**Abstract**

While many genetic variations are benign, some have been linked to various types of diseases. Databases such as dbVar provide a curated list of common variants vs. variants linked to clinical phenotypes. Previous studies have identified specific variants that produce disease phenotypes. Characterizing the types and locations of genomic variations associated with specific diseases is beneficial for informing ongoing studies, with the ultimate goal of identifying disease causal variants. This web user interface (GUI) was developed using Python to calculate a distribution of variant type, length, and genomic location for each phenotype.

1. **Introduction**

Genetic variations are the differences in DNA among individuals or the differences between populations. Structural variation (SV) is a location on a genome assembly, marked by start and stop coordinates, representing a submitter's assertion of a region containing observed structural variation. Specific variations produce disease phenotypes.1 A phenotype is the observable characteristics or traits of an organism that are produced by the interaction of the genotype and the environment. The need is to be able to inform ongoing studies to identify disease causal variants. The existence of dbVar, which is a database of human genomic structural variation where users can search, view, and download data from submitted studies, allows us to extract data necessary to be able to study and experiment with genetic variations.

The identification of variants and genes for complex, polygenic traits has always been a difficulty within research. This difficulty arises as polygenic traits may have tens to hundreds of risk variants where each variant explains only a small proportion of disease heritability. For researchers to accurately accomplish these, several factors are included, such as a series of genetic association studies on polygenic traits to develop a better resolved human genome, a map of common genetic variants, and sufficient sample sizes to gain statistical power. When causal variant identifications are conducted they typically involve searching through sizable regions of genomic DNA near disease-associated SNPs for sequence variants in functional elements, including protein-coding, regulatory, and structural sequences. Prioritization of these searches is greatly aided by knowledge of the location of functional sequences in the human genome.2

Goals:

The goal of this project is to create an interactive site that enables the user to select a phenotype, plot a distribution of these associated variant characteristics and visualize all relevant genetic variations of the phenotype. The target users for this application are people curious about phenotypes and their variants. If the application is successful, it will increase the user's understanding of phenotypes and their variants regarding location within the genome.

Related Work:

Genomic structural variants have long been implicated in phenotypic diversity and human disease but dissecting

the mechanisms by which they exert their functional impact has proven elusive. Recently however, developments in high-throughput DNA sequencing and chromosomal engineering technology have facilitated the analysis of structural variants in human populations and model systems in unprecedented detail. In this Review, we describe how structural variants can affect molecular and cellular processes, leading to complex organismal phenotypes, including human disease. We further present advances in delineating disease-causing elements that are affected by structural variants, and we discuss future directions for research on the functional consequences of structural variants. 3

