Kenneth Nguyen
QBIO 490
10/23/22

Mid-Semester Project

Part 1: Code
Submitted separately

Part 2: Scientific paper

Introduction

Cancer is a condition where damage to cell regulation mechanisms causes proliferation of cells at an uncontrolled rate, causing cancerous growths such as tumors. Breast cancer specifically is a subtype of cancer affecting hundreds of thousands of women and killing tens of thousands of women every year, although it can affect men as well, so understanding how breast cancer works and how it can be treated will allow many breast cancer patients to have more successful outcomes. Mutations in the TP53 gene are a common and well-known cause of breast cancer (Tchelebi et al., 2014) since this gene produces the p53 protein, which regulates the cell cycle. To investigate the TP53 gene further, a multiomic analysis (incorporating multiple biological fields such as clinical data, genomics, transcriptomics, and proteomics). Specifically, accession code TGCA-BRCA data from The Cancer Genome Atlas, an NIH program that provides cancer data publicly in order to support cancer research efforts, was collected and analyzed on RStudio in order to determine the correlation between the TP53 and MKI67 genes as well as what the data suggested about the viability of radiotherapy for breast cancer patients with TP53 mutations. This analysis supports the existence of a biological pathway or mechanism creating an indirect linkage in TP53 and MKI67 expression, although it discourages the use of radiotherapy as a treatment method for breast cancer patients with a TP53 gene mutation.

Methods

The TCGA analysis was conducted by querying clinical and genetic data from the TCGA database (accession code TCGA-BRCA). The R packages BiocManager, maftools, TCGAbiolinks, ggplot2, SummarizedExperiment, survival, and survminer were used to generate all plots in this analysis. Specifically, the clinical dataframe was used to generate the Kaplan-Meier plot for radiation therapy survival, while the clinical radiation dataframe was used to produce the radiation dosage data. The clinical MAF (mutation annotation format) data was used to produce an oncoplot comparing mutation data between the TP53 and MKI67 genes. Lastly, the RNA genes dataframe, derived from the RNA SE (summarized experiment) data, was used to produce a Draftsman plot of four different RNA gene counts, 5-year survival plots for the individual TP53 and MKI67 genes, and a scatterplot comparing TP53 counts with MKI67 counts.

Results

Upon comparing counts for genes related to breast cancer (specifically TP53, TTN, MKI67, and PTGS2), the Draftsman plot (Figure 1) indicates a poor correlation between all gene pair counts except for those of TP53 and MKI67, which shows a moderate positive linear correlation. This linear correlation is clarified by the linear line of best fit on the scatterplot comparing MKI67 counts and TP53 counts (Figure 4). However, for both of these genes, the counts did not depend on the 5-year survival status of breast cancer patients as demonstrated by the paired boxplots (Figures 2 and 3). Furthermore, the oncoplot comparing TP53 mutations and MKI67 mutations shows that there is little alignment in mutations between both genes, TP53 is 17 times as likely to be mutated as MKI67, and both genes have missense mutations as the most common mutation type (Figure 5). The analysis of radiation dosage data from the clinical radiation dataframe shows that the median radiation dosage for radiotherapy is 5040mSv, while the mean radiation dosage for radiotherapy is 4667mSv (Figure 6). Additionally, the Kaplan-Meier plot modeling survival probability over time for patients with and without

radiation therapy indicates that there is no statistically significant difference between survival probability over time for the two treatment groups (Figure 7).

