Initiation sur genologin

TP note - Nextflow.

Auteur : Sandra Dailhau, M2 BIOINFORMATIQUE ET BIOLOGIE DES SYSTEMES

Introduction

Genologin est un serveur sécurisé capable d'accueillir des centaines d'utilisateurs en simultanée. Il peut être utilisé pour différent besoin :

- Pour servir les environnements de développement
- Pour tester son script avant l'analyse des données
- Pour lancer des travaux sur les nœuds du cluster
- Pour récupérer les résultats des données

Le serveur est lié à un cluster nommé SLURM – Simple Linux Utility for Resource Management - qui possède environs 3 000 cœurs. Le cluster est un ensemble d'ordinateur de calcul.

Exercice 1 :

Le travail de ce TP se trouve sur le compte chardon, plus précisément au chemin : /work/chardon/nextflow_note/. On peut y retrouver les données à télécharger dans le dossier data, le fichier rnaseq_traitement.sh.

Exercice 2 :

Le fichier bash créé se nomme donc rnaseq_traitement.sh. Il comporte l'ensemble des éléments nécessaires pour lancer un job sur le cluster de calcul genologin appartenant à genotoul. Ce fichier possède pour première ligne :

```
#!/bin/bash
#SBATCH -J SandraDailhau # Nom du job
#SBATCH -p workq # Partition a utiliser
#SBATCH --time=1-0 # Duree maximale du job est de 1 jour
#SBATCH --mem=6G # Memoire du job
#SBATCH -e rnaseq_traitement.err
#SBATCH --cpus-per-task=4
```

SBATCH est une commande permettant de lancer un job au cluster Slurm, il est possible de mettre des attributs pour personnaliser le job. Pour cette étude, les attributs utilisés sont les suivants :

-J SandraDailhau ; Cela permet de donner un nom au job lorsqu'il sera transmis. La valeur par défaut est le nom du script batch, ou simplement "sbatch" si le script est lu sur l'entrée standard de sbatch.

-p workq ; Cela indique le nom de la partition à utiliser pour l'allocation des ressources, workq est la partition choisie par défaut :

Partitions (queues)	Access	Priority	Max time	Max threads
workq	everyone	100	4 days (96h)	5272
unlimitq	everyone	1	180 days	3072
interq (runVisusession.sh)	on demand		2 days (48h)	64
smpq	on demand		180 days	96
wflowq	specific software	200	180 days	3072

--time=1-0 ; C'est le temps limite de l'exécution du job, ici cela correspond à 1 jour.

--mem=6G ; Correspond à la mémoire réelle requise par nœud. Les unités par défaut sont les mégaoctets. Différentes unités peuvent être spécifiées en utilisant le suffixe [K|M|G|T].

-e rnaseq_traitmeny.err ; Indique à Slurm de connecter l'erreur standard du script batch directement au nom de fichier spécifié Par défaut, la sortie standard et l'erreur standard sont dirigées vers le même fichier.

--cpu-per-tasks=4 ; Cela permet d'indiquer au contrôleur Slurm que les étapes suivantes de la tâche nécessiteront 4 processeurs par tâche. Sans cette option, le contrôleur essaiera simplement d'allouer un processeur par tâche.

Le fichier se poursuit avec :

```
# mise a zero des modules
module purge

#Chargement du Pipeline Nextflow
module load system/singularity-3.5.3
module load bioinfo/Nextflow-v21.04.1

# Lancement du pipeline
nextflow run nf-core/rnaseq \
-r 3.4 \
-profile genotoul \
--input data/fastq.csv \
--outdir result_rnaseq \
--gtf data/ITAG2.3_genomic_Ch6.gtf \
--fasta data/ITAG2.3_genomic_Ch6.fasta
```

La commande module fait référence au package Environment Modules qui permet la modification dynamique de l'environnement d'un utilisateur via des fichiers du package.
La commande module modifie l'environnement du shell de l'utilisateur :

- ajoute la commande dans le PATH
- ajouter le chemin vers les dépendances
- ajouter le chemin vers des librairies spécifiques

Les modules peuvent être chargés et déchargés dynamiquement. Il est possible de chercher des modules via un mot-clé à l'aide de search_module. Le module purge permet d'enlever tous les modules téléchargés par l'utilisateur et donc de remettre à zéro l'environnement de celui-ci. Le module load charge un module. Il est indispensable de télécharger tous les modules requis avant l'exécution d'un logiciel.

Le logiciel utilisé lors de cette étude est rnaseq à l'aide de Nextflow est un gestionnaire de flux qui assure un déploiement et une reproductibilité efficace des pipelines d'analyse computationnelle. nf-core est un dossier de pipeline d'analyse prêt pour la production et construit avec Nextflow.

Plusieurs attributs ont été passés pour l'appel de rnaseq :

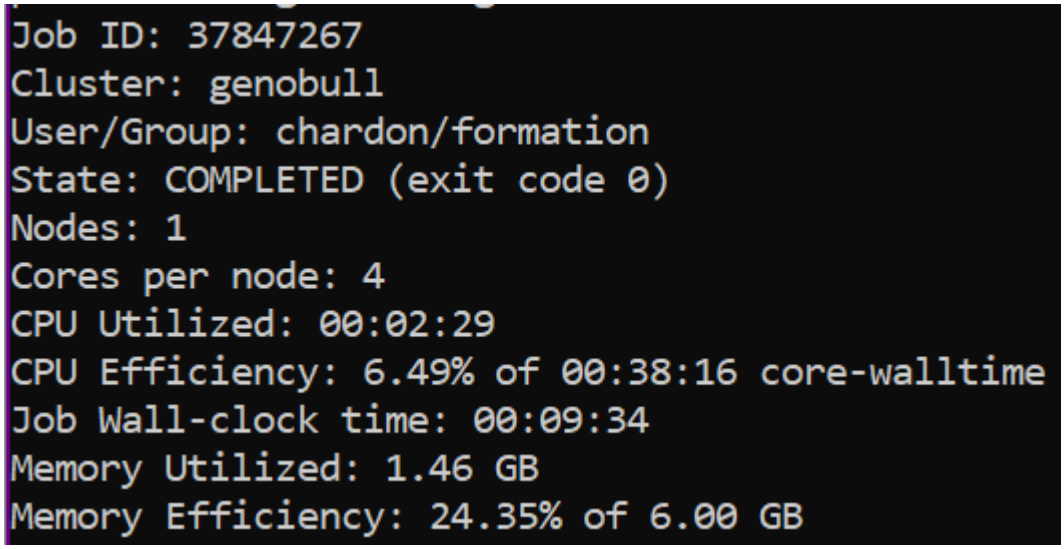
- r 3.4 ; Cela spécifie que la version de pipeline à utiliser est la version 3.4. Cela garantit qu'une version spécifique du code et du logiciel du pipeline est utilisée lorsque le pipeline est exécutée.
- profile genotoul ; On choisit un profil de configuration. Les profils peuvent donner des préréglages de configuration pour différents environnements de calcul.
- input ; Il nous permet de fournir les fichiers d'entrées dont on souhaite le traitement. Il renvoie à une fichier csv :

```
sample,fastq_1,fastq_2,strandedness
1,data/MT_rep1_1_Ch6.fastq.gz,data/MT_rep1_2_Ch6.fastq.gz,unstranded
2,data/WT_rep1_1_Ch6.fastq.gz,data/WT_rep1_2_Ch6.fastq.gz,unstranded
```

