Variant Annotation

Malay (malay@uab.edu)

November 16, 2016

Contents

1 Variant Annotation **1**

2 TABIX **1**

2.1 What is the mutation frequency of P53 gene in normal human population 2

3 BCFtools **2**

4 VCFtools **2**

1 Variant Annotation

The VCF files contains information how does a set of sequence data differs from the reference genome. By itself it does not show information about the type of changes or genes that are affected by those changes. Variant annotation is a process where we annotate those genome. Two programs are very common in annotating raw VCF files:

1. Annovar
2. SNPeff

We will use SNPeff in our class.

Download SNPeff from http://snpeff.sourceforge.net/.

```
java -jar snpEff/snpEff.jar databases | grep -i sapiens
```

We will `hg19` for annotation. To annotate a vcf use.

```
java -Xmx4g -jar snpEff/snpEff.jar hg19 snpEff/examples/test.chr22.vcf >test.chr22.ann.vcf
```

2 TABIX

A useful tool to manipulate tab-delimited files.

https://sourceforge.net/projects/samtools/files/tabix/

Index a vcf file using tabix.

```
# Compress the vcf file
bgzip TCGA_test.vcf

# Index using tabix
tabix -p vcf TCGA_test.vcf.bgz

# See all the chromoses
tabix -l TCGA_test.vcf.bgz
```

```
# Extract all line from a location
tabix TCGA_test.vcf.bgz 1:15000-50000
```

Let's do something more interesting.

2.1 What is the mutation frequency of P53 gene in normal human population

TP53 gene is on chromosome 17 in location 7571720:7590868. We will download this portion of the variation from 1000 genome data using `tabix`.

Once `tabix` is installed. We can download this portion of the file using the following command.

```
tabix -fh ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/\
ALL.chr17.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz \
17:7571720-7590868 >p53.vcf
```

3 BCFtools

Download and install BCFtools using the standard procedure.

```
wget https://github.com/samtools/bcftools/releases/download/1.6/bcftools-1.6.tar.bz2

# Extract only the PRIMARY column for the TCGA VCF and
# remove all line wihtout the alternate allete

bcftools view -s PRIMARY -a TCGA_test.vcf.gz

# filter snv where PRIMARY has DP > 10
bcftools view -s PRIMARY -a TCGA_test.vcf.gz | \
bcftools filter -i "FORMAT/DP>10"
```

4 VCFtools

** VCFtools is currently broken ** Software suite to manipulate VCF files. Website https://vcftools.github.io/index.html. There is a set of individual perl command and also a single tool.

```
wget https://github.com/vcftools/vcftools/tarball/master
tar -xvzf master
cd vcftools-vcftools-c4e2bf4/
cd src/perl
export PERL5LIB=`pwd`
cd ../..
./autogen.sh
./configure
make
cd src/perl
export PATH=$PATH:`pwd`
cd ..
cd cpp
export PATH=$PATH:`pwd`
```

You can do all sorts of operations using `vcftools`. One very comnon one is to extract one sample from multi sample vcf file. One of the column in a tcga vcf is "PRIMARY". Let's extract that column:

```
# Basic file stat
vcftools --vcf input.vcf

# Applyin filter
./vcftools --vcf input_data.vcf --chr 1 --from-bp 1000000 --to-bp 2000000

# Apply a quality filter
vcf-annotate --filter Qual=10/MinDP=20


vcf-subset -c PRIMARY TCGA_test.vcf
```