Bru Annabelle
M2_BBS

Nextflow

Le répertoire données_tomates est le dossier où sont stockées les données du TP.
Ce sont 2 réplicats du chromosome 6 pour 2 conditions différentes. La première condition est le WT (Wild Type) et la 2e est le MT(Mitochondrie).
Le répertoire results : résultats du sbatch. En réalité le répertoire se nomme donees_tomates mais j'ai pas osé changé le nom pour pas tout faire planter.

Création de rnaseq.sh dans work/cobee/nextflow qui va être notre fichier d'exécution.
Création du fichier sample.csv dans work/cobee/nextflow/donees_tomates pour nf-core/rnaseq.
Puisqu'on utilise la 3.4 il faut respecter le format demander c'est-à-dire :

```
sample,fastq_1,fastq_2,strandedness
CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,unstranded
CONTROL_REP1,AEG588A1_S1_L003_R1_001.fastq.gz,AEG588A1_S1_L003_R2_001.fastq.gz,unstranded
CONTROL_REP1,AEG588A1_S1_L004_R1_001.fastq.gz,AEG588A1_S1_L004_R2_001.fastq.gz,unstranded
```

Donc dans notre cas le ficher sample.csv est rempli de la manière suivante :

```
cobee@genologin2 /work/cobee/nextflow $ more donees_tomates/sample.csv
sample,fastq_1,fastq_2,strandedness
1,donees_tomates/WT_rep1_1_Ch6.fastq.gz,donees_tomates/WT_rep1_2_Ch6.fastq.gz,forward
2,donees_tomates/MT_rep1_1_Ch6.fastq.gz,donees_tomates/MT_rep1_2_Ch6.fastq.gz,forward
```

Au départ le chemin des fastq n'était pas précisé dans le sample.csv car je pensais que le fait que le fichier sample.csv soit dans le même répertoire suffisait puisque j'indiquais le chemin de ce dernier dans mon rnaseq.sh.

```
ERROR: Please check input samplesheet -> Read 1 FastQ file does not exist!
WT_rep1_1_Ch6.fastq.gz
```

Cette erreur est obtenue sans le chemin vers le fichier.
Sinon dans le sample.csv on retrouve les fichiers fastq qui n'ont pas besoin d'être dézippés. Les WT sont sur la même ligne et les MT sur une autre. Le sens des brins a été déterminé arbitrairement en forward. Avec les résultats obtenus détaillés plus loin on pourrait mettre "unstranded". Ayant testé avec ça les résultats ne changent pas pour le reste mis à part que du coup on a plus

Le rnaseq.sh comporte les lignes suivantes :

```
cobee@genologin2 /work/cobee/nextflow $ more rnaseq.sh
#!/bin/bash
#SBATCH -p workq
#SBATCH --time=24:00:00
#SBATCH -J AnnabelleBru
#SBATCH --mem=6GB


#To erase remaining module
module purge


#Load binaries
module load bioinfo/nfcore-Nextflow-v21.04.1
module load bioinfo/sratoolkit.3.0.0

nextflow run nf-core/rnaseq -r 3.4 -profile genotoul --input ./donees_tomates/sample.csv --fasta donees_tomates/ITAG2.3_genomic_Ch6.fasta --
gtf donees_tomates/ITAG2.3_genomic_Ch6.gtf --outdir results/
```

2e ligne : Partition sur laquelle on travaille.
3e ligne : Temps maximum alloué au job

4e ligne : Nom du job
5e ligne : Mémoire allouée pour ce job.

Le module purge permet d'éviter des conflits de module par rapport à la compatibilité. Si on a plusieurs jobs à faire, c'est une bonne idée de nettoyer pour éviter tout souci.
Les –fasta et –gtf sont le génome de référence et l'annotation du chromosome 6 de la tomate.

J'ai mis quelques tentatives à faire fonctionner le sbatch rnaseq.sh à cause dans un premier temps du nom du répertoire de données qui se nomme donees_tomates et non donnees_tomates puis les «/» qui n'étaient pas à mettre avant le nom du dossier sauf pour l'input.

```
Completed at: 30-Sep-2022 11:24:08
Duration    : 11m 26s
CPU hours   : 3.0
Succeeded   : 80
```

Seff :

```
cobee@genologin2 /work/cobee/nextflow $ seff 37375772
Job ID: 37375772
Cluster: genobull
User/Group: cobee/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:03:03
CPU Efficiency: 24.86% of 00:12:16 core-walltime
Job Wall-clock time: 00:12:16
Memory Utilized: 1.89 GB
Memory Efficiency: 31.47% of 6.00 GB
```

Job ID : comme son nom l'indique, c'est l'identifiant du job exécuté.
User/Group : Le nom de l'utilisateur et son groupe
State : Etat du job, permet de voir s'il est actif (running) ou s'il y a eu une erreur pendant le run.
Le CPU efficiency est donc l'efficacité du CPU est calculée comme le ratio du temps réel de tous les cœurs divisé par le nombre de cœurs demandés divisé par le temps d'exécution. Le job a utilisé 24,86 % CPU. La mémoire utilisée pour ce job est de 1,89 GB soit 31,47 % des 6GB disponibles.
Intérêt du resume : De ne pas recommencer à 0. Quand un job a une erreur et qu'il y a une modification du fichier .sh cela permet de reprendre le job sans avoir à tout relancer. C'est un gain de temps.

Résultats obtenus : on obtient différents dossiers :

genome comprenant les fichiers fasta et gtf utilisés pour la référence ainsi qu'en format bed et sizes qui sont des fichiers intermédiaires générés par la pipeline. Ce dossier comporte aussi 2 dossiers contenant des informations sur le génome et le chromosome.
fastaqc → Résultats des fastqc pour les 4 fastq c'est-à-dire le contrôle qualité pour chaque fastq.
Multiqc → On retrouve le fichier html ainsi que les fichiers utilisés pour alimenter ce html.
pipeline_info → Dossier où l'on retrouve le rapport de l'exécution de la commande nextflow en format html et txt. On a les détails du temps d'exécution des différents outils (Salmon, STAR, Fastqc,...), l'état des différentes tâches ("Completed").
star_salmon → C'est le dossier où sont stockés les fichiers d'alignement. Par défaut, c'est star_salmon qui est utilisée. Ce dossier permet donc de garder les résultats de ces alignements si un autre run est effectué (–resume) avec un autre outil d'alignement.
trimgalore → Contient des fichiers qui résument les actions effectuées sur les fastq/
Contient le total de reads traités

Interprétation du multiqc

Multiqc est un rapport qui permet de réunir tous les résultats obtenus des différents logiciels/outils utilisés.
Dans un premier temps nous obtenons un tableau récapitulatif composé de 8 lignes et 28 colonnes

