Université Paul Sabatier
Master 2 BBS - 2022-2023
RACOUPEAU Martin

Rapport nextflow

1. Préparation

Le dossier « projet_nextflow » a été créé dans le répertoire « iris » et les dossiers telechargés à l'aide d'un wget.

```
wget -rl 2 http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/
```

2. Script de lancement

Le fichier de lancement a été crée sous le nom de launch_pipeline.sh afin de lancer sur la worq un job n'utilisant pas plus de 6gb de mémoire.

```
iris@genologin1 /work/iris/projet_nextflow $ more launch_pipeline.sh
#!/bin/bash
#SBATCH --time=24:00:00
#SBATCH -J projetNextflowM2bioinfo
#SBATCH -e LesProblemesCommencent.out
#SBATCH --mem=6G

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

nextflow run nf-core/rnaseq --reads "TP_TOMATES/WT*{1,2}*.fastq.gz" --fasta TP_TOMATES/ITAG2.3_genomic_Ch6.fasta --gtf TP_TOMATES/ITAG2.3_genomic_Ch6.gtf -profile genotoul
iris@genologin1 /work/iris/projet_nextflow $
```

Le script a été lancé avec sbatch.

```
iris@node109 /work/iris/projet_nextflow $ seff 38068611
Job ID: 38068611
Cluster: genobull
User/Group: iris/formation
State: COMPLETED (exit code 0)
Cores: 1
CPU Utilized: 00:01:06
CPU Efficiency: 13.33% of 00:08:15 core-walltime
Job Wall-clock time: 00:08:15
Memory Utilized: 1.33 GB
Memory Efficiency: 22.21% of 6.00 GB
```

Seff permet de contrôler l'avancement du job, ainsi que l'utilisation du/des CPU et de la RAM :

L'intérêt du resume est de reprendre le job au même endroit ou il à été interrompu, peut importe la raison, ce qui accélère grandement le debugging car les portions fonctionnantes normalement ne sont pas réexécutées à chaque lancement du pipeline.

3. Résultats

Les résultats sont contenus dans le dossier « results » :

```
iris@genologin1 /work/iris/projet_nextflow $ tree -d results/
results/
├── dupradar
│   ├── box_plots
│   ├── gene_data
│   ├── histograms
│   ├── intercepts_slopes
│   └── scatter_plots
├── fastqc
│   └── zips
├── featureCounts
│   ├── biotype_counts
│   ├── gene_counts
│   └── gene_count_summaries
├── markDuplicates
│   └── metrics
├── MultiQC
│   ├── multiqc_data
│   └── multiqc_plots
│       ├── pdf
│       ├── png
│       └── svg
├── pipeline_info
├── preseq
├── qualimap
│   ├── MT_rep1_1_Ch6Aligned.sortedByCoord.out
│   │   ├── css
│   │   ├── images_qualimapReport
│   │   └── raw_data_qualimapReport
│   └── WT_rep1_1_Ch6Aligned.sortedByCoord.out
│       ├── css
│       ├── images_qualimapReport
│       └── raw_data_qualimapReport
├── rseqc
│   ├── bam_stat
│   ├── infer_experiment
│   ├── inner_distance
│   │   ├── data
│   │   ├── plots
│   │   └── rscripts
│   ├── junction_annotation
│   │   ├── data
│   │   ├── events
│   │   ├── junctions
│   │   └── rscripts
│   ├── junction_saturation
│   │   └── rscripts
│   ├── read_distribution
│   └── read_duplication
│       ├── dup_pos
│       ├── dup_seq
│       └── rscripts
├── STAR
│   └── logs
├── stringtieFPKM
│   ├── ballgown
│   │   ├── MT_rep1_1_Ch6Aligned.sortedByCoord.out_ballgown
│   │   └── WT_rep1_1_Ch6Aligned.sortedByCoord.out_ballgown
│   ├── cov_refs
│   └── transcripts
└── trim_galore
    ├── FastQC
    └── logs

61 directories
```

Le dossier dupradar contient les plots et autres comptes-rendus de la qualité du jeu de données RNA-seq majoritairement l'estimation du taux d'artefact introduits par l'amplification PCR.

Le dossier fastqc contient des données sur les lectures, le nombre, leur longueur ainsi que la qualité mais permet aussi de comparer les read avec des adaptateurs connus pour détecter des potentielles contaminations.

Le dossier featureCounts contient combien de read ont été mappés sur des éléments génomiques (gènes, exon, promoteurs…).

