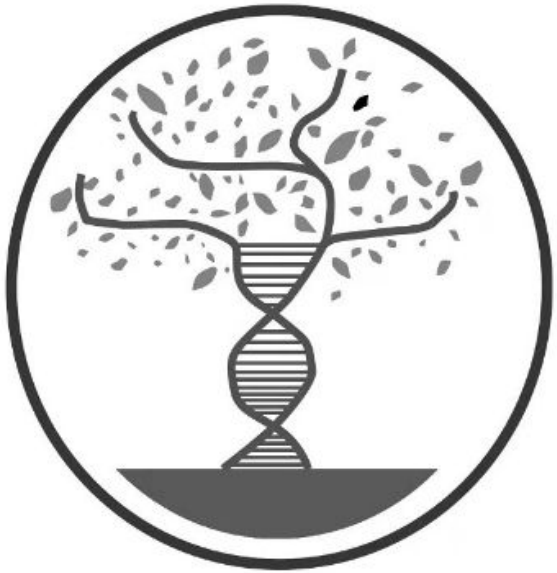
SNP Global

By: Gabriel, Ervin, Zainab, and Zibo

Overview

- Software Architecture
- Website Structure
- Data Collection
- Database Schema
- Statistics
- Technologies Utilized
- Demonstration of SNPGlobal
- Opportunities for Future Development

The aim of SNPGlobal is to provide biologists and researchers a resource to obtain information on single nucleotide polymorphism in humans. SNPGlobal allows users to obtain basic information and summary statistics for SNPs by searching based on gene name, Gene ID, rsID, or genomic position.



Software Architecture

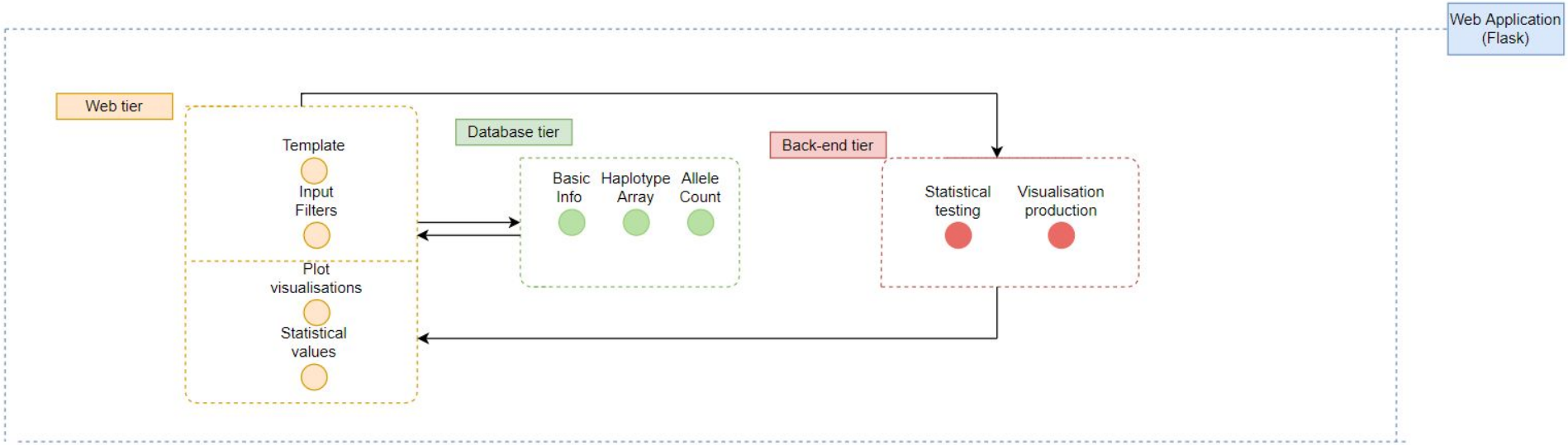


Figure 1: Software architecture visualised with three specific tiers and their respective layers that form the Web Application.