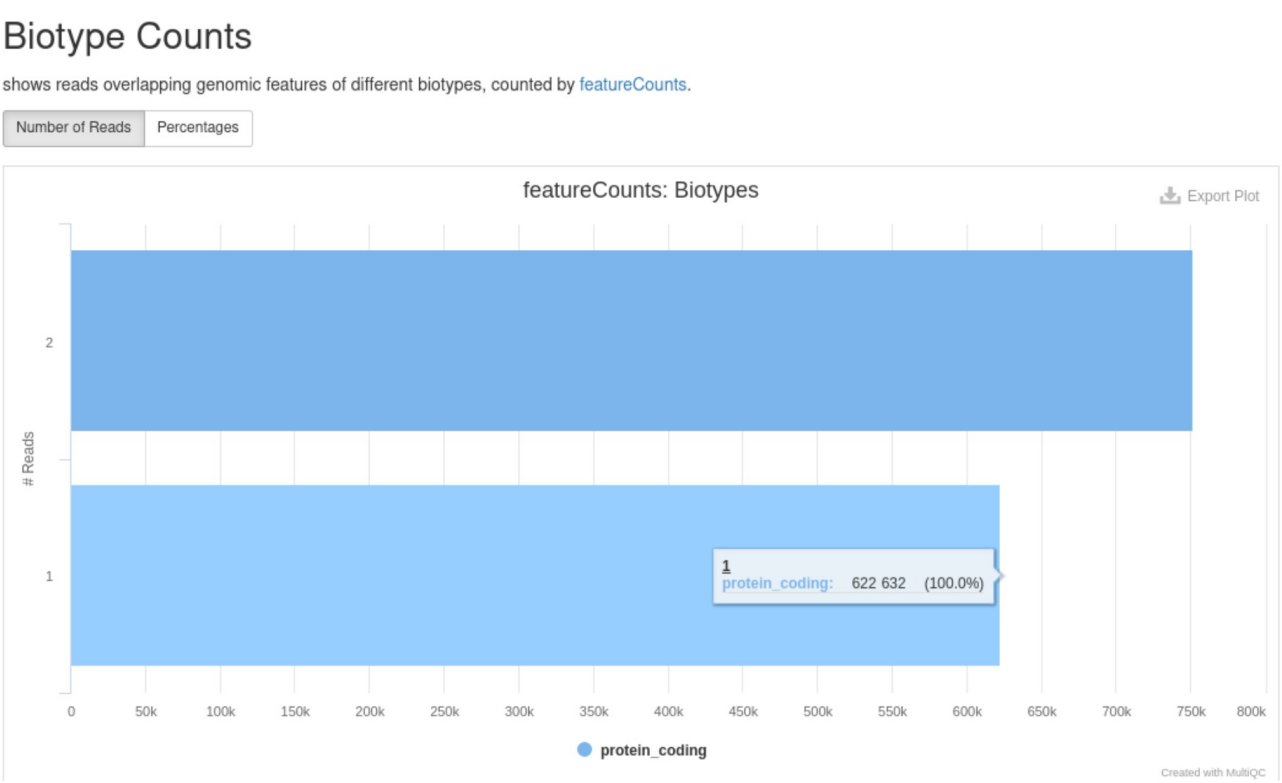
Sample Name	M Reads	M Reads Mapped	% rRNA	dupInt	% Dups	5'-3' bias	M Aligned	% Proper Pairs	Error rate	M Non-Primary
1	2.7	2.7	0.00%		18.3%	1.43	1.3	79.1%	0.16%	0.0
1.0				0.00%						
1_1										
1_2										
2	3.2	3.2	0.00%		17.3%	1.44	1.6	80.5%	0.16%	0.0
2.0				0.00%						
2_1										
2_2										

La première colonne correspond au nom de l'échantillon. En entrée nous avons donné 4 fastq au

M Reads Mapped	% Mapped	% Proper Pairs	% MapQ 0 Reads	M Total seqs	% Aligned	M Aligned	% Dups	% GC	Length	% Failed
2.6	100.0%	100.0%	0.0%	2.6	98.0%	1.3				
							48.5%	42%	101 bp	9%
							48.2%	42%	101 bp	9%
3.2	100.0%	100.0%	0.0%	3.2	98.1%	1.6				
							49.7%	42%	101 bp	9%
							49.2%	41%	101 bp	9%

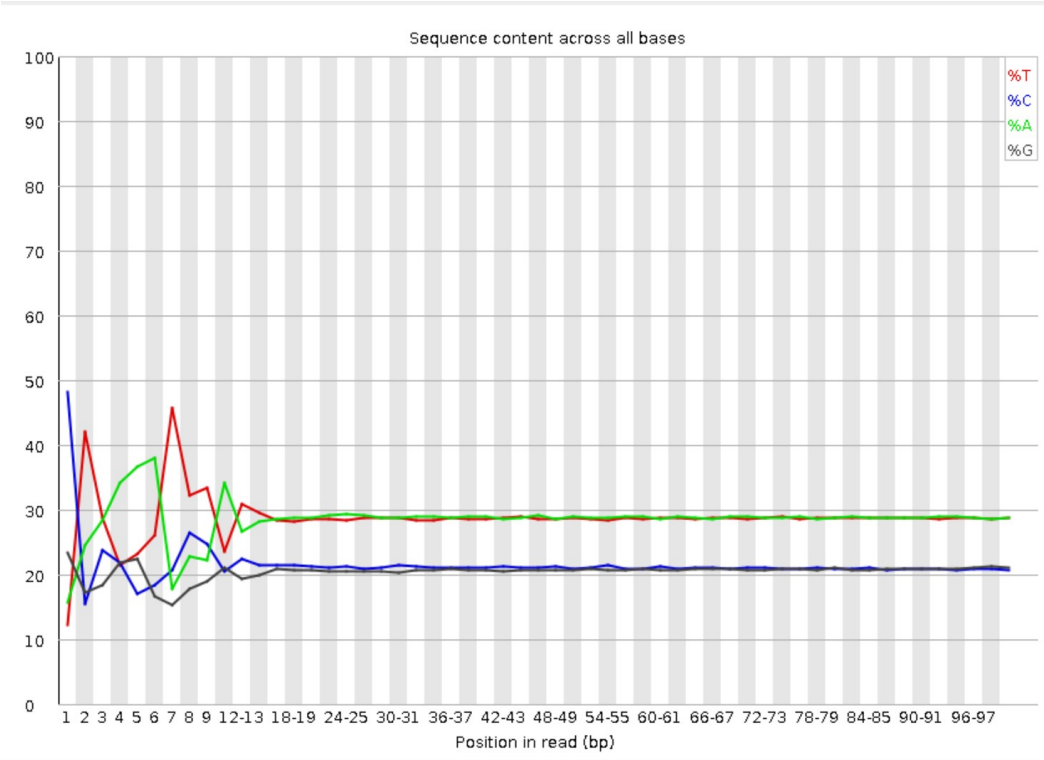
pipeline. Le 1_1correspond au WT réplicat 1, le 1_2au WT réplicat 2, le 2_1au MT réplicat 1 et 2_2au MT réplicat 2. Le 1 va correspondre aux 2 fastq WT fusionnés et le 2 aux 2 fastq MT.
M Reads Mapped est obtenu par Samtools et correspond au nombre de reads mappés dans le fichier BAM donc 2,7 millions et 3,2 millions pour le WT et MT respectivement.

Le %rRNA traduit le pourcentage de contamination par des rRNA. Ce n'est pas le cas dans ce résultat, aucun rRNA de présent. Grâce à l'outil featureCounts qui permet de quantifier pour résumer la distribution de reads mappés on a un contrôle qualité supplémentaire pour mettre en évidence ou non la présence de rRNA.

