Evaluation finale Du-Bii modules 4 et 5

Control qualité et alignement au génome de référence des données de reséquençage génomique de bactéries du sol - étude d'un cas

Viviana MARIN-ESTEBAN

2020-09-01

Contents

Objectif

Faire une analyse simple, de données de reséquençage d'un génome bactérien. Le séquençage a été réalisé en paired-end et les deux fichiers des données seront récupérés en format fastq. Ces données sont issus du travail publié dans l'article : "Complete Genome Sequences of 13 Bacillus subtilis Soil Isolates for Studying Secondary Metabolite Diversity" (doi:10.1128/MRA.01406-19)

Ce rapport devra être mis à notre disposition dans un dépôt public GitHub. Les analyses devront pouvoir être rejouées sur le cluster de l'IFB.

Etapes

Création des dossiers de travail

Sur le serveur de l'IFB, créer les dossiers pour les différentes étapes des analyses à faire :

```
ssh [USER_NAME]@core.cluster.france-bioinformatique.fr

mkdir M4_Eval

mkdir -p ~/M4_Eval/FASTQ
mkdir -p ~/M4_Eval/CLEANING
mkdir -p ~/M4_Eval/MAPPING
mkdir -p ~/M4_Eval/QC
mkdir -p ~/M4_Eval/INPUT_genome_files
mkdir -p ~/M4_Eval/gene_trmNF
```

Récupération des données

- Récupération des données brutes de séquençage (format fastq non compressé), à partir du site European Nucleotide Archive, avec l'outil sra-toolkit

```
module load  sra-tools
srun fasterq-dump -S -p SRR10390685 --outdir ~/M4_Eval/FASTQ --threads 1
```

- Compresser les fichiers FASTQ, se positionner dans le dossier FASTQ

```
cd ~/M4_Eval/FASTQ
gzip *.fastq
```

```
ls -ltrh ~/M4_Eval/FASTQ/

total 1.3G
-rw-rw-r-- 1 vmarinesteban vmarinesteban 627M Aug 28 18:21 SRR10390685_2.fastq.gz
-rw-rw-r-- 1 vmarinesteban vmarinesteban 617M Aug 28 18:21 SRR10390685_1.fastq.gz
```

Control qualité des fichiers FASTQ

Cette analyse permet un aperçu des donnés non traitées pour identifier des problèmes de qualité et de biais de séquences qui pourraient impacter ou fausser les analyses en aval.

```
cd ~
module load fastqc
srun --cpus-per-task 8 fastqc ~/M4_Eval/FASTQ/SRR10390685_1.fastq.gz -o ~/M4_Eval/QC/ -t 8
srun --cpus-per-task 8 fastqc ~/M4_Eval/FASTQ/SRR10390685_2.fastq.gz -o ~/M4_Eval/QC/ -t 8
```

```
ls -ltrh ~/M4_Eval/QC/

total 1.7M
-rw-rw-r-- 1 vmarinesteban vmarinesteban 291K Aug 28 19:06 SRR10390685_1_fastqc.zip
-rw-rw-r-- 1 vmarinesteban vmarinesteban 548K Aug 28 19:06 SRR10390685_1_fastqc.html
-rw-rw-r-- 1 vmarinesteban vmarinesteban 312K Aug 28 19:08 SRR10390685_2_fastqc.zip
-rw-rw-r-- 1 vmarinesteban vmarinesteban 561K Aug 28 19:08 SRR10390685_2_fastqc.html
```

Inspection des résultats :

Voir les rapports HTML générés par l'outil fastq : fichiers SRR10390685_1_fastqc.html et SRR10390685_2_fastqc.html

Pour SRR10390685_1_fastqc on trouve :

