Variant Analysis

CB2-201 – Computational Biology and Bioinformatics
February 27, 2015

Emidio Capriotti
http://biofold.org/emidio

Biomolecules Folding and Disease

Division of Informatics
Department of Pathology

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Variant Call Format

The final result of the variant calling procedure is a VCF file.

```
##fileformat=VCFv4.1
##tcgaversion=1.1
##reference=<ID=hg19,source=.>
##phasing=none
##geneAnno=none
##INFO=<ID=VT,Number=1,Type=String,Description="Variant type, can be SNP, INS or DEL">
##INFO=<ID=VLS,Number=1,Type=Integer,Description="Final validation status relative to non-adjacent Normal, ......”>
##FILTER=<ID=CA,Description="Fail Carnac (Tumor and normal coverage, tumor variant count, mapping quality, ......”>
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth at this position in the sample">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Depth of reads supporting alleles 0/1/2/3...">
##FORMAT=<ID=BQ,Number=.,Type=Integer,Description="Average base quality for reads supporting alleles">
##FORMAT=<ID=SS,Number=1,Type=Integer,Description="Variant status relative to non-adjacent Normal,0=wildtype, ......">
##FORMAT=<ID=SSC,Number=1,Type=Integer,Description="Somatic score between 0 and 255">
##FORMAT=<ID=MQ60,Number=1,Type=Integer,Description="Number of reads (mapping quality=60) supporting variant">
#CHROM     POS        ID    REF   ALT   QUAL  FILTER     INFO            FORMAT                      NORMAL                  PRIMARY
1          10048      .     C     CCT   .     CA         VT=INS;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:66:.,0:.:0:.:0      0/1:32:.,2:.:2:.:0
1          10078      .     CT    C     .     CA         VT=DEL;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:25:.,0:.:0:.:0      0/1:13:.,2:.:2:.:0
1          10177      .     A     AC    .     CA         VT=INS;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:57:.,0:.:0:.:0      0/1:22:.,2:.:2:.:0
. . . . .
. . . . .
1          900505     .     G     C     .     PASS       VT=SNP;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/1:188:.,89:26:1:.:81  0/1:210:.,113:24:1:.:100
. . . . .
. . . . .
1          1991007    .     G     T     .     PASS       VT=SNP;VLS=5    GT:DP:AD:BQ:SS:SSC:MQ60     0/0:222:.,1:2:0:.:1     0/1:88:.,41:25:2:50:34
. . . . .
```

File Content

The file contains information about single nucleotide variants and indels of single or multiple samples.

For each variant the number of supporting reads for reference and alternative alleale

The original VCF does not contain any information functional effect of the variants.

Main data sources

Single genetic variants are collected in different databases:

- dbSNP - variation from all species. http://www.ncbi.nlm.nih.gov/SNP/
- EVS - specific for human. http://evs.gs.washington.edu/EVS/
- ClinVar - Variants and human health. http://www.ncbi.nlm.nih.gov/clinvar/
- Cosmic - Somatic mutation in cancer. http://cancer.sanger.ac.uk/

This information is important for variant calling but useless for capturing the complexity of genotype/phenotype relationship. The VCF more informative because we can analyze co-occurring events. The major sources are:

- 1000 Genomes: WGS data of individuals http://www.1000genomes.org/
- TCGA: Cancer Genomes https://tcga-data.nci.nih.gov/

Most common tools

The most common tools for the manipulation of vcf files are:

- tabix: fast indexer for tab separated file distributed with samtools
  http://samtools.sourceforge.net/

- vcftools: package designed for working with VCF files
  http://vcftools.sourceforge.net/

Tabix with SAM and VCF

Tabix works with bgzip files. To work we need to have an object file bgzipped and an index file

```
> bgzip $file.sam
> tabix -p sam $file.sam.gz
> tabix $file.sam.gz chr:pos1-pos2
```

How to get the variants found TP53 present in 1000 Genomes?

TP53 = chr17:7571720-7590868

```
> tabix -h $ftpfile.gz chr:pos1-pos2
```

chr17: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/ALL.chr17.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz

vcftools

Set of tools for the manipulation of vcf files

```
> vcf-merge $file1.vcf.gz $file2.vcf.gz   #indexed file
> cat $file.vcf | vcf-tstv
> vcf-query $file.vcf.gz chr:pos1-pos2
```

Select particular samples in multisample VCF

```
> vcf-subset -c sample1,sample2 $file.vcf.gz
```

Variant Annotation

There are different tools for variant annotation among the most used Annovar and snpEff.

