**Analyse de données NGS**

***(Module3 session 2)***

**# 1ère étape:** création d'un dictionnaire de ton génome de référence (utile pour bwa)

## Par exemple le nom du fichier de mon génome de référence est:

**genomeReference.fasta**

**Commande à exécuter: bwa index genomeReference.fasta**

**# 2ème étape:** alignement du ou des fichiers fastq contre le génome de référence selon ta situation

**## 1er cas:** je n'ai qu'un seul fichier fastq, exemple **Ech.fastq.** Le fichier de sortie sera un fichier **.sam**

**Commande à exécuter: bwa mem -t 2 genomeReference.fasta Ech.fastq > alignement.sam**

***Nb: ### le numéro 2 après -t, est le nombre de cœurs de votre ordinateur que vous devez utiliser pour faire votre alignement. Plus vous mettez de coeurs, plus l'alignement est rapide.***

***Attention !!!! votre ordinateur a un nombre de coeurs bien précis. Vous ne devez pas mettre un chiffre qui est égale ou supérieur au nombre de coeurs de votre ordinateur.***

***### Pour connaitre le nombre de coeurs de ton pc utilise la commande:***

***nproc***

**## 2ème cas:** j'ai mes deux fichiers fastq (R1 et R2), exemple Ech\_R1.fastq et Ech\_R2.fastq

**Commande à exécuter: bwa mem -t 2 genomeReference.fasta Ech\_R1.fastq Ech\_R2.fastq > alignement.sam**

**##3ème cas:** j'ai mes deux fichiers fastq (R1 et R2) qui sont compréssés (.gz ou .zip), exemple Ech\_R1.fastq.gz et Ech\_R2.fastq.gz

**Commande à exécuter : bwa mem -t 2 genomeReference.fasta '<zcat Ech\_R1.fastq' '<zcat Ech\_R2.fastq' > alignement.sam**

**# 3ème étape:** transformation de ton fichier sam en bam

**Commande à executer:** **samtools view -b -S alignement.sam > alignement.bam**

**# 4ème étape:** le trie de ton fichier bam

**Commande à exécuter : samtools sort alignement.bam -o alignement\_sorted.bam**

**# 5ème étape:** l'indexation de ton fichier bam trié

**Commande à exécuter : samtools index alignement\_sorted.bam**

## Si vous avez réussi toutes les étapes précédemment, vous pouvez utiliser le logiciel tablet téléchargeable en ligne pour regarder l'alignement.

## Tablet aura besoin du fichier **alignement\_sorted.bam**, fichier **alignement\_sorted.bam.bai** et de **genomeReference.fasta**

## Si votre alignement est bon, vous n’aurez plus besoin de certains fichiers qui sont volumineux (le .sam et .bam). Vous ne les utiliserez plus jamais, vous pouvew donc les supprimer en exécutant la commande suivante :

**rm alignement.sam**

**rm alignement.bam**

**# 6ème étape:** appel de variants par l’outil samtools

La première tâche consistera à **indexer** notre séquence de référence **genomeReference.fasta**

**Commande à exécuter : samtools faidx genomeReference.fasta**

La deuxième tâche consistera à la genèse de variants et pour cela il faut juste lancer la commande suivante :

**Commande à exécuter : samtools mpileup -g -f genomeReference.fasta alignement\_sorted.bam > alignement\_sorted.bcf**

L’option **-g** va générer les vraisemblances des génotypes dans le fichier de sortie .bcf

L’option **-f** fait appel à la séquence de référence **genomeReference.fasta**

**# 7ème étape:** appel de **SNP** avec l’outil BCFtools

Pour faire appel au single nucleotid polymorphism (SNP) nous utiliserons l’outils BCFtools et la commande à exécuter est la suivante :

**Commande à exécuter : bcftools call -O b -vc alignement\_sorted.bcf >**

**Alignement\_SNP.bcf**

-O c’est le type de fichier de sortie (b=BCF)

-v sort uniquement les sites qui contiennet les variants

-c utilise la méthode originale d'appel au consensus

**# 8ème étape:** filtration de **SNP** avec l’outil vcfutils.pl

Dans cette étape nous allons filtrer les différents SNP et la commande à exécuter est la suivante :

**Commande à exécuter : bcftools view Alignement\_SNP.bcf | vcfutils.pl varFilter - > alignement\_var\_final.vcf**

**# 9ème étape:** Annotation de variants en utilisant l’outil **SnpEff**

Le pipeline présenté ici s’arrête après l’étape de **filtrage.** Ceci dit, une fois nos SNPs obtenus, nous pouvons également les comparer à des bases de données afin de les annoter. Pour ce cas précis nous allons utiliser SnpEff. SnpEff est inclus dans l’outil GATK avec la commande suivante

**Commande à executer: java -jar snpeff.jar -c alignement\_var\_final.vcf -no-downstream -no-upstream -onlyProtein Pf3D7v3 > annotated.vcf**

**Ainsi s’achève cette analyse. L’interprétation et la validation dépendront de la question biologique posée au départ.**