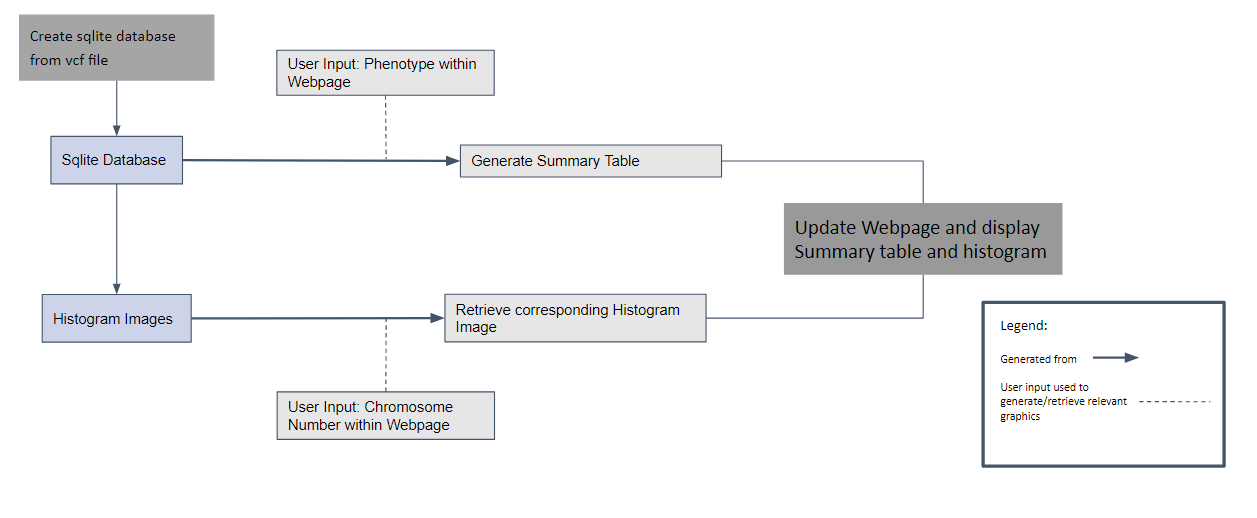
# DataSet

The data set used in this project was VCF files with structural variants found in GRCh38. To access this data the main website with dbVar data was <https://www.ncbi.nlm.nih.gov/dbvar/>. The direct location of files (Benign and clinical) was from<https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/GRCh38/vcf/>. Once the database was obtained, it was used to create the SV\_PHENOTYPES database needed to create the GUI.

# Methods and Implementation

Implementation of the proposed webpage project is done through Python and PHP script. The vcf databased was retrieved and stored within a local folder. Python scripts were utilized to convert the vcf database into an SQLite database and generate histogram images of the phenotype by chromosome.

The results are then stored within a local folder. A PHP script was utilized to create a user-interactive webpage for the user to select the desired phenotype and chromosome number. Once the user selects the desired phenotype, the webpage will generate an HTML table of all relevant variants of the phenotype. Once a chromosome number is selected, the webpage retrieves the relevant histogram image from the local folder. The flow chart below depicts the relationship between each script and its function.

****

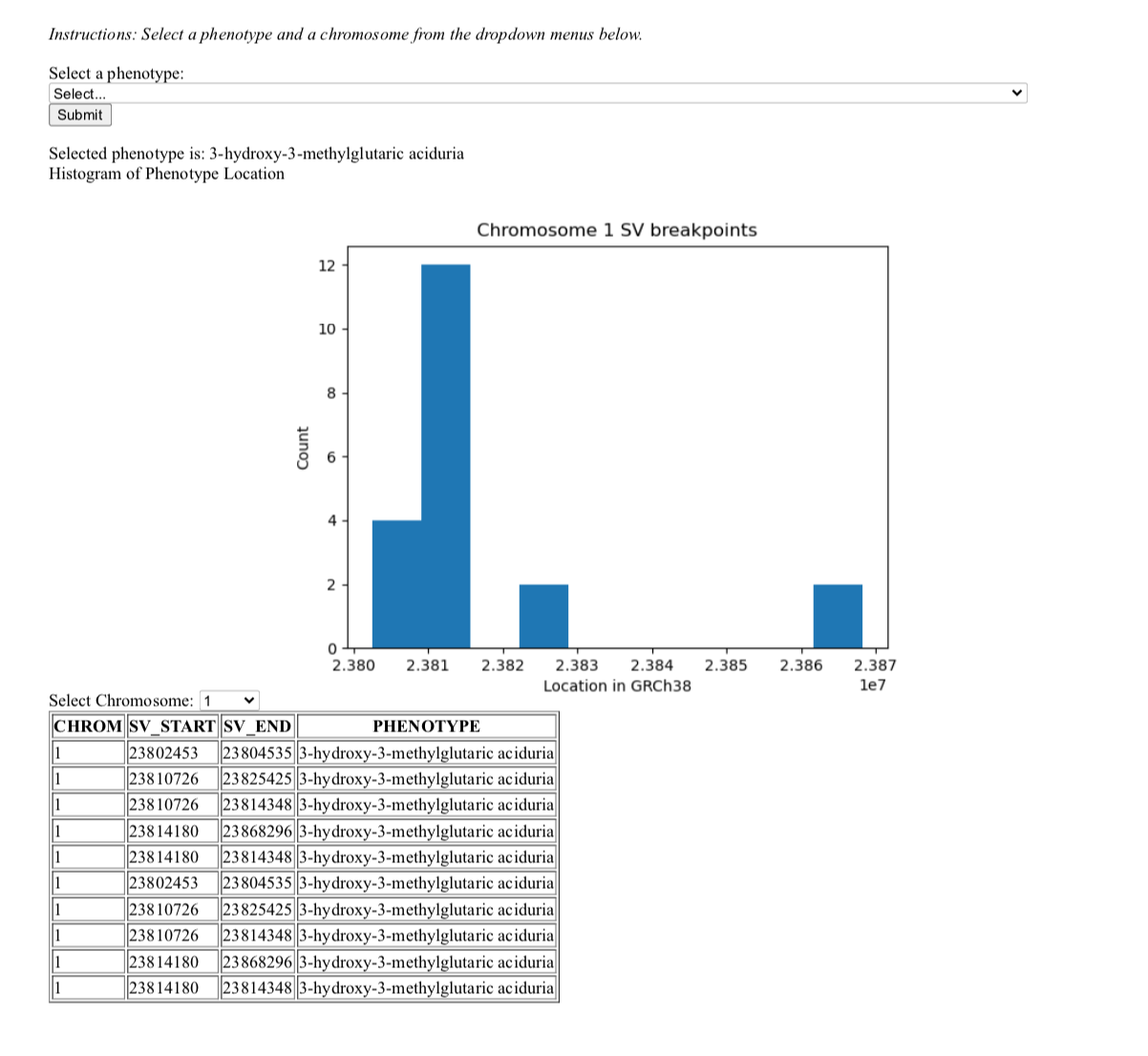
**Figure 1. Functionality flowchart for GUI**

