SOFTWARE DEVELOPMENT GROUP PROJECT 2024

Team Mint

BY: Amin Shire, Bishan Rai, Christina Yiangou, Namit Nagar, Peng Zhou

Introduction

With the help of this web tool, users may easily explore the SNPs on chromosome 1 in 3929 samples that come from various communities worldwide.

In particular, it enables users to cluster data using Principal Component Analysis and Admixture Analysis, two population genetic structure assessments on the given data set.

In addition, a search function allows the user to obtain allele and genotype frequencies along with other pertinent data, including clinical details for the SNPs of interest in the populations of interest.

Software Architecture

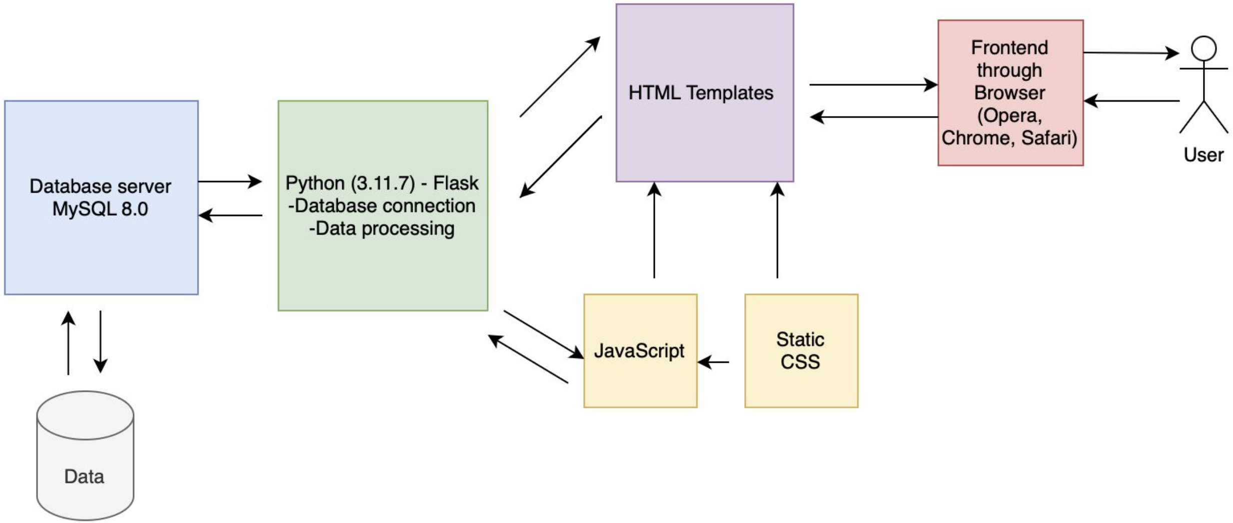


Figure 1: Graphic representation of the application's software architecture. This application is based on a MySQL 8.0 database that contains the relevant information about SNPs on chromosome 1 in 3929 individuals. The software was developed using Flask through Python 3.11.7. The database and Flask interact through the mysql.connector package to run the code of the app based on what the user chooses. Flask uses Javascript, HTML templates and CSS to define the visual representation of the website. This image was created using Draw.io.

