Projet Bioinformatique

Système - HAI724I

Anna-Sophie Fiston-Lavier

1 Motivation biologique

Depuis 2005, de nouvelles technologies de séquençage ("NGS" ou *Next Generation Sequencing*) d'ADN ont émergé. Ces technologies ont pour but de faciliter l'analyse de séquences d'espèces proches ou d'individus d'une même espèce. Ces analyses de séquences sont des analyses clefs aussi bien dans le domaine de la Santé que de l'Ecologie.

Une approche classique de l'analyse de séquences consite à aligner les séquences afin d'identifier les différences (voir Figure 1). Ces différences peuvent correspondre à une différence d'une base ou de plusieurs bases. On utilise les termes de polymorphisme ou variant pour qualifier respectivmement ces différences.

```
Coor      12345678901234   5678901234567890123456789012345
Ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
r001/1          TTAGATAAAGGATA*CTG
r002          aaaAGATAA*GGATA
r003        gcctaAGCTAA
r004                        ATAGCT..............TCAGC
r003                               ttagctTAGGC
r001/2                                          CAGCGGCAT
```

```
@SQ SN:ref LN:45
r001  99    ref  7   30  8M2I4M1D3M  =  37  39   TTAGATAAAGGATACTG  *
r002  0     ref  9   30  3S6M1P1I4M  *  0   0    AAAAGATAAGGATA     *
r003  0     ref  9   30  5S6M        *  0   0    GCCTAAGCTAA        *
r004  0     ref  16  30  6M14N5M     *  0   0    ATAGCTTCAGC        *
r003  2064  ref  29  17  6H5M        *  0   0    TAGGC              *
r001  147   ref  37  30  9M          =  7   -39   CAGCGGCAT          *  NM:i:1;
```

Figure 1: Alignements de reads et fichier de mapping (SAM)

Cependant, ces technologies NGS encore couramment utilisées ne peuvent lire que des petits fragments de séquences d'ADN d'intérêt. Ces courtes lectures (ou *reads*) vont de actuellement de 150 à 300 paires de bases, cela dépend des machines de séquençage. Les reads peuvent correspondre à la lecture d'une

extremité (*single-end*; ; voir Figure 1 : R002) ou de deux extrémités (*paired-ends*; voir Figure 1 : R001/1 -R001/2) d'une séquence d'intérêt.

2 SAM, le format de fichier de mapping

Un format de fichier appelé "SAM" pour "**S**equence **A**lignment/**M**ap" créé spécifiquement pour stocker ces données (Figure 1). Le format SAM est un fichier plat tabulé. Les premières lignes du fichier commençant par le caractère "@" correspondent à l'entête (ou *header*). Elles contiennent les informations générales sur le fichier comme les noms des séquences de références et l'outil d'alignement qui a permis de générer ce fichier (Figure 2).

Le corps du fichier SAM contient les informations d'alignement. Ces informations sont organisées et donc accessible (Figure 2).

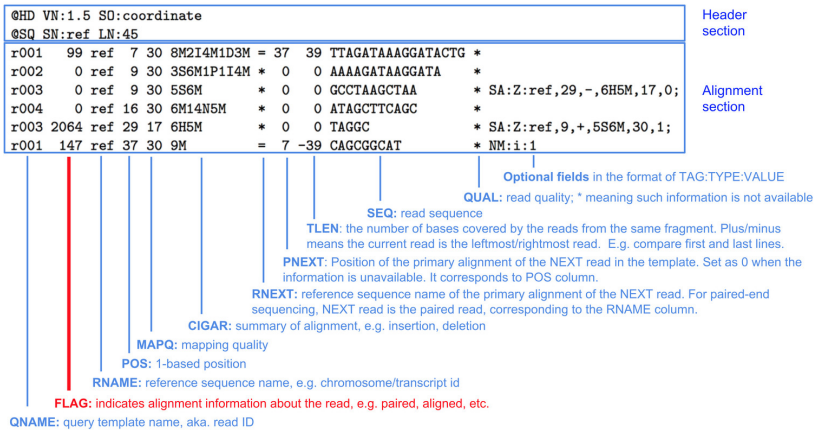

Figure 2: SAM format

3 Lecture, Stockage et Parcours des données de mapping

Le but du projet consiste à implémenter un script Python qui affiche un résumé des informations des alignements (ou *mapping*) de lectures (ou *reads*) issues du séquençage d'ADN sur une séquence dite de référence (Figure 1: Ref). Les séquences de référence peuvent correspondre aux séquences d'ADN entières ou partielles des chromosomes. Pour ce faire, vous devrez *parser*, c'est-à-dire lire, stocker et parcourir les données de mapping [1].

L'étape suivante consitera à extraire les reads "mal mappés", c'est-à-dire les reads:

- non-mappés,
- partiellement mappés,
- les paires de reads où un seul read de la paire est entièrement mappé et l'autre non mappé,
- les paires de reads où un seul read de la paire est entièrement mappé et l'autre partiellement mappé.

Pour ce faire, vous allez devoir sélectionner les lignes du fichier SAM en fonction du FLAG [1].

La dernière étape sera d'afficher le résumé des résultats obtenues, comme le nombre de reads et paires de chaque catégorie. Les sorties graphiques ne sont pas nécessaires.

4 Rendus

Chaque groupe devra rendre sur moodle un fichier compressé contenant 3 documents:

- Un rapport de 5 pages. Ce rapport devra être organisé de manière à avoir:
  - une introduction,
  - une partie de présentation des données SAM,
  - une exemple d'application (motivation biologique),
  - une partie expliquant votre projet/script,
  - une discussion sur les aspects positifs et négatifs de votre script.
- Un fichier README. Ce fichier doit contenir toutes les informations pour installer et utiliser votre script.
- et logiquement, le script Python dans un dépôt git.

References

[1] *Sam format: https://www.samformat.info/sam-format-flag.*