TP : Preprocessing et Variant Calling avec GATK

Ecole de bioinformatique Roscoff 2013 : Initiation au traitement des données de génomique obtenues par séquençage à haut débit

Sylvain Marthey

1. Import des données

- Importer dans votre historique le dataset correspondant au fichier bam d'alignement de vos reads sur le génome
- Importer votre génome/séquence de référence
- Importer les fichiers vcf de référence suivants :
    - Fichier VCF contenant l'ensemble des variants connus
    - Fichier VCF les indels connus

2. Preprocessing

a. Retrait des duplicats PCR

Utiliser l'outil **Mark Duplicate Reads** (catégorie NGS : Picard (beta)) pour éliminer les duplicats présents dans votre fichier d'alignement bam.

- cocher l'option '**Remove duplicates from output file**'
- cocher l'option '**Assume reads are already ordered'**

Analyse du résultat :

- Quel est le niveau de duplication de votre banque ?
- Combien de reads vous reste-t-il ?
- Quel est maintenant le niveau de couverture moyen de votre génome ?

b. Ajout des reads groups

Utiliser l'outil **Add or Replace Groups** (catégorie NGS : Picard (beta)) pour ajouter les reads groups correspondant à votre librairie.

c. Réalignement autour des indels

1. Détection des zones à réaligner

Utiliser l'outil **Realigner Target Creator** (catégorie GATK Tools (beta)) pour détecter les régions sur lesquelles seront effectués les réalignements des reads.

- Ajouter le fichier VCF contenant les indels de référence dans l'option '**Binding for reference-ordered datas**'
    - Selectionner '**Binding Type'** : INDELs

Aide : Il se peut que votre génome/séquence de référence n'apparaisse pas dans la liste des fichiers utilisable dans l'option **Using reference file/genome**.

Dans ce cas, assurez-vous que l'attribut **Database/Build** du fichier BAM que vous souhaitez utiliser a bien la valeur unspecified (?), car ceci permet d'utiliser comme référence les fichiers fasta importés dans votre historique.

Remarque : le fait de modifier mettre la valeur unspecified à l'attribut Database d'un fichier bam inactive dans Galaxy la fonction permettant de le visualiser directement dans l'UCSC ou IGV. Penser à remodifier cet attribut pour réactiver ces options.

Analyse des résultats :

- Utiliser le 'dataset display' pour ouvrir le fichier d'intervals que vous venez de créer
  - ⇨ Quel est la nomenclature utilisée pour décrire les intervals ?
- Utiliser IGV pour visualiser les alignements de vos reads sur quelque un de ces intervals :
  chr12:1016878-1016882
  chr12:1041953-1042147
  chr12:2760951-2760966

2. Réalignement des lectures

Utiliser l'outil **Realigner Target Creator** (catégorie GATK Tools (beta)) pour détecter les régions sur lesquelles seront effectués les réalignements des reads.

- Dans l'option '**Restrict realignment to provided intervals** ', sélectionner le fichier d'intervals que vous venez de créer à l'étape précédente
- Ajouter le fichier VCF contenant les indels de référence dans l'option '**Binding for reference-ordered datas**'
  - Selectionner '**Binding Type'** : INDELs

Remarque : Il est très important ici d'utiliser les mêmes fichiers BAM et d'indels que ceux utilisés à l'étape **Realigner Target Creator**

Analyse des résultats :

- Utiliser IGV pour visualiser les alignements de vos reads sur quelque un des intervals de réalignement sur lesquels les reads ont été réalignés.
  - ⇨ Observez-vous des différences ?

d. Recalibration

1. Calcul des covariables

Utiliser l'outil **Count Covariates** (catégorie GATK Tools (beta)) pour calculer les tables de covariation.