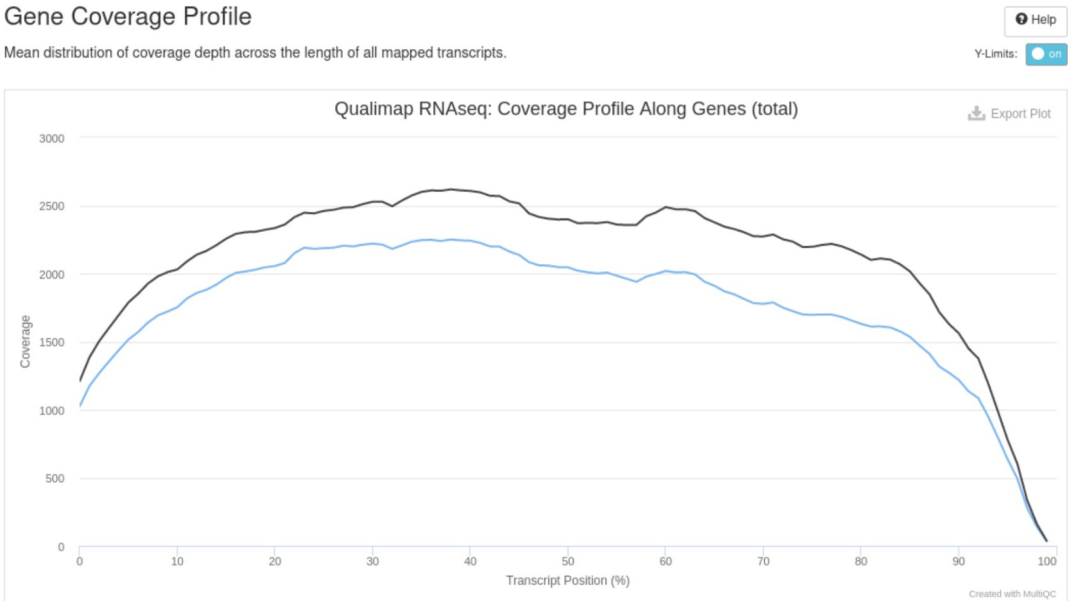


Pour les 2 échantillons 100 % sont des protéines et ça confirme la non contamination des échantillons.

Le 5'-3' biais se traduit par si les données présentent des biais en 5' ou en 3'. Ce biais peut être est causé par l'amorçage aléatoire lors de la préparation des échantillons car en théorie il y a une sélection d'hexamères aléatoire qui devrait avoir la même fréquence dans le mélange et s'amorcer avec la même efficacité alors qu'en réalité certains hexamères sont favorisés pendant l'étape d'amorçage. Cela amène à de potentielles augmentations d'erreurs d'amorçage. Il peut aussi être causé par la dégradation de l'ARN. Ici le biais 5'-3' de nos 2 échantillons est égal à 1,43 environ. Dans l'idéal ce biais devrait être égal à 1 ce qui voudrait dire qu'il n'y a aucun biais. On a quand même la présence de légers biais qu'on peut visualiser avec le contenu en bases de la séquence.



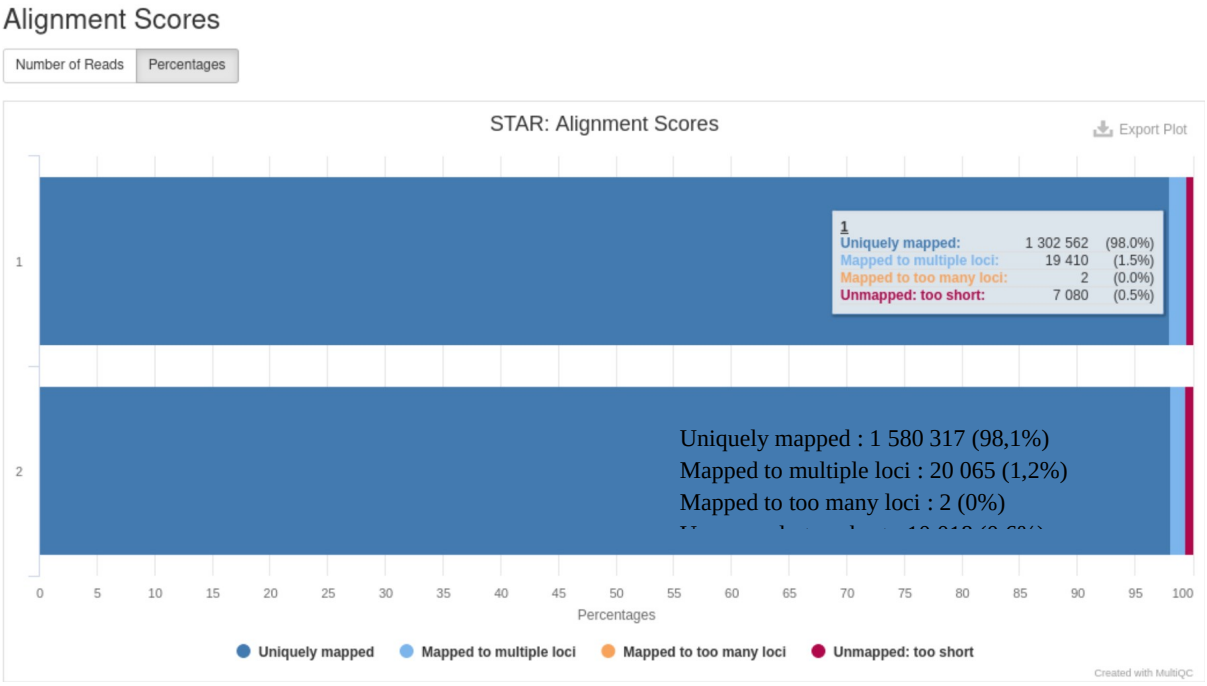
Le biais se voit ici au début, au niveau des 12 premières bases. Comme mentionné précédemment ceci peut être dû à l'étape d'amorçage. Sinon avec ces premières positions les proportions se stabilisent. Une autre figure pour visualiser de potentiels biais avec le profil de couverture en fonction de la position du transcrit.



On s'attend à une couverture à peu près égale pour les 2 échantillons ainsi qu'à une faible couverture en début et fin de transcrit. Nous avons donc ici une bonne couverture ainsi qu'une confirmation de la bonne qualité des échantillons.

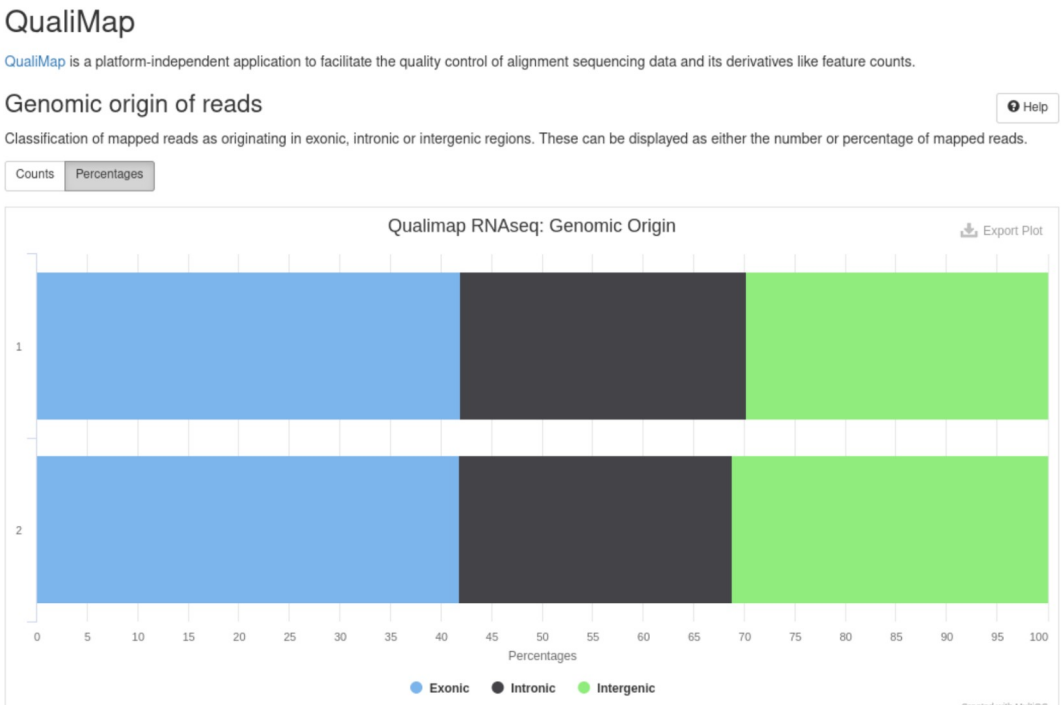
Le taux d'erreur représente le nombre de mismatches/bases mappés.

