TP annotation de variants constitutionnels

Ghislain Bidaut, Plateforme Cibi, CRCM, Aix-Marseille Université

24/03/2022

1 Introduction

Lors de ce TP, nous allons poursuivre l'analyse des variants détectés lors du TD précédent sur les données générées par Nikitin *et al.*: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273971/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7273971/)).

Nous allons maintenant utiliser les **bases de données d'annotations publiques** de variants pour les **annoter**, ce qui permettra de les interpréter.

La collection de l'information d'annotation biologique connue dans les bases publiques se fait grâace à des programmes spécialisés. Nous allons en explorer 2 dans ce TD: **snpEff** et **Annovar**.

Pour rappel, les variants ont été stockés dans un fichier au format gVCF disponible sur le cluster IFB:

```
/shared/projects/form_2022_09/data/DNA-seq_PRJNA588789_Variants/all.sample.filtered.vcf.gz
```

Le fichier est également disponible sur MyCore:

https://mycore.core-cloud.net/index.php/apps/files/?dir=/partage_data_formation_bioinfo_ngs/TP_Annotation_VCF&fileid=2340766861 (https://mycore.core-cloud.net/index.php/apps/files/?dir=/partage_data_formation_bioinfo_ngs/TP_Annotation_VCF&fileid=2340766861)

Nous allons travailler sur le cluster IFB pour des raisons de simplicité. J'ai déjà téléchargé les **bases de données** pour l'utilisation des programmes d'annotation **snpEff** et **Annovar**.

Je propose de travailler dans l'arborescence suivante:

```
/shared/home/<username>/formation_amu/DNA-seq_PRJNA588789_VariantAnnot
```

1.1 Rappel sur les données

Nikitin *et al.* ont publié une étude portant sur le séquençage ciblé des gènes du pathways **MMR** (MLH1, MSH2, MSH6, EPCAM, et PMS2) sur 711 patients atteint d'un cancer du sein héréditaire et 60 atteint d'un cancer du sein sporadique.

Le syndrome de Lynch est provoqué par des mutations constitutionelles dans les gènes du pathways MMR et peux provoques de nombreux types de cancers, donc le cancer du sein.

1.2 Analyse du fchier VCF avant annotation

1.2.1 TD:

1. Récupérer le fichier `all.sample.filtered.vcf.gz` dans un répertoire de travail et le décompresser.

Examiner le fichier avec Excel. Que peut-on dire ?

2. Utiliser l'utilitaire `vt` sur ce fichier. Faire un `vt peek` .

Que peut-on dire ?

3. Utiliser l'utilitaire `bcftools`

- Imprimer l'entête du fichier VCF
- Faire des statistiques sur les variants. Puis en tirer des conclusions.
- Extraire un fichier VCF pour l'échantillon `SRR10426968`

4. Utiliser `bcftools` pour extraire un VCF par échantillon.
5. Que peut-on dire du variant `chr2` position `47416318` ?
6. Donnez un exemple d'INDELS parmi les mutations listées.

1.2.2 Solution

1.3 Visualisation

TD:

Visualiser le fichier VCF multi-échantillons conjointement avec les fichiers BAM grâce au programme IGV.

1.4 Annotation

L'annotation consiste à collecter des annotation biologique à partir des bases publiques pour les variants que nous avons détectés.

Pour cela, nous allons utiliser le programme **SnpEff** dans une première partie de TP, puis **Annovar** pour la seconde moitié du TP.

1.5 Annotation des variants obtenus par SnpEff

SnpEff est un programme permettant de prédire l'effet d'un variant sur les gènes ou protéines (changement dans les acides aminés par exemple).

- Il supporte 38 000+ génomes
- Il supporte le format standard d'annotation **ANN**
- Notation **HGVS**
- Support de Sequence Ontology
- Compatible **GATK** en natif.

1.5.1 TP

1. Parcourir la documentation de SnpEff : http://pcingola.github.io/SnpEff/ (http://pcingola.github.io/SnpEff/) . Installer la base de données d'annotation pour l'humain `hg38` .

2. Annoter notre fichier VCF de manière à obtenir un nouveau fichier annoté.
3. Analyse du rapport **HTML** généré par **snpEff**.

Récupérer le rapport HTML généré par SnpEff **en local** (Commande `scp`) puis interprétez le.

Analyser le rapport et expliquez les différents éléments qui le constituent.

