Natalie Fortunato
QBIO 490
Midterm
March 10, 2023

Part 1

General Concepts

1. What is TCGA and why is it important?

The Cancer Genome Atlas is a database of multi omic data collected from over 20,000 samples of 32 different types of cancer. It is important because it has led to improvements in cancer diagnoses, treatment, and prevention.

2. What are some strengths and weaknesses of TCGA?

The greatest strength of TCGA is that it is publicly accessible on the internet. This gives an opportunity for many different groups to analyze the data. One weakness is that it only contains data samples of 32 different types of cancer. This is because there was a very specific protocol that needed to be followed in order for the samples to be viable for the TCGA database. Additionally, many values are missing which can make it difficult to perform analysis.

Coding Skills

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

```
git add
git status
git commit - m “”
git push
```

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

```
if (!require(name_of_package)){
  install.packages("name_of_package")
}
library(name_of_package)
```

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 = "3.16")
```

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

R is a vector language, so when the ifelse() function is used, the computer runs down a column or across a row and notes if each value satisfies the argument or not. A boolean vector called a mask is then created with each entry corresponding to each element in the row/column. The dataframe is then indexed using the mask. Only the "TRUE" rows/columns are included in the new dataframe. This is most useful for filtering out data like NA values. It is also helpful in categorizing data. For example, a boolean vector can be used add a column to a dataframe with patients categorized into young and old based on their ages in the dataframe.

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

students dataframe

Student ID #	Age	Ethnicity	Major	Dorm
01	19	White	QBIO	Illium
02	20	Hispanic/Latino	CSCI	Illium
03	18	Black	HBIO	McCarthy
04	20	Black	QBIO	Illium

```
Age_mask <- ifelse(students$age < 20, T, F)
Age mask = [TRUE FALSE TRUE FALSE]
```

