Rafique, Talha

Introduction:

With the advent of modern sciences and public data, the research community has put together resources that let us analyze data from real patients to come up with various findings and conclusions. One such resource is the public dataset known as TCGA. TCGA stands for "The Cancer Genome Atlas" and is a genomics program that has published over 20,000 primary cancers and matched normal samples spanning 33 cancer types. Using this data which contains information such as treatments, to genes present in cancer patients, and even their lifespans, we can draw conclusions that could help understand the way cancer affects patients.

Within TCGA is the documentation of patients with breast cancer. Breast cancer is an invasive cancer whose new cases are estimated to be in the 280,000 range per year in women. Within the realm of breast cancer there are many factors that impact the recovery and survivability of patients including the genes at play, the age of patients, types of treatment, and many others. By deciphering the many factors that contribute to breast cancer we can hope to find modes of predicting, preventing, or treating breast cancer.

Within this paper we analyze particular cancer related genes that are highly common within patients, and look at which types of mutations they most commonly cause. From this we hope to draw conclusions that could indicate what mutations could occur depending on a particular patient's gene count of particular genes. In turn this could help provide patients with the appropriate medication when a certain gene is flagged within their diagnosis.

Methods:

To begin we need to identify some of the genes that are present across the board within breast cancer patients. We installed three separate libraries: BiocManager, TCGAbiolinks, and maftools in order to work with our data. We also pulled in the "Gene Expression Quantification" data from the TCGA-BRCA database and utilized this data to determine which genes were present in breast cancer patients and which types of mutations they caused. Lastly we created a MAF object to hold the maf data to run our analysis on.

Once all the necessary packages and data were loaded into the workspace we began to collect the information. First in order to identify the genes that were present in the patients we ran an oncoplot on our maf object. We displayed this in an Oncoplot as well to then know the percentage rate at which mutations were occurring. This in turn provided us with the genes most commonly expressed within breast cancer patients, along with the percentage of mutations.

Based on the information returned we then ran an analysis on each of the top 8 genes with the highest rates of mutations. These genes in order of the highest rate of mutations to the lowest are presented below in the results section. To analyze each of these genes and what type of mutations they resulted in, we used lollipop plots to show the dispersion of the types of mutations. Through this information we were able to draw conclusions on the types of mutations present for each type of gene.