- 7066055 reads, de taille entre 35 et 151 nucléotides, avec 43 % de contenu GC.
- Per base quality: superior ou égale à 30 dans tout le long de la séquence
- Per sequence quality scores : ne montre pas une distribution anormale des la qualité des reads. La courbe commence à se décoller du sol à une qualité de 24, et plus de 90% des séquences ont une qualité > 28.
- Per base content : MODULE AVEC FAIL car une distribution anormale dans la position 2 et 3 des reads est signalé. Dans ces positions la distribution de T est de ~45 % et ~35 % dans l'ensemble des reads, respectivement. Cependant, ce problème est bien localisé dans un but des séquences. Ce problème sera possible de fixer dans l'étape de nettoyage.
- Per séquence GC content : le contenu GC sur toute la longueur de reads n'est pas anormal, par rapport au modèle calculé.
- Per base N content : l'assignation globale de nucléotides dans les reads est fiable.
- Sequence Length Distribution : MODULE AVEC AVERTISSEMENT car distribution des longueurs des reads pas suffisamment homogène. Cependant, la plus part des reads ont autour de 148 à 152 nucléotides. Ce problème sera fixé dans l'étape de nettoyage.
- Sequence Duplication Levels : MODULE AVEC AVERTISSEMENT, les séquences non uniques représentent entre 20% et 50% du total. La perte globale attendue de séquences sera de 42.7 % si la bibliothèque est dédupliquée.
- Overrepresented sequences : MODULE AVEC AVERTISSEMENT. Au moins une séquence d'au moins 20 bp représente 0,1% à 1% des séquences. Dans ce cas c'est une séquence NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN qui représente 0.12 % des séquences. Aucune source possible de contaminant est proposée.
- Adapter Content : le pourcentage cumulé des séquences d'adaptateurs est 0%.

Pour SRR10390685_2_fastqc on trouve des résultats similaires à ceux de SRR10390685_1_fastqc :

- 7066055 reads, de taille entre 130 et 151 nucléotides, avec 43 % de contenu GC.
- Per base quality: superior ou égale à 30 dans tout le long de la séquence
- Per sequence quality scores : ne montre pas une distribution anormale des la qualité des reads. La courbe commence à se décoller du sol à une qualité de 22, et plus de 90% des séquences ont une qualité > 28.
- Per base content : MODULE AVEC FAIL car une distribution anormale dans la position 2 et 3 des reads est signalé. Dans ces positions la distribution de T est de ~45 % et ~35 % dans l'ensemble des reads, respectivement. Cependant, ce problème est bien localisé dans un but des séquences. Ce problème sera possible de fixer dans l'étape de nettoyage.
- Per séquence GC content : le contenu GC sur toute la longueur de reads n'est pas anormal, par rapport au modèle calculé.
- Per base N content : l'assignation globale de nucléotides dans les reads est fiable.
- Sequence Length Distribution : MODULE AVEC AVERTISSEMENT car distribution des longueurs des reads pas suffissament homogène. Cependant, la plus part des reads ont autour de 150 à 152 nucléotides, avec un pic à 151 nucléotides. Ce problème sera fixé dans l'étape de nettoyage.
- Sequence Duplication Levels : MODULE AVEC AVERTISSEMENT, les séquences non uniques représentent entre 20% et 50% du total. La perte globale attendue de séquences sera de 39.01% si la bibliothèque est dédupliquée.
- Overrepresented sequences : MODULE AVEC AVERTISSEMENT. Au moins une sequence d'au moins 20 bp représente 0,1% à 1% des sequences. Dans ce cas c'est une sequence GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG qui représente 0.11 % des séquences. Aucune source possible de contaminant est proposée.
- Adapter Content : le pourcentage cumulé des séquences Illumina Universal Adapter est <5% dans les reads de >110 nucléotides. Ce problème sera certainement fixé dans l'étape de nettoyage.

Réponses aux questions posées :

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

- *En vue des résultats de l'analyse de qualité avec FASTQC (fichiers SRR10390685_1_fastqc.html et SRR10390685_2_fastqc.html), les séquences des fichiers fastq sont globalement de bonne qualité. Cependant parmi les 9 grands modules évalués, nous avons vu 3 signaux d'avertissement et un signal de FAIL.*
- *L'avertissement que je trouve le plus critique est le niveau de duplications de séquences (séquences identiques), qui est aux alentours de 40%, ce qui ramènerait à une réduction de la taille des librairies de ~7 millions de reads ver ~4 millions. Cette profondeur reste encore bonne, en considérant que l'objectif de l'étude semble être d'avoir de bonnes couverture génomique et qualité de séquençage pour caractériser des isolats différents d'une espèce bactérienne dans le sol.*
- *Les autres problèmes soulevés par l'analyse FASTQC me semblent gérables dans l'étape de nettoyage. Cependant, plus tard nous verrons que le nettoyage que j'ai fait n'a pas été très astringent et seulement peu de séquences ont été enlevées et les élagages des extrémités ont fait le compromis de ramené les autres problèmes à des niveaux acceptables sans les éliminer totalement, pour garder la meilleur taille et le nombre plus fort de reads (voire le fichier multiqc_report.html)*

