École de bioinformatique - Roscoff – Novembre 2016

Atelier Détection de Variants

TP Filtrage de Variants

Nicolas Lapalu

Objectif

Filtrer les variants générés par les TP précédents, sur les données Exome-seq de la publication Pickrell 2012, avec la suite GATK. Comparer les variants obtenus avec une autre méthode de variant calling (Samtools Mpileup / Varscan) sur les données RNA-seq de Pickrell 2012. (Données fournies non générées par les TP précédents).

Fichiers de départ

Vous allez trouver ci-dessous, la liste des fichiers nécessaires à ce TP. Vous avez soit déjà importé ou généré l'ensemble de ces fichiers dans les TPs précédents, à l'exception du fichier **varscan_chr12.vcf**. Vous trouverez ci-dessous le lien web pour récupérer ce fichier, ainsi que les autres fichiers nécessaires dans le cas où vous les auriez supprimés.

Solution 1 : importez tous les fichiers de l'historique :

http://sigenae-workbench.toulouse.inra.fr/galaxy/u/nlapalu/h/tp-variant-eba2016---filtering

Solution 2: récupérez les fichiers manquants :

- varscan_chr12.vcf (à récupérer - dispo ici)
- gatk_chr12_targeted.vcf (généré au TP précèdent – dispo ici)
- dbsnp137.hg19_chr12.vcf (importé au début du TP – dispo ici)
- chr12.fa (importé au début du TP – dispo ici)

Question : Combien de variants ont été détectés pour les données exome GATK (gatk_chr12_targeted.vcf) et Varscan (varscan_chr12.vcf)?

Réponse :

Filtrage des variants exome issus de GATK

Nous allons réaliser trois filtres de façon simultanée. Les trois filtres seront appliqués sur les données initiales ce qui permettra de visualiser l'action séparée de chaque filtre (colonne FILTRE remplie avec potentiellement plusieurs filtres). Nous allons filtrer sur la qualité du variant (colonne QUAL), sur la qualité de mapping et la profondeur du variant (colonne INFO, Tags MQ et DP).

Tool : GATK3 – Variant Filtration on VCF files

Choose the source for the reference list: **History**

Using reference file: **chr12.fa**
Select interval subset to operate on? **No**
Analysis Type: **VariantFiltration**
Variant files: **gatk_chr12_targeted.vcf**
Configure Optional Parameters : **Yes**
One or more expressions used with INFO fields to filter.
QUAL<40.0;MQ<30.0;DP<20
Names to use for the list of filters.
VariantQuality;MappingQuality;Cov20

Renommez le fichier vcf de sortie **: gatk_chr12_filtered.vcf**

Question : Le tag DP est aussi présent dans la colonne FORMAT. Comment faire pour que le filtre s'applique au niveau de l'échantillon (FORMAT) plutôt qu'au niveau du variant (INFO) ?

Réponse :

Nous souhaitons maintenant supprimer les variants qui ont été filtrés pour ne garder que ceux qui nous intéressent.

Tool: GATK3 – Select Variant from VCF files

Choose the source for the reference list: **History**
Using reference file: **chr12.fa**
Select interval subset to operate on? **No**
Analysis Type: **SelectVariants**
Variant files: **gatk_chr12_filtered.vcf**
Configure Optional Parameters : **Yes**
Don't include filtered sites : **Yes**

Renommez le fichier vcf de sortie **: gatk_chr12_filtered_only.vcf**

Question : Combien de variants reste-il ?

Réponse :

Filtrage des variants RNA-Seq issus de VARSCAN

Les données de variants issues de VARSCAN ont été générées indépendamment des TP précédents. Elles sont issues d'un variant calling fait par l'outil samtools Mpileup / Varscan avec un mapping préalable par TopHat.

Question : Quels sont les différences que vous pouvez potentiellement observées entre des données Exome-Seq et RNA-Seq. (cf : Chee-Seng Ku, 2012, Expert Reviews)

Réponse :

Les Tags présents dans le fichier VCF Varscan sont différents de GATK. On se propose ici de filtrer sur le tag ADP.

Question : A quoi correspond le tag ADP?

Réponse :

Il est possible d'utiliser directement l'outil Select Variant pour filtrer les données. Cela peut permettre de limiter la taille du fichier de variants obtenu en le limitant aux données désirées.

Importez le fichier **varscan_chr12.vcf** dans votre historique. Editez les attributs du fichier et associez le à la database '**human – hg19**'.

Tool: GATK3 – Select Variant from VCF files

Choose the source for the reference list: **History**
Using reference file: **chr12.fa**
Select interval subset to operate on? **No**
Analysis Type: **SelectVariants**
Variant files: **varscan_chr12.vcf**
Configure Optional Parameters : **Yes**
One or more criteria to use when selecting the data…
ADP>30
Don't include filtered sites : **Yes**

Renommez le fichier vcf de sortie **: varscan_chr12_filtered_only.vcf**

Question : Combien de variants reste-il ?