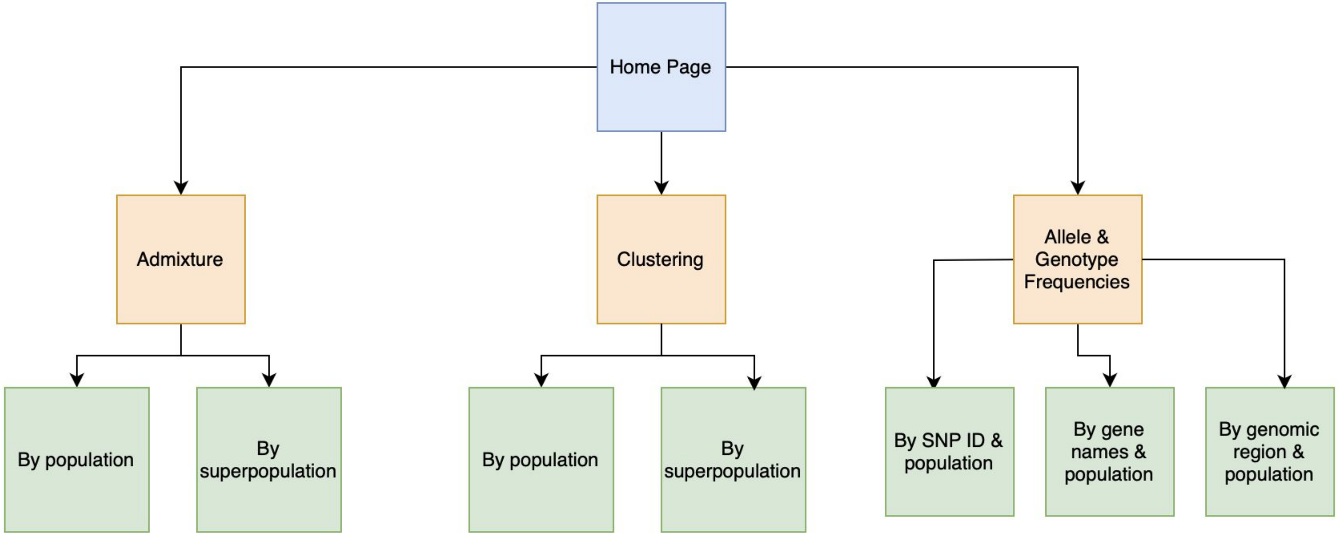


Figure 2: Detailed view of frontend of the web application showing all the possible analyses available to the user. Admixture and Clustering analysis can be done based on populations or superpopulations. Allele and genotype frequencies can be retrieved for SNPs of interest by directly choosing the SNPs of interest, the gene names or a genomic region on chromosome 1. This image was created using Draw.io.

Data Processing and Collection

To convert the existing IDs to their corresponding rsIDs, we used bcftools to annotate the chr1.vcf.gz file with SNP ids extracted through the FTP site of Broad Institute (Li, 2011).

Information on clinical relevance for SNPs was retrieved from Clinvar as clinvar.vcf.gz (Landrum et al., 2017).

The names of genes and their positions were gathered from Ensembl using a data-mining tool called BioMart (Kinsella et al., 2011).

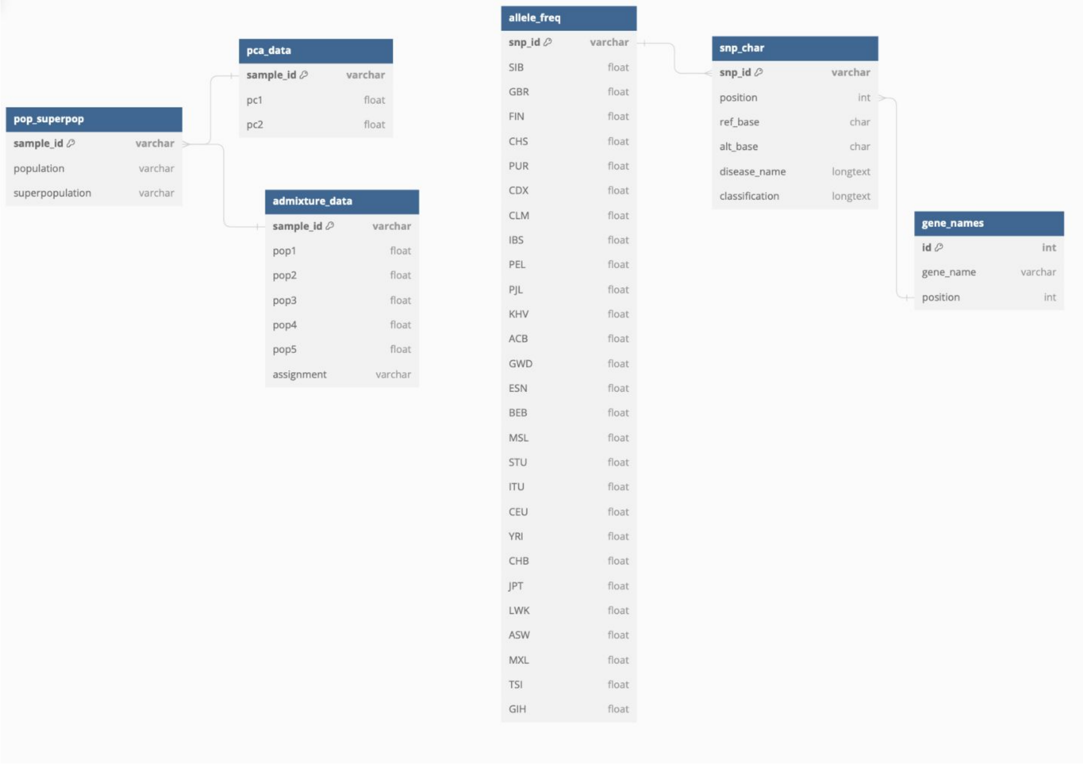
The corresponding geographical superpopulations for each population were retrieved from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015).