- Cocher l'option '**Use the standard set of covariates in addition to the ones selected**'
- Dans l'option '**Covariates to be used in the recalibration',** cocher les covariables suivantes : ReadGroupCovariate, QualityScoreCovariate, CycleCovariate, DinucCovariate, HomopolymerCovariate, GCContentCovariate
- Ajouter le fichier VCF contenant les variants de référence dans l'option '**Binding for reference-ordered datas**'
  - Selectionner '**Binding Type'** : dbsnp

2. Analyse des covariables avant recalibration

Utiliser l'outil **Analyze Covariates** (catégorie GATK Tools (beta)) pour créer des graphiques permettant d'analyser les covariables.

Analyse des résultats :

- Utiliser le 'dataset display' pour ouvrir le fichier HTML que vous venez de créer
  - Télécharger et ouvrir les pdfs 'CycleCovariate*'
    - ⇨ Observez-vous une différence entre les qualités reportées et empiriques ?
    - ⇨ Actuellement les qualités par cycle sont elles surévaluées ou sous-évaluées ?
  - Télécharger et ouvrir les pdfs 'DinucCovariate*'
    - ⇨ La balance Dinucléotidique est elle homogène dans les données ?
  - Télécharger et ouvrir les pdfs 'QualityScore*'
    - ⇨ Les échelles de score reportés et empiriques sont elles similaires ?

3. Recalibration

Utiliser l'outil **Table Recalibration** (catégorie GATK Tools (beta)) pour recalibrer les fichiers d'alignement.

- Dans l'option '**Covariates table recalibration file**', sélectionner le fichier de covariables que vous venez de créer à l'étape précédente
- Dans l'option '**BAM file**', sélectionner le fichier BAM réaligné

4. Analyse des covariables après recalibration

Utiliser l'outil **Analyze Covariates** (catégorie GATK Tools (beta)) pour créer des graphiques permettant d'analyser les covariables.

Analyse des résultats :

⇨ Comparer ces nouveaux graphiques avec ceux observés précédemment. La recalibration les a-t-elle modifiés ?

3. Variant Calling

a. Calling global

Utiliser l'outil **Unified Genotyper** (catégorie GATK Tools (beta)) pour effectuer le SNP/INDELS calling..

- Dans l'option '**BAM files** ', sélectionner le fichier de réaligné & recalibré généré à l'étape précédente.
- Ajouter le fichier VCF contenant les indels de référence dans l'option '**Binding for reference-ordered datas**'
  - Selectionner '**Binding Type'** : dbSNP

- Dans l'option '**Genotype likelihoods calculation model to employ:** ', sélectionner **SNP**

Vous venez de lancer le SNP calling, relancer maintenant le traitement précedent 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 en simultanément en sélectionnant le Genotype likelihoods calculation model **BOTH**

Analyse des résultats :

- Utiliser le 'dataset display' 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

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

1. Créer un fichier au format 'interval' de GATK contenant les coordonnées de la région sur laquelle sera effectué le variant calling.

Aide : Utiliser la nomenclature vue lors de la création des régions à réalignées.

2. Utiliser l'outil **Unified Genotyper** (catégorie GATK Tools (beta)) pour effectuer le SNP/INDELS calling..

- Cliquer sur le boutton '**Basic or Advanced GATK options**' pour faire apparaitre des options supplémentaires.
- Dans l'option '**Operate on Genomic intervals'** , sélectionner le fichierd 'intervals que vous venez de créer.

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

Analyse des résultats :

- Utiliser le 'dataset display' pour ouvrir les fichiers de metrics que vous venez de créer
  - ⇨ Combien de SNPS avez-vous obtenus ? Combien d'Indels ?
- Utiliser le 'dataset display' pour ouvrir les fichiers VCF que vous venez de créer.
  - ⇨ Essayez de regarder dans IGV quelque variants et de visualiser la signification des différents TAG du VCF (GT:AD:DP:GQ:PL)

4. Workflow

Créer le workflow qui effectuera le realignement et la recalibration d'un fichier BAM.

Ce workflow prendra entrée les fichiers suivants :

- Génome de référence
- Fichier BAM
- Fichier VCF d'INDELs
- Fichiers VCF contenant tous les variants

Il devra enchainer les outils suivants :

Realigner Target Creator -> Indel Realigner -> Count Covariates -> Table Recalibration