Atelier Variant Introduction

Olivier RUÉ - INRA
Guillaume ROBERT-SIEGWALD - Inovarion
Bastien JOB - INSERM / Gustave Roussy

Maria BERNARD - INRA
Elodie GIRARD - Institut Curie

École de bioinformatique AVIESAN-IFB 2018

Objectifs

Processus d'analyse de données de séquences, des filtres de qualité à la détection de variants :

- SNVs et indels de petite taille, à l'aide de 3 outils : GATK, Mpileup/Varscan et discoSNP (Olivier, Guillaume)
- **V**ariations **S**tructurales (SV) et **V**ariations du **N**ombre de **C**opies (CNVs) (Guillaume, Bastien)
- Utilisation de R pour visualiser des métriques/résultats obtenus (Elodie)
- Automatisation des traitements → Workflow (Maria)

Cluster de l'IFB

doc cluster

L'Institut Francais de Bioinformatique met à disposition de la communauté un cluster de calculs

Your turn! Se connecter au cluster

 # Sous Windows avec MobaXterm

Session : ssh

Host : core.cluster.france-bioinformatique.fr

Specify username : coché et complété

 # Sous Mac avec Cyberduck

Open connexion : SFTP

Server : core.cluster.france-bioinformatique.fr

Username/Password : à compléter

Cluster de l'IFB

Tout le monde est au départ connecté sur le même noeud. Nous allons maintenant lancer une session interactive sur le cluster

Your turn ! Connectez vous sur un des noeuds

```
# Nous aurons besoins au cours du TP de ressources CPU et mémoire
$ sinteractive -J <userName>_TPVariant --cpus=4 --mem=16G

# Chargement de l'environnement dédié à l'atelier variant
$ module load conda
$ . activate eba2018_variant_calling_python3
```

Cluster de l'IFB

Rappels sur l'utilisation de session interactive

```
# Nous aurons besoins au cours du TP de ressources CPU et mémoire
$ sinteractive -J <userName>_TPVariant --cpus=4 --mem=16G
```

```
# Pour quitter sans terminer sa session
$ screen -d
# Récupération du noeud sur lequel tourne la session
$ squeue -u <userName>
```

```
# Pour réouvrir sa session le lendemain
$ ssh <nodeName>
$ screen -ls
$ screen -r <sessionName>
```

```
# Pour terminer définitivement la session
$ exit
```

Jeux de données #1 : SNVs/Indels

Depuis que l'homme fait de l'élevage, il essaie de faire en sorte de toujours améliorer sa **production, que ce soit en quantité ou en qualité**.

Les technologies de génotypage permettent maintenant de **sélectionner les mâles reproducteurs en fonction du fond génétique** qu'ils vont pouvoir transmettre à leur descendance.

Chez le bovin, il existe un locus de caractères quantitatifs (**QTL**) lié à la production de lait, situé sur le **chromosome 6**, et plus exactement sur une région de 700 kb, composée de 7 gènes.

Jeux de données #1 : SNVs/Indels

Les échantillons **QTL+** sont caractérisés par **une diminution de la production en lait** et une augmentation des concentrations en protéine et lipide.

Vous aurez à votre disposition :

- Un extrait des données de séquences d'un échantillon du projet 1000 génomes bovins, phénotypé comme **QTL-** : SRR1262731
- Les résultats du variant calling pour deux échantillons phénotypés **QTL+** : SRR1205992 et SRR1205973

Your turn !

Quelle mutation est responsable de ce QTL ?

Miri Cohen-Zinder et al, Genome Research 2005, DOI:10.1101/gr.3806705

Jeux de données #2 : SVs

Zymoseptoria tritici : Champignon ascomycète, pathogène du blé tendre, responsable d'une maladie foliaire (septoriose).

- Principale maladie du blé (jusqu'à 50% de perte de rendement).
- Haploïde, génome de 40 Mb séquencé en 2011 : 13 chromosomes essentiels + 8 chromosomes accessoires
- Souche séquencée avec **deux technologies : Illumina et Minlon**

Your turn !

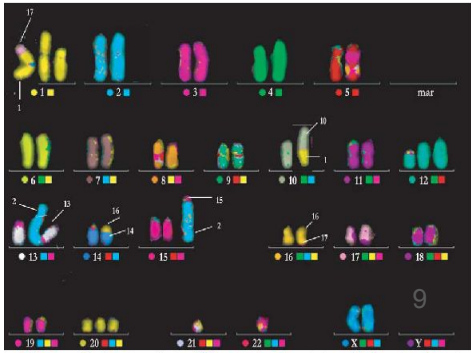
Retrouvez les délétions de grande taille

Jeux de données #3 : CNVs

Oncogenèse : perte de contrôle de la structure chromatinienne engendrant des **remaniements affectant le nombre de copies** de certaines régions du génome.

Ces altérations sont favorisées pendant l'évolution tumorale en fonction de leur impact sur la survie de la cellule tumorale (gain des facteurs/gènes à effet positif, perte de ceux à effet négatif).

Ces profils sont alors utiles à la caractérisation des tumeurs malignes, dans l'ensemble de leur pathologie (analyse descriptive) comme à titre individuel (médecine personnalisée).



Jeux de données #3 : CNVs

Vous aurez à votre disposition :

- Les séquences alignées (BAM) de la tumeur du sein et de la référence normale en WES d'une patiente du projet TCGA (The Cancer Genome Atlas), limitées aux chr11, chr17 et chr18.
- Les données complètes pré-processées de la même source.
- *Les données complètes pré-processées de trois autres patientes.*
- Les résultats des mêmes patientes analysées sur microarray (Affymetrix snp6.0)

Your turn !

Quelles sont les anomalies du nombre de copies dans ces données ?

Emplacement des données brutes

- Jeux de données #1 : SNVs/Indels

  → `/shared/home/mbernard/atelier_variant/tp_variant`

- Jeux de données #2 : SVs

  → `/shared/home/grobertsiegwald/cours/SV/data/`

- Jeux de données #3 : CNVs

  → /shared/home/bjob/TP_CNV/

Cheatsheet :
https://zerkalo.curie.fr/partage/tp_variant/itmo_variants_final.html
https://zerkalo.curie.fr/partage/tp_variant/itmo_variants_final.pdf