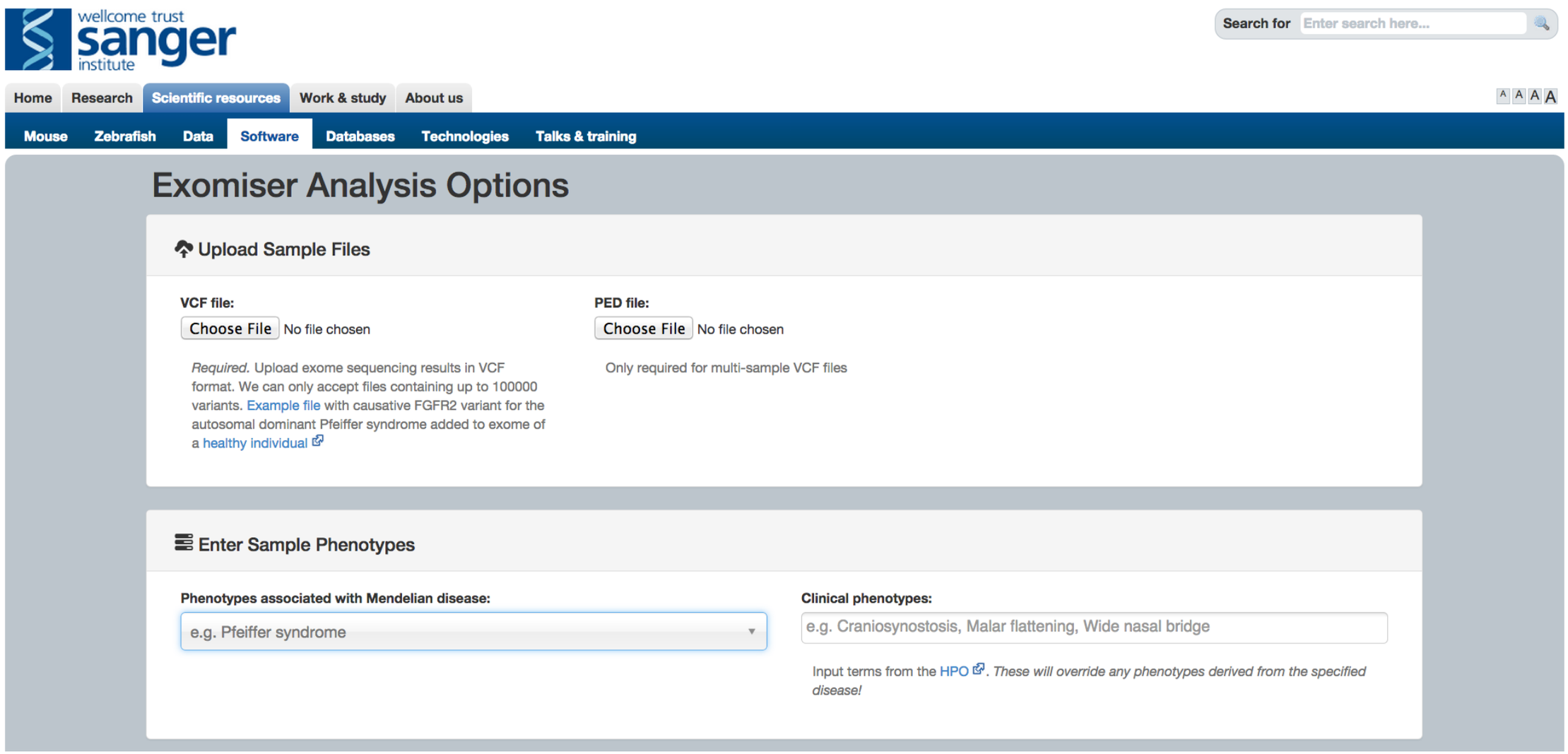
```
# Annotation
>java -jar snpEff.jar $db $file.vcf >$file.snpeff.vcf

# Filtering
> java -jar  SnpSift.jar extractFields -s "," -e "."
$file.anno.vcf CHROM POS REF ALT "ANN[*].EFFECT"
“ANN[*].GENE" "ANN[*].HGVS_P"

# Remove 0/0
> cat $file.snpeff.vcf |java -jar snpEff.jar RmRefGen
```

All in One

Exomizer is a variant analysis tools that tests presence of variants associated to specific phenotypes



http://www.sanger.ac.uk/resources/software/exomiser/submit/

Problem

Write a shell script that takes in input:

- genomic location chr:start-end
- Sample ID

and annotates the returning portion of genome.

Calculate for the number of missense variants for two samples of your choice.