Formation "Cycle NGS": Module 1 DNAseq

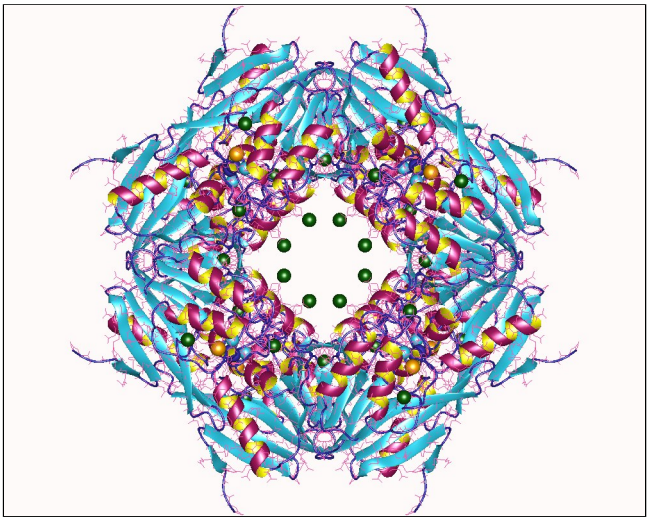
Mapping: étude d'un plasmide d'Escherichia coli

Contexte de l'étude

Dans ce TP, nous allons apprendre à mapper des reads contre une séquence de référence. Nous utiliserons l'instance Galaxy France usegalaxy.fr sur laquelle vous pouvez vous logger. Ensuite, pour le TP nous utiliserons ce lien, ressources propres à la formation.

Les jeux de données utilisés dans ce TP proviennent d'une étude sur la caractérisation d'une souche clinique d'Escherichia coli au Japon [Takayama et al. (2020)]. Quatre entrées de séquençage sont disponibles sur SRA: 1 entrée PacBio et 3 entrées Illumina MiSeq, une pour le chromosome et 3 autres qui correspondent aux 3 plasmides de la souche.
Les données Illumina sont des données paired-end (Forward-Reverse). **Attention**, ce ne sont pas les données brutes qui sont disponibles, elles ont été modifiées (trimming/Filtrage) par les auteurs avant le dépôt dans SRA.

La New Delhi métallo-β-lactamase (NDM) est une enzyme qui confère aux bactéries qui la synthétisent une résistance aux antibiotiques de la famille des carbapénèmes, habituellement réservés au traitement des infections multirésistantes. Elle est codée par un gène dénommé NDM-1, situé sur un plasmide.



Octamère de métallo-β-lactamase. Source: https://fr.wikipedia.org/

La NDM-5, qui diffère de la NDM-1 par la substitution de deux acides aminés, montre une résistance encore plus forte aux carbapénèmes. Dans cette étude, les auteurs ont caractérisé une souche d'Escherichia coli (isolate KY1497) produisant la NDM-5 qui a été isolée d'un patient présentant une infection urinaire au Japon.

Dans ce TP, pour réduire le temps d'exécution de certains programmes, nous allons travailler uniquement sur les données issues du plasmide1.
Note: pour les besoins du TP, les données disponibles sur SRA ont été modifiées.

Import des données

En plus des fichiers FASTQ contenant les reads, nous allons utiliser la séquence d'un plasmide d'E. coli contenant le gène blaNDM-5 comme séquence de référence.
Vous pouvez télécharger les données qui sont disponibles ici:

- reads_R1.fastq: les reads forward

- reads_R2.fastq: les reads reverse
- REF_plasmid.fasta: la séquence de référence au format fasta
- REF_plasmid.gb: la séquence de référence au format genbank

Dans le menu de gauche de Galaxy, tout en haut, cliquez sur le bouton *Upload Data*. Sélectionnez ensuite les 3 fichiers: reads_R1.fastq, reads_R2.fastq et REF_plasmide.fasta et cliquez sur *start*.

Mapping des reads avec Bowtie2

Nous allons commencer par réaliser un premier mapping des reads contre la séquence de référence en utilisant les paramètres par défaut de bowtie2 afin de se faire une première idée sur les données.

