Chris Ahn
QBIO 490
Midterm Project Part 1

Part 1: Review Questions

General Concepts

1. TCGA (The Cancer Genome Atlas) is a cancer genomic project by the NCI and National Human Genome Research Institute which makes cancer data publicly available for analysis. This data includes genomic, epigenomic, transcriptomic, and proteomic data. It has data from over 20,000 different samples and 33 cancer types which is why this data is very important since it is so large. Analyses performed on this data could lead to correlations which can help researchers find possible treatments for these cancers. Long-term, this data can be used to help prognosis and overall public health for those diagnosed with cancer and those at-risk for these types of cancer.
2. Some strengths of the TCGA is that the data is publicly available, so anyone can access this data and perform their own analyses of it. It also has huge data sizes which increases the significance of these results since the sample sizes are very large. Lastly, the data are from actual cancer patients which means that the results of analysis are directly applicable to the general population. Some weaknesses of the TCGA are that the data is hard to work with sometimes, which means that you have to filter it and clean it before you can use it most of the time. For example, some patients don't have data in certain categories which means you must filter them out before performing the analysis. Also, it sometimes takes a very long time to run a code on the data because the files are very large. Lastly, the data is not very easy to look at, which means that it is hard to see possible trends in the data before you run an analysis of it.

Coding Skills

3. The commands used to save a file to your github repository:
    a. git status (Tells you if there are files which are already pushed to your github)
    b. git add (Saves all files)
    c. git commit -m "<Message>" (Commit the file for saving)
    d. git push (Push changes to github)
4. You should include an if statement that checks to see if you already have a package. Then use the function install.packages("<Name>") to install the package. Lastly, the library(<Name>) function saves the package in your library so you can use the package.
5. Boolean indexing (mask) is a technique used to filter a list from a dataset. The idea behind it is that you make a list that directly matches each element in a particular column which can either have a value of "TRUE" or "FALSE". The true and false depends on the criteria that you are filtering out. You can then create a subset of the dataframe which only includes the rows that have a value for "TRUE" in the boolean mask which means

that you have gotten rid of all of the rows that did not pass your criteria. Some other applications is that you can quickly filter out all the rows that have a value of NA in it by making your condition in the ifelse statement some form of the is.na() function, which returns true if the element is NA in that column.

6.

Student	**Favorite Color**	**Year**
Student A	Yellow	NA
Student B	Blue	Freshman
Student C	Green	Senior
Student D	Yellow	Senior

a. NA_mask <- ifelse(!is.na(df$Year), T, F)

    i. The ifelse has three arguments. The first is the condition (the filter that you are using). In this case, I want to filter out the students that have NA for their year. The second argument is the value that I want to give to the boolean mask if the condition is true. In this case, if the student does not have an NA for their year, they will get the value of TRUE in the boolean mask. The last argument is the value I will give to the student if they do not fulfill the condition, in this case, they have an NA for their year.

b. subset_df <- df[NA_mask,]

    i. The boolean mask is the NA_mask. What I am doing with this line of code is making a subset dataframe which removes the students that have NA in their year (so just Student A in this case). The boolean mask is the same length as the number of rows in the dataframe, so for each student it has a value of TRUE or FALSE. The subset dataframe only includes the rows that have a TRUE value in the boolean index.

Part 2

Is there a correlation between patients that have positive breast carcinoma receptor status and survival rate and certain mutations?

Part 3:

Breast Carcinoma Estrogen Receptor Status in Cancer Patients

Intro:

Breast cancer is the most frequently diagnosed cancer in women worldwide (Sharma, G. et al., 2010). About 80% of the breast cancer patients are over the age of 50, and the death rate has increased in the past three decades (Łukasiewicz, S. et al., 2021). Some risk factors for breast cancer are family history, physical activity level, alcohol intake, density of breast tissue, etc. (Łukasiewicz, S. et al., 2021). Breast cancer can be classified as estrogen receptor (ER) positive or negative. A positive cell can use the hormone estrogen to grow and this is a useful marker of breast cancer since about 50-80% of breast carcinoma cases are positive for ER (Łukasiewicz, S. et al., 2021). This study aims at finding possible correlations between certain clinical variables like age and vital status and gene mutation status with breast carcinoma estrogen receptor status to see if there are underlying explanations for why ER status is a good marker for breast carcinoma.

