MSBI 32400 – LAB 7
LARRY HELSETH, PHD AND JASON EDELSTEIN

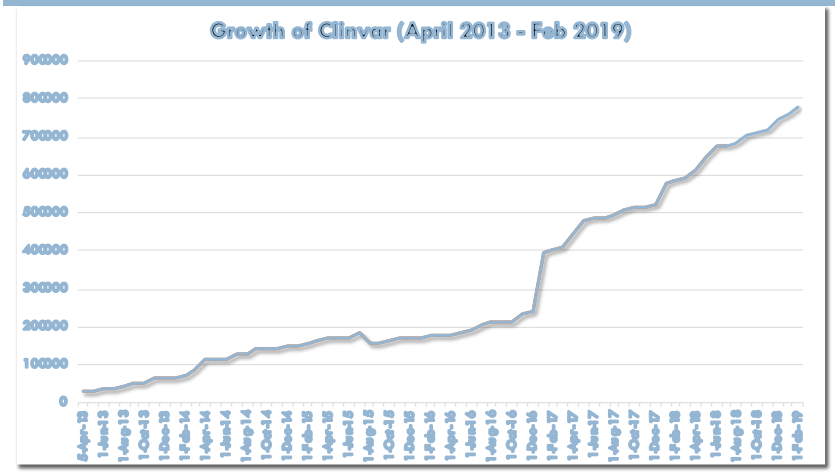
February 27, 2019

What is the "ultimate truth"?

- Jennifer L. Yen, Sarah Garcia, Aldrin Montana, Jason Harris, Stephen Chervitz, Massimo Morra, **John West**, **Richard Chen** and **Deanna M. Church**, "A variant by any name: quantifying annotation discordance across tools and clinical databases" Genome Med. 2017 Jan 26;9(1):7. doi: 10.1186/s13073-016-0396-7. PMID: 28122645
- Compared performance of snpEff, VEP and NCBI Variation Reporter on known set of problematic variants
- Used Jan 2016 GRCh37 Clinvar

Growth of ClinVar

3



Source: ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/release_notes/

Today's mission:

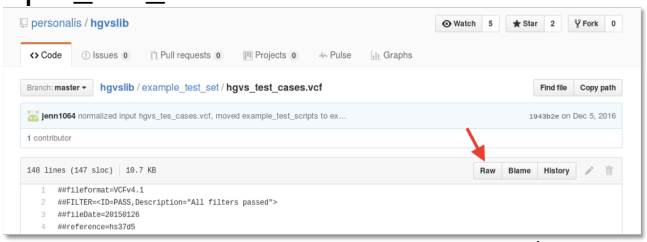
4

- Download Yen's test VCF from their GitHub site
- Download the current Clinvar database to your VM (if you didn't already do so for Lab 5)
- Run snpEff against your hg19 genome + Clinvar per lab 5 notes
- Upload the unannotated version of Yen's test VCF to VEP, wANNOVAR, and SeattleSeq on-line
- Extract your snpEff results using SnpSift
- Compare with on-line results and Yen's results

Let's get the test data set

5

- See Yen, et al. (PMID: 28122645) note on Availability of data and materials for GitHub URL, then download "hgvs_test_cases.vcf" from their example_test_set folder



- Click the link, choose "RAW", then File/Save (Ctrl-S)
- Move that to your /data/lab7/data folder

Analyze with snpEff + Clinvar

6

- From your /data/lab7/data folder:
  - java -Xmx2G -jar /data/snpEff/snpEff.jar eff \
    -canon -noLog hg19 hgvs_test_cases.vcf > \
    /data/lab7/results/hgvs_test_cases_snpEff.vcf
  - java -Xmx2G -jar /data/snpEff/SnpSift.jar annotate \
    -noLog
    /data/snpEff/data/hg19/clinvar/**clinvar_20190211.vcf.gz** \
    /data/lab7/results/hgvs_test_cases_snpEff.vcf > \
    /data/lab7/results/hgvs_test_cases_snpEff.clinvar.vcf