1.5.2 Solution:

1.6 Annotation de variants par Annovar

En prenant en entrée une liste de variants associés à leur position chromosomiques, Annovar peut, **pour le génome spécifié**:

- Sortir l'information liée aux gènes concernés. Exemple: Est ce que le variant provoque un changement dans le codage de la protéine ?
- Identifier l'information documentée dans des bases de données, spécifiques, par exemple **dbSNP**, **The 1000 genome project** ou **gnomAD**. La liste des bases de données est disponible ici: https://annovar.openbioinformatics.org/en/latest/user-guide/filter/ (https://annovar.openbioinformatics.org/en/latest/user-guide/filter/)

Annovar n'est **pas directement disponible** dans conda, il faut l'installer à partir du site https://annovar.openbioinformatics.org/en/latest/ (https://annovar.openbioinformatics.org/en/latest/)

Lien de téléchargemment:
http://www.openbioinformatics.org/annovar/download/0wgxR2rIVP/annovar.latest.tar.gz (http://www.openbioinformatics.org/annovar/download/0wgxR2rIVP/annovar.latest.tar.gz)

1.6.1 TD

Pour des facilités d'analyse, j'ai installé Annovar dans le répertoire partagé `/shared/projects/form-2021-018/data/annovar`.

Il est installé avec les bases de données suivantes:

- refGene
- ExAC version 3
- dbSNP version 147
- 1000 Genomes Project Aug 2015
- GnomAD

Comme précédemment, nous travaillons avec le fichier VCF : `/shared/projects/form_2022_09/data/DNA-seq_PRJNA588789_Variants/all.sample.filtered.vcf.gz`.

1. Parcourir la documentation d'Annovar: https://annovar.openbioinformatics.org/en/latest/ (https://annovar.openbioinformatics.org/en/latest/). Notamment, comment faire une conversion de notre fichier VCF multi-échantillons vers un format compatible ?
2. Faire un tour des bases de données, comprendre leur contenu et la manière d'y accéder. Faire un petit tableau récapitulatif.
3. Formatter un appel à Annovar dans un script bash pour interroger **refGene, ExAC3, dbSNP147,1000 Genomes Project Aug 2015 et GnomAD**.
4. Importer le fichier de variants annoté avec Annovar dans Excel, Que peut-on dire du variant positionné `chr2` position `47416318` ?

- Sur quel type de région génomique est localisé ce variant ?
- Sur quel gène est localisé ce variant ?
- Quelle type de mutation ce variant provoque t-il dans la protéine correspondante ?

- Est-ce que que ce variant est nouveau, rare ou commun dans la population générale ?
- Quel est l'impact prédit de ce variant ?
- Quel est l'identifiant dnSNP de ce variant ? L'utiliser pour voir s'il existe des connaissances cliniques sur ce variant (base de données Clinvar).

5. Quelle base de données pourrait nous donner des résultats supplémentaires intéressants si on était dans une analyse tumorale ?

1.6.2 Solution

1.7 Annexe:

1.7.1 Installation d'annovar

Attention: il faut préciser que l'on est en HG38 lors de l'installation des bases de données.

```
wget http://www.openbioinformatics.org/annovar/download/0wgxR2rIVP/annovar.latest.tar.g
z
tar xzf annovar.latest.tar.gz
cd annovar
./annotate_variation.pl -buildver hg38 -downdb -webfrom annovar refGene humandb/
./annotate_variation.pl -buildver hg38 -downdb -webfrom annovar 1000g2015aug humandb/
./annotate_variation.pl -buildver hg38 -downdb -webfrom annovar exac03  humandb/
./annotate_variation.pl -buildver hg38 -downdb -webfrom annovar avsnp147  humandb/
./annotate_variation.pl -buildver hg38 -downdb -webfrom annovar gnomad_genome humandb/
./annotate_variation.pl -buildver hg38 -downdb -webfrom annovar ljb26_all  humandb/
```

Voici un exemple d'utilisation avec les bases de donneés `exac03` et `avsnp` , et l'annotation par RefSeq.:

```
/shared/projects/form-2021-018/data/annovar/table_annovar.pl SRR10426968.trim.target.vc
f /shared/projects/form-2021-018/data/annovar/humandb/
-buildver hg38 -out SRR10426968 -remove -protocol refGene,exac03,avsnp147
-operation gx,f,f -nastring . -csvout -polish -xref /shared/projects/form-2021-018/dat
a/annovar/example/gene_fullxref.txt
```

1.7.2 Exemple d'utilisation d'Annovar

Arguments de la ligne de commande:

- `<chemin>/humandb/` : répertoire contenant les bases de données
- `-buildver hg38` : version du génome
- `-out SRR10426968` : nom de base des fichiers de sortie
- `-remove` : supprime les fichiers temporaires générés lors du téléchargement
- `-protocol refGene,exac03,avsnp147` : protocoles: bases de données à interroger
- `-operation gx,f,f` : opération à faire pour chacun des protocoles (g=gene, r=region, gx=gène avec référence croisée sur le fichier donné en `-xreffile`)
- `-xreffile` : fichier d'annotation de gènes