Formation à l'alignement de séquences et à la recherche de polymorphismes

- EXERCICES -

"**FastQC** is a quality control tool for high throughput sequence data."
http://www.bioinformatics.bbsrc.ac.uk/

"**Burrows-Wheeler Aligner** (**BWA**) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome." http://bio-bwa.sourceforge.net

"**SAM** (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments."
http://samtools.sourceforge.net

"The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated datasets. It supports a wide variety of data types including sequence alignments, microarrays, and genomic annotations."
http://www.broadinstitute.org/igv

"The **Genome Analysis Toolkit** or **GATK** is a software package developed at the Broad Institute to analyse next-generation resequencing data. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size."
http://www.broadinstitute.org/gatk/

"A set of tools (in Java) for working with next generation sequencing data in the BAM format." http://picard.sourceforge.net/

Objectifs :

Cette formation a pour objectif de vous aider à traiter les séquences issues des SGS avec l'environnement GALAXY. Vous y découvrirez les formats de séquences, d'alignements et de variants et mettrez en œuvre des logiciels dédiés à l'alignement de reads sur génome de référence, à la recherche et au filtre et à l'annotation de polymorphismes.

Pour vous connecter à l'instance SIGENAE de GALAXY :
http://sigenae-workbench.toulouse.inra.fr

Vous pouvez utiliser vos identifiants et mots de passe de votre compte sur la plateforme genotoul/bioinfo, ou bien utiliser un des comptes disponibles le temps de la formation:

- *Logins:* anemone, aster, bleuet, cobee, capucine, coquelicot, arome, clematite, chardon, camelia
- *Password:* … demandez aux formateurs ...

Récupération des données

À partir de NG6:

- Parcourir le projet «Galaxy training» et le run «Galaxy - DNAseq SNP» disponible dans NG6.
- Créer les liens symboliques de l'ensemble des fichiers du run dans votre compte Genotoul.
- A partir de Galaxy uploader ces fichiers dans votre historique de travail.

*ou à partir de l'historique publié « **TP_SNP_Data_Fév2018** » :*

- Importer cet historique.

Doivent être présents dans votre historique de travail les datasets:

- variants_BosTaurus.vcf
- Bos_taurus_incl_consequences-chr25.vcf
- SRR1425152-chr25_1.fastq.gz
- SRR1425152-chr25_2.fastq.gz
- SRR1425153-chr25_1.fastq.gz
- SRR1425153-chr25_2.fastq.gz
- SRR1425154-chr25_1.fastq.gz
- SRR1425154-chr25_2.fastq.gz
- ensembl_bos_taurus_genome-chr25.fa

Exercice n°1 : Analyse de la qualité

Afin d'obtenir plus d'informations sur les jeux de données, rechercher les entrées, sur le site de l'EBI, correspondant aux identifiants: «SRR1425152», «SRR1425153» et «SRR1425154».

- Quel est le type de séquenceur utilisé pour chacune de ces 3 entrées ?
- Lancer FastQC sur un des fichiers FASTQ.
- Explorer le rapport généré.
- Pour le run «SRR1425152» en forward, quel est le nombre de lectures contenant un ou plusieurs 'N' ?

Aide :
http://regexr.com/ permet d'écrire et de tester les expressions régulières.

Exercice n°2 : Alignement des séquences

Quelques liens:

- BWA : http://bio-bwa.sourceforge.net

Alignement des lectures avec les outils «BWA» :

- Le fichier FASTA de référence sur lequel nous allons aligner nos lectures est disponible dans l'historique: ensembl_bos_taurus_genome-chr25.fa.
- Aligner les FASTQ de «SRR1425152» sur la référence. Tester éventuellement les différents algorithmes.

