BLG 564 – Advanced Database Systems
Take Home Final Exam
Spring 2023

1 Background

1.1 DNA

DNA is a long, double-stranded (two chains) molecule that carries the genetic information required for the development, functioning, and reproduction of all known living organisms. It is made up of four different nucleotide bases (adenine, guanine, cytosine, and thymine) that are arranged in a specific order and held together by chemical bonds. The sequence of these bases determines the specific genetic information contained in a DNA molecule.

1.2 Genetic Variants

A genetic variant is a difference in the DNA sequence of an individual's genome (the complete set of an individual's DNA). These variations can occur at the level of a single base pair, or they can involve larger changes, these variations are called structural variants and can be of different types such as deletions, insertion. Genetic variants can be inherited from a person's parents or can arise spontaneously during the course of an individual's life. Some genetic variants have no effect on an individual's health or traits, while others can have significant impacts. For example, certain genetic variants are associated with an increased risk of developing certain diseases or disorders, while other variants may affect an individual's physical characteristics or response to medications. It used to be a very challenging and expensive task to map out an individual's genome, but with technological innovations it is becoming a simpler and cheaper task and more individual’s whole genomic sequences are being mapped out allowing scientists to discover many secrets about human genetics (for example clinical implications of unknown variants). But how are genetic sequences mapped out, it is done through a process called variant calling.

1.2.1 Variant Types: SNPs

SNPs are the most common type of genetic variants, and they make up most of the variant data on our database, these variants occur when a single nucleotide in the DNA (A, C, G or T base) change from one to another, SNPs have can sometimes have no effect on the function of the gene, make it lose its function or give a new function to it.

1.2.2 Variant Types: Structural Variants

There are several types of structural variants out there as well; indels (insertions or deletions), copy number variants (CNVs), inversions, translocations or complex rearrangements, ALUs (Alu repeat) and SVAs (Satellite variant).

- Indels occur when one or more DNA bases are added to or removed from the DNA, these types of variants can have significant effects on gene function depending on the size and location of the variation.
- CNVs involve the duplication or deletion of large DNA segments. CNVs can have particularly significant effects in the case that they remove or duplicate whole genes.
- Inversions, translocations and complex rearrangements involve large scale changes to the structure of the genome and can have impact on the genetic function and cause genetic disorders.
- Alu elements are short repetitive stretches of DNA that are present in the genome and can occur in different places in the genome (sort of move around in it). ALU variations are the changes in these elements.
- Satellite DNA elements are a type of DNA element that is found in many organisms and is characterized by a high degree of sequence repetition. SVAs are variations like insertion, deletion or substitutions that occur in these elements.

Please also see https://en.wikipedia.org/wiki/Human_genetic_variation for more details.

1.3 Variant Browsers

Variant browsers are tools that are used by scientists to search and analyze genetic variant data. They have different functionalities depending on the vendor and help scientists find clinical implications of variants. They typically allow users to view and explore genetic variations in a genome, such as single nucleotide polymorphisms (SNPs) and structural variations, as well as associated annotations and functional consequences. These annotations can highly vary, some annotation examples are, variations frequencies in different ethnic groups, genetic zone of the variant (intron, exon), name of the gene variant occurred in, type of the variant (SNP, insertion, deletion, duplication etc.), clinical implications of the variant, reference and alternative alleles and so on. The browsers can also have visualizers that allow users to move along the genome while observing the changes compared to the reference. Depending on the genome sample specific data can also be observed from these browsers.

2 Data Description

You are provided a database dump of a custom variant browser which allows users to search the variants based on combinations of different criteria along with several visualization capabilities. The database is a relational one implemented in Postgres (PostgreSQL 13.0).

2.1 Database Design

Entities:

- **Variants**: The Variant entity represents a specific genetic variant or mutation observed in the population. It may refer to different types of variations, such as single nucleotide polymorphisms (SNPs), insertions, deletions, duplications, or other genomic alterations. The Variant entity typically includes attributes such as variant ID, chromosome location, position, reference, alternative and other relevant information about the genetic variation.

- **Samples**: The Sample entity represents an individual or a group of individuals from the population who have been analyzed for genetic variants. Each Sample is associated with specific genetic variant data, including the variants observed in that sample. It may include attributes such as sample ID, demographic information, clinical data, and a collection of observed variants.

- **Genes**: The Gene entity represents a specific gene within the genome. It may contain information about the gene, including its identifier, name, position, and other relevant details. Genes play a crucial role in understanding the impact of genetic variants and their potential effects on various biological processes.

- **ReferenceSequence**: The ReferenceSequence entity represents the reference genome sequence against which genetic variants are compared. It serves as a baseline for identifying and annotating variations. The ReferenceSequence entity includes attributes such as reference ID, version, sequence data, and associated metadata.

- **GeographicRegion**: This entity represents the geographic region that an individual lives in.

