Examen final Modules 4 et 5

DUBii 2021

Nicolas Dechamp

14 April, 2021

Contents

Consignes

Complétez ce document en remplissant les chunks vides pour écrire le code qui vous a permis de répondre à la question. Les réponses attendant un résultat chiffré ou une explication devront être insérés entre le balises html `code`. Par exemple pour répondre à la question suivante :

`La bioinfo c'est : <code>MERVEILLEUX</code>.`

N'hésitez pas à commenter votre code, enrichier le rapport en y insérant des résultats ou des graphiques/images pour expliquer votre démarche. N'oubliez pas les **bonnes pratiques** pour une recherche **reproductible** ! Nous souhaitons à minima que l'analyse soit reproductible sur le cluster de l'IFB.

Introduction

Vous allez travailler sur des données de reséquençage d'un génome bactérien : *Bacillus subtilis*. Les données sont issues de cet article :

- Complete Genome Sequences of 13 Bacillus subtilis Soil Isolates for Studying Secondary Metabolite Diversity

Analyses

Organisation de votre espace de travail

Création de l'architecture du dossier

```
mkdir -p data
mkdir -p src
mkdir -p results
mkdir -p results/FASTQC
mkdir -p results/CLEANING
mkdir -p results/QC
```

Téléchargement des données brutes

Récupérez les fichiers FASTQ issus du run **SRR10390685** grâce à l'outil sra-tools [1]

Réservation de ressources de calcul pour l'analyse interactive via salloc Récupération du module SRA-TOOLS pour utiliser fasterq-dump afin de recuperer la séquence Lancement de la commande sur le cluster avec srun

```
cd data

salloc --cpus-per-task=10 --mem=1G
module load sra-tools
fasterq-dump -h
srun --cpus-per-task=6 fasterq-dump --split-files -p SRR10390685 --outdir results
```

Combien de reads sont présents dans les fichiers R1 et R2 ?

```
echo $(cat SRR10390685_1.fastq | wc -l)/4|bc
```

Les fichiers FASTQ contiennent chacuns 7066055 reads.

Téléchargez le génome de référence de la souche ASM904v1 de *Bacillus subtilis* disponible à cette adresse

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/045/GCF_000009045.1_ASM904v1/GCF_000009045.1_
```

Quelle est la taille de ce génome ?

1ère ligne

```
zcat GCF_000009045.1_ASM904v1_genomic.fna.gz | awk ' { if (NR>1) { print$0} } ' | wc -l
```

La taille de ce génome est de 52696 paires de bases.

Téléchargez l'annotation de la souche ASM904v1 de *Bacillus subtilis* disponible à cette adresse

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/045/GCF_000009045.1_ASM904v1/GCF_000009045.1_
```

Combien de gènes sont connus pour ce génome ? 9ème colonne donne les informations sur les gènes. On sépare avec les délimiteurs ; pour ne récupérer que ID=gene

```
gunzip GCF_000009045.1_ASM904v1_genomic.gff.gz
cut -f 9 GCF_000009045.1_ASM904v1_genomic.gff | cut -d ";" -f 1 | grep "ID=gene" | sort -u | wc -l

# autre possibilité sans dézipper :
#zcat GCF_000009045.1_ASM904v1_genomic.gff.gz |cut -f 9  | cut -d ";" -f 1 | grep "ID=gene" | sort -u |
```

4536 gènes sont recensés dans le fichier d'annotation.

Contrôle qualité

Lancez l'outil fastqc [2] dédié à l'analyse de la qualité des bases issues d'un séquençage haut-débit

```
cp *fastq ../results/FASTQC
cd FASTQC

module load fastqc

srun --cpus-per-task 8 fastqc SRR10390685_1.fastq -o QC/ -t 8
srun --cpus-per-task 8 fastqc SRR10390685_2.fastq -o QC/ -t 8
```

La qualité des bases vous paraît-elle satisfaisante ? Pourquoi ?

- ☒ Oui
- ☐ Non

car la qualité de séquence par base reste au dessus de 30 même si elle décroit légèrement à partir de 100pb.
La quasi totalité a une longueur de plus de 144pb avec plus de 2M de reads
comme le montre

Lien SRR8082143_1_fastqc.html Lien SRR8082143_2_fastqc.html

Est-ce que les reads déposés ont subi une étape de nettoyage avant d'être déposés ? Pourquoi ?

- ☐ Oui
- ☒ Non

car la taille de la séquence est toujours de 7066055 et les séquences ont une taille comprises entre 35 et 151 pb pour un sens et 130-151pb pour l'autre

Quelle est la profondeur de séquençage (calculée par rapport à la taille du génome de référence) ?

```
echo "la Profondeur de séquençage est de : $((7066055/52696))"
```

La profondeur de séquençage est de : 134 X.

