



Atelier Variant Introduction

Nadia BESSOLTANE - INRAE
Olivier RUÉ - INRAE
(Elodie GIRARD - Institut Curie)

Maria BERNARD - INRAE
Olivier Quenez - INSERM
Bastien JOB - Gustave Roussy

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 R., Nadia)
- Variations Structurales (SV) (Olivier Q.)
- Utilisation de R pour visualiser des métriques/résultats obtenus (Maria)
- Automatisation des traitements → Workflow (Maria)

Cluster de l'IFB



L'Institut Français 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

- **!!** Connexion initiale : tout le monde sur le noeud maître sur lequel il ne faut pas travailler **!!**
- Lancement de “jobs” ou d’une session interactive sur le cluster
- [Vidéo] : [The 5 minutes IFB Core Cluster tutorial](#)

Remember : Tous les jobs doivent être lancés sur un noeud du cluster !

```
# Chargement de l'environnement dédié à chaque outil (exemple pour varscan)
$ module avail -l | grep varscan
$ module load varscan/2.4.3          # ou module load varscan

# Nous aurons besoin au cours du TP de ressources CPU et mémoire (RAM)
$ sbatch --cpus=4 --mem=16G -J toolName_<user_name> --wrap="tool command line"

# Pour suivre vos “jobs” soumis sur le cluster, 2 solutions

$ squeue -u <user_name>                $ scontrol show job <job_id>
```

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 ?

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



Emplacement des données brutes

- Jeux de données #1 : SNVs/Indels

→ /shared/projects/ebai2020/atelier_variant/data/variants

- Jeux de données #2 : SVs

→ /shared/projects/ebai2020/atelier_variant/data/sv

Cheatsheet :

→ Version [html](#) :

/shared/projects/ebai2020/atelier_variant/cheatsheet_variant.html