Mini-Projet Nextflow

Ce tp est réalisé sur le cluster de calcul genotoul avec une connexion ssh sur le compte formation Cosmos.Tout le travail est réalisé dans le répertoire /work/cosmos/NEXTFLOW du compte formation cosmos.

I) PRÉPARATION DE L'ESPACE DE TRAVAIL

Un répertoire data est créé dans le répertoire de travail NEXTFLOW. Ce répertoire contiendra tous les fichiers relatifs aux données. Pour récupérer les données, la commande wget suivie des liens contenant les données, disponibles à l'adresse suivante : http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/

II) FICHIER BASH DE LANCEMENT NEXTFLOW

Dans le répertoire Nextflow est créé le fichier rnaseq.sh contenant le script permettant de lancer la pipeline nf-core/rnaseq sur le cluster.

```
cosmos@genologin2 /work/cosmos/NEXTFLOW $ more rnaseq.sh
#!/bin/bash
#SBATCH --time=00:24:00
#SBATCH -J NassimDuprat
#SBATCH --mem=6G

module purge

module load bioinfo/nfcore-Nextflow-v21.04.1

nextflow run nf-core/rnaseq -r 3.4 --input /work/cosmos/NEXTFLOW/data/sampletab.csv --fasta /work/cosmos/NEXTFLOW/data/genome_ch6.fasta --gtf /work/cosmos/NEXTFLOW/data/genome_Ch6.gtf -profile genotoul
cosmos@genologin2 /work/cosmos/NEXTFLOW $
```

La première ligne de ce fichier indique que ce script bash est fichier de type texte contenant une suite de commandes shell, exécutable par l'interpréteur. Les lignes commençant par #SBATCH décrivent une sorte d'argument optionnel spécifique au cluster de calcul. Ici, on précise le nom du job par l'option -J, le temps maximum attribué au job avec –time. Enfin, la mémoire réservée à ce job. Les lignes qui ne commencent pas par un # sont des commandes bash. On retrouve module load qui permet de nettoyer les modules précédemment utilisés, comme une réinitialisation des modules chargés. bioinfo/nf-core-Nextflow-v21.04.1 est le module qui permet de charger la version 21.04.1 de nextflow. Nextflow nf-core est une sorte de gestionnaire de script bioinformatique qui supervise l'exécution des scripts jusqu'à ce que la pipeline soit terminé. C'est un outil très précieux en bioinformatique car il permet de gagner énormément de temps. Il permet également de reproduire des workflows. Ainsi, on utilise la pipeline rnaseq de Nextflow nf-core pour traiter les données ARN.

La commande de run contient beaucoup d'arguments:

- **-r 3.4** : permet de sélectionner la révision 3.4
- **–input** : indique le chemin du fichier .csv contenant les échantillons fastq de la forme : sample,fastq_1,fastq_2,strandedness
- **–fasta** : indique le chemin du fichier fasta du génome de référence
- **–gtf** : indique le chemin du fichier d'annotation du génome

- **-profile genotoul :** utilise le profil de configuration de genotoul(télécharge et lancera le `genotoul.config` qui a été pré-configuré avec une configuration adaptée au cluster Bioinfo Genotoul), une image docker contenant tous les logiciels requis sera téléchargée

Le script est lancé depuis le genologin frontal avec la commande `sbatch rnaseq.sh`. Hormis une erreur de chemin dans l'option –fasta, il n'y a pas eu d'erreur. Une fois le job fini, on obtient deux répertoires work et results ainsi qu'un fichier texte de type slurm-[jobid].out qui contient ce qu'affiche la sortie sur le terminal.

```
cosmos@genologin2 /work/cosmos/NEXTFLOW $ tail -n 25 slurm-37375110.out
[2f/10a090] process > NFCORE_RNASEQ:RNASEQ:DUPRAD... [100%] 4 of 4 ✔
[80/2e020f] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[29/5e7fc5] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[64/fac68f] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[08/9fa0e0] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[9b/30aaca] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[35/f2300b] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[d2/99e4ff] process > NFCORE_RNASEQ:RNASEQ:RSEQC:... [100%] 4 of 4 ✔
[11/c88093] process > NFCORE_RNASEQ:RNASEQ:MULTIQ... [100%] 1 of 1 ✔
[98/bfdd7e] process > NFCORE_RNASEQ:RNASEQ:CUSTOM... [100%] 1 of 1 ✔
[70/fd3101] process > NFCORE_RNASEQ:RNASEQ:MULTIQ... [100%] 1 of 1 ✔
-[nf-core/rnaseq] 4/4 samples passed STAR 5% mapped threshold:
    97.63%: WT2
    97.58%: WT1
    97.98%: MT2
    98.0%: MT1
-
-[nf-core/rnaseq] Pipeline completed successfully-
Waiting files transfer to complete (1 files)
Completed at: 30-Sep-2022 11:12:47
Duration    : 6m 11s
CPU hours   : 2.7
Succeeded   : 144
```

Voici un aperçu de la fin du fichier qui nous confirme que le job est terminé et sans erreur (au niveau de la commande en tout cas).