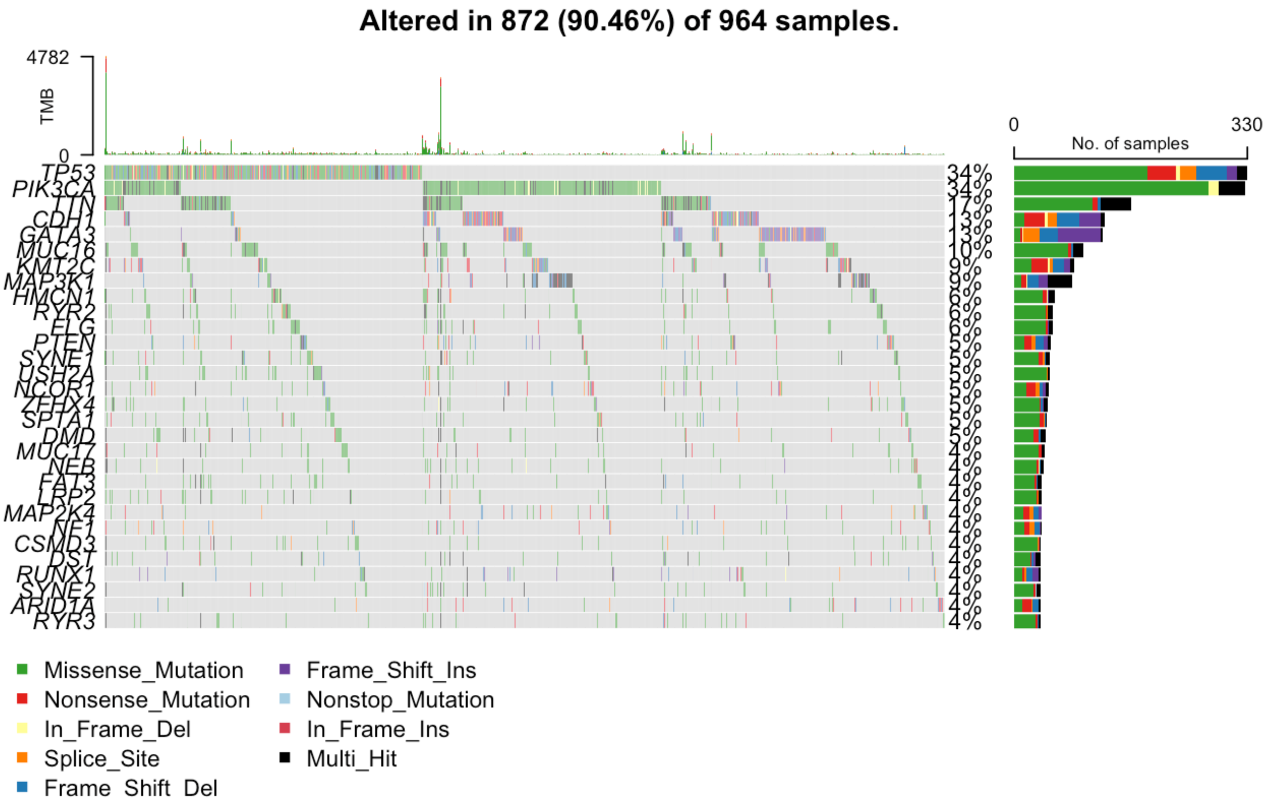


Fig 1: The figure above was generated based on an oncoplot analysis run on ourmaf object. The genes with the highest mutation rates are ordered as the following: TP53 - 34%, PIK3CA - 34%, TTN - 17%, CDH1 - 13%, GATA3 - 13%, MUC16 - 10%, KMT2C - 9%, and MAP3K1 - 9%

Based on the results from the oncoplot above we registered the 10 most common genes that mutate within breast cancer patients, as well as the rate they mutate at. With this data we can analyze each gene individually and map what kinds of mutations occur and how often they occur across genes.