Réponse :

Comparaison des résultats GATK Exome-Seq vs VARSCAN RNA-Seq

Il est possible de pouvoir identifier les variants détectés spécifiquement par une ou l'autre méthode. Pour cela utilisez l'outil CombineVariants. Nous allons en profiter pour combiner nos résultats avec la base dbSNP. Nous pourrons alors identifier les variants détectés connus ou inconnus de la base.

Tool: GATK3 – CombineVariants

Choose the source for the reference list: **History**
Using reference file: **chr12.fa**
Select interval subset to operate on? **No**
Analysis Type: **CombineVariants**
Variant files (VCF format):
- gatk_chr12_filtered_only.vcf
- varscan_chr12_filtered_only.vcf

- dbsnp137.hg19_chr12.vcf

Renommez le fichier vcf de sortie **: combined_chr12.vcf**

La colonne INFO du nouveau fichier contient un nouveau tag « **set** » avec la valeur **Intersection**, **variant**,**variant2,variant3,variant-variant2,variant-variant3,variant2-variant3**. Il est donc possible de filtrer sur ce nouveau tag pour ne garder que les résultats d'une des 2 méthodes. Pour savoir à quel fichier correspond chaque valeur variant, variant1, variant2, regardez la ligne de commande dans les commentaires.

Nous allons maintenant sélectionnés seulement les variants découverts à la fois par GATK et VARSCAN, et présents dans dbSNP :

Tool: GATK3 – Select Variant from VCF files

Choose the source for the reference list: **History**
Using reference file: **chr12.fa**
Select interval subset to operate on? **No**
Analysis Type: **SelectVariants**
Variant files: **combined_chr12.vcf**
Configure Optional Parameters : **Yes**
One or more criteria to use when selecting the data…
set=='Intersection'
Don't include filtered sites : **Yes**

Vous obtenez maintenant un fichier avec des informations renseignées dans la 3eme colonne du VCF (ID) avec les identifiants connus dans dbSNP.

Question : Combien de variants sont retrouvés par dbSNP, GATK et varscan ? Comment feriez-vous pour récupérer les variants spécifiques de GATK, VARSCAN, GATK+dbSNP

Réponse :

Nous allons tester un autre filtre pour mettre en évidence des variants qui montrent un biais de brin (SNP trouvé sur un seul brin), dû à une séquence particulière en amont (cf Meacham (2011) *BMC Bioinformatics*). Pour cela nous pouvons filtrer sur le tag FS (p-value du test de Fisher en Phred-score).

Tool: GATK3 – Select Variant from VCF files

Choose the source for the reference list: **History**
Using reference file: **chr12.fa**
Select interval subset to operate on? **No**
Analysis Type: **SelectVariants**

Variant files: **combined_chr12.vcf**
Configure Optional Parameters : **Yes**
One or more criteria to use when selecting the data…
FS>60.0
Don't include filtered sites : **Yes**

Regardez le variant à la position **chr12-4461695** et connu dans dbSNP avec l'identifiant **rs75781974**. Dans la colonne INFO, ce variant obtient un score de FS (p-value du test de Fisher en Phred-score) supérieur à 60, soit p=0,000001. A la vue du résultat obtenu pour ce variant, que concluez-vous ? Vous pouvez consulter la fiche de ce variant sur http://www.ncbi.nlm.nih.gov/snp?term=rs75781974 , cela conforte-t-il votre hypothèse ? Vous pouvez aussi ouvrir IGV à la position du variant et regarder les résultats de mapping des TPs précédents.

Quelques conseils pour vos analyses

Si vous n'êtes pas limités par la taille de l'espace disk de vos machines, préférez le format gVCF qui peut vous donner aussi l'information de profondeur de couverture pour l'ensemble des bases du génome de référence. Trop souvent l'absence d'un polymorphisme à une base donnée est interprétée comme une homologie avec la séquence de référence. Dans certains cas cela peut être faux et simplement dû à une trop faible profondeur de couverture pour réaliser le 'calling' du snp. Le format gVCF vous permettra aussi de pouvoir facilement rajouter des échantillons par la suite.

Filtrez aussi l'ensemble des positions qui peuvent poser problème : éléments répétés-transposable, zones de faible complexité. Si vous travaillez sur des séquençages de génome entier, définissez un intervalle de couverture de base autour de la couverture moyenne observée (recommandé 2x écart-type ou 4x racine carrée de la moyenne). En fonction de la finalité de vos analyses GWAS, cartographie, génomique des populations, etc…, il n'est pas nécessaire de garder tous les variants. Il est souvent préférable de ne conserver des variants bien filtrés pour éviter d'introduire des faux-positifs.