**Merge tool Documentation**

06/08/2021

**Purpose:**

The purpose of this tool is to merge a phenotype database (ClinVar) with a genotype database (gnomAD) to generate a flat file of variants.

**Overview:**

The merge tool was developed to obtain estimates of allele frequencies for the pathogenic and likely pathogenic variants that will later be used as input parameters for a decision model. It pulls phenotype data from ClinVar and genotype data from gnomAD to generate a dataset of variants for the model that fulfill the desired phenotype criteria. The tool eliminates the need for manual curation of variants and allows for further output specifications that are not available from the online manual curation platforms. This tool aims to simplify this process while also still being flexible to the specific needs of a project.

The tool uses python to pull ClinVar databases, gsutils to pull gnomAD databases, and bash script and BCFtools to format and merge the databases. This can all be completed using a Linux/Unix-based platform.

**Database Background:**

ClinVar is a public archive of reports of relationships between human variations and phenotypes, including clinical significance, submitter details, and other related data [1]. These reports are aggregated and downloadable via FTP. This database allows us to assess the clinical validity and significance of genetic variants for certain diseases to decide which variants to include in our model input parameters. It is updated monthly.

GnomAD (Genomic Aggregation Database) is a public archive of exome and whole-genome sequences from disease-specific and population genetic studies [2]. It contains genotype information, including allele frequency estimates of the overall population and by specific subpopulations. There are two gnomAD databases that are available: genome and exome.

For the purposes of our projects, we chose to pull gnomAD exome version 2.1.1. We chose gnomAD v2 database because the Broad Institute recommends using it for coding region analyses, and because gnomAD v2 has a much larger number of observations. For non-coding regions, the Broad recommends using the gnomAD v3 database (<https://gnomad.broadinstitute.org/faq>). GnomAD v3 is currently only available as a genome database, not an exome database. The most recent version available for the genome database is v3.1.1, updated November 2021.

The gnomAD exome database is linked to GR37. While GR38 is the newer, more updated genomic reference, the most updated gnomAD exome database is linked to GR37. Therefore, we have decided to use ClinVar GR37 and gnomAD GR37 exome databases for now. As these databases are updated, we will update the Merge Tool to reflect the best databases available.

The ClinVar and gnomAD databases can be downloaded as VCF files. VCF files are formatted in such a way that there are different subfields and specific ways to assess these fields. The merge tool uses BCFtools, a program for querying, sorting, and manipulating VCF files, to perform operations on these genomic databases [3]. BCFtools package is part of a larger genomic program called HTSLib [4]. BCFtools is a well-documented and maintained program that makes VCF file handling clear and manageable. The primary commands we use in the merge tool are view, query, and annotate. For further information on BCFtools, visit the source code and documentation sites (<https://www.htslib.org/>, <https://github.com/samtools/bcftools>).

**Setup:**

To use this tool, you need access to a Linux/Unix-based computing platform. Download the github repository and install the required programs (https://github.com/graceannobrien/Merge-tool). Additional edits can be made to customize it to your project’s specific needs. See **Extended Doc** for more details on customizing the tool.

**Input formats:**

This tool works for VCF files only.

**Output formats:**

The output is a flat file of AF estimates for the gene of interest.

**Default Settings:**

The default settings for the merge tool phenotype filtering and conditions are as follows:

1. Variants must be pathogenic or likely pathogenic (P/LP). This means that the CLNSIG variable must contain one of the following strings: “Pathogenic”, “Pathogenic/Likely\_pathogenic”, “Likely\_pathogenic”.
2. Variants must be 2 star or higher. This means the CLNREVSTAT variable must contain one of the following strings: “reviewed\_by \_expert\_panel” or “criteria\_provided,\_multiple\_submitters,\_no\_conflicts”.

If you wish to alter these, see the Data Dictionaries for each database, located in the git repository, to see the variable options. Note: the code will need to be altered to include those variables as additional columns.