Le dossier markDuplicates contient les fichiers BAM où les read dupliqués (issus du même du même fragment d'ADN) sont localisés et marqués.

Le dossier pipeline_info contient les informations sur l'exécution du pipeline : la version, un diagramme du pipeline et les fichiers de rapport d'exécution.

Le dossier preseq contient des informations sur la complexité des données de séquençage.

Le dossier qualimap contient des résultats du contrôle qualité des données mappés, ce qui permet de retirer les biais introduits lors de l'alignement sur le génome.

Le dossier rseqc contient les résultats d'un contrôle qualité similaire à celui de fastqc mais spécifiquement orienté vers les données rnaseq.

Le dossier STAR contient les logs du logiciel d'alignement des reads rna-seq sur un génome de référence.

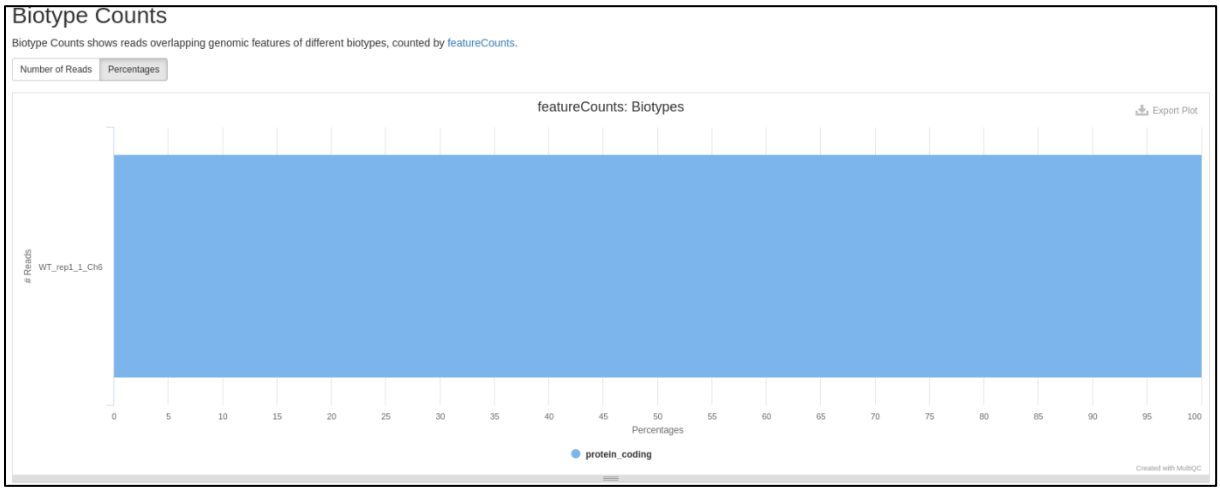
Le dossier stringtieFPKM contient les résultats du logiciel stringtie qui assemble les reads en potentiels transcrits.

Le dossier trim_galore contient les données fastqc car trim galore est une enveloppe de cutadapt et fatqc, mais ce dossier ne contient pas de fichier intermédiaire de cutadapt par défaut.

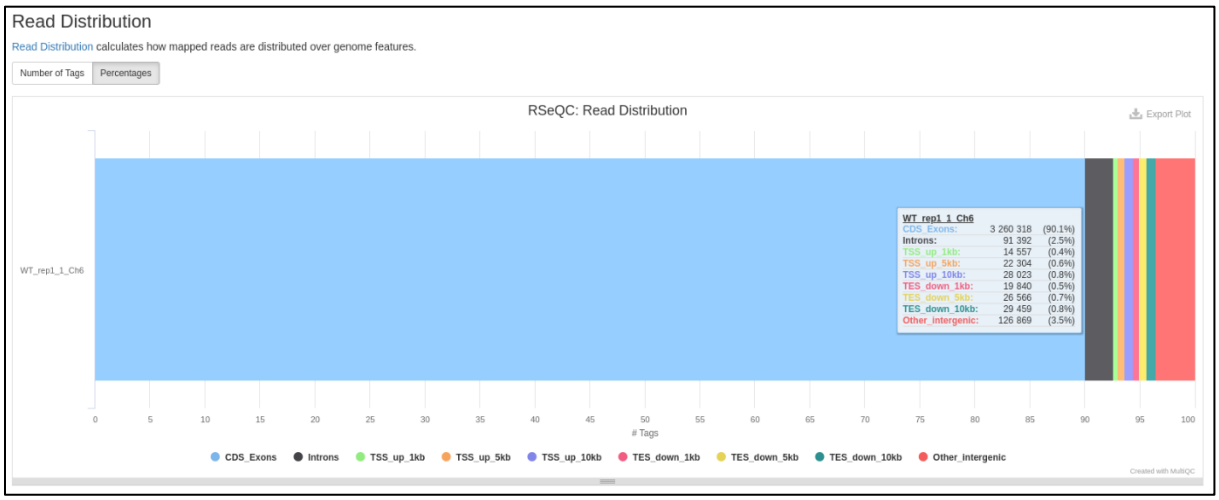
Le dossier MultiQC contient l'agrégat de tous les résultats produits par les différents outils du pipeline : il permet d'observer les différentes statistiques de chaque échantillon dans un seul fichier.

