genome analytics

Thursday 3 March 2022

SNP Browser

Christian

Eri

Rym

Yasemin

Table of Contents

About

With the continuous development and improvements of gene sequencing techniques and technologies, genomic data is becoming increasingly available. This allowed for large scale genomic studies, which often form as a large collection of genomic information from various populations, often displayed in variant call format (VCF). Subsequently, the field of population genomics, which investigates the genetic diversity both within and between populations, became progressively noticeable. Importantly, population genomics can provide insights into how and why the observed allele and genotype frequencies change over time.[1] However, most popular genomic databases, such as ENSEMBL, NCBI and UCSC Genome Browser lack a variety of tools to compute population summary statistics.[2,3,4] Therefore, there is a present need for an inclusive platform that not only allows the querying and extraction of genomic information but the opportunity to calculate essential population summary statistics.

SNP Browser is a fast and accurate software that allows for genomic data retrieval of SNPs located on chromosome 22 in conjunction with genotype and allele frequencies across populations. A variety of population summary statistics, which include, Nucleotide diversity, Homozygosity, Tajima's D, and fixation index (FST) can be computed, which are returned in a tabular format and plotted along sliding windows, with the additional opportunity to download the results as a text file. SNP Browser was developed by Genome Analytics, a group composed of four MSc Bioinformatics students from the Queen Mary University of London under the guidance of Professor Conrad Bessant and Dr Matteo Fumagalli.

Design Philosophy

SNP Browser was designed to provide a platform for those who are interested in population genetics; to query SNP data and compute statistics for their queries. SNP Browser allows users to have an uninterrupted and effortless time exploring the genetic differences and similarities between populations. Some central aspects in the design of the application included:

Target

It was critical to design the software tailored to a specific target audience, which included scientists of all levels, from students to professionals. The application was designed to be easily accessible for all; the clean and simple appearance allows for easy search queries and analysis of results.

SNP Data

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans and have been linked to the prevalence of certain diseases, responses to therapeutics, and track inheritance of genetic diseases. Apart from showing the genetic differences between populations, SNP data can also be used to highlight distinct populations with higher disease risks and drug responses.

Population

SNP Browser aimed to provide data on an array of diverse populations; these included the Gujarati, Gambian Mandinka, Japanese, Puerto Rican, and Tuscan populations, which encompass West Africa, East Asia, North America, Southern Europe, and the Indian subcontinent. These populations span the globe and can be inferred as genetically diverse groups. Drawing genetic diversity comparisons from populations from different areas can aid in interpreting whether the genomic area is indeed very conserved or genetically diverse (which may be due to population).

Ease of Use

Simplicity

Data was returned to the user in a simple tabular format, allowing for effortless interpretation of the results. The scope of the searches allows users to query into specific regions of interest along chromosome 22, by either genomic coordinates, gene name and alias or rsID or easily explore the chromosome less specifically.

Navigation

Moving between multiple routes allowed for easy navigation of the software, all search bars are easily accessible both in the home page content and in the navigation bar. The inclusion of the navigation bar provides users easy access to alternate search bars in the instance that they wish to search based on a different parameter. Easy navigation leads to less frustrated users and a more satisfying experience.

Technologies

Flask, a micro web framework written in Python, was used to build the architecture of the software, this included the routes and navigation between them. Flask was used due to numerous benefits, some include its impressive customisability and flexibility paired with its simplicity when implementing code, which allows for a distinguished environment for a beginner. It also implements Werkzeug, an application library that provides a debugger so errors can be easily rectified, even by beginners. Ultimately, Flask provides the ability to produce multiple solutions for problems permitting more freedom. The Jinja template engine exploited by Flask dynamically builds HTML pages creating better integration between flask and HTML. Additionally, Flask supports extensions and packages as if they were implemented in flask itself, some packages used within Flask include:

- WTForms

WTForms was used within Flask to take user input from the user interface and query the database. Flask allowed for the seamless transfer of the user input from the user interface to the HTML and finally to the Flask application where the query function returned the results.

- Pandas

The Pandas package was used within Flask to manipulate the data frames containing the results so that the user would be able to interpret the results more clearly. Flask was able to send these manipulated data frames to HTML templates for easy rendering.

- Other packages

Other packages included Flask_Bootstrap, Flask_APScheduler, scikit-allel, plotly, FLASK-WTF, and numpy.