Lancez *bowtie2* avec ces paramètres:

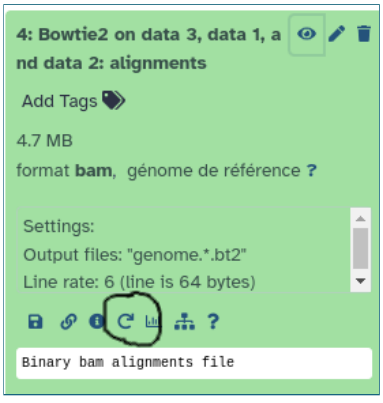
- "Is this single or paired library": `Paired-end`
- "FASTA/Q file #1 :" `reads_R1.fastq`
- "FASTA/Q file #2: " `reads_R2.fastq`
  Attention, l'ordre est important.
- "Will you select a reference genome from your history or use a built-in index?" `Use a genome from the histoty and build index`
- "Select reference genome *" `REF_plasmid.fasta` à partir de votre historique
- “Save the bowtie2 mapping statistics to the history”: `Yes`

Visualisez le **fichier contenant les statistiques de mapping**.
Par défaut, bowtie2 fait-il de l'alignement global ou de l'alignement local?
Combien de reads ont été mappés de façon concordante exactement une fois?
Combien de reads ont été mappés de façon concordante plus d'une fois?
Parmi les reads alignés de façon condordante 0 fois, combien sont alignés de façon discordante une fois?
Parmi les reads non-alignés ou alignés de façon discordante, quel est le pourcentage de reads non-alignés?
Quel est le taux global d'alignement?
Est-ce que les résultats de ce mapping vous semble corrects?
Le taux de mapping des reads est faible, quelles sont les causes qui pourraient expliquer un nombre si important de reads non-mappés?

Mapping des reads avec bowtie2 local

Nous allons maintenant réaliser un mapping des reads contre la même séquence de référence, mais en utilisant un alignement local.

Dans galaxy, il est possible de relancer un job en cliquant sur la flèche "Run Job Again" dans l'historique situé dans la partie droite.



Relancez le mapping avec les mêmes fichiers mais en choisissant dans la partie "Do you want to use presets?" `Fast local (--fast-local)` .

Visualisez le **fichier contenant les statistiques de mapping**.
Combien de reads ont été mappés de façon concordante exactement une fois?
Combien de reads ont été mappés de façon concordante plus d'une fois?
Parmi les reads non-alignés ou alignés de façon discordante, quel est le pourcentage de reads non-alignés?
Quel est le taux global d'alignement?
Comparez ces résultats avec ceux issus de l'alignement global, sont-ils différents?

Visualisez maintenant les **fichiers contenant les alignements**.
Que contiennent ces fichiers?
En regardant, la colonne contenant le CIGAR, remarquez-vous une différence entre l'alignement global et l'alignement local?
Que siginifie le caractère **S** qui apparaît dans le CIGAR de l'alignement local?

L'alignement local permet d'aligner les reads en ne tenant pas compte des extrémités des reads. A votre avis, qu'indiquent ces résultats?

Les extrémités des reads ne s'alignent pas sur la séquence de référence, nous allons donc vérifier la qualité des données avec FastQC.

Contrôle qualité avec FastQC

Lancez FastQC sur les reads R1 avec ces paramètres:

- "Raw read data from your current history *" `reads_R1.fastq`

Lancez maintenant FastQC sur les reads R2:

- "Raw read data from your current history *" `reads_R2.fastq`

Vous pouvez agréger ces résultats avec MultiQC. Lancer MultiQC avec ces paramètres:

- "Which tool was used generate logs?" `FastQC`
- Cliquez sur insert FastQC output
- "Type of FastQC output? *" `Raw data`
- Dans "FastQC output *", sélectionnez les fichier "raw data" de FastQC pour les reads R1 et R2 en maintenant la touche Ctrl.

Que pouvez-vous dire sur la qualité des reads R1 et R2?
Que pouvez-vous dire sur la distribution de taille des reads R1 et R2?
Existe-t-il des séquences sur-représentées?
Des adaptateurs sont-ils présents dans les jeux de données?

Les reads ont besoin d'être nettoyés, pour cela nous allons utiliser *cutadapt*.

Nettoyage des données avec cutadapt

Lancez cutadapt avec ces paramètres:

- "Single-end or Paired-end reads?" `Paired-end`
- "FASTQ/A file #1 *" `reads_R1.fastq`
- "FASTQ/A file #2 *" `reads_R2.fastq`
- Dans **Read 1 Options** cliquez sur "Insert 3' (End) Adapters"
  - Choisissez "Enter custom sequence"
  - "Custom 3' adapter sequence": `AGATCGGAAGAGCACACGTCTGAACTCCAGTCA`
- Dans **Read 2 Options** cliquez sur "Insert 3' (End) Adapters"
  - Choisissez "Enter custom sequence"
  - "Custom 3' adapter sequence": `AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT`
- Dans **Filter Options**:

- "Minimum length (R1)": `50`

- Dans **Read Modification Options**
  - "Quality cutoff *": `20`
  - "Trim Ns": `Yes`
- "Outputs selector": Report: `Cutadapt's per-adapter statistics. You can use this file with MultiQC`

Renommez les fichiers de sortie contenant les reads nettoyés: cutadap_R1 et cutadapt_R2.

