counts_scaled.tsv         salmon.merged.transcript_tpm.tsv
```

C'est un logiciel qui permet de quantifier l'expression de transcrits à l'aide de données RNA-seq. Il se réalise en deux étapes: l'indexation et la quantification. Il suffit de lui fournir les fichiers obtenus en sortie du logiciel STAR: un fichier FASTA du transcriptome et un fichier .sam ou .bam contenant un ensemble d'alignements.

Les dossiers **Controle Rep 1** et **Controle Rep 2** contiennent aussi des fichiers relatif à Salmon comme:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/star_salmon/CONTROL_REP1 $ ls
aux_info  cmd_info.json  libParams  logs  quant.genes.sf  quant.sf
```

- aux_info: Versions et le nombre de lectures mappées.
- cmd_info: Informations sur la commande, la version et les options de quantification de Salmon.
- libParams : Contient le fichier **flenDist.txt** pour la distribution de la longueur des fragments.
- logs : Contient le fichier **salmon_quant.log** donnant un enregistrement de la quantification de Salmon.
- quant.genes.sf : Quantification au niveau du gène de l'échantillon, y compris la longueur des caractéristiques, la longueur effective, le TPM et le nombre de lectures.
- quant.sf : Quantification de l'échantillon au niveau de la transcription, y compris la longueur des caractéristiques, la longueur effective, le TPM et le nombre de lectures.

Samtools stats:

```
CONTROL_REP2.markdup.sorted.bam.flagstat  CONTROL_REP2.markdup.sorted.bam.stats  CONTROL_REP2.sorted.bam.idxstats
CONTROL_REP2.markdup.sorted.bam.idxstats  CONTROL_REP2.sorted.bam.flagstat       CONTROL_REP2.sorted.bam.stats
```

Un ensemble d'outils qui manipulent les alignements dans les formats SAM (Sequence Alignment/Map), BAM et CRAM. Il permet de convertir entre différents formats et effectue le tri, la fusion et l'indexation, et peut même récupérer rapidement les lectures dans toutes les régions. On peut aussi faire des statistiques.

Dans ce dossier on retrouve les séquences indexées et triées mais aussi les statistiques réalisées dessus.

Samtools flagstat compte le nombre d'alignements pour chaque type FLAG (catégorie). Chaque catégorie de la sortie est divisée en QC réussi et QC échoué.

Samtools idxstats fait les statistiques du fichier d'index correspondant au fichier d'entrée. La sortie est délimitée par des tabulations, chaque ligne étant constituée du nom de la séquence de référence, de la longueur de la séquence, du nombre de segments de lecture mappés et du nombre de segments de lecture non mappés.

Picard metrics:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/star_salmon/picard_metrics $ ls
CONTROL_REP1.markdup.sorted.MarkDuplicates.metrics.txt  CONTROL_REP2.markdup.sorted.MarkDuplicates.metrics.txt
```

Picard est un ensemble d'outils permettant de manipuler des données et des formats de séquençage à haut débit. Ici on a utilisé nos fichiers BAM issu de samtools qu'on a analysé avec l'option **DuplicationMetrics** consistant à calculer différents paramètres:

```
--showHidden false --USE_JDK_DEFLATER false --USE_JDK_INFLATER false
## htsjdk.samtools.metrics.StringHeader
# Started on: Fri Sep 30 10:33:12 GMT 2022

## METRICS CLASS        picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED     SECONDARY_OR_SUPPLEMENTARY_RDS  UNMAPPED_READS  UNPAIRED_READ_DUPLICATES
Unknown Library 617     1321355 46649   0       460     241599  631     0.182973        3164270
```

Library: La bibliothèque sur laquelle le marquage dupliqué a été effectué.

Unpaired reads examined: Le nombre de lectures mappées examinées qui n'avaient pas de paires mappées.

Read pairs examined: Le nombre de paires de lecture mappées examinées.

Secondary or supplementary rds: Le nombre de lectures secondaires ou supplémentaires.

Unmapped reads: Le nombre total de lectures non mappées examinées.

Unpaired read duplicates: Le nombre de fragments marqués comme doublons.

Read pair duplicated: Le nombre de paires de lecture qui ont été marquées comme doublons.

Read pair optical duplicates: Le nombre de duplications de paires de lecture causées par la duplication optique.

Percent duplication: La fraction de la séquence mappée qui est marquée comme doublon.

Estimated library size: Le nombre estimé de molécules uniques dans la bibliothèque en fonction de la duplication PE.

Bigwig:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/star_salmon/bigwig $ ls
CONTROL_REP1.forward.bigWig  CONTROL_REP1.reverse.bigWig  CONTROL_REP2.forward.bigWig  CONTROL_REP2.reverse.bigWig
```

Contient les fichiers au format bigwit obtenu par conversion à partir des fichiers bedGrapah. Le format bigWig est utilisé pour les données denses qui seront affichées dans GenomeBrowser sous forme de graphiques.

Les fichiers bedGraph ont été obtenus à l'étape précédente du pipeline: On a d'abord fusionner nos fichiers BAM grâce à **bedtools** (combine des entités qui se chevauchent dans un fichier d'intervalles en une seule entité qui couvre toutes les entités combinées.). Ensuite avec **bedtools genomecov** qui calcul la couverture au format BEDGRAPH uniquement pour les régions du génome avec ou sans couverture.

Stringtie:

```
clematite@genologin2 /work/clematite/Nextflowtp/results/star_salmon/stringtie $ ls
CONTROL_REP1.ballgown  CONTROL_REP1.coverage.gtf  CONTROL_REP1.gene.abundance.txt  CONTROL_REP1.transcripts.gtf
```

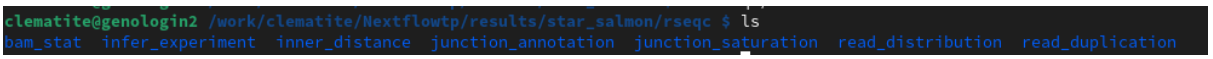
C'est un assembleur rapide et très efficace d'alignements RNA-Seq en transcrits potentiels. Il peut prendre en entrée des alignements de lectures courtes ou longues. Afin d'identifier les gènes différentiellement exprimés entre les expériences, la sortie du logiciel sera utilisée sur d'autres logiciels tels que DESeq2. Il produit les différents fichiers que l'on retrouve dans le répertoire:

- Un fichier GTF contenant les transcriptions assemblées .
- L'abondance de gènes au format délimité par des tabulations .
- Transcriptions entièrement couvertes qui correspondent à l'annotation de référence, au format GTF .
- Fichiers (tableaux) requis en entrée de Ballgown, qui les utilise pour estimer l'expression différentielle .
- En mode fusion, un fichier GTF fusionné à partir d'un ensemble de fichiers GTF .