Software architecture

Figure 1 describes the key components and the interconnections that form the overall software. The data source layer represents the raw data used in the project; this was collected from the 1000 genomes project in VCF file format. The final database which holds all essential SNP data, allele, and genotype frequencies (created with Pandas) is shown in the data access layer. The backend/service layer of the software signifies the Flask application, in which the user input is treated and filtered to output queried SNP information and can also compute the population summary statistics requested as well as interactive plots. Finally, the visualization layer contains all the different HTML, CSS, Bootstrap, and JavaScript used for the front-end of the application.

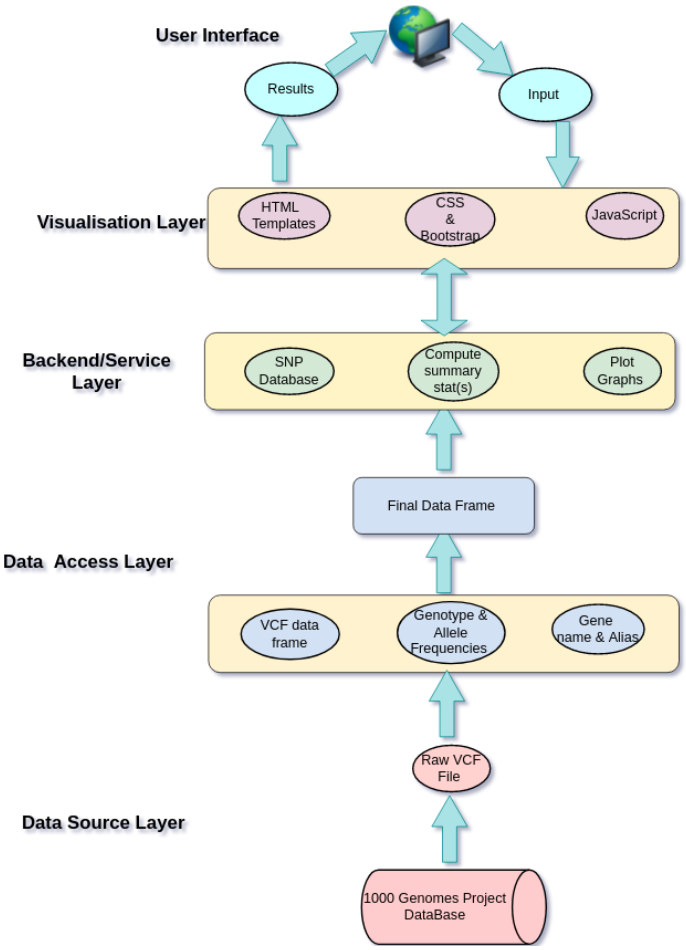

Figure 1. SNP Browser software architecture and integration. Diagram was produced using Diagrams.net.

Website schematic

Figure 2 shows an overview of the website's paths and their interconnections. The Home page has direct links to the three search engines (rsID, genomic co-ordinates and gene name or alias). Each search bar generates a page with various information, including rsID, co-ordinates, gene names and alias, and the observed population allele and genotype frequencies. In addition, the genomic co-ordinate and gene search tools will direct the user to the Stats tool. The stats tool allows the user to compute a variety of population summary statistics and plot the respective graphs. All summary statistics and plots can be downloaded by the user. Internal links are also present within the gene and genomic co-ordinate search, redirecting the user to the rsID profile; external links also redirect the user to the ENSEMBL and NCBI websites for further information.[1,2]

