

SOFTWARE DEVELOPMENT PROJECT: DEEPIN GENETICS HUB

This project was developed by MSc Bioinformatics students from Queen Mary University of London.

Group members: Harry Towner, Lydia Obeng and Shakkthi Pougoujali

Supervisors: Dr Matteo Fumagalli and Professor Conrad Bessant

Table of Contents

ABOUT DEEPIN GENETIC HUB SOFTWARE	3
DESIGN PHILOSOPHY	4
Target	4
SNP Data.....	4
Populations	4
Simplicity	4
Navigation	4
WEB DEVELOPMENT SOFTWARE	5
Flask	5
HTML	5
Pandas	5
Plotly	5
ADMIXTURE	5
PLINK.....	5
SciPy.....	5
Bcftools	5
SOFTWARE ARCHITECTURE.....	6
DATA COLLECTION.....	7
SNP Data.....	7
Allele Frequencies	7
Genotype Frequencies	7
DATA ANALYSIS	8
Statistical tests	8
Principal Component Analysis (PCA)	9
ADMIXTURE	9
DATA SCHEMA	10
FRONTEND.....	11
Website Features.....	11
Running the website.....	13
Design and Structure of the Website	14
LIMITATIONS AND FUTURE DEVELOPMENT	14
REFERENCES	15

ABOUT DEEPIN GENETIC HUB SOFTWARE

A group of MSc Bioinformatics students have collaborated to design Deepin Genetics Hub, a web-based software tool that explores the intricate and complex dimensions of human population genetics. The study of human population genetics has encompassed various scientific and practical implications, highlighting its importance in investigating the ancestral patterns and genetic diversity, all essential applications in multiple fields such as medicine and public health. Pivotal research carried out by Alexander et al in 2009 [1], “Genome Research”, has inspired our team to support a clinician in comprehending genetic data that has been collected from a diverse range of human populations. The foundational understanding of clustering analysis such as Principal Component Analysis (PCA) and ADMIXTURE (Alexander et al, 2009) were utilised as an approach to the population structure analysis.

We were provided with a dataset containing genotyped markers and whole genome sequencing data of more than 4000 individual and for this project our focus was only on chromosome 1. The information was sourced from Eurasian samples including Siberian population. The aim of this research is to be able to provide an understanding of the genetic relationships between the Eurasian samples and other populations. Through structure analyses, we collected genotype and allele frequencies helping us to examine and explore the genetic variations. We have designed our web application in a way that it is able to perform clustering and admixture analyses and retrieve genetic information for regions specified by the user, including SNP IDs and Gene IDs and present informative visualisations. Besides performing analyses, our web application also provides clinical relevance for specific SNP and Gene IDs that a user is particularly interested in. As we progress, we are committed to improve the usefulness of the software tool in terms of building a connection between research in genetics and the practical applications for clinical settings.

DESIGN PHILOSOPHY

Deepin Genetics Hub is a web tool that is particularly developed for analysing and investigating on population genetic data. Several tools, including GenPop[2] and TEGA [3], embody a similar ideology, providing platforms for geneticists and epidemiologists to explore the study of human population genetics. Much like these platforms, our design philosophy is dedicated to enable a diverse target audience to perform easy and manageable analyses.

Target

Our web tool encompasses a wide range of target audience including clinicians, geneticist and students who are looking to use a powerful and intuitive software that facilitates exploration and analysis of human genetic diversity. We have designed the application in a simple and efficient way, so that users are able to search SNP queries and relevant information easily.

SNP Data

Single nucleotide polymorphisms (SNPs) represent the most widespread type of genetic variation among humans and play a crucial role in the predisposition to various diseases, influencing how individuals react to medical treatments, and tracing the transmission of hereditary illnesses. Beyond illustrating the genetic disparities among different human populations, data on SNPs can be instrumental in identifying specific groups that may have increased susceptibility to diseases or differing reactions to pharmaceuticals.

Populations

Our Genetics Hub contains many types of data on twenty-seven diverse populations from around the world with populations from all the five super populations: Africa, Mixed American, East Asian, European and South Asian. We have included our collaborators Siberian dataset into the European super population due to the shared ancestry with Slavic speaking eastern Europeans but there is a possibility some of the samples within the data could have ancestry of Siberian Tatars and are a mixture between Europeans and Asians. We have chosen to label them as European based on this ancestry rather than their geography.

