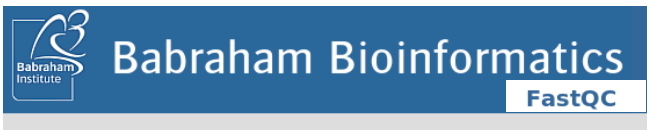
Formation à l'alignement de séquences issues des SGS et à la recherche de polymorphismes

- EXERCICES -



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

BWA

"**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

SAMtools

"**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/

Picard

"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 (Illumina). Vous y découvrirez les nouveaux formats de séquences et d'alignement, les biais connus et mettrez en œuvre des logiciels d'alignement sur génome de référence, de recherche de polymorphismes et de visualisation d'alignement.

Pré-requis : savoir utiliser un environnement Unix.

Pour réaliser l'ensemble de ces exercices, se connecter sur « genotoul » (en utilisant éventuellement « putty » sous windows).

Pour les traitements « lourds » utiliser le cluster avec la commande « qlogin ».

Sur « genotoul », créer dans le répertoire work un répertoire de travail.

Exercice n°1 : Analyse de la qualité

Quelques liens :

- NCBI : http://www.ncbi.nlm.nih.gov
- FastQC : http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

FastQC est déjà installé sur « genotoul ».
Penser à utiliser la complétion en utilisant la touche 'Tab'. (exemple : fast + 'Tab')

Récupération des données au NCBI :

- Sur le site du NCBI rechercher les entrées correspondant aux identifiants : « SRR1425152 », « SRR1425153 » et « SRR1425154 » (en une seule requête).
- Explorer les entrées : organisme ? échantillon ? séquenceur utilisé ? séquençage pairé ? ...
- Les séquences s'alignant sur le chr25 bovin sont disponibles à l'url suivante : http://genoweb.toulouse.inra.fr/~formation/16_SGS-SNP/DATA/. Copier les 6 fastq (read1/read2) dans votre répertoire de travail.

Afin d'optimiser l'espace disque, il n'est plus possible d'accéder directement au fichier « fastq », le NCBI a mis en place le format « sra ».
La conversion en « fastq » d'un fichier « sra » est effectuée grâce à la commande suivante (disponible sur « genotoul » au travers du « sra toolkit ») :

```
fastq-dump [...].sra
```

Manipulation des « fastq », pour le run SRR1425152 :

- Extraire et comparer les 5 premiers identifiants de chaque fastq du run.
- Quel est le nombre de fragments ?
- Combien de lectures contiennent un ou plusieurs « N » ?

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

Statistiques avec FastQC :

- Exécuter « fastqc » sur l'ensemble des fastq.
- Explorer les rapports générés.

Exercice n°2 : Alignement des séquences

Quelques liens :

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

Alignement des lectures avec « BWA » :

- Copier la séquence de référence (chromosome 25 bovin) à partir de l'url suivante : http://genoweb.toulouse.inra.fr/~formation/16_SGS-SNP/DATA/.
- Familiarisation :
  - Afficher l'aide pour visualiser les commandes disponibles
    `bwa`
  - Afficher l'aide pour visualiser une des commandes
    `bwa index`
- Indexer la séquence de référence
  `bwa index ensembl_bos_taurus_genome-chr25.fa`
- Effectuer l'alignement en utilisant l'algorithme mem

Attention utiliser les paramètres suivant :
`-t 4` : « number of threads » afin d'accélérer le traitement
`-M` : « mark shorter split hits as secondary »
`-R '…'` : « read group header line such as '@RG\tID:foo\tSM:bar' »

Exercice n°3 : Formats, manipulations et conversions

Quelques liens :

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

A partir des trois fichiers SAM issues de BWA :

- Visualiser les trois premières lignes de chaque SAM dans un terminal et repérer les différents champs
- Pour le run SRR1425152 :
  - Quels sont les différents « flags » ?
    Que signifient-ils (http://picard.sourceforge.net/explain-flags.html) ?
  - Combien de lectures ont pour « CIGAR » 100M ?
  - Combien de lectures sont **parfaitement** alignées ?
  - Pourquoi les réponses aux deux questions précédentes diffèrent-elles ?
  - Quelle est la « mapping quality » maximum ?
    Combien de lectures ont cette valeur de qualité ?
  - En utilisant le champ « CIGAR », quelles sont les tailles de délétions et leurs effectifs ?
  - En utilisant le champ « MD », quel est le nombre de mismatch ?