Commande seff et l'option -resume :

```
cosmos@genologin2 /work/cosmos/NEXTFLOW $ ls
data  results  rnaseq.sh  slurm-37375110.out  work
cosmos@genologin2 /work/cosmos/NEXTFLOW $ seff 37375110
Job ID: 37375110
Cluster: genobull
User/Group: cosmos/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:02:22
CPU Efficiency: 34.80% of 00:06:48 core-walltime
Job Wall-clock time: 00:06:48
Memory Utilized: 1.95 GB
Memory Efficiency: 32.56% of 6.00 GB
cosmos@genologin2 /work/cosmos/NEXTFLOW $
```

La commande seff s'exécute avec l'id du job, elle permet de récupérer des informations relatives à l'état d'avancement du job, sa consommation en ressource, son temps d'exécution et la l'efficacité de mémoire (mémoire utilisée par rapport à la mémoire réservée) dans le cas où il est terminé. On peut voir également le cluster utilisé (ici genobull, ce qui est étrange car c'est un serveur frontal).

L'option resume est utilisée lors du redémarrage d'une pipeline, elle permet de reprendre les résultats cachés de chaque étape du pipeline où les entrées sont les mêmes, cela permet de consommer moins de ressource et gagner un peu de temps au lieu de reproduire les mêmes résultats.

III) PRINCIPAUX RÉPERTOIRE ET FICHIER DE SORTIE

Comme dit plus haut, deux répertoires sont créés : work et results en plus des fichier log cachés de nextflow. Pour faire simple, le répertoire work contient les fichiers de travail de nextflow et results contient les résultats du job. Ci-dessous, un aperçu des deux répertoires.

```
cosmos@genologin2 /work/cosmos/NEXTFLOW/work $ ls -l
total 54
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:09 01
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:10 02
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:08 03
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:10 05
drwxr-xr-x 4 cosmos formation 4096 Sep 30 11:11 06
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:11 08
drwxr-xr-x 4 cosmos formation 4096 Sep 30 11:09 0e
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:10 0f
drwxr-xr-x 4 cosmos formation 4096 Sep 30 11:11 11
drwxr-xr-x 4 cosmos formation 4096 Sep 30 11:10 15
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:10 17
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:11 18
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:07 1a
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:10 1b
drwxr-xr-x 4 cosmos formation 4096 Sep 30 11:10 1e
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:10 28
drwxr-xr-x 3 cosmos formation 4096 Sep 30 11:11 29
drwxr-xr-x 4 cosmos formation 4096 Sep 30 11:10 2a
```

```
cosmos@genologin2 /work/cosmos/NEXTFLOW/results $ ls -l
total 3
drwxr-xr-x  2 cosmos formation 4096 Sep 30 11:08 fastqc
drwxr-xr-x  4 cosmos formation 4096 Sep 30 11:08 genome
drwxr-xr-x  3 cosmos formation 4096 Sep 30 11:12 multiqc
drwxr-xr-x  2 cosmos formation 4096 Sep 30 12:26 pipeline_info
drwxr-xr-x 17 cosmos formation 4096 Sep 30 11:11 star_salmon
drwxr-xr-x  3 cosmos formation 4096 Sep 30 11:08 trimgalore
```

Results :

Pour commencer on voit que ce répertoire contient 6 sous-répertoires. Pour savoir de quoi il s'agit, la première chose qui vient en tête est d'ouvrir ce répertoire

```
cosmos@genologin2 /work/cosmos/NEXTFLOW/results/fastqc $ ls -l
total 4208
-rw-r--r-- 1 cosmos formation 658458 Sep 30 11:08 MT1_fastqc.html
-rw-r--r-- 1 cosmos formation 416526 Sep 30 11:08 MT1_fastqc.zip
-rw-r--r-- 1 cosmos formation 654774 Sep 30 11:08 MT2_fastqc.html
-rw-r--r-- 1 cosmos formation 412536 Sep 30 11:08 MT2_fastqc.zip
-rw-r--r-- 1 cosmos formation 658630 Sep 30 11:08 WT1_fastqc.html
-rw-r--r-- 1 cosmos formation 415675 Sep 30 11:08 WT1_fastqc.zip
-rw-r--r-- 1 cosmos formation 654982 Sep 30 11:08 WT2_fastqc.html
-rw-r--r-- 1 cosmos formation 412866 Sep 30 11:08 WT2_fastqc.zip
cosmos@genologin2 /work/cosmos/NEXTFLOW/results/fastqc $
```