Simplicity

The website was designed to be easy to use and shouldn't require a deep understanding of genetics to understand and query. Simply selecting what populations you wish to visualise for ADMIXTURE and PCA Graphs produces the output and for more in-depth analysis the user only needs to have an SNP id or gene name they want to look at to compare allele and genotype frequencies between populations.

Navigation

The navigation bar supports this desire for simplicity. The user just needs to move between tabs to select what they wish to compare and view. There are separate tabs for Clustering Analysis (PCA) and ADMIXTURE analysis and then finally a tab for the SNP browser where the user can search their desired SNP or Gene and the results displayed on the same page with the option to download the results into a text file.

WEB DEVELOPMENT SOFTWARE

Our web platform is built using several different types of software, each chosen for its strength and versatility.

Flask

At the core, Flask, the lightweight and versatile web framework written in Python, links the back-end architecture. Selected for its simplicity and flexibility, Flask streamlines route management and view rendering, making web development more intuitive.

HTML

The structure and content of the web pages are crafted using HTML, the standard markup language for documents designed to be displayed in a web browser. It provides the scaffolding for the user interface, ensuring that information is presented in a logical, user-friendly manner. When combined with Flask's rendering capabilities, HTML becomes a dynamic canvas for our web applications.

Pandas

Data manipulation is handled by Pandas, a data analysis library that excels in structuring and organizing complex datasets. Within our Flask application, Pandas is used to transform data into easily understandable formats, facilitating the presentation of results to end-users. By converting data frames into clear, interpretable forms.

Plotly

For visual analytics, we incorporate Plotly, an interactive graphing library that allows users to engage with data visually. Plotly integrates smoothly with our data processing pipelines and converting numerical insights into engaging heatmaps PCA graphs and stacked bar charts.

ADMIXTURE

We utilize ADMIXTURE, a software tool optimized for deducing the ancestry composition of individual samples based on genetic variation data.

PLINK

PLINK, the whole-genome association analysis toolset, is used for its comprehensive statistical methods. It provides an multitude of functionalities to perform genetic variant analysis, fitting our need for in-depth genetic data examination.

SciPy

Python-based open-source software for mathematics, science, and engineering, complements our analytical toolset. It extends the capabilities of our platform by providing additional modules for optimization, integration. Its statistical subpackage is particularly useful for tasks that require robust statistical tests.

Bcftools

Bcftools, a suite of utilities for variant calling and manipulating VCFs and is an essential tool for working with genomic data. It ensures that our platform can handle data efficiently, allowing for the processing, filtering, and analysis of genetic variants.

By using this selection of technologies, we have crafted a web platform that is powerful yet user-friendly, capable of sophisticated data analysis while providing an intuitive interface for users to interact with complex datasets.