```
students <- students[, Age_mask]
```

students dataframe

Student ID #	Age	Ethnicity	Major	Dorm
01	19	White	QBIO	Illium
03	18	Black	HBIO	McCarthy

Introduction

The Cancer Genome Atlas is publically available data set that was collected through the combined efforts of the National Cancer Institute and the National Human Genome Research Institute. It includes more than 20,000 samples of 32 cancer types. The dataset is limited because each sample was collected following a strict protocol. The dataset contains genomic, epigenomic, transcriptomic, and proteomic data which was necessary for multi-omic data analysis. Specifically, this allows for a comprehensive analysis of the correlations between different levels of data such as gene expression which combines genomic and transcriptomic data.

Breast cancer is the most common form of cancer globally and the most common form of cancer diagnosed in American women. In 2020, 2.3 million women globally were diagnosed with breast cancer, 685,000 of whom died. That equates to approximately one diagnosis every fourteen seconds. Because of its global impact, it's important to study the potential causes or identifiers of breast cancer so that improved treatments can be developed.

This research project sought to find a correlation between hormone and chemotherapy treatments compared to other factors such as gene expression, gene mutations, and survival. Several R packages were used to plot data in order to visualize the correlation between different variables. No correlation was found between TP53 and RYR2 gene expression. Patients who received hormone therapy had a higher chance of survival until day 200 when their survival rates remained below that of the chemotherapy group for later dates. The location of RYR2 gene mutations in patients receiving hormone and chemotherapy were virtually the same. No genes were significantly up or down regulated when comparing the hormone and chemotherapy patients.

Methods