Le %Alignment a été réalisé par STAR qui est un outil permettant d'aligner les données RNaseq. En règle générale, nous pouvons considérer qu'un bon échantillon à un pourcentage supérieur à 75 %. Évidemment cela peut varier selon l'espèce étudiée. Pour les 2 échantillons nous avons 98 % d'alignement donc nous pouvons considérer qu'ils sont de bonnes qualités.

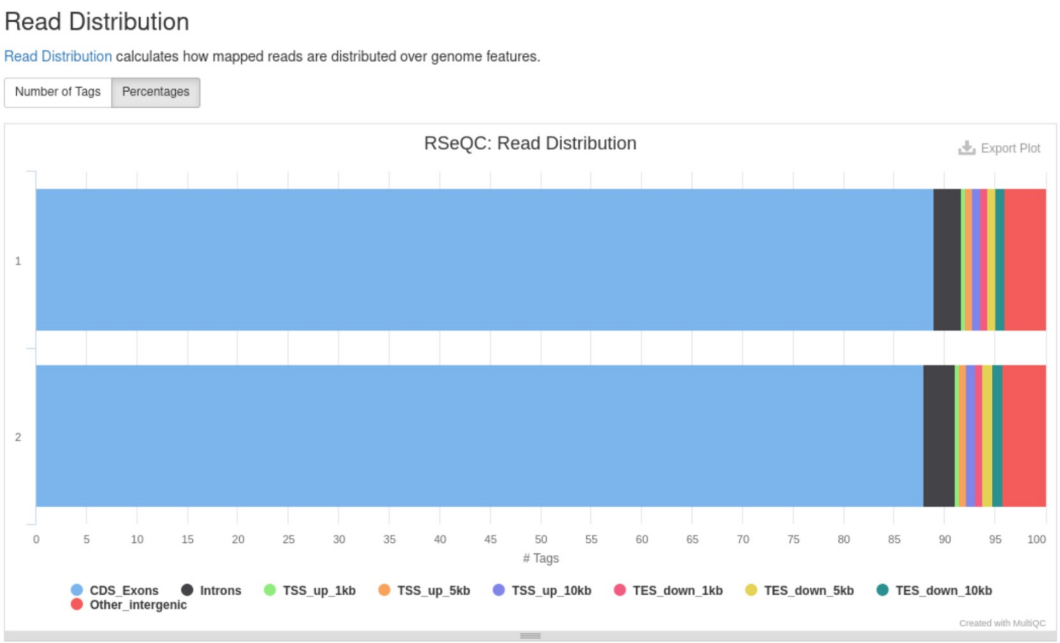


Cette figure donne les détails de l'alignement. On observe un bon alignement puisque 98% des reads sont alignés de manière unique.

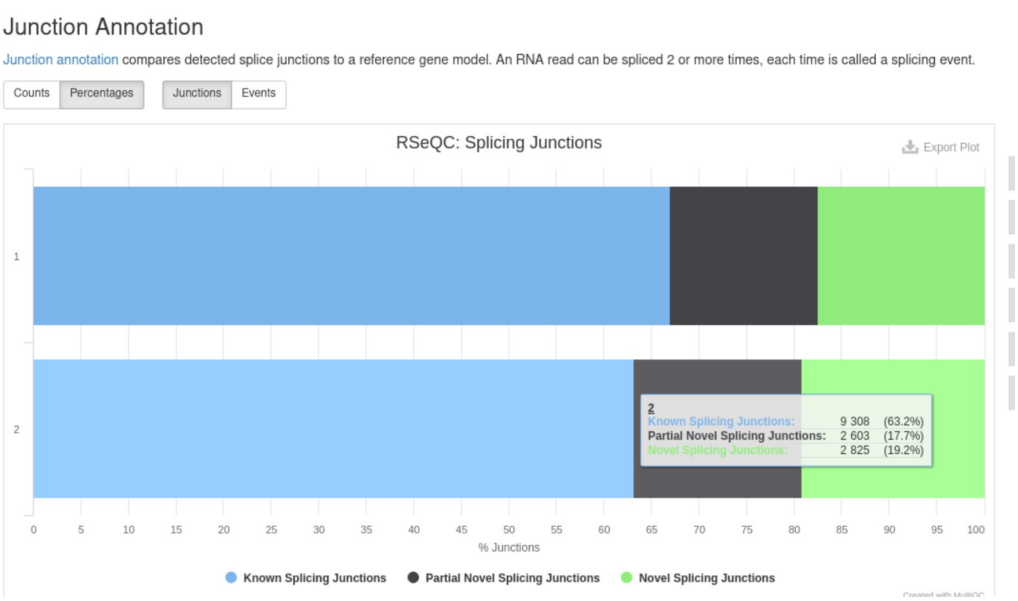
Ensuite, toujours afin de vérifier la qualité des échantillons, l'outil Qualimap permet de produire un plot qui va identifier des problèmes de contamination en s'intéressant au pourcentage de reads qui sont des exons, introns ou inter-géniques.



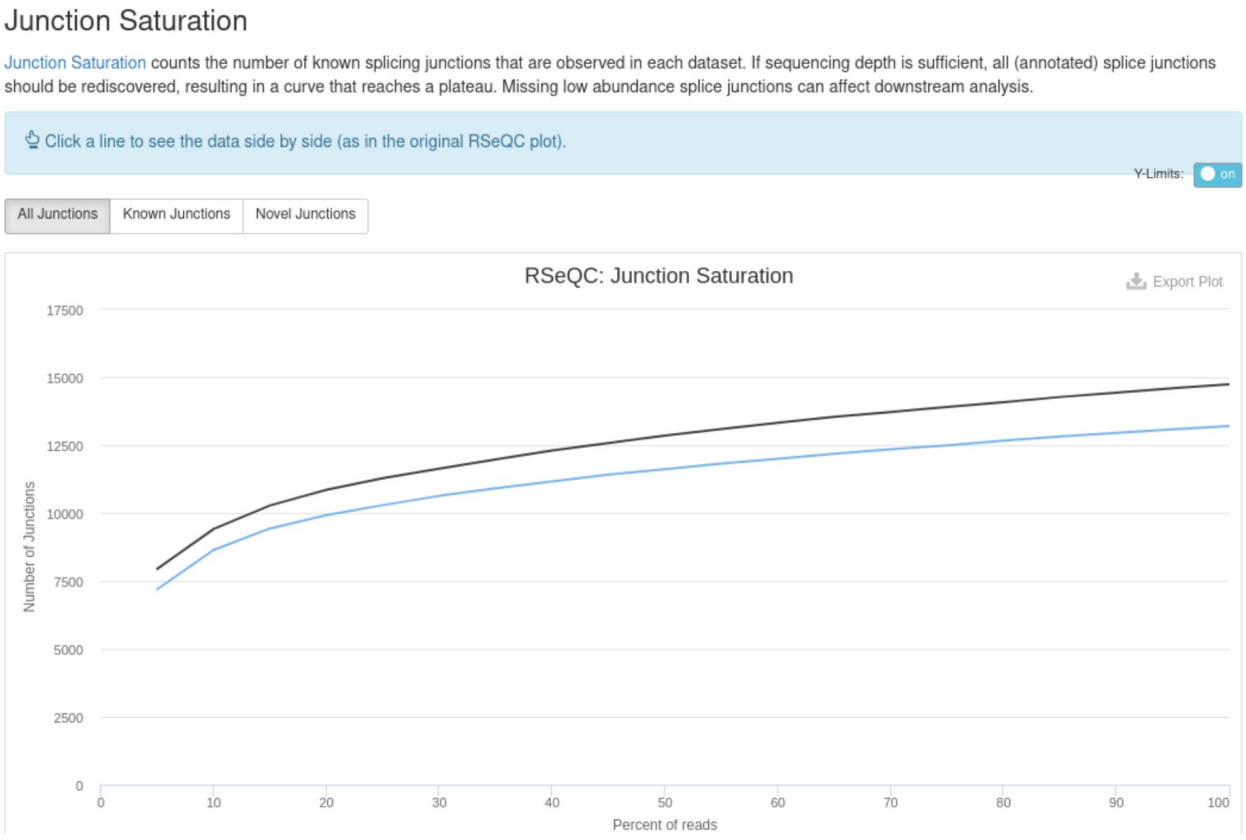
42 % d'exons et 28 % d'introns et 30 % de régions inter-géniques. Généralement quand le pourcentage de régions inter-géniques >30 % c'est qu'il y a contamination par conséquent ici on peut supposer une légère contamination des échantillons.