  Celui-ci fait référence au chemin d'accès des différents fichiers à analyser.

- outdir ; Cela correspond au nom du dossier qui accueillera toutes les sorties de la commande.
- gtf ; Permet à l'utilisateur de transmettre le génome de référence au format GTF.
- fasta ; Donne une autre possibilité de fournir le génome de référence au format FASTA.

Lors du lancement du job, des erreurs ont amené à faire des modifications. Le nombre par défaut du CPU qui est de 1 ne suffisait pas, donc l'attribut --cpu-per-task a été ajouté pour demander 4 cœurs, ce qui était conseillé par le message d'erreur. Puis une deuxième erreur à vu le jour, le module singularity était inconnu. Alors un nouveau module load a été ajouté pour charger le module singularity. Enfin, le lancement du job ne donne plus d'erreur. Il est possible de voir son déroulement avec seff :

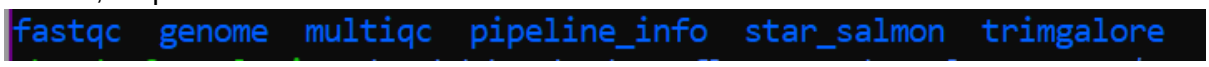


Cette commande vous indique les ressources ayant réellement été utilisées pour exécuter votre job, par rapport à la réservation de ressources initiales. On y retrouve l'identifiant du job, le cluster sur lequel il est exécuté, l'utilisateur qui a lancé le job et le groupe auquel il appartient, le statut de l'exécution, le nombre de nœuds utilisés, le nombre de cœurs par nœuds utilisés, le temps d'utilisation du CPU, l'efficacité du CPU, le temps d'exécution du job, la mémoire utilisée et l'efficacité de la mémoire.

Un job peut aussi avoir un attribut -resume. En utilisant l'attribut -resume, les tâches terminées avec succès sont sautées et les résultats précédemment mis en cache sont utilisés dans les tâches en aval. Ainsi, lors d'une erreur sur un job, le resume permet de ne pas reprendre dès le début du job, mais seulement à partir de la tâche qui a échoué.

Exercice 3 :

Le job fini nous fournit un dossier nommé result_rnaseq (passé par l'attribut outdir). Dans celui-ci, on peut retrouver différents dossiers :



FastQC

Avant d'analyser les résultats pour en tirer des conclusions biologiques, il est nécessaire d'effectuer quelques contrôles de qualité simples pour s'assurer que les données brutes sont bonnes et qu'il n'y a pas de problèmes ou de biais dans vos données.

FastQC a pour but de fournir un rapport de contrôle de qualité qui permet de repérer les problèmes qui proviennent soit du séquenceur, soit de la bibliothèque de départ. Pour cela, il donne des mesures de qualité générales sur les lectures séquencées. Il permet une visualisation graphique des différentes métriques d'intérêt.

Genome

Lors de l'exécution du pipeline, des fichiers sont créés en faisant référence au génome pour de futurs traitements. L'ensemble des fichiers créés sur le génome sont enregistrés dans ce dossier. La création de ces fichiers peut être longue donc cela permet de les réutiliser en économisant de l'espace.

Pipeline_info

Grâce à Nextflow, il est possible de générer divers rapports relatifs au fonctionnement et à l'exécution du pipeline. Cela permet de résoudre les problèmes liés à l'exécution du pipeline et fournit d'autres informations telles que les commandes de lancement, les temps d'exécution et l'utilisation des ressources.

Trimgalore

Ce dossier possède l'ensemble des fichiers générés par l'outil du même nom. Il effectue la détection et le découpage des adaptateurs sur les fichiers FastQ.

Star_Salmon

STAR pour Spliced Transcripts Alignment to a Reference est un aligneur de lecture conçu pour les données de séquençage de l'ARN.Star_salmon est l'option d'alignement par défaut, Salmon est un quantificateur de transcription de données RNA-seq. Il a besoin d'un ensemble de transcrits cibles (provenant d'un assemblage de référence ou de-novo) afin d'effectuer la quantification. Tout ce dont a besoin Salmon est d'un fichier FASTA contenant les transcriptions de référence et un ensemble de fichiers FASTA/FASTQ/BAM contenant les lectures. Les fichiers BAM au niveau du transcriptome généré par STAR sont fournis à Salmon pour la quantification en aval.

MultiQC

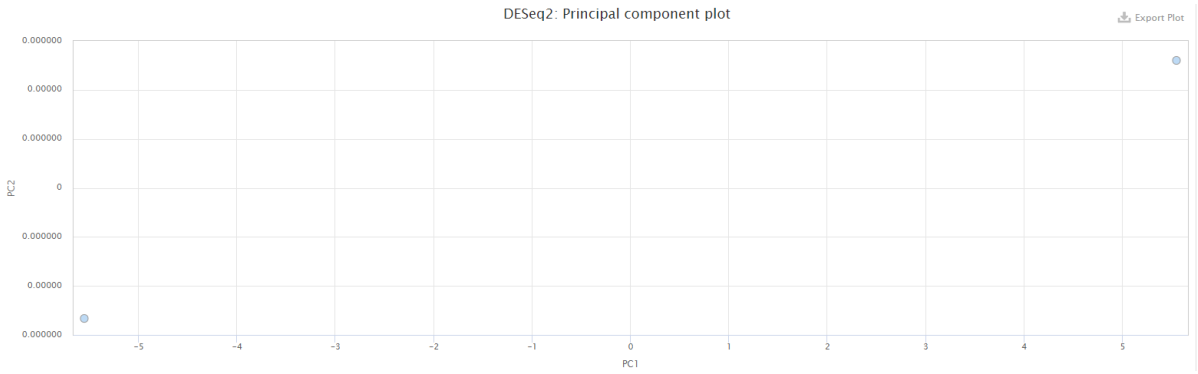
Cet outil permet à l'utilisateur d'observer tous les échantillons d'un job. Dans ce fichier HTML, on peut y retrouver la plupart des résultats du contrôle qualité du pipeline et d'autres statistiques sont disponibles dans le répertoire.