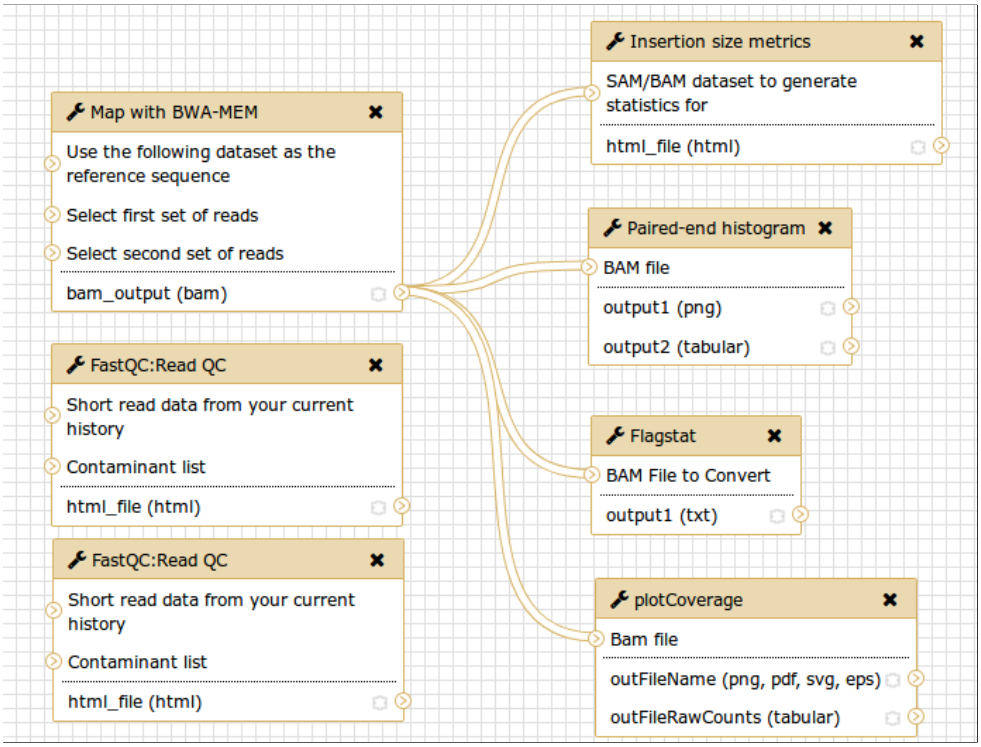
> **Attention** :
> Bien regarder les paramètres que les outils proposent (Single/Paired)...

- Comparer les formats de sortie de chaque outil et le taux d'alignement grâce à l'outil Flagstat.
- Transformer les fichiers BAM en SAM (si besoin) et tenter de comprendre les différentes informations du format SAM.

> **Attention** :
> Le header fait partie intégrante du format SAM !

Construire un workflow réalisant l'alignement BWA-MEM et quelques statistiques suivant le modèle ci-dessous.

L'appliquer aux trois couples de FASTQ.



Aller plus loin avec Galaxy :

- le nommage des datasets
- la notion de datasets collections

À l'issue de cet exercice, vous devez au minimum avoir les fichiers BAM obtenus par l'alignement des fichiers SRR1425152 et SRR1425153 *avec BWA-MEM.*

Exercice n°3 : Formats, manipulations et conversions

Quelques liens:

- SAMtools: http://samtools.sourceforge.net
- Picard: http://picard.sourceforge.net