La distribution de reads quant à elle montre de bons résultats. Tous les reads sont mappés sur des régions exoniques.



Les jonctions exons-introns contiennent des séquences nucléotidiques bien conservées et qui vont être importantes pour l'épissage. L'annotation des jonctions compare les jonctions d'épissage qui ont été détectées à un modèle de gènes de références (le fichier au format gtf). Il y a 2 niveaux d'annotations : un au niveau de l'événement de l'épissage et l'autre au niveau de la jonction. Le but est de séparer les jonctions détectées en 3 catégories : « connues », « nouvelles » ou « partiellement connues ».

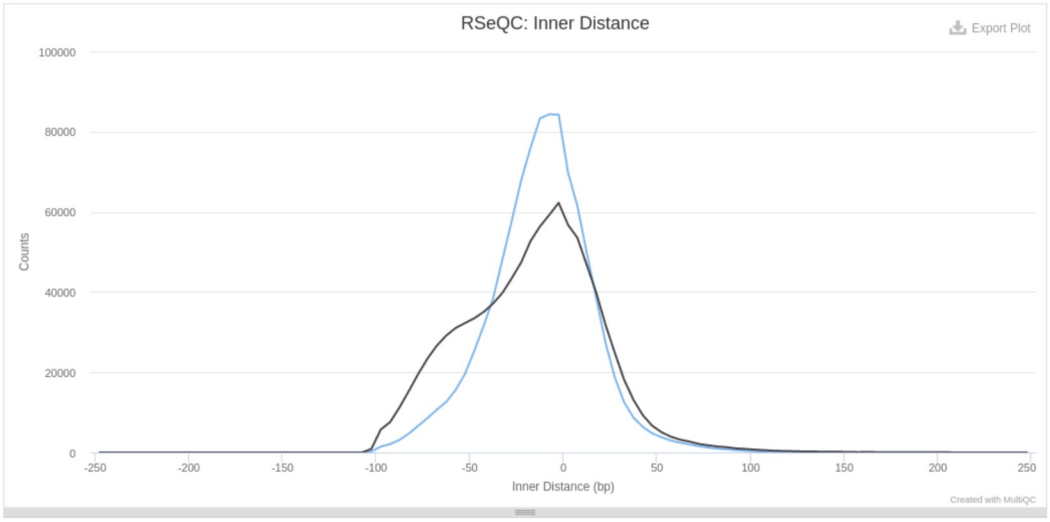


Cette figure montre le nombre de sites d'épissage détectés dans les données à différent niveau de sous-échantillonnage. Un bon échantillon est un échantillon qui atteint un plateau avant d'arriver à 100 % des données indique que toutes les jonctions ont été détectées et que la poursuite du séquençage ne donnera pas d'autres observations. Ici on remarque que les 2 courbes se rapprochent de ce cas là.

Inner Distance

Inner Distance calculates the inner distance (or insert size) between two paired RNA reads. Note that this can be negative if fragments overlap.

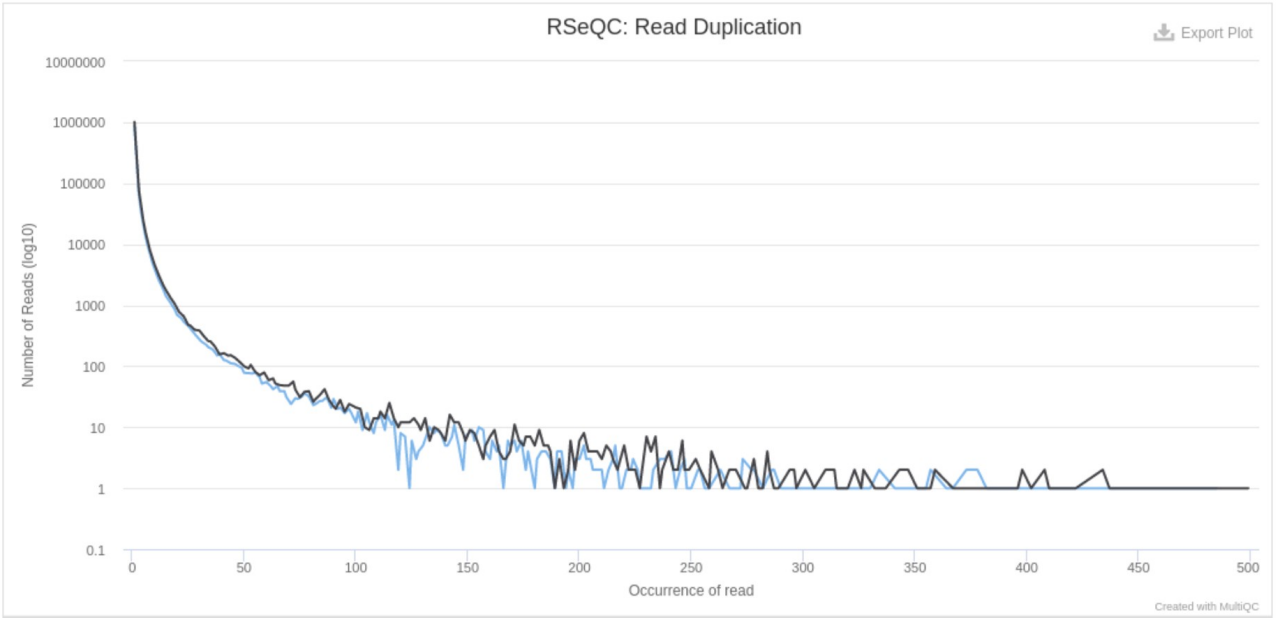
Counts | Percentages



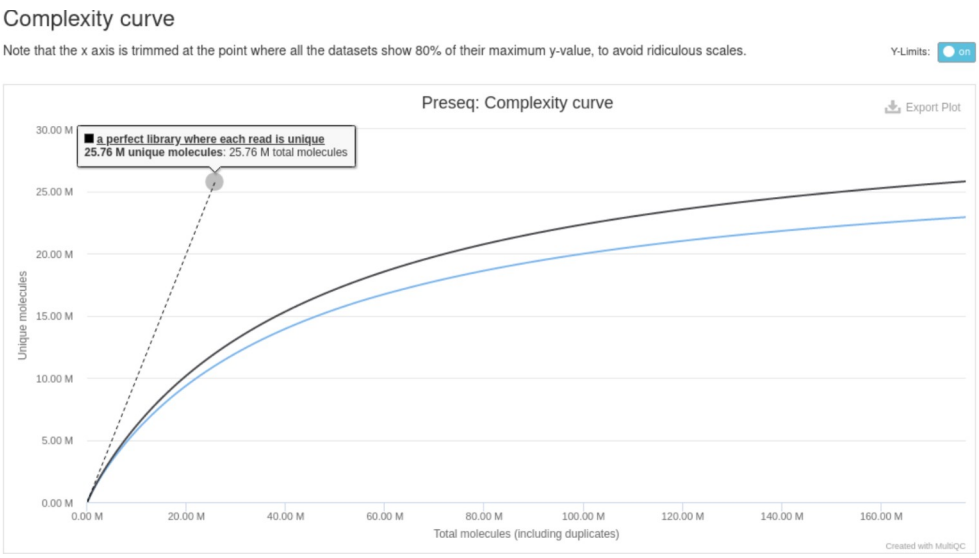
Ce plot n'est pas généré pour des données single-end donc cela veut dire que nous sommes en présence d'un séquençage paired-end. La distance mesurée est celle qui sépare la fin d'un read avec le début du suivant. Si la distance est courte c'est que la qualité est mauvaise. La qualité est moins bonne sur l'échantillon des MT.