Figures

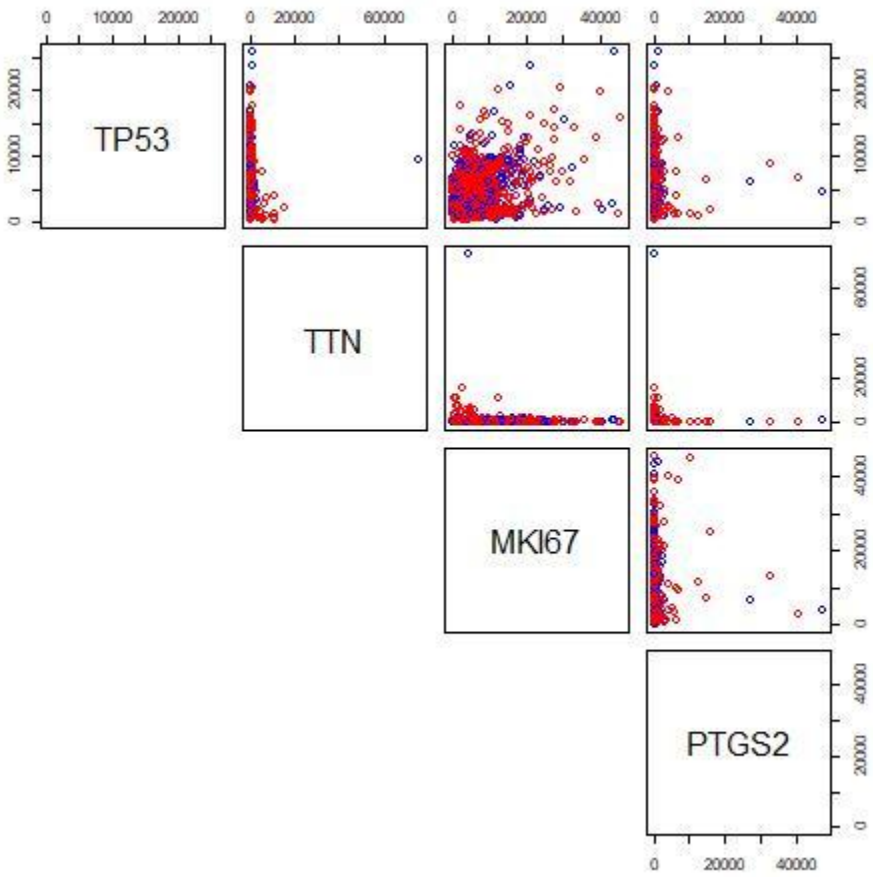

Figure 1. Draftsman plot comparing counts among 4 commonly-mutated breast-cancer genes (TP53, TTN, MKI67, and PTGS2). The TP53 vs. MKI67 plot suggests a moderate, positive linear correlation.