Featurecounts:

Est un programme qui compte le nombre de lectures correspondant aux caractéristiques telles que les gènes, les exons, les promoteurs, les corps de gènes, les bacs génomiques et les emplacements chromosomiques.Il peut aussi servir à compter les chevauchements avec différentes classes de caractéristiques génomiques que l'on retrouve dans le fichier **TSV** et le fichier **summary** qui contient les statistique sur les lectures. Il fournit aussi un fichier **Biotype** qui permet de vérifier quelles caractéristiques sont les plus abondantes dans l'échantillon et pour mettre en évidence les problèmes potentiels tels que la contamination par l'ARNr.

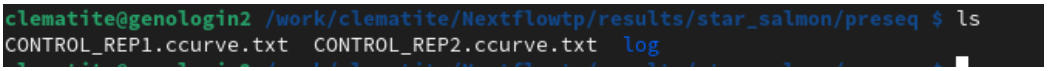
Rseq:



Un ensemble de scripts conçus pour évaluer la qualité des données RNA-seq dont les fichiers de sortie seront utilisés dans le rapport MultiQC. Ce dossier comporte plusieurs sous-dossier qui sont les résultats des scripts rseq citons:

- bam stat: Résumé des statistiques de mappage d'un fichier BAM ou SAM.
- infer experiment: Devine comment les lectures ont été bloquées pour les données ARN-seq spécifiques à un brin, en comparant le blocage des lectures avec la « stabilité des transcriptions ».
- inner distant: calcul la distance interne entre les paires de reads.
- junction_annotation: Un fichier d'alignement au format BAM ou SAM et un modèle de gène de référence au format BED permet de comparer les jonctions d'épissage détectées au modèle de gène de référence.
- junction saturation: Vérifier si la profondeur de séquençage actuelle est suffisamment profonde pour effectuer des analyses d'épissage alternatives.
- read distribution: Un fichier BAM/SAM et un modèle de gène de référence ce qui permettra de voir comment les lectures cartographiées ont été distribuées sur les caractéristiques du génome (comme l'exon CDS, l'exon 5'UTR, l'exon 3'UTR, l'intron, les régions intergéniques).
- read duplication: Déterminer le taux de duplication des lectures.

Preseq:



Le package vise à prédire et à estimer la complexité d'une bibliothèque de séquençage génomique(estimer le nombre de lectures redondantes à partir d'une profondeur de séquençage donnée et le nombre attendu d'un séquençage supplémentaire à l'aide d'une expérience de séquençage initiale). Les résultats sont utilisés dans MultiQC.
On a deux types de fichier ici, la sortie d'erreur standard de la commande: **log** et les fichiers **curve** qui calcule le rendement attendu des lectures distinctes pour les expériences plus petites que l'expérience d'entrée dans un fichier .bed ou .bam par rééchantillonnage**.**

Qualimap:

Une application qui facilite le contrôle qualité des alignements séquencés. Il examine les données d'alignement de séquençage en fonction des caractéristiques des lectures cartographiées et de leurs propriétés génomiques. Il fournit une vue globale des données qui aide à détecter les biais dans le séquençage et/ou la cartographie des données et facilite la prise de décision pour une analyse plus approfondie.

Le dossier contient un rapport HTML des figures d'analyse de qualité, les résultats en version texte ainsi qu'un ensemble de fichiers dans la partie CSS nécessaires au bon fonctionnement du rapport.

Dupradar:

Une bibliothèque Bioconductor qui donne différentes données et graphiques qui relient le taux de duplication aux niveaux d'expression génique afin d'identifier les expériences avec une technique de duplication élevée.

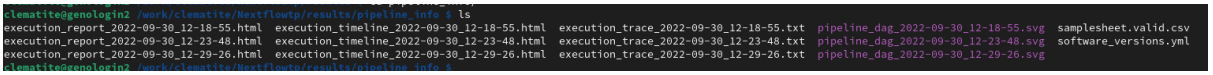
Dans le répertoire on peut retrouver des histogrammes, des scatter plot, des box plot mais aussi des métriques par gène dans **gene data.**

Répertoire DESq2:

Un package pour effectuer une analyse d'expression différentielle pour les ensembles de données RNA-seq. Dans le script du pipeline on a normalisé le nombre de lectures sur tous les échantillons fournis afin de créer un tracé PCA et une heatmap groupée montrant les distances euclidiennes par paires entre les échantillons de l'expérience. On a différents dossiers:

- plots.pdf : Fichier contenant les graphiques PCA et de clustering hiérarchique.
- dds.RData : fichier contenant l'objet R DESeqDataSet généré par DESeq2, avec un test rlog ou vst stockant les données stabilisées par la variance.
- pca.vals.txt : Matrice des valeurs des 2 premières composantes principales.
- sample.dists.txt : Exemple de matrice de distance.
- R_sessionInfo.log : fichier contenant des informations sur R, le système d'exploitation et les packages attachés ou chargés.
- size_factor: Fichiers contenant DESeq2 sizeFactors par échantillon.

Répertoire Pipeline info:



Un ensemble d'informations relatives au déroulement du pipeline que ce soit les commandes utilisées, les durées d'exécution, les erreurs..etc. Plusieurs fichiers:

- execution_report.html, execution_timeline.html: Rapports générés par Nextflow
- pipeline_report.html,software_versions.yml: Rapports générés par le pipeline.
- samplesheet.valid.csv: Fichiers d'échantillons formatés utilisés comme entrée du pipeline

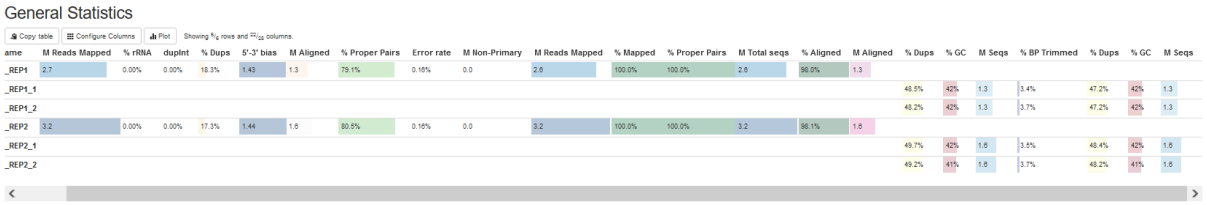
Répertoire MutliQC:

Outil de visualisation qui génère un seul rapport HTML résumant tous les résultats du pipeline. Les résultats générés par MultiQC rassemblent le QC du pipeline à partir d'outils pris en charge, c'est-à-dire FastQC, Cutadapt, SortMeRNA, STAR, RSEM, HISAT2, Salmon, SAMtools, Picard, RSeQC, Qualimap, Preseq et featureCounts, dupRADAR,DESq.

Ce dossier contient le rapport ainsi qu'un dossier **multiqc_data** contenant les statistiques analysées des différents outils utilisés dans le pipeline.

