Report on Particular Contributions to the NGS Esophogeal Cancer Project

Ross Eldridge

Abstract:

Ranking and identifying the most likely gene targets was the end goal of this project. To that effect I prepared an initial research presentation analyzing various analysis tools and methods for ranking genes. After that, I spent most of my time organizing the data into a database where we could select for the features we wanted. After one of the tools, wANNOVAR, required VCF files as input, I modified the database to include the VCF data, including a matching of the VCF files with the Excel files, so that accurate VCF files of only the genes we desired could be generated for input into wANNOVAR and other tools.

Research:

A first look at the current field of gene ranking revealed a difference in both methods as well as an amalgamation of tools available for use. Statistical and data mining methods, such as Hidden Markov Models and decision trees, and structural-based methods, characterizing a modified amino acid based on accessibility and affect on protein secondary structure, were explored. Various tools to explore each of these potential techniques were assembled, and the idea of an "ensemble" method, similar to our final choices of Provean and wANNOVAR, was presented as a sort of ideal.

Initial Database:

The first database was set-up based on the *.xlsx files we received, based on output from SAMTools and VCFTools. Given that each file was for a particular patient, each spreadsheet had an extra column added in with a patient ID based on the initial file name, identical for all rows. This was then saved as a *.csv file for import into a mySQL database. The schema for the initial database is shown in **Figure 1** in the supplemental materials.

The data was initially brought into a temporary table to gather all of the patients' information in one location. This was then distributed into the two main tables via REPLACE and INSERT statements. A "mutation ID" was based on a concatenation of the chromosome name, start and end locations, and the initial and modified sequence. The schema was designed so that every distinct mutation would have its own record in the Mutation table, while the PatientMutations table would keep track of which patients had which mutations, and how accurate the reading for that particular patient was (using snp_quality and the depth variables). This database could then survey various combinations of genes and mutations quickly and efficiently with easy-to-write SQL queries.

This database was then, with the help of the BINF department, uploaded to GMU servers for use by the entire group when needed. As well, for those members unfamiliar with SQL syntax and coding, I wrote a short tutorial on SQL using one example from our database to help the other group members write their own queries easily.

Ranking Parameters:

As a group, we then determined a set of parameters and restrictions to begin lowering the number of candidates. Genic regions were restricted to exonic and exonic-related regions, MAF < 0.05, and mutations with commonalities between patients both in exact mutations as well as simply between genes. Later on, given the evidence that this type of cancer is isolated and unique to the region, we also tried removing every mutation that has a pre-existing dbSNP entry to find particularly unknown candidates.

Modified Database:

When we discovered that some tools, such as wANNOVAR, needed a VCF input, I went about modifying the database to include the VCF information in the database, so that data could be output in a VCF format for easy input into those tools. Given that the reference and alternate sequences were treated differently between the VCF and Excel files, I was forced to write a script to concatenate the VCF file onto the Excel file, seemingly in order. To verify that each line was matching correctly, I devised a formula to match the difference in sequence length and start position between files. In the final concatenations, no mismatches were reported, and the sequences were matched accurately. The data were then entered into a database with schema shown in Figure 2 in the supplemental materials.

The columns from the VCF file were distributed amongst the tables accordingly, based on whether the information was unique to the patient or not. Even columns left blank in our VCF files, such as ID and FILTER, were included so that a proper VCF file could later be generated. Given this, we could then write a script to output the same mutations we desired, but as a VCF file we could put into wANNOVAR for further analysis.