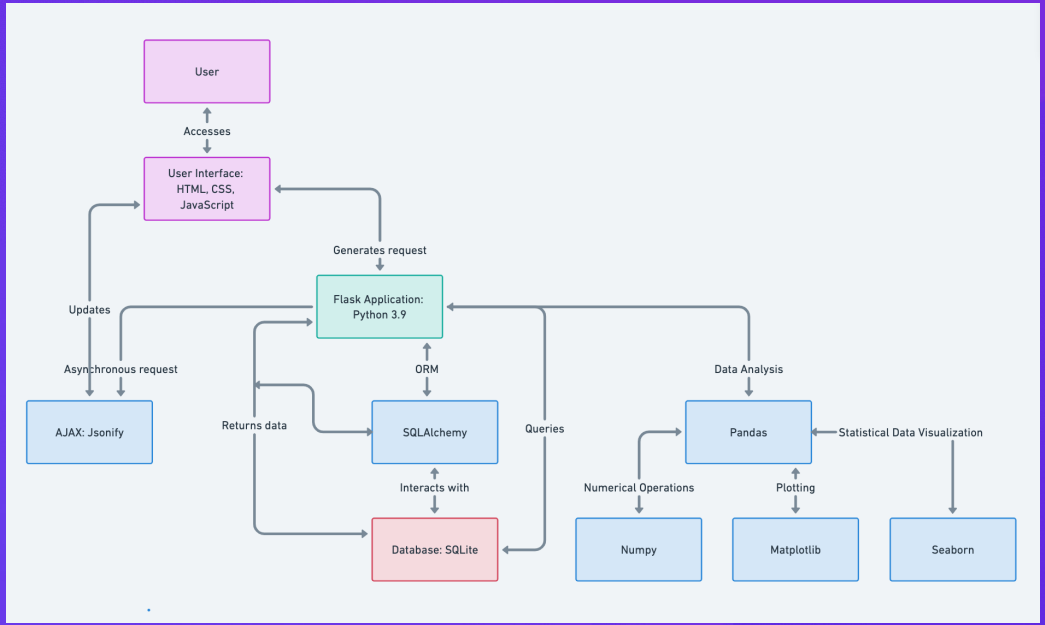
WELCOME TO OPENSUSE SOFTWARE DEVELOPMENT PROJECT – WEB APPLICATION

Bruna Plateroti, Hanna Abby, Elma Handzic, Ryan Beadle.

Introduction to Our Software

Streamlining Genetic Analysis

- **Specialised Genetic Analysis**: Analysis of human population genetics, focusing on clustering, admixture analyses, and SNP examination.
- **Comprehensive Backend**: Utilizes SQLite and SQL Alchemy for database management, with Python and PLINK/ADMIXTURE for analytical computations.
- **User-Friendly Frontend**: Developed with Flask, enhanced by JavaScript and AJAX for efficient interaction.
- **Impact and Significance**: Facilitates fast, precise genetic analysis, making complex data accessible and interpretable for researchers.

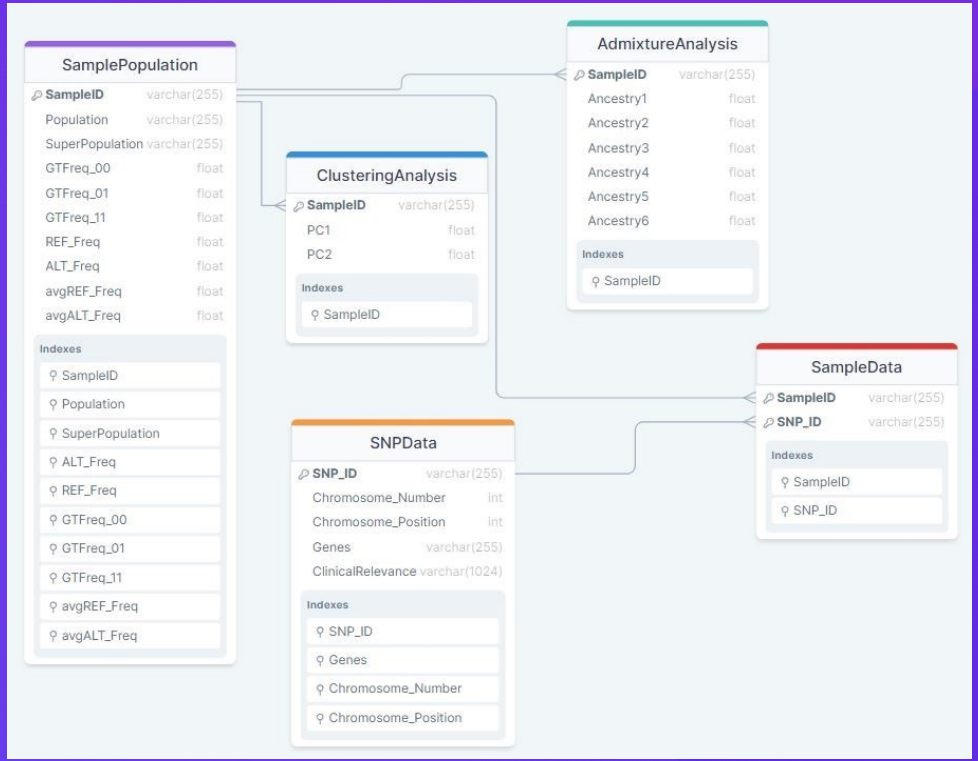


Structure of the web-application for data processing and visualization. Pink signifies front end technologies, blue illustrates business logic, red shows backend technology and green visualises integration of technologies between all layers.