The default settings for the merge tool output files contain the following variables:

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

If you wish to alter these, see the Data Dictionaries for each database, located in the git repository, to see the variable options. Note: the code will need to be altered to include those variables as additional columns. Examples of variables you may want to add include AC, AN, AF for popmax and for specific populations like African Americans, Europeans, etc.

**Default output variables**

|  |  |  |
| --- | --- | --- |
| **Output variable** | **Description** | **Database** |
| CHROM | Chromosome | ClinVar, gnomAD |
| POS | Position | ClinVar, gnomAD |
| ALLELEID | Allele ID | ClinVar |
| GENEINFO | List of pairs of gene symbol and NCBI GeneID at the location of the variation. The gene symbol and ID are separated by a colon (:) and each pair is delimited by a vertical bar (|). Example: SYMBOl1:GeneID1|SYMBOl2:GeneID2| . | ClinVar |
| REF | Reference allele | ClinVar, gnomAD |
| ALT | Alternate allele | ClinVar, gnomAD |
| nhomalt | Count of homozygous individuals in samples | gnomAD |
| CLNDC | A string consisting of the disease name used by the database specified by CLNDISDB | ClinVar |
| CLNREVSTAT | Integer that represents ClinVar Review Status. One of the following values may be assigned: No assertion; No criteria; Criteria provided single submitter; Criteria provided multiple submitters no conflict; Criteria provided conflicting interpretations; Reviewed by expert panel; Practice guideline. | ClinVar |
| CLNSIG | A string that describes the variant's clinical significance.  One of the following values may be assigned: 0 unknown, 1 untested, 2 nonpathogenic, 3 probable-nonpathogenic, 4 probable-pathogenic, 5 pathogenic, 6 drug-response, 7 histocompatibility, 255 other. | ClinVar |
| CLNVC | Variant type | ClinVar |
| 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 de-novo, 64 biparental, 128 uniparental, 256 not tested; 512 tested inconclusive, 1073741824 other. | ClinVar |
| AC | Alternate allele count for samples | gnomAD |
| AN | Total number of alleles in samples | gnomAD |
| AF | Alternate allele frequency in samples | gnomAD |

**Flowchart:**

**Diagram

Description automatically generated**

**General Pipeline:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Steps** | **Description** | **Command** | **Output** |
| **Curate gene list** | Generate list of genes and the chromosomes they are on that you want to run the merge tool on. | | |
| **Download ClinVar** | Use FTP to download database. | sbatch clinvar\_pull.sh | clinvar\_GR37\_YEAR-MONTH-DATE.vcf.gz |
| **Download gnomAD** | Use gsutils to copy database to merge tool directory. Depending on storage levels, likely will need to do by individual chromosome. | gsutils cp gs://gcp-public-data--gnomad/release/2.1.1/vcf/exomes/ gnomad.exome.r2.1.1.  sites.CHROMOSOME /location of Merge Tool/ gnomad.exome.r2.1.1.sites.CHROMOSOME\_YEAR\_MONTH\_DAY.vcf.bgz | gnomad.exome.r2.1.1.  sites.CHROMOSOME\_YEAR\_MONTH\_DAY.vcf.bgz |
| **Perform merge tool** | For the merge tool, three input arguments are necessary: phenotype database, genotype database, and gene of interest. One input argument is optional: disease of interest. | sbatch master\_merge.sh gnomad.exomes.r2.1.1.sites.CHROMOSOME\_YEAR-MONTH-DATE.vcf.bgz clinvar\_GR37\_YEAR-MONTH-DATE.vcf.gz GENE DISEASE | **Temp files:** temp\_merge\_1.vcf, temp\_merge\_2.tsv, temp\_merge\_3.tsv, temp\_merge\_4.tsv    **Final files:** clinvar\_gnomad\_exome\_GENE \_YEAR-MONTH-DAY.tsv  clinvar\_gnomad\_exome\_GENE\_DISEASE\_YEAR-MONTH-DAY.tsv |
| **Export dataset** | Use FileZilla or other file transfer software to move final output file onto desktop. | | |
| **Collapse into single AF per gene** | Collapse the merge tool outputs for a gene into a single estimate of AF of P/LP variants for a gene. Briefly, we constructed a beta distribution with parameters alpha = (sum of AF\_exome \* average of AN\_exome), and beta = average of AN\_exome – alpha. This provides a distribution around the probability of a P/LP variant in the gene in the population. The conversion from probability of variant alleles to probability of individuals with a variant is performed in the simulation model code itself using Hardy Weinberg. | | |
| **Repeat for all genes** | Run merge tool for all other genes from curated list in Step 1. Add collapsed AF estimates for genes in same excel spreadsheet for upload to model parameters. | | |

