Joshua Conte
7/23/17

DA5020 – Collect, Store, Retrieve Data
Term Project – Project Proposal

I am planning to make a tool designed to extract all of the fields from a VCF file to a database, which will make the data more convenient to work with in downstream analyses.

VCF stands for Variant Call Format. It is a standardized text file format for representing genomic data such as SNP, indels, and structural variation calls and contains the following columns: chromosome, position, identifier, reference allele, alternate allele, quality, filter, info, format, and sample(s). The VCF format is very explicit about the exact type and sequence of variation as well as the genotypes of multiple samples for this variation. With this in mind, there are two big problems with VCF files, there is too much metadata at the beginning of the file and how the data is stored makes it difficult to search and analyze the information.

In order to make VCF files easier to analyze, the tool that I will make will do the following:

1. Remove all metadata and leave only the genomic data with the headers.

2. Organize the data for all of the columns making it easy to search and filter the data for analysis. For example, the database will only include variants that passed the filters under the FILTER column.

3. The tool will also reformat the data for improved analysis. For example, FORMAT and SAMPLE have valuable information. Below is an example of this information is currently presented in a VCF file:

FORMAT	SAMPLE
GT:AD:DP:GQ:PL	1/1:0,2:2:40:86,6,0

   Where:
   GT : The genotype of the sample at the specified site.
   AD and DP : Allele depth and depth of coverage.
   PL : "Normalized" Phred-scaled likelihoods of the possible genotypes.
   GQ : Quality of the assigned genotype.

   The tool will present it like this:

UNIQUE_ID	FORMAT	SAMPLE
1	GT	1/1
1	AD	0,2
1	DP	2
1	GQ	40
1	PL	86,6,0

   With this data in a table, it will make searching and analyzing this information quicker and easier. This concept will also apply to the INFO column in the VCF file.

   Note: Most of the VCF files I use have one sample (I analyze one exome at a time), however, there are some VCF files that have more than one sample. For those files, my tool will omit the FORMAT and SAMPLE columns and only have the first 9 columns. This is common practice with other VCF parsers too.

The final deliverable will put the VCF data into a normalized 3NF database, optimized for parsing the data. I will also include analysis of the data too.