**Variant Functional Annotation on whole exome sequencing variants**

**Aim**

The aim of this standard operating procedure (SOP) is to provide guidance on annotations of variants detected from whole exome sequencing samples for diagnostic purpose. The annotation file generated from this procedure is the final results from bioinformatics analysis pipeline. These annotations aim to provide information to decide which variants are more likely to be disease causing variants.

The analysis is performed by the bioinformatician.

**Tools**

The functional annotation is performed by the bioinformatic tool: ANNOVAR (version 2012 May 25). ANNOVAR is an efficient software to utilize up-to-date information to functionally annotate genetic variants detected from diverse genome.

The extraction of variants from candidate gene list and annotation of frequency in in-house database is done by BEDtools. The BEDtools utilities allow one to address common genomics tasks such as finding feature overlaps and computing coverage. The utilities are largely based on four widely-used file formats: [BED](http://genome.ucsc.edu/FAQ/FAQformat.html#format1), [GFF/GTF](http://genome.ucsc.edu/FAQ/FAQformat.html#format3), [VCF](http://www.1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcf4.0), and [SAM/BAM](http://samtools.sourceforge.net/).

**Input**

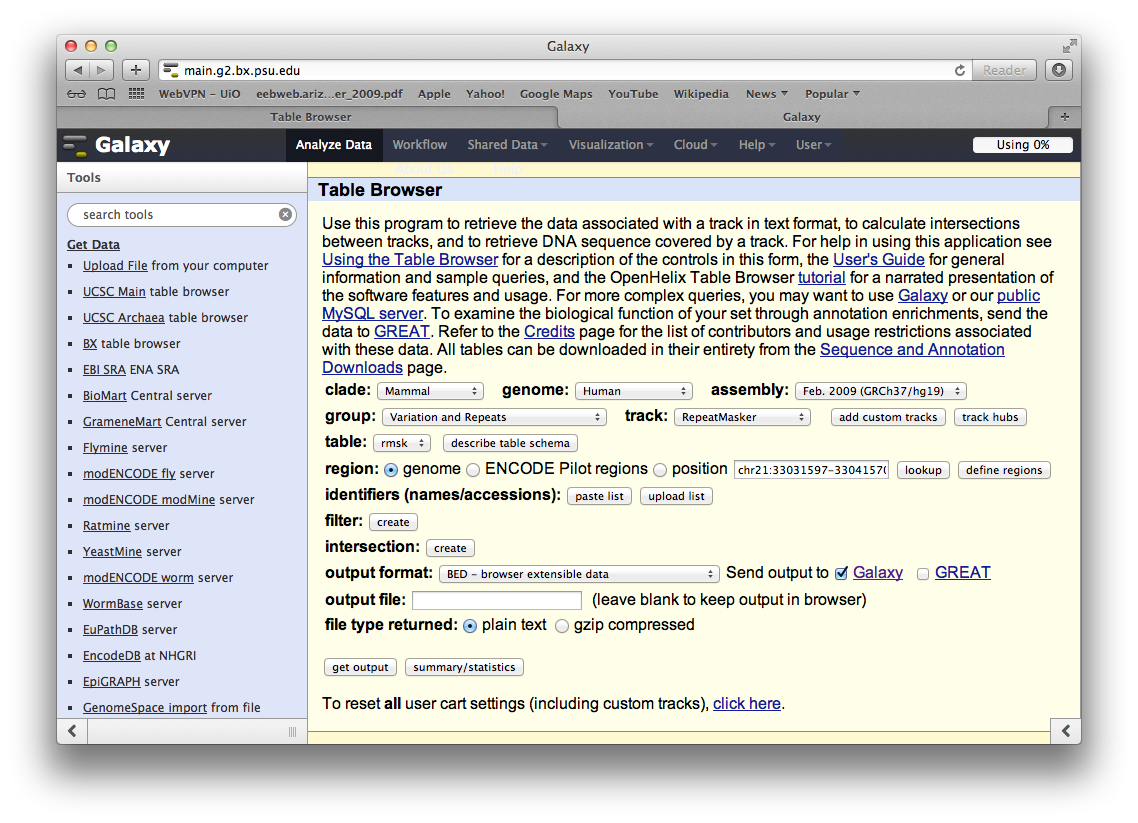
The input is VCF file generated from variant quality filtration and its index file.

**Procedure**

**1, Preparation of repeat database**

The aligner is hard to decide the exact genomic locations of reads which mapping to repeat regions, because similar sequences are appeared repetitively there. Therefore, some of the variants (especially small indels) located in repeat regions are mapping artifacts.

