Examen final Modules 4 et 5

DUBii 2021

Marika Kapsimali

09 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

```
ssh -XY mkapsimali@core.cluster.france-bioinformatique.fr
# go to dir projects/dubii2021/mkapsimali and create dir projet_M45
cd ../../
cd projects/dubii2021/mkapsimali
mkdir projet_M45
cd projet_M45
mkdir QC
mkdir FASTQ
mkdir Cleaning
mkdir Mapping
ls
```

Téléchargement des données brutes

Récupérez les fichiers FASTQ issus du run **SRR10390685** grâce à l'outil sra-tools @sratoolkit

```
#recover FASTQ files
module load sra-tools
srun --cpus-per-task=6 fasterq-dump --split-files -p SRR10390685 --outdir FASTQ
#verify files are in dir FASTQ
cd FASTQ
ls
#zip files fastq
srun gzip *.fastq
#visualize format with first 8 lines
zcat SRR10390685_1.fastq.gz | head -8
zcat SRR10390685_2.fastq.gz | head -8
```

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

```
#calculate number of reads
zcat SRR10390685_1.fastq.gz | echo $((`wc -l`/4))
zcat SRR10390685_2.fastq.gz | echo $((`wc -l`/4))
```

The FASTQ files contain 7066055 reads.

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

```
#recover ref. genome in dir projet_M45
cd ../
srun wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/045/GCF_000009045.1_ASM904v1
/GCF_000009045.1_ASM904v1_genomic.fna.gz
# verify ref. genome file
ls
```

Quelle est la taille de ce génome ?

```
#unzip ref. genome file to calculate size
gunzip GCF_000009045.1_ASM904v1_genomic.fna
#visualize start and end to calculate nucleotides per line in the file
head -n 3 GCF_000009045.1_ASM904v1_genomic.fna
tail -n 3 GCF_000009045.1_ASM904v1_genomic.fna
#80 nucleotides per line except last line 6 nucleotides. First line:name of the sequence
#Total number of lines:
```

```
wc -l GCF_000009045.1_ASM904v1_genomic.fna
#52697
#calculate number of nucleotides:
echo "$((52695 * 80 + 6))"
```

The genome size is 4215606 base pairs.

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

```
srun wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/045/GCF_000009045.1_ASM904v1
/GCF_000009045.1_ASM904v1_genomic.gff.gz
```

Combien de gènes sont connus pour ce génome ?

```
#unzip
gunzip GCF_000009045.1_ASM904v1_genomic.gff
head GCF_000009045.1_ASM904v1_genomic.gff
cut -f 3 GCF_000009045.1_ASM904v1_genomic.gff | grep "gene" |wc -l
cut -f 3 GCF_000009045.1_ASM904v1_genomic.gff | grep "pseudogene" |wc -l
```

4536 genes, of which 88 pseudogenes are found in the annotation file.

Contrôle qualité

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

```
module load fastqc
fastqc --version
#FastQC v0.11.9
srun --cpus-per-task 8 fastqc FASTQ/SRR10390685_1.fastq.gz -o QC/ -t 8
srun --cpus-per-task 8 fastqc FASTQ/SRR10390685_2.fastq.gz -o QC/ -t 8
```

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

- [x] Oui

because the boxplots and mean values are mostly above score 30 shown in the per base quality graph However,there are overepresented sequences (N or G) above 0.1% shown in over-represented sequence table and low adapter contamination shown in adapter content graph of SRR10390685_2

Lien vers le [rapport MulitQC] (https://github.com/mkapsimali/DuBii2021/blob/master/SRR10390685_1_fastqc.html) (https://github.com/mkapsimali/DuBii2021/blob/master/SRR10390685_2_fastqc.html)

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

- [x] Oui

Partially. Although N content accross all bases is zero, there are overepresented sequences N for SRR10390685_1 and adapter content is not always zero for SRR10390685_2

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

The number of reads is 7066055. The number of bases is approximately 150 based on the 2 QC sequencing reports. The reference genome size is 4215606bp

```
echo "$(((7066055 * 150 * 2)/4215606))"
```

La profondeur de séquençage est de : 502 X.

