

1 Introduction

2 Données

3 Contrôle qualité

4 Alignement

5 Comptage

TD Analyse RNA-seq sous Linux: Contrôle qualité, alignement et comptage

Ghislain Bidaut, Plateforme Cibi, CRCM, Aix-Marseille Université

24/03/2022

Formation
Bioinformatique
Ghislain BIDAUT
Aix-Marseille Université



1 Introduction

Nous allons analyser des données générées par Yusenko MV *et al*: Expression profiling by high throughput sequencing of THP-1 cells treated with Monensin.

Ces données sont disponibles sur le serveur Gene Expression Omnibus (GEO) à l'adresse

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130657>
(<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130657>)

Pour des raisons de rapidité d'analyse et d'exécution des programmes de traitement, les FASTQ ont été sous-échantillonnés à 10^6 reads avec **seqtk**.

Les données sont disponibles directement sur le cluster IFB (**IFB-Core**) dans le chemin suivant:
`/shared/projects/form_2022_09/data/RNA-seq_GSE130657_Subsampled`

2 Données

2.1 Tableau récapitulatif des données

SRA	GEO	Treatment	Sample
-----	-----	-----------	--------

SRA	GEO	Treatment	Sample
SRR9005674	GSM3746500	THP0	1
SRR9005675	GSM3746501	THP0	2
SRR9005676	GSM3746502	THP0	3
SRR9005677	GSM3746503	M	1
SRR9005678	GSM3746504	M	2
SRR9005679	GSM3746505	M	3

- **THP0**: Pas de traitement
- **M**: Cellules traités avec Monensin.

2.2 TD: Préparation de l'environnement de travail sur le cluster IFB

- Se connecter au cluster **IFB-Core**.
- Créer une arborescence de travail dans votre **répertoire personnel**:
`formation_ngs/data_rna_seq_GSE130657/subsampling_analysis`
- Copier les données fastq.gz, ainsi que le fichiers d'annotation `gencode.v38.annotation.gtf`.
- Copier les scripts d'exécution du pipeline dans ce répertoire.
- Activer Conda puis **créer** et **activer** l'environnement `rnaseq`
- Installer les utilitaires suivants dans cet environnement:
- `fastqc`
- `multiqc`

2.2.1 Solution:

```
ssh <username>@core.cluster.france-bioinformatique.fr

mkdir formation_ngs
cd formation_ngs
mkdir data_rna_seq_GSE130657y
cd data_rna_seq_GSE130657
mkdir subsampling_analysis
cd subsampling_analysis

cp /shared/projects/form_2022_09/data/RNA-seq_GSE130657_Subsampled/*fastq.gz .
cp /shared/projects/form_2022_09/data/RNA-seq_GSE130657_Subsampled/gencode.v38.annotation.gtf .
cp -r /shared/projects/form_2022_09/data/RNA-seq_GSE130657_Subsampled/*sh .

conda create -n rnaseq
conda activate rnaseq
conda install fastqc multiqc
```

3 Contrôle qualité

3.1 TD: Controle Qualité FastQC

- Faire le contrôle qualité des données avec `fastqc` en utilisant **SLURM** sur le compte `form_2022_09` avec le script `fastqc_all.sh`.
- Grouper les résultats en un rapport unique avec `multiqc`
- Rapatrier le rapport `multiqc` en "local".
- Ouvrir ce rapport et commenter les résultats obtenus pour obtenir un bilan de la qualité des données.

3.1.1 Solution:

```
cp /shared/projects/form_2022_09/data/RNA-seq/subsampling-fastq/ .
nano fastqc_all.sh
multiqc .

#Sur la machine locale, ouvrir un Terminal et rapatrier le fichier `multiqc_report.html`
`ainsi que le dossier `multiqc_data`.

scp -rp <username>@core.cluster.france-bioinformatique.fr:/shared/home/<username>/formation_ngs/data_rna_seq_GSE130657/subsampling_analysis/multiqc_report.html .

# Puis à partir du bureau, double-cliquer sur "Répertoire Personnel", puis "multiqc_report.html".
```

3.2 Présentation de Trimmomatic

Maintenant que nous avons examiné nos librairies, nous allons faire du "nettoyage de reads" de manière à supprimer les séquences adaptatrices et assurer une qualité minimale à nos séquences. Pour cela, nous allons opérer un *trimming* avec l'utilitaire **Trimmomatic**.

