Formation Bioinformatique pour le traitement de données de séquençage

Récupération et prétraitement des données

Annabelle Haudry

Contents

Programme:

- Analyse qualité des lectures
- manipuler les outils et les lancer en ligne de commande
- transfert de données

Objectifs:

- identifier et récupérer de jeux de données sur la base SRA
- format fastq
- analyse qualité et interprétation
- Filtrage et nettoyage des données

Récupération de données

base de données ENA (European Nucleotide Archive) https://www.ebi.ac.uk/ena C'est l'option la plus directe, car on récupère les fichiers au format fastq... ou presque, car ils sont compressés, donc fastq.gz

exercice

Récuperez le jeu de données suivant de séquencage de génome de *Staphylococcus aureus* : en HiSeq4000 (*SRR7748059*), dans un dossier "data" que vous devez créer. Une fois le jeu de données identifié, vous devez récupérer le lien url de téléchargement: dans le tableau, dans la colonne "Generated FASTQ files: FTP", cochez les fichiers à récupérer puis cliquez sur "Get download script".Les lignes de commandes pour le téléchargement des fichiers ont été téléchargées dans un fichier ".sh". Copiez ces lignes ("wget...").

Ouvrez une fenêtre de terminal et connectez vous à votre machine virtuelle: **Sur votre machine locale:**

```
ssh ubuntu@XXX.XXX.XXX.XX # remplacer les X par l'adresse IP de votre VM
```

Sur votre machine virtuelle:

```
cd data/mydatalocal
wget -nc https://ftp.sra.ebi.ac.uk/vol1/fastq/SRR774/009/SRR7748059/SRR7748059_1.fastq.gz
wget -nc https://ftp.sra.ebi.ac.uk/vol1/fastq/SRR774/009/SRR7748059/SRR7748059_2.fastq.gz
```

S'il y a un problème avec cette fonction, vous pouvez telecharger les fichiers sur votre machine locale, puis les copier sur la machine virtuelle: **Sur votre machine locale:**

```
cd #chemin du dossier contenant les fichiers fastq.gz
scp *.gz ubuntu@134.158.XXX.XX:~/data/mydatalocal
```

pour décompresser les fichiers: **Sur votre machine virtuelle:**

```
gzip -d *.fastq.gz
```

Contrôle de la qualité des données

données

Nous allons contrôler la qualité du jeu de données de *S. aureus* que vous venez de télécharger.

Nous allons tout d'abord jeter un oeil au format des lectures (deux alternatives, mais ne pas ouvrir le fichier en entier, avec la commande nano par exemple, car ce sont de gros fichiers).

```
less SRR7748059_1.fastq
head SRR7748059_1.fastq
```

Analyse de la qualité des reads

Le programme communément utilisé est FastQC https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Nous allons l'installer puis afficher les options proposées.

```
sudo apt install fastqc
fastqc --h
```

Que fait l'option *-t* ?

Faites l'analyse qualité en utilisant la ligne de commande suivante

```
mkdir fastQC  # créer un dossier pour les fichiers de sortie
fastqc SRR7748059_1.fastq SRR7748059_2.fastq -o fastQC --noextract -t 2
```

Les analyses sont résumées dans les sorties (fichiers .html) produites par FastQC. Récupérez ces fichiers de sortie sur votre machine locale. Pour cela, ouvrez une autre fenetre de terminal et tapez ces commandes:

```
#tapez le chemin ou voulez coller vos resultats
mkdir -p formation_NGS/jour2/data
cd formation_NGS/jour2/data
scp ubuntu@134.158.XXX.XX:~/data/mydatalocal/fastQC/SRR7748059_*_fastqc.html .
```

Ouvrez le(s) fichier(s) html en double-cliquant dessus.

Que pouvez-vous dire de la qualité de ce jeu de données?

On remarque:

- les tailles des lectures font ~150pb pour HiSeq4.
- per base quality: la qualité est plus faible en fin de lectures.
- per base sequence content: Il y a un probleme de biais en debut de reads

Globalement, les reads plutot de bonne qualité, mais besoin de trimmer les lectures aux extrêmités.

Filtrage et nettoyage des données

Plusieurs outils sont disponibles. Nous allons utiliser FastP https://github.com/OpenGene/fastp#readme. Nous allons tout d'abord installer l'outil à l'aide de Bioconda sur la machine virtuelle.

Sur votre machine virtuelle:

```
cd /home/ubuntu/
conda init
conda install -c bioconda fastp
```

Répondre "yes" pour finaliser l'installation.

Puis lancer le programme ainsi afin de voir les options proposées :

```
fastp --h
```

Trimmer les reads du jeu de données. Le signe "\\" signifie que la ligne de commande se poursuit à la ligne suivante (dans votre console, ne faites pas de retour à la ligne).

```
cd ~/data/mydatalocal
fastp -i SRR7748059_1.fastq.gz -I SRR7748059_2.fastq.gz \\
-o SRR7748059_1trimmed.fastq.gz -O SRR7748059_2trimmed.fastq.gz \\
-l 80 -c
conda deactivate
```

Quelles options a-t-on choisies ? Analysez la qualité des reads après l'étape de nettoyage à l'aide de FastQC. Qu'en dites vous ?