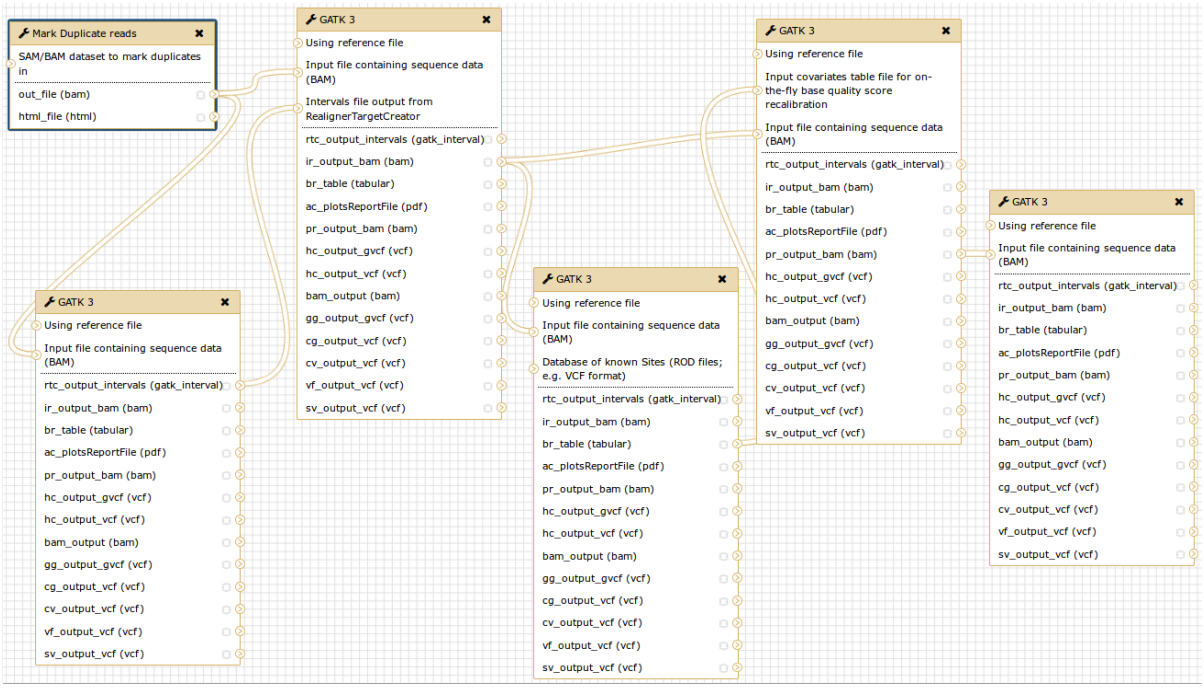
Afin de se familiariser avec le format SAM et les SAMtools, à partir du fichier SAM «SRR1425152» :

- Repérer les différents champs du SAM.
- Quels sont les différents «flags» présents ? Que signifient-ils ?

> **Aide** :
> « Traduire » la valeur d'un flag : [explain-flags]
> « Traduire » des valeurs de qualité : [explain-qualities]

- Combien de lectures ont pour «CIGAR» 100M ?
- Combien de lectures sont **parfaitement** alignées ?
- Pourquoi les réponses aux deux questions précédentes diffèrent-elles (rappel: la longueur des «reads» est de 100 pb) ?
- Générer l'histogramme des qualités de mapping
- Extraire les 1000 premières bases de notre référence
- Extraire les reads alignées sur les 1000 premières bases de notre référence

A partir des 2 (ou 3) fichiers BAM générés au cours de l'exercice 2 :
Importer le workflow partagé « SNP_2_VariantCalling_GATK3.5 ».
Exécuter ce workflow sur chacun des 2 (ou 3) fichiers BAM.



Exercice n°4 : Visualisation

Quelques liens:

- Interactive Genomics Viewer – IGV: http://www.broadinstitute.org/igv

Interactive **G**enomics **V**iewer – IGV :

- Se rendre à l'url suivante: http://www.broadinstitute.org/igv/download
- Deux possibilités pour lancer IGV :
  - Téléchargement sur le PC de formation et exécution en double cliquant sur le fichier igv_win.bat (Note: il est alors possible de modifier la mémoire allouée en éditant ce fichier : -Xmx1g par exemple)
  - Lancement «webstart»
- Importer le génome de référence «ensembl_bos_taurus_genome-chr25.fa» que vous avez au préalable téléchargé sur votre machine depuis votre instance Galaxy.
- Faire de même pour les fichiers BAM/BAI.
- Explorer l'interface (déplacement, zoom, étiquettes d'informations, clic droit, ...)
- Tester les sessions: Enregistrer votre session, supprimer l'ensemble des pistes, recharger votre session sauvegardée en utilisant «Open session».