Analyse du rapport MultiQC:

General Statistics:

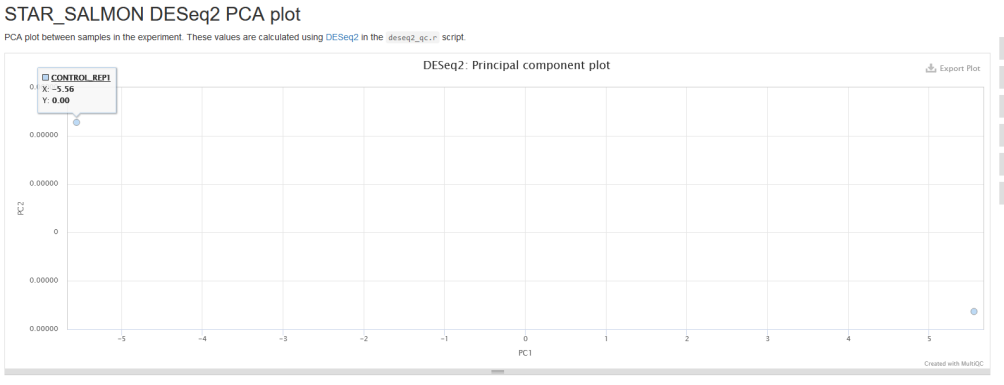


Cela montre un aperçu des valeurs clés extraites de tous les modules. L'objectif du tableau est de rassembler les statistiques de chaque échantillon. Même si les données vont être plus détaillées par la suite, certaines peuvent être déjà intéressantes à observer:

- M seqs: nombre total de lectures brutes.
- %Aligned: pourcentage de lecture mappé unique (un bon échantillon à majoritairement des lectures uniques).
- %Dups: pourcentage de duplication (s'assurer qu'il n'y a pas un trop grand nombre de duplication).
- 5'-3' bias: regarder si nos données ont un biais (généralement, nous devrions explorer davantage nos données si nous avons des biais approchant 0,5 ou 2).
- %GC: s'assurer que les % de GC est semblable entre échantillons.

On a un bon pourcentage de lecture unique, un GC équilibré entre les séquences, il ne semble pas y avoir particulièrement beaucoup de biais et le taux de duplication est assez faible.

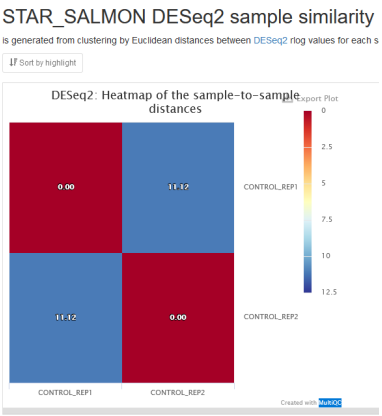
PCA plot:



L'analyse en composantes principales (ACP) est une technique de réduction de dimension et de visualisation qui est utilisée ici pour projeter le vecteur de données multivariées de chaque échantillon dans un tracé en deux dimensions. De sorte que la disposition spatiale des points dans le tracé reflète les données globales "similarité/dissimilarité" entre les échantillons.

On remarque une grande différence de variance entre le **controle rep 1 (WT)** et le **controle rep 2 (MT)**. Le mutant à une plus grande variance qui montre bien une différence d'expression entre le Mutant et le Wild type.

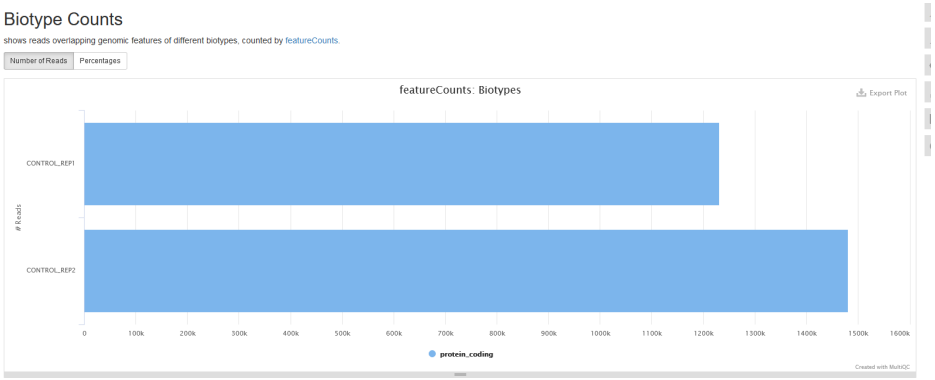
Heatmap:



Dans les heatmaps, les données sont affichées dans une grille où sont représentés les gènes et les échantillons. La couleur et l'intensité des cases sont utilisées pour représenter les changements de l'expression des gènes.

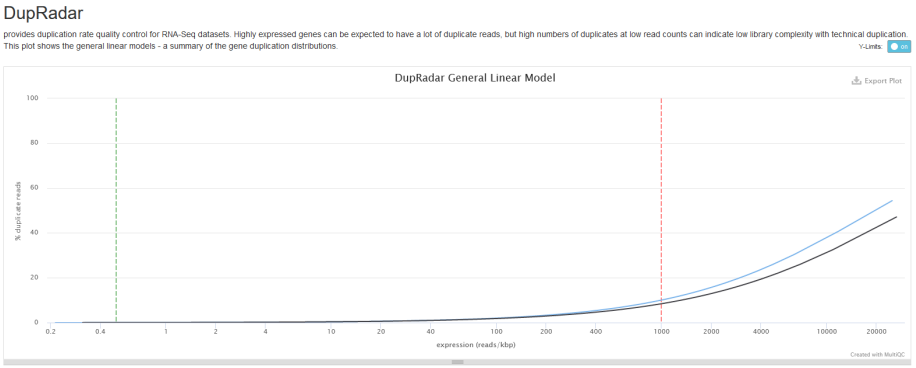
On confirme ce que nous avions vu plus tôt, il y a bien une différence d'expression et le Mutant présente une surexpression par rapport au Wild Type.

Biotype Count:



Permet de compter les chevauchements avec différents types de données. Cela donne une bonne idée de l'endroit où aboutissent les lectures alignées et peut montrer des problèmes potentiels tels que la contamination.
Il ne semble pas y avoir de problème de contamination (car 100%).