- Utilisation des ***SAMtools*** :
  - Familiarisation :
    - Afficher l'aide pour visualiser les commandes disponibles
      `samtools`
    - Afficher l'aide d'une des commandes
      `samtools view`
  - Convertir les fichiers SAM en BAM
  - Afficher les premières lignes d'un des fichiers BAM créés précédemment
  - Une partie du fichier SAM est absente, pourquoi ? Modifier la commande précédente afin de l'afficher ?
  - Trier les BAM
  - Indexer les BAM triés
  - Utiliser la commande faidx des samtools pour extraire les 1 000 premières bases du chr25
  - Combien de lectures pour le run SRR1425152 s'alignent entre les coordonnées 1 et 1000 de la référence ?
  - Calculer les statistiques d'alignement

Exercice n°4 : Visualisations

Quelques liens :

- SAMtools tview : http://samtools.sourceforge.net/tview.shtml
- 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 (il est nécessaire d'avoir le fichier la référence sur votre PC local).
- Importer les fichiers BAM (il est nécessaire d'avoir les fichiers BAM et BAI pour chaque run sur votre PC local).
- Explorer l'interface (déplacement, zoom, étiquettes d'informations, clic droit, ...)
- Récupération des annotations au format GFF.
- Chargement des annotations au format GFF.
- Repérer et marquer quelques régions d'intérêt. Exporter ces régions, effacer et recharger les.
  Menu « Regions »
- Tester les sessions : Enregistrer votre session, supprimer l'ensemble des pistes, recharger votre session sauvegardée en utilisant « Open session ».
  Menu « File »

Exercice n°5 : Recheche SNPs / Indels

Quelques liens :

- GATK : http://www.broadinstitute.org/gatk/ (gatk_path = /usr/local/bioinfo/src/GATK/latest/)
- Picard : http://picard.sourceforge.net (picard_path = /usr/local/bioinfo/src/picard-tools/current/)

Parcourir brièvement les « best practices » GATK : https://www.broadinstitute.org/gatk/guide/best-practices.php

Préprocess Picard/GATK :

- Marquage des duplicats
  - Pour le run SRR1425152
    - Combien de lectures non pairées et de lectures pairées sont dupliquées ?
    - Vérifier ce comptage en utilisant le champ « FLAG » du BAM obtenu.
- Réalignement autour des zones d'INDEL
  - En utilisant le champ « CIGAR », produire un histogramme des tailles de délétions ?
  - En utilisant le champ « MD », quel est le nombre de mismatch ?
- Recalibration de la qualité des bases

Variant calling :

- Effectuer le calling par run
- Joindre les fichiers GVCF obtenus

Exercice n°6 : VCF, annotation, filtre...

Quelques liens :

- SNPeff : http://snpeff.sourceforge.net/index.html (snpeff_path = /usr/local/bioinfo/src/SnpEff/current/)
- SNPsift : http://snpeff.sourceforge.net/SnpSift.html (snpsift_path = /usr/local/bioinfo/src/SnpEff/current/)
- VCF format : http://vcftools.sourceforge.net/specs.html

Repérer les différents champs du fichier VCF obtenu précédemment.
Combien de variants contient le fichier VCF ?
En utilisant SNPeff :

- Afficher l'aide de la commande SNPeff
- Une annotation est-elle disponible pour notre génome d'intérêt ? Si oui, laquelle ?
- Lancer l'annotation
- Repérer les ajouts dans le VCF (champ INFO)

En utilisant SNPsift :

- Afficher l'aide de la commande SNPsift
- Ajouter l'information « snp connus » dans le champ ID du VCF

Grace à SNPsift filter :

- Combien de variants ont une qualité >=30 ?
- Combien de variants sont homozygotes pour SRR1425152 et hétérozygotes pour les deux autres ?
- Combien de variants sont inconnus et ont un impact « HIGH » ?
- …

Ajouter le fichier VCF dans IGV :

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)