Quelle sera la taille minimum des reads nettoyés?
Quel est le pourcentage des reads R1 contenant un adaptateur?
Quel est le pourcentage des reads R2 contenant un adaptateur?
Combien de reads ont été filtrés car ils étaient trop courts?
Combien de reads reste-t-il après trimming et filtrage?

Nous allons maintenant vérifier la qualité des données après cette étape de nettoyage.

Lancez FastQC sur les reads R1 avec ces paramètres:

- "Raw read data from your current history *" `Cutadapt_R1`

Lancez maintenant FastQC sur les reads R2:

- "Raw read data from your current history *" `Cutadap_R2`

Comment la qualité a-t-elle évoluée?
Est-ce que les adaptateurs ont bien été trimmés?
Reste-t-il des séquences sur-représentées?

Nous allons maintenant mapper les reads nettoyés sur la séquence de référence.

Mapping des reads nettoyés

Lancez *bowtie2* avec ces paramètres:

- "Is this single or paired library": `Paired-end`
- "FASTA/Q file #1 :" `Cutadapt_R1`
- "FASTA/Q file #2: " `Cutadapt_R2`
- "Write unaligned reads (in fastq format) to separate file(s)": `Yes`
- "Will you select a reference genome from your history or use a built-in index?" `Use a genome from the histoty and build index`
- "Select reference genome *" `REF_plasmide.fasta`
- “Save the bowtie2 mapping statistics to the history”: `Yes`

Visualisez le **fichier contenant les statistiques de mapping**.
Combien de reads ont été mappés de façon concordante exactement une fois?
Combien de reads ont été mappés de façon concordante plus d'une fois?
Parmi les reads non-alignés ou alignés de façon discordante, quel est le pourcentage de reads non-alignés?
Quel est le taux global d'alignement?
Est-ce que les résultats de ce mapping vous semblent corrects?

En cochant l'option "write unaligned reads", en plus du fichier d'alignement et du fichier de statistiques, deux autres fichiers sont générés.
A quoi correspondent ces deux fichiers?

Il est souvent intéressant d'étudier les reads non-mappés. En effet, ils peuvent apporter des informations importantes comme la présence d'un contaminant ou encore une insertion dans le génome étudié. Pour avoir une idée rapide, il est possible d'utiliser *Kraken2* qui permet d'assigner un taxon à un ensemble de reads.

Lancez *kraken2* avec ces paramètres:

- "Single or paired reads": `Paired`
- "Forward strand *" `unaligned reads (L)`
- "Reverse strand *" `unaligned reads (R)`
- "Print scientific names instead of just taxids": `Yes`
- Dans **Create report** "Print a report with aggregrate counts/clade to file": `Yes`
- "Select a Kraken2 database *": `Prebuild Refseq indexes: PlusPF-16`

Visualisez le fichier **Kraken report**.
De quel organisme proviennent la plupart des reads non-mappés?
Cela vous semble-t-il cohérent?
Pour quelle raison ces reads n'ont pas été mappés sur la séquence du plasmide?
Quels sont les autres organismes identifiés dans l'échantillon?

Taille d'insert

Nous allons étudier la distribution des tailles d'insert à partir du fichier d'alignement.

Lancez *CollectInsertSizeMetrics* avec ces paramètres:

- "Select SAM/BAM dataset or dataset collection *": `Bowtie2 alignments`
- "Load reference genome from": `History`
- "Use the folloing dataset as the reference sequence *": `REF_plasmid.fasta`

Visualisez le fichier **PDF**.
Quelle est la distribution des tailles d'insert?
Que pouvez-vous en déduire?

Il est possible de lancer bowtie2 en changeant la taille d'insert.

Relancez *bowtie2* avec une taille d'insert maximale de 700 bp (dans la partie "Do you want to set paired-end options?").

Quelle est la taille d'insert maximale par défaut?
Est-ce que les stastiques de mapping changent?

Etude des alignements

Nous allons étudier quelques alignements. Pour cela, ouvrez les alignements générés lors du dernier mapping avec bowtie2. Pour décoder les SAM flags, vous pouvez utiliser le site du Broad Institute.

Etudions la paire de reads **DRR184077.32629**.
Quel sont les flags associés à cette paire de reads? A quoi correpondent-ils?
Que pouvez-vous en déduire?

Quels sont les flags pour la paire de reads **DRR184077.59916**?
Que pouvez-vous en conclure? Le mapping est-il concordant ou discordant?