To do this, data from The Cancer Genome Atlas (TCGA) was uploaded to RStudio to perform multi-omic analysis tests on them. The TCGA was founded by the NIH and National Human Genome Research Institute which provides publicly available data from cancer patients. Multi-omic analysis refers to correlating data from multiple levels of analysis including genomic data, transcriptomic data and proteomic data. An advantage to looking at multiple levels of analysis is that correlations can be drawn which otherwise would not have been found by only looking at one -omic.

A box plot was made to compare the ages of patients that are positive for ER status and negative for ER status. Next, a Kaplan-Meier plot was made to compare the overall survival status for between patients that were positive and negative for ER status. After that, a co-oncoplot was made to compare the rates of certain gene mutations in these patients, and lastly, a volcano plot was made comparing the statistical significance of breast carcinoma receptor status and the genes which are mutated while keeping vital status and stage of cancer constant.

Methods:

For the box plot, the dataframe about the clinical variables of the breast carcinoma patients was used. The patient population was split into two groups (postive and negative for ER status [patients that had NA or Indeterminate for this variable were filtered out) and their ages were compared side by side (Fig.1). The clinical data was accessed from the TCGA with the accession code ("TCGA-BRCA") and the R package ggplot2 was used to make the box plot. Next, a Kaplan-Meier plot was made with the patients also being separated in these same groups (Fig. 2). This plot used the same TCGA data as the box plot and the R package survminer and survival were used to accomplish this. Next, the mutational data dataframe from the TCGA (accession code: "TCGA-BRCA") was used to make a co-oncoplot between patients that had positive ER status and patients with negative ER status (Fig. 3). The R package maftools was used to perform this analysis. Lastly, a volcano plot was made using the transcriptomics data from the TCGA (accession code: "TCGA-BRCA") to compare the rates of breast carcinoma receptor status while controlling for vital status and stage of the cancer (Fig. 4). The R packages DESeq2 and Enhanced Volcano were used to make these analyses.

Results:

Figure 1 showed that on average, the patients with a positive ER status were older than those that were negative (Fig. 1). However, Figure 2 showed that those with negative ER status had a lower survival rate until about 75% when it leveled off. There was a significant difference in rates of gene expression since positive ER patients had PIK3CA mutations 39% of the time (the most common mutation for this group of patients) while negative ER patients only had these muations 16% of the time (Fig. 3). The most common mutation in negative ER patients was TP53 at 89% occurence rate while positive ER patients only had these mutations 21% of the time. Lastly, the volcano plot (Fig. 4) shows which gene mutations are statistically significant between positive ER patients and negative ER patients. Most of the genes are statistically different between these two groups of patients using the p-value of 0.05, which shows that there is a significant difference between the gene mutations in these two groups of patients (Fig. 5).

Figures:

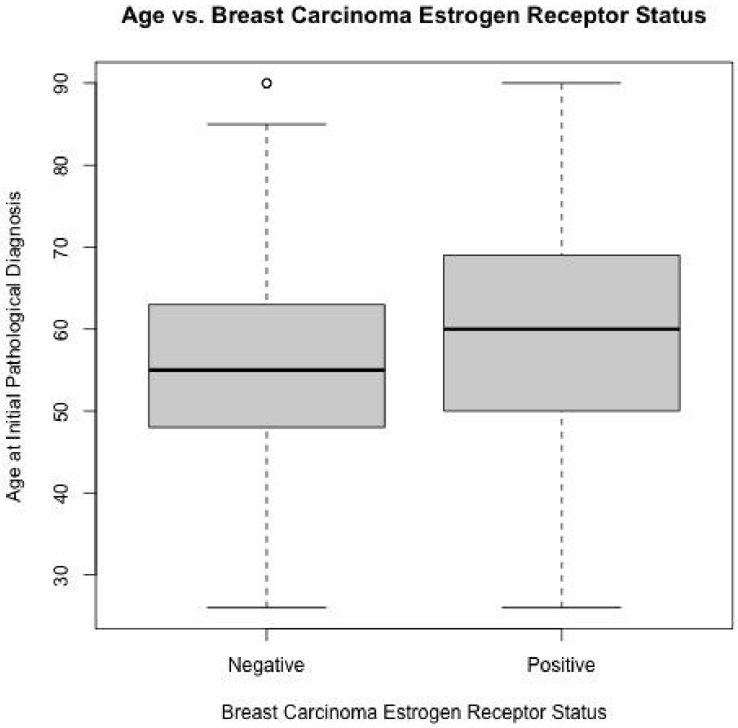

Figure 1. Box plot comparing the age distribution between patients with a negative breast carcinoma estrogen receptor status and patients that are postive for this variable.