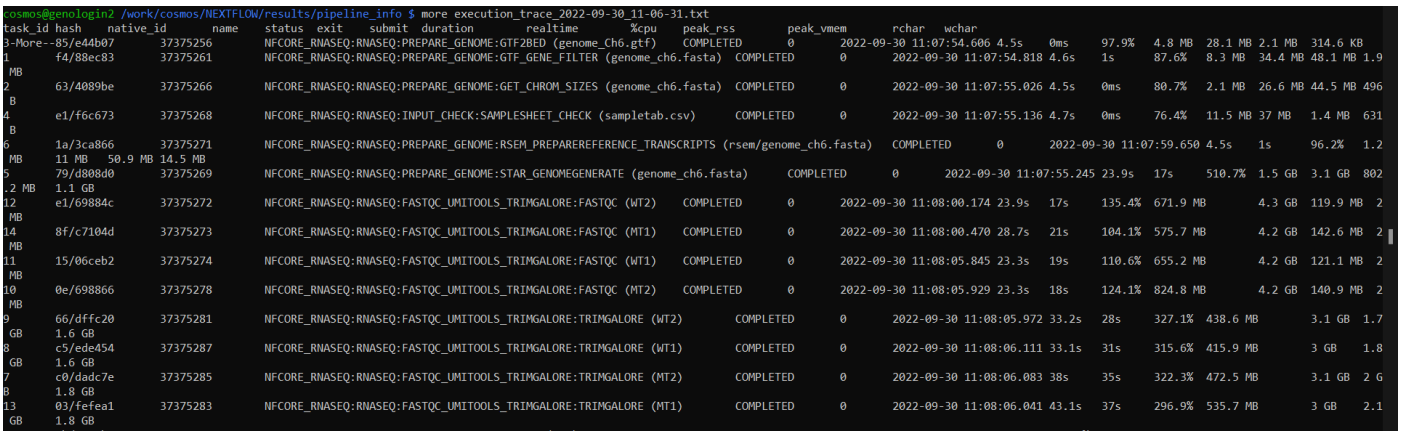
fastqc :
On a un fichier html et un fichier .zip pour chaque échantillon. Le fichier html renvoie vers une page html qui affiche plusieurs graphiques qui permettent de contrôler la qualité de la séquence d'entrée de l'échantillon.

genome :
Ce répertoire contient l'ensemble des fichiers et répertoires relatifs à l'indexation du génome et son annotation. On y retrouve le genome.bed pour l'annotation et le genome.fai pour l'indexation.

multiqc :
Il s'agit de l'agrégation des fichiers de sorties obtenues avec fastqc.

trimgalore :
Ce répertoire contient les fichiers fastqc réduit avec l'outil cutadapt qui enlève les adaptateurs des échantillons. Par conséquent, on trouve dans ce répertoire un sous-répertoire fastqc contenant les html et .zip des échantillons réduits. En plus, il y a des fichiers texte résumant les paramètres et résultats du programme cutadapt. On trouve donc un fichier texte par échantillons, 4 dans notre cas.

pipeline_info :
Le nom du répertoire est explicite, il contient toutes les informations de la pipeline, notamment les échantillons (fichier.csv qui suit –input dans la commande), un fichier texte, et html.