Read Duplication

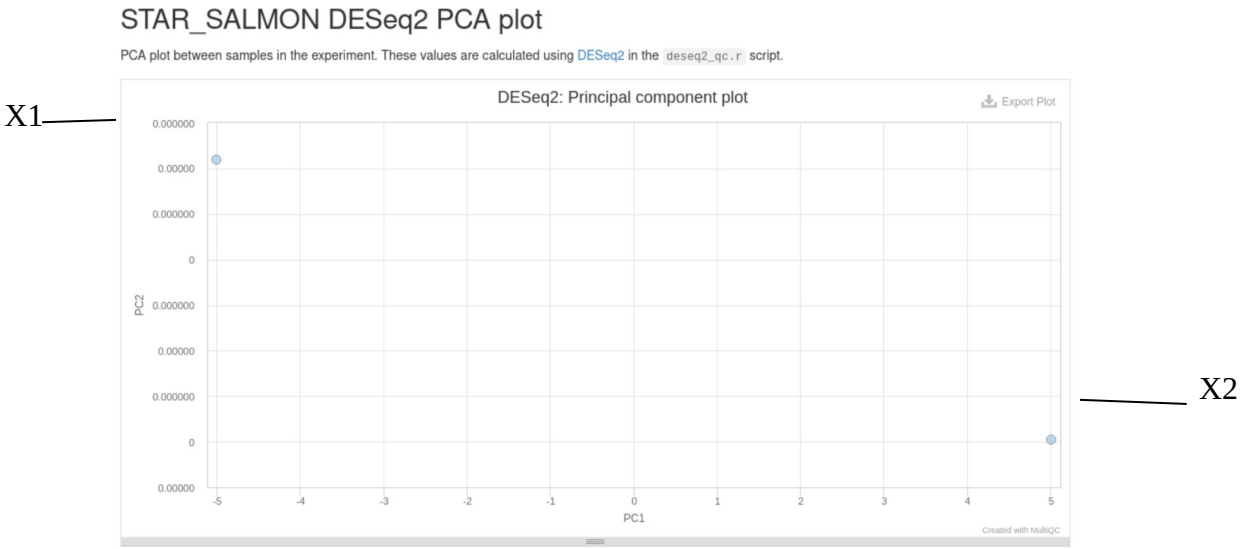
read_duplication.py calculates how many alignment positions have a certain number of exact duplicates. Note - plot truncated at 500 occurrences and binned.



Normalement il devrait y avoir un faible taux de doublons pour la majorité des reads ce qui a l'air d'être le cas donc au niveau de la préparation et du séquençage il n'y a pas eu de duplication technique excessive.

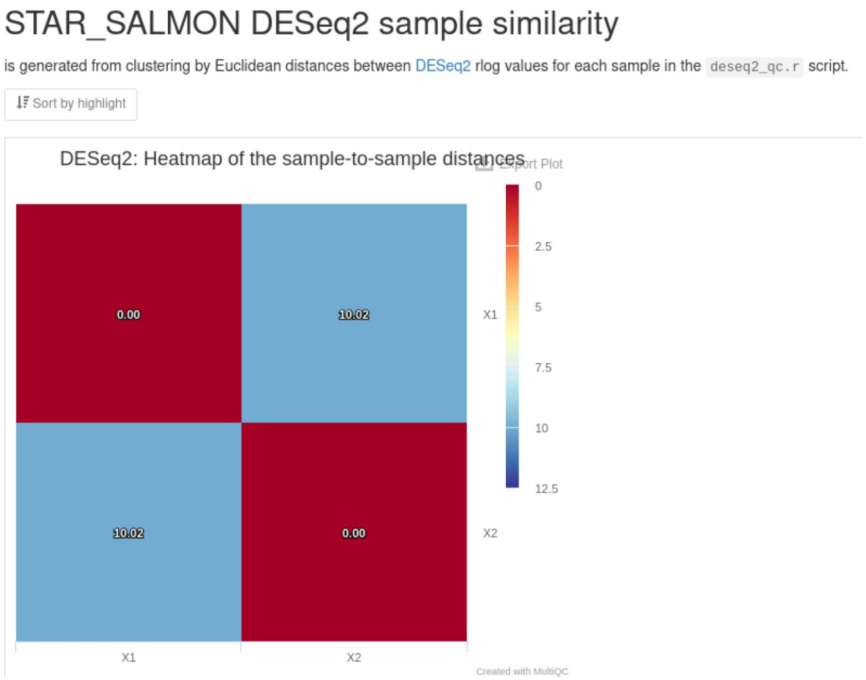


Fait la prédiction et l'estimation du nombre de reads redondants. Cela permet d'examiner si besoin d'un séquençage supplémentaire ou d'optimiser la profondeur du séquençage dans le but d'éviter des échantillons de faible complexité. Les courbes sont peu profondes c'est-à-dire qu'il y a une saturation de complexité et donc le séquençage effectué est suffisant et un autre n'ajoutera pas de nouveaux reads uniques.



X1

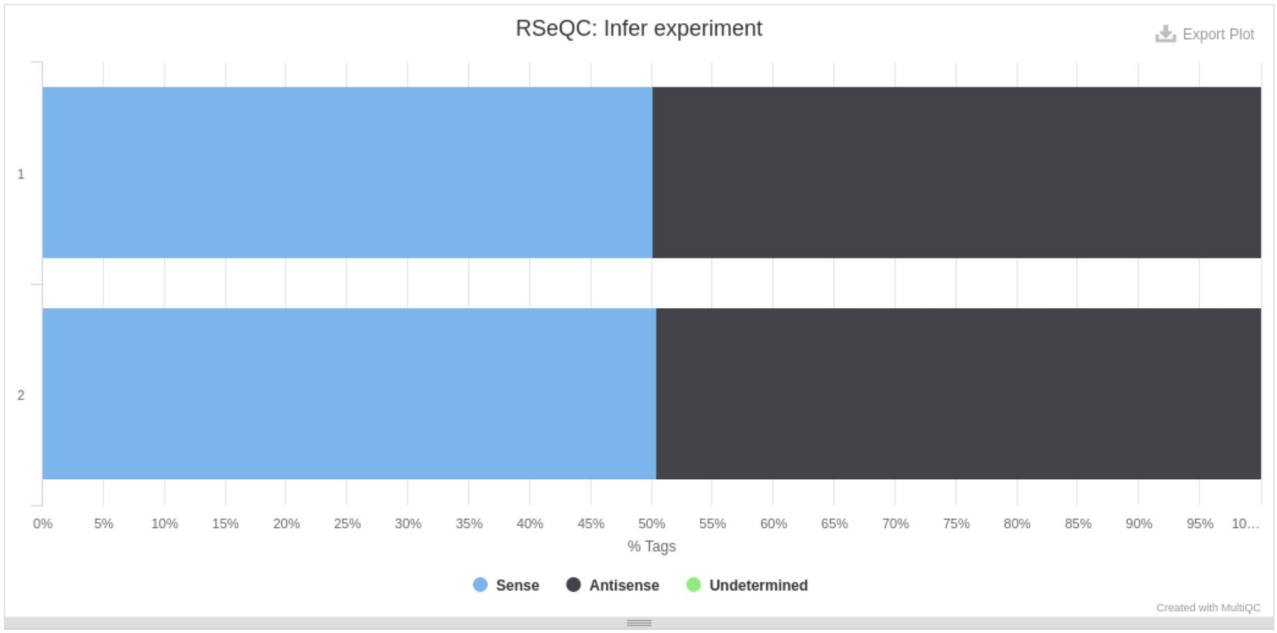
X2



Avec plus de données c’est plus intéressant de voir les différences et les similitudes, là cela nous confirme juste que les 2 échantillons sont différents.