BiocManager was used to install the maftools and TCGAbiolinks libraries. These libraries were then used to query all of the necessary data (clinical, drug, radiation, etc.). After masking the RNA counts data of RYR2 and TP53 expression, a scatterplot was created using the built in R plot() function. A Kaplan Meier plot of the survival rates of patients receiving hormone and chemotherapy using the survival and survminer packages after combining the clinical and drug data. The lollipop plot of RYR2 mutation locations of hormone and chemotherapy patients was created using the ggplot2 library and subsetting hormone and chemotherapy data from the clinical MAF file. Finally, the DESeq2 library was utilized to create a volcano plot of gene regulation.

Results

We found no correlation between the expression of RYR2 and TP53 genes as shown in Figure 1. There is a clear clustering of the data, but no significant linear relationship exists.

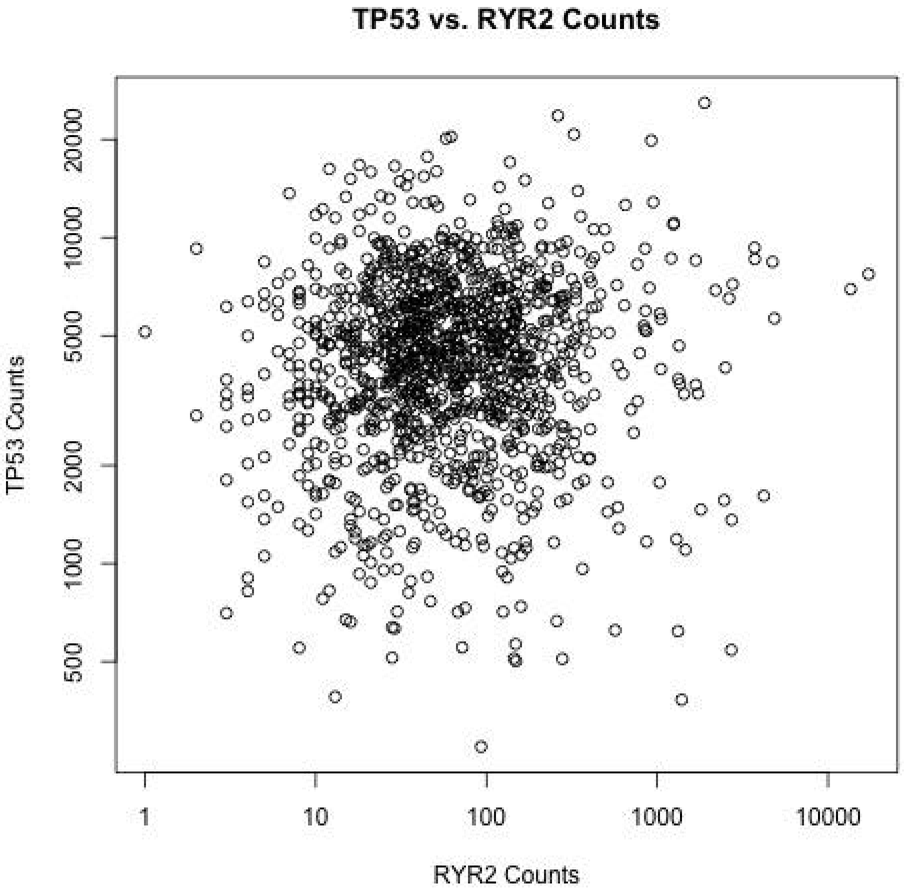

Figure 1. Scatter plot showing no correlation between TP53 and RYR2 gene counts.

Patients receiving chemotherapy had higher survival rates than those receiving hormone therapy until day 200. As shown in Figure 2, after day 200, patients who receive hormone therapy were more likely to survive.

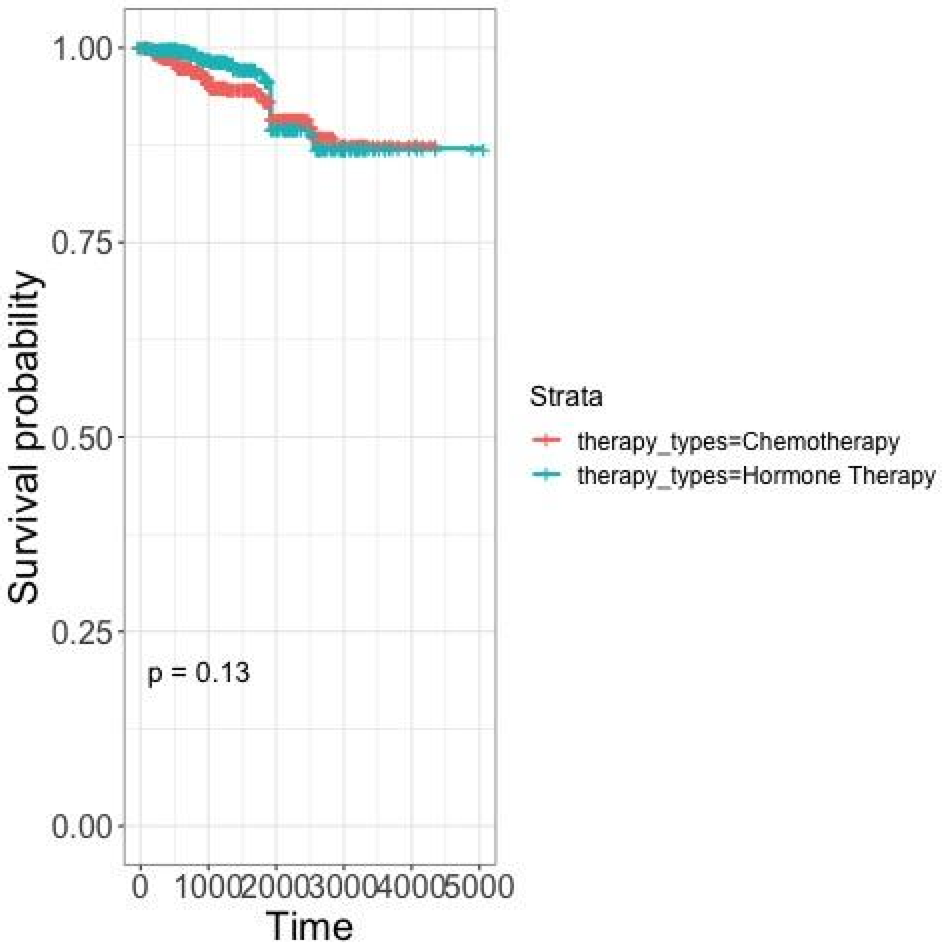

Figure 2. Kaplan Meier plot depicting the survival rates of patients who receive chemotherapy treatments as compared to those who received hormone therapy. P-value of 0.13 indicated statistically insignificant results.

Figure 3 displays a visual representation of the location of RYR2 gene mutations in hormone and chemotherapy patients. No conclusion was able to be drawn from this data as the colollipop plot showed similar results for both groups.

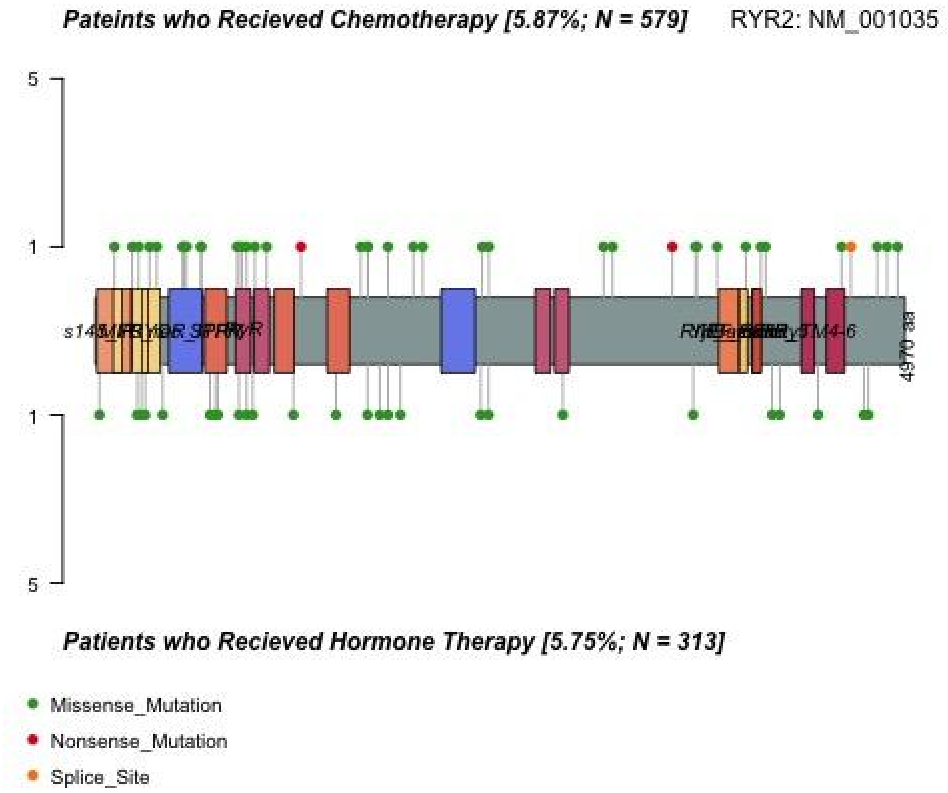

Figure 3. Colollipop plot showing the location of RYR2 mutations in hormone and chemotherapy patients.

No significant up or down regulation of any gene in the hormone therapy patients compared to the chemotherapy patients was observed in Figure 4.

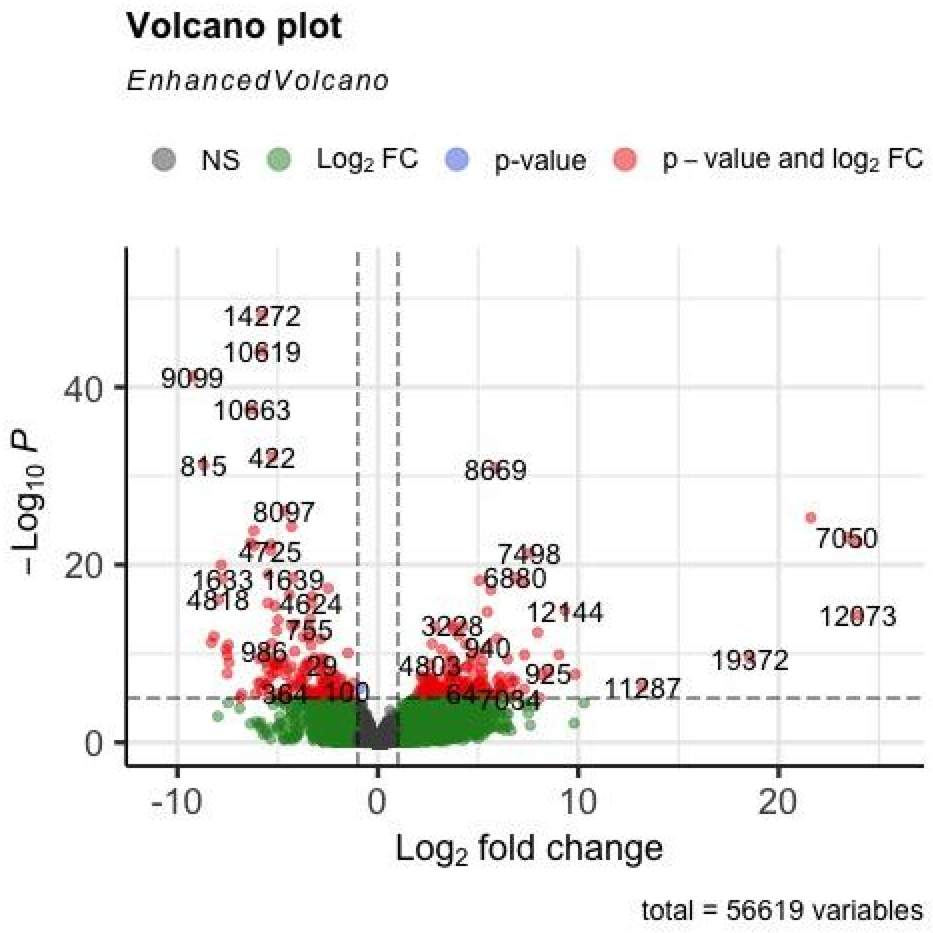

Figure 4. Volcano plot depicting gene regulation in chemotherapy patients compared to hormone therapy patients.

Discussion