DupRadar linear model:

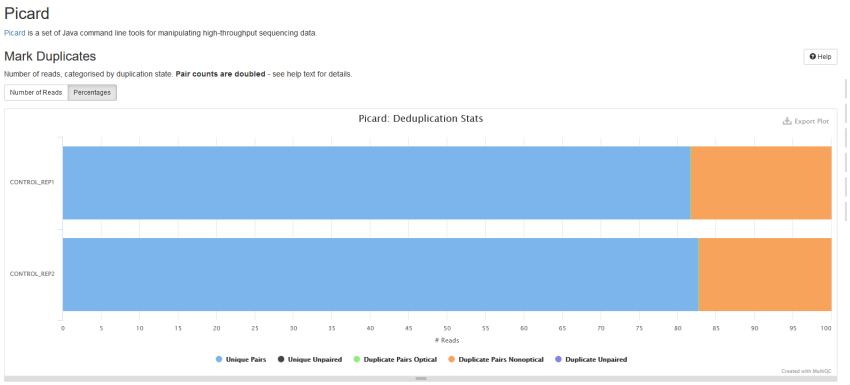


Le nombre de lectures par base attribuées à un gène dans un ensemble de données RNA-Seq idéal devrait être proportionnel à l'abondance de ses transcrits dans l'échantillon.

Ce plot représente le taux de duplication par rapport à l'expression pour chaque gène. Un bon échantillon avec peu de duplication ne montrera qu'un nombre élevé de doublons pour les gènes hautement exprimés.

Ici on voit bien sur le modèle linéaire que l'on a une forte expression seulement pour les gènes hautement exprimés, on a donc de bon échantillons.

Picard Mark Duplicates:

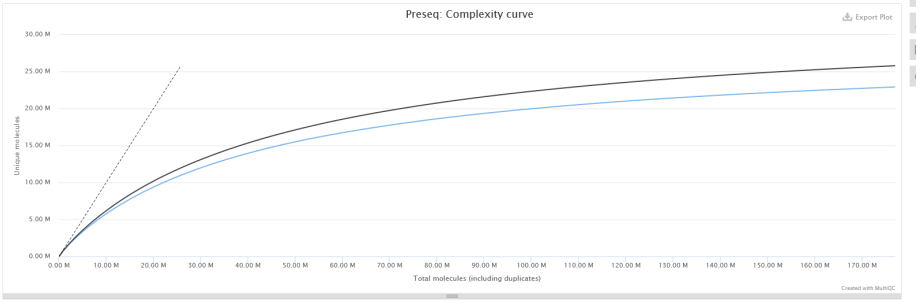


Permet d'identifier le niveau global de duplication dans vos échantillons. Les doublons optiques se produisent lorsqu'un cluster unique est reconnu comme deux clusters adjacents lors de la lecture. Les deux séquences ont des séquences très similaires. En général, si N bases sont identiques, elles sont classées comme doublons optiques.

Picard MarkDuplicates ne fait pas de distinction entre les différents types de duplication biologique ou PCR. Mais cela peut aider à identifier si un ensemble de données de séquençage a un problème avec les doublons optiques permettant ainsi de procéder avec prudence.

On voit bien que nos séquences présentent un peu de duplications optiques, il faut donc faire attention car ce ne sont pas de vraies duplication.

Preseq:

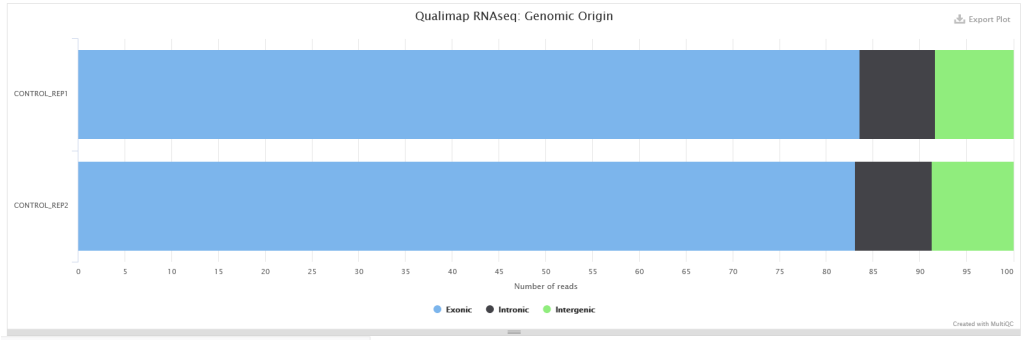


Permet d'estimer la complexité d'une bibliothèque de séquençage génomique, (nombre de lectures redondantes à partir d'une profondeur de séquençage donnée) On cherche à évaluer l'utilité d'un séquençage supplémentaire, optimiser la profondeur de séquençage ou pour cribler plusieurs bibliothèques afin d'éviter les échantillons de faible complexité. Une courbe peu profonde indique que la bibliothèque a atteint la saturation de la complexité et qu'un séquençage supplémentaire n'ajoutera probablement pas d'autres lectures uniques.

Ici on voit bien qu'on a pas encore atteint la profondeur de séquence maximale pour les deux échantillons. On peut donc optimiser encore le séquençage.

Qualimap:

Genomic origin of reads:

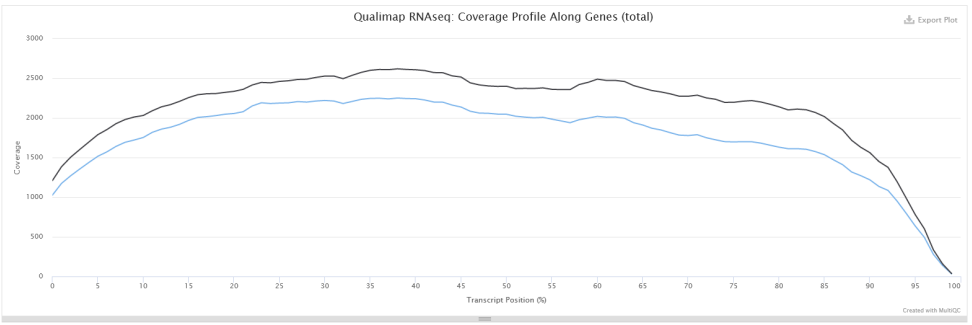


Rapporte combien d'alignements tombent dans des régions exoniques, introniques et intergéniques avec un certain nombre d'alignements introniques/intergéniques chevauchant des exons.

On s'attend à avoir une proportion de lectures cartographiées des régions exoniques > 60 % et à des taux de cartographie introniques plus faibles (15 à 30 %).
Un taux de cartographie intronique plus élevé est attendu pour l'élimination de l'ARNr par rapport à la sélection polyA. Les lectures introniques proviennent probablement de transcrits immatures qui comprennent soit des molécules de pré-ARNm pleine longueur, soit des transcrits naissants où l'ARN polymérase ne s'est pas encore attachée à l'extrémité 3 'du gène.
Une répartition à peu près égale des lectures cartographiant les régions introniques, exoniques et intergéniques suggère qu'il existe une contamination par l'ADN.

Sur nos échantillons on a une répartition correcte des proportions pour les différentes catégories.