Nous allons utiliser le manuel (en ligne) pour la rédaction d'une ligne de commande pour Trimmomatic: http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf)

En particulier, il prend en entrée les arguments suivants:

- `in.fq.gz` : Fichier FASTQ d'entrée
- `out.fq.gz` : Fichier FASTQ "trimmé"

3.3 TD: Trimming

- Installer l'utilitaire Trimmomatic dans conda
- Trouver le fichier d'adaptateurs `TruSeq3-SE.fa` avec la commande `find` dans notre répertoire d'installation de Conda. Le copier dans le répertoire de travail.
- Editer le script `trim_all.sh`. Comprendre les différentes parties et repérer la ligne d'appel à **Trimmomatic**. Y ajouter les bonnes options pour que Trimmomatic fasse les actions suivantes:
- **ILLUMINACLIP** : Suppression des adaptateurs (nombre de mis-matches **2**, seuil de qualité cas PE **30**, seuil de qualité cas SE **10**)
- **LEADING** : Suppression des débuts de reads de faible qualité (seuil de qualité **3**)
- **TRAILING** : Suppression des fins de reads de faible qualité (seuil de qualité **3**)

- SLIDINGWINDOWS : Suppressions des zones de faible qualité sur fenêtre glissante à partir de l'extrémité 5' (seuil de qualité **15**, taille de fenêtre **4**)

3.3.1 Solution:

```
# Installation de Trimmomatic sur le cluster IFB-Core.

conda install trimmomatic

# Recherche du fichier d'adaptateurs et le copier dans le répertoire de travail:

find /shared/home/<username>/miniconda3 -name TruSeq3-SE.fa

/shared/home/<username>/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/envs/rnaseq/share/t
rimmomatic-0.39-2/adapters/TruSeq3-SE.fa

cp /shared/home/<username>/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/envs/rnaseq/shar
e/trimmomatic-0.39-2/adapters/TruSeq3-SE.fa .
```

```
# Voici un script bash qui exécute Trimmomatic sur tout les fichiers fastq avec les opt
ions choisies.

#!/bin/bash

#SBATCH --job-name=trimming          # Job name
#SBATCH --ntasks=1                  # Run on a single CPU
#SBATCH --mem=5gb                    # Job memory request
#SBATCH --time=05:00:00              # Time limit hrs:min:sec
#SBATCH --output=trimming_%j.log     # Standard output and error log

for fastqgz in $(ls *sub*.fastq.gz)
do
    echo "process $fastqgz"

    trimmed=${fastqgz/.fastq.gz/.trim.fastq.gz}
    log=${fastqgz/.fastq.gz/.trim.log.txt}
    summ=${fastqgz/.fastq.gz/.trim.summary.txt}
    trimmomatic SE -trimlog $log -summary $summ $fastqgz $trimmed ILLUMINACLIP:TruSe
q3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15

done
```

Détail des options:

```

#- `ILLUMINACLIP:TruSeq3-SE.fa:2:30:10`: Suppression des séquences adaptatrices correspon-
dant au kit Illumina *TruSeq Stranded Total RNA*.
#- `LEADING:3`: Supprime les bases à partir du début, si elles sont en dessous de la qu-
alité 3.
#- `TRAILING:3` Supprime les bases à partir de la fin, si elles sont en dessous de la qu-
alité 3.
#- `SLIDINGWINDOW:4:15`: supprime les bases aux extrémités dont la qualité est inférieu-
re à 15 sur une fenetre glissants de taille 4.

# Pour couper systématiquement le début et la fin des séquences, on peut utiliser les o-
ptions:

#- `CROP:n` Coupe n bases depuis la fin du read
#- `HEADCROP:n` Coupe n bases depuis le début du read

# Enregistrer ce script dans le fichier `trimall.sh` et l'exécuter:

sbatch -A form_2022_09 trimall.sh
watch squeue -u <username>

```

Après avoir exécuté Trimmomatic, on peut refaire tourner `fastqc+multiqc` pour vérifier **l'effet du trimming**.

```

fastqc *trim.fastq.gz
multiqc .