Quels sont les flags pour la paire de reads **DRR184077.32494**?
Que pouvez-vous en conclure? Le mapping est-il concordant ou discordant?

Visualisation du mapping dans IGV

Pour pouvoir visualiser le mapping dans IGV, nous allons exporter les résultats du dernier alignement généré avec bowtie2. Pour cela, cliquez sur l'icône *disquette [Download]*. Deux fichiers vous sont proposés:

- Download Dataset : le fichier de mapping (format .bam)
- Download bam_index : l'index du mapping (format .bai)

Téléchargez et enregistrez ces deux fichiers **dans le même répertoire**.

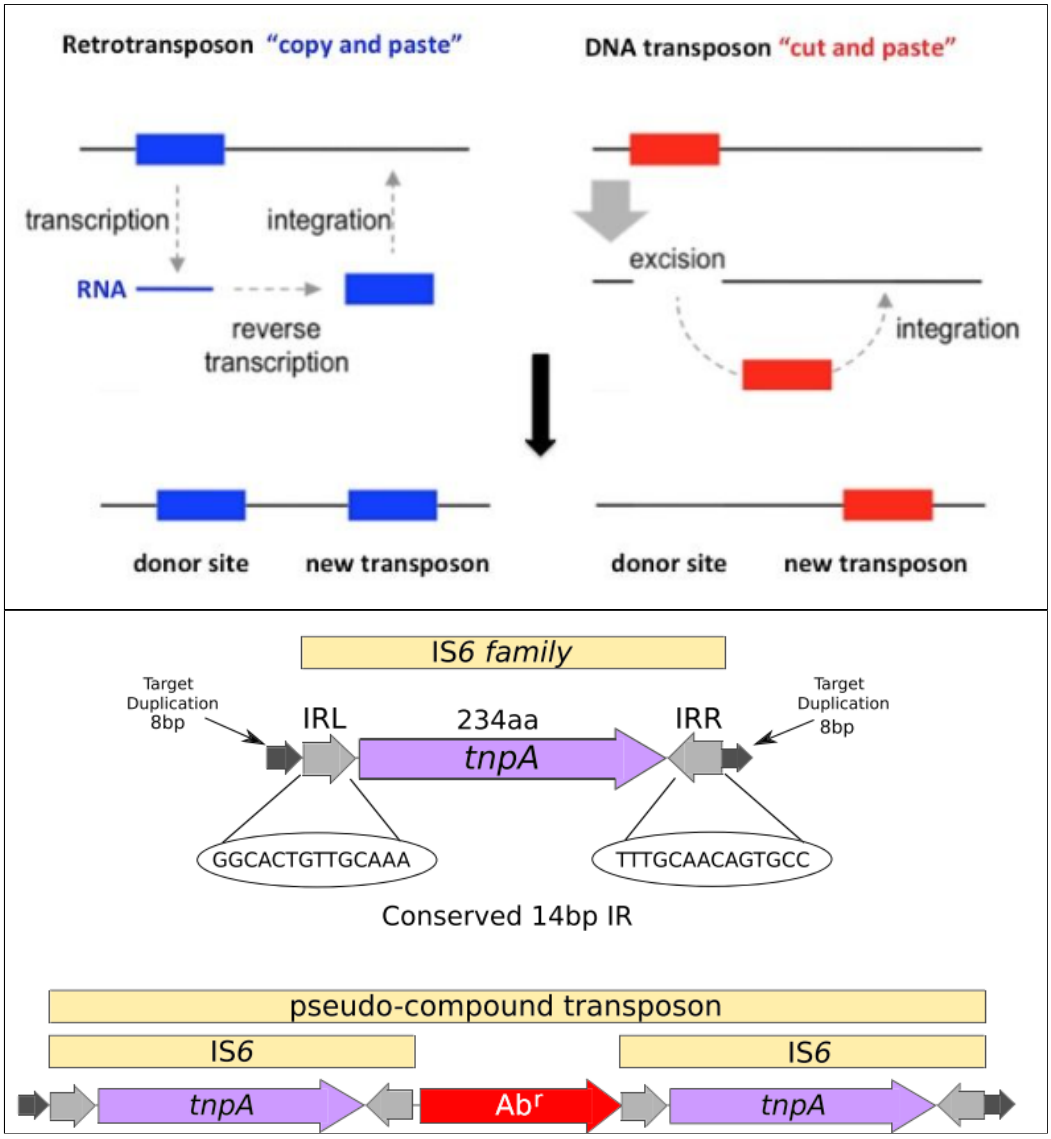
Ouvrez IGV. Dans le menu *Genomes* cliquez sur *Genome Load from file* et choisissez le fichier contenant la séquence de référence **au format genbank** [*REF_plasmid.gb*].
Ensuite dans le menu *File* cliquez sur *Load from file* et chargez le fichier d'alignement qui vient d'être téléchargé de Galaxy au format .bam [*Galaxy.bam*].
Pour voir les reads s'afficher, il faut zoomer à l'aide du **+** en haut à doite.
Par défaut, IGV considère les reads comme non-pairés. Faites un click droit sur un read, puis dans le menu cochez *View as pairs*.
Vous pouvez-maintenant explorer les alignements dans les différentes parties de la séquence de référence.