Genome Coverage profile:



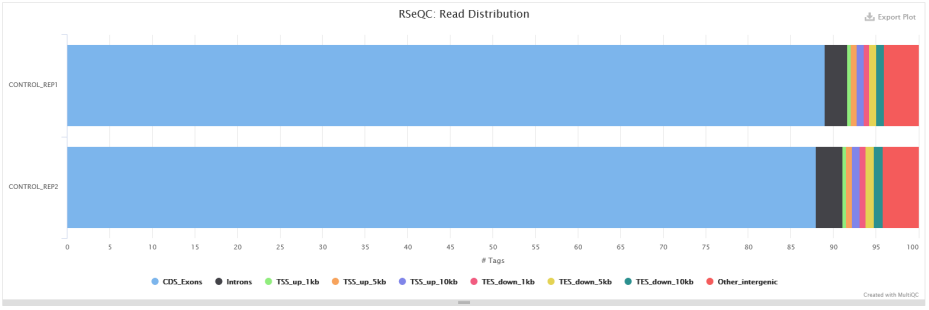
Le profil fournit des rapports entre la couverture moyenne dans la région 5', la région 3' et l'ensemble de la transcription.

Dans une expérience de séquençage parfaite, vous vous attendez à voir un rapport de biais 5'-3' élevé au milieu avec une faible couverture aux deux extrémités de la transcription. Cela suggérerait qu'aucun biais n'est présent.
Si les lectures s'accumulent principalement à l'extrémité 3 'des transcrits dans des échantillons cela pourrait indiquer une faible qualité d'ARN dans la matière de départ. Un biais vers l'extrémité 3' des gènes pourrait indiquer une dégradation de l'ARN.

On a des valeurs de couverture un peu élevées en 5', pour ce qui du biais 5'-3' et 3' cela semble normal. On a peut-être un ARN dégradé.

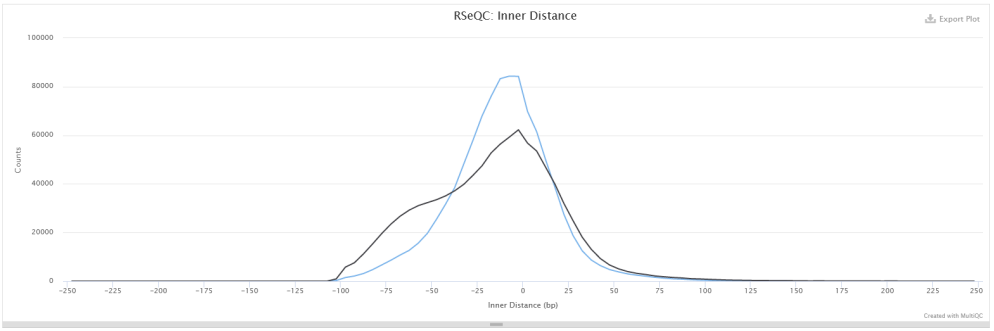
RSeQC:

Read distribution:



Calcule la répartition des lectures mappées sur les caractéristiques génomiques. Un bon résultat pour une expérience RNA-seq est généralement d'avoir autant de lectures exoniques que possible. Une grande quantité de lectures introniques peut indiquer une contamination de l'ADN dans l'échantillon (sauf si ARN total).
Notre séquence présente un très grand nombre de lecture exonique ce qui semble bon.

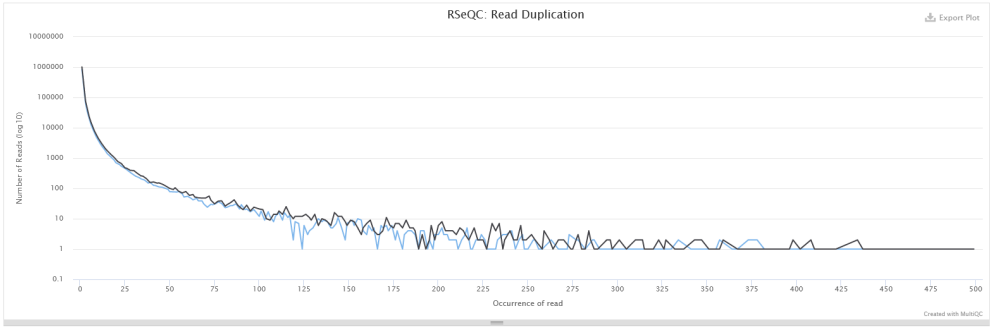
Inner distance:



Calcul la distance interne entre deux lectures appariées. Des distances intérieures très courtes sont souvent observées dans des échantillons anciens ou dégradés et les valeurs peuvent être négatives si les lectures se chevauchent systématiquement.

On remarque qu'une partie des lectures se chevauchent ainsi que des lectures négatives. Cela semble donc corrélé avec ce qui était observé sur le Genome Coverage de Qualimap, on a peut être une séquence dégradée.

Read duplication:

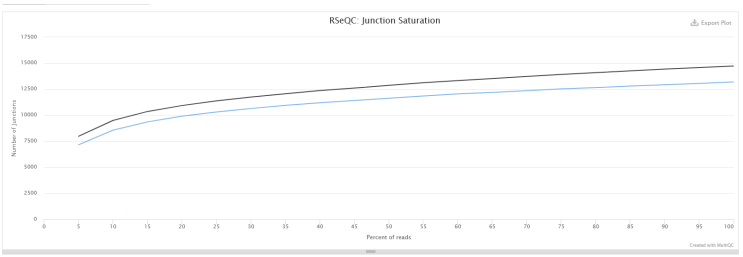


Montre le nombre de lectures (axe y) avec un nombre donné de doublons exacts (axe x). La plupart des lectures dans une bibliothèque RNA-seq doivent avoir un faible

nombre de doublons exacts. Les échantillons qui ont de nombreuses lectures avec de nombreux doublons (une grande zone sous la courbe) peuvent souffrir d'une duplication technique excessive. Le taux de duplication de lecture est très spécifique au type d'échantillon. Ainsi, cela dépend de la source d'origine de l'ARN séquencé. Le taux de duplication doit être aussi faible que possible mais la qualité de l'échantillon original influence les données.

Cela corrèle ce que nous avions vu précédemment avec Picard, nos échantillons on des duplication mais pas tant que ça.