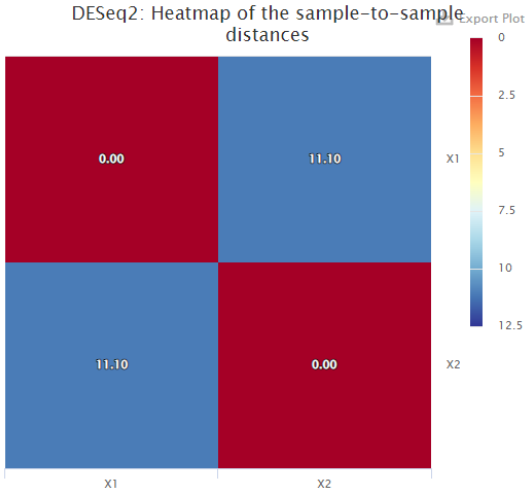
General Statistics

Copy table | Configure Columns | Plot | Showing 8/8 rows and 22/28 columns.

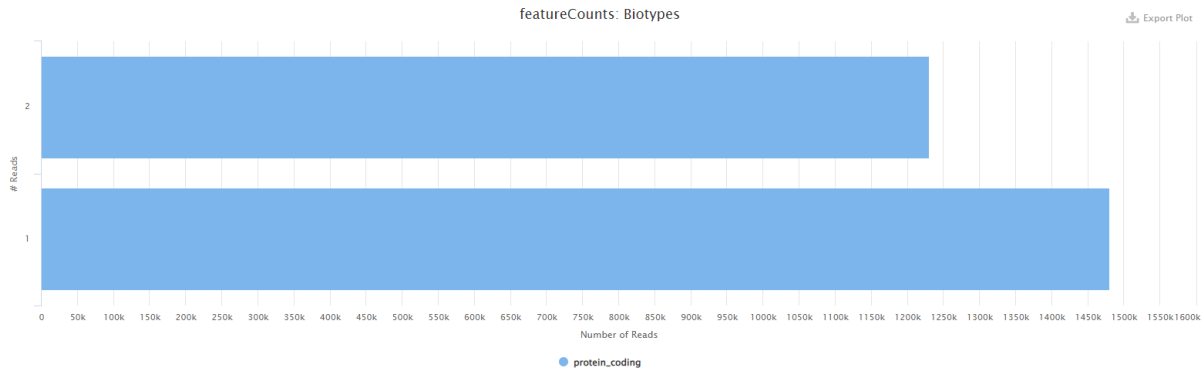
Sample Name	M Reads Mapped	% rRNA	dupInt	% Dups	5'-3' bias	M Aligned	% Proper Pairs	Error rate	M Non-Primary	M Reads Mapped	% Mapped	% Proper Pairs	M Total seqs	% Aligned	M Aligne
1	3.2	0.00%		17.3%	1.44	1.6	80.5%	0.16%	0.0	3.2	100.0%	100.0%	3.2	98.1%	1.6
1.0			0.00%												
1_1															
1_2															
2	2.7	0.00%		18.3%	1.43	1.3	79.1%	0.16%	0.0	2.6	100.0%	100.0%	2.6	98.0%	1.3
2.0			0.00%												
2_1															
2_2															

Un tableau entame la lecture du fichier. C'est un tableau récapitulatif de l'ensemble des statistiques retrouvées dans ce fichier.