- *La stratégie de séquençage est de paired-ends, ce qui permettra plus de fiabilité au moment d'aligner les séquences sur le génome de référence. Chaque séquençage à été fait avec une profondeur globale de plus de 7 millions de reads.*

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

- *La taille du génome de référence est de 4215606 paires de bases. Ainsi, si on fait une approximation grossière (avant nettoyage) de 7 millions de read par séquençage, avec une taille moyenne de 140 nucléotides par read, il y aurait ~980 millions de bases séquencées.*
- *Par rapport aux ~4,2 million de bases du génome, chaque séquençage aurait une profondeur de* **245 équivalents génome**, *ce qui équivaut à une bonne profondeur de séquençage, avec, en théorie, 245 chances de que chaque nucléotide soit séquencé dans des reads différents, chevauchants. Le séquençage paired-end augmente la profondeur de séquençage et la certitude / qualité des alignements.*

Nettoyage des reads avec FASTP

Cette étape permet de corriger des problèmes de qualité et de biais de séquences déjà identifiés dans l'étape précédente. Par exemple : enlever des séquences d'adaptateurs, des séquences surreprésentés, de couper les extrémités des reads avec contenu distribution anormal de nucléotides.

```
module load fastp
cd ~/M4_Eval/

srun --cpus-per-task 8 fastp \
 --in1 FASTQ/SRR10390685_1.fastq.gz \
 --in2 FASTQ/SRR10390685_2.fastq.gz \
 -l 100 \
 --out1 CLEANING/SRR10390685_1.cleaned_filtered.fastq.gz \
 --out2 CLEANING/SRR10390685_2.cleaned_filtered.fastq.gz \
 --unpaired1 CLEANING/SRR10390685_singles.fastq.gz \
 --unpaired2 CLEANING/SRR10390685_singles.fastq.gz \
 -w 1 \
 -j CLEANING/fastp.json \
 -h CLEANING/fastp.html \
 -t 8
```

```
ls -ltrh ~/M4_Eval/CLEANING/

total 1.2G
-rw-rw-r-- 1 vmarinesteban vmarinesteban 132K Aug 28 22:49 fastp.json
-rw-rw-r-- 1 vmarinesteban vmarinesteban 475K Aug 28 22:49 fastp.html
-rw-rw-r-- 1 vmarinesteban vmarinesteban  15M Aug 28 22:49 SRR10390685_singles.fastq.gz
-rw-rw-r-- 1 vmarinesteban vmarinesteban 609M Aug 28 22:49 SRR10390685_2.cleaned_filtered.fastq.gz
-rw-rw-r-- 1 vmarinesteban vmarinesteban 604M Aug 28 22:49 SRR10390685_1.cleaned_filtered.fastq.gz
```

Inspection des résultats du nettoyage avec fastp

Un rapport résumé est affiché à l'écran à la fin de la procédure et un autre rapport plus complet est généré dans un fichier .json et un autre .html. Au total, 97.015110% des reads ont été retenus. Le nettoyage a été équilibré dans les deux fichiers. Au total, en partant de 14132110 reads, 421828 ont été exclues, c'est à dire moins de 3% des reads (2.985%). Les reads ont été élaguées, passant d'une longueur moyenne avant nettoyage de 149bp (fichier _1) et 150bp (fichier _2) à une longueur moyenne de 141bp dans les deux cas.