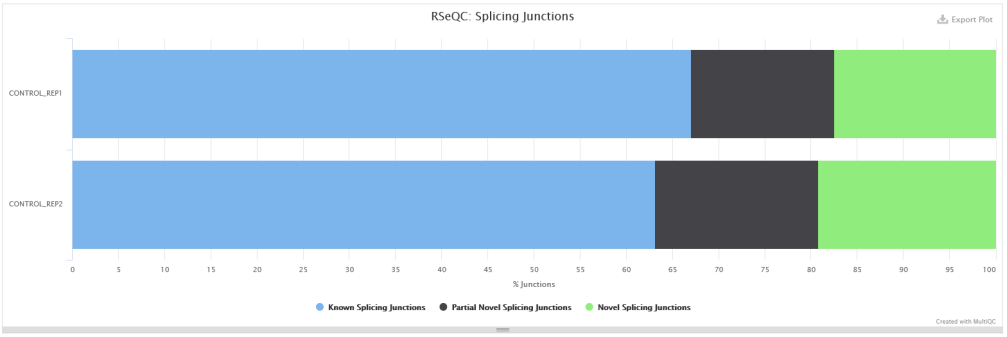
Junction saturation:



Montre le nombre de sites d'épissage détectés au niveau des données à différents niveaux de sous-échantillonnage. Un échantillon qui atteint un plateau avant d'atteindre 100 % des données indique que toutes les jonctions de la bibliothèque ont été détectées et qu'un séquençage supplémentaire ne produira pas plus d'observations. Un échantillon de bonne qualité devrait approcher un tel plateau.

Cela reprend ce que nous avions vu avec Preseq, nous n'avons pas de plateau on peut donc encore améliorer le séquençage.

Junction Annotation:

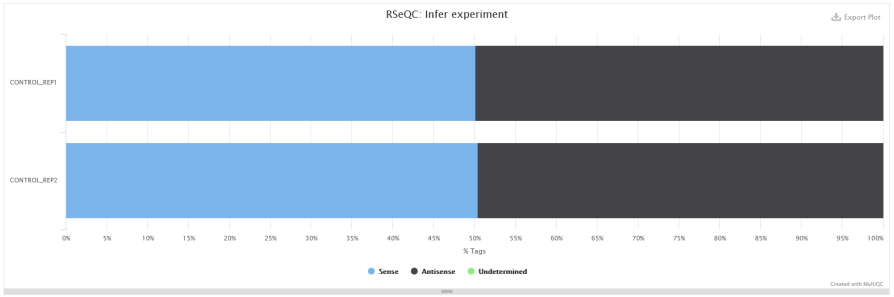


Compare les jonctions d'épissage détectées à un modèle de gène de référence. L'annotation d'épissage est effectuée à deux niveaux : niveau d'événement d'épissage et niveau de jonction d'épissage. Elle sont regroupées en 3 catégories:

- Known: la jonction fait partie du modèle de gène. Les deux sites d'épissage, le site d'épissage 5' (5'SS) et le site d'épissage 3' (3'SS) peuvent être annotés par un modèle de gène de référence.
- Novel: nouvelle jonction complète. Aucun des deux sites d'épissage ne peut être annoté par un modèle de gène
- Partial Novel : l'un des sites d'épissage (5'SS ou 3'SS) est nouveau, tandis que l'autre site d'épissage est annoté (connu)

Le Mutant (rep2) comporte des nouveaux sites d'épissages par rapport au Wild Type, or nous avions vu qu'il présentait une surexpression par rapport à WT , donc cela peut en être une raison.

Infer experiment:



Ce script prédit le mode de préparation de la bibliothèque (brin sens ou brin antisens) en fonction de la façon dont les lectures alignées superposent les caractéristiques du gène dans le génome de référence.

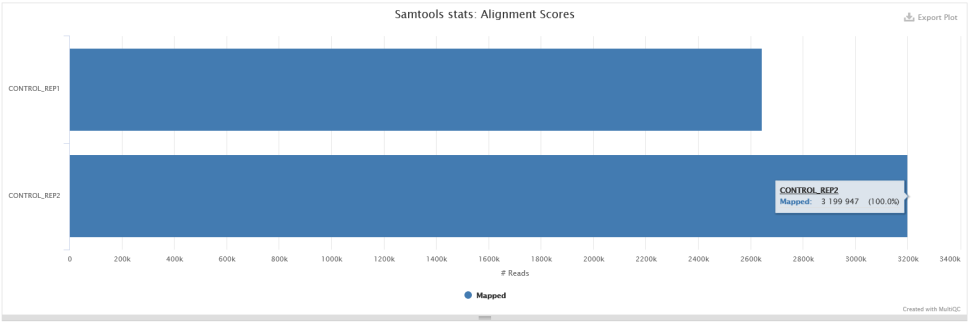
On peut voir que les prédictions semblent les mêmes pour les deux échantillons dont les alignements ne sont pas aberrants car le sens des échantillons est correct.

BAM Stat :

Permet de calculer les statistiques de mappage des lectures à partir du fichier BAM fourni. Ce script détermine les "lectures mappées de manière unique" à partir de la qualité du mappage, quelle qualité correspond à la probabilité qu'une lecture soit mal placée.

Samtools:

Percent mapped: Nombre de read mappés. On peut voir qu'il ne semble pas y avoir de problème de contamination

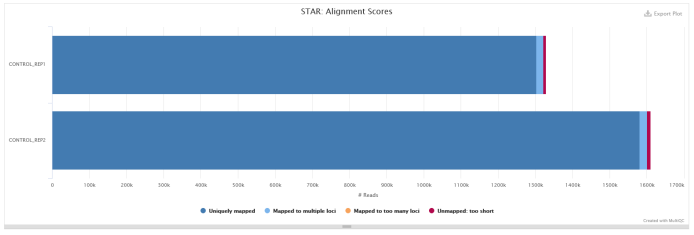


Alignment metrics: Un ensemble de mesure sur les alignements.

Samtools flagstat: compte le nombre d'alignements pour chaque type FLAG(catégorie). Chaque catégorie de la sortie est divisée en QC réussi et QC échoué. Permet de voir ce qui est mappé et si c'est réaliste.

Samtools idxstats: Vérifiez le nombre de lectures mappées sur chaque chromosome. Il peut aider à évaluer la qualité de l'échantillon.

STAR:



Donne des informations sur les lectures de mappage uniques; Un échantillon de bonne qualité aura au

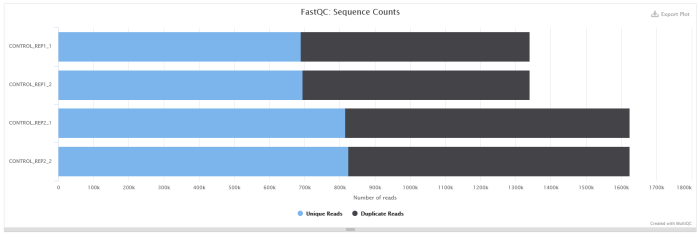
moins 75 % des lectures mappées de manière unique. Plus le nombre de lectures de mappage unique est faible, plus le nombre de lectures mappées à plusieurs emplacements est élevé.

On a des bons échantillons car on a très peu de mappage multiple et de séquence non mappée.