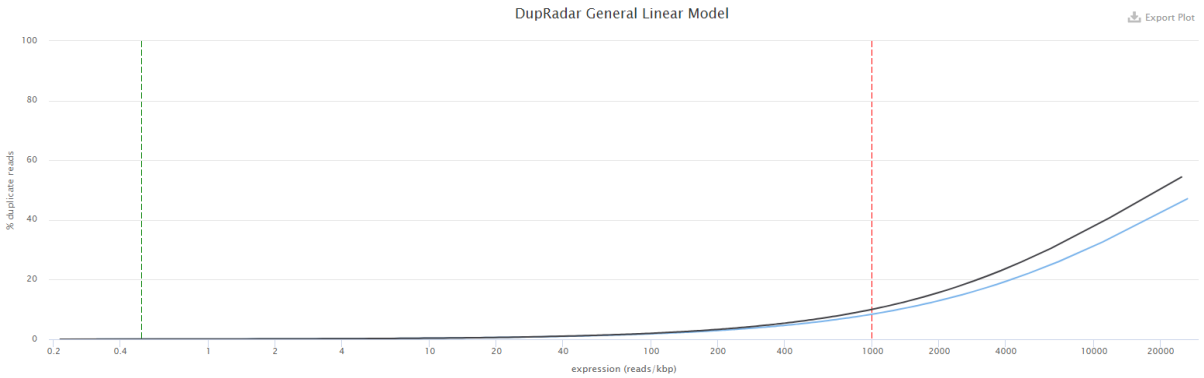




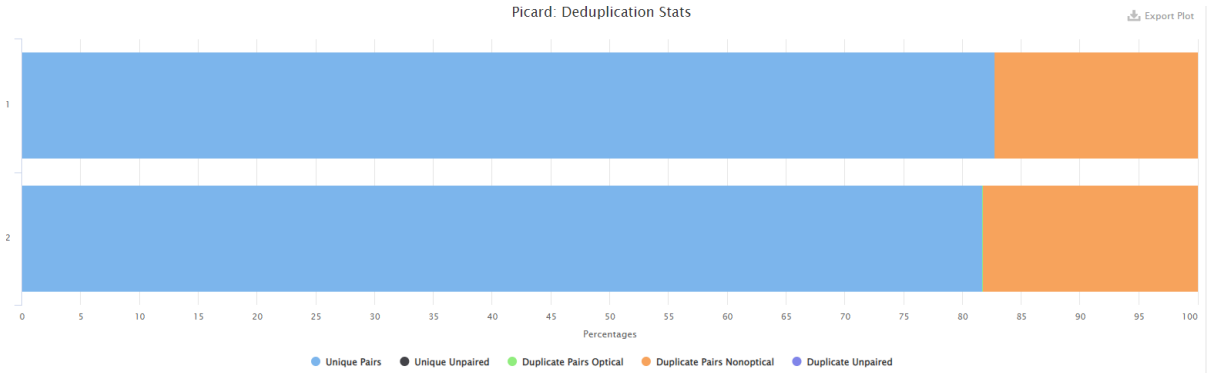
Par la suite, on trouve une ACP réalisée par DESeq2. Ce logiciel est utilisé pour effectuer des analyses d'expression différentielle des donnes RNA-seq. Il peut être utilisé pour avoir une information sur la reproductibilité entre les échantillons. Ici, il normalise le nombre de lectures dans tous les échantillons fournis afin de créer une ACP et une heatmap montrant les distances euclidiennes par paire entre les échantillons de l'expérience.
Cela permet de montrer la similarité entre les groupes d'échantillons et peut révéler les effets de lot et d'autres problèmes potentiels de l'expérience.
Pour notre étude, les deux échantillons sont bien différents.



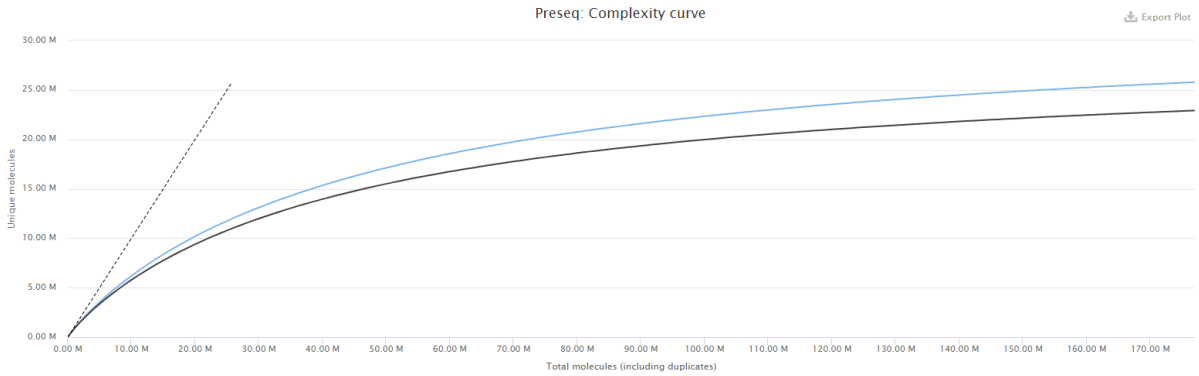
Avec featureCounts provenant de Subread, il est possible de résumer la distribution des lectures. Nos deux échantillons sont à 100% des protéines, le premier échantillon possède plus de reads.



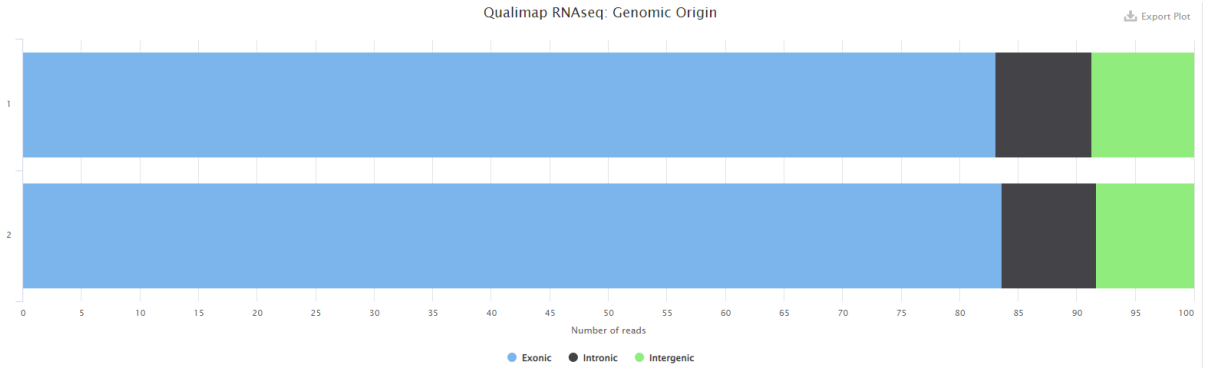
Nous poursuivons avec un graphe obtenu à l'aide de dupRadar. Cet outil de Bioconductor fournit un contrôle qualité du taux de duplication pour l'ensemble des données RNA-seq. On s'attend à retrouver une plus forte duplication sur les gènes les plus exprimés. Donc les résultats montre que les échantillons sont corrects. Mais un trop haut nombre de duplications peut aussi indiquer une faible complexité de la bibliothèque.

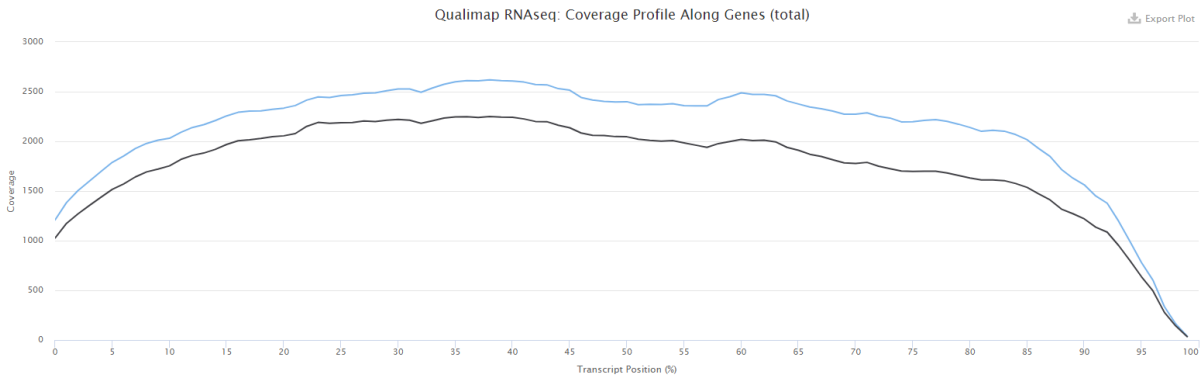


Le logiciel Picard MarkDuplicate marque les lectures dupliquées parmi les alignements afin d'évaluer le niveau global de duplication dans chaque échantillon.
Il est normal de retrouver des duplications en RNA-seq liées aux gènes les plus exprimés.