Si on ouvre le pdf MultiQC_report.html, on constate que tous les outils précédemment décrits sont présents :

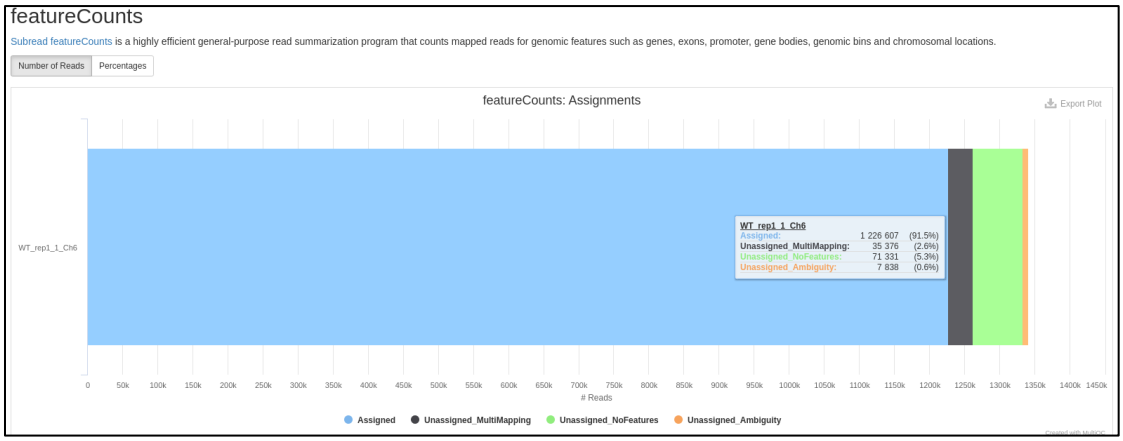
On peut ainsi voir que les reads correspondent exclusivement à des protéines codantes :



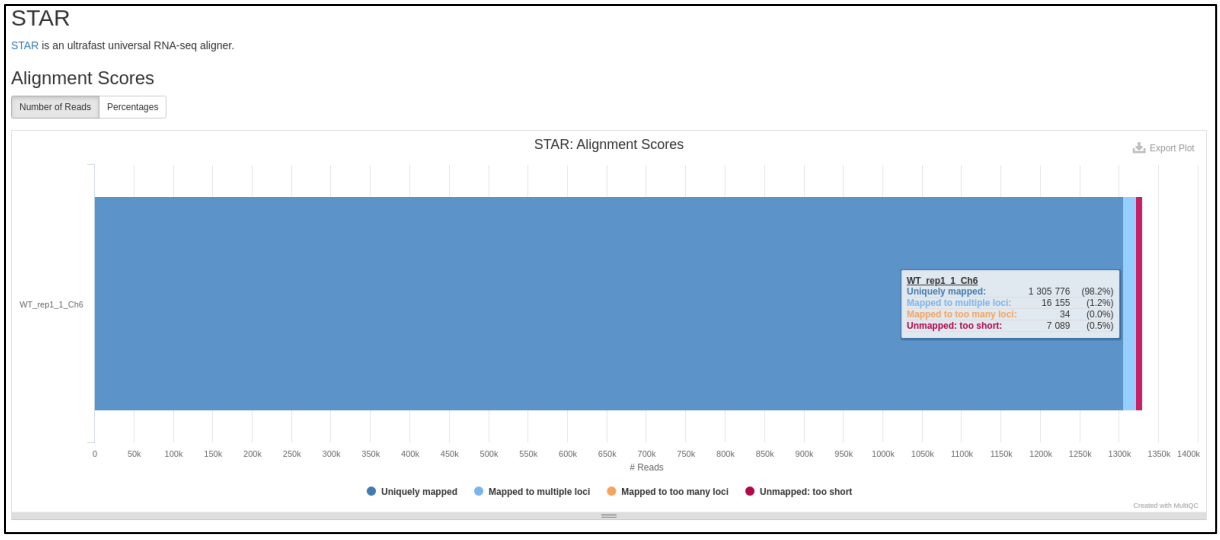
Ce qui est cohérent avec les résultats RSeQC qui montre une majorité de read dans des exons :