- Ceci est le rapport affiché à l'écran à la fin du nettoyage :

Read1 before filtering: total reads: 7066055 total bases: 1056334498 Q20 bases: 989425011(93.6659%) Q30 bases: 950850058(90.0141%)

Read2 before filtering: total reads: 7066055 total bases: 1062807718 Q20 bases: 975887037(91.8216%) Q30 bases: 933776181(87.8594%)

Read1 after filtering: total reads: 6855141 total bases: 972593291 Q20 bases: 918080212(94.3951%) Q30 bases: 885808239(91.0769%)

Read2 aftering filtering: total reads: 6855141 total bases: 972626525 Q20 bases: 906379990(93.1889%) Q30 bases: 872179928(89.6726%)

Filtering result: reads passed filter: 13710282 reads failed due to low quality: 356126 reads failed due to too many N: 4092 reads failed due to too short: 61610 reads with adapter trimmed: 352768 bases trimmed due to adapters: 5499206

Duplication rate: 1.17555%

Insert size peak (evaluated by paired-end reads): 250

JSON report: CLEANING/fastp.json HTML report: CLEANING/fastp.html

- Et celui-ci est le résumé initial dans le rapport fastp.html généré suite au processus de nettoyage :

fastp version: 0.20.0 sequencing: paired end (151 cycles + 151 cycles) mean length before filtering: 149bp, 150bp mean length after filtering: 141bp, 141bp duplication rate: 1.175552% Insert size peak: 250

Before filtering total reads: 14.132110 M total bases: 2.119142 G Q20 bases: 1.965312 G (92.740923%) Q30 bases: 1.884626 G (88.933448%) GC content: 43.657172%

After filtering total reads: 13.710282 M total bases: 1.945220 G Q20 bases: 1.824460 G (93.791981%) Q30 bases: 1.757988 G (90.374782%) GC content: 43.539890%

Filtering result reads passed filters: 13.710282 M (97.015110%) reads with low quality: 356.126000 K (2.519978%) reads with too many N: 4.092000 K (0.028955%) reads too short: 61.610000 K (0.435958%)

Réponses aux questions posées :

1. Quel pourcentage de reads sont filtrés et pourquoi ?

Una réponse rapide à cette question est donné dans le dernier paragraphe du résumé initial du fichier fastp.html (voir dernier paragraphe avant cette section) :

Le pourcentage de reads filtrées (exclues) a été de ~2.985%. En détail, 2.52% des reads ont été exclues par qualité basse, 0.029% de reads par un pourcentage fort de nucléotides incertains (N) et 0.436% en raison d'une langueur trop petite.

Un rapport MultiqQC

Regroupe les informations / rapports des procédures FastQC, Fastp (et d'autres Trimmomactic, Cutadapt, etc..).

Le rapport généré (multiqc_report.html) permet de voir avec le même format les analyses réalisés avec FASTQ pour les librairies avant et après le nettoyage.

```
cd ~/M4_Eval
module load multiqc
# export LANG=en_US.UTF-8 #ligne à ajouter si problème avec multiqc du à UTF-8 locale
multiqc -d . -o CLEANING
```

```
ls -ltrh ~/M4_Eval/CLEANING/

total 1.3G
-rw-rw-r-- 1 vmarinesteban vmarinesteban  15M Aug 28 22:49 SRR10390685_singles.fastq.gz
-rw-rw-r-- 1 vmarinesteban vmarinesteban 609M Aug 28 22:49 SRR10390685_2.cleaned_filtered.fastq.gz
-rw-rw-r-- 1 vmarinesteban vmarinesteban 604M Aug 28 22:49 SRR10390685_1.cleaned_filtered.fastq.gz
-rw-rw-r-- 1 vmarinesteban vmarinesteban 132K Aug 28 22:49 fastp.json
-rw-rw-r-- 1 vmarinesteban vmarinesteban 475K Aug 28 22:49 fastp.html
-rw-rw-r-- 1 vmarinesteban vmarinesteban 1.2M Aug 28 23:57 multiqc_report.html
drwxrwxr-x 2 vmarinesteban vmarinesteban 1.1M Aug 28 23:57 multiqc_data
```

Inspection des résultats du rapport MultiqQC

Alignement au génome de référence / Mapping

Alignement des reads sur le génome de référence. Les alignements sont accompagnés d'une évaluation statistique qu'informe de la probabilité d'un alignement correct.

Télécharger les fichiers du génome de référence (formats fasta et gff)

```
cd ~/M4_Eval

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

Indexation du génome et alignement

Je vais commencer par dezipper le fichier fasta du génome (.fna), dont j'aurais besoin plus tard, avec les outils bedtools. Ensuite, il faut indexer le génome de référence pour pouvoir utiliser BWA. Ensuite, l'alignement est réalisé. Enfin, je réalise l'alignement des read du séquençage sur le génome de référence.