Infer experiment

Infer experiment counts the percentage of reads and read pairs that match the strandedness of overlapping transcripts. It can be used to infer whether RNA-seq library preps are stranded (sense or antisense).



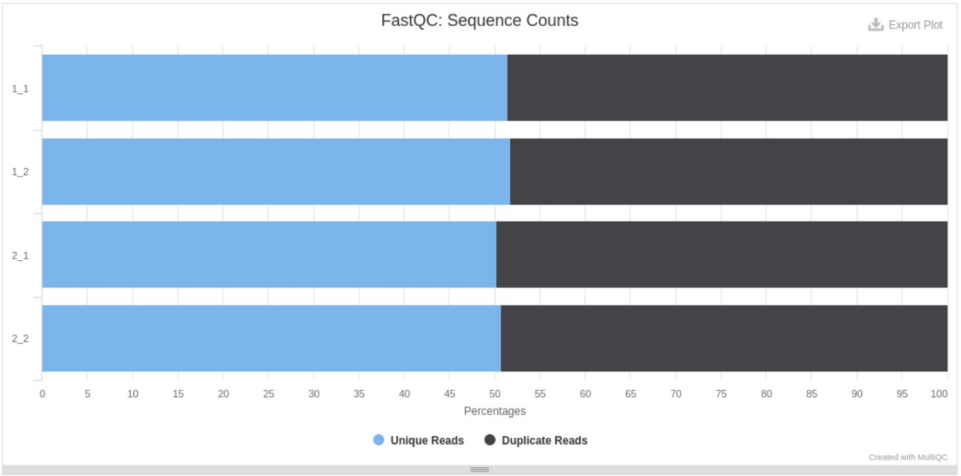
Les données n’ont pas de sens déterminés puisque les reads se sont alignés autant dans le sens forward que dans le sens reverse.

Sequence Counts

Help

Sequence counts for each sample. Duplicate read counts are an estimate only.