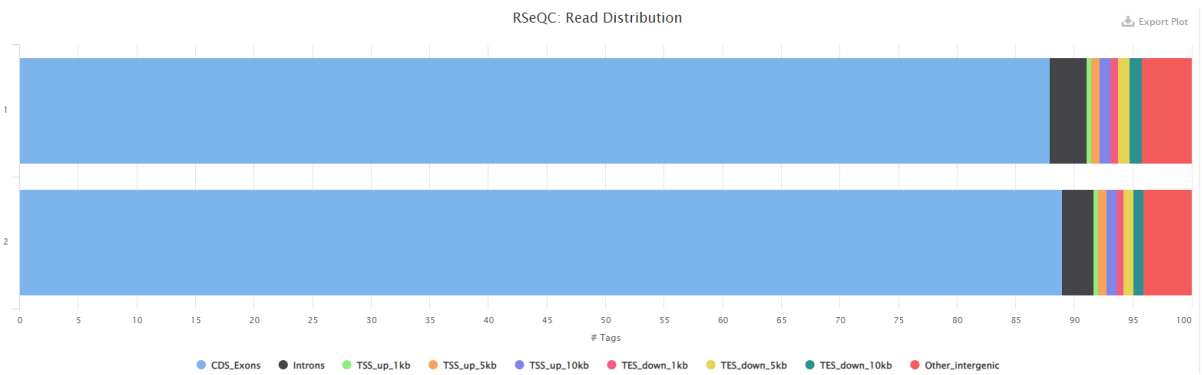


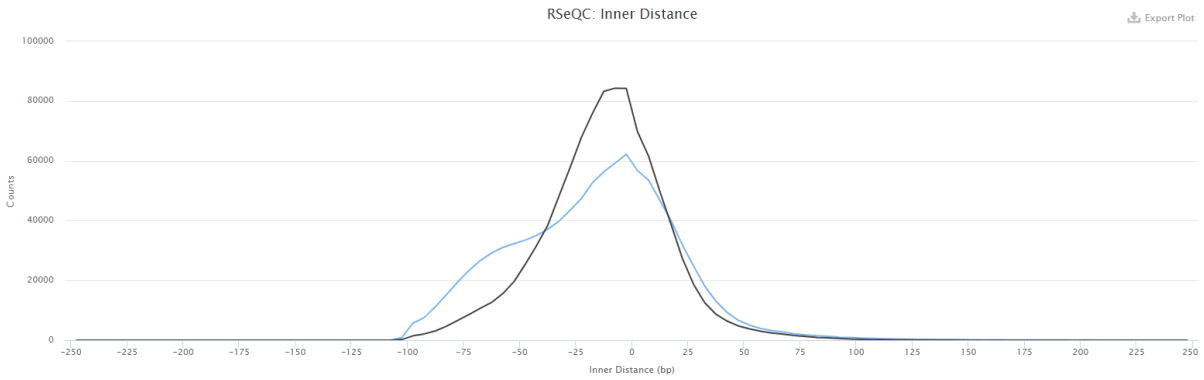
La complexité de la bibliothèque de séquençage peut être prédite et estimée avec l'outil Preseq. La ligne en pointillés montre une bibliothèque parfaitement complexe. Donc la bibliothèque manque de la complexité.



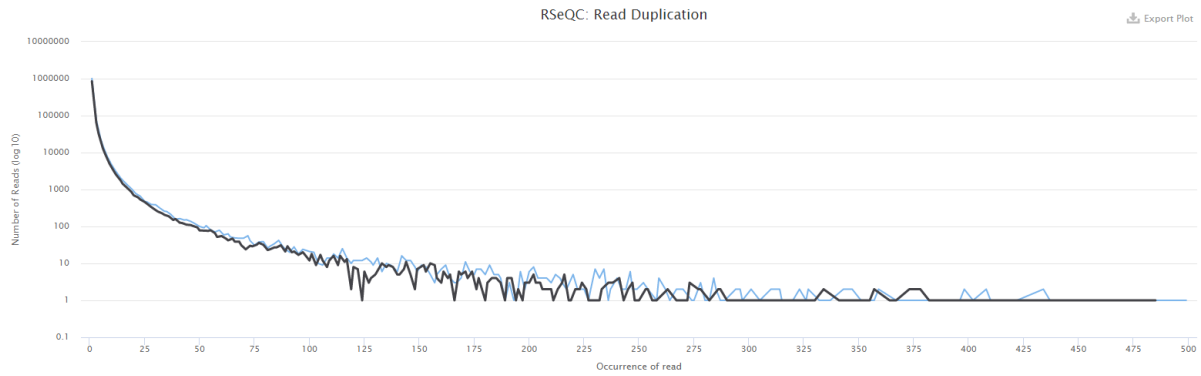


Afin d'évaluer la cartographie générale, l'outil QualiMap est utilisé. Cela classe les lectures selon la région dont elle provient. En majorité, nous retrouvons des exons ce qui est attendu en expérience RNA-seq.