SOFTWARE ARCHITECTURE

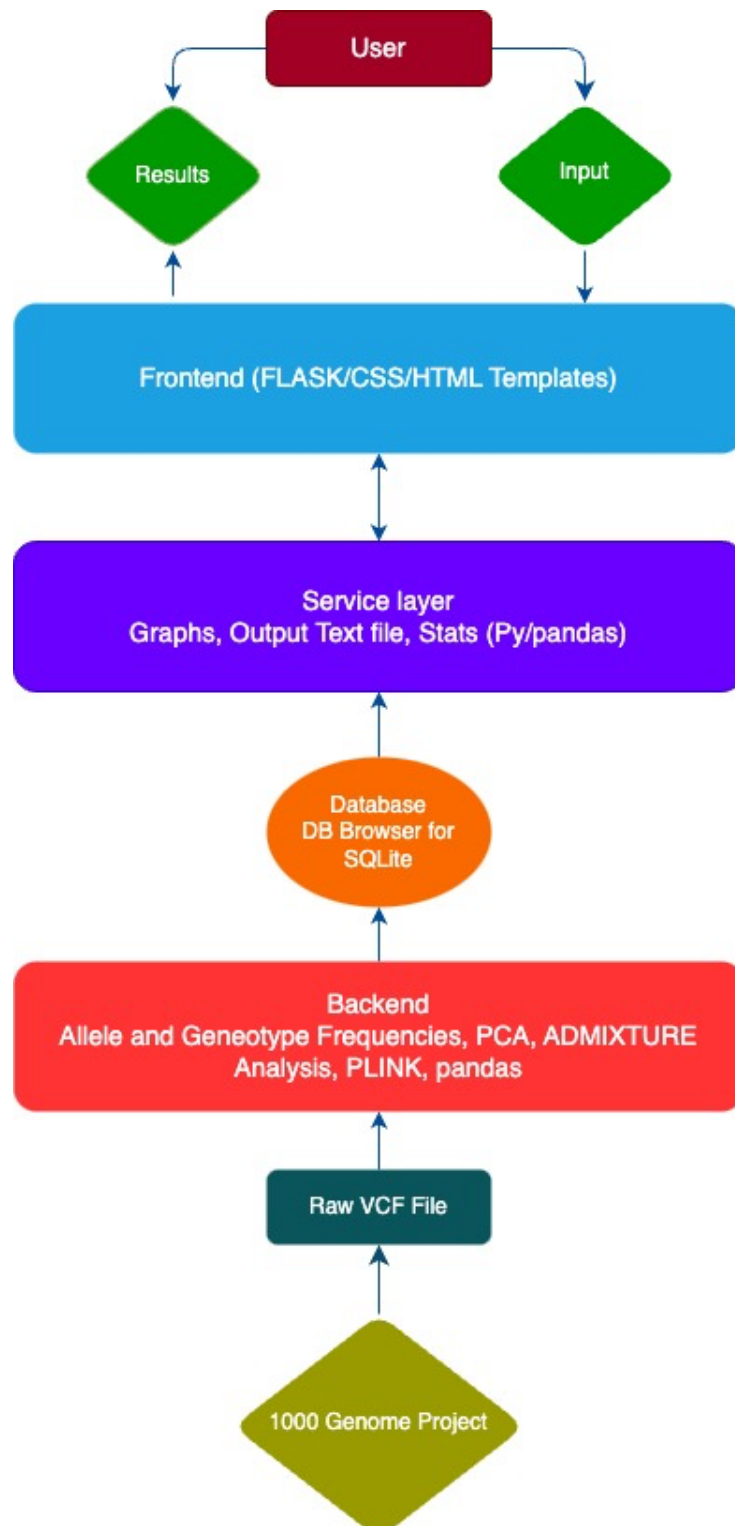


Figure 1. Deepin genetics hub's software architecture and integration. This diagram was produced using drawio.com.

We were given data from a collaborator for a Siberian population as well as whole-genome sequencing data of samples from several human populations from the 1000 Genomes Project database in the form of a raw VCF which was processed to produce TSV files which formed the basis of our database. The final database which holds all essential SNP data, allele, and genotype frequencies. The service layer signifies FLASK application, graph generation and statistical tests performed on the queries drawn from the database and finally the front end which signifies the different HTML templates used to visualise the application.

DATA COLLECTION

SNP Data

We were provided the input data in the form of a VCF file from our collaborator. They obtained genetic data for several markers for 726 human samples from an unspecified location in Siberia the file also contained whole-genome sequencing data of samples from twenty-seven human populations. This was obtained from the 1000 Genomes Project database. In total we have genetic data for almost 4000 individuals, which are specific to chromosome 1 only. Our initial VCF file was first filtered using bcftools so only variants that passed quality control filters were retained. This ensured the retainment of high-quality variants. Some data was still lacking from our initial input VCF file containing SNP and Genotype (GT) data for the 27 populations and almost 4000 individuals. Gene names were sourced and appended to the filtered VCF file using SnpEff (specifically data from GRCh37.87 was used as a reference), roughly half of the SNP IDs were assigned a gene name. SnpEff is designed primarily to predict the effects of genetic variations, particularly single nucleotide polymorphisms (SNPs), insertions, deletions, and other genomic variations, on genes and proteins as published by Cingolani et al. in 2012 [5]. However, in this project, we have utilised SnpEff to add gene name and positions to our initial filtered VCF file. The VCF file was then used to create a new TSV file containing the position of the SNP, SNP IDs and gene names which was inputted into the database.

Allele Frequencies

Allele frequency is the quantitative measure of the relative frequency of an allele on a genetic locus in each population. It is expressed as a proportion of all alleles for the same gene in the population. The allele frequency was calculated by dividing the number of copies of a specific allele by the total number of all alleles for the gene in the population. We used Plink [4] to do this from the filtered VCF file to produce a .bam, .bed, .fam files. Plink was used again to calculate the frequencies. The results from this were exported into a TSV format and super population codes added and incorporated into the database.

