**ABSTRACT**

Dean, Taylor. *Final Programming Project*. Master of Science in Human Genetics and Genomic Data Analytics, KGI, Claremont, California.

**BACKGROUND:** Studies show that variants of unknown significance (VUS) have significantly higher detection rates than the detection of pathogenic variants (1). With more direct to consumer devices and the rapidly declining cost of next generation sequencing, we are accumulating a lot of data that does not have known significance. This has strong clinical significance for the treatment of disease. Diseases such as breast cancer have multiple susceptibility genes and with increased rates of VUS, many patients have genetic information without a physician’s knowledge of how to use it for treatment decisions (2). Missense mutations, or the substitution of one base for another that changes the amino acid sequence, are routinely detected in clinical screenings (3) but are useless without pathogenicity. Computational approaches have been used to clinically classify these mutations as benign, pathogenic, or potentially pathogenic (4). This project sought to create a computational analysis of a DNA transcript, randomly mutate it to mimic random mutagenesis, and return the altered amino acid sequence. Once the amino acid sequence was returned, this project sought to classify this mutation as a silent mutation and likely benign or a mutation that would be likely pathogenic (missense or nonsense), and wanted to discover how detrimental this mutation would be based on this information. In this report, we analyze the results of this project.

**METHODS:** This project used Python 3.7 to create a user input driven model for determining variant classification. To create the script, a series of interactive steps were created. First, the script creates a prompt for a user-given sequence, which it will translate into a protein sequence. It will then administer a prompt to randomly mutate the sequence at one base and return the new protein sequence using a series of loops. Once the user informs the program that the sequence changed or did not change, it will use another loop to tell the use that it is likely pathogenic or likely benign mutation and likely pathogenic mutations will interface with the format cyvcf2 to determine potential detrimental qualities.

**RESULTS:** Overall, the results of this project was successful. The python script correctly runs the looping series and user interface. Inputs proceeded correctly. Importing cyvcf2 to my computer did not work on multiple internet connections over a series of days and attempts through pip install in bash and anaconda prompt, and thus the second part of the project had to be scrapped. With more time and troubleshooting, implementing the second half of this script to determine the detrimental aspects of a protein coding sequence change would have potential implications for disease mechanisms and would have been very interesting.

**CONCLUSIONS**: Knowing the clinical significance of variants is important for disease treatment and decisions and characterizing variants of unknown significance will provide much needed information to physicians. This project is an attempt to create a computational algorithm to determine pathogenicity and detrimental qualities. Future work can be done on this project to create a clinically useful bioinformatics tool, though at present it is in its basic form. Limitations of this study include the three-week time constraint and my limited experience (one semester) of coding experience. With more python experience (rather than just one semester), I believe this would be an interesting system to continue to play and work with as a variant classification algorithm.

1. van Marcke C, Collard A, Vikkula M, Duhoux FP. Prevalence of pathogenic variants and variants of unknown significance in patients at high risk of breast cancer: A systematic review and meta-analysis of gene-panel data. Crit Rev Oncol Hematol. 2018 Dec;132:138–44.

2. Ricks-Santi L, McDonald JT, Gold B, Dean M, Thompson N, Abbas M, et al. Next Generation Sequencing Reveals High Prevalence of BRCA1 and BRCA2 Variants of Unknown Significance in Early-Onset Breast Cancer in African American Women. Ethn Dis. 2017;27(2):169–78.

3. Guidugli L, Carreira A, Caputo SM, Ehlen A, Galli A, Monteiro ANA, et al. Functional assays for analysis of variants of uncertain significance in BRCA2. Hum Mutat. 2014 Feb;35(2):151–64.

4. Schulz WL, Tormey CA, Torres R. Computational Approach to Annotating Variants of Unknown Significance in Clinical Next Generation Sequencing. Lab Med. 2015;46(4):285–9.