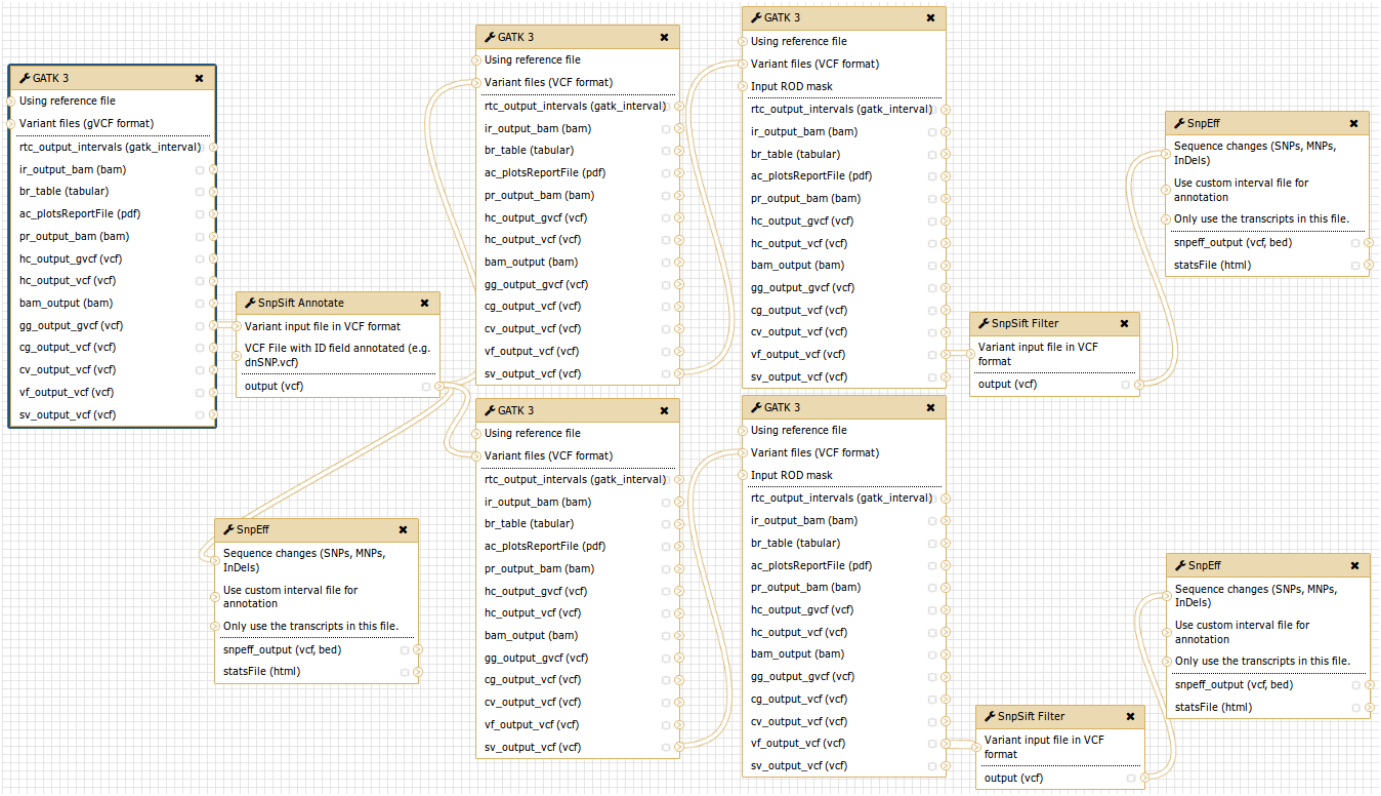
Il est possible de se déplacer sur la référence en «sautant» de «feature» en «feature»:
- Cliquer sur la piste qui contient les «features»
- Se déplacer avec :'Ctrl+F' (suivant) et 'Ctrl+B' (précédent)

Exercice n°5 : Filtre et annotation de variants

Quelques liens:

- SNPeff: http://snpeff.sourceforge.net/
- SNPsift: http://snpeff.sourceforge.net/SnpSift.html

Importer et exécuter le workflow « SNP_3_Filter_GATK3.5 » sur l'ensemble des fichiers GVCF générés lors de l'exercice 3.



Explorer le workflow et les différents datasets obtenus.

Combien de SNP ont été supprimé suite au « HardFilter GATK » ?
Dans le fichier final combien de SNP ont un effet « high », combien tombent dans un gène ?

Afin de se familiariser avec l'outil de filtre SnpSift :

- Extraire les « nouveaux » SNPs
- Parmi ces « nouveaux », combien sont homozygote pour SRR1425152 et hétérozygote pour SRR1425153 ?
- Extraire les SNPs ayant un effet « High »
- Existe-t-il un « nouveau » SNP ayant un effet « High » ?
- Si oui quel « animal » est impacté ?