```
module load bwa
module load samtools

cd ~/M4_Eval/INPUT_genome_files/
gunzip GCF_000009045.1_ASM904v1_genomic.fna
srun bwa index GCF_000009045.1_ASM904v1_genomic.fna

cd ~/M4_Eval/
srun --cpus-per-task=32 bwa mem \
  INPUT_genome_files/GCF_000009045.1_ASM904v1_genomic.fna \
  CLEANING/SRR10390685_1.cleaned_filtered.fastq.gz \
  CLEANING/SRR10390685_2.cleaned_filtered.fastq.gz \
  -t 32 \
```

```
  | \
  samtools view -hbS - > MAPPING/SRR10390685.bam
```

```
ls -ltrh ~/M4_Eval/MAPPING/

total 1.4G
-rw-rw-r-- 1 vmarinesteban vmarinesteban 1.4G Aug 29 02:20 SRR10390685.bam
```

Tri et indexation du fichier d'alignements avec SAMTOOLS (.bam)

Etape non nécessaire sauf pour le résumé statistique obtenue avec l'outil flagstat de samtools.

```
module load samtools

# Sort BAM file
samtools sort MAPPING/SRR10390685.bam -o MAPPING/SRR10390685.sorted.bam

# Index sorted BAM file
samtools index MAPPING/SRR10390685.sorted.bam

# Get some statistics
samtools flagstat MAPPING/SRR10390685.sorted.bam
```

Le rapport affiché dans l'écran suite au calcul des statistiques (samtools flagstst) :

flagstat MAPPING/SRR10390685.sorted.bam 13726125 + 0 in total (QC-passed reads + QC-failed reads) 0 + 0 secondary 15843 + 0 supplementary 0 + 0 duplicates 12969706 + 0 mapped (94.49% : N/A) 13710282 + 0 paired in sequencing 6855141 + 0 read1 6855141 + 0 read2 12887940 + 0 properly paired (94.00% : N/A) 12911596 + 0 with itself and mate mapped 42267 + 0 singletons (0.31% : N/A) 0 + 0 with mate mapped to a different chr 0 + 0 with mate mapped to a different chr (mapQ>=5)

Réponses aux questions posées :

1. Quel est le % de reads pairés alignés ?

- *Comme indiqué par les informations issues des statistiques du fichier .bam, obtenues avec la fonction flagstat de samtools (voir paragraphe ci-dessous),*

Ensemble des fichiers dans le dossier MAPPING

```
ls -ltrh ~/M4_Eval/MAPPING/

[vmarinesteban@clust-slurm-client M4_Eval]$ ls -ltrh ~/M4_Eval/MAPPING/
total 2.2G
-rw-rw-r-- 1 vmarinesteban vmarinesteban 1.4G Aug 29 02:20 SRR10390685.bam
-rw-rw-r-- 1 vmarinesteban vmarinesteban 766M Aug 29 02:42 SRR10390685.sorted.bam
-rw-rw-r-- 1 vmarinesteban vmarinesteban  14K Aug 29 02:44 SRR10390685.sorted.bam.bai
```

Extraire dans un fichier BAM les reads chevauchant à au moins 50% le gène trmNF

Retrouver les lignes du fichier .gff contenant le nom du gène trmNF

```
cd ~/M4_Eval/INPUT_genome_files/
gunzip GCF_000009045.1_ASM904v1_genomic.gff.gz

cd ~/M4_Eval
grep trmNF INPUT_genome_files/GCF_000009045.1_ASM904v1_genomic.gff \
| awk '($3 == "gene")'\
> gene_trmNF/trmNF_gene.gff

cat gene_trmNF/trmNF_gene.gff
#NC_000964.3     RefSeq  gene    42917   43660   .       +       .       ID=gene-BSU_00340;Name=trmNF;g
```

Inspection du fichier .gff du gene trmNF

Je vérifie la ligne sélectionnée. Ceci permet de confirmer que le gène trmNF a été bien choisi et nous permet de savoir des informations telles que la taille du gène (744 bp, coordonnées de début et fin inclues : 42917 à 43660 sur le génome) et qu'il est codé par le brin + du génome.

Récupérer la séquence du gène trmNF avec bedtools

Pour obtenir cette séquence on croise le fichier de séquence du génome en format fasta (option -fi de getfasta) (fichier INPUT_genome_files/GCF_000009045.1_ASM904v1_genomic.fna) avec le fichier des annotations du gène trmNF (fichier gene_trmNF/trmNF_gene.gff). Ce dernier fichier (.gff) informe la position qui ont sur le génome les nucléotides du début et de la fin de la séquence du gène. Apporte donc la tranche de nucléotides à extraire de la séquence du génome (option -bed de getfasta).