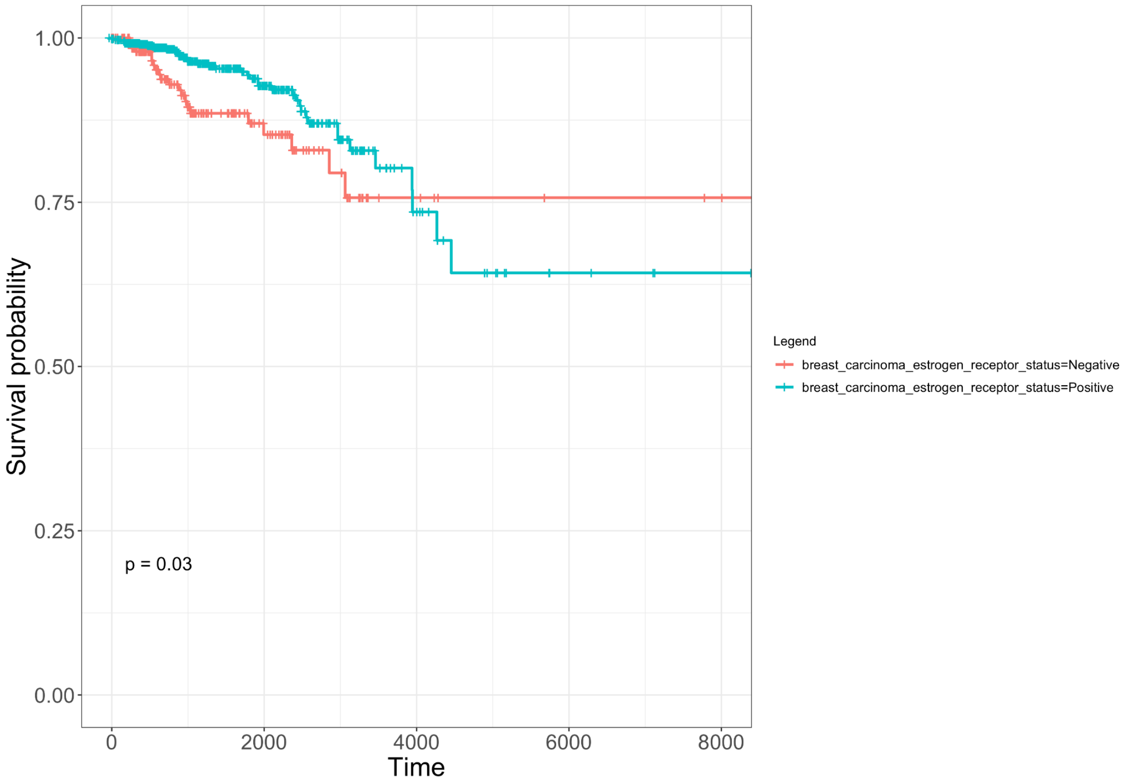

Figure 2. Kaplan-Meier plot showing that patients with a negative ER status (red) had a lower survival rate compared to patients with a positive ER status (blue). However, the survival rate stayed the same in these groups because the data stopped being collected for these groups. The p-value of 0.03 indicates that these results show statistical significance.