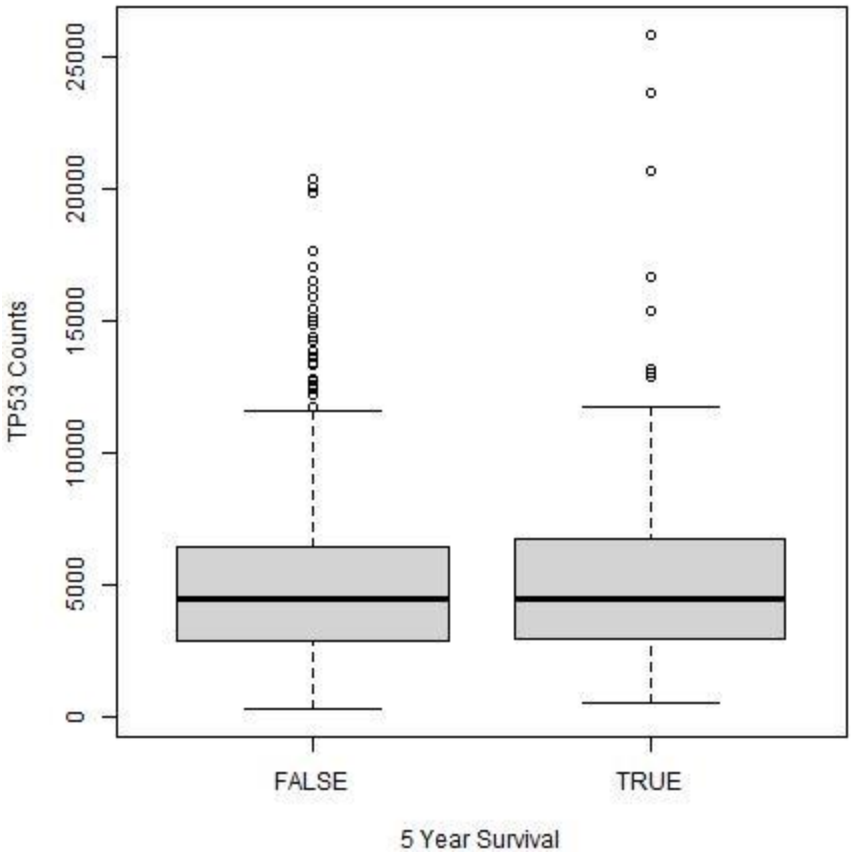

Figure 2. Boxplots comparing TP53 counts for patients who survived breast cancer for at least 5 years and patients who didn't. The large overlap between both boxplots suggests that there is no statistically significant difference in TP53 counts between the two survival groups.

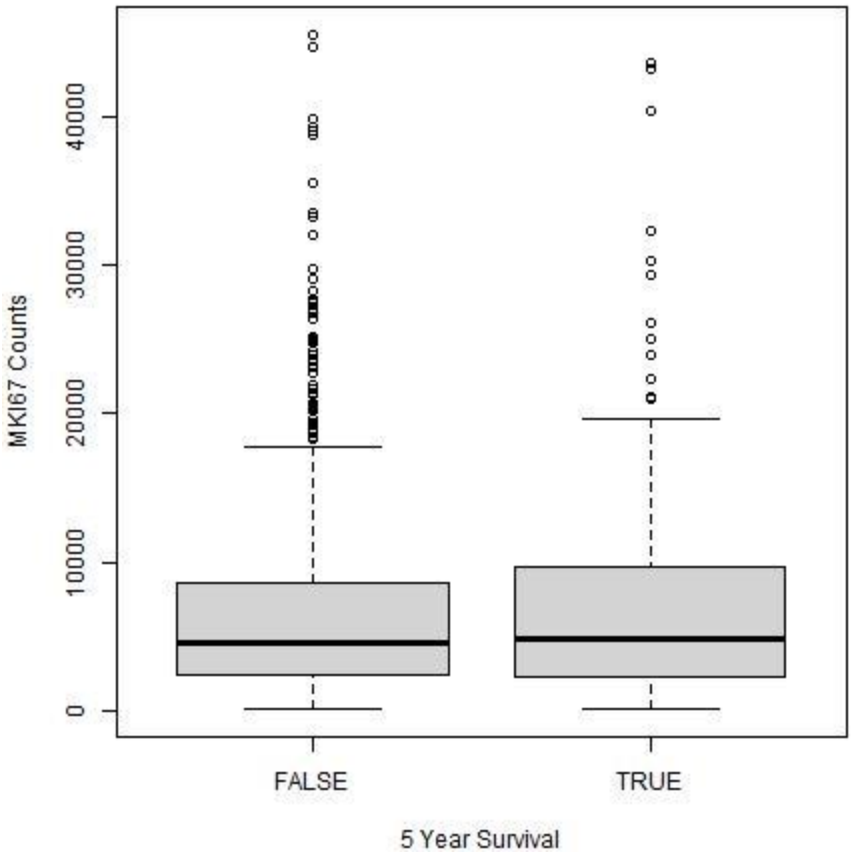

Figure 3. Boxplots comparing MKI67 counts for patients who survived breast cancer for at least 5 years and patients who didn't. The large overlap between both boxplots suggests that there is no statistically significant difference in MKI67 counts between the two survival groups.