N-tier architecture style, dividing logical layers and physical tiers

Website Structure

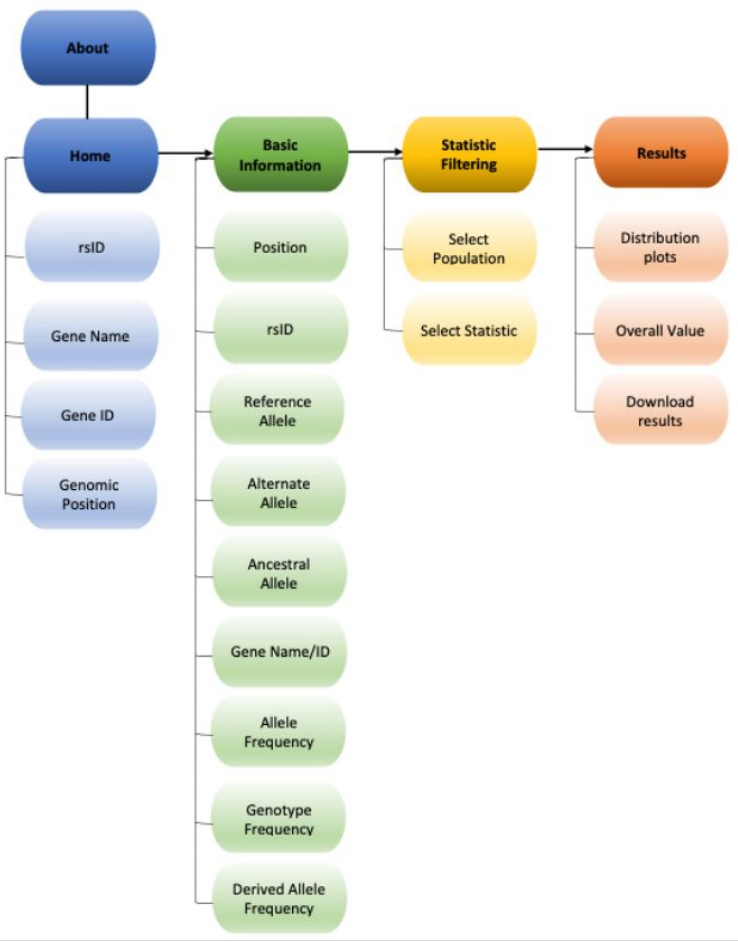
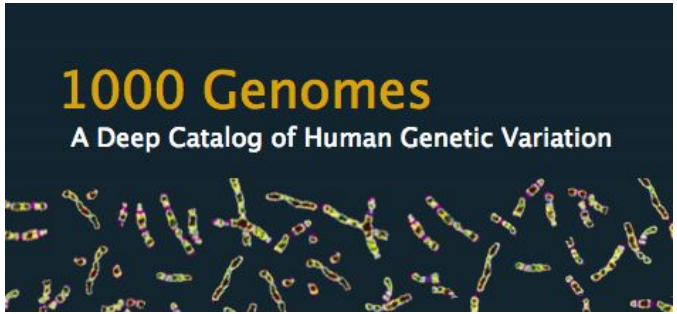


Figure 2: A schematic diagram of the website with the pages and the information on each page. The arrows indicate directional relationship from one page to the next. The line connects the information on each page.

Data collection

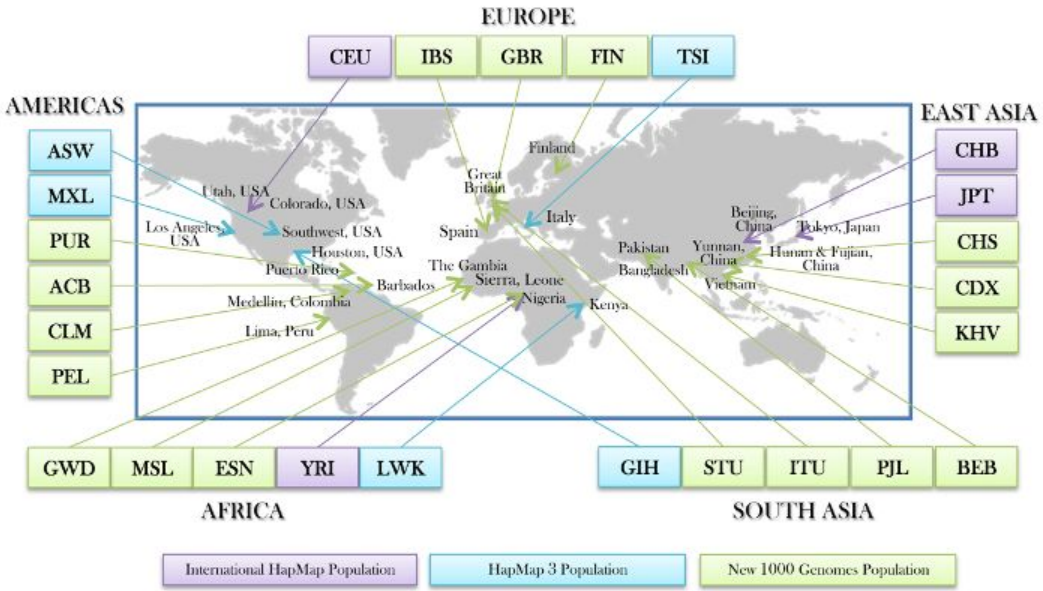
- The International Genome Sample Resource:
  1000 Genomes 30x on GRCh38
  - Phased VCF of chromosome 22
  - Annotation text file
  - Sample information files