**Merge tool Example:**

In this example, we want to produce a list of variants on the SCN5A gene that are related to Long QT in the gnomAD exome v2.1.1 database. Since SCN5A is on chromosome 3, we will only pull the gnomAD exome chromosome 3 database.

1. **Download databases**

**ClinVar:**Using FTP python code

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

**gnomAD:** Using gsutils in transfer directory

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

1. **Run Merge tool**

[o2\_username@login03 ~] sbatch master\_merge.sh

gnomad.exome.r2.1.1.sites.3\_2021-04-23.vcf.bgz clinvar\_GR37\_2021-04-23.vcf.gz SCN5A longQT

This will output the following flat TSV files:

clinvar\_exome\_ SCN5A \_2021-04-23.tsv

clinvar \_exome\_ SCN5A \_longQT\_2021-04-23.tsv

clinvar \_genome\_ SCN5A \_2021-04-23.tsv

clinvar \_genome\_ SCN5A \_longQT\_2021-04-23.tsv

1. **Export flat files using FileZilla or other file transfer software.**
2. **Manual curation of variants if needed. This may be disease-specific and vary by project. Document additional steps as needed.**
3. **Collapse merge tool output to an estimate of one AF per gene**
   1. Let AF = sum of AF\_exome
   2. Let Sample Size = Average(AN\_exome) where AN = allele number
      1. Note the MAX sample size = 2 x # people in gnomAD
      2. # people in gnomAD v2\* exomes only = 125,748 🡪 max = 251,496
   3. Let alpha = AF\*sample size
   4. Finally, let beta = sample size – alpha

**Additional Capabilities:**

The gnomAD genome database can be used as the phenotype database instead of gnomAD exome database in the merge tool. It follows all the same steps as the gnomAD exome and ClinVar merge. We have created an R function to combine gnomAD exome and ClinVar merge tool output with gnomAD genome and ClinVar merge tool output of the same gene. This allows users to compare the variants in each gnomAD database for that gene and combine the variants’ allele frequencies from the genome and exome into one allele frequency. It combines them by using their allele counts and allele numbers. As the gnomAD genome database adds more data and more closely represents the general population’s allele frequencies, this R function will become more useful.

**Limitations:**

There are three main limitations of the merge tool. Due to the size of gnomAD, we are unable to pull the entire database, instead we need to do it by chromosome. Additionally, this step of pulling gnomAD databases is not automated and needs to use a different browser than the rest of the tool. It is not able to be downloaded via FTP. We plan on developing this further and working on incorporating the gsutils command into a python script.

Additionally, we are limited by how often the databases are updated and how often the VCF files we download are updated to reflect those changes. The ClinVar VCF file is updated on the first Thursday of the month. GnomAD releases updates approximately once a year in October/November. Since the ALFA database was first released in March 2020, there has only been one update since then, and they have not stated a set release schedule yet.

The merge tool is also limited in its ability to use genotype databases other than gnomAD. We are currently working on adapting the tool to be able to merge ClinVar with the Allele Frequency Aggregator (ALFA) database.

1. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Res, 2018. **46**(D1): p. D1062-D1067.

2. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.

3. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2).

4. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 2011. **27**(21): p. 2987-93.