Genotype Frequencies

Genotype frequency is the proportion of individuals in a population that possess a specific genotype. This can be calculated by dividing the number of individuals with a specific genotype (either Homozygous reference, heterozygous, or homozygous alternate) by the total number of individuals within that population. Again plink was used to count the numbers off individuals with a specific genotype at each SNP, due to the fact that plink doesn't link up population IDs to each SNP the .bed, .bam and .fam files were first split up into 27 new files (one for each population) and then genotypes were counted using plink, and frequencies calculated using pandas for each of the 27 files and population and super population codes added before being re-joined and incorporated into the database.

DATA ANALYSIS

Statistical tests

The SNP browser enables users to compare allele and genotype frequencies of selected SNPs or genes across two chosen populations. Utilizing pandas for data manipulation and SciPy for statistical computations, it calculates p-values using Fisher's Exact Test in real-time. Given the extensive number of alleles and genes in the dataset, storing all possible comparisons in a database was deemed impractical. Instead, the test is run on demand when a user initiates a query.

For each pair of selected populations or super populations, the allele frequency of one population is multiplied by its population size to generate count data. These counts are used to construct a 2x2 contingency table, considering both the allele frequencies and population sizes, which is necessary for Fisher's Exact Test. The test assesses the significance of the difference in allele counts between the two populations, with the resulting p-value indicating the likelihood of such a deviation if no actual difference existed.

The SNP browser promptly displays the results on the frontend, and users have the option to download the findings as a text file for further analysis.

The FST Comparison Tool offers a comprehensive method for analysing genetic differentiation across various populations. By using Python's pandas and NumPy libraries, it provides a robust platform for calculating Weir and Cockerham's FST analysis, adjusted for population sizes, thus facilitating a deeper understanding of genetic structure and diversity.

Upon loading the SNP allele frequency data, the tool lists the selected populations involved in the study. It uses importance of sample size by incorporating a pre-defined dictionary of sample sizes for each population, ensuring that the FST calculations are weighted appropriately.

This function computes the allele frequencies and their variances while factoring in the sample sizes to adjust for within-population variance. It performs a pairwise comparison of populations, generating a matrix that reflects the genetic variance between each unique pair.

To prevent redundancy and maintain computational efficiency, the tool calculates FST values only for unique population pairs. It then averages these values across the selected loci to provide a single, representative FST statistic that encapsulates the genetic divergence between selected populations.

The final output is a symmetrical matrix, presented in a Heatmap for easy interpretation, which maps out the genetic differentiation of each population pair. This matrix is not only viewable within the Python environment but is also exportable as a text file for further analysis or publication. The FST Comparison Tool provides researchers with a valuable resource for investigating population structure and evolutionary dynamics.

Statistical Test	Description	values
Fishers Exact Test	Statistical significance test used to determine if there are non-random associations between two allele or genotype frequencies between different populations.	A small p-value (typically <0.05) suggests that the observed association is unlikely to have occurred by chance
Fixation Index (FST)	The fixation index, denoted as F_{ST} , is a measure used in population genetics to quantify genetic differentiation or genetic variation among populations. It compares the genetic variability within populations to the total genetic variability across all populations.	The F_{ST} value ranges from 0 to 1, where 0 indicates that the populations are genetically identical, and 1 indicates complete genetic differentiation with no shared genetic variants between the populations.

Table 1. Summary of statistical tests.

Principal Component Analysis (PCA)

One of the features of the website is to produce a PCA plot from the populations selected by the user. This is done using Plotly and the data in the PCA tsv file (table 2). Principle components 1 and 2 were precalculated for the entire dataset using PLINK and stored in the database. Each point on the graph represents an individual within the selected population. The genotype frequency for each SNP is the dimension being reduced. Principle component one and two for each sample were produced directly by using PLINK.

ADMIXTURE

ADMIXTURE was run for a range of different k values between $k=1$ and $k=5$ (see figure 2). The k value with the lowest cv error was used. In our case $k=5$ ($CV = 0.04071$) and as such was implemented into our database with data for 5 ancestries.