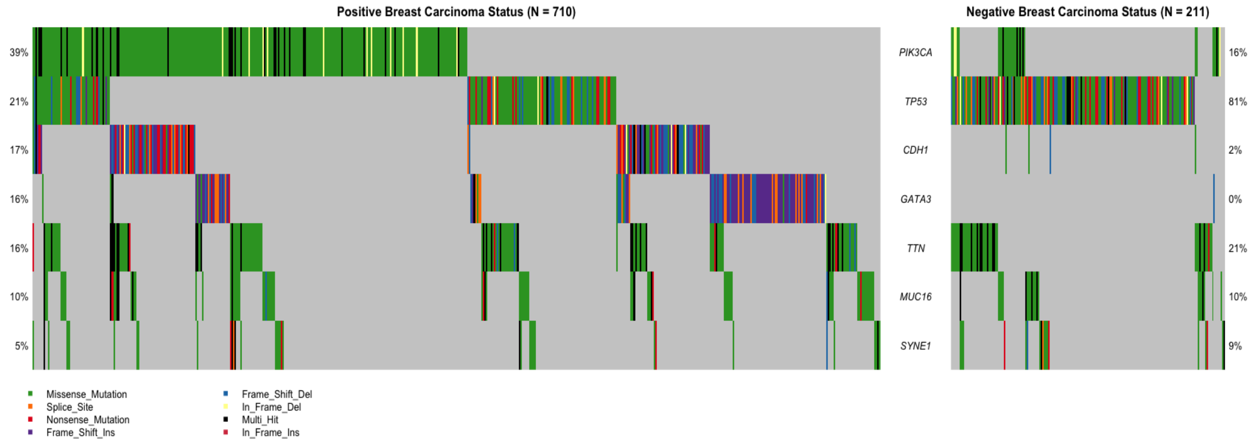

Figure 3. Co-oncoplot which compares the seven most common gene mutations between patients that are positive and negative in their ER status. The most commonly mutated gene in positive patients was PIK3CA (39%) while the most commonly mutated gene in negative patients was TP53 (81%).

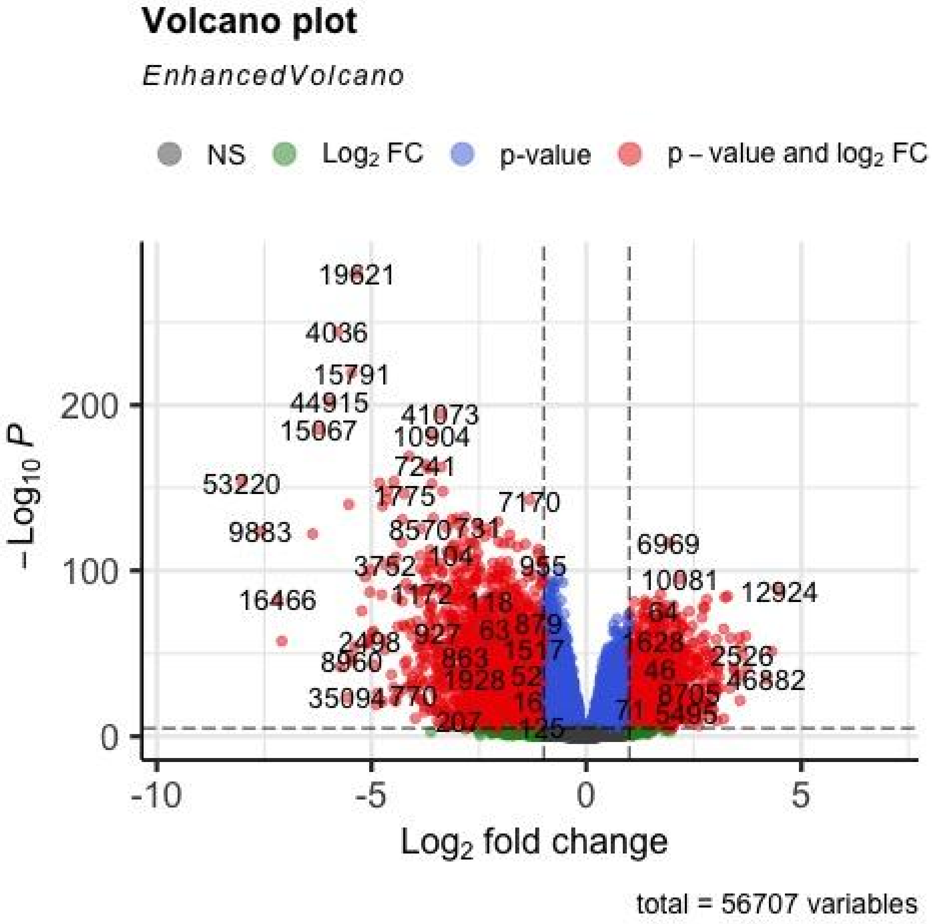

Figure 4. Enhanced Volcano plot showing all of the statistically significant (red) genes between patients that are postive for breast carcinoma estrogen receptor status and negative for breast carcinoma ER status. The blue dots are neither up or down regulated and the green dots are the genes that are not statistically significant.