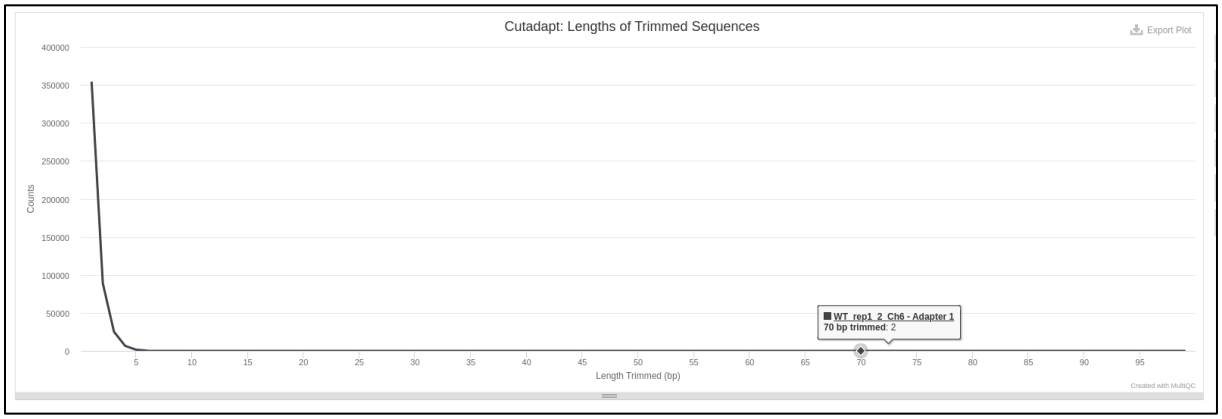
Ce qui explique le résultat similaire produit par featureCounts, une majorité assignée correspondant aux exons :



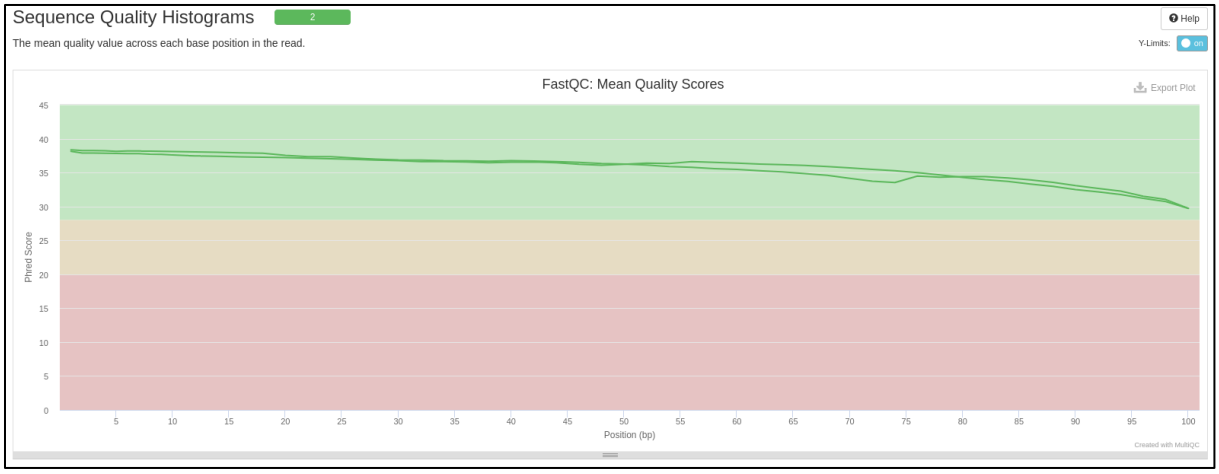
Les résultats de l'alignement des reads sur le génome sont satisfaisant, la grande majorité étant assignée uniquement sur une portion du génome :