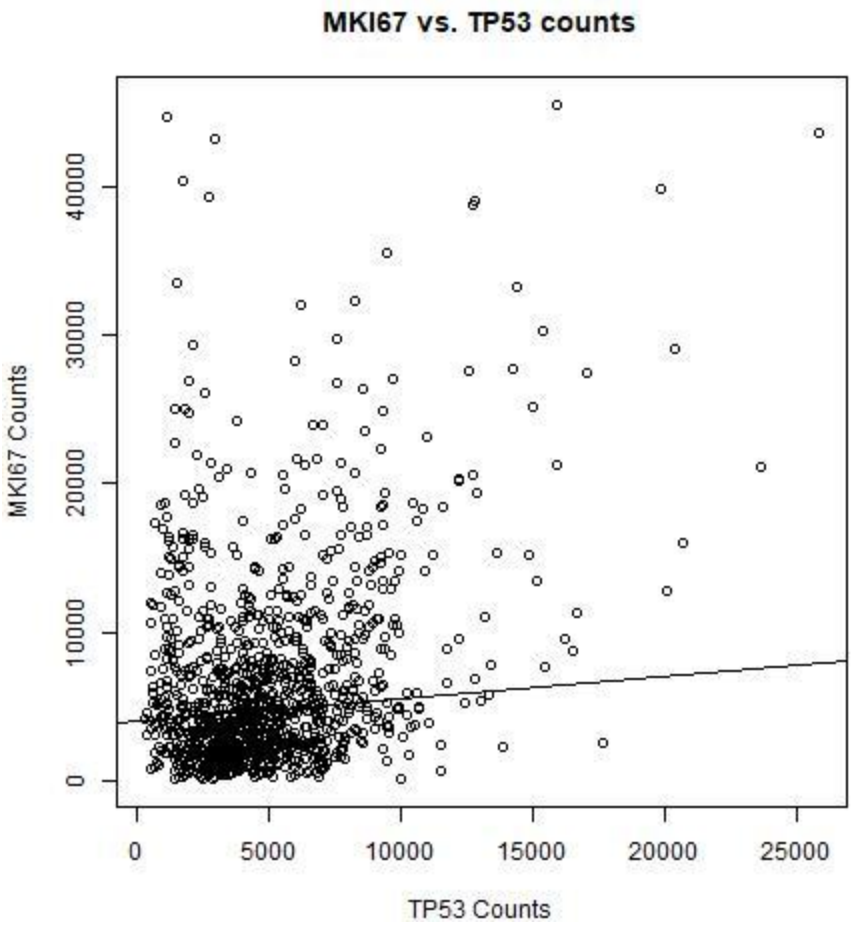

Figure 4. A scatterplot comparing gene TP53 counts and gene MKI67 counts for breast cancer patients. Although there are outliers, there is a moderate but noticeable positive correlation between the counts for both genes.