* Go to the following website from web browser: <https://main.g2.bx.psu.edu>
* Click “Get Data” on the “Tools” frame ->Click “UCSC Main”. On the right frame, choose options in the same way as the following figure:



* Click “get output” and Click “Send query to Galaxy” in the next page
* Wait until the file on the “History” frame is finished processing. Click “Convert Formats” on the “Tools” frame and choose “BED-to-GFF converter”. Be sure that the file name under “Convert this query” is the right file and click “Execute” button.
* A new file will appear in the “History” frame. Wait until the file is finished processing, download the file and change the file name to “repeatMasker\_hg19\_all.ori.gff3”
* Extend 2 bp on two ends of each repeats and saved the file with the name “repeatMasker\_hg19\_all.gff3” under “/Volumes/data.odin/software/mac/annovar/annovar\_2012May25/humandb/”

**2, Format conversion from VCF format to standard format of ANNOVAR**

perl convert2annovar.pl VCF\_file –format vcf4 –includeinfo –allallele > all.filtered.avinput 2>errConvert2annovarAll

**3, Functional Annotation**

The functional annotation includes:

1, Gene-based annotation: identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected.

2, Region-based annotations: identify variants in specific genomic regions, for example, conserved regions among 44 species, etc.

3, Filter-based annotation: identify variants that are reported in dbSNP, or identify the subset of common SNPs (MAF > 1%) in the 1000 Genome project, etc.

(Please see detail about information in different columns in 4.4 Annotering av detekterte variantene from SOP: Strategi for analyse av varianter fra HTS data.).

perl summarize\_annovar\_repeat.pl all.filtered.avinput –buildver 135 annovarDir/humandb –outfile sumAll 2>errSummarizeAnnovar

**4, Extraction of variants and their annotations in the candidate gene regions**

perl /Volumes/data.odin/script/variantCalling/GATK2.0/getAnnotationAll.pl 010\_qualityFiltration/all.filter.vcf 020\_annovarAnnotation path/to/candidate\_geneList\_sampleID.txt

A new folder “030\_inCand” is generated under the script running folder. Within it, all variants in the candidate gene regions is save in file “all.filter.inCand.vcf” and all csv files are saved under folder “annovarAnnotation”.

Manually check whether the annotated transcript is the one provided in the candidate gene list. If not, check the file:

050\_postVarCalProcess/gatk/020\_annovarAnnotation/sumAllTestSeparate.exonic\_variant\_function

and change the annotation (ExonicFunc) in csvs into the right transcript.

**5, Annotation of variants with frequency from in-house database**

intersectBed -wa -a /Volumes/data.odin/common/inHouseDB/variant.combine.filter.vcf.freq.txt -b all.filter.inCand.vcf > all.filter.inCand.freq.vcf

This gives the frequency for each variant appeared in the in-house database.

intersectBed -v -a all.filter.inCand.vcf -b /Volumes/data.odin/common/inHouseDB/variant.combine.filter.vcf.freq.txt

This gives the variants which are not appeared in the in-house database. Add all the variants showing up here. You need to the file “all.filter.inCand.freq.vcf” with the same format, but the freq is 0, and numbers are 0/44.

(See attachments for detail about the options of annovar scripts and BEDtools.)

**Output**

1, Annotation files:

sumAll.exome\_summary.inCand.csv and sumAll.genome\_summary.inCand.csv

The “exome\_summary” csv file contains all variants which are mapped to coding exons or canonical splicing site (2 bp away from splice site) in the candidate gene regions. The “genome\_summary” csv file contains all variants in the candidate gene regions. The files contain both variant properties and annotations. Please see detail about information in different columns in 4.4 Annotering av detekterte variantene from SOP: Strategi for analyse av varianter fra HTS data.

2, Files with variant frequency in in-house database:

all.filter.inCand.freq.vcf

The file contains all variants in the candidate gene regions. The variants are ordered by the variant frequency in in-house database from high to low. The columns in the file are the first five columns in the VCF file plus quality filters (pass or not pass quality filter), frequency in absolute value format (how many samples in the database have the variants/how many samples in the database) and fraction.

3, err Outputs for procedure 2 and 3

Err\* files are logging files for each step. if all processes running successfully:

* The err output for procedure 2, there should be a line says
* The err output for procedure 3, there should be a line says

**Variation**

The whole procedure needs to be rerun from the first failed step.

Storage

See “SOP for storage and security of high-throughput sequencing data”

**Reference**

1, Wang K, Li M, Hakonarson H. **ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data.** *Nucleic Acids Research*, 38:e164, 2010

2, Quinlan AR and Hall IM, 2010. **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics*. 26, 6, pp. 841-842

Appendix

1, convert2annovar.pl options

2, summarize\_annovar\_repeat.pl options

3, intersectBed options

4, Summarize\_annovar\_repeat.pl script

5, getAnnotationAll.pl script