Dans cette capture d'écran, on peut voir une partie du fichier texte qui ressemble au stdout du run nf-core/rnaseq. C'est -à -dire qu' on voit les jobs qui sont lancés simultanément, avec un programme spécifique. Il apparaît la consommation de cpu pour chaque job et leur temps d'exécution.
Dans ce répertoire, on trouve aussi un fichier .yml qui stocke les versions de tous les logiciels utilisés pour ce pipeline.

star_salmon :
Ce répertoire contient beaucoup de sous-répertoires et de fichiers de sorties obtenues après traitement à chaque étape du pipeline. Star_salmon est un workflow de traitement de donné RNA-seq :
1 : lecture et contrôle des séquences input avec Fastqc
2 : réduction si nécessaire avec Cutadapt (par exemple)
3 : compter les lectures qui se chevauchent avec les gènes, par exemple en utilisant features Counts
4 : analyse différentielle de l'expression des gènes (avec deseq2 par exemple)

Voici dans les captures d'écrans, à gauche le dossier star_salmon et à droite la pipeline d'analyse rnaseq de nf-core.

```
cosmos@genologin2 /work/cosmos/NEXTFLOW/results/star_salmon $ ls -l
total 383976
drwxr-xr-x 2 cosmos formation      4096 Sep 30 11:11 bigwig
drwxr-xr-x 3 cosmos formation      4096 Sep 30 11:11 deseq2_qc
drwxr-xr-x 7 cosmos formation      4096 Sep 30 11:11 dupradar
drwxr-xr-x 2 cosmos formation      4096 Sep 30 11:11 featurecounts
drwxr-xr-x 2 cosmos formation      4096 Sep 30 11:09 log
drwxr-xr-x 5 cosmos formation      4096 Sep 30 11:09 MT1
-rw-r--r-- 1 cosmos formation 108019359 Sep 30 11:11 MT1.markdup.sorted.bam
-rw-r--r-- 1 cosmos formation     73920 Sep 30 11:11 MT1.markdup.sorted.bam.bai
drwxr-xr-x 5 cosmos formation      4096 Sep 30 11:09 MT2
-rw-r--r-- 1 cosmos formation 105492420 Sep 30 11:10 MT2.markdup.sorted.bam
-rw-r--r-- 1 cosmos formation     74656 Sep 30 11:10 MT2.markdup.sorted.bam.bai
drwxr-xr-x 2 cosmos formation      4096 Sep 30 11:11 picard_metrics
drwxr-xr-x 3 cosmos formation      4096 Sep 30 11:10 preseq
drwxr-xr-x 6 cosmos formation      4096 Sep 30 11:11 qualimap
drwxr-xr-x 8 cosmos formation      4096 Sep 30 11:10 rseqc
-rw-r--r-- 1 cosmos formation    167740 Sep 30 11:11 salmon.merged.gene_counts_length_scaled.rds
-rw-r--r-- 1 cosmos formation    256263 Sep 30 11:10 salmon.merged.gene_counts_length_scaled.tsv
-rw-r--r-- 1 cosmos formation    133972 Sep 30 11:11 salmon.merged.gene_counts.rds
-rw-r--r-- 1 cosmos formation    174386 Sep 30 11:11 salmon.merged.gene_counts_scaled.rds
-rw-r--r-- 1 cosmos formation    256280 Sep 30 11:10 salmon.merged.gene_counts_scaled.tsv
-rw-r--r-- 1 cosmos formation    142756 Sep 30 11:10 salmon.merged.gene_counts.tsv
-rw-r--r-- 1 cosmos formation    200376 Sep 30 11:10 salmon.merged.gene_tpm.tsv
-rw-r--r-- 1 cosmos formation    150192 Sep 30 11:11 salmon.merged.transcript_counts.rds
-rw-r--r-- 1 cosmos formation    196196 Sep 30 11:10 salmon.merged.transcript_counts.tsv
-rw-r--r-- 1 cosmos formation    253816 Sep 30 11:10 salmon.merged.transcript_tpm.tsv
-rw-r--r-- 1 cosmos formation    160341 Sep 30 11:10 salmon_tx2gene.tsv
drwxr-xr-x 2 cosmos formation      4096 Sep 30 11:11 samtools_stats
drwxr-xr-x 6 cosmos formation      4096 Sep 30 11:11 stringtie
drwxr-xr-x 5 cosmos formation      4096 Sep 30 11:09 WT1
-rw-r--r-- 1 cosmos formation  89559937 Sep 30 11:10 WT1.markdup.sorted.bam
-rw-r--r-- 1 cosmos formation     65224 Sep 30 11:10 WT1.markdup.sorted.bam.bai
drwxr-xr-x 5 cosmos formation      4096 Sep 30 11:09 WT2
-rw-r--r-- 1 cosmos formation  87657997 Sep 30 11:10 WT2.markdup.sorted.bam
-rw-r--r-- 1 cosmos formation     68520 Sep 30 11:10 WT2.markdup.sorted.bam.bai
cosmos@genologin2 /work/cosmos/NEXTFLOW/results/star_salmon $
```

Pipeline overview

The pipeline is built using Nextflow and processes data using the following steps:

- Preprocessing
  - cat - Merge re-sequenced FastQ files
  - FastQC - Raw read QC
  - UMI-tools extract - UMI barcode extraction
  - TrimGalore - Adapter and quality trimming
  - BBSplit - Removal of genome contaminants
  - SortMeRNA - Removal of ribosomal RNA
- Alignment and quantification
  - STAR and Salmon - Fast spliced aware genome alignment and transcriptome quantification
  - STAR via RSEM - Alignment and quantification of expression levels
  - HISAT2 - Memory efficient splice aware alignment to a reference
- Alignment post-processing
  - SAMtools - Sort and index alignments
  - UMI-tools dedup - UMI-based deduplication
  - picard MarkDuplicates - Duplicate read marking
- Other steps
  - StringTie - Transcript assembly and quantification
  - BEDTools and bedGraphToBigWig - Create bigWig coverage files
- Quality control
  - RSeQC - Various RNA-seq QC metrics
  - Qualimap - Various RNA-seq QC metrics
  - dupRadar - Assessment of technical / biological read duplication
  - Preseq - Estimation of library complexity
  - featureCounts - Read counting relative to gene biotype
  - DESeq2 - PCA plot and sample pairwise distance heatmap and dendrogram
  - MultiQC - Present QC for raw reads, alignment, read counting and sample similiarity
- Pseudo-alignment and quantification
  - Salmon - Wicked fast gene and isoform quantification relative to the transcriptome
- Workflow reporting and genomes
  - Reference genome files - Saving reference genome indices/files
  - Pipeline information - Report metrics generated during the workflow execution

Juste en comparant le nom des sous répertoires avec le nom des outils (https://nf-co.re/rnaseq/3.8.1/output) on peut voir que ce répertoire contient les résultats de chaque progiciel utilisé dans le workflow.

VI) INTERPRÉTATION DU REPORT MULTIQC

La première étape est d'ouvrir le fichier html du répertoire multiqc. Des problèmes ont été rencontrés pour l'ouverture depuis une connexion ssh depuis un terminal shell (git bash). Une copie du fichier html avec la commande scp est utilisée pour ouvrir le fichier en local.

Ce fichier comporte plusieurs analyses issues des différents échantillons qui ont été fournies dans la pipeline.

Dans un premier temps, on a un tableau qui fournit des statistiques très générales de l'analyse notamment le pourcentage de de read qui sont mappés au génome de référence. Ensuite, un outil permet d'estimer le sens des read (RSeQC infer_experiment.py) est utilisé en première étape. Il compare le sens fourni par l'utilisateur et celui calculé par cet outil. Comme je ne me suis pas renseigné sur le sens des read, j'ai mis forward par défaut lors de la déclaration des données. Ici, le tableau indique "unstranded" ce qui signifie qu'il n'a pas trouvé le sens des lectures (si on regarde les valeurs pour le premier échantillon par exemple, il indique que la moitié sont sens et l'autre antisens; par conséquent il n'a pas pu départagé pour la totalité des reads).

DESeq2

Le premier outil d'analyse est deseq2, afin de réaliser des analyses en composantes principales entre les différents échantillons. On voit que les échantillons mutants et sauvages sont bien séparés, comme on s'y attendait.

Par la suite, une heat-map sur la base d'un partitionnement des différents échantillons. C'est une autre manière que l'ACP de visualiser la séparation des échantillons sauvages et mutants.

Biotypes Counts

Ce programme quantifie les reads mappés sur différentes régions du génome pour chaque échantillon en utilisant featureounts. On peut voir que tous les read (tous les échantillons) sont mappés à des régions codantes du génome (protein_coding). De plus, il y a plus de reads issues des échantillons mutants mappés sur le génome que de read issue des échantillons sauvages. Ainsi, on compte environ 700k reads des échantillons mutants mappés sur régions codant pour des protéines contre 600k pour les reads des échantillons sauvages.

Dupradar

Cet outil fournit un moyen facile de distinguer la fraction de lectures provenant de la duplication naturelle due à une expression élevée de la fraction induite par des artefacts. dupRadar évalue la fraction de lectures en double par gène en fonction du niveau d'expression. On voit ici qu'il y a corrélation positive entre le % de duplication des reads et le niveau d'expression génique, c'est attendu, cette analyse ne révèle pas d'artefact.