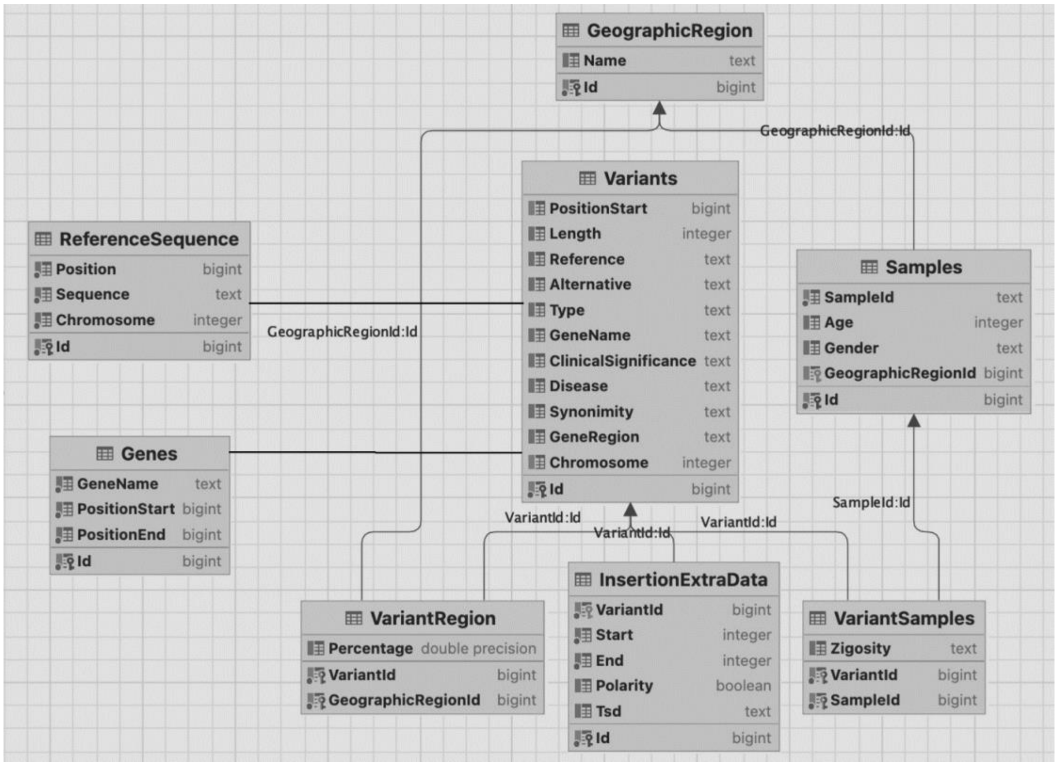

Figure 2.3: Database Design

2.2 Data

The database is populated with data that are obtained from the 1000 Genomes Project (https://www.internationalgenome.org/) to populate the variants table.

Dataset Content:

- The dataset used for evaluating query performance consists of 13 million genetic variants obtained from multiple populations.
- The dataset includes different types of variants. Distribution of the variants are as follows: 11.453.684 SNPs, 1.566.849 INDELS, 4.101 INSs, 4.178 DELs, 475 CNs, 61 INVs.
- The dataset contains relevant attributes for each variant, such as chromosome, position, reference allele, alternate allele, variant type, and additional annotations.
- The dataset contains 4000 genes.
- The dataset contains 3 samples.

3 Project Description

The variant browser executes a number of SQL queries to perform users' search and visualization requests. You are provided a set of most common queries grouped into four different categories. The goal of this project is to improve the query execution performance (in terms of query execution time) of these queries by applying different techniques that you would be designing. To this end:

- You will first measure the baseline performance of these queries on the provided databases.
- Then, try to improve this baseline performance as much as you can. You are encouraged to apply multiple techniques in sequence, and make sure that you are reporting the corresponding improvement after applying each technique.
- You may move the data to another data management system (relational or non-relational) and do a similar study over there. In that case, your baseline will be the initial performance values that you observe on your selected system without any enhancement.
- Please report the improvement amount in terms of percentages as well along with actual values.
- Please use visualizations (e.g., bar charts) as much as possible to report your results.

As a deliverable, please prepare a report that will discuss your applied techniques and experimental results. Please make sure that you also discuss your experimental setting in terms of hardware and software. For each query category, please report the maximum, minimum, median, and average execution times. It would be more proper if you repeat your experiments multiple times and report their averages.

Files:

- Postgres DB dump: https://drive.google.com/file/d/1DTVLG_6W_2M1bbQAcwViJdSEupohywB5/view?usp=drivesdk
- Test Queries: https://drive.google.com/file/d/15iUeNKu6SOgebfB0nL7stjWPzfwDuoTE/view?usp=drivesdk

Submission:

- Please submit a report summarizing your work and results on Turnitin by June 16, 10:00 am.
- You will also give a short demo on June 16 at 12:00 via a Zoom session.