H3ABioNet

Pan African Bioinformatics Network for H3Africa

WELLCOME GENOME CAMPUS

CONNECTING SCIENCE

ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Next Generation Sequencing Bioinformatics Course 2021

Pathogen Variant Calling

Variant annotation

H3ABioNet

Pan African Bioinformatics Network for H3Africa

WELLCOME GENOME CAMPUS

CONNECTING SCIENCE

ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Overview

Investigating resistance in Mycobacterium tuberculosis

- What mutations are causing the resistance phenotype?
- Using SnpEff
  - Using databases
  - Interpreting the output
  - Filtering based on your research goals

SnpEff

- Provides information on the effects of variants
  - Within gene
    - Mutation effect
      - Silent, missense, nonsense
  - Up / downstream effects
    - Relevant for regulation (TF binding sites)

SnpEff

- Using existing databases
  - Need to use the same assembly that matches the annotation
  - Find using grep & SnpEff database command

```
~/iCloud/programs — -bash                                                    ~/Downloads
[snpEff$ java -jar snpEff.jar databases | grep tuberculosis | grep h37rv
Mycobacterium_tuberculosis_h37rv                        Mycobacterium_tuberculosis_h37rv
ttps://snpeff.blob.core.windows.net/databases/v5_0/snpEff_v5_0_Mycobacterium_tuberculosis_h37rv.zip
Mycobacterium_tuberculosis_h37rv_gca_000277735          Mycobacterium_tuberculosis_h37rv_gca_000277735
ttps://snpeff.blob.core.windows.net/databases/v5_0/snpEff_v5_0_Mycobacterium_tuberculosis_h37rv_gca_000277735.zip
Mycobacterium_tuberculosis_h37rv_gca_000667805          Mycobacterium_tuberculosis_h37rv_gca_000667805
ttps://snpeff.blob.core.windows.net/databases/v5_0/snpEff_v5_0_Mycobacterium_tuberculosis_h37rv_gca_000667805.zip
Mycobacterium_tuberculosis_h37rv_gca_000831245          Mycobacterium_tuberculosis_h37rv_gca_000831245
ttps://snpeff.blob.core.windows.net/databases/v5_0/snpEff_v5_0_Mycobacterium_tuberculosis_h37rv_gca_000831245.zip
Mycobacterium_tuberculosis_h37rvsiena                   Mycobacterium_tuberculosis_h37rvsiena
ttps://snpeff.blob.core.windows.net/databases/v5_0/snpEff_v5_0_Mycobacterium_tuberculosis_h37rvsiena.zip
snpEff$
```

SnpEff

- Building your own database
  - Use custom annotations
  - Useful when assembling genomes
  - Annotations not found in the database
  - From the creators of SnpEff:
    - “Most people do NOT need to build a database, and can safely use a pre-built one. So unless you are working with an rare genome you most likely don't need to do it either.”

SnpEff

- How?
  - Create folder in SnpEff directory for the genome (for example newBacGenome)
  - Move the fasta reference into the new folder and rename it "sequences.fa"
  - Move annotation file (GFF) to the folder and rename it "genes.gff"
  - Add "newBacGenome.genome: newBacGenome" to the bottom of the snpEff.config file where "newBacGenome" is the folder name
- https://pcingola.github.io/SnpEff/se_buildingdb/

SnpEff

- Output of snpEff annotation:

```
# Here is how the output looks like
$ head examples/test.chr22.ann.vcf
##SnpEffVersion="4.1 (build 2015-01-07), by Pablo Cingolani"
##SnpEffCmd="SnpEff  GRCh37.75 examples/test.chr22.vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele |
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effect
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
22  17071756    .   T   C   .   .   ANN=C|3_prime_UTR_variant|MODIFIER|CCT8L2|ENSG
22  17072035    .   C   T   .   .   ANN=T|missense_variant|MODERATE|CCT8L2|ENSG000
22  17072258    .   C   A   .   .   ANN=A|missense_variant|MODERATE|CCT8L2|ENSG000
22  17072674    .   G   A   .   .   ANN=A|missense_variant|MODERATE|CCT8L2|ENSG000
```

SnpEff

- Output vcf ANN entry:
  - ANN=T|missense_variant|MODERATE|CCT8L2|ENSG00000198445|transcript|ENST00000359963|protein_coding|1/1|c.1406G>A|p.Gly469Glu|1666/2034|1406/1674|469/557||,T|downstream_gene_variant|MODIFIER|FABP5P11|ENSG00000240122|transcript|ENST00000430910|processed_pseudogene||n.*397G>A||||3944|
  - Filter by:
    - cat sample_filtered_annotated.vcf | grep HIGH
  - Count by:
    - cat sample_filtered_annotated.vcf | grep start_lost | wc -l
  - More info on format here:
    - https://pcingola.github.io/SnpEff/se_inputoutput/

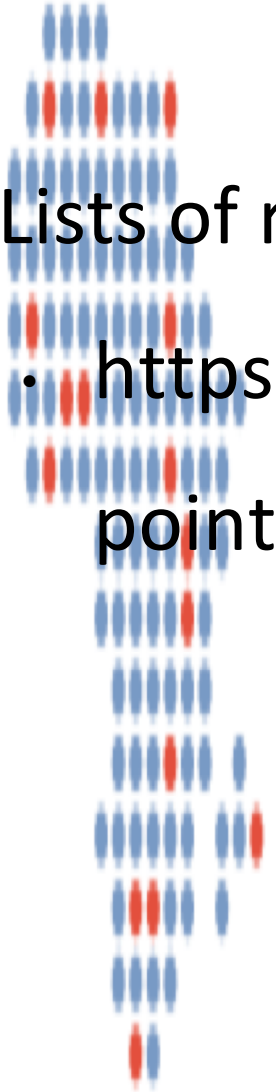
SnpEff

- Filtering:
  - Effect of mutation
    - High, moderate, low
  - Type of mutation
    - Indel, frameshift_variant, missense_variant
  - Genes of interest
    - Search by gene name to find all related variants
    - Includes upstream / downstream

SnpEff

Lists of resistance genes for different microbes

- https://bitbucket.org/genomicepidemiology/pointfinder_db/src/master/



H3ABioNet

Pan African Bioinformatics Network for H3Africa