- Ensembl
  - VCF of chromosome 22
  - Gene names and Alias



(Byrska-Bishop et al., 2021) and (Howe et al., 2021)

Populations



- Luhya in Webuye, Kenya (LWK)
- Southern Han Chinese, China (CHS
- Colombian in Medellin, Colombia (CLM)
- Toscani in Italy (TSI)
- Sri Lankan Tamil in the UK (STU)

(Byrska-Bishop et al., 2021)

Database Schema

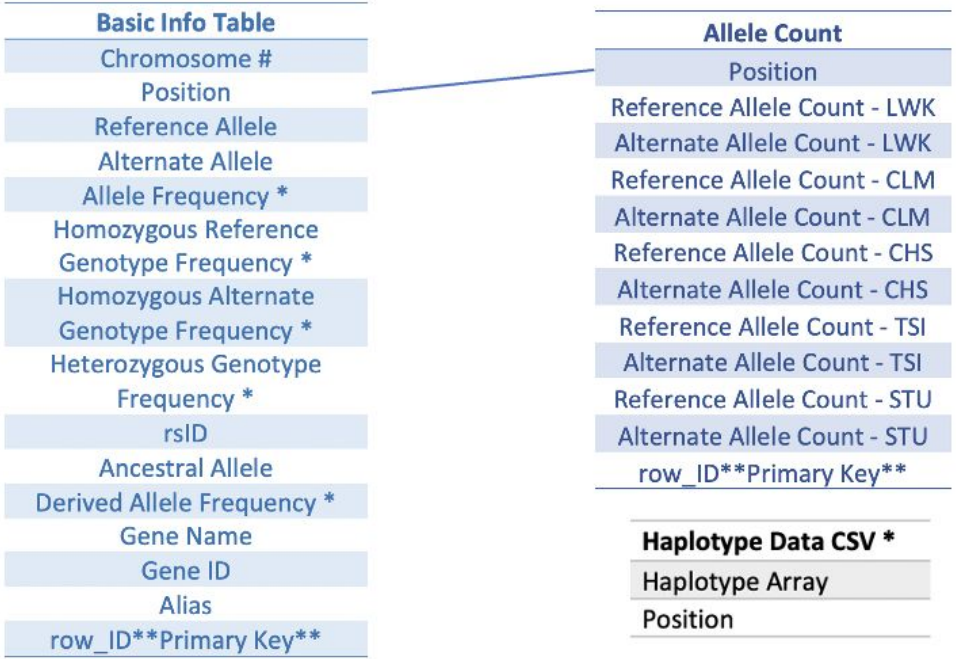


Figure 3: Database schema showing the Basic Information sql table, Allele count sql table, and the haplotype csv. The line connects the basic Information and Allele count table based on position. The * indicates the value was repeated for each of the 5 populations.

Statistics

FST

Tajima’s D

Watterson’s estimator

Haplotype diversity

FST

The variance of allele frequencies between populations or the probability of Identity by descent.

FST value ranges from 0 to 1. Higher value means bigger genetic distance.

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}(1-\bar{p})}$$

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

Tajima’s D

Tajima’s D uses the difference of the average number of pairwise differences and the number of segregating loci to explain selection history.

