Ethnicity and Breast Cancer: Multiomic Analysis of Factors Affecting Survival, with Focus on Differential Mutation Rate by Race in Genes TP53 and PIK3CA.

Brandon Ye

QBIO 490 Mid-Semester Project

October 23, 2022

I. Introduction

Breast cancer (BC), a highly heterogenous malignancy developing more commonly in women than men, is the most frequently diagnosed cancer indication worldwide and is among the leading causes of cancer death (Sung, H. et al., 2021). The association between ethnicity, BC incidence, and ultimately survival has been extensively studied on a multitude of levels, including but not limited to biological, pathological, and socioeconomic. However, this association is less studied within a genomic context, specifically in relation to genes TP53 and PIK3CA. TP53 and PIK3CA are among the most commonly mutated genes in BC patients and are widely understood to have antagonistic effects on patient survival outcomes. Whereas TP53 mutations are associated with poorer overall survival, especially in late-stage or metastatic BC patients, (Meric-Bernstam, F. et al., 2018), PIK3CA mutations have been shown to present with clinically favorable features, including lower tumor grade, HER2 negativity, and older age at diagnosis (Kalinsky, K. et al., 2009). Thus, the advancement of personalized therapies for cancer comes with a need to investigate the association between genes pivotal for breast cancer incidence and prognosis (among them TP53 and PIK3CA) and ethnicity to ultimately improve patient outcomes.

The current large-scale study of BC and other cancer indications is hindered by a lack of high-quality data, standardized study design, and agreeable parameters, leading to oftentimes confounding results. These shortcomings are largely addressed by The Cancer Genome Atlas (TCGA), an NIH-funded public-data project which aims to create a comprehensive collection of multi-omic cancer data via large-scale genome sequencing and various other forms of integrated multi-dimensional analyses (Tomczak, K. et al., 2015). For this study of BC, TCGA provides valuable clinical and demographic as well as genomic mutation data for various analyses.

Herein, we discuss differential BC survival outcomes by ethnicity with a special focus on the differential mutation rates in TP53 and PIK3CA by race. The goal of this study, therefore, is to provide novel insights into the association between ethnicity and BC incidence and outcome that may potentially be useful in guiding the future of personalized therapy.

II. Methods

TCGA analysis requires the use of primarily two datasets: clinical and mutation annotation format (MAF) clinical data. Each dataset was queried separately prior to conducting any analysis. The clinical dataset was queried using GDCquery with the project “TCGA-BRCA” and data category “Clinical”, downloaded using

GDCdownload(), and finally prepared for analysis using GDCprepare_clinic. The clinical MAF dataset was accessed from a larger MAF object, which was queried similarly using GDCquery with the project "TCGA-BRCA" and data category "Simple Nucleotide Variation", prepared using GDCprepare, and read into a dataframe using read.maf(). Necessary packages to prepare the aforementioned datasets include "maftools", "BiocManager", and "TCGAbiolinks".

Simple visualizations such as countplots were constructed using the package "ggplot2", while survival-related visualizations, including Kaplan-Meier plots, necessitated the "survival" and "survminer" packages. The integration of clinical drug data requires the use of merge(). Prior to any survival analysis being conducted, survival time and survival status columns were