Discussion:

Although negative breast carcinoma estrogen receptor status patients on average were younger than their positive counterparts, this group of patients were found to have a lower survival rate (Fig. 1, Fig. 2). Negative ER status has already been known to mean that the cancer does not need estrogen to grow and thus is harder to treat (Li, Y. & Brown, P. 2009), so it makes sense that there would be a lower survival rate. Moreover, younger age has been correlated to a worse prognosis in breast cancer patients and this may be due to the increased rates of ER negative status in younger patients (Fu, J. et al., 2019).

Negative ER status also had a much higher rate of TP53 mutation (81%) compared to positive ER status (21%) and this may be the underlying reason for why negative ER status leads to lower survival rates (Fig. 3). This is because TP53, an extensively researched gene, functions to express a tumor suppressing protein that can induce apoptosis, cell cycle arrest, DNA repair, etc. when it detects cellular stresses in the expression of specific genes (Ozaki, T., & Nakagawara, A., 2011). A mutation in this gene can render the protein functionless, which can greatly increase the rates of cancer. So negative ER status is highly correlated to TP53 mutations which is a known cause of cancer. PIK3CA is also known to be correlated to cancer since PIK3CA is a gene that encodes a group of lipid kinases which play a role in cell survival and motility. A mutation in this gene can increase the PI3K (Phosphatidylinositol 3-kinase) signaling pathway which can promote growth factor-independent growth which leads to cancer and metastasis. So greater rates of mutation in this gene have already been correlated to cancer occurence (Ligresti, G. et al., 2009), which is interesting since positive breast carcinoma estrogen receptor status is correlated to higher rates of PIK3CA mutations while negative ER status is correlated to TP53 mutations (Fig. 3). This may suggest that positive breast carcinoma ER status is caused by a mutation in different genes than negative breast carcinoma ER status, which could explain why the survival rates between these two groups of patients are different even though both groups had breast cancer. These findings are also supported by the volcano plot (Fig. 4) since most of the genes are statistically different (p-value < 0.05) between these two groups of patients which could point to the role of differences in gene mutation and expression as the main determinants of breast carcinoma ER status in breast cancer patients.

Future research can focus on why mutations in genes like PIK3CA lead to positive breast carcinoma ER status which leads to a better survival rate early on in the prognosis and why mutations in TP53 lead to negative ER status. This may show the causes for negative ER status, which unfortunately is harder to treat than positive ER status (Bae, S. et al., 2015), and these causes and pathways could lead to possible treatments to help these kinds of patients.

References:

Bae, S. Y., Kim, S., Lee, J. H., Lee, H.-chul, Lee, S. K., Kil, W. H., Kim, S. W., Lee, J. E., &amp; Nam, S. J. (2015). Poor prognosis of single hormone receptor- positive breast cancer: Similar outcome as triple-negative breast cancer. *BMC Cancer*, 15(1). https://doi.org/10.1186/s12885-015-1121-4

Fu, J., Wu, L., Xu, T., Li, D., Ying, M., Jiang, M., Jiang, T., Fu, W., Wang, F., & Du, J. (2019). Young-onset breast cancer: a poor prognosis only exists in low-risk patients. *Journal of Cancer*, 10(14), 3124–3132. https://doi.org/10.7150/jca.30432

Li, Y., & Brown, P. H. (2009). Prevention of ER-negative breast cancer. *Recent results in cancer research. Fortschritte der Krebsforschung. Progres dans les recherches sur le cancer*, 181, 121–134. https://doi.org/10.1007/978-3-540-69297-3_13

Ligresti, G., Militello, L., Steelman, L. S., Cavallaro, A., Basile, F., Nicoletti, F., Stivala, F., McCubrey, J. A., & Libra, M. (2009). PIK3CA mutations in human solid tumors: role in sensitivity to various therapeutic approaches. *Cell cycle (Georgetown, Tex.)*, 8(9), 1352–1358. https://doi.org/10.4161/cc.8.9.8255

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. *Cancers*, 13(17), 4287. https://doi.org/10.3390/cancers13174287

Ozaki, T., & Nakagawara, A. (2011). Role of p53 in Cell Death and Human Cancers. *Cancers*, 3(1), 994–1013. https://doi.org/10.3390/cancers3010994

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010). Various types and management of breast cancer: an overview. *Journal of Advanced Pharmaceutical Technology & Research*, 1(2), 109–126.