Picard

Cet outil permet en quelque sorte de marquer les lectures en double identifiées parmi les alignements. Il permet d'évaluer le niveau global de duplication dans les échantillons. On peut voir que pour chaque échantillon, on a environ autant de reads uniques que de reads dupliqués. (cependant, je ne sais pas à quel pourcentage on s'attend)

Preseq

Preseq donne un aperçu sur la complexité des librairies, on voit un graphique montrant le nombre de read unique séquencé en fonction du nombre de read de la librairie. Cette courbe permet de voir le nombre de read redondant pour une profondeur de séquençage donnée. La courbe en pointillé indique à quoi ressemblerait une librairie parfaite où chaque read est unique. On constate que plus la taille de la librairie est grande, plus la fraction de

reads uniques diminue jusqu'à un ce que la pente de la droite stagne. Cela indique la saturation de la librairie, c'est-à-dire que toutes les combinaisons de reads ont été vues, et augmenter la taille de la librairie n'augmentera pas le nombre de read unique. Ici, pour une librairie de 12M de reads, on a plus ou moins atteint la saturation pour chaque échantillon.

QualiMap

Cet outil est utilisé pour faciliter le contrôle de la qualité des alignements des données de séquençage.
Il fournit deux visualisation :

Genomic origin of reads : Visualisation de l'origine (intron, exon, intergénique) des lectures mappées sur le génome. On retrouve les même proportions chez les différents échantillons. Les comptages sur MT1 par exemple indiquent environ 690 000 read mappés sur des exons. En fait, on retrouve plus ou moins le nombre de read mappés sur des régions codant pour des protéines (analyse Biotypes Counts).

Gene Coverage Profile : Fournit une vue globale des données qui aide à détecter les biais dans le séquençage et la cartographie des données. La distribution moyenne de la profondeur de couverture des reads est en fonction de la position de l'ensemble transcripts mappés. Les extrémités 5' (100 premières pb) et 3' (100 dernières pb) sont à gauche et à droite respectivement de l'axe X. On s'attend à avoir une faible profondeur aux extrémités du transcript pour un séquençage sans biais.

RSeQC

Cet outil fournit des analyses similaires à celles vues précédemment. En effet, on retrouve dans un premier temps la distribution des reads sur les différentes caractéristiques génomiques (CDS, intron, intergénique). Au vu des chiffres, cette première analyse prend en compte les read dupliqués. Une seconde analyse montre le nombre de read ayant x nombre de duplicatas. On voit que un read peut avoir plus de 400 occurrences. IL n'y a pas de différences notables entre les échantillons mutants et sauvages. Ensuite, L'annotation de jonction compare les jonctions d'épissage détectées à un modèle de gène de référence. Ici, on a environ 70/80% de jonctions qui sont connues. La courbe de saturation des jonctions compte les jonction connues dans chaque échantillons, ici un plateau est atteint alors que les échantillons ne sont pas à 100% de donnée, cela indique toutes les jonctions de la bibliothèque ont été détectées et qu'un séquençage supplémentaire ne produira plus d'observations.

Samtool

La première analyse proposée est à propos des scores d'alignements, chaque échantillon est à 100%. On voit ensuite les statistiques générales fournies après l'analyse. Samtool indique aussi le nombre de read mappés par chromosome, ici tous les reads sont sur le chromosome SL2.40ch06.

STAR

Fastqc

Sur les données brute, on retrouve des analyses déjà effectuées notamment le comptage unique/dupliqué des reads; on y retrouve les mêmes résultats.
Le graphique suivant nous montre la qualité des reads, par par contenu de séquence de base (%A/T/G/C). On constate que les read n'ont pas été contaminés.
Un analyse complémentaire révèle la composition en nucléotide de chaque position pour différents échantillons.
Ensuite, on peut voir la moyenne du contenu en GC des read, chaque échantillon contient environ 43% de GC. L'analyse teneur en base N permet d'indiquer la fiabilité du séquençage. On regarde la proportion de base N pour chaque position, on s'attend à une droite plate. En effet, si cette proportion dépasse quelques pour cent, cela suggère que le pipeline d'analyse n'a pas été en mesure d'interpréter les données suffisamment bien pour effectuer des appels de base valides
Le degré de duplication est montré par la suite, pas de différence notable entre échantillon et on voit que on a 16% des librairies (pour chaque échantillon) qui contient des read avec plus de 10 duplications.

Cutadapt

Cet outil enlève les adaptateurs sur les reads. Le seul résultat disponible montre que tous les reads ont été filtrés par cet outil, autrement dit, tous les read possédaient des adaptateurs. La longueur des adaptateurs va de 1 à 6 paires de base.