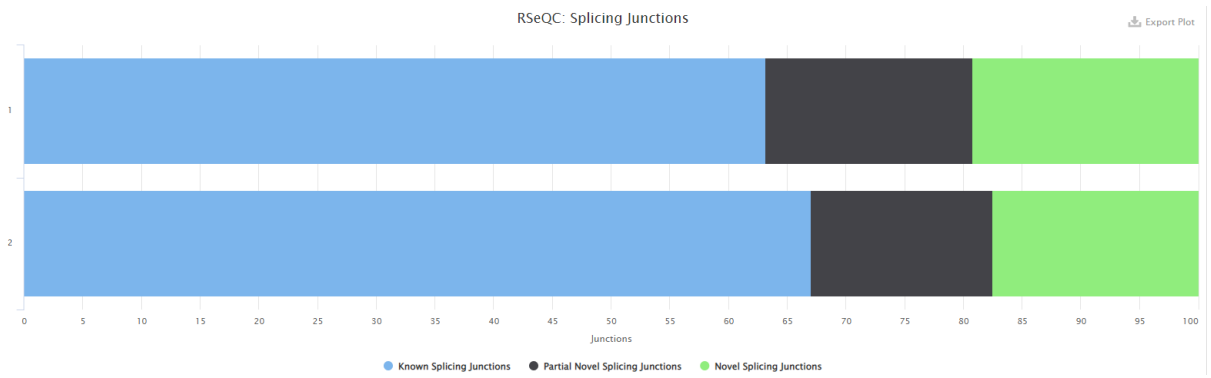


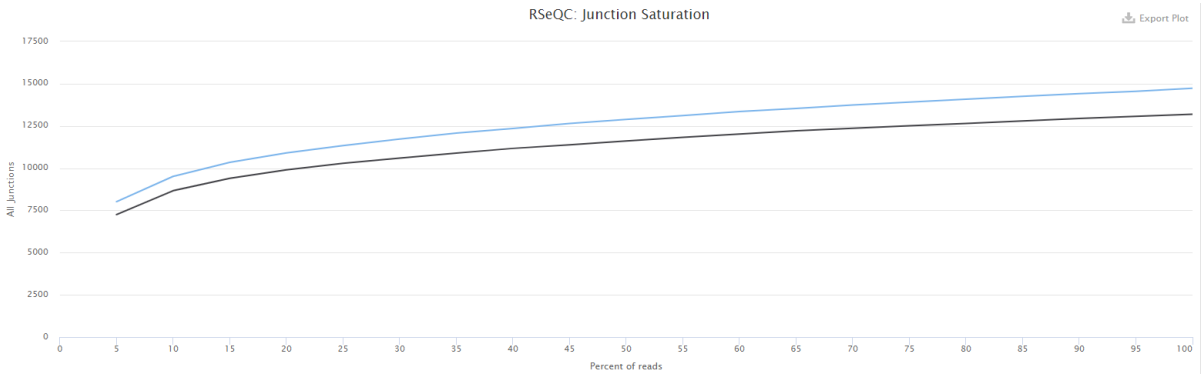


Les lectures sont aussi classées en caractéristiques génomiques par RSeQC. On remarque que les résultats est lié au précédent puisque la majorité sont des CDS donc des séquences codantes retrouvées dans les régions exoniques. Un grand nombre de lectures introniques pourrait indiquer une contamination par l'ADN.
Avec la Inner distance, qui calcule la distance entre deux lectures appariées, on remarque que beaucoup se chevauchent.

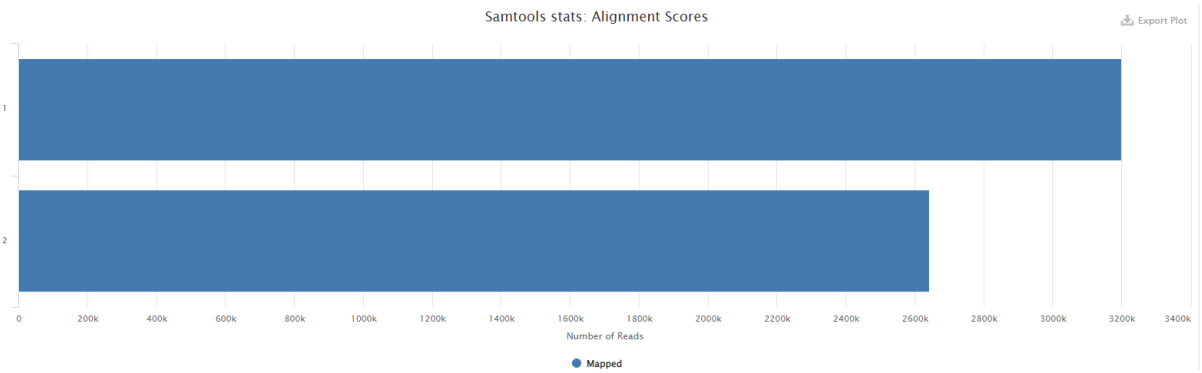


Ce même package permet de vérifier les doublons de lectures. On voit le nombre de lectures en fonction du nombre de doublons. Les échantillons ont un faible taux de doublons, ce qui est attendue dans une RNA-seq.

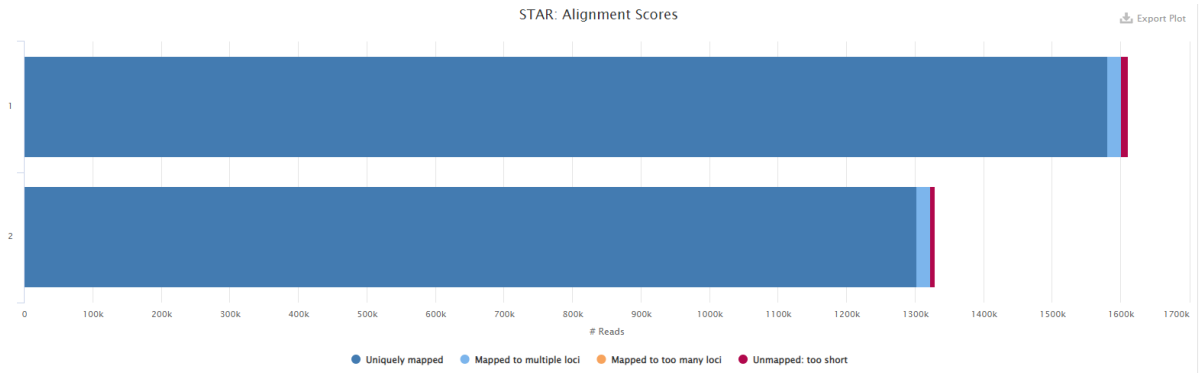




Par la suite, une étude des jonctions est faite. L'annotation des jonctions compare les jonctions d'épissage détectées à un modèle de gène de référence. L'annotation de l'épissage est effectuée à deux niveaux : le niveau de l'événement d'épissage et le niveau de la jonction d'épissage.