```

4 Alignement

4.1 Indexation du génome de référence

Nous utilisons l'utilitaire **STAR**, spécialement conçu pour les alignements RNA-seq.

Dans le cadre de son intégration dans un pipeline bioinformatique, **STAR** fonctionne en 2 étapes:

- **Génération d'un index du génome de référence.** A cette étape, l'utilisateur doit apporter la séquence du génome de référence. STAR va pouvoir générer les indexes qui seront utilisés lors de la seconde partie, l'alignement proprement dit. Cet index a besoin d'être généré une seule fois et peut être réutilisé (uniquement avec STAR).
- **L'alignement sur le génome.** A cette étape, l'utilisateur doit apporter les indexes générés, les séquences à aligner et le programme va aligner les fragments et générer un fichier BAM, ainsi que des statistiques.

4.2 Option 1: pas de TD indexage.

On récupère le génome indexé directement:

```

cp /shared/projects/form_2022_09/data/RNA-seq_GSE130657_Subsampled/hg38.p13.chr8 .

```

4.3 Option TD.

On Récupère les données brutes et on indexe le génome nous même:

Obtention d'un génome de référence. Il existe plusieurs sites majeurs.

- UCSC: <https://hgdownload.soe.ucsc.edu/downloads.html>
(<https://hgdownload.soe.ucsc.edu/downloads.html>)

Exemple: **Fichier Fasta pour HG38 patch release 13:**

<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz>
(<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz>)

- Genecode: <https://www.gencodegenes.org/> (<https://www.gencodegenes.org/>)

Ce site se concentre sur les génomes humain et murins et contient à la fois les séquences des génomes (Fichier **FA**) et annotations (Fichier **GTF**)

- **Fichier FA:** [GRCh38.p13.genome.fa.gz](http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.p13.genome.fa.gz)
(http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.p13.genome.fa.gz)
- **Fichier d'annotation GTF:** [gencode.v38.annotation.gtf.gz](http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.annotation.gtf.gz)
(http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.annotation.gtf.gz)

Le format GTF est défini sur le site de l'UCSC: (<http://genome.ucsc.edu/FAQ/FAQformat>
(<http://genome.ucsc.edu/FAQ/FAQformat>)).

4.3.1 TD: Indexation avec STAR

- Récupérer le génome de référence humain sur le site de Genecode
- Récupérer un fichier d'annotation de génome humain sur le site Genecode (Format GTF)
- Installer **STAR** et **samtools** sous Conda.
- Faire une extraction du chromosome 8 du génome de référence avec `samtools`
- Créer un script `star_chr8_index.sh` pour indexer le génome (ici limité au chr8) avec **STAR**. Voir les options de STAR sur la doc
https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf
(https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf)
- Faire tourner ce script sur le cluster avec `sbatch`.

4.3.2 Solution

```
# Installation de `star`, `samtools` et récupération des données:

conda install star samtools

mkdir hg38.p13.chr8
wget http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.p13.genome.fa.gz
mv GRCh38.p13.genome.fa.gz hg38.p13.chr8

wget http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.annotation.gtf.gz
```

```

# Extraction du chromosome 8 du génome humain et indexation:

cd hg38.p13.chr8
gunzip GRCh38.p13.genome.fa.gz
samtools faidx GRCh38.p13.genome.fa chr8 > GRCh38.p13.chr8.fa
rm GRCh38.p13.genome.fa
cd ..

# Mettre le code suivant dans 'star_chr8_index.sh':

#!/bin/bash

#SBATCH --job-name=starchr8index          # Job name
#SBATCH --ntasks=20                      # Run on 8 threads
#SBATCH --mem=10gb                       # Job memory request
#SBATCH --time=05:00:00                  # Time limit hrs:min:sec
#SBATCH --output=starchr8index_%j.log    # Standard output and error log

STAR --runThreadN 20 --runMode genomeGenerate --genomeDir hg38.p13.chr8 --genomeFastaFiles hg38.p13.chr8/GRCh38.p13.chr8.fa --sjdbGTFfile gencode.v38.annotation.gtf --sjdbOverhang 74

# Puis
sbatch -A form_2022_09 ./star_align_all.sh

```

4.4 TD: Alignement des séquences:

Une fois le génome de référence obtenu et indexé, écrire un script pour aligner nos séquences trimmées avec **STAR** et stocker les **BAM** résultants dans `star_aligned`. Générer les indexes des fichiers `bam` avec **samtools**. Le script doit se lancer sur le cluster.

