Bio Comp

UF2.2 - Bioinformatique pour la génomique

Alignement de séquences et recherche de polymorphisme

Apprendre à traiter les séquences issues des NGS
pour la recherche de polymorphisme

Travail en Autonomie

Philippe Bardou - Cédric Cabau
Décembre 2024

Pour ce travail en autonomie nous allons rechercher les variants chez *Drosophila melanogaster* à partir de deux couples de FASTQ. L'ensemble des données d'entrée est disponible à cette adresse :
https://web-genobioinfo.toulouse.inrae.fr/~formation/16_SGS-SNP/Data_TA-BioComp/
Avec :

- Génome : GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna
- Jeu de variants connus : 7227_GCA_000001215.4_current_ids.vcf.gz + index
- Données FASTQ :
    - ERR10673070_R1.fastq.gz et ERR10673070_R2.fastq.gz
    - ERR10673071_R1.fastq.gz et ERR10673071_R2.fastq.gz

Les commandes doivent être exécutées sur un noeud de calcul et non directement sur le frontal genobioinfo (`srun --pty bash`).

Les étapes :

- Constituer 4 groupes
- Travail commun aux 4 groupes :
    - Ensemble des traitements jusqu'à la génération du BAM recalibré
- Travail par groupe :
    - Se répartir chacun de ces 4 callers : GATK HaplotypeCaller, Freebayes, Mpileup et DeepVariant
    - Produire un VCF annoté (snpEff et snpSift) : CALLER.vcf.gz
- Evaluation lors de la dernière séance (5 minutes par groupe) :
    - Une présentation en 2 à 3 slides avec :
        - Le répertoire de travail (attention aux droits en lecture)
        - Le taux de couverture pour chaque couple de FASTQ
        - Les métriques d'alignement (%mapped et %properly paired)
        - Le nombre de Read marquées comme duplicats
        - Le nombre total de variants
        - Le nombre de variants "nouveaux"
        - Les positions des 5 "meilleurs nouveaux" SNP qui ont un effet "HIGH"
        - Une capture d'IGV du "meilleur nouvel" SNP qui a un effet "HIGH"
        - Une capture d'IGV d'un SNP qui discrimine les deux échantillons
    - Envoi par mail de la présentation (avant le 2 décembre).

Philippe Bardou - Cédric Cabau
Décembre 2024

Notes autour de l'utilisation des callers non vus en cours/TPs :

```
## Freebayes https://github.com/freebayes/freebayes ##
module load bioinfo/freebayes/1.3.6

cat > 20_freebayes.jobs
freebayes -f GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna
--min-alternate-fraction 0.1 --min-mapping-quality 1 ERR10673070.md.recal.bam
| bgzip -c > ERR10673070.md.recal_freebayes.vcf.gz
freebayes -f GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna
--min-alternate-fraction 0.1 --min-mapping-quality 1 ERR10673071.md.recal.bam
| bgzip -c > ERR10673071.md.recal_freebayes.vcf.gz

sarray -J 20_freebayes -e LOGS/%x_%j.err -o LOGS/%x_%j.out --mem=15G
20_freebayes.jobs


## mpileup https://samtools.github.io/bcftools/bcftools.html#mpileup ##
module load bioinfo/Bcftools/1.17

cat > 30_mpileup.jobs
bcftools mpileup --fasta-ref
GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna
ERR10673070.md.recal.bam | bcftools call --output-type v --multiallelic-caller
| bcftools view --output-file ERR10673070.md.recal_mpileup.vcf.gz
--output-type z -i 'count(GT=="RR")==0'
bcftools mpileup --fasta-ref
GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna
ERR10673071.md.recal.bam | bcftools call --output-type v --multiallelic-caller
| bcftools view --output-file ERR10673071.md.recal_mpileup.vcf.gz
--output-type z -i 'count(GT=="RR")==0'

sarray -J 30_mpileup -e LOGS/%x_%j.err -o LOGS/%x_%j.out --mem=15G
30_mpileup.jobs
```

Philippe Bardou - Cédric Cabau
Décembre 2024

```
## deepvariant https://github.com/google/deepvariant ##
module load containers/singularity/3.9.9

singularity pull --name deepvariant-1.6.0.sif
docker://google/deepvariant:1.6.0

cat > 40_deepvariant.jobs
singularity exec --bind /work/project $PWD/deepvariant-1.6.0.sif
/opt/deepvariant/bin/run_deepvariant --model_type WGS --ref
$PWD/GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna --reads
$PWD/ERR10673070.md.recal.bam --output_vcf
$PWD/ERR10673070.md.recal_deepv.vcf.gz --output_gvcf
$PWD/ERR10673070.md.recal_deepv.g.vcf.gz --model_type WGS --num_shards 12
--intermediate_results_dir $PWD/DeepV_intermediate_70
singularity exec --bind /work/project $PWD/deepvariant-1.6.0.sif
/opt/deepvariant/bin/run_deepvariant --model_type WGS --ref
$PWD/GCF_000001215.4_Release_6_plus_ISO1_MT_genomic_Renamed.fna --reads
$PWD/ERR10673071.md.recal.bam --output_vcf
$PWD/ERR10673071.md.recal_deepv.vcf.gz --output_gvcf
$PWD/ERR10673071.md.recal_deepv.g.vcf.gz --model_type WGS --num_shards 12
--intermediate_results_dir $PWD/DeepV_intermediate_71

sarray -J 40_deepv -e LOGS/%x_%j.err -o LOGS/%x_%j.out --mem=50G
--cpus-per-task 12 2440_deepvariant.jobs
```

Notes pour créer un VCF contenant les variants de plusieurs échantillons :

```
search_module bcftools
module load bioinfo/Bcftools/1.17
bcftools merge ERR*.md.recal.vcf.gz -o CALLER.vcf.gz
```

Philippe Bardou - Cédric Cabau
Décembre 2024