Tajima’s D ＞0: population shrinkage

Tajima’s D = 0: no evidence of selection

Tajima’s D ＜0: population expansion

$$D = \hat{\theta}_{\pi} - \hat{\theta}_{S}$$

$$E(S) = a_1 M$$

Watterson’s Estimator

Watterson estimator is a method to estimate the genetic diversity of a population by counting the number of polymorphic loci.

Values vary with population size.

$$\widehat{\theta}_w = \frac{K}{a_n}$$

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Haplotype Diversity

Haplotype diversity refers to the frequency at which two different haplotypes are randomly selected from a sample.

Value roughly between 0.2 and 0.8 overall.

$$H = \frac{N}{N-1}(1 - \sum_i x_i^2)$$

Population Genomics Tools

- Scikit-allel package used for all summary statistics
- Provides numerous statistical functions for genetic variations
- Almost every statistic shared same parameters
- Similar outputs to be used for visualisation

Diversity Visualisations

- Plotly package used to visualise all summary statistics
- Embedded onto webpage using HTML
- Interactive plots with zoom and scroll features
- Downloadable images of plots in jpeg format

User input

- User searches & selection options registered with Flask-WTF

- Input forms basis for Database Querying and Data Manipulation using Python

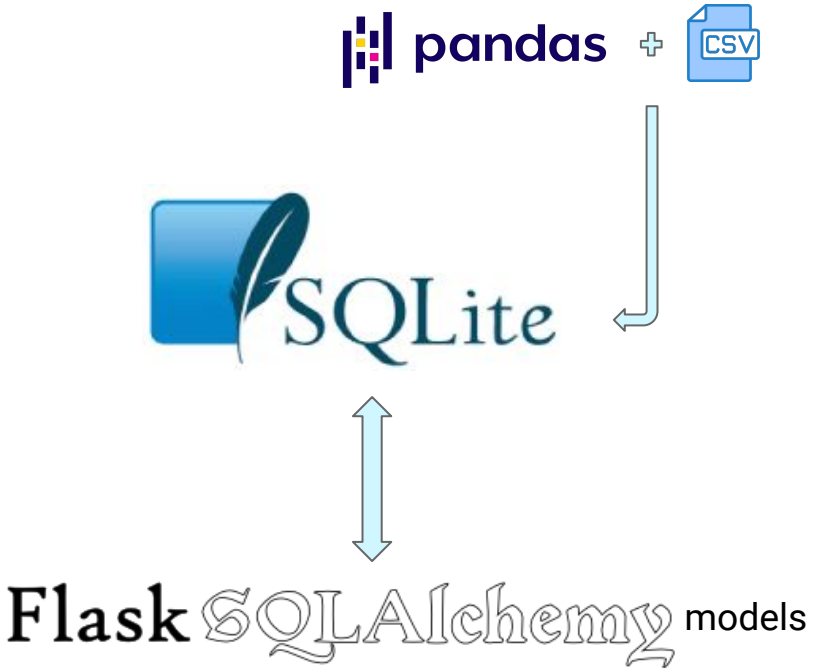
Database creation + integration

Creation

- Database tables created using SQLite3
- Tables populated using the Pandas library + CSVs

Integration

- Model declaration with Flask SQLAlchemy facilitates communication between user search and Database



Website demonstration

http://127.0.0.1:5000/

Limitation & Improvement

Expanding the database

Increasing Statistical Tests

Links to external websites

Option for the sliding window size

Flow of the Website

Thank you!

Any Questions?

References

Bayer, M., 2012. SQLAlchemy. In A. Brown & G. Wilson, eds. The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks. aosabook.org. Available at: "http://aosabook.org/en/sqlalchemy.html"

Byrska-Bishop, M. et al (2021). High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios. SSRN Electronic Journal,.

Howe, K., Achuthan, P., Allen, J., Allen, J. and Bennet, R., 2021. Ensembl 2021. Nucleic Acids Research, [online] 49(D1), pp.D884–D891. Available at <https://academic.oup.com/nar/article/49/D1/D884/5952199.

Grinberg, M., 2018. *Flask web development: developing web applications with python*, " O'Reilly Media, Inc."