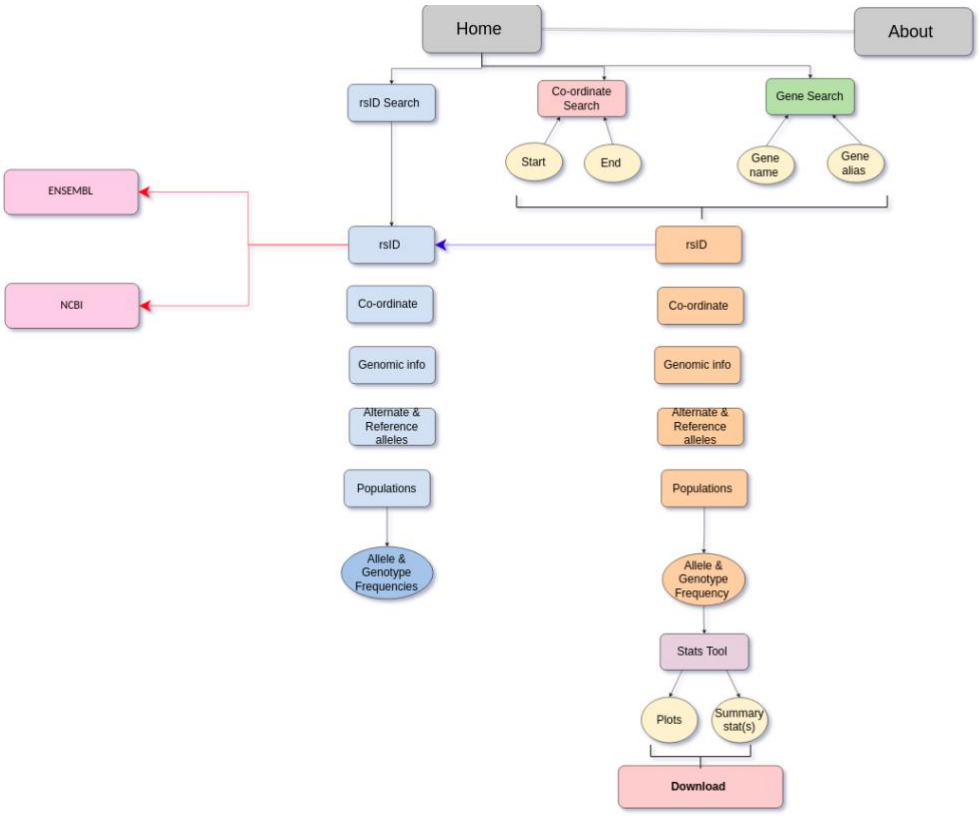

Figure 2. SNPs browser website schematic: the internal hyperlink is represented by the blue arrow whereas the external hyperlinks are represented by the red arrows. Diagram was produced using Diagrams.net.

Running SNP Browser

SNP Browser can be run in the command line, Visual Studio Code and PyCharm, but the command line approach will be followed. To run SNP Browser from your local machine, using the command line move to your Desktop directory and download the project repository. Move into the Software_Development_group folder and install all the required packages from the requirements.txt file (pip install -r requirements.txt). Then move into the website directory and type the following command “python3 application.py”. Copy and paste the URL into your browser (we recommend Google Chrome or Safari), which will direct you to the homepage of SNP Browser.

Structure and Design

A combination of HTML, CSS and Bootstrap 5 was used to deliver a professional user-friendly website. Bootstrap 5 is an HTML, CSS, and JavaScript framework which is often used for developing responsive projects on the web, thus our website exhibits many responsive features which means it is compatible with smaller screens and effectively maintains its usability. Using the Bootstrap framework also allows for fewer cross-browser bugs and supports most major browsers. In some instances, additional CSS was used to format various aspects of the website, for example, producing search fields and forms that appear professional and formatting of the results tables returned to the user. HTML tags were used to define different components of the website.

Data Collection

SNP Data

The 1000 Genomes Project provides a comprehensive description of common human genetic variation across diverse individuals from multiple populations. The initial SNP data was gathered from the NCBI FTP site (https://ftp.ncbi.nlm.nih.gov/), which contains all the variation data obtained by the 1000 Genomes Project in VCF format. Specifically, the phase 3 variant data for chromosome 22 was obtained and utilised to form the basis of the SNP databases, which is available within the 1000genomes section of the NCBI FTP site (https://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz). The VCF file contained variation data, for bi-allelic SNPs, multi-allelic SNPs, indels and structural variants, across all individuals in all populations. Bi-allelic SNP variation data for five populations (Gujarati, Gambian Mandinka, Japanese, Puerto Rican, and Toscani) were isolated using BCFtools, populations with large sample sizes (100+) were chosen to avoid any influence on the reporting of population genetic summary statistics and provide a suitable statistical power.[5] The data contained all vital information, rsID, genomic position, reference, and alternate alleles as well as the phased genotype data.