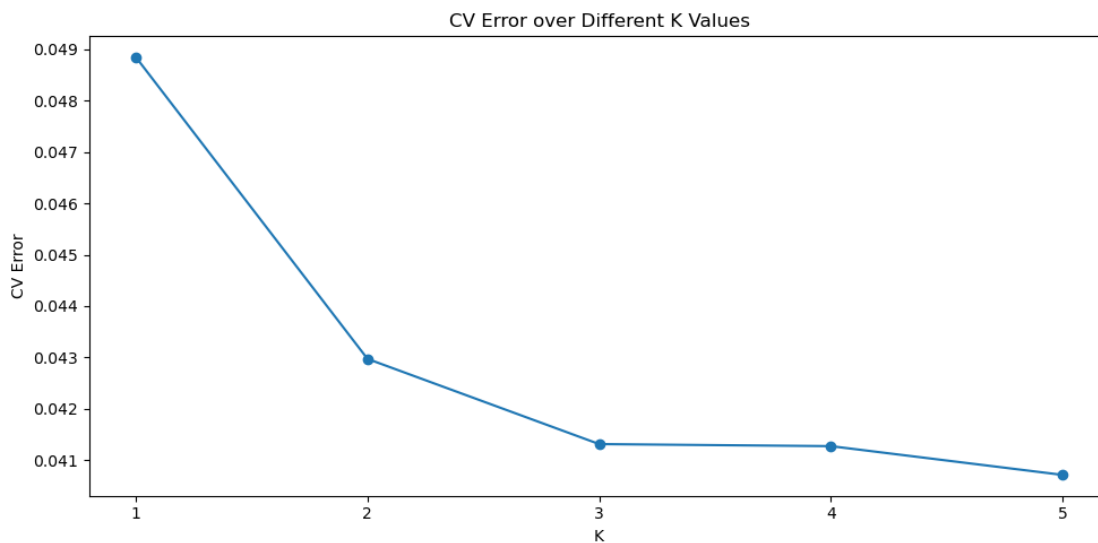


Figure 2. CV error plot of CV error against K Values. We have chosen $k=5$ as it produced the lowest cv error.