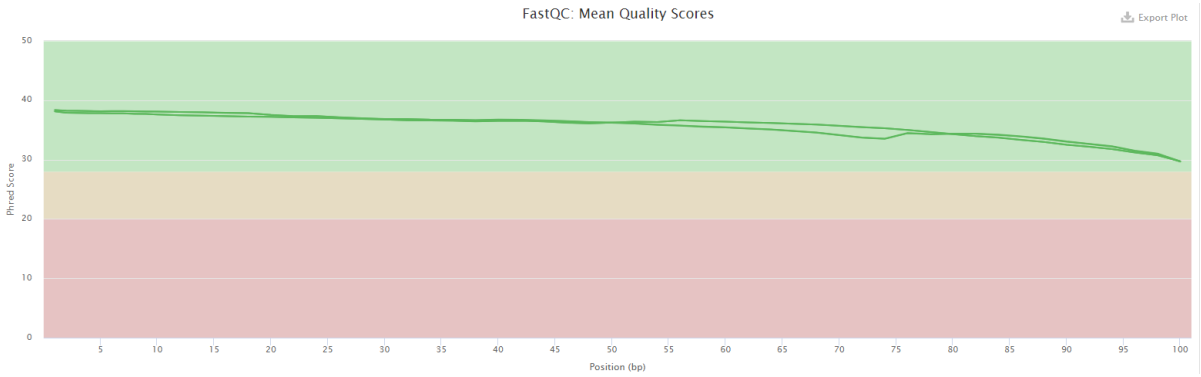


Le pipeline continue avec l'utilisation de SAMtools. Celui-ci propose les fichiers BAM contenant l'alignement pour trier par coordonnées et indexer les séquences sur le génome de référence. Il fournit alors des statistiques de cartographie de lectures.
Pour nos données, l'ensemble des reads des échantillons ont été mappées, c'est-à-dire indexé sur le génome de référence.



Un autre outil d'indexation des lectures sur le génome a été utilisé. Il s'agit de STAR provenant de RSEM. Pour cet outil, plusieurs informations sont données. Premièrement, la majorité des reads ont été mappées une unique fois, moins de 2 % des lectures ont plusieurs possibles indexations et moins de 1% ne sont pas mappés. Cela reste cohérent avec le résultat retrouvé via SAMtools.

Les prochains résultats proviennent de l’outil FastQC présentaient précédemment.



Tout d’abord, un graphe montre une vue d'ensemble des valeurs de qualité à travers toutes les bases à chaque position dans le fichier.

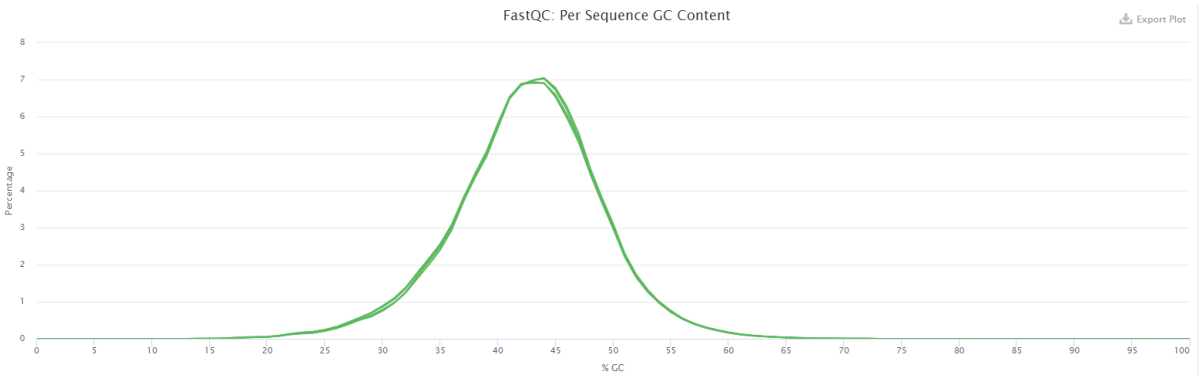
On y retrouve en ordonnées le score de qualité, plus il est haut meilleur sont les séquences à cet indice. Cet axe est divisé en trois couleurs : vert est la zone où les bases sont considérées de très bonnes qualités, orange pour une qualité raisonnable et rouge pour une mauvaise qualité.

La ligne rouge centrale représente la valeur médiane. La boîte jaune représente l'écart inter-quartile. Les moustaches supérieure et inférieure représentent les points de 10% et 90%. La ligne bleue représente la qualité moyenne

La qualité des bases sur la plupart des plates-formes se dégradent au fur et à mesure de la lecture, il est donc fréquent de voir les appels de base tomber dans la zone orange vers la fin d'une lecture. Les 2 échantillons montrent donc de très bonnes qualités des bases à toutes les positions.

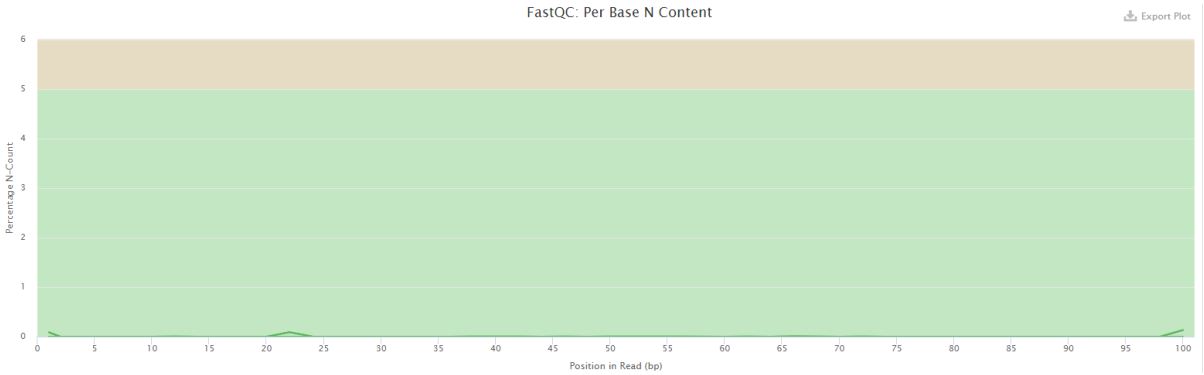
Le rapport sur les scores de qualité par séquence permet de voir si un sous-ensemble des séquences a des valeurs de qualité universellement faibles. Il est fréquent qu'un sous-ensemble de séquences ait une qualité universellement faible, souvent parce qu'elles sont mal représentées, mais elles ne devraient représenter qu'un faible pourcentage du total des séquences.

Pour nos données, la distribution du score de qualité sur toutes les séquences est correcte pour l’ensemble de nos fichiers.