Gene and Alias Data

Some specific information (i.e. gene and alias names) was lacking from the VCF file. To gather this essential data, the modified VCF file was annotated using the software SnpEff, a genetic variation annotation and functional effect prediction toolbox.[6] The VCF was annotated based on the GRCh37 human reference genome, as the phase 3 data variation data was based on this build. SnpEff appends its annotations within the INFO field of the VCF, consequently, the gene names and their corresponding positions were extracted with the utilisation of sci-kit allel and stored as a data frame. A list of the extracted gene names was input into the HGNC multi-symbol checker, any gene names identified as a previous symbol or alias name were substituted for its current HGNC accepted gene name.[7] To obtain gene alias data, human gene names and aliases were obtained from the NCBI FTP site, (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz). This data was stored in a comma-separated text format and contained all genes and alias data for humans, thus genes and aliases relevant to chromosome 22 were extracted and

mapped accordingly to the HGNC gene list to acquire alias names for the respective HGNC gene name. The gene name and alias data were stored in a CSV file for easy integration into the final database.

Allele frequency data

Allele frequency is a measure of the relative frequency of an allele on a genetic locus within a population and can display the presence of genetic diversity, or equivalently, the richness of its gene pool. The allele frequency was obtained for all SNPs for integration into the final database rather than implementing a calculation during the software runtime for a faster return time of initial results. The allele frequency for each SNP was calculated separately for every population using VCFtools and the results were exported into a CSV format, prepared for the final database.[8] In this instance, as the focus was on bi-allelic SNPs, the reference allele and alternate allele frequency were calculated.

Genotype frequency data

Genetic differences and variations of a given population can also be assessed by measuring the genotype frequencies observed in a population. Genotype frequency can be defined as the number of samples within the population expressing a certain genotype divided by the total population size.[9] The genotype frequency data was computed for all populations prior, for easy integration into the final database. VCFtools was utilised to calculate the observed genotypes with the -hardy function.[8] These generated results were subsequently read into Python and divided by the relevant population size, which was extracted by using sci-kit allel functions on the VCF files.

Database structure

Pandas was used due to its ability and efficiency for data loading, manipulation, information retrieval and format transformation. One main data frame was created containing all bi-allelic SNP data with the observed allele and genotype frequencies for each SNP across each population, as seen in *figure 3*. This data frame was converted into a compressed CSV file and uploaded into Flask as the database.

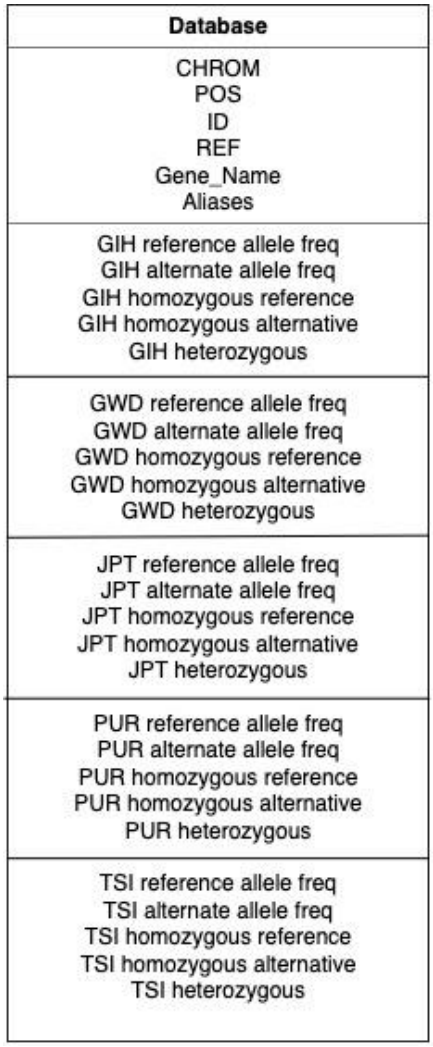

Figure 3.. SNP Browser software Database. Each square represents a section of the database: Genomic information, genotype, and allele frequency for populations; Gujarati (GIH), Gambian Mandinka (GWD), Japanese (JPT), Puerto Rican (PUR), Toscani (TSI). Diagram was produced using Diagrams.net.

Website features

Searches:

The search function allows for the user to query the SNP data in three different ways, one is to query by rsID (SNP ID), gene name/gene alias (due to the ever-updating nature of gene names), and by a set of genomic co-ordinates.

rsID search:

The rsID search allows for the user to query the data based on the input string, the query searches through the ID section of the database extracting the row where the ID matches the string. The output returns tables on SNP information as well as tables with observed allele and genotype frequencies for each population. This search function would be useful to those who know already which SNPs they want to search for allowing for a quick and direct gathering of information.