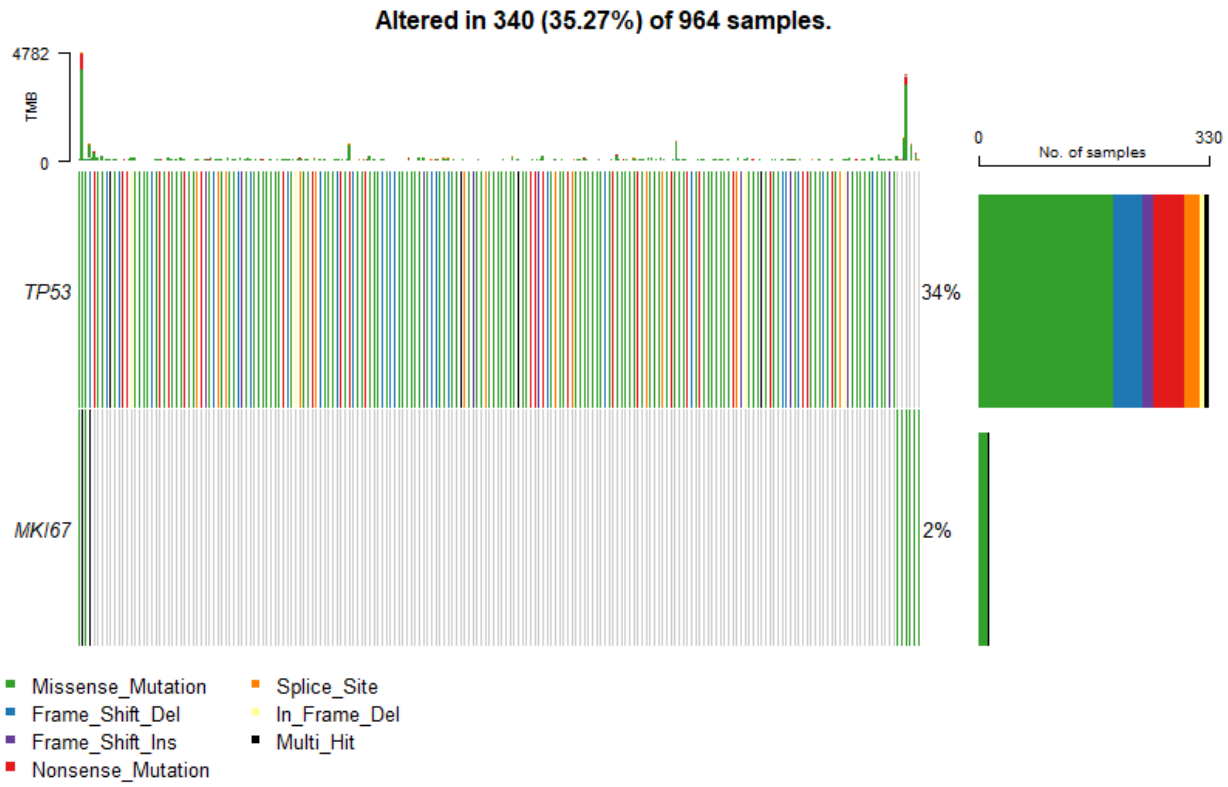

Figure 5. An oncoplot for genes TP53 and MKI67. Mutations between both genes do not align well (few patients have mutations in both genes), and TP53 has a much higher mutation rate (34%) than MKI67 (2%) in breast cancer patients. However, both genes have missense mutation as the most common mutation.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
10	3000	5040	4667	6040	37400	103

Figure 6. A summary of radiation dosage data from the clinical radiation dataframe. The high mean and median radiation dosage shows that the radiation levels used in radiotherapy are very dangerous.