Backend Database Architecture

Optimising Genetic Data Management

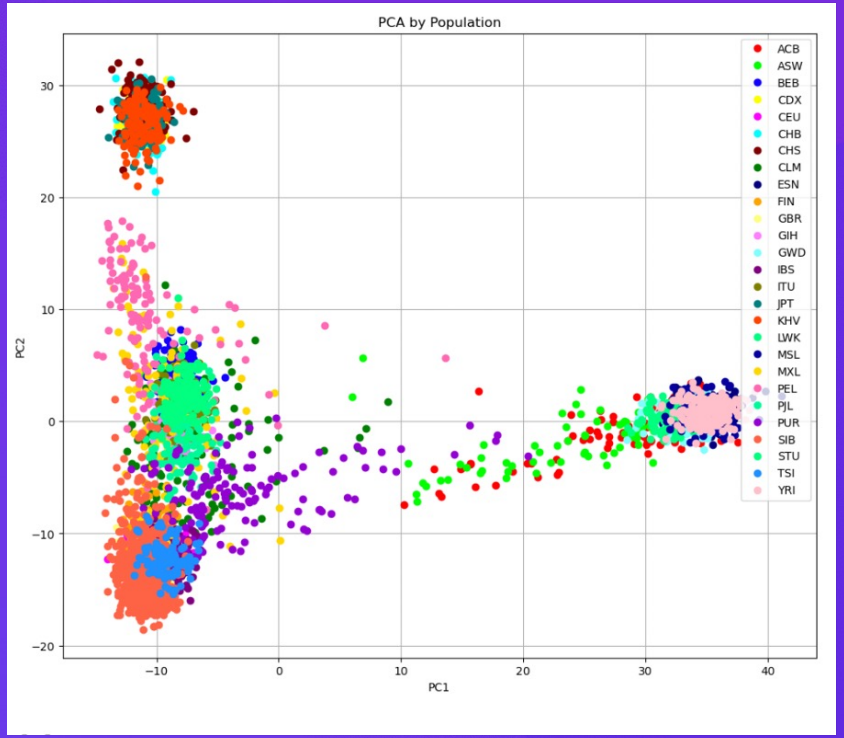
- **Database Schema Design**: Incorporates tables like AdmixtureAnalysis, ClusteringAnalysis, SNPData, SamplePopulation, and SampleData, optimized for efficient data querying and storage.
- **SQLite & SQL Alchemy Integration**: Chosen for simplicity and compatibility, enabling streamlined database management and interaction.
- **Comprehensive Data Processing**: Data filtering and structuring using bcftools and Pandas, ensuring integrity and readiness for analysis.
- **Enhanced Query Performance**: Tables are indexed to facilitate faster execution and enhance data quality, significantly improving query performance.



Database schema which consists of five tables, each containing specific fields corresponding to the data it holds

Principal Component Analysis (PCA)

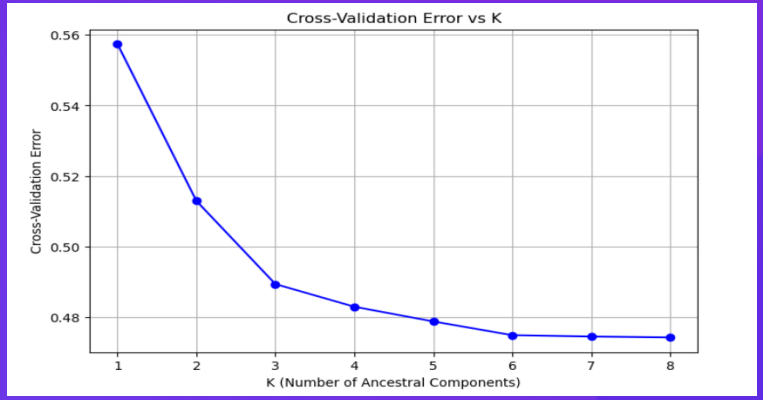
- **Why PCA ?** Chosen for its proficiency in managing high dimensional data from SNPs unlike MDS/UMAP whilst providing more interpretable results.
- **Data Preparation & Standardisation**: Involves cleaning genetic data from the VCF and standardizing SNPs for PCA, as-well as numerical conversion of genotype data.
- Use of Scikit-learn for the PCA and number of component= 2 (PC1 and PC2 capture the most variance)
- **Visualisation and Storage**: Utilizes **Matplotlib** for clear visualization of PCA outcomes, stored in TSV files in the database for efficient access and analysis
- From **Database → SQL Query → perform_pca function → PCA plot**