Gene search:

The gene search allows the user to search the data based on gene names, the query looks through the gene names section of the database returning SNPs and their respective allele and genotype frequencies for each population. Due to the difficult standardised gene name nomenclature, alias names have arisen. Thus, users can query for alias names with a string search within the same form of the gene search. This is useful for those studying specific genes allowing them to view all SNPs within the gene of interest. The alias search adds a greater variety of inputs that could return results and those that reference the same gene in different ways are still able to receive the same information.

Position search:

The position search allows the user to query with a set of genomic co-ordinates, which are input as integer searches into the two separate search fields. The range is utilised to return all SNP data for all populations within the set of genomic co-ordinates input. This feature is useful if the user wishes to explore a larger area of the chromosome than a gene region, with full control of which area they query. The searches are set up to give the user a customisable scope to search if they know what they are looking for in an easy manner. For those who are just browsing for curiosity, the searches enable them to do so freely.

Hyperlinks:

Internal:

Another feature of the application is the use of internal hyperlinks one such example is the transitions between routes via for example the navigation bar and the submit buttons. Another example of an internal hyperlink is in the tables returned via the gene and position searches where the rsIDs are highlighted and once clicked takes the user to the results page for the rsID search of the SNP selected.

External:

To provide users with a comprehensive range of information, which in some cases the website may not contain, external hyperlinks were implemented to provide this. Access to external hyperlinks is provided within the returned rsID results; the position of the rsID contains the external link for the ENSEMBL position viewer, which allows the user to view the area surrounded by the SNP visually. Links to the summary NCBI and ENSEMBL pages are also provided for the relevant SNP.

Stats Tool

Population genomics refers to the study of genetic diversity, both within and between different populations. Drawing analysis from any observed similarities or differences

between individuals from different populations can aid in the inference of historical events, selection, and some aspects of epidemiology. The SNP Browser provides users with the opportunity to compute a variety of inter- and intra- population summary selections, these include an index of genetic diversity, an index of haplotype diversity, one test against neutrality, and one measure of population differentiation.

Data store

For the calculation of all the summary statistics, the phased genotype data for each population was converted to a haplotype array using a sci-kit allel and stored as compressed pickle files. This resulted in a total of five pickle files containing the haplotype data which were then used as the source of data for calculating the statistics. Pickle files are commonly used for object serialisation and are considerably faster and smaller compared to standard CSV files. Thus, less disk storage is used while running the application, which subsequently led to a quicker runtime when calculating the statistics- the runtime was approximately halved when using the pickle files as a data store compared to the CSV files.

Intra-population summary statistics

The Stats Tool provides users with the opportunity to analyse three intra-population statistics. Nucleotide diversity, a measure of genetic variation data, was chosen as an index of genetic variation due to its very simple and intuitive measure, allowing users to draw clear conclusions from the results. To add, nucleotide diversity is accurately estimated, even when calculated with very small sample sizes.[10] Users can also draw inferences from the haplotype diversity with the measure of homozygosity, which was chosen due to its easy calculation as the genotype data obtained from the VCF was already phased. There has been an increasing interest in detecting whether a population is naturally evolving or under selective pressure. Tajima's D has arguably become one of the most famous neutrality tests used for detecting selection, accordingly, this measure is offered for users to calculate.[11] Sci-kit allel was utilised to accurately estimate these summary statistics for the region specified, users are able to select any number of populations or summary statistics at one time and all results are returned. Table 1 outlines the intra-population statistics, a description of the calculation and the inferences that can be made from the values returned.

Table 1. Overview of intra-population summary statistics.

Summary Statistic	Description	Value inference
Nucleotide diversity	The average pairwise differences between all possible pairs of individuals within the population.	Nucleotide diversity values can range from 0 to 1. Values nearer to 0 suggest a lack/no nucleotide diversity present within the region while values closer to 1 suggest a high presence of nucleotide diversity.
Homozygosity	The frequency of homozygous haplotypes within the population, whether this is homozygous for the reference or alternate allele.	Homozygosity values range from 0 to 1. Values nearing 1 suggest a lack of haplotype diversity as the most individuals in the population within the region exhibit a homozygous haplotype. Values nearer to 0 suggest increased haplotype diversity due to the presence of heterozygous haplotypes.
Tajima's D	Compares the average number of pairwise differences among base pairs in a population sample with the total number of variant sites in that sample.	Negative Tajima's D values signifies an excess of low frequency polymorphisms relative to expectation, this indicates population size expansion and/or positive selection. Positive Tajima's D values signifies both high and low frequency polymorphisms, indicating a decrease in population size and/or balancing selection. As a guide, values exceeding +2 or less than -2 are likely to be significant.

