QBIO490 Mid-Semester Report (Colin Yeo)

Introduction: Multi-omic data analysis is growing to become one of the most powerful tools in bioinformatics, particularly from a clinical perspective. In broad terms, multi-omic analysis considers data from areas like the genome (DNA), proteome (protein makeup), and transcriptome – specifically, it examines these areas within groups of individuals to recognize patterns in disease and genetic mutation. One of the most applicable areas of multi-omic data analysis, clinically-speaking, is breast cancer research. By examining the current literature on this disease, some risk factors for breast cancer become apparent: living in developed countries, being a part of certain racial and ethnic groups, and hereditary factors are all potential contributors to breast cancer incidence (Momenimovahed et al., 2019).

Notably, this cancer is a significant cause of premature mortality among women, accounting for the most cancer deaths other than lung cancer in this demographic (Coughlin, 2019). Given the advancement of multi-omic data analysis, though, some researchers are advocating for increased individualized breast cancer patient care, as this type of analysis allows clinicians to begin tailoring treatments for patients based on considerations like their genome.

In this study, trends between certain breast cancer mutation types, different cancer manifestations, procedure/treatment types, and survival times were observed, in hopes of elucidating patterns between these variables. The public dataset TCGA, a cancer genomics program consisting of real-life, clinical data, was utilized in this study's analyses to answer the question: are mutation type, procedure type, and survival rates related in breast cancer patients? Ultimately, the most significant finding in this study was the relationship between procedure type and survival time – specifically, that lumpectomies typically correlate with a higher survival probability in the early stages of treatment, but that simple mastectomies are associated with higher long-term survival rates.

Methods: To examine the relationships between the selected variables, clinical data and MAF data related to breast cancer patients was gathered from the TCGA dataset using accession code "TCGA-BRCA". A total of seven RStudio libraries were downloaded and utilized in this study's analyses: BiocManager, TCGAbiolinnks, SummarizedExperiment, maftools, ggplot2, survival, and survminer.

First, the three types of surgical procedures were chosen from the clinical data; there were five total procedures listed (Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, "Other", and a blank procedure), but the "other" and blanks were removed using simple Boolean masking. Next, a Kaplan-Meier plot was created that traces the survival rates for each of the three valid procedures using functions "Surv", "surv_fit", and "ggsurvplot".

Next, two bar graphs were generated for this study's analyses. The first showed the average survival time of individuals who fell under each of the three procedure categories, and was created using simple sum and average calculations with ggplot, from the clinical dataset. The second showed the spread of drug usage across the three procedure categories. To simplify the graphic, only the top four drugs used were included for each type of procedure – the patients who received a valid procedure but did not receive a drug that was within the top four drugs used were masked out of this analysis. Like the first barplot, this barplot also used ggplot, but used a stacked bar format for the chart, rather than a simple barplot, drawing data from the clinical drugs dataframe and the clinical patients dataframe.

Lastly, a simple oncoplot visualizing the association between certain mutation types and the three selected surgical procedures was created, specifically to see if any clustering patterns were present over any surgical procedure type. This analysis was performed using MAF data.

Results:

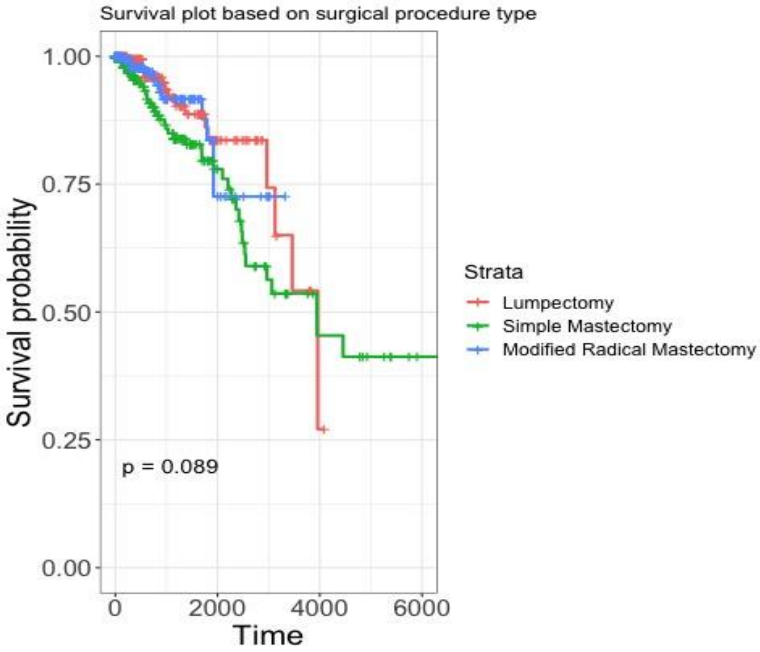

Figure 1. Kaplan-Meier survival plot showing the relative survival probabilities following each of the three selected surgical procedures. P-value of 0.089 is slightly above the significance cutoff of 0.05, which indicates that the results may not hold much statistical significance.

The Kaplan-Meier plot (Figure 1) created for the three surgical procedures and their survival rates pointed out two noteworthy trends -- it can be seen that lumpectomies exhibit better survival at the beginning of treatment plans (indicated by the red line remaining the highest in the first half of the plot), but that modified radical mastectomies show better long-term survival (as the green line persists for the longest, while the other two lines drop off the graph quickly). However, note that this may be due to a gap in the data (e.g. the blue and red lines stop much sooner than the green line, which indicates that there were not enough patients included in this analysis to accurately represent all three surgical procedures).

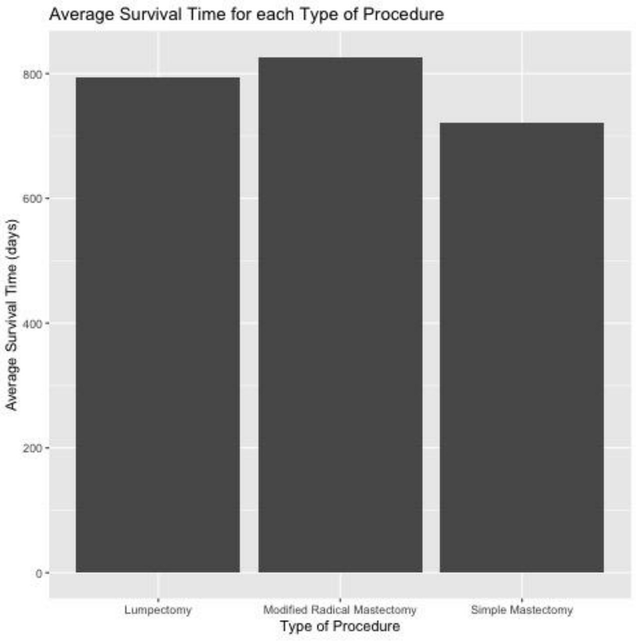

Figure 2. Barplot showing the average survival time post-procedure for each of the three selected procedure types.

From Figure 2, it can be seen that the patients who received a modified radical mastectomy demonstrated a longer survival time on average than the other two procedures, and that individuals who received a simple mastectomy exhibited the shortest survival time on average. However, the average survival times for all three treatments tended to hover between 700-850 days, without any significant differences between them.

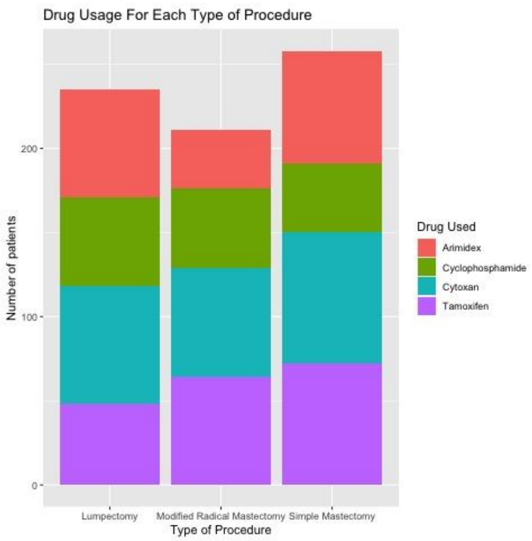

Figure 3. Stacked barplot showing the usage of four different breast cancer drugs post-surgical procedure, for each of the three selected procedures.

From Figure 3, there are some slight differences in drug choice for individuals who got differing surgical procedures to treat their breast cancer. For example, Arimidex was chosen least of the four drugs to treat individuals after a modified radical mastectomy, and Tamoxifen was chosen to treat patients who received a simple mastectomy more than the other two procedure groups.

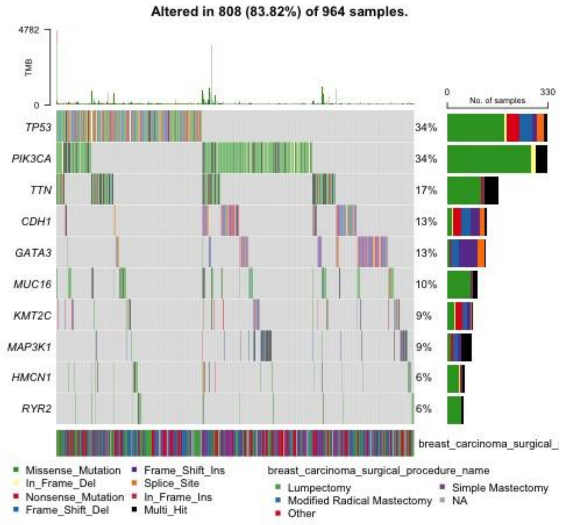

Figure 4. A basic oncoplot showing the clustering of certain genes and their corresponding surgical procedure choices.

In Figure 4, there doesn’t seem to be significant clustering of any one category (e.g. lumpectomy or simple mastectomy) in relation to specific genes. However, there is a bit of clustering within the TP53 gene and the mastectomy procedures (if the simple and modified radical mastectomy procedures are considered a single factor).

Discussion: One of the strongest trends found in this study’s analyses was the association between procedure type and survival time (Figure 1). Though the p-value was slightly higher than the threshold of $p < 0.05$, it was still shown that lumpectomies exhibit better outcomes towards the early stages of treatment and breast cancer progression. This finding was corroborated by a study published in *Cancer*, which found that breast-conserving therapies (i.e. lumpectomies fall under this category, as they don’t remove the whole breast) were associated with improved disease-specific survival compared to full mastectomies (Hwang et al., 2013).

Furthermore, another study found that lumpectomy combined with irradiation treatment was much more effective at preventing recurrent cancerous tumors compared to lumpectomy without irradiation – this wasn’t a metric that the clinical data included for this study but might be an area worth exploring in the future (Fisher et al., 2002). Though these findings are contradicted slightly by Figure 2, as it shows that individuals who received certain types of mastectomies had a long survival time on average, the differences between these survival times were not significantly different, differing by less than a few months.

Though this study analyzed a few different genes, the one of most interest, from the figures that were generated is TP53 – this gene is responsible for a significant fraction of breast cancer diagnoses in clinical settings. Specifically, it was found that mastectomies (both simple and radical) were used more often to treat individuals with TP53 mutations (see the clustering in Figure 4), which was corroborated by a review study that found mastectomies were recommended over lumpectomies in TP53 mutation carriers (Schon et al., 2017). Furthermore, the use of tamoxifen in patients who received mastectomies was most prevalent, compared to other drug choices (Figure 3). This finding is slightly divisive when examining current literature, as some studies suggest that tamoxifen may be responsible for recurring breast cancer, but a recent study showed that for individuals with TP53 mutations, tamoxifen showed superior survival rates (Shahbandi et al., 2020). Connecting these figures, one of the main conclusions that this report found was that individuals with TP53 mutations tend to receive mastectomies, and that those who received mastectomies were treated often with tamoxifen, in favor over other drugs – a conclusion that is echoed by current literature on the topic.

Given the analyses that were performed in this study and the results that were observed, one area of potential future exploration lays within the more nuanced treatment plans that breast cancer patients embark on as of recent. For example, comparing different types of treatments with or without irradiation or chemotherapy therapeutics, for different mutation types, might yield interesting trends that may influence clinical practices in the future. Futhermore, performing the same analysis in this study, but on a larger set of patients (e.g. using a different data source) might be worthwhile in confirming and refining the conclusions of this study.

References

Coughlin, S. (2019). Epidemiology of Breast Cancer in Women. *Adv Exp Med Biol* 1152, 9-29. DOI: 10.1007/978-3-030-20301-6_2.

Fisher, B., Anderson, S., Bryant, J., Margolese, R., Deutsch, M., Fisher, E., Jeong, J. & Wolmark, N. (2002). Twenty-Year Follow-up of a Randomized Trial Comparing Total Mastectomy, Lumpectomy, and Lumpectomy plus Irradiation for the Treatment of Invasive Breast Cancer. *The New England Journal of Medicine* 347, 1233-1241. DOI: 10.1056/NEJMoa022152.

Hwang, E., Lichtensztajn, D., Gomez, S., Fowble, B. & Clarke, C. (2013). Survival after lumpectomy and mastectomy for early stage invasive breast cancer. *Cancer* 119, 1402-1411. DOI: 10.1002/cncr.27795.

Momenimovahed, Z. & Salehiniya, H. (2019). Epidemiological Characteristics of and Risk Factors for Breast Cancer in the World. *Breast Cancer (Dove Medical Press)*. DOI: 10.2147/BCTT.S176070.

Schon, K & Tischkowitz, M. (2017). Clinical implications of germline mutations in breast cancer: TP53. *Breast Cancer Research and Treatment* 167, 417-423. DOI: 10.1007/s10549-017-4531-y

Shahbandi, A., Nguyen, H. & Jackson, J. (2020). TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines. *Trends in Cancer* 6, 98-110. DOI: 10.1016/j.trecan.2020.01.007

Review Questions:

General Concepts

1. What is TCGA and why is it important?

   TCGA is a cancer genomics program that makes patient data publicly available. The data that is accessible from TCGA includes genome, epigenome, proteome data, etc. It also covers a variety of cancer types (e.g. breast cancer). It is important because it allows researchers to access a wealth of real-life clinical data in a way that isn't costly or difficult to do.

2. What are some strengths and weaknesses of TCGA?

   One of the biggest strengths of TCGA is its accessibility, as it is a publicly accessible dataset that researchers and scientists can utilize to do many different types of analyses. Additionally, TCGA, unlike many other sources of data for scientific articles, doesn't require any costly subscription fees or payment. However, TCGA did initially face concerns about the ethical implications of releases health data to the public – they responded to these concerns in an appropriate manner.

Coding Skills

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

   Git add, Git commit, Git push.

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

   Install.packages("package_name"), then library("package_name").

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

   BiocManager::install("package_name"), then library("package_name").

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

Generating a column that is filled with true and false values (Booleans) based on some quality of another column in your dataframe. It allows you to parse and sift out values you don't want in a quick way, to make analysis more efficient (and in the case of NA values, it makes the analysis possible to begin with).

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.

name	age
Colin	20
Karla	19
Sabrina	21

Dataframe named "roster".

a. an ifelse() statement

older_than_19_mask <- ifelse(roster$age > 19, T, F)

The above creates a vector of boolean containing "T, F, T" as its values, corresponding to whether each person in that row is older than 19 or not.

b. boolean indexing

older_roster <- roster[, older_than_19_mask]

The above creates a new dataframe with just the data from "Colin" and "Sabrina" rows, and excludes the row with "Karla", because she is not over 19. This is done using the mask created in part a.