DATA SCHEMA

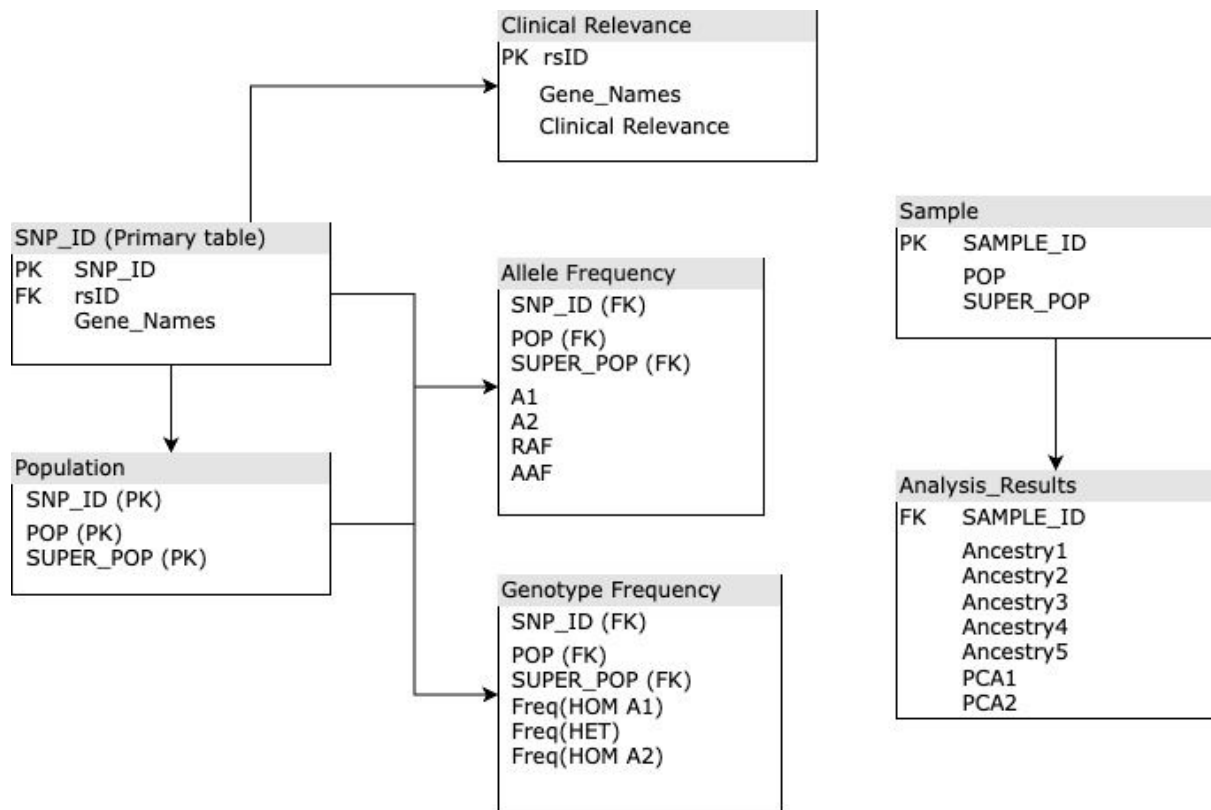


Table 2. Deepin Genetic Hub's Software Database. Each square represents a table of the database and the arrows represent their links. This diagram was drawn using draw.io.

FRONTEND

Website Schematic

Figure 3 shows the overview and connectivity of the website. The structure of the website is designed to be user-friendly and logical. Each section leads to the next, ensuring a smooth navigational flow and a cohesive user experience.

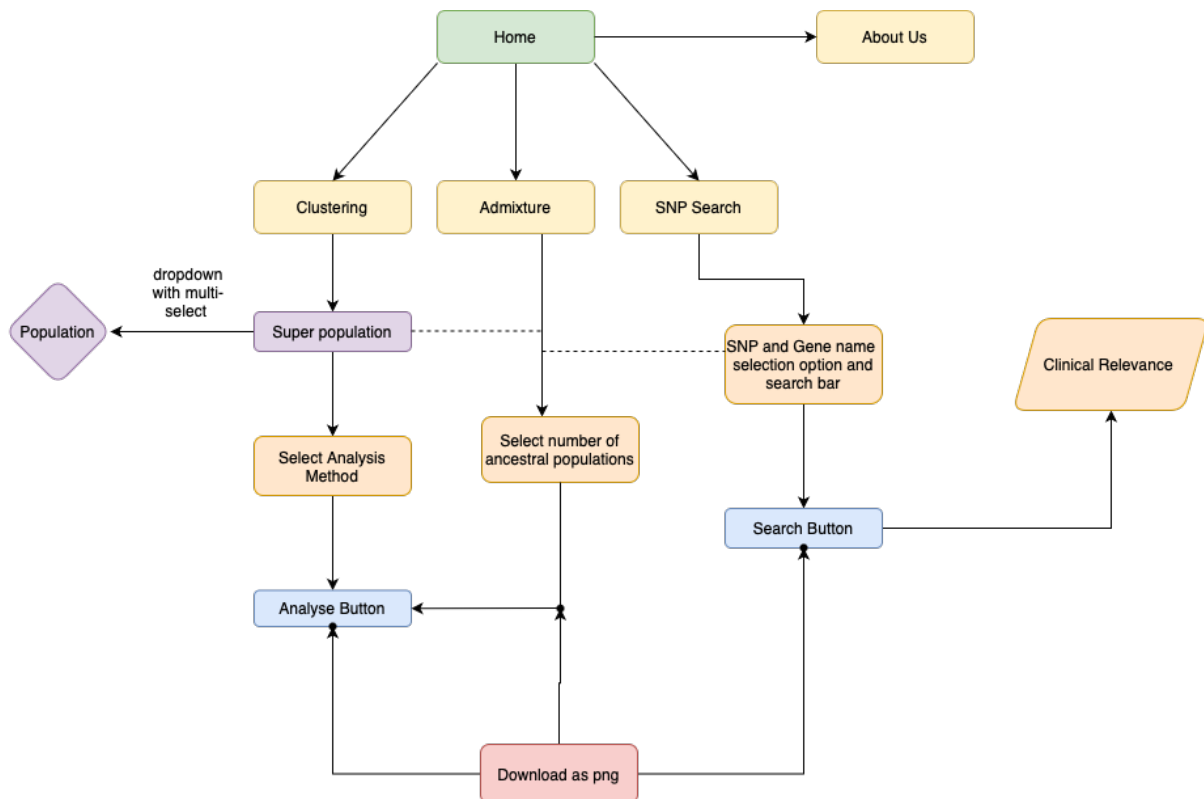


Figure 3. Genetic Analysis Hub website schematic demonstrating the link between webpages and search functions. The dash lines indicate similar feature on the different analysis pages. The flow chart was designed on Diagrams.net.