Number of reads | Percentages

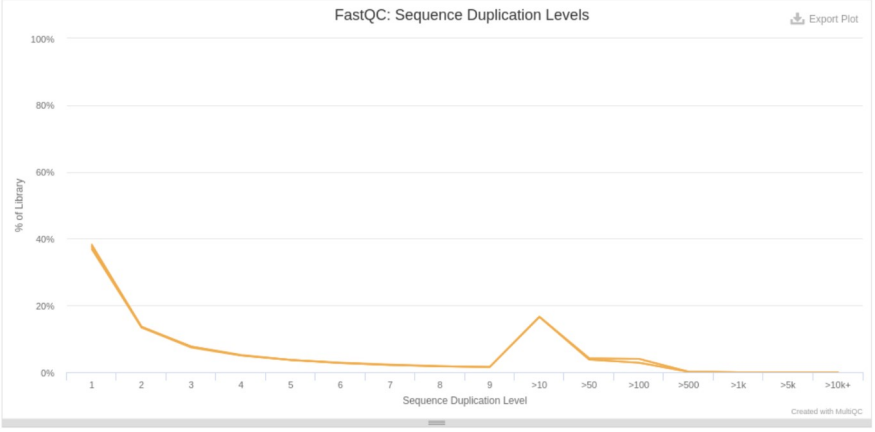


Sequence Duplication Levels

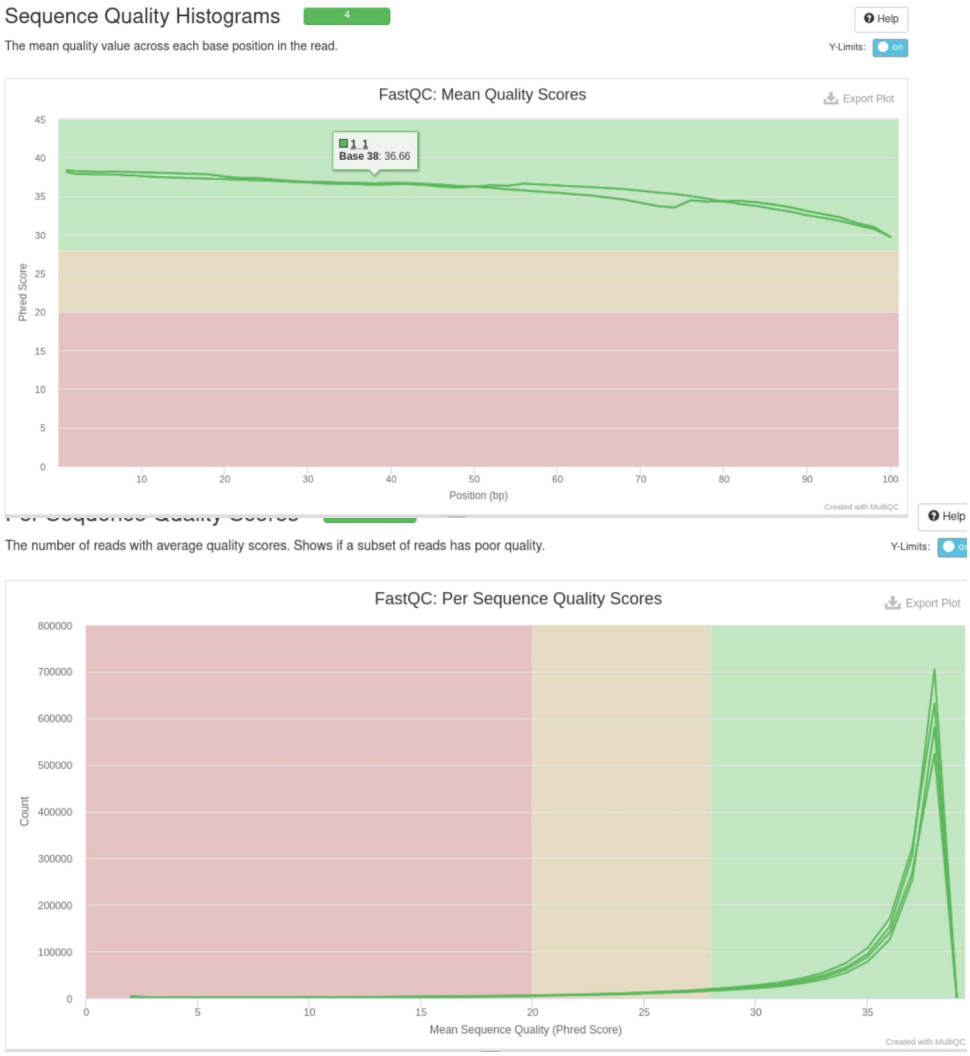
Help

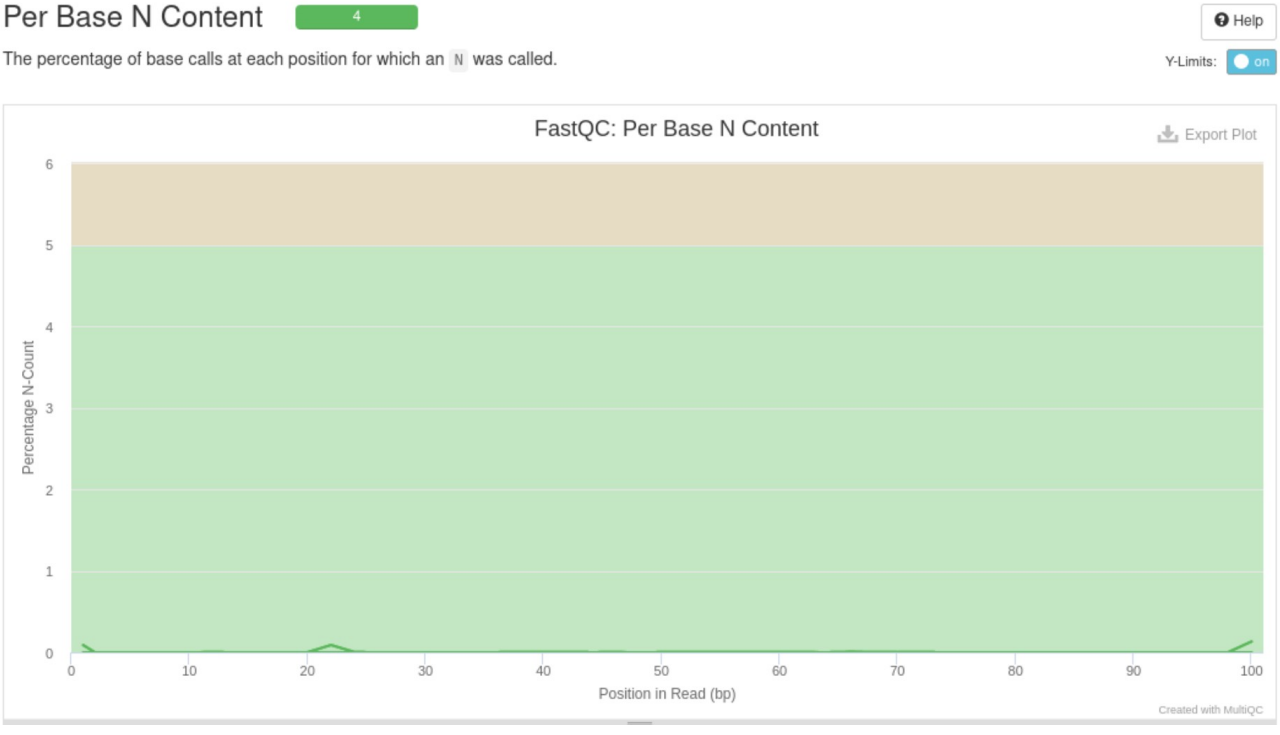
Y-Limits: on

The relative level of duplication found for every sequence.



Présence de beaucoup de reads dupliqués dans la RNAseq mais cela peut être normal dû à la présence de gènes qui sont fortement exprimés.





Per Sequence GC Content

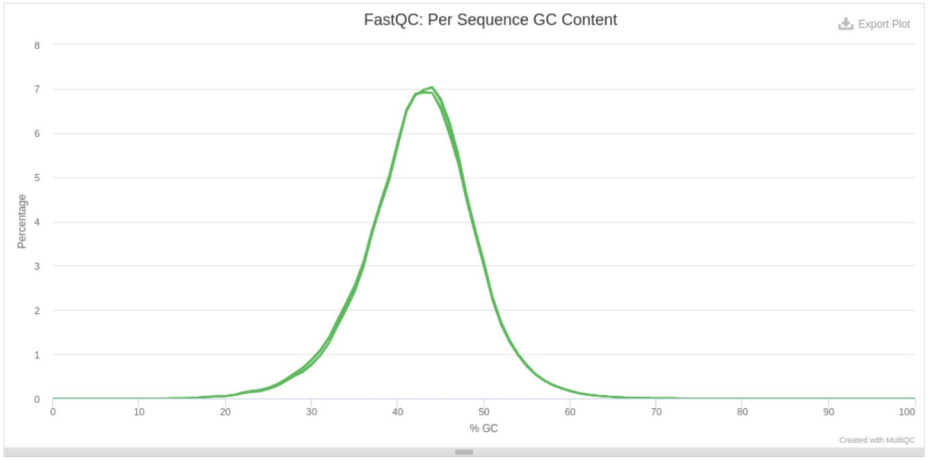
4

Help

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

Y-Limits: on

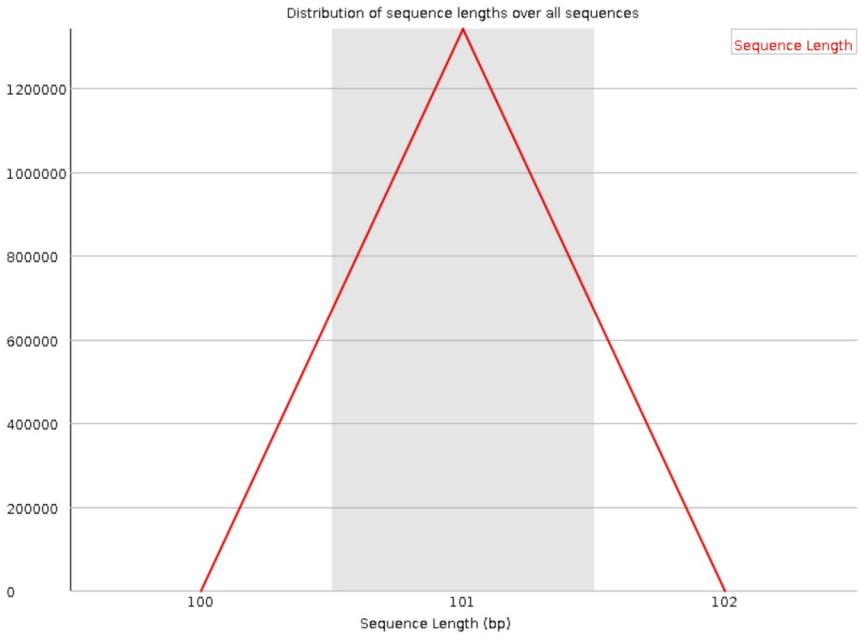
Percentages | Counts



Sequence Length Distribution

4

All samples have sequences of a single length (101bp).



Exemple de la distribution de la taille des séquences avec le chromosome 6 du WT (rep1) présent dans fastq.

Overrepresented sequences

4

Help

The total amount of overrepresented sequences found in each library.

4 samples had less than 1% of reads made up of overrepresented sequences

Adapter Content

4

Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

No samples found with any adapter contamination > 0.1%

On a une bonne distribution de la qualité par read tout le long du séquençage, un bon pourcentage de GC et pas de présence de N c'est-à-dire que le séquenceur a toujours su déterminer quelle base à mettre. On a pas non plus de séquences surreprésentées (c'est plus le cas pour l'ADN). Il n'y a pas de contamination de l'adaptateur.
Ensuite le logiciel va effectuer un cutadapt pour enlever les séquences provenant de l'adaptateur. Cependant comme on vient de le voir précédemment aucun read s'est aligné avec l'adaptateur donc on a de nouveaux de nouveaux fichiers de qualité mais avec les mêmes résultats voire moins bon. Pour éviter cela on peut mettre dans les paramètres l'option pour passer cette étape avec –skip_trimming.

Test de la commande avec cette fois l'alignement avec Hisat2. Résultat similaire avec STAR sauf qu'il n'y pas de quantification et par conséquent pas d'ACP ou de matrice de similarité pour comparer les échantillons.