```
module load bedtools

cd ~/M4_Eval

srun bedtools getfasta \
-fi INPUT_genome_files/GCF_000009045.1_ASM904v1_genomic.fna \
-bed gene_trmNF/trmNF_gene.gff \
> gene_trmNF/trmNF_gene.fasta

# Regarder le fichier
cat gene_trmNF/trmNF_gene.fasta

# longueur de la ligne la plus longue du fichier
wc -L gene_trmNF/trmNF_gene.fasta
#744 gene_trmNF/trmNF_gene.fasta
```

Inspection du fichier .gff du gène trmNF

Une inspection du fichier crée avec les commandes cat montre un fichier de deux lignes. Dans la première ligne du fichier fasta les informations de la séquence. Elle correspond aux coordonnées 42916-43660 du

fichier originel du génome (C_000964.3). Dans la seconde ligne on voit une séquence de nucléotides. Une inspection de cette seconde ligne avec la commande wc -L permet de vérifier que la longueur de la séquence est de 744nucléotides, comme attendu.

Explorer la couverture et profondeur de séquençage du gène trmNF

Récupérer les reads totales (+ et -) de la librairie SRR10390685 mappée (SRR10390685.bam) dont au moins 50% de la séquence (option -f 0.50) chevauche le gène trmNF

```
cd ~/M4_Eval

# Sélection des reads chevauchant le gene trmNF
srun bedtools intersect \
-a MAPPING/SRR10390685.bam -b gene_trmNF/trmNF_gene.gff -f 0.50 \
> gene_trmNF/SRR10390685_on_trmNF_50.bam

# Organiser les reads sélectionnées
samtools sort gene_trmNF/SRR10390685_on_trmNF_50.bam  -o gene_trmNF/SRR10390685_on_trmNF_50.sorted.bam

# Transformer le fichier .bam en .bed pour manipuler plus facilement les intervalles
srun bedtools bamtobed -i gene_trmNF/SRR10390685_on_trmNF_50.sorted.bam \
-split \
> gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed

# Comptage de reads sélectionnées
wc -l gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed
# 2848 gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed

# Nombre de reads sélectionnées dans la direction sens
grep + gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed | wc -l
#1402

# Nombre de reads sélectionnées dans la direction anti-sens
grep - gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed | wc -l
#1446


# Pour information, la même procédure sans filtrage du pourcentage de chevauchement du gène trmNF sélec
srun bedtools intersect \
-a MAPPING/SRR10390685.bam -b gene_trmNF/trmNF_gene.gff \
> gene_trmNF/SRR10390685_on_trmNF.bam

srun bedtools bamtobed -i gene_trmNF/SRR10390685_on_trmNF.bam \
-split \
> gene_trmNF/SRR10390685_on_trmNF.bed

wc -l gene_trmNF/SRR10390685_on_trmNF.bed
# 3418 gene_trmNF/SRR10390685_on_trmNF.bed
```

Couverture du gène trmNF par les reads sélectionnées, explorée avec bedtools coverage

L'outil coverage calcule à la fois la profondeur et l'étendue de la couverture des entités du fichier B sur les entités du fichier A. L'option -hist rapporte un histogramme de couverture pour chaque entité individuelle dans A ainsi que pour l'ensemble des entités dans A. Sortie (délimitée par des tabulations) après chaque fonction dans A: 1) profondeur 2) # bases en profondeur 3) taille de A 4)% de A en profondeur

```
cd ~/M4_Eval

# Histogramme de couvertures
srun bedtools coverage \
-a gene_trmNF/trmNF_gene.gff \
-b gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed \
-hist

# Couverture par position
srun bedtools coverage \
-a gene_trmNF/trmNF_gene.gff \
-b gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed \
-d

# Couverture moyenne
srun bedtools coverage \
-a gene_trmNF/trmNF_gene.gff \
-b gene_trmNF/SRR10390685_on_trmNF_50.sorted.bed \
-mean

