Merge tool Example MYH7 Cardiomyopathy

PreEMPT group (internal use)

04/15/2021

**Overview:**

Below is a detailed example of steps to run merge tool and making simple edits to include more information in output.

For this example, we are interested in generated a flat file of variants on gene MYH7 on chromosome 14 with the following phenotypic characteristics:

|  |  |  |  |
| --- | --- | --- | --- |
| **INFO Tag** | **Public Term** | **Description** | **Values filtered** |
| CLNDN | Clinical Disease Name | A string consisting of the disease name used by the database specified by CLNDISDB. | Cardiomyopathy |
| ORIGIN | Clinical Origin | A string that describes the origin of the variant allele.  One or more of the following values may be assigned: 0 unknown, 1 germline, 2 somatic, 4 inherited, 8 paternal, 16 maternal, 32 denovo, 64 biparental, 128 uniparental, 256 nottested; 512 testedinconclusive, 1073741824 other. | Germline=1  Germline,somatic=1+2=3  Germline,somatic,maternal=1+2+16=19 |
| CLNREVSTAT | Clinical Review Status | A string that describes the variant’s ClinVar Review Status. | Criteria provided multiple submitters no conflict, Reviewed by expert panel |
| CLNSIG | Clinical Significance | A string that describes the variant's clinical significance. | Pathogenic, Likely\_pathogenic/pathogenic,  Likely\_pathogenic |

We want to output the **Default Settings**\* variables and the following additional variables:

|  |  |
| --- | --- |
| **ID** | **Description** |
| AC\_popmax | "Allele count in the population with the maximum AF"> |
| AN\_popmax | "Total number of alleles in the population with the maximum AF"> |
| AF\_popmax | "Maximum allele frequency across populations (excluding samples of Ashkenazi)> |
| AC\_afr | "Alternate allele count for samples of African-American/African ancestry"> |
| AN\_afr | "Total number of alleles in samples of African-American/African ancestry"> |
| AF\_afr | "Alternate allele frequency in samples of African-American/African ancestry"> |

**\*Default Settings** variables include the following:

CHROM, POS, ALLELEID, GENEINFO, REF, ALT, nhomalt, CLNREVSTAT, CLNSIG, CLNVC, ORIGIN, AC, AN, AF

**Steps:**

1. Download scripts from github.
2. Download ClinVar database using pull\_clinvar python script. Note: This pulls the most recent version of ClinVar GR37 and adds date of download to vcf file name. If need to pull ClinVar GR38, you need to go into the pull\_clinvar python script and change it.

**Command line:**

[o2\_username@login03 ~] sbatch pull\_clinvar.sh

**Output:**

clinvar\_GR37\_YEAR-MONTH-DATE.vcf.gz

1. Downlaod gnomAD exome file. The easiest way to download this directly to O2 is via gsutils. For this example, we only need to download Chromosome 14. See O2 internal wiki page for more information (<https://wiki.rc.hms.harvard.edu/display/O2/File+Transfer>).

**Command line:** This will prompt you to provide password and DUO authentication to switch over to transfer directory.

[o2\_username@login03 ~] ssh o2\_username@transfer.rc.hms.harvard.edu

**Command line:** Once successfully in the transfer directory, copy gnomAD exomes Chr14 into your merge tool directory location.

[o2\_username@transfer01 ~]$ gsutil cp gs://gcp-public-data--gnomad/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.14.vcf.bgz /home/o2\_username/merge tool location/gnomad.exomes.r2.1.1.sites.14\_$(date +%F).vcf.bgz

\*Note: if you want to explore gnomAD options, in command line:

[o2\_username@transfer01 ~]$ gsutil ls gs://gcp-public-data--gnomad/

Once transfer is complete, switch back over to regular O2 by either opening up new O2 session or submit the command below. This will prompt you to provide password and DUO authentication to switch over to default directory.

[o2\_username@transfer01 ~]$ ssh o2\_username@login.rc.hms.harvard.edu

**Output:**

gnomad.exomes.r2.1.1.sites.14\_YEAR\_MONTH\_DAY.vcf.bgz

1. For this example, we are going to show how to add more output variables and add another filter (uncomment out #5.B). If using default settings, skip to #5. Edits to code are shown in RED.

Open master\_merge shell script. Edits to code are shown in RED.

**Master\_merge shell script:**

**```````````````````````````````````````````````````````````**

# 1. Tabix both files to provide index file for the merge. Output is tabixed VCF files (.tbi).

tabix -p vcf $1

tabix -p vcf $2

# 2. Merge genotype and phenotype files by "annotating" ClinVar with the variables from gnomAD that want. Output is VCF file.

bcftools annotate -a $1 -c CHROM,POS,REF,ALT,INFO/AC,INFO/AN,INFO/AF,INFO/nhomalt,INFO/AC\_popmax,INFO/AN\_popmax,INFO/AF\_popmax,INFO/AC\_afr,INFO/AN\_afr,INFO/AF\_afr $2 > temp\_merge\_1.vcf