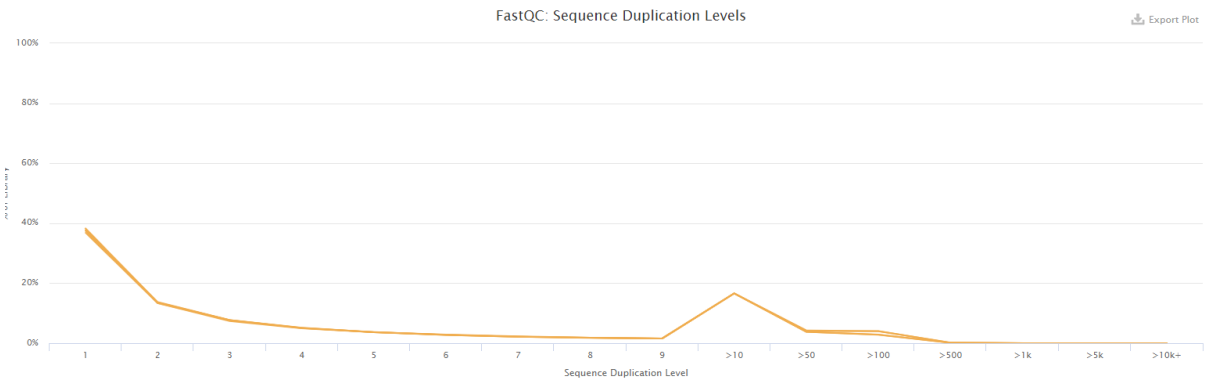
Ce module mesure le contenu GC sur toute la longueur de chaque séquence. Dans une bibliothèque aléatoire normale, on s'attendrait à voir une distribution à peu près normale du contenu GC où le pic central correspond au contenu GC global du génome sous-jacent. Une distribution de forme inhabituelle pourrait indiquer une bibliothèque contaminée ou un autre type de sous-ensemble biaisé. Une distribution normale qui est décalée indique un biais systématique indépendant de la position des bases.

Nos données ont un contenu en GC qui suit une distribution normale.



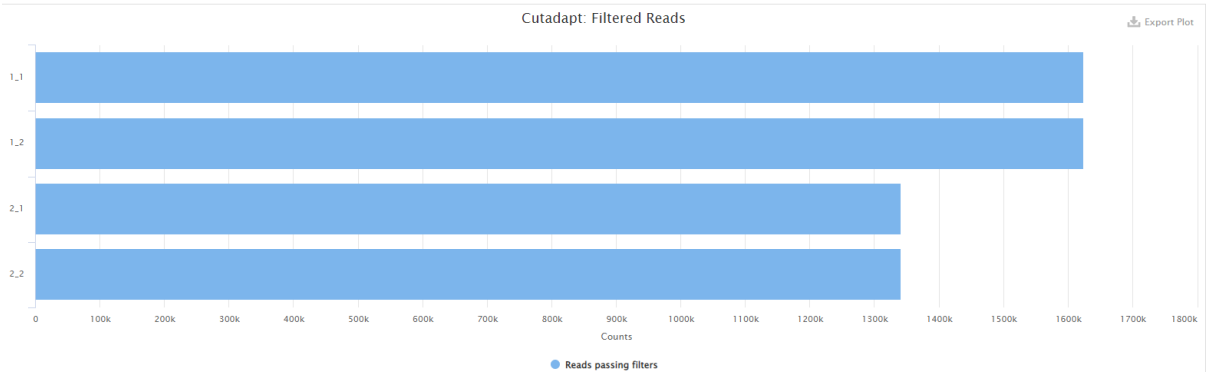
Si un séquenceur est incapable de faire un appel de base avec suffisamment de confiance, il substituera normalement un N plutôt qu'un appel de base conventionnel. Ce module indique le pourcentage d'appels de base à chaque position pour laquelle un N a été appelé.

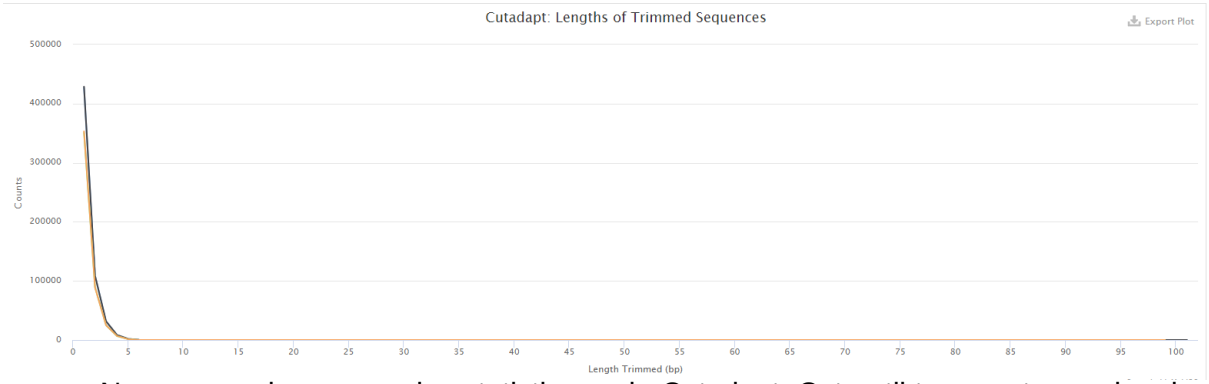
Le taux de N est très faible pour les deux échantillons, les résultats sont donc de bonnes qualités.



Ce module compte le degré de duplication pour chaque séquence dans une bibliothèque et crée un graphique montrant le nombre relatif de séquences avec différents degrés de duplication. Dans une bibliothèque correctement diversifiée, la plupart des séquences devraient se trouver à l'extrême gauche du graphique.

C’est ce qui est retrouvé ici, mais il y a 20% de la librairie qui est fortement dupliquée (>10).





Nous poursuivons avec les statistiques de Cutadapt. Cet outil trouve et supprime les séquences adaptatrices, les amorces, les queuew poly-A et d'autres types de séquences indésirables.

Nous pouvons observer que toutes les lectures ont passé le filtre et qu’elles ont toutes subies la suppression des adaptateurs ayant une longueur inférieur à 5 paires de bases.

Conclusion

Pour résumer l'interprétation des fichiers obtenus par nf-core/rnaseq, il a tout d'abord réalisé une fusion des fichiers reséquencés avec cat, cela donne deux fichiers à analyser. Les fichiers passent l'analyse FastQC qui permet d'avoir plusieurs statistiques sur la qualité des données recueillies. Par la suite, le choix de l'outil d'alignement et de quantification a été celui par défaut (STAR Salmon). Un tri et une indexation des alignements obtenus est réalisé par SAMtools. Cela abouti à la création d'un fichier de couverture. Enfin, un contrôle qualité sur ces résultats est réalisés par plusieurs modules qui sont rassemblés dans le fichier MultiQC.