#NC_000964.3     RefSeq  gene    42917   43660   .       +       .       ID=gene-BSU_00340;Name=trmNF;g
```

Inspection des résultats de la couverture de séquençage du gène trmNF

La sélectionné des reads qui ont été alignées sur le gène trmNF a été fait avec un filtre pour choisir les reads dont au moins 50% de la longueur du read chevauché avec le gène. Ainsi, 28548 reads on été sélectionnées. Si on élimine ce filtre, 3418 reads auraient été sélectionnées.

Parmi les 28548 reads sélectionnées, 1402 vient du séquençage en direction sens et 1446 du séquençage en direction anti-sens.

Les données de couverture générées par bedtools coverage -hist pour l'ensemble de nucléotides révèlent des profondeurs par nucléotide entre 262 et 581. Avec l'option -d on peut savoir la couverture de chaque nucléotide du gène d'intérêt. L'option -mean indique que le profondeur moyenne de séquençage pour chaque nucléotide du gène a été de 516.4314575.

Organisation du repretoire du projet

```
tree ~/M4_Eval/
```

```
[vmarinesteban@clust-slurm-client M4_Eval]$ tree
.
├── CLEANING
│   ├── fastp.html
│   ├── fastp.json
│   ├── multiqc_data
│   │   ├── multiqc_data.json
│   │   ├── multiqc_fastp.txt
│   │   ├── multiqc_fastqc.txt
│   │   ├── multiqc_general_stats.txt
│   │   ├── multiqc.log
│   │   └── multiqc_sources.txt
│   ├── multiqc_report.html
│   ├── SRR10390685_1.cleaned_filtered.fastq.gz
│   ├── SRR10390685_2.cleaned_filtered.fastq.gz
│   └── SRR10390685_singles.fastq.gz
├── FASTQ
│   ├── SRR10390685_1.fastq.gz
│   └── SRR10390685_2.fastq.gz
├── gene_trmNF
│   ├── SRR10390685_on_trmNF_50.bam
│   ├── SRR10390685_on_trmNF_50.sorted.bam
│   ├── SRR10390685_on_trmNF_50.sorted.bed
│   ├── SRR10390685_on_trmNF.bam
│   ├── SRR10390685_on_trmNF.bed
│   ├── trmNF_gene.cov.bedg
│   ├── trmNF_gene.fasta
│   └── trmNF_gene.gff
├── INPUT_genome_files
│   ├── GCF_000009045.1_ASM904v1_genomic.fna
│   ├── GCF_000009045.1_ASM904v1_genomic.fna.amb
│   ├── GCF_000009045.1_ASM904v1_genomic.fna.ann
│   ├── GCF_000009045.1_ASM904v1_genomic.fna.bwt
│   ├── GCF_000009045.1_ASM904v1_genomic.fna.fai
│   ├── GCF_000009045.1_ASM904v1_genomic.fna.pac
│   ├── GCF_000009045.1_ASM904v1_genomic.fna.sa
│   └── GCF_000009045.1_ASM904v1_genomic.gff
├── MAPPING
│   ├── SRR10390685.bam
│   ├── SRR10390685.sorted.bam
│   └── SRR10390685.sorted.bam.bai
└── QC
    ├── SRR10390685_1_fastqc.html
    ├── SRR10390685_1_fastqc.zip
    ├── SRR10390685_2_fastqc.html
    └── SRR10390685_2_fastqc.zip

7 directories, 37 files
```

Fichiers de résultats disponibles dans le dossier github

1. SRR10390685_1_fastqc.html
2. SRR10390685_2_fastqc.html

3. fastp.html
4. multiqc_report.html
5. trmNF_gene.fasta
6. SRR10390685_on_trmNF_50.sorted.bd

Conclusion et remerciements

Ce travail m'a permis de réviser les étapes pour le contrôle de qualité des fichiers fastq d'un séquençage, l'alignement sur un génome de référence et l'inspection de couverture et profondeur de séquençage d'un gène individuel. La qualité et profondeur de séquençage de cet exercice sont très bonnes (à mon avis) et n'ont pas posé de grands soucis. Merci pour cette formation du DU-Bii qui a été très dense, appliquée et intéressante :-).