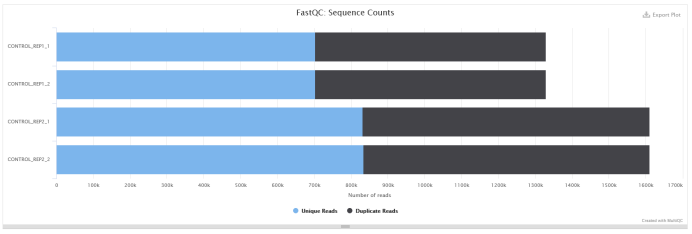
FastQC:

Sequence count:

Avant réduction:



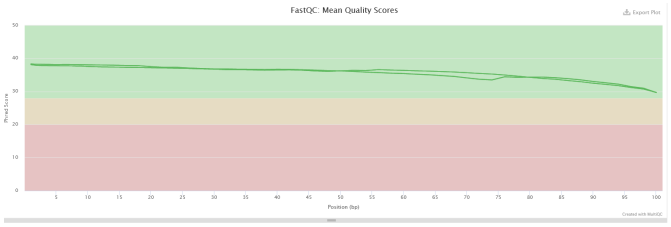
Après réduction



Sur ce graphique le bleu représente les lectures uniques et le noir représente les lectures en double. L'axe des abscisses est le nombre de lectures. On ne remarque pas de changement avant et après réduction.
On a un ratio de 1:1 pour lecture double et unique ce qui suggère que la bibliothèque n'a pas été séquencé au maximum(car souvent on a beaucoup plus de lecture double si on a séquencé au maximum de la profondeur).

Sequence quality

Avant réduction:



Après réduction

Montre des valeurs de qualité à chaque position de base dans le fichier. L'axe des Y sur le graphique montre les scores de qualité. Plus le score est élevé, meilleur est le séquençage de la base.
La qualité de de séquençage est bonne sur l'ensemble et semblable avant et après réduction.

Per base sequence content:

Proportion de chaque base de l'ADN dans le fichier. La quantité relative de chaque base doit refléter la quantité globale des bases dans le génome. Si c'est déséquilibré les unes par rapport aux autres cela peut refléter un problème de séquençage ou une contamination.

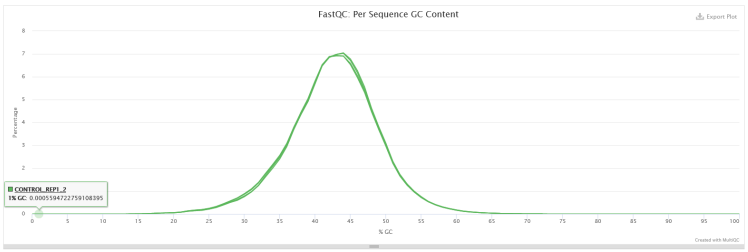
Avant réduction:

Après réduction:

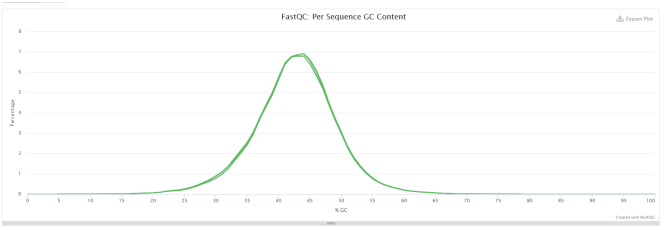
On remarque dans tous les cas un déséquilibre dans le contenu en base sur le début de la séquence. Dans les analyses précédentes nous n'avions pas trouvé de contamination cependant on avait remarqué un biais dans la partie 5' de la séquence qui laissait suggérer un problème sur les séquences sur cette partie(échantillon abimé) et donc peut expliquer ce déséquilibre si la partie est dégradée. On remarque qu'après réduction un déséquilibre apparaît en fin de séquence. C'est sûrement lié au fait qu'on a retiré les adaptateurs avec Trim galore qui en fonction de l'endroit où il a coupé a créé un déséquilibre sur les bases.

Per sequence GC content:

Avant réduction:



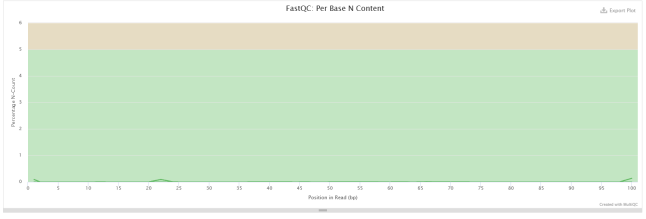
Après réduction:



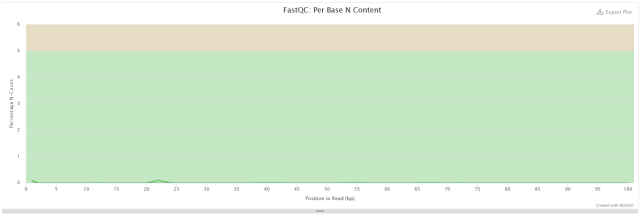
Contenu de GC à chaque position de base.Si c'est déséquilibré cela peut refléter un problème de séquençage ou une contamination.
On remarque que le contenu en GC est semblable pour les échantillons et ne change pas après réduction.

Per base N content:

Avant réduction:



Après réduction:

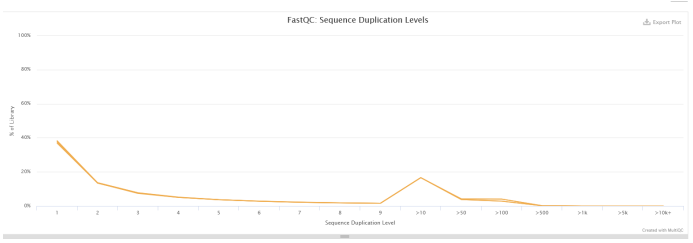


Pourcentage de base à chaque position pour laquelle un N était placé. Ne doit pas dépasser quelques pour cent sinon cela suggère qu'il n'a pas était possible d'interpréter les données suffisamment bien pour être valides.

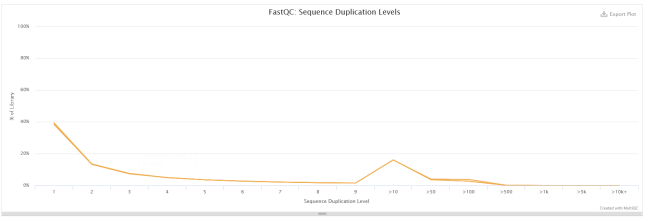
Le pourcentage de N est très faible est à même diminué après réduction. Les données sont donc valides.