Website Features

Home

The homepage has a modern and streamlined interface which features a prominent, interactive banner by providing a starting point to access the various analytical tools (Clustering, Admixture and SNP search).

Navigation Bar

The navigation bar is a fixed component present at the top of every page. This ensures users can easily and quickly navigate to any part of the webpage without the need to backtrack to the homepage.

Analysis Types (Clustering, Admixture and SNP Search)

A banner is added to each analysis page for visual aid. The interface for the various analysis types on the website allows the user to select populations under the superpopulation with the option to select multiple populations for analysis using checkboxes. The dropdown arrow allows easier navigation on the website. The button at the end of each form will submit the

selection/input from the user and query the data from our database to generate an output in the form of a graph and clinical information (for SNP search).

Below are the visual representation of our website's analysis webpages.

Clustering Analysis

Deepin Genetic Analysis Hub



CLUSTERING
ANALYSIS

Clustering Analysis

Select Africa Populations

Select America Populations

Select East Asian Populations

Select European Populations


Select South Asian Populations

Select Analysis Method

☐ PCA

Analyse

Admixture Analysis



ADMIXTURE
ANALYSIS

Admixture Analysis

Select Africa Populations

Select America Populations

Select East Asian Populations


Select European Populations

Select South Asian Populations

Number of Ancestral Populations (K)

submit for analysis

SNP Search



SINGLE NUCLEOTIDE POLYMORPHISM (SNP)

SEARCH

SNP Search

Search By

☒ SNP IDs

☐ Gene Names

☐ Chromosome Position

SNP IDs

Select Africa Populations

Select America Populations

Select East Asian Populations

Select European Populations

About Us

The about us section entails information about the project ensuring the user understand the purpose of the website.

Running the website

To run the website on your local machine, clone the BIO727P-Team-Deepin repository and install all the packages in the requirements.txt using “`pip install -r requirements.txt`” The website can be accessed on your local machine’s terminal by navigating into the website’s directory or if you are using Visual Studio (VS) code, upload the BIO727P-Team-Deepin. To run the website, use “`flask run`” in the command line which will generate a URL link – copy and paste the link in your web browser (Google or Safari).

Design and Structure of the Website

The website's interface is designed using HTML for defining the core structure ensuring the content is well organised for easier accessibility. We used CSS to apply style to the layout and visual elements which provide smoother experience for the user. Using JavaScript added interactivity to the website by handling tasks like form validation.

LIMITATIONS AND FUTURE DEVELOPMENT

A possible limitation to this website would be that only half of the SNPs were assigned gene names, which restrains the functionality of the databases to perform other analysis such as gene set enrichment analysis (GSEA). Moreover, SNP search is only focused at comparing populations or super populations, which may not be entirely representative in terms of researching and analysing data. Additionally, in terms of FST, the sample sizes are hard coded, therefore updating the database would be quite difficult because the code has to be changed each time. This would result in the database being inefficient. With the current software architecture and the query optimisation, growth in the database cannot scale up efficiently hence making the software quite slow in terms of giving out results for the user. Another limitation to our data is that the schema in Table 2 demonstrates a lot of repetition leading to large size file. This reduces the time for data query and retrieval making the software less quicker.

For future developments in our website, integration of advanced statistical analyses such as Tajima's D statistical test would help us observe the populations even deeper and give us more knowledge on the biological evolution. We would need to improve on our design responsivity by reducing recurrent data and compressing larger files to increase the speed of our website.

REFERENCES

1. 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>
2. Mahesh, B. A., Kannan, E., Davis, G. D. J., Venkatesan, P., & Ragunath, P. K. (2020). GenPop-An Online Tool to Analyze Human Population Genetic Data. *Bioinformatics*, 16(2), 149–152. <https://doi.org/10.6026/97320630016149>
3. Elias, D. E., & Rueda, E. C. R. (2020). Tools for Evolutionary and Genetic Analysis (TEGA): A new platform for the management of molecular and environmental data. *Genetics and molecular biology*, 43(2), e20180272. <https://doi.org/10.1590/1678-4685-GMB-2018-0272>
4. PLINK-1.9, Package : PLINK [version], Authors : Shaun Purcell, Christopher Chan, URL: www.cog-genomics.org/plink/1.9/ .Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4.
5. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. *Fly (Austin)*. 2012 Apr-Jun;6(2):80-92. PMID: 22728672