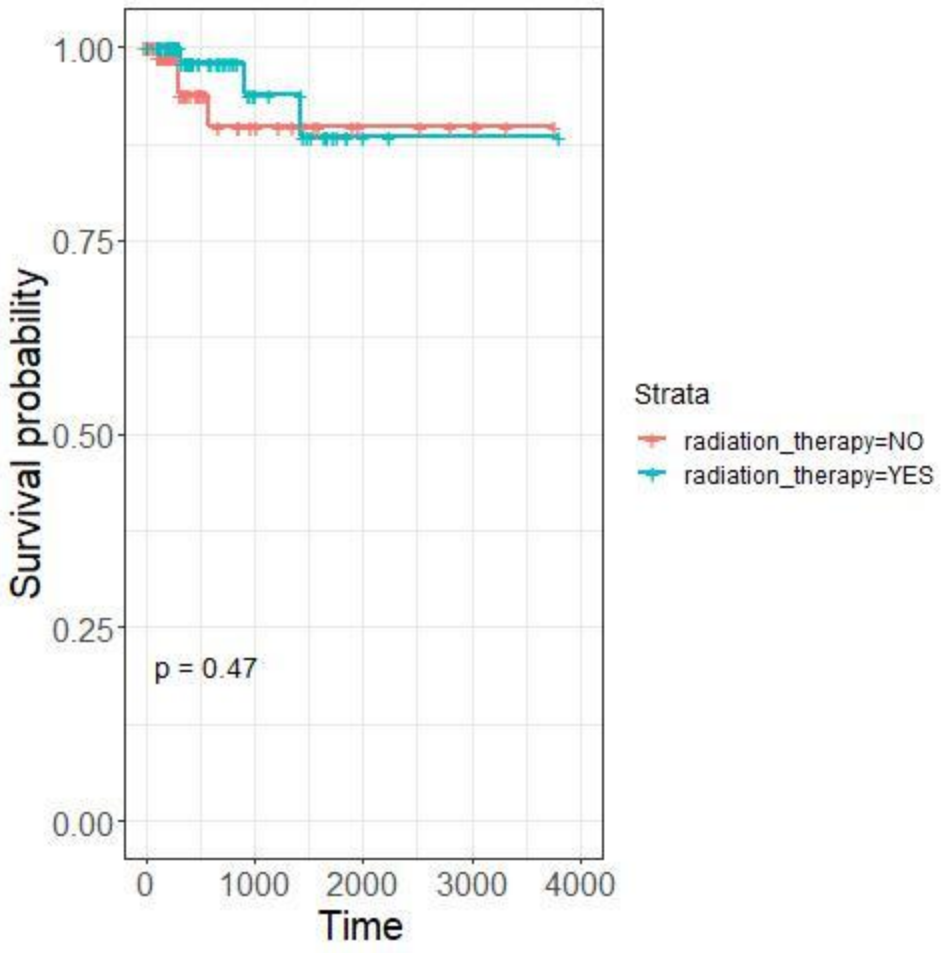

Figure 7. A Kaplan-Meier plot modeling survival probability over time for breast cancer patients who did or did not undergo radiation therapy. The p-value of 0.47 indicates that there is no statistically significant difference in survival between breast cancer patients who did undergo radiotherapy and those who didn't.

Discussion

The moderate positive linear correlation between TP53 counts and MKI67 counts in the Draftsman plot (Figure 1) was an anomaly since the other five gene pairs in the plot clearly had no correlation. Furthermore, the fact that the correlation was positive supports the idea that both genes have some sort of linkage or shared biological pathway that cause both genes to proliferate simultaneously. This correlation has a biological basis, as p53 protein (which is produced due to the TP53 gene) indirectly downregulates transcription of MKI67 (Uxa et al., 2021). This correlation is important because it suggests a possible mechanism for the onset of

breast cancer. This is because a loss in Ki-67 protein (which is produced due to the MKI67 gene) causes greater defects and mititic damage to chromosomes when p53 is absent (Garwain et al., 2021). This mechanism is supported by the fact that most of the breast cancer patients in the TCGA-BRCA dataset have low amounts of both TP53 and MKI67 gene counts, with only a few outlier data points having a higher count for either gene (Figure 4). The RStudio analysis of the TCGA-BRCA provides evidence supporting the biological mechanism that low expression of both the TP53 and MKI67 genes increases risk of breast cancer.

Interestingly, however, the counts for neither gene contribute to a decreased breast cancer survival rate. The paired boxplots in Figures 2 and 3 demonstrate that there is no statistically significant difference in counts for either gene between the group that survived breast cancer for at least 5 years and the group that didn't. This suggests that while there is a well-defined pathway linking a reduction in gene expression for both TP53 and MKI67, this mechanism does not have a major impact on the five-year lethality of the breast cancer.

Upon taking a closer look at the oncoplot comparing TP53 and MKI67 mutations in breast cancer patients (Figure 5), there are not many genetic similarities. Relatively few patients have mutations in both genes, and TP53 is mutated in 34% of the breast cancer patients as compared to 2% of breast cancer patients for MKI67. However, for both genes, missense mutations are the most common type of mutation among the breast cancer patients. This high prevalence of missense mutations is explained by the fact that mutated p53 can take a form that not only reduces tumor suppression, but also gains tumorgenesis functionality to rapidly produce cancer cells (Zhu et al., 2020). While other mutations types for TP53 will result in weakened tumor suppression ability for p53, the missense mutation specifically will also provide p53 with the ability to promote cancer growth, leading to more severe outcomes for breast cancer. While the missense mutation type dominates for the MKI67 gene (over 90% of mutations for this gene are missense mutations), it is only a narrow majority for TP53 gene (around 60% of mutations for this gene are missense mutations). This suggests while there may

be genetic mechanisms causing missense mutations to predominate over other mutation types for each of the two genes, these genetic mechanisms are unlikely to be the same or even just linked.

