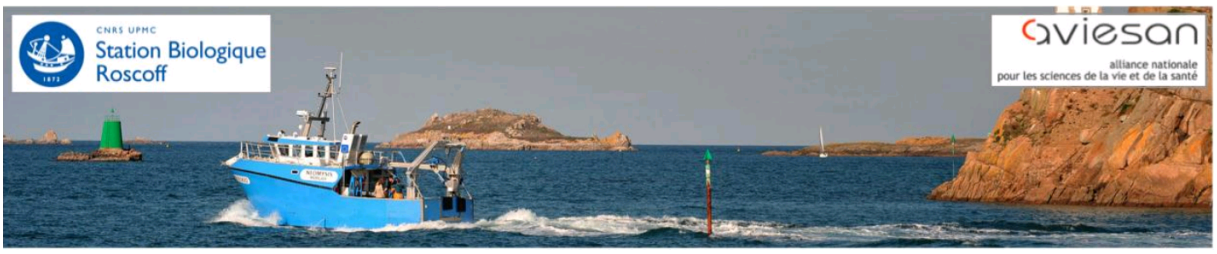
TP : Variant Calling avec GATK



Ecole de bioinformatique Roscoff 2014 : Initiation au traitement des données de génomique obtenues par séquençage à haut débit

Olivier Sand

1. Objectif

Découvrir les variants à partir de l'alignement généré aux TPs précédents (données WES Pickrell 2012) avec les variant callers de la suite GATK, Unified Genotyper et Haplotype Caller.

2. Fichiers nécessaires

- dataset correspondant au fichier bam d'alignement de vos reads sur le génome, réaligné et recalibré (TP précédent 7/10/2014)
- génome/séquence de référence : chr12.fa
- fichier vcf de référence : dbsnp137.hg19_chr12.vcf
- fichier bed de région cible : targeted_regions.bed

3. Variant Calling

a. Calling global avec Unified Genotyper

Utiliser l'outil **Unified Genotyper** (catégorie GATK2 Tools) pour effectuer le SNP/INDELs calling.

- Dans l'option **Choose the source for the reference list:**, choisir *History*
- Dans l'option **BAM files**, sélectionner le fichier BAM réaligné & recalibré généré au TP précédent
- Dans l'option **Using reference file**, sélectionner le fichier *chr12.fa*
- Dans l'option **Provide a dbSNP Reference-Ordered Data (ROD) file**, sélectionner *Set dbSNP*
- Dans l'option **dbSNP ROD file**, sélectionner le fichier *dbsnp137.hg19_chr12.vcf*
- Dans l'option **Genotype likelihoods calculation model to employ:**, sélectionner *SNP*
- Laisser le reste inchangé
- Lancer l'exécution

Vous venez de lancer le SNP calling, relancer maintenant le traitement précédent en modifiant l'option **Genotype likelihoods calculation model to employ** à *INDEL* pour effectuer le calling des INDELs

Remarque : Vous pouvez effectuer le SNP calling et l'INDEL calling simultanément en sélectionnant le **Genotype likelihoods calculation model** à *BOTH*

Pour une documentation complète sur l'Unified Genotyper, voir https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php

Analyse des résultats :

- Utiliser le **View data** (oeil) pour ouvrir les fichiers de metrics que vous venez de créer
  - ⇨ Combien de SNPs avez-vous obtenus ? Combien d'INDELs ?

b. Calling sur une région cible avec Unified Genotyper

Le but ici est de reproduire l'étape précédente sur une zone d'intérêt.

Utiliser l'outil **Unified Genotyper** (catégorie GATK2 Tools) pour effectuer le SNP/INDELs calling.

- Dans l'option **Choose the source for the reference list:**, choisir *History*
- Dans l'option **BAM files**, sélectionner le fichier BAM réaligné & recalibré généré au TP précédent
- Dans l'option **Using reference file**, sélectionner le fichier *chr12.fa*
- Dans l'option **Provide a dbSNP Reference-Ordered Data (ROD) file**, sélectionner *Set dbSNP*
- Dans l'option **dbSNP ROD file**, sélectionner le fichier *dbsnp137.hg19_chr12.vcf*
- Dans l'option **Genotype likelihoods calculation model to employ:**, sélectionner *SNP*
- Sélectionner *Advanced* dans le menu **Basic or Advanced GATK options** pour faire apparaitre des options supplémentaires.
- Dans l'option **Operate on Genomic intervals**, cliquer *Add new Operate on Genomic intervals* et sélectionner le fichier d'intervalles *targeted_regions.bed* dans le menu qui apparait
- Laisser le reste inchangé
- Lancer l'exécution

Vous venez de lancer le SNP calling, relancer maintenant le traitement précédent en modifiant l'option **Genotype likelihoods calculation model to employ** à *INDEL* pour effectuer le calling des INDELs

Vous venez de lancer l'INDEL calling, relancer maintenant le traitement précédent en modifiant l'option **Genotype likelihoods calculation model to employ** à *BOTH* pour effectuer le calling des SNPs et des INDELs ensembles

Renommer le vcf de sortie en pe_gatk_chr12_targeted.vcf

Remarque : Tous les outils de GATK peuvent être utilisés sur un intervalle particulier en utilisant cette option.

Analyse des résultats :

- Utiliser le **View data** (oeil) pour ouvrir les fichiers de metrics que vous venez de créer
  - ⇨ Combien de SNPs avez-vous obtenus ? Combien d'I?NDELs ?
- Utiliser le **View data** pour ouvrir les fichiers VCF que vous venez de créer.
  - ⇨ Essayez de regarder dans IGV quelques variants et de visualiser la signification des différents tags du VCF (GT:AD:DP:GQ:PL)

c. Calling sur une région cible avec Haplotype Caller

Utiliser l'outil **Haplotype Caller** (catégorie GATK2 Tools) pour effectuer le SNP/INDELS calling.

- Dans l'option **Covariates table recalibration file**, choisir le fichier *Base Recalibrator... (Covariate file)* généré au TP précédent
- Dans l'option **Choose the source for the reference list:**, choisir *History*
- Dans l'option **BAM files**, sélectionner le fichier BAM réaligné & recalibré généré préalablement
- Dans l'option **Using reference file**, sélectionner le fichier *chr12.fa*
- Dans l'option **Provide a dbSNP Reference-Ordered Data (ROD) file**, sélectionner *Set dbSNP*
- Dans l'option **dbSNP ROD file**, sélectionner le fichier *dbsnp137.hg19_chr12.vcf*
- Sélectionner *Advanced* dans le menu **Basic or Advanced GATK options** pour faire apparaitre des options supplémentaires.
- Dans l'option **Operate on Genomic intervals**, cliquer *Add new Operate on Genomic intervals* et sélectionner le fichier d'intervalles *targeted_regions.bed* dans le menu qui apparait
- Laisser le reste inchangé
- Lancer l'exécution

Pour une documentation complète sur l'Haplotype Caller, voir https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php

Analyse des résultats :

- Cliquer le nom du dataset obtenu pour lire les détails
    - ⇨ Combien de variants avez-vous obtenus ? Est-ce différent du nombre obtenu avec Unified Genotyper (option *BOTH*) ?
- Utiliser le **View data** (oeil) pour ouvrir le fichier VCF que vous venez de créer
    - ⇨ Y a-t-il plus d'INDELs découverts avec HC ?
    - ⇨ Regarder dans IGV le fichier Bam de départ aux positions de quelques variants et évaluer leur fiabilité (ex: positions 6777069 et 55038408)