Advanced Practical Computer Concepts for Bioinformatics
Final Project Narrative – Genetic variation search tool –

Background

The purpose of this project was to develop a bioinformatics tool that helps users explore gene variants and their associated phenotypes using data from NCBI Clinvar. I chose to develop this tool because it aligned well with my interest in genetics and thought it would be extremely cool to work on a large dataset. The complexity of gene variant data and my desire to create an accessible tool that simplifies the process of querying gene-phenotype relationships for researchers that need quick and effective data exploration was a driver for me to work on this tool.

Project Development

The database schema was designed to include the necessary data fields, consisting of the tables: genes, variants, variant_references, phenotypes and variant_phenotype (**Figure 1**) all of which are aimed to represent the data effectively. The backend uses MySQL to store and query the collected data. To begin the process, a tab-delimited file was parsed and processed using python which allowed the data to be loaded into the normalized tables.

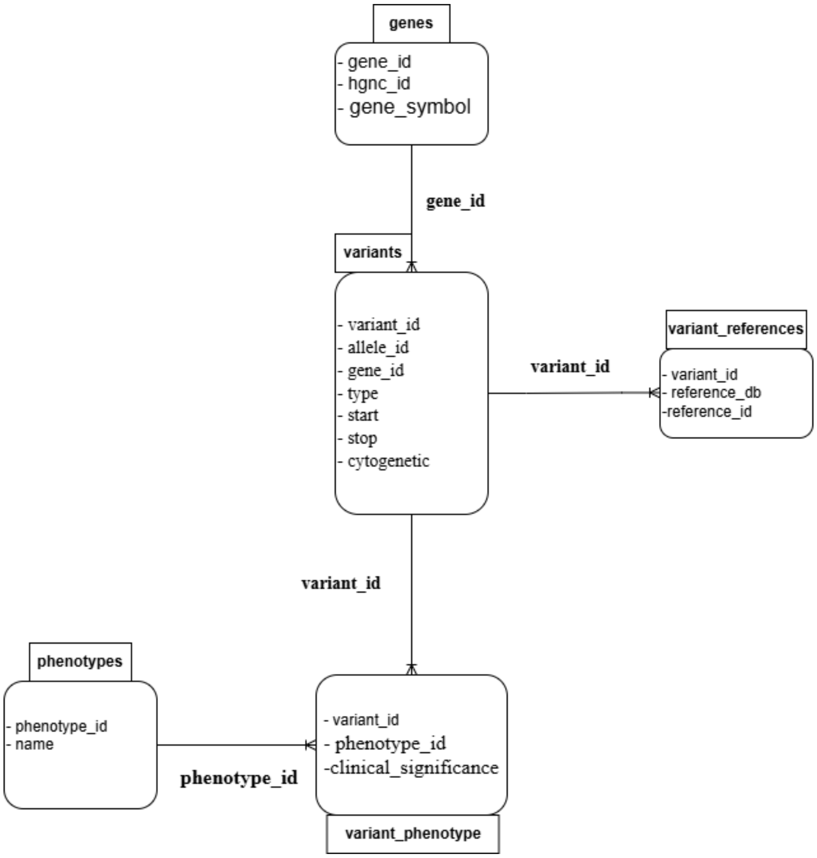

Figure 1. database schema and tables relationships.

The web interface was built using python’s CGI module to make the tool accessible through a web browser where users can either input a gene or a phenotype name to query the database. The returned search results include the following details which are displayed using Jinja2:

- Gene Symbol
- HGNC ID
- Allele ID
- Type
- Start
- Stop
- Cytogenetic Location
- Reference Info
- Phenotype
- Clinical Significance
- Description

JavaScript was integrated into the tool so that the scripts dynamically show warnings based on the user selection from the dropdown menu which ensures that the user is informed about specific considerations when searching based on phenotype selection. Another functionality that was employed using JavaScript was the dynamic fetching of descriptions from the MONDO API and updating the page with it.

Challenges

One of the main challenges that I faced was parsing the initial data file where it was not organized that well and had missing values, compound fields and overall inconsistent formatting. This issue was tackled by trial and error of taking a small subset and running the parsing python script on it then optimizing that script to fix any encountered issue(rinse and repeat until the script ran through the entire data. It took a lot of effort.)

Handling large datasets also was a challenge especially when it came to database queries. When I finished loading the data into my database and started doing test queries, the run time was slow due to the large volume of data in the database. The solution to this problem was to add indexes to frequently queried fields like gene_symbol, phenotype_name and variant_id. I also had to adjust the use of wildcard searches, initially, the wildcard search pattern used was “%s% which matches the search term anywhere within the specified fields leading to inefficient scanning of rows. I changed it to “s%” so it only matches to values that start with the search term (this significantly improved querying the table). Pagination was another addition I made to ensure that only a number of results were displayed at a time to optimize performance (due to the large volume of data loaded per query).

Managing redundant output due to the many to many relationships between variants, phenotypes, references and clinical significance was another issue. This redundancy made the search results cluttered so my solution was to use the group_concat function to properly aggregate related data and avoid duplicates.

Conclusion

In conclusion, the tool developed Is simple and user-friendly, but there are future improvements to be made. One planned enhancement is adding an autocomplete feature which will allow users to easily find relevant information based on partial inputs, making the tool even more intuitive and user friendly. Overall, this project gave me valuable experience in working with large datasets and integrating database design with a web based front end