Database Populating

VCF files were firstly converted into Zarr files. This step is highly beneficial when working with large VCF files in terms of saving memory and fast data extraction.

We then executed make_database.ipynb which is designed to build the database structures and populate the database with data extracted from zarr files and other datasets.

Database Schema



pop_superpop	
sample_id	varchar
population	varchar
superpopulation	varchar

pca_data	
sample_id	varchar
pc1	float
pc2	float

admixture_data	
sample_id	varchar
pop1	float
pop2	float
pop3	float
pop4	float
pop5	float
assignment	varchar

allele_freq	
snp_id	varchar
SIB	float
GBR	float
FIN	float
CHS	float
PUR	float
CDX	float
CLM	float
IBS	float
PEL	float
PJL	float
KHV	float
ACB	float
GWD	float
ESN	float
BEB	float
MSL	float
STU	float
ITU	float
CEU	float
YRI	float
CHB	float
JPT	float
LWK	float
ASW	float
MXL	float
TSI	float
GIH	float

snp_char	
snp_id	varchar
position	int
ref_base	char
alt_base	char
disease_name	longtext
classification	longtext

gene_names	
id	int
gene_name	varchar
position	int

Figure 3: Schema of the database used in this application called ‘final’. The database has 6 tables; pop_superpop, pca_data, admixture_data, allele_freq, snp_char and gene_names. Pca_data and admixture_data tables are linked to the pop_superpop table through their sample_id columns. Allele_freq table is linked to the snp_char table through the snp_id column while gene_names table is linked to the snp_char table through the position column. This image was created using dbdiagram.io.

Clustering

- Principal Components Analysis (PCA)
- Popular as the first step in population genetic studies (McVean, 2009; Conomos, Miller & Thornton, 2015; Elhaik, 2022)
- PLINK (Purcell et al, 2007)
- PC1 & PC2 visualized in a scatterplot using matplotlib
- As the number of samples increases the proportion of variance explained by the first 2 PCs gets smaller and less accurate (Gaspar & Breen, 2019)
- Only the first 2 PCs are visualised which makes it harder to identify individuals with ancestries from more than one population (Gaspar & Breen, 2019)

Admixture Analysis

- ADMIXTURE (Alexander, Novembre & Lange, 2009)
- Faster than STRUCTURE (Alexander, Novembre & Lange, 2009)
- k = 5
- Data visualized in a stacked bar plot using matplotlib
- Does not account for linkage disequilibrium (LD) (Alexander, Novembre & Lange, 2009)

Allele and Genotype Frequencies

- PLINK to calculate minor allele frequencies = q (Purcell et al, 2007)
- Per SNP per population
- 1-q = p = reference allele frequency
- Genotype frequencies using HWE: $p^2 + 2pq + q^2 = 1$ (Andrews, 2010)
- User can search a list of SNP IDs, gene names or a genomic position
- Populations to include in search

Fst Matrix

- Fixation Index (Fst) using alt allele frequencies (Bhatia et al., 2013)
- $Fst = N/D$
  - $N = q1(p2 - p1)+q2(p1 - p2)$
  - $D = q1p2+p1q2=N+q1p1+q2p2$
- Range from 0 to 1
  - 0 is no difference
  - 1 is completely different
- Visualised using matplotlib
- Lighter the squares the more different the population

Limitations and Further Development

- Decrease code size
- Improve database schema
- Enhance memory usage, especially in graph plotting functions
- Additional features, e.g. option to customise the name of the downloaded matrix text file
- Enrichment of the dataset

Live demo

http://127.0.0.1:5000/

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Andrews, C. A. (2010). *The Hardy-Weinberg Principle*. Nature.com. https://www.nature.com/scitable/knowledge/library/the-hardy-weinberg-principle-13235724/

Bhatia, G. et al. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology*, *39*(4), 276–293.

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports*, *12*(1), 14683. https://doi.org/10.1038/s41598-022-14395-4

Gaspar, H. A., & Breen, G. (2019). Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-2680-1

Kinsella, R. J. et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, *2011*(0), bar030–bar030. https://doi.org/10.1093/database/bar030

Landrum, M. J. et al. (2017). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis. *PLoS Genetics*, *5*(10). https://doi.org/10.1371/journal.pgen.1000686

Purcell, S. et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

The 1000 Genomes Project Consortium. (2015). A Global Reference for Human Genetic Variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393