With TP53 being the most prevalent mutated gene in breast cancer patients (Tchelebi et al., 2014), it is important to look into treatment options for patients with TP53 mutations specifically. Radiotherapy is an option for many cancers (including breast cancer), but cancer researchers have been hesitant about using radiotherapy to treat breast cancer caused by TP53 mutations out of fear that the radiation will actually worsen the cancer by damaging cell regulatory mechanisms. In fact, some studies have shown that mutant p53 either has no effect on or increases cellular sensitivity to chemotherapy and radiotherapy (Tchelebi et al., 2014), and so this can make it challenging to develop a TP53-specific treatment plan. Thus, an analysis of radiation data was performed on breast cancer patients to examine the viability of radiotherapy for breast cancer patients affected by mutated TP53.

An analysis of summary data from the clinical radiation dataframe (Figure 6) shows high mean (4667mSv) and median (5040mSv) levels of radiation. These levels of radiation are sometimes lethal, and supports the idea that radiation can potentially damage cell regulatory pathways. Furthermore, the p-value of 0.47 for the Kaplan-Meier plot modeling survival probability over time for radiotherapy breast cancer patients and non-radiotherapy breast cancer patients (Figure 7) indicates that radiotherapy does not cause a statistically significant difference in survival for breast cancer patients. Due to the fact that radiotherapy exposes patients to a dangerous amount of radiation (which could compromise and worsen cell regulation) and fails to increase survival probability for breast cancer patients, the RStudio analysis of TGCA-BRCA data suggests that radiotherapy is not an ideal method to treat breast cancer caused by TP53 gene mutations.

Going forward, it is important to analyze whether any other genes interact with the TP53 gene in order to produce cancerous outcomes. The TCGA-BRCA analysis as well as other

cancer research studies have demonstrated that TP53 and MKI67 are linked, and exploring other pathways may help us determine ways to control TP53 gene expression and TP53 mutations. Furthermore, to understand exactly how radiotherapy affects breast cancer patients with TP53 mutations, it would be useful to obtain data specifically on patients who undergo radiotherapy so that a more direct analysis can be conducted.

Conclusion

While the TGCA-BRCA analysis of breast cancer data on RStudio did not yield any new findings, it did confirm old ones. The analysis supported the linkage between TP53 and MKI67 genes by demonstrating a correlation in counts for both genes and showing that breast cancer patients have low expression for both genes. Similarly, it supports the idea that missense mutations are most common for breast cancer patients with TP53 gene mutations due to the added mutant gene functionally of tumorgenesis. The analysis does also suggest that treating breast cancer patients with the TP53 gene mutation using chemotherapy is risky due to the insignificant survival benefit and the risk of damaging cells further with radiation. These findings are important to establishing a scientific consensus for the mechanisms of TP53 genes in breast cancer that can lay the groundwork for further research that can help prevent or even treat this spefic type of breast cancer.

References

Garwain, O., Sun, X., Iyer, D. R., Li, R., Zhu, L. J., & Kaufman, P. D. (2021). The chromatin-binding domain of Ki-67 together with P53 protects human chromosomes from mitotic damage. *Proceedings of the National Academy of Sciences*, *118*(32). https://doi.org/10.1073/pnas.2021998118

Tchelebi, L., Ashamalla, H., & Graves, P. R. (2014). Mutant P53 and the response to chemotherapy and radiation. *Subcellular Biochemistry*, 133–159. https://doi.org/10.1007/978-94-017-9211-0_8

