Coding assessment for the Van Allen Laboratory

Complete the following programming assignment using this dataset. If you are familiar with Python or R, please use one of these languages; otherwise, use a language of your choice. This exercise should not take more than a couple of hours. Please send your source code with your conclusion to Jihye Park (jpark@broadinstitute.org).

In this exercise you will explore cancer genomic data (tumor/normal whole exome sequencing) from 50 patients that received the same type of treatment, half of whom responded.

Overall question: Can you discover any mutations that are associated with treatment response?

1. **Download the dataset linked to above and load the Mutation Annotation Format (MAF) files found in the data/mafs/ folder.** Each of these 50 files contains the genomic mutations observed in a different patient's tumor, obtained by biopsy and sequenced with whole-exome sequencing. Each row in a MAF file corresponds to a different mutation.

2. **Subset for mutations that are not of the Variant Classification "Silent"**. For the purposes of this analysis, we will restrict ourselves to substitutions which result in changes to the produced protein ("nonsynonymous mutations").

3. **Find the 10 most common mutations.** Genes names are included in the column Hugo_Symbol and protein changes are stored in the column Protein_Change.

4. **Perform a statistical test to explore if any mutated genes are enriched in patients who either responded or not.** Response labels for individual patients are found in the file **data/sample-information.tsv**.

5. **Create a scatter plot of genes with the number of mutated patients on the x-axis and your results from question 4 on the y-axis.** Can the axes be scaled or transformed in any way to improve readability? If so, recreate the plot using your suggestion(s).

6. **How many samples are wild-type versus mutant with respect to the most significantly enriched gene from Question 4? Plot the number of nonsynonymous mutations per megabase in the mutant vs. wild-type samples. Is there a significant difference in the number of mutations between the two groups?** Information on the number of nonsynonymous mutations per megabase for each patient can be found in the file **data/sample-information.tsv**.

7. **Write conclusions have made based on your analysis. How might this analysis be improved or expanded upon? Please include all requested figures in your report**.