TP Introduction à l'analyse de données NGS (des données brutes au mapping)

INTRODUCTION

Les précédents exercices ont permis :

- d'introduire les commandes de base pour se déplacer dans l'arborescence et la modifier
- de manipuler les fichiers de données, de filtrer des lignes d'un fichier, de les trier
- de lier des commandes et de rediriger les sorties des commandes

L'objet de ce TP2 est d'utiliser linux ***de manière plus avancée*** en utilisant les commandes vues en cours et en lançant quelques programmes bio-informatiques et scripts.

Durant ce TP, vous allez réaliser en ligne de commande les premiers traitements réalisés lorsque l'on reçoit des données Illumina. Dans le répertoire Illumina, vous avez les 2 fichiers issus d'un séquençage RNA-seq de différents accessions sauvages et cultivés de riz (technologie illumina) ainsi que le fichier contenant la référence (qui va servir pour l'étape de "mapping") et un fichier contenant les primers/adaptateurs (qui servira à l'étape de "cleaning").

Formation bio-informatique, Module Linux
Alexis, Dereeper, Christine Tranchant-Dubreuil

Exercice 1: Vérification rapide de la validité des séquences et du format

1. *Visualiser le contenu des 2 fichiers de séquences dans le répertoire ~/Data/Illumina.*

   head, tail

2. *Quel est le format des fichiers?*

3. *En observant les scores de qualité, quel est le format utilisé pour coder la qualité?*

4. *Valider rapidement les 2 fichiers format fastq en vérifiant qu'il y a le même nombre de lignes dans les 2 fichiers de séquences. Le transfert des fichiers de séquence sur le serveur s'est il bien déroulé et a-t-il été complet?* ***wc -l***

Exercice 2: Mise en place de l'arborescence au niveau du dossier Illumina

1. Créer le répertoire 0_fastq et déplacer les 2 fichiers de séquences illumina dans ce nouveau répertoire ***mkdir, cp, rm***

2. *Créer le répertoire Bank et déplacer le fichier reference.fasta dans ce répertoire*

   mkdir, cp, rm

Exercice 3: Contrôle qualité - *logiciel fastqc*

1. *Créer le répertoire 1_fastqc* ***mkdir***

2. *Lancer successivement le logiciel fastqc sur les 2 fichiers fastq brutes reçus du de la boîte de séquençage*

Ligne de commande du logiciel fastqc

```
fastqc -o 1_fastqc/ nom_fichier_fastq_a_analyser
```

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

3. *Transférer le dossier 1_fastqc du serveur sur votre poste de travail pour analyser les résultats du logiciel fastqc* ***filezilla (protocole sftp)***

Regardez les 2 fichiers fastqc_report.html générés par fastqc pour chaque fichier fastq brute. Quel est le nombre de séquence et la taille moyenne des séquences? Quel est le format utilisé pour encoder la qualité?

Exercice 4: Traitement de la qualité/trimming - *logiciel cutadapt*

1. *Créer le répertoire 2_cutadapt* ***mkdir***

2. *Lancer le logiciel cutadapt*

```
cutadapt –O 7  -b sequence1  -B sequence1  -q 30,30  -m 35
     -o 2_cutadapt/sequence_1.CUTADAPT.fastq -p
     2_cutadapt/sequence_2.CUTADAPT.fastq sequence_1_a_traiter.fastq
     sequence_2_a_traiter.fastq
```

https://code.google.com/p/cutadapt/

Quelques explications pour mieux comprendre les options de cutadapt :

-b sequence1 : séquence de l'adapteur à enlever en 5'
-B sequence1 : séquence de l'adapteur à enlever en 3'
-q 30,30 : seuil de qualité minimale de la séquence 1, seuil de qualité minimale de la séquence 2
-O 7: la séquence de l'adapteur doit s'aligner avec la séquence d'au moins 7 pb
-m 35 : après le retrait des extrémités de mauvaise qualité, les séquences inférieures à 35 pb sont éliminées
-e 0.1 : permet un mismatch de 1 dans 10 bases entre la séquence du read et de l'adaptateur

3. *Lancer le logiciel fastqc sur les 2 fichiers fastq générés par cutadapt.*

Regardez les 2 fichiers fastqc_report.html générés par fastqc pour chaque fichier fastq brute. Quel est le nombre de séquences et la taille moyenne des séquences? Quel est le format utilisé pour encoder la qualité?