The scripts used to create the webpage were created by team members. The javascript to select the desired chromosome was based on an example created by Ahmet Sacan, which was implemented in our code to assist with the histogram population.

# Experiments and Results

The user interactive webpage inputs two dropdown sections for phenotype and chromosome number.

When selecting a choice of phenotype and specifying chromosome count, a histogram and table will be shown. The table will display the selected phenotype and all its variants. The table contains the chromosome, SV\_Start, SV\_End, and the phenotype. When the chromosome dropdown is selected, a histogram of the desired phenotype location on the desired chromosome is displayed. Figures 2 and 3 demonstrate the functionality of the webpage. In Figure 2, the phenotype 3-hydroxy-3-methylglutartic aciduria and chromosome 1 was selected. The webpage, as a result, display a histogram and Summary table for 3-hydroxy-3-methylglutartic on chromosome 1. In Figure 3, the phenotype neurodevelopmental disorder and chromosome 2. The Histogram displays the phenotype and the SV breakpoints on chromosome 2. The Summary Table displays the phenotype across all chromosomes in the database.



**Figure 2 Interactive Webpage** with phenotype and chromosome count selected. The output of the histogram showing the location of the phenotype in GRCh38 Vs. the count.



**Figure 3. Multiple Chromosomes** for the phenotype selected are shown in the table in an organized fashion as well as the histogram shows the location of the phenotype in GRCh38 Vs. the count.

# DISCUSSION

The purpose of this webpage project is to visualize the desired phenotype and its variants. Summarizing the vcf database with a comprehensible technique, allows the user to identify variants that can be investigated as a potential source of the same disease. Biologically, the results of the webpage match expectations as the

webpage visualizes the vcf database information by phenotype and genome location. One of the limitations of our project revolves around our database source. Our results are based on the information that has already

been published and analyzed to be related to the selected phenotype. However, this webpage project has the potential to be customized to any user’s specific needs. In the future, we would like to identify phenotype variants across multiple database sources. Ultimately creating a tool that allows the user to explore the desired phenotype and their variants and access information regarding the variant directly from the webpage. This

allows the webpage to become a primary navigation page for researchers and laypersons to discover,

identify, and explore disease phenotypes and their variants. Thus meeting the ultimate goal of creating a better understanding of the genetic cause of disease.

# REFERENCES

1. “Home - Dbvar - NCBI.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, <https://www.ncbi.nlm.nih.gov/dbvar/>.
2. Kai Wang, et al. “A Useful Tutorial.” *Quick Start-up Guide - ANNOVAR Documentation*, <https://annovar.openbioinformatics.org/en/latest/user-guide/startup/>.
3. “Mondo Disease Ontology.” *Mondo.monarchinitiative.org*, https://mondo.monarchinitiative.org/.
4. “Overview of Structural Variation.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/dbvar/content/overview/.
5. Stankiewicz, Paweł, and James R Lupski. “Structural Variation in the Human Genome and Its Role in Disease.” *Annual Review of Medicine*, U.S. National Library of Medicine, <https://pubmed.ncbi.nlm.nih.gov/20059347/>.
6. Weischenfeldt, Joachim, et al. “Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease.” *Nature News*, Nature Publishing Group, 18 Jan. 2013, https://www.nature.com/articles/nrg3373.