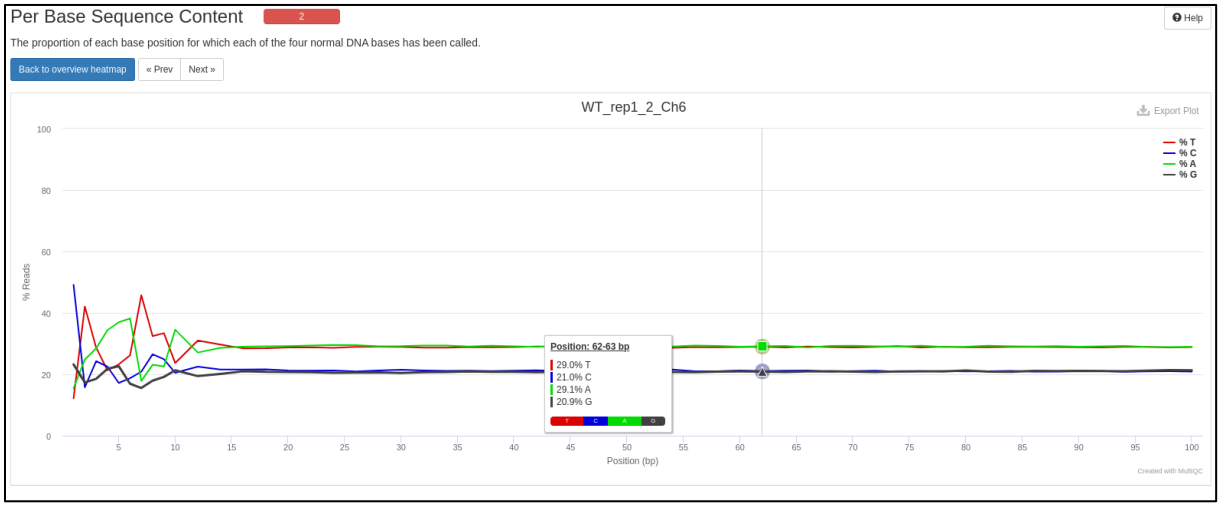
La répartition de séquence retirée par cutadapt semble cohérente et relativement réduite, ce qui est attendu :



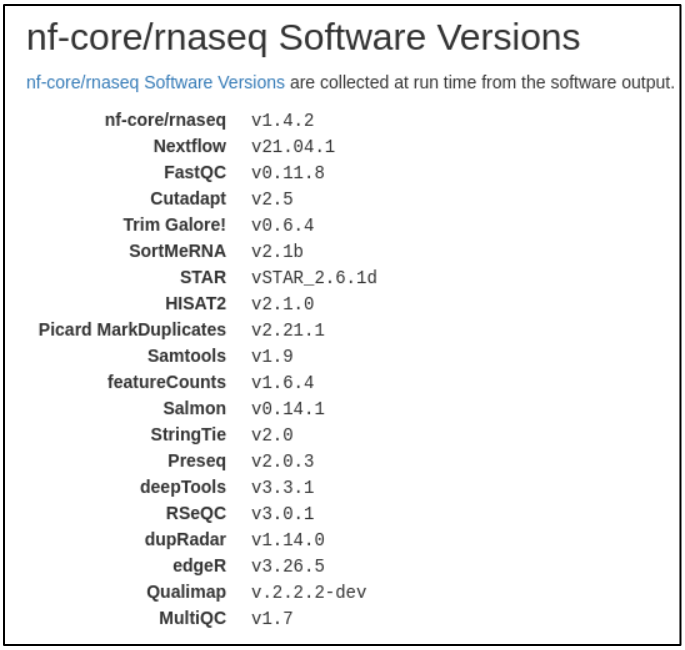
Le compte rendu fastQC montre une qualité de séquençage très satisfaisante avec une légère diminution attendue en fin de séquence :



En regardant le résultat du contenu par base, on remarque une instabilité jusqu'à 12-13 pb qui peut être expliquée par la présence d'adaptateur, conclusion supportée par les résultats de cutadapt. Le reste des positions semble stable ce qui donne une certaine confiance en l'absence de contamination :



Enfin on peut voir les différentes versions des logiciels utilisés :



nf-core/rnaseq Software Versions

nf-core/rnaseq Software Versions are collected at run time from the software output.

nf-core/rnaseq	v1.4.2
Nextflow	v21.04.1
FastQC	v0.11.8
Cutadapt	v2.5
Trim Galore!	v0.6.4
SortMeRNA	v2.1b
STAR	vSTAR_2.6.1d
HISAT2	v2.1.0
Picard MarkDuplicates	v2.21.1
Samtools	v1.9
featureCounts	v1.6.4
Salmon	v0.14.1
StringTie	v2.0
Preseq	v2.0.3
deepTools	v3.3.1
RSeQC	v3.0.1
dupRadar	v1.14.0
edgeR	v3.26.5
Qualimap	v.2.2.2-dev
MultiQC	v1.7