Vous pouvez changer les options d'affichage des alignements à partir du menu obtenu en faisant en click droit sur un read. Par défaut l'alignement est coloré par la taille d'insert et l'orientation de la paire.

Changez la coloration de l'alignement pour avoir la couleur en fonction de la **taille d'insert**.
En cliquant sur un read vous obtenez les données d'alignement pour la paire de reads sélectionnée.
A quoi correspondent les couleurs utilisées (rouge et bleu)?
Sélectionnez maintenant la couleur en fonction de **l'orientation de la paire**. A quoi correspondent les couleurs vert, bleu et turquoise?

Observez la région comprise entre 6Kb et 20 kb. Vous pouvez revenir à la vue globale en cliquant sur l'icône *maison* puis sélectionnez la zone d'intérêt sur la première piste.
Que remarquez-vous sur cette partie de la séquence de référence?
Quelle est la signification biologique?
Quels sont les gènes impliqués? (vous pouvez cliquer sur les annotations de la dernière piste)

Observez maintenant la région comprise entre 0 et 10 kb. Nous allons étudier la qualité de mapping (MAPQ). Pour cela accédez aux options d'affichage (click droit sur un read), cliquez sur *Shade alignments by* et sélectionnez *mapping quality low*.
Quelle est la signification du MAPQ?
Remarquez-vous des zones avec un faible MAPQ?
A quoi correspondent ces zones?



source: Varani et al. (2021) https://www.judithrecht.com/blog/jumping-pieces-in-our-

Les séquences d'insertion (IS) sont des fragments d'ADN courts qui codent une enzyme, la transposase, permettant le mouvement d'une partie d'ADN d'un site donneur vers un site accepteur. Les IS sont présentes dans les génomes procaryotes en plusieurs copies, représentant donc des séquences répétées. La famille des IS6 à un rôle important dans la disséminitation des gènes de résistance aux antibiotiques.

Observez la région comprise entre les positions 27,429 et 27,577. Vous pouvez directement entrer les coordonnées, *CP023959.1:27,429-27,577*, dans la boîte en haut à coté du génome.

Que remarquez-vous sur cette partie de la séquence?
En cliquant sur la piste de la couverture, vous accédez au décompte de chacun des quatre nucléotides alignés à cette position.
Quelle est la profondeur de séquençage à la position 27,462? Quelle est la répartition des bases à cette position?
Quelle est la profondeur de séquençage à la position 27,463? Quelle est la répartition des bases à cette position?

Observez la région *CP023959.1:37,614-38,348*.
Cette région correpond-elle à un gène ? Si oui lequel?
Le gène TraT code une protéine qui intervient dans une diminution de la réponse immunitaire de l'hôte et induit une résistance aux traitements antibiotiques.
Le gène est-il sur le brin direct ou indirect? Il est possible, en faisant un click droit sur la piste de la séquence en bas, d'afficher la traduction de la séquence en cochant *show translation*. Est-ce que l'une des phases de lecture est correcte?
Vous pouvez changer la direction du brin en cliquant sur *flip strand* afin d'obtenir la séquence du brin indirect. Pouvez-vous maintenant obtenir la séquence protéique associée?
Qu'observez vous en position 37,914?
Pensez-vous que cette variation peut avoir un impact sur la protéine?
Il est possible d'extraire la séquence consensus de la partie affichée. Pour cela, faites un click droit dans la piste de l'alignement et choisissez "copy consensus sequense", ce qui permet ensuite d'étudier cette séquence à l'aide d'autres logiciels.

Conclusions

Le mapping consitue une étape très importante de l'analyse NGS. Il n'est pas une finalité, mais constitue la base d'analyses ultérieures telles que la détection de variant, l'analyse RNAseq ou encore le ChipSeq. Obtenir un fichier d'alignement correct, en utilisant les bons outils et les bons paramètres, est donc une étape cruciale dans beaucoup d'analyses NGS.