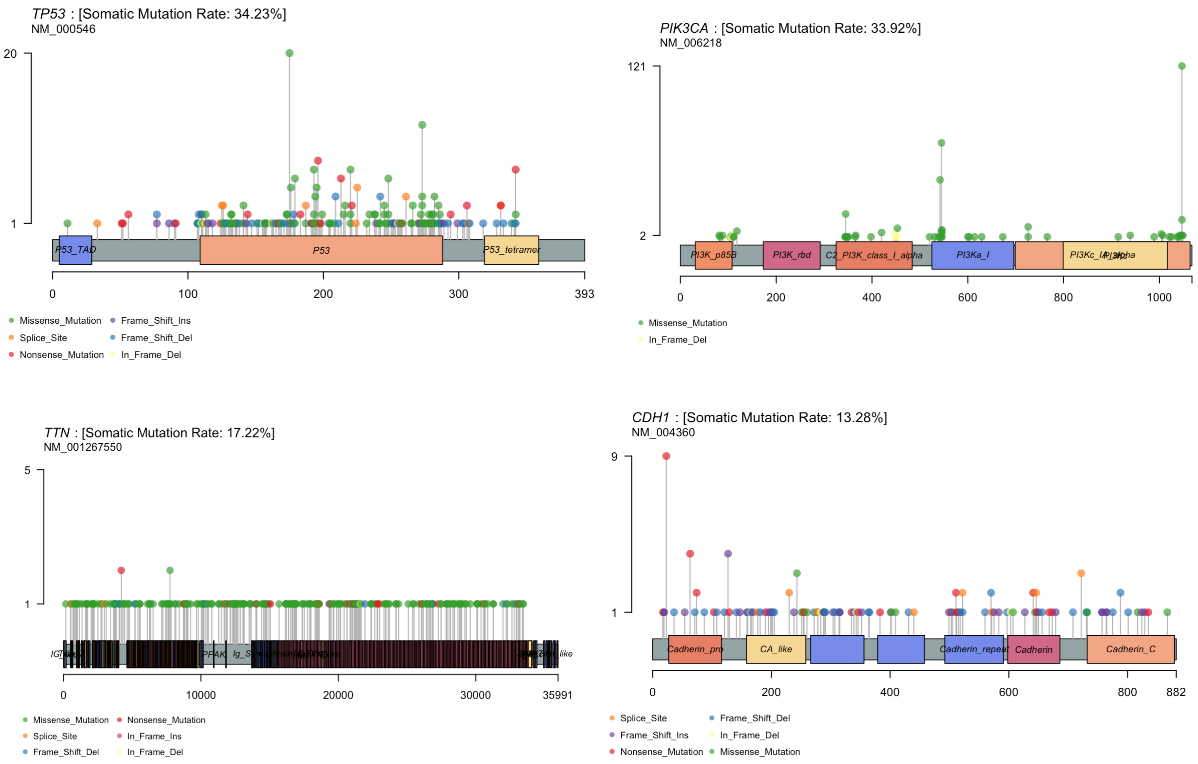


Fig 2: Represents the top 4 genes distribution of mutations. Top Left - TP53, Top Right - PIK3CA, Bottom Left - TTN, Bottom Right - CDH1.

From the figure 2 we can already see a trend of mutation rates. The TP53 gene has the highest rates of mutations, along with the most types of mutation. These gene can be flagged as having a large role in breast cancer occurence. Between PIK3CA and TTN we see a similarity of mostly missense mutations but at different rates and occurrences. And lastly in CDH1 we see fewer missense mutations compared to other types of mutations.

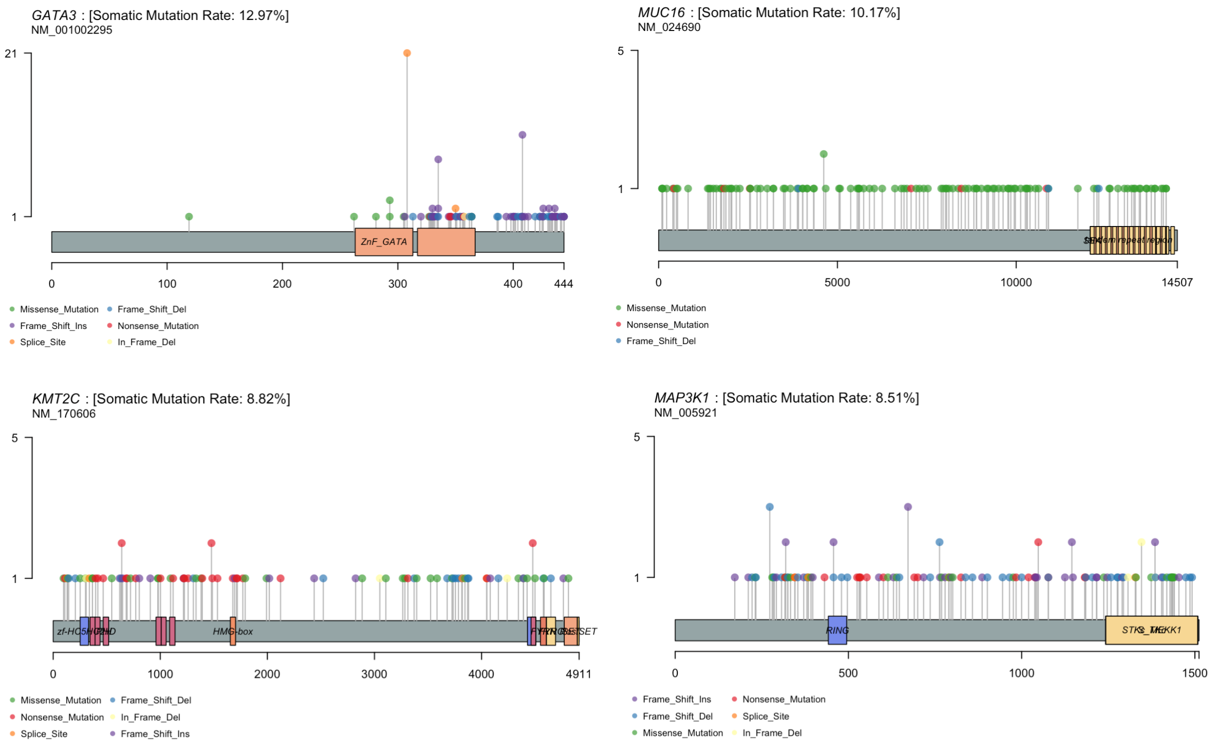


Fig 3: Represents the top 4 genes distribution of mutations. Top Left - GATA3, Top Right - MUC16, Bottom Left - KMT2C, Bottom Right - MAP3K1.

