**Abram Roesner**

**AS.410.712.81.SU23**

**07/23/23**

**Final Project: Proposal**

Background

The field of pharmacogenomics has evolved significantly over the past few decades, making it possible to accurately predict the intersection between gene and drug. This offers profound impacts on personalized medicine, especially in fields like psychiatry and psychotherapy, where medication effectiveness can vary widely among individuals and where time and money are often spent in excess while drug stabilization occurs on long timelines.

Despite this potential, accessing personal pharmacogenomic data has proven to be a challenge for many. Existing resources for obtaining this data are either too costly, use proprietary methods, or are difficult for an average person to access. This has created a barrier to the wide adoption of pharmacogenomics in personal medicine. Recognizing this gap, this project aims to provide an accessible and cost-effective solution for individuals to retrieve personalized pharmacogenomic data from currently available resources.

The PharmGKB/CPIC database was developed by the Pharmacogenomics Research Network (PGRN) and funded by the National Institutes of Health (NIH). The database collects, curates, and disseminates knowledge about the impact of human genetic variations on drug responses. It includes clinical guidelines, drug labels, potentially actionable gene-drug associations, and genotype-phenotype relationships. Nonclinical programs such as 23andme, Ancestry.com, or MyHeritage.com offer cheap alternatives to the problem of expensive sequencing. The tool will then match this data to the relevant annotations in the PharmGKB/CPIC database leveraging the Pharmcat program stack to generate a report for users. This report will provide users with personalized insights about their genetic makeup's potential impact on drug responses, which could be used to guide their future healthcare decisions.

A screen shot of a computer

Description automatically generated

Figure 1 Clinvar representation of gene ATK3 and its clinical annotations. Region: 243651535-244014381 GRCH38

Functionality

The tool accepts raw genetic data files from popular platforms, including 23andme, Ancestry.com, and MyHeritage.com. Their format is a .txt file with lists of relevant SNV RSIDs and genotype information. It incorporates an easy-to-use interface, prompting users to upload their genetic data file and select the corresponding file type. After uploading, the backend processes the user data, first converting the provided GRCh37 data to the GRCh38 alignment. This conversion serves as a crucial step to ensure that the user's genetic data corresponds accurately to the standard genome structure.

The aligned data is then converted into a Variant Call Format (VCF) file. VCF is a widely accepted format in bioinformatics, storing gene sequence variations in a standard and compact manner, which allows for effective downstream analysis.

Once the VCF file is created, it is then processed to normalize the data, preparing it for further steps. This preprocessing step includes the cleaning, transformation, and integration of the genetic data to ensure compatibility with the PharmGKB/CPIC database. The processed data is then fed into a modified version of the PharmCAT tool, which matches the provided VCF to named alleles in the PharmGKB/CPIC database. This crucial step bridges the gap between the user's raw genetic data and the clinical/pharmaceutical annotations available in the PharmGKB/CPIC database. Following the PharmGKB/CPIC matching, the tool then associates these named alleles with their corresponding phenotypes. The tool collates all this information, associates it with the relevant clinical guidelines, and constructs a comprehensive report for the user back on the front end.