Remember to "sanitize" the above when you copy to the VM

SnpSift commands

7

```
[student@MSBI32400Lab1 data]$ java -Xmx2G -jar /data/snpEff/SnpSift.jar
SnpSift version 4.3i (build 2016-12-15 22:33), by Pablo Cingolani

Usage: java -jar SnpSift.jar [command] params...
Command is one of:
        alleleMat     : Create an allele matrix output.
        annotate      : Annotate 'ID' from a database (e.g. dbSnp). Assumes entries are sorted.
        annMem        : Annotate 'ID' from a database (e.g. dbSnp). Loads db in memory. Does not assume sorted entries.
        caseControl   : Compare how many variants are in 'case' and in 'control' groups; calculate p-values.
        ccs           : Case control summary. Case and control summaries by region, allele frequency and variant's functional effect.
        concordance   : Concordance metrics between two VCF files.
        covMat        : Create an covariance matrix output (allele matrix as input).
        dbnsfp        : Annotate with multiple entries from dbNSFP.
        extractFields : Extract fields from VCF file into tab separated format.
        filter        : Filter using arbitrary expressions
        geneSets      : Annotate using MSigDb gene sets (MSigDb includes: GO, KEGG, Reactome, BioCarta, etc.)
        gt            : Add Genotype to INFO fields and remove genotype fields when possible.
        gtfilter      : Filter genotype using arbitrary expressions.
        gwasCat       : Annotate using GWAS catalog
        hwe           : Calculate Hardy-Weimberg parameters and perform a godness of fit test.
        intersect     : Intersect intervals (genomic regions).
        intervals     : Keep variants that intersect with intervals.
        intIdx        : Keep variants that intersect with intervals. Index-based method: Used for large VCF file and a few intervals to retrieve
        join          : Join files by genomic region.
        op            : Annotate using an operator.
        phastCons     : Annotate using conservation scores (phastCons).
        private       : Annotate if a variant is private to a family or group.
        rmRefGen      : Remove reference genotypes.
        rmInfo        : Remove INFO fields.
        sort          : Sort VCF file/s (if multiple input VCFs, merge and sort).
        split         : Split VCF by chromosome.
        tstv          : Calculate transiton to transversion ratio.
        varType       : Annotate variant type (SNP,MNP,INS,DEL or MIXED).
        vcfCheck      : Check that VCF file is well formed.
        vcf2tped      : Convert VCF to TPED.

Options common to all SnpSift commands:
        -d                  : Debug.
        -download           : Download database, if not available locally. Default: true.
        -noDownload         : Do not download a database, if not available locally.
        -noLog              : Do not report usage statistics to server.
        -h                  : Help.
        -v                  : Verbose.
```


SnpSift extractFields

8

- Extract specified VCF fields

```
[student@MSBI32400Lab1 data]$ java -Xmx2G -jar /data/snpEff/SnpSift.jar extractFields
SnpSift version 4.3i (build 2016-12-15 22:33), by Pablo Cingolani

Usage: java -jar SnpSift.jar extractFields [options] file.vcf fieldName1 fieldName2 ... fieldNameN > tabFile.txt

Options:
        -s     : Same field separator. Default: '       '
        -e     : Empty field. Default: ''
[student@MSBI32400Lab1 data]$
```

Extract data from VCF using SnpSift

9

- java -Xmx2G -jar /data/snpEff/SnpSift.jar extractFields -s ',' -e '.' /data/lab7/results/hgvs_test_cases_snpEff.clinvar.vcf CHROM POS REF ALT ID "ANN[*].ALLELE" "ANN[*].EFFECT" "ANN[*].IMPACT" "ANN[*].GENE" "ANN[*].FEATURE" "ANN[*].FEATUREID" "ANN[*].BIOTYPE" "ANN[*].RANK" "ANN[*].HGVS_C" "ANN[*].HGVS_P" "ANN[*].CDNA_POS" "ANN[*].CDNA_LEN" "ANN[*].AA_LEN" "ANN[*].DISTANCE" "LOF[*].GENE" "LOF[*].NUMTR" "LOF[*].PERC" CLNREVSTAT RS CLNDNINCL ORIGIN MC CLNDN CLNVC CLNVI AF_EXAC AF_ESP CLNSIG CLNSIGINCL CLNDISDB GENEINFO CLNDISDBINCL AF_TGP CLNHGVS SSR > /data/lab7/results/hgvs_test_cases_snpEff.clinvar.Extracted

Upload test VCF to on-line sites

10

- wANNOVAR:
  - http://wannovar.wglab.org/
- VEP:
  - http://grch37.ensembl.org/Homo_sapiens/Tools/VEP/
- SeattleSeq:
  - http://snp.gs.washington.edu/SeattleSeqAnnotation138/

In each case, upload the original (unannotated) VCF, use hg19/GRCh37, ask for HGVS notation if possible

Compare results

11

- Pick two SNPs on different chromosomes from VCF and compare results from each method (including your local snpEff + SnpSift)
  - Cut -f9,14-16 your .Extracted file OR
  - Open your .Extracted file in text editor (or OpenOffice Calc) to visualize. Be careful to SaveAs to a different file
- What is the p. (and c.) notation for each variant you compared?
- Compare these with recommendations for reporting Protein and DNA variants as described on http://varnomen.hgvs.org

DEMO

12

- If there's time, Larry will demo some of the things you can do with VEP using the Ensembl VM
  - Current Ubuntu VM can be downloaded from https://www.ensembl.org/info/data/virtual_machine.html but requires >10 GB of data files for use.

Homework

13

- Upload to Canvas or e-mail Jason (jasone@uchicago.edu) the README with the file information requested above before next class with "**Lab #7**" in the subject line