4.4.1 Solution

Voici un script bash qui exécute `star` sur tout les fichiers fastq trimmés:

```
#!/bin/bash

#SBATCH --job-name=staralign          # Job name
#SBATCH --ntasks=20                  # Run on 8 threads
#SBATCH --mem=40gb                    # Job memory request
#SBATCH --time=05:00:00               # Time limit hrs:min:sec
#SBATCH --output=staralign_%j.log     # Standard output and error log

for fastqgz in $(ls *sub*.trim.fastq.gz)
do
    echo "process $fastqgz"

    pref=${fastqgz/.fastq.gz/_}

    bam=${fastqgz/.fastq.gz/Aligned.sortedByCoord.out.bam}

    STAR --runThreadN 20 --genomeDir hg38.p13.chr8 -- readFilesIn $fastqgz --readFiles
Command gunzip -c --outSAMtype BAM SortedByCoordinate --outFileNamePrefix star_aligne
d/$pref

    samtools index -@ 20 $bam

done
```

On enregistre ce script sous `staralign.ah` . On l'exécute par slurm par:

```
sbatch -A form_2022_09 staralign.sh
watch squeue -u <username>
```

4.5 Visualisation des BAMs sous IGV

IGV (*Integrative Genomics Viewer*) (Ne pas confondre avec IGB...) est un programme graphique de visualisation de données génomiques sous forme de *pistes* (tracks). Il permet de naviguer de manière linéaire sur des coordonnées chromosomiques.

Il permet de **visualiser les fragments de lecture alignés (format BAM)**. **Important:** Il faut que les fichiers BAM **soient triés et indexés**.

Un algorithme permet de visualiser les mutations (SNPs et Indels) par un code couleur.

IGV peut également afficher des **reads appairés**.

On peut le télécharger pour les OS les plus courants à l'adresse <https://software.broadinstitute.org/software/igv/> (<https://software.broadinstitute.org/software/igv/>) ou l'installer sous conda.

Il permet également de superposer d'autres données (variants, bigWig, BED, etc...).

4.5.1 TD: Visualisation des fichiers BAM sous IGV

Installer, puis lancer IGV et visualiser 2 échantillons BAM **en local (pas sur le cluster)**.

4.5.2 Solution


```
# On lance un terminal sur notre machine en local. On active conda et on installe IGV:
conda install igv

# On récupère les fichiers `bam`et `bai` en 'local'

mkdir bam
scp -rp <username>@core.cluster.france-bioinformatique.fr:/shared/home/<username>/formation_ngs/data_rna_seq_GSE130657/subsampling_analysis/star_aligned/SRR*ba* bam

# Enfin, on lance IGV et on ouvre les échantillons récupérés.

igv
```

5 Comptage

5.1 TD: Comptage des reads

- Installer l'utilitaire `featurecounts` (avec `s`) dans l'environnement Conda `rnaseq` (package `subread`)
- Appliquer `featureCounts` sur les fichiers BAM obtenus par STAR pour obtenir un fichier de comptage en format texte délimité par des tabulation.

5.1.1 Solution

```
# Sur le cluster IFB, installer **subread**:

conda install subread

# Ecrire le script suivant et le stocker dans `featurecount.sh`.

#!/bin/bash

#SBATCH --job-name=featurecount          # Job name
#SBATCH --ntasks=1                      # Run on 1 core
#SBATCH --mem=5gb                       # Job memory request
#SBATCH --time=1:00:00                 # Time limit hrs:min:sec
#SBATCH --output=featurecount_%j.log    # Standard output and error log

featureCounts -a gencode.v38.annotation.gtf -o counts_GSE130657_subsampled.txt star_aligned/*bam

# Puis faire tourner la commande suivante:

sbatch -A form_2022_09 featurecount_all.sh

watch squeue -u <username>
```

5.2 TD: MultiQC final

- Refaire un rapport multiQC avec l'ensemble des analyses
- Importer ce rapport en local, l'analyser et le commenter

5.2.1 Solution

Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International (CC BY-NC-ND 4.0).