No statistically significant results were derived from this experiment to support a correlation between hormone and chemotherapy and transcriptomic data or survival rates. Other research has shown that machine learning can be used to predict survival rates of patients who received hormone and chemotherapy by analyzing tumor gene expression signatures (Mucaki et. al.). Their technology was able to predict survival outcomes with over 73% accuracy which increased for patients who received treatment as opposed to those who elected not to receive treatment. TCGA lacks data on chemotherapeutic changes on tumor tissue, but this information can be helpful in the management of breast cancer, particularly analyzing the effectiveness of a treatment (Kennedy et. al.). The cost effectiveness of hormone and chemotherapy is important to analyze as many people struggle to find affordable health care. Cost effective treatments are also important to allow treatment centers to help the maximum number of patients. The cost effectiveness of hormone and chemotherapy treatments has been shown to rely on the expression of the HER2 gene and age (Diaby et. al.). In future studies, TCGA data can be used to support findings from other studies which analyzed data from other databases. Since there is evidence that gene expression and therapy type correlates to survival outcomes, more studies can be conducted to determine effectiveness of different treatments.

References

Diaby, V., Tawk, R., Sanogo, V., Xiao, H., & Montero, A. J. (2015). A review of systematic reviews of the cost-effectiveness of hormone therapy, chemotherapy, and targeted therapy for breast cancer. *Breast câncer research and treatment*, *151*, 27-40.

Kennedy, S., Merino, M. J., Swain, S. M., & Lippman, M. E. (1990). The effects of hormonal and chemotherapy on tumoral and nonneoplastic breast tissue. *Human pathology*, *21*(2), 192-198.

Mucaki, E. J., Baranova, K., Pham, H. Q., Rezaeian, I., Angelov, D., Ngom, A., ... & Rogan, P. K. (2016). Predicting outcomes of hormone and chemotherapy in the molecular taxonomy of breast cancer international consortium (METABRIC) study by biochemically-inspired machine learning. *F1000Research*, *5*.