Inter-population summary statistics

The Stats tool also provides the user to calculate an inter-population summary statistic, in this instance the fixation index (F_{ST}). F_{ST} measures the divergence between two populations based on genetic structures such as SNPs. F_{ST} scores can vary from 0 to 1, with 0 indicating genetic similarity between the two populations and 1 indicating that all genetic variation is explained by population structure. There are currently many estimators and methods of calculating F_{ST}, these include Weir and Cockerham and Hudson's estimator. In this instance,

Hudson's estimator was elected to be the summary statistic offered to users as it is not dependent of population size, does not exaggerate F_{ST}, and is considered stable in comparison to other estimators.[12] Sci-kit allel was once again used to compute F_{ST}. The user has the choice of selecting more than 2 populations (the user can pick up to 5 populations) and computing fixation indexes of each pair of combinations separately. The function outputs the F_{ST} average over each variant of each population pair. The user can return F_{ST} simultaneously for all population combinations into a readable table.

Data visualisation

To provide users further insights into the population summary statistics returned, and to gain a deeper comprehension of what is occurring within the specified region, plots are generated over sliding windows for all population summary statistics. The plots were generated by utilising the windowed statistic function provided by sci-kit allel and integrating these results into the python package, Plotly.

Interactivity

Plotly provides a range of features that deliver interactive elements to the plots. In this case, range sliders and selectors were selected to obtain a sliding x-axis, which allows zooming of the sliding window positions and a sliding y-axis, which zooms on the summary statistic values. Additionally, various populations or population combinations can be plotted instantaneously within the same graph with their respective colour code, subsequently, individual populations can be isolated with the selector feature.

Download feature of plots

The user also has the option of downloading the plots generated as a png file. Furthermore, the download feature presents the option of capturing a specific sub-region of the plot and downloading this as a png file. This feature is certainly beneficial if the user wishes to utilise the plots within their research, use or further analysis.

Download of summary statistics

Saving of summary statistic results

Statistical test results are automatically saved into the download folder as a text file named "Results.txt". The text file is overwritten once a new set of statistical tests are run. This feature keeps our application files more organised and prevents excessive memory usage for multiple statistical result files. Thus, making the application more memory efficient.

Download features

The user has the option of downloading the "Results.txt" file containing the summary statistics they selected using the download button located at the bottom of the summary statistics page. After the file has been downloaded, the "Results.txt" file is automatically deleted from the download function. This feature prevents any unnecessary memory usage due to its storage, thus making the software more memory efficient.

Auto-deletion

The auto-delete feature deletes all saved summary statistics in the "Results.txt" file that remains in the download folder after 30 minutes from the time they were generated. This was performed using the package Flask_APScheduler, due to its compatibility with Flask. Thus, this feature, prevents an increase of memory usage for the software, allowing more resources for the further execution of the summary statistics.

Limitations

Alias Search

In some cases, the 'Aliases' column of the database contains multiple alias names separated by commas, a query search was unable to identify any of these alias names accurately. A string search overcame this issue by querying the 'Aliases' data for partial-string matches. This allowed for data to be returned based on all aliases present, however, this meant that any partial match could be unintentionally returned. Specifically, if the user was to query 'L', all aliases which contain the string 'L' would be returned.

VCF data

The 1000 genomes VCF data obtained from the NCBI FTP site was built on the GRCh37 human reference genome. The most current build is the GRCh38 build, which has seen several updates over the previous build, some include the repair of incorrect reads, the addition of alternate loci and the inclusion of model centromere sequences. Although the GRCh37 build and insights are offered on numerous other platforms, i.e., ENSEMBL, and meaningful conclusions can still be drawn within this build, using the most updated build of the human genome would provide more useful.

Technical solutions

User experience

Considering the user's experience while using the software is essential to consider during the design. Initially, the user could not navigate between pages or search engines, which would provide for an increasingly difficult experience while using the software. Thus, a fully-working navigation bar was implemented, as well as a variety of internal hyperlinks, allowing for an easier journey.

Design – responsivity

The SNP Browser software can effectively be used on smaller devices and screens while remaining fully functional. The pages exhibit many responsive components, including the home page display, navigation bar, tabular results, and search engines, which aid in providing a software suitable for all devices.