Nettoyage des reads

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

```
module load fastp
fastp --version
#fastp 0.20.0
srun --cpus-per-task 8 fastp --in1 FASTQ/SRR10390685_1.fastq.gz --in2 FASTQ/SRR10390685_2.fastq.gz
--out1 Cleaning/SRR10390685_1.cleaned_filtered.fastq.gz
--out2 Cleaning/SRR10390685_2.cleaned_filtered.fastq.gz
--html Cleaning/fastp.html --thread 8 --cut_mean_quality 30 --cut_window_size 8
--length_required 100 --cut_tail --json Cleaning/fastp.json
#calculate loss of reads after filtering
echo "$(((100*(7066055-6777048))/7066055))"
```

The following parameters are chosen :

Parameter | Value | Explanation |

1)Cut_mean_quality 30 : to keep only very good quality of bases. 2)Length_required 100 : reads shorter than 100 are discarded (smaller size too short for good quality mapping). 3)Cut_tail, it moves a sliding window from tail (3') to front, drops the bases in the window if its mean quality < threshold, stops otherwise. 4)Cut_window_size 8: the window size option for cut_tail (sequencing of this extremity can be of bad quality)

These parameters allow keeping 6777048 paired reads, with a loss of 4% of raw reads.

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 @bwa et samtools @samtools.

```
#move ref.genome file in Mapping dir for simplicity
mv GCF_000009045.1_ASM904v1_genomic.fna Mapping/GCF_000009045.1_ASM904v1_genomic.fna
cd Mapping
module load bwa
#Index FASTA file with bwa index
srun bwa index GCF_000009045.1_ASM904v1_genomic.fna
#Map reads with bwa mem
srun --cpus-per-task=32 bwa mem GCF_000009045.1_ASM904v1_genomic.fna
../Cleaning/SRR10390685_1.cleaned_filtered.fastq.gz
../Cleaning/SRR10390685_2.cleaned_filtered.fastq.gz -t 32
> SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sam

module load samtools
samtools --version
#samtools 1.10
#convert SAM file to BAM file with samtools view
srun --cpus-per-task=8 samtools view --threads 8 SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sam
-b > SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.bam
#Sort the BAM file with samtools sort
srun samtools sort SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.bam -o
SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam
#Index the BAM file with samtools index
srun samtools index SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam
```

Combien de reads ne sont pas mappés ?

```
srun samtools flagstat SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam
>SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam.flagstat
#view stats
cat SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam.flagstat
#calculate
echo "$((13571369-12826829))"
```

744540 reads are not mapped.

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 @bedtools:

```
#go to Projet_45 dir where there is GCF_000009045.1_ASM904v1_genomic.gff
cd ../
#find line with trmNF gene
grep trmNF GCF_000009045.1_ASM904v1_genomic.gff
#Choose line with trmNF where 3rd column is 'gene' and write to output file.
grep trmNF GCF_000009045.1_ASM904v1_genomic.gff |awk '$3=="gene"' > trmNF_gene.gff
module load bedtools

#Get genomic sequence of the gene with bedtools getfasta
srun bedtools getfasta -fi Mapping/GCF_000009045.1_ASM904v1_genomic.fna -bed trmNF_gene.gff
> trmNF_gene.fasta
#to verify presence of trmNF_gene.fasta
ls
#to have a look at the sequence
head trmNF_gene.fasta
#Calculate the number of reads of which at least 50% overlaps with the gene trmNF with
#bedtools intersect and write to output file
srun bedtools intersect -f 0.50 -b trmNF_gene.gff -a
Mapping/SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam
> result_intersection.bam
#sort the output file and see statistics
srun samtools sort result_intersection.bam -o result_intersection.sort.bam
srun samtools flagstat result_intersection.sort.bam > result_intersection.sort.bam.flagstat
cat result_intersection.sort.bam.flagstat

#For visualisation necessary:
samtools index result_intersection.sort.bam
```

2801 reads overlap with the gene of interest.

Visualisation :

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

See picture for trmNF result_intersection https://github.com/mkapsimali/DuBii2021/blob/master/result_intersection.png

It was obtained by uploading on IGV the genome (fna and fnai),

SRR10390685_on_GCF_000009045.1_ASM904v1_genomic.sort.bam and bai,

and the (trmNF) result_intersection.sort.bam and bai.

See also tree https://github.com/mkapsimali/DuBii2021/blob/master/tree.png

References

1. toolkit NS. NCBI sra toolkit. NCBI, GitHub repository. 2019.
2. Andrews S. FastQC a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 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.