Magdalena Szczuka
M2 BIOINFORMATIQUE

PROJET NEXTFLOW

Nextflow est un logiciel gratuit et accessible à tout le monde, qui permet de lancer les outils diverses pour la recherche bioinformatique. Pendant ce projet, nous avons utilisé surtout le module nf-core de Nextflow, pour effectuer la recherche rnaseq de données biologiques de chromosome 6 de tomate (Solanum lycopersicum).

MATÉRIEL & MÉTHODES

Après avoir téléchargé les données biologiques : séquence fasta, le fichier gtf et 4 fichiers en format fastq, il fallait préparer le script Bash sous format .sh pour pouvoir lancer nextflow sur le cluster Genotoul. Le script contient entre outre:

```
#!/bin/bash
#SBATCH -p workq
#SBATCH --time=1-0
#SBATCH -J MagdalenaSzczuka
#SBATCH -e error_cluster.out
#SBATCH --mem=6G
#SBATCH --cpus-per-task=8

#purge of previous modules
module purge

#load modules
module load bioinfo/Nextflow-v21.04.1
module load system/singularity-3.5.3

#nextflow run command
nextflow run nf-core/rnaseq -r 3.4 -profile genotoul \
--fasta /work/lavande/nextflow_tp/data/ITAG2.3_genomic_Ch6.fasta \
--gtf /work/lavande/nextflow_tp/data/ITAG2.3_genomic_Ch6.gtf \
--input /work/lavande/nextflow_tp/data/tomates_reads.csv
```

- Le temps maximal permis pour effectuer le travail (1 jour)
- Le mémoire réservé sur le cluster (6G)
- CPUs disponibles pour le job (8 CPUs)
- Le chargement de modules nécessaires pour effectuer le travail
- On utilise le module Nextflow v21.04.1 qui n'est pas le plus récent mais il a permis le lancement correcte.
- Module singularity-3.5.3 était nécessaire pour bien lire le contenu de fichiers données en entrée.
- commande principale pour lancer nextflow : dans les paramètres il fallait donner la version de Nextflow utilisé (3.4), le profile genotoul et les fichiers de données biologiques en entrée (fasta, gtf et le fichier csv contenant un Samplesheet de

fichiers fastq. Pour que la commande marche bien, il fallait donner le chemin absolue de fichiers fastq dans un fichier csv).

Après avoir lancé le script .sh dans SBATCH, c'était possible de regarder l'avancement de job en regardant seff du numéro de travail donné par le cluster.
Mon job a obtenu le ID 37374173 et le résultat de la commande $ seff 37374173 est le suivant :

```
Job ID: 37374173
Cluster: genobull
User/Group: lavande/formation
State: COMPLETED (exit code 0)
Nodes: 1
Cores per node: 8
CPU Utilized: 00:02:48
CPU Efficiency: 2.97% of 01:34:24 core-walltime
Job Wall-clock time: 00:11:48
Memory Utilized: 1.44 GB
Memory Efficiency: 24.07% of 6.00 GB
lavande@genologin1 /work/lavande/nextflow_tp $
```

Le job était lancé sur un noeud de cluster, il n'a pas utilisé toutes les CPUs donnés, ni toute la mémoire. On pourrait observer le pourcentage de CPUs utilisés et le temps qu'on a eu besoin pour finir des calculs.

Il existe une autre commande permettant de contrôler l'avancement de job : resume. Si le job n'est pas exécuté à cause d'un erreur et on a modifié le script sh, cette commande permet de reprendre le travail depuis le moment où elle s'est arrêtée, sans devoir tout relancer dès le début. Cela permet d'économiser les ressources et le temps.

RÉSULTATS

Nexflow crée les répertoires suivantes:

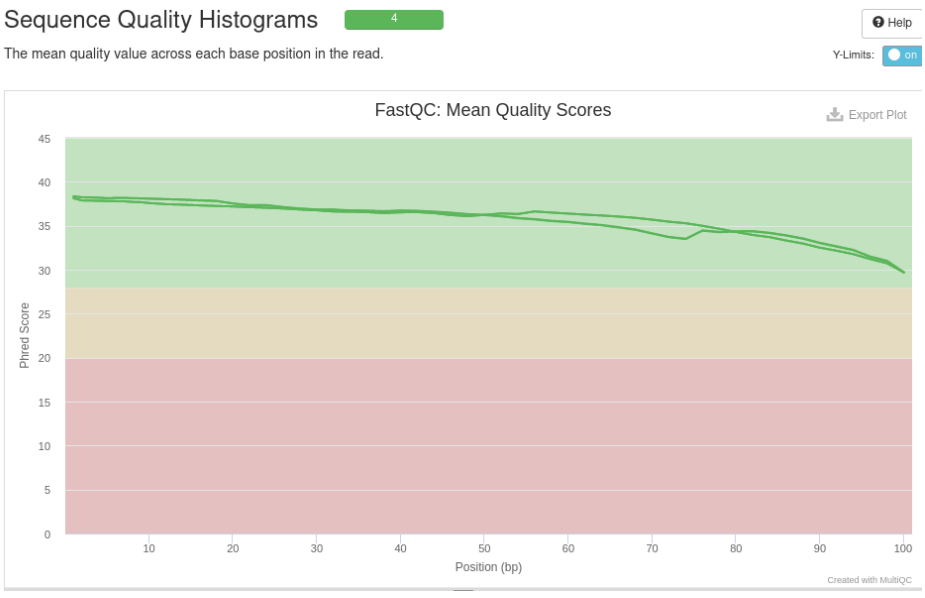
- Work : contient des nombreuses répertoires contenants des informations pour lancer le run correctement.
- Results : et ses répertoires suivants:
  - fastqc : qui contient les résultats sous forme fastq et html pour chacun de fichiers fastq données en entrée dans un fichier csv.
  - genome : contient les fichiers de données génomiques sous le format gtf, bed, fai, etc. Le fichier fai contient l'information par rapport le nom de contig (SL2.40ch06), le nombre de bases dans le contig (46041636), byte index de

fichier contenant le début de la séquence de contig (12), nombre de bases par ligne dans fichier FASTA (80) et bit par ligne dans le fichier FASTA (81).

- pipeline_info : on trouve dedans les informations par rapport à l'execution de notre pipeline.
- star_salmon et trimgalore : ils semblent de contenir les outils et les résultats qu'on visualise avec multiQC
- multiqc : on trouvé dedans une répertoire star_salmon qui contient multiqc_report.html et multiqc_data.

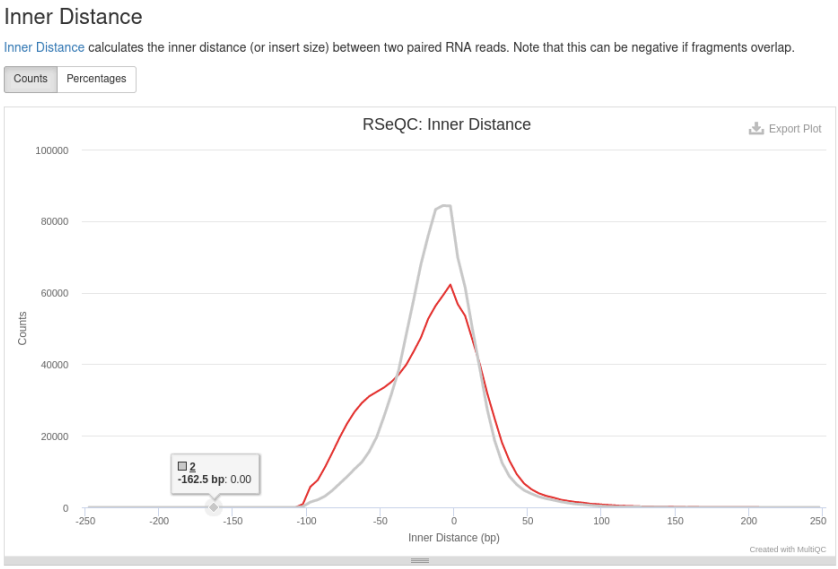
MULTIQC

C'est un outil bioinformatique qui rassemble les analyses d'échantillons multiples dans un seul rapport, visualisé dans un navigateur. On trouve dedans des différents statistiques de % rRNA, le nombre de reads mapped, les informations par rapport à l'alignement, le marge d'erreur, etc. On y trouve également des analyses statistiques comme ACP, Heatmap de similarité entre des échantillons, le modèle linéaire de duplicats de reads, la classification de mapped reads par rapport leur provenance (nos échantillons sont en 83% environ exonic, 8% intronic et 9% intergenic), ou la moyenne de score de qualité de séquence :



En plus, c'est possible de visualiser les résultats propres pour chacune d'échantillons sur chaque plot. Par exemple, pour le plot de FastQC: Mean Quality Scores on a deux courbes - chacun pour une échantillon donnée en entrée. Si on veut visualiser les données d'une échantillon, il faut spécifier le

nom d'échantillon dans un onglet 'Highlight Samples' en haut de la page. Toutes les données par rapport à l'échantillon choisie seront soulignées dans les plots, comme ci-dessous :



SOMMAIRE D'OUTILS

Pour effectuer une analyse et fournir des résultats présentés dans le rapport MultiQC, Nextflow a utilisé des outils suivants :

- DESeq2 PCA plot : Analyse statistique en composantes principales entre des échantillons, permettant de trouver une similarité entre des variables des échantillons.
- DESeq2 sample similarity : donne une distance euclidienne obtenue après le clustering d'échantillons.
- Biotype Counts : visualise les reads qui chevauchent les caractéristiques génomiques
- DupRadar : permet de visualiser le contrôle qualité de taux de duplicatons de datasets. Pour nos échantillons le pourcentage de duplicats augmente jusqu'à 50-55% pour plus que 10000 reads/kbp.
- Picard : outils Java qui permet de manipuler les données du séquençage.
- Preseq : donne une estimation de complexité des libraires et montre combien des reads uniques sont séquences en plus pour augmenter total read count. Cet outil montre également une saturation de complexité.
- QualiMap (Genomic origin of reads, Gene Coverage Profile) : pour une contrôle qualité d'alignement des données séquencées.