Exercice 5: Mapping des reads contre une référence - *logiciel bwa*

1. *Aller dans le répertoire bank. La première étape va être de créer les fichiers index de la référence nécessaires à bwa.*

Ligne de commande du logiciel bwa index

```
bwa index reference.fasta
```

Lister le contenu du répertoire Bank. Que remarquez vous? ***ls -lt***

2. *Créer le répertoire 3_bwa dans le répertoire NGS* ***mkdir***

3. *Lancer l'analyse de mapping dans ce répertoire à l'aide du logiciel bwa*

   a) *Lancer la commande* ***bwa aln*** *sur chaque fichier de read (forward et reverse) Cette première commande génère les alignements pour chaque read (indépendamment du read "pairé") contre la réference (fichier .sai généré).*

Ligne de commande du logiciel bwa aln

```
bwa aln -f 3_bwa/sequence_1.sai reference.fasta
sequence_1.CUTADAPT.fastq
bwa aln -f 3_bwa/sequence_2.sai reference.fasta
sequence_2.CUTADAPT.fastq
```

   b) *Vérifier en listant le contenu du répertoire que les fichiers d'extension .sai ont bien été créés et qu'ils ne sont pas vides* ***ls -lt***

   c) *Lancer la commande* ***bwa sampe***

Cette deuxième commande permet de "pairer" les deux reads et de produire l'alignement dans le format SAM.

Ligne de commande du logiciel bwa sampe

```
bwa sampe -f 3_bwa/all_seq.sam reference.fasta 3_bwa/sequence_1.sai
3_bwa/sequence_2.sai sequence_1.CUTADAPT.fastq
sequence_2.CUTADAPT.fastq
```

d) *Vérifier en listant le contenu du répertoire que le fichier d'extension .sam a bien été créé et qu'il n'est pas vide* **ls –l**

Exercice 6: Manipulation des fichiers .sam - *logiciel samtools*
http://samtools.sourceforge.net/samtools.shtml

1. *Afficher les premières lignes du fichier .sam* **head**

2. Convertir le fichier sam en fichier bam et créer l'index du fichier bam généré

samtools view

```
samtools view –S 3_bwa/all_seq.sam –b -o 3_bwa/all_seq.bam
```

OPTIONS :

-b output BAM

-S input is SAM

index ?

3. Récupérer des statistiques du mapping avec le logiciel samtools flagstat

samtools flagstat

```
samtools flagstat 3_bwa/all_seq.bam
```

4. On souhaite extraire du SAM uniquement les reads mappés correctement au niveau de la séquence forward ET reverse. Le logiciel samtools view permet de filtrer sur la colonne 2 "flag" du fichier sam (cf. rappel format sam juste au dessus).

a) Vérifier au niveau du site ***http://picard.sourceforge.net/explain-flags.html*** *que le flag avec la valeur 0X2 est correct pour effectuer ce filtre.*

b) Conversion du fichier SAM en fichier BAM (format compressé utilise par les logiciels) en séparant les reads mappés et non mappés **samtools view**

Ligne de commande du logiciel samtools view

```
samtools view 3_bwa/all_seq.bam -f 0X2 –b –o 3_bwa/all_seq.MAPPED.bam
```

OPTIONS :

-f INT Only output alignments with all bits in INT present in the FLAG field. INT can be in hex in the format of /^0x[0-9A-F]+/ [0]

c) Vérifier en listant le contenu du répertoire que le fichier a bien été créé et qu'il n'est pas vide ***ls -lt***

Exercice 7 : Tri du fichier bam

Beaucoup de logiclels d'analyses utilisant les fichiers bam demande un fichier bam trié par chromosome et par position. Nous allons réaliser cette étape avec le programme ***samtools sort.***

1. *Lancer le programme* ***samtools sort*** *sur notre fichier bam créé à l'étape précédente*

Ligne de commande du logiciel samtools sort

```
samtools sort 3_bwa/all_seq.MAPPED.bam prefixe_fichier_bam
```

Par exemple, le préfixe utilisé peut être `all_seq.SORT`

2. Lister le contenu du répertoire et regarder si des nouveaux fichiers ont été créés. ***ls -lt***

Exercice 8 : Création des indexes du fichier bam

De nombreux logiciels (utilisant les fichiers bam) demandent également la création de fichiers index permettant d'accéder aux informations relatives aux "read" plus rapidement. Nous allons réaliser cette étape avec le programme ***samtools index.***

1. *Lancer le programme* ***samtools index*** *sur notre fichier bam créé à l'étape précédente*

Ligne de commande du logiciel samtools index

```
samtools index all_seq.SORT.bam
```

2. Lister le contenu du répertoire et regarder si des nouveaux fichiers ont été créés. ***ls -lt***

Exercice 9 : Recherche de SNP/Indel avec *samtools mpileup*

1. La première étape est de créer un index de la référence nécessaire pour la suite de l'analyse.

Ligne de commande du logiciel samtools faidx

```
samtools faidx reference.fasta
```

2. L'étape suivante est de lancer samtools mpileup pour générer le fichier bcf (format vcf compressé).

Ligne de commande du logiciel samtools mpileup

```
samtools mpileup -u -f ../Bank/reference.fasta all_seq.SORT.bam   >
all_seq.SORT.bcf
```

3. La dernière étape va être de générer le fichier vcf avec le logiciel bcftools

Ligne de commande du logiciel samtools mpileup

```
bcftools view -v -c -g all_seq.SORT.bcf > all_seq.SORT.bam.vcf &
```

Regarder le contenu du fichier généré. Combien de SNP ont été détectés?