Looking at figure three now with the last four genes we see a different dispersion as well. The GATA3 gene has particularly high frameshift mutations, while the MUC 16 gene is more similar to that of the TTN gene in figure 2 with many more missense mutations. KMT2C as well as MAP3K1 have higher nonsense mutations as well as frameshift deletions.

From the results presented in both figure 2 and figure 3 we can see that the mutations that occur the most often are missense mutations. These also tend not to always have the worst effects but within breast cancer genes could prove detrimental.

Discussion:

Looking at the implications of these genes and the rates of mutations we can see how they can be detrimental and cause higher or more serious consequences of breast cancer.

Starting with the gene with the highest rate of mutation we know TP53 is also the gene that provides instructions for a tumor protein which acts as a suppressor, meaning it regulates cell division. Typically meant for keeping cells from growing and proliferating too quickly, when mutated can result in allowing cells to mutate at an uncontrollable pace.

We can also draw the conclusion that missense mutations occur the most out of any other mutation amongst the 8 most common mutations in breast cancer patients.

Part 3: Review Questions

General Concepts

1. What is TCGA and why is it important?

TCGA stands for “The Cancer Genome Atlas” and is a genomics program that has publicly sequenced and published over 20,000 primary cancers and matched normal samples spanning 33 cancer types. These were all collected from real patients and helped us gain a better understanding of the biology/pathology of various tumors. Specifically in our use case with breast cancer.

2. What are some strengths and weaknesses of TCGA?

The greatest strength is that it is a large patient sample that may be otherwise difficult to procure through independent clinical research. This is also inherently a weakness because the data for various projects is then this same sample. Another weakness is that the data is spotty since almost all the samples are untreated with scattered follow up data. There is much missing data as well. Another weakness is the quality of patients can’t be measured in turn, sometimes not capturing the full picture of the patient's history.

1. What commands are used to save a file to your GitHub repository?

   1. cd into the local repo
   2. git status - to check files changed
   3. git add “name of file or folder”
   4. git commit -m “informative message”
   5. git push

2. What command(s) must be run in order to use a standard package in R?

```
        if (!require(standard package name)){
  install.packages("standard package name")
}
```

3. What command(s) must be run in order to use a Bioconductor package in R?

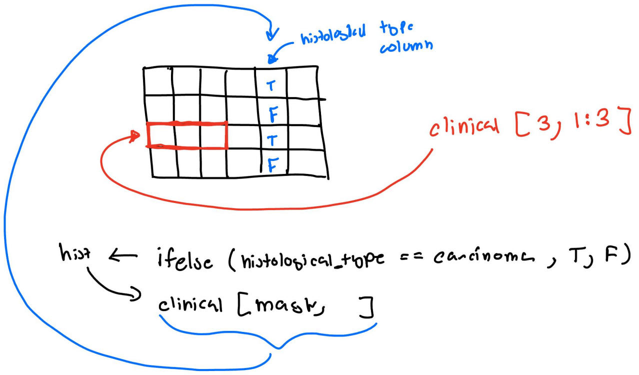
```
        if (!require("BiocManager", quietly = TRUE))
                install.packages("BiocManager")
                BiocManager::install(version = "3.15")
        library(BiocManager)
```

4. What is boolean indexing? What are some applications of it?

Boolean indexing is the practice of using the indices through rows and columns to access the values in a dataframe. By using an indexing approach we can filter the dataframe in many ways as well allowing us access to specific rows and columns.

5. Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.

   a. an ifelse() statement
   b. boolean indexing


histological type column
T
F
T
F
clinical [3, 1:3]
hist ← ifelse (histological_type == carcinoma , T, F)
clinical [mask,]