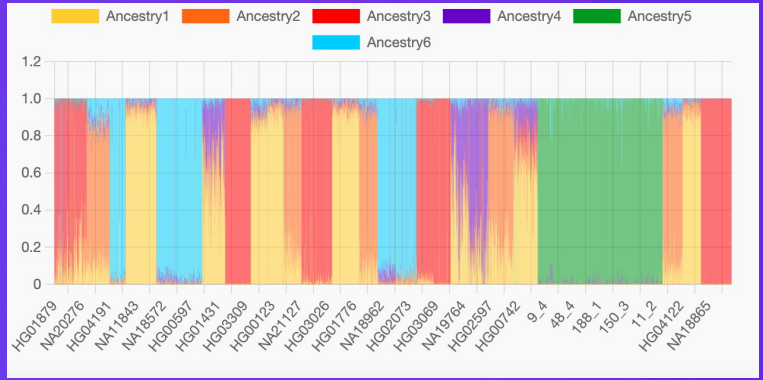


Principal Component Analysis (PCA) of genetic data by population. This scatter plot visualizes the genetic variation across multiple populations, with each dot representing an individual's genetic makeup projected onto the first two principal components (PC1 and PC2).

Admixture Analysis Overview

- **Data Processing and Conversion with PLINK**: Our VCF file was transformed into PLINK's binary format for compatibility with ADMIXTURE.
- **ADMIXTURE Software**: Chosen for ease-of-use as well as speed and integrated cross-validation ability.
- **Optimal K Value Determination**: Cross-validation errors were analysed to determine the optimal K value (number of genetic clusters).
- **Interactive Visualization**: Results integrated into the SQL database and retrieved using SQL alchemy and jsonify before being displayed with chart.js.





Data Retrieval and FST Analysis

Key Genetic Metrics Uncovered

- **Clinical relevance and genes for SNPs:**
  - EBI GWAS Catalog – Clinical relevance
  - Ensembl – Genes

- **Genotype and Allele Frequencies**:
  - Computes average allele and genotype frequencies per SNP per population

- **FST Matrix Calculation**: Fixation Index (Fst)
  - Per populations as selected by the user

- **Visualisation Tools**:
  - Matplotlib - heatmap

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$

1.Calculate average allele frequencies for reference allele and alternate allele across selected populations

2. Calculates variance of allele frequencies among populations for both alleles

3. Calculates Fst for reference allele and alternate allele using their respective variances and average frequencies

4. Average values of reference and alternate alleles to get single fst value per population

Frontend Logic and User Interaction

Enhancing User Experience

- **HTML, CSS, and JavaScript Integration**:
  - HTML templates
  - CSS for styling
  - JavaScript content updates

- **AJAX for Asynchronous Updates**:
  - Real-time data retrieval

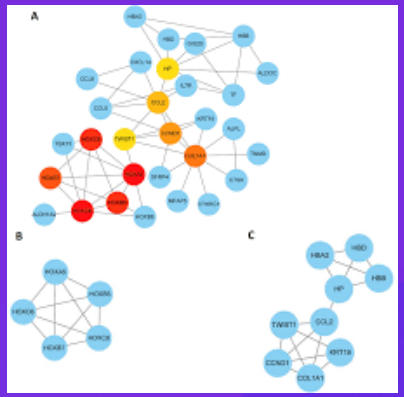
- **Visualisation with Chart.js**
  - Chart.js - rendering interactive charts

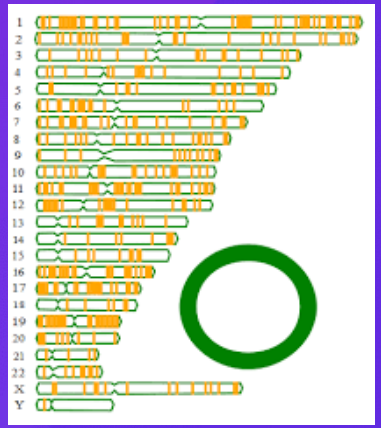
- **Blueprints for Modular Code**
  - Flask Blueprints - Structure

Limitations and Future Directions

Evolving Our Genetic Analysis Software

- **Data Scope Limitation**: Currently focused on chromosome 1, missing critical genetic variations across other chromosomes. Future expansions to include comprehensive genetic data coverage.
- **Database Scalability**: Utilises SQLite; transitioning to more robust systems like PostgreSQL could enhance scalability and performance for larger datasets.
- **Real-time Data Analysis**: Relies on pre-calculated results for admixture and PCA. Integrating tools like Chromo-Map could enable dynamic, real-time analysis capabilities and allow the user to look at the clustering / admixture data under different parameters.
- **Bioinformatics Network Analysis**: Future integration with comprehensive bioinformatics software, such as **Cytoscape**, could provide enriched network analysis and uncover relationships between genes.





PostgreSQL