Data validation

SNP Browser utilises the validator parameter for the form classes created this means that the user is unable to put in the incorrect type of input for example, an integer to a string search field. Also, there where conditions placed within SNP Browser that intervened when a the user entered an invalid input the error message and reloaded the webpage rather than taking the user to an error page.

Generated results text file

To allow the users to format and display their statistical results, the results from the statistical test are converted into a text file. This file can be downloaded by the users for further analysis or displayed with R or python.

Controlling application memory usage

To make the application more memory efficient, two measures were implemented to the software. After the "Results.txt" file is downloaded, there is no need to store it within the software, thus this first measure automatically deleted the "Results.txt" file once it's downloaded by the user. The second measure automatically deletes the "Results.txt" file remaining in the download folder 30 minutes after it was generated.

Reducing memory burden from population and database files

Although the size of the database and population haplotype data files would be considered small relative to large scale genomic data, the size of these data stores translate into a longer wait time for the user. Thus, the population haplotype data was serialised into pickle files and compressed, reducing the file size considerably and reducing the run time. The database was also compressed to reduce the memory and disk space required.

Future developments

Updating of the VCF data

As the basis of the database is extracted from VCF data grounded on the GRCh37 build, using more recent VCF data provided by the 1000 Genomes project which is from the GRCh38 build is certainly a reasonable future development to implement for the software. Keeping up to date with human reference builds will keep up with other large databases containing SNP data. Additionally, another advancement could include the user's ability to upload novel SNP data to the database. Following publication checks and validation of the SNP data, this can be implemented to the database available for all users.

Summary statistics expansion

Although SNP Browser offers numerous population summary statistics, providing insights into intra- and inter-population genetic variation, a wider implementation of summary statistics could be provided in future expansions. These could include additional neutrality tests (such as Fu and Li), genetic diversity indices (including Watterson's estimator), and alternative haplotype diversity tests. Other estimators of FST, for example, Weir and Cockerham and Nei estimators, could also be implemented as this could allow the user to

draw multiple concise interpretations. In addition, incorporating another form within the stats tool that also allows the input of window size by variants could be executed, allowing for more personable results.

Refining of Download function

Expansion of downloading function - Every time a statistical test is run, an individual results file would be created, named after the abbreviations of their corresponding statistical test. These then will be displayed as links to be downloaded when clicked. Partial implementation of this development can be observed in the 'del' function, in which rather than looking for Results.txt file only, it looks for all the files within the download folder.

Upgrade the database

Although the current database fulfils the requirement for the software project, the size of the files takes a toll on the querying speed. Thus, it was decided that the software database will be upgraded into an SQLite3 database to improve the speed of querying and to increase the scale by adding other chromosomes.

References:

1. Amorim A. Population Genetics. In: Aaronson S, Abedon S, Adams D, Adhya, S, Aguilera A, Ahn Y et al., ed. by. Brenner's Encyclopedia of Genetics [Internet]. 2nd ed. Academic Press; 2013 [cited 2 March 2022]. p. 407-411. Available from: https://www.sciencedirect.com/science/article/pii/B9780123749840011955
2. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res [Internet]. 2021;49(D1):D884–91. Available from: http://dx.doi.org/10.1093/nar/gkaa942
3. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res [Internet]. 2022;50(D1):D20–6. Available from: http://dx.doi.org/10.1093/nar/gkab1112
4. Ucsc Genome Browser ; Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.
5. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021;10(2). Available from: http://dx.doi.org/10.1093/gigascience/giab008
6. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. 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: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) [Internet]. 2012;6(2):80–92. Available from: http://dx.doi.org/10.4161/fly.19695
7. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.Org: The HGNC and VGNC resources in 2021. Nucleic Acids Res [Internet]. 2021;49(D1):D939–46. Available from: http://dx.doi.org/10.1093/nar/gkaa980
8. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011;27(15):2156–8. Available from: http://dx.doi.org/10.1093/bioinformatics/btr330
9. Lachance, J., 2008. A Fundamental Relationship Between Genotype Frequencies and Fitnesses. *Genetics*, 180(2), pp.1087-1093.
10. Li WH, Sadler LA. Low nucleotide diversity in man. Genetics [Internet]. 1991;129(2):513–23. Available from: http://dx.doi.org/10.1093/genetics/129.2.513
11. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics [Internet]. 2013;14(1):289. Available from: http://dx.doi.org/10.1186/1471-2105-14-289
12. Bhatia, G., Patterson, N., Sankararaman, S. and Price, A., 2013. Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9), pp.1514-1521.