Ungerleider, N. A., Rao, S. G., Shahbandi, A., Yee, D., Niu, T., Frey, W. D., & Jackson, J. G. (2018). Breast cancer survival predicted by TP53 mutation status differs markedly depending on treatment. *Breast Cancer Research*, *20*(1). https://doi.org/10.1186/s13058-018-1044-5

Uxa, S., Castillo-Binder, P., Kohler, R., Stangner, K., Müller, G. A., & Engeland, K. (2021). Ki-67 gene expression. *Cell Death & Differentiation*, *28*(12), 3357–3370. https://doi.org/10.1038/s41418-021-00823-x

Zhu, G., Pan, C., Bei, J.-X., Li, B., Liang, C., Xu, Y., & Fu, X. (2020). Mutant p53 in cancer progression and targeted therapies. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.595187

Part 3: Review Questions
Attach the answers to the following questions after your paper references.
General Concepts
1. What is TCGA and why is it important?

TCGA, or the Cancer Genome Atlas, is a cancer genomics program founded in December 2005. It is important because it publicly provides thousands of epigenomic, transcriptomic, and proteomic databases that scientists can use to conduct cancer research and analysis without having to go to the clinic to obtain this data from patients themselves.

2. What are some strengths and weaknesses of TCGA?

TCGA is beneficial to cancer researchers because they can use cancer data already obtained by others to perform their cancer research and analysis. This saves them a lot of time, effort, and cost that would otherwise go to finding patients to get cancer data from, getting consent from each patient to obtain that data, and waiting for the production of time-dependent data (ex. days until death). Furthermore, the fact that the data is publicly available means that cancer researchers who want to verify another cancer researcher's results can do so easily due to the availability of the data.

However, using TCGA data collected by someone else means that cancer researchers who use the data may not recognize certain biases that are affecting it. For instance, some areas or hospitals may have more of a certain demographic of patients compared to the national distribution (ex. more higher-income patients at private hospitals or more people of color in a disadvantaged community). Additionally, lack of patient consent and state laws regarding public research data may make it more challenging for data to be sent to the TCGA databases from certain regions, which can introduce more regional biases.

Coding Skills

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

```
cd [repositorypath]
git status
git add [filename.filetype]
git commit -m “[message]”
git push
```

- Note: the [ ] brackets are not actually included

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

```
if (!require([package])) {install.packages([package])}
library([package])
```

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

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager") {BiocManager::install(version = "3.15")}
library(BiocManager)
if (!require([package])) {BiocManager::install(“[package]”)}
library([package])
```

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

Boolean indexing is a method of making a vector or dataframe column full of “TRUE” or “FALSE” values based on whether the data in a corresponding vector or dataframe column meet a user-specified condition. This can be used to make a boolean mask that can be used to subset out unwanted data, allowing for he analysis of a smaller set of data containing only the desired data points.

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.

dataframe dicerolls

	number	frequency	frequency_adjusted
1	1	154	169.4
2	2	203	182.7
3	3	183	201.3
4	4	204	183.6
5	5	188	206.8
6	6	169	185.9

a. an ifelse() statement

```
dicerolls$frequency_adjusted = ifelse(dicerolls$frequency < 190, (dicerolls$frequency * 1.1), (dicerolls$frequency * 0.9))

#This line of code creates a new column in dicerolls called frequency_adjusted that multiplies the frequency value in dicerolls$frequency by 1.1 if it is less than 190 and multiplies the frequency value in dicerolls$frequency by 0.9 if it is greater than or equal to 190.
```

b. boolean indexing

```
even_mask = ifelse(dicerolls$number % 2 == 0, TRUE, FALSE)

#This line of code creates a boolean mask that sets each value to true if the corresponding value of dicerolls$number is even and false if the corresponding value of dicerolls$number is odd.

dicerolls_even = dicerolls[even_mask, ]

#This line of code creates a new dataframe called dicerolls_even with only the rows that meet the condition (even dicerolls$number value) set by the mask.
```

dataframe dicerolls_even

	number	frequency	frequency_adjusted
2	2	203	182.7
4	4	204	183.6
6	6	169	185.9