- RSeQC (Read distribution, Inner distance, Read duplication, Junction Annotation Junction Saturation, Infer experiment, Bam Stat)
- Samtools (Percent Mapped, Alignment metrics, Flagstat, Mapped reads per contig)
- STAR : permet un alignement ARN universel très rapide
- FastQC (Sequence counts, Sequence quality histograms, Per sequence quality scores, Per base sequence content, per sequence GC content, Per base N content, Sequence length distribution, Sequence duplication levels, Overrepresented sequences, Adapter Content, Status checks)
- Cutadapt (Filtered Reads, Trimmed Sequence Lengths)

Le contenu de répertoire multiqc_data :

```
lavande@genologin1 /work/lavande/nextflow_tp/results/multiqc/star_salmon $ ls multiqc_data/
multiqc.log           multiqc_fastqc_1.txt        multiqc_rseqc_infer_experiment.txt     multiqc_samtools_idxstats.txt
multiqc_cutadapt.txt  multiqc_general_stats.txt   multiqc_rseqc_junction_annotation.txt  multiqc_samtools_stats.txt
multiqc_data.json     multiqc_picard_dups.txt     multiqc_rseqc_read_distribution.txt    multiqc_sources.txt
multiqc_fastqc.txt    multiqc_rseqc_bam_stat.txt  multiqc_samtools_flagstat.txt          multiqc_star.txt
```

Dans les fichiers on trouve des données pour 4 fichiers fastq données en entrée. Ces données sont visualisées dans un rapport de multiqc, mais là on a l'accès à la version numérique. Par exemple le contenu de fichier multiqc_cutadapt.txt est le suivant :

```
lavande@genologin1 /work/lavande/nextflow_tp/results/multiqc/star_salmon/multiqc_data $ more multiqc_cutadapt.txt
Sample  r_processed     r_with_adapters r_written       bp_processed    quality_trimmed bp_written      percent_trimmed
1_1     1624613 578951  1624613 164085913       4900627 158377334       3.4790183359616007
1_2     1624613 580657  1624613 164085913       5182934 158091519       3.65320452585104
2_1     1340546 473901  1340546 135395146       3974452 130763062       3.4211595739185507
2_2     1340546 476095  1340546 135395146       4300469 130434981       3.663473282860524
```

INTERPRÉTATION DE RÉSULTATS

Les résultats affichées dans un rapport multiqc font référence à deux échantillons données en entrée (deux fichiers fastqc). Dans les statistiques générales on peut lire que la première échantillon a 17.3% de duplicats et la deuxième l'en a 18.3%. Sur la figure DupRadar on peut observer que le nombre de duplicats augmente avec le taux d'expression et ce nombre est significativement plus grande pour le taux d'expression au delà de 1000 (reads/kbp). Pour les deux échantillons le taux d'erreur est bas - 0.16%. Le nombre de Mapped reads pour la première échantillon est 3.2M et pour la deuxième échantillon 2.6M. La première échantillon est donc plus grande que la deuxième. Un outil DESeq2 permet de visualiser la distance euclidienne ou la

similarité de 11.12 entre deux échantillons. On peut dire que les deux échantillons ne sont pas très similaires. Cela confirme le PCA plot, où deux échantillons se trouvent aux cotés opposés par rapport à deux axes. La distribution de reads dans les caractéristique de génomes est pareil pour deux échantillons, ce qui visualise la figure de RSeqQC - ReadDistribution. La quality de score semble d'etre bonne - chaque pic de qualité sur la figure FastQC: Per Sequence Quality Scores atteint son max à 38 de moyenne de qualité du score.

CONCLUSION

Ce projet m'a permis d'approfondir les connaissances et d'entrainer de travailler sur un cluster ce qui est très important dans la vie professionnelle d'un bioinformaticien. Il m'a montré l'utilité de création de script SBATCH ce que je vais certainement implémenter en futur.

Ce qui m'a posé un problème, c'était l'interprétation de résultats biologiques obtenues, mais avec la recherche personnelle j'ai essayé de le faire le mieux possible, en espérant que cette connaissance viendra avec de la pratique.