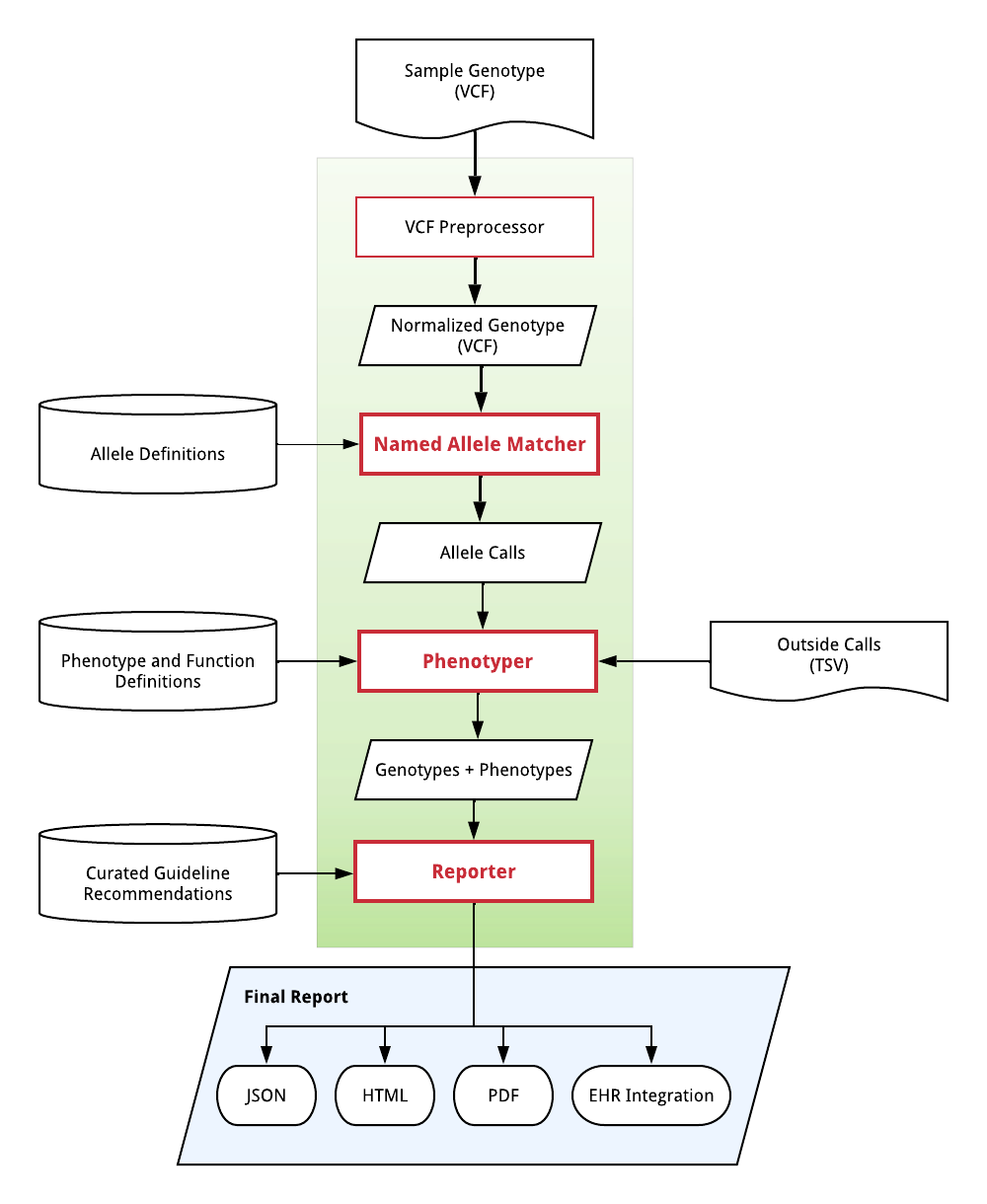


Figure 2 PharmCat function stack (Stanford University & University of Pennsylvania (NHGRI U24HG010862))

Elements

1. **PharmCAT System** – The PharmCAT system is utilized as the primary data retrieval and processing module. It retrieves and processes pharmacogenomic data from the APIs of the Clinical Pharmacogenetics Implementation Consortium (CPIC) and the Dutch Pharmacogenomics Working Group (DPWG), along with PharmGKB, which curates data from these sources.
2. **Python-based back-end** – This forms the necessary elements to bring a tool like Pharmcat to users. It's responsible for managing user data, converting it into a suitable format for processing by PharmCAT. Following the data retrieval and processing the Python back-end organizes this data and creates a detailed HTML report.
3. **CSS/HTML/JavaScript-based Graphic User Interface (GUI)** – This is the front-end of the tool, responsible for facilitating user interactions. Users can upload their genetic data through an easy-to-use interface, which is then sent to the Python back-end for processing through the VCN converter and sent into the Pharmcat pipeline. JavaScript is used for handling file uploads and communicating with the Python back-end, while CSS and HTML structure and style the webpage. The final report generated by the Python back-end is displayed in a user-friendly manner on this front-end GUI.

For data security, it implements stringent measures to ensure user data privacy. All user data is processed immediately upon upload and deleted after the report has been generated, maintaining the privacy and security of user data. For the purposes of this course we will be operating with test data from The Personal Genome Project. **my.pgp-hms.org**

Structure

1. **Web Interface (Frontend)**: The frontend of the tool is deliberately kept simple and user-friendly. It is built using an HTML/CSS/JavaScript stack and consists of an upload button and a dropdown menu. The user can upload their genomic data file and select the file type (23andme, Ancestry.com, or MyHeritage.com) via the dropdown menu. This minimalist design keeps the user interface clean and ensures that the user's interaction with the tool is as straightforward as possible.
2. **File Upload and Temporary Storage**: Once a file is uploaded, it is temporarily stored in a backend folder. This process is managed by the frontend JavaScript code, which handles the file transfer from the frontend to the backend.
3. **File Conversion**: The uploaded file is then transformed into a VCF format suitable for input into PharmCAT. This is achieved using a Python script developed specifically for this purpose. During this process we will also shift our alignment from GRCh37 to GRCh38
4. **Data Preprocessing and Analysis with PharmCAT**: The converted VCF file is then fed into the PharmCAT pipeline using Python. Depending on the options selected by the user, PharmCAT processes the VCF file, matches named alleles to the PharmGKB database, and determines the associated phenotypes.
5. **Report Generation and Presentation**: The processed data is then packaged into a comprehensive report using a combination of HTML, CSS, and JavaScript. This report adopts the PharmCAT native format and includes sections like a genotype summary, prescribing recommendations, allele matching details, and disclaimers. The genotype summary itself consists of three columns - 'Drugs', 'Gene', and 'Genotypes', with subsets for 'Allele' and 'Phenotype'. This is determined specifically by the annotations – and does not require any form of clinical determination on the program’s end.

A screenshot of a medical report

Description automatically generated

Figure 3 example native Pharmcat HTML report

References:

Proposed liftover method for alignment shifts: <https://docs.google.com/document/d/15rxe0iG2kruEWsvBCLyNGof-YRo5T10zuQiBJkUbyJ0/edit#heading=h.uz5vvqa6zm10>

Pharmcat: <https://pharmcat.org/>

The Personal Genome Project: <https://my.pgp-hms.org/>

PharmGKB: <https://www.pharmgkb.org/>

CPIC: <https://cpicpgx.org/>

Note:

This is the idea I’ve had for a useful personal project.

Sadly, this doesn’t completely utilize the database portion of the project recommendations.

Attached is a TSV file that could be used for a much simpler project where user RSID’s are matched to the annotation file from PharmGKB and a list of corresponding links to resources are spit out.

This is a much less fancy/fun project but I understand if this is a better alternative for the course.