# 3. Query variables that you want. Add additional variables here. Output is TSV file.

bcftools query -f '%CHROM\t%POS\t%REF\t%ALT\t%INFO/ALLELEID\t%INFO/AC\t%INFO/AN\t%INFO/AF\t%INFO/nhomalt\t%INFO/AC\_popmax\t%INFO/AN\_popmax\t%INFO/AF\_popmax\t%INFO/AC\_afr\t%INFO/AN\_afr\t%INFO/AF\_afr\t%INFO/CLNDN\t%INFO/CLNREVSTAT\t%INFO/CLNSIG\t%INFO/CLNVC\t%INFO/GENEINFO\t%INFO/ORIGIN\n' > temp\_merge\_1.vcf > temp\_merge\_2.tsv

# 4. Add header in. Make sure column names align with the order that they were queried. Output is TSV file.

echo -e "CHROM\tPOS\tREF\tALT\tALLELEID\tAC\tAN\tAF\tnhomalt\tAC\_popmax\tAN\_popmax\tAF\_popmax\tAC\_afr\tAN\_afr\tAF\_afr\tCLNDN\tCLNREVSTAT\tCLNSIG\tCLNVC\tGENEINFO\tORIGIN\n" | cat - temp\_merge\_2.tsv > temp\_merge\_3.tsv

# 5.A Filter file to:

#a. Exclude variants that are missing genotype info (chose AC to filter here)

#b. Include clinical review status of CLNREVSTAT= criteria provided, multiple submitters, no conflicts OR reviewed by expert panel

#c. Include clinical significance of CLNSIG= pathogenic, likely pathogenic OR likely pathogenic/pathogenic (this code grabs any CLNSIG value that contains "athogenic")

awk 'NR==1 || !($6 == ".") && !($6 == "0") && (($17=="criteria\_provided,\_multiple\_submitters,\_no\_conflicts") || ($17== "reviewed\_by\_expert\_panel")) && ($18 ~/athogenic/)' temp\_merge\_3.tsv > temp\_merge\_4.tsv

# OPTIONAL 5.B Filter by origin- currently commented out because easier to check for germline manually in final output. Can use this filtration step instead.

# This code will include only the following ORIGIN values:

#a. germline=1

#b. germline,somatic=1+2=3

#c. germline,somatic,maternal=1+2+16=19

awk 'NR==1 || (($21==1) || ($21==3) || ($21==16))' temp\_merge\_4.tsv > temp\_merge\_4.tsv

# 6. Generate the final output file.

#a. Include gnomAD database type (exome vs genome) in final output filename

#b. Filter by GENEINFO provided in command line. Include GENEINFO name in final outuput filename

#c. Filter by CLNDN (clinical disease name) if provided in command line. Include CLNDN name in final output filename

if [[ $1 =~ ^gnomad.exome ]]; then

for i in $3;

do

awk 'NR==1 || ($20 ~/'$i'/)' temp\_merge\_5.tsv > clinvar\_gnomad\_exome\_${i}\_$(date +%F).tsv;

done

if (( $4 = 1 )); then

for j in $4;

do

awk 'NR==1 || ($16 ~/'$j'/)' clinvar\_gnomad\_exome\_${i}\_$(date +%F).tsv > clinvar\_gnomad\_exome\_${i}\_${j}\_$(date +%F).tsv;

done

fi

fi

if [[ $1 =~ ^gnomad.genome ]]; then

for i in $3;

do

awk 'NR==1 || ($20 ~/'$i'/)' temp\_merge\_4.tsv > clinvar\_gnomad\_genome\_${i}\_$(date +%F).tsv;

done

if (( $4 = 1 )); then

for j in $4;

do

awk 'NR==1 || ($16 ~/'$j'/)' clinvar\_gnomad\_genome\_${i}}\_$(date +%F).tsv > clinvar\_gnomad\_genome\_${i}\_${j}\_$(date +%F).tsv;

done

fi

fi

**```````````````````````````````````````````````````````````**

Save this shell script with different name so you do not overwrite the original default master merge shell script. For this example, shell script was saved as “master\_merge\_add.sh”

1. Run merge on command line.

**Command line:**

sbatch master\_merge\_add.sh gnomad.exomes.r2.1.1.sites.14\_YEAR-MONTH-DATE.vcf.bgz clinvar\_GR37\_YEAR-MONTH-DATE.vcf.gz MYH7 cardio

**Output:**

Temp files: temp\_merge\_1.vcf, temp\_merge\_2.tsv, temp\_merge\_3.tsv, temp\_merge\_4.tsv

Final files: clinvar\_gnomad\_exome\_MYH7 \_YEAR-MONTH-DAY.tsv

clinvar\_gnomad\_exome\_MYH7 \_cardio\_YEAR-MONTH-DAY.tsv

1. Export final file via FileZilla to get dataset onto your desktop. See O2 wiki page for details (https://wiki.rc.hms.harvard.edu/display/O2/File+Transfer).

**Additional Notes:**

1. Important dates: gnomADv3.1 exome database is expected to be released in **October/November 2021**. This version will have genome reference GR38, so we will need to merge it with ClinVar GR38 instead of GR37. We will also need to see if variable names/definitions change with this release.
2. We should keep tabs on the **gnomAD genome database**- when it becomes a more representative sample, we will eventually want to switch over to this instead of the gnomAD exome database. I do not think this will happen anytime soon though.
3. If we want to use different genotype and phenotype databases, the merge tool code will need to be tweaked. If the two files to be merged are VCFs and have the same genome reference and index file, it should not be very difficult to do. Merging ALFA with ClinVar is more difficult because they do not use the same index file- we have that only partially done (see **clinvar-ALFA-merge** directory).