Sequence duplication levels:

Avant réduction:



Après réduction:

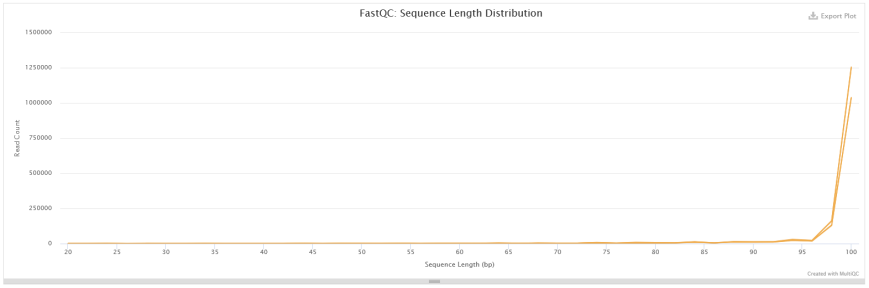


Montre le nombre relatif de séquences avec différents degrés de duplication. Un haut de niveau de duplication peut indiquer une sorte de biais d'enrichissement (par exemple PCR sur amplification).

Sans surprise on retrouve les résultats des analyses précédentes (picard) , nos séquences contiennent des duplications optiques(20%). Cela ne semble pas changer avant et après réduction.

Sequence length distribution:

Après réduction:



Montre la distribution des tailles de fragment dans le fichier.

Avant réduction tous les échantillons ont des fragments de taille 101 bp mais cela change après réduction on a des fragments de plus grande taille c'est sûrement lié à l'endroit où ont étaient supprimés les adaptateurs.

Overrepresented sequences:

Répertorie toutes les séquences qui représentent plus de 0,1 % du total.

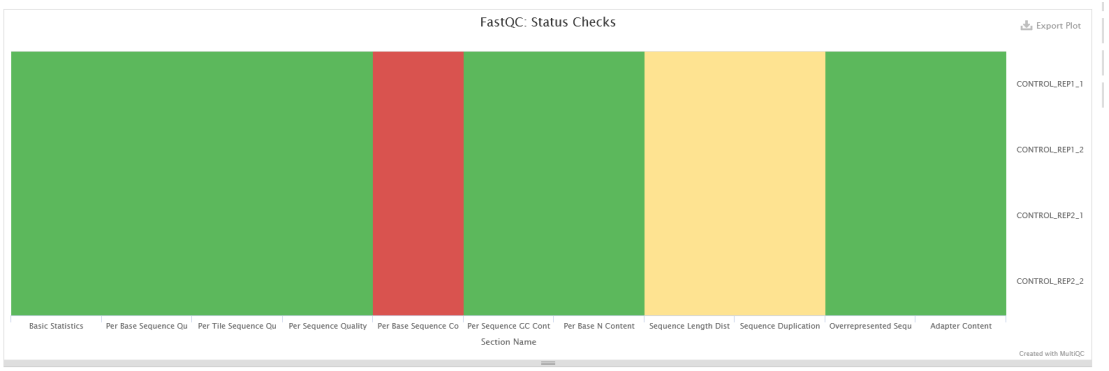
Que ce soit avant ou après réduction, 4 échantillons avaient moins de 1 % de lectures composées de séquences surreprésentées. Les échantillons sont équilibrés.

Adaptateur content:

Analyse de tous les Kmers pour trouver ceux qui n'ont pas une couverture uniforme sur toute la durée des lectures.La présence de séquences surreprésentées (telles que des dimères adaptateurs) entraînera la domination du tracé par les Kmers que contiennent ces séquences.

Aucun échantillon trouvé avec une contamination de l'adaptateur > 0,1 % pour les échantillons avant et après réduction.

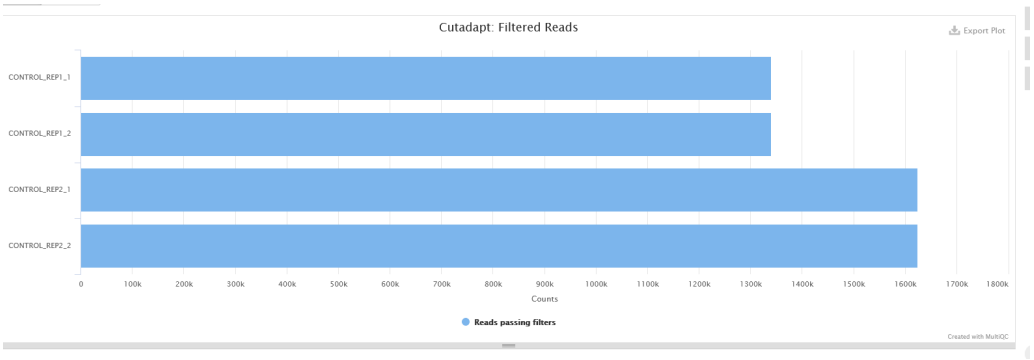
Statuts Checks:



Statut pour chaque analyse FastQC indiquant si les résultats semblent tout à fait normaux (vert), légèrement anormaux (orange) ou très inhabituels (rouge), c’est que nous avons analysé plus tôt.

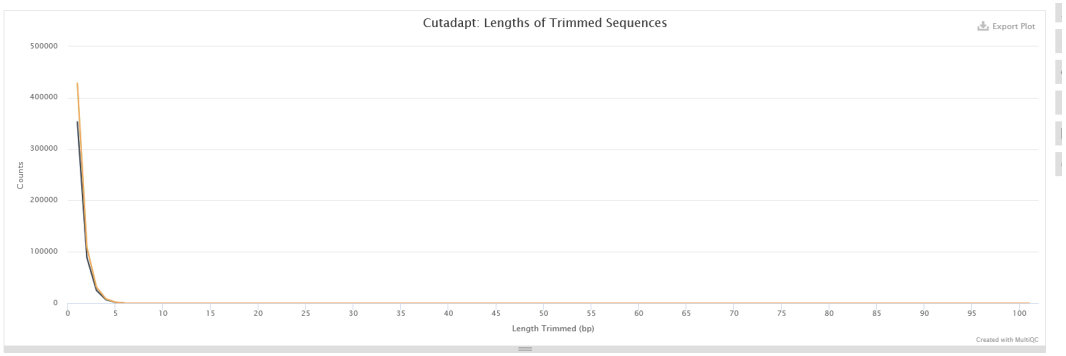
Cutadapt:

Filtered reads:



Le graphique montre le nombre de lectures (SE) / paires (PE) supprimées par Cutadapt. On voit que 100% des reads restent après filtrage et réduction.

Trimmed Sequence Lengths:



Ce graphique montre le nombre de lectures et la longueur d'adaptateur coupés. On voit une partie coupée de séquences. La longueur de la séquence est de 5pb. Il a quelques lectures coupées tout le long qui peuvent provenir d'erreurs acquises lors du séquençage ou de la génération d'échantillons

Software version:

Process Name	Software	Version
BEDTOOLS_GENOMECOV	bedtools	2.30.0
CUSTOM_DUMPSOFTWAREVERSIONS	python	3.9.5
	yaml	5.4.1

Rapport généré par le pipeline qui montre les versions des logiciels utilisés lors du pipeline ainsi que le nom des processus utilisés.

Workflow summary:

Rapport généré par nextflow qui compile toutes les informations sur pipeline, le job lancé mais aussi les fichiers d’input.