Nettoyage des reads

Vous voulez maintenant nettoyer un peu vos lectures. Choisissez les paramètres de fastp [3] qui vous semblent adéquats et justifiez-les.

```
module load fastp

srun --cpus-per-task 8 fastp --in1 SRR10390685_1.fastq --in2 SRR10390685_2.fastq --out1 ../CLEANING/SRR
eaned_filtered.fastq --out2 ../CLEANING/SRR10390685_2.cleaned_filtered.fastq --html CLEANING/fastp.html

echo $(cat ../CLEANING/SRR10390685_1cleaned_filtered.fastq | wc -l)/4|bc
```

Les paramètres suivants ont été choisis :

Parametre	Valeur	Explication
–cut_mean_quality	30	quality moyenne >= 30
–cut_window_size	8	sur fenêtre glissante de 8
–cut_tail		move a sliding window from tail (3') to front
–length_required	100	longueur de seq >= 100 pb

Ces paramètres ont permis de conserver 6777048 reads pairés, soit une perte de moins de 5% des reads bruts.
Lien sortie_fastp

Alignement des reads sur le génome de référence

Maintenant, vous allez aligner ces reads nettoyés sur le génome de référence à l'aide de bwa [4] et samtools [5].

```
cd ../data
gunzip *.gz
mv*fna * gff ../results

module load bwa
srun bwa index GCF_000009045.1_ASM904v1_genomic.fna

srun --cpus-per-task=8 bwa mem GCF_000009045.1_ASM904v1_genomic.fna CLEANING/SRR10390685_1.cleaned_filt
```

Combien de reads ne sont pas mappés ?

```
srun --cpus-per-task=8 samtools view --threads 8 SRR10390685_on_GCF_000009045.1.sam -b > SRR10390685_on

srun samtools sort SRR10390685_on_GCF_000009045.1.bam -o SRR10390685_on_genomic.sort.bam
srun samtools index SRR10390685_on_genomic.sort.bam

srun samtools idxstats SRR10390685_on_genomic.sort.bam > SRR10390685_on_genomic.sort.bam.idxstats
srun samtools flagstat SRR10390685_on_genomic.sort.bam > SRR10390685_on_genomic.sort.bam.flagstat
```

744540 reads ne sont pas mappés.

Croisement de données

Calculez le nombre de reads qui chevauchent avec au moins 50% de leur longueur le gène *trmNF* grâce à l'outil bedtools [6]:

```
module load bedtools

grep trmNF GCF_000009045.1_ASM904v1_genomic.gff | awk '$3=="gene"' > trmNF_gene.gff3

srun bedtools intersect -a SRR10390685_on_GCF_000009045.1.sort.bam -b trmNF_gene.gff3 -f 0.50 -r > SRR1
srun samtools index SRR10390685_on_trmNF.bam
srun samtools idxstats SRR10390685_on_trmNF.bam > SRR10390685_on_trmNF.bam.idxstats
srun samtools flagstat SRR10390685_on_trmNF.bam > SRR10390685_on_trmNF.bam.flagstat
```

2801 reads chevauchent le gène d'intérêt.

Visualisation

Utilisez IGV [7] sous sa version en ligne pour visualiser les alignements sur le gène. Faites une capture d'écran du gène entier.

References

1. toolkit NS. NCBI SRA toolkit. NCBI, GitHub repository. 2019.

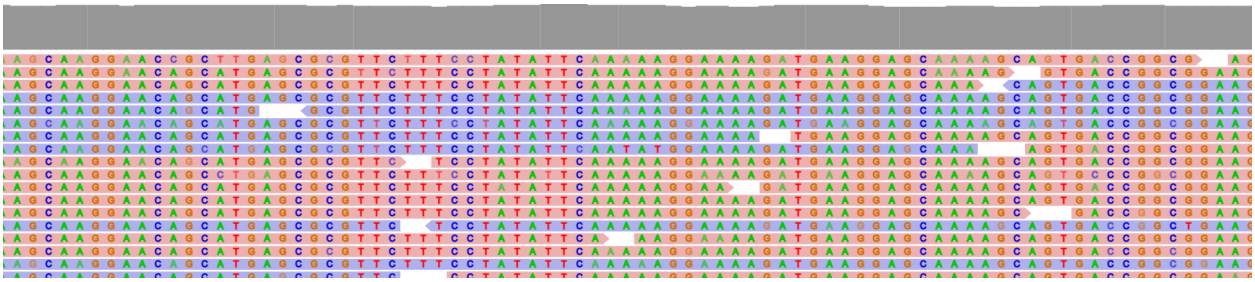

Figure 1: Capture d'écran du gène entier

2. Andrews S. FastQC a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/pro http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

3. Zhou Y, Chen Y, Chen S, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. doi:10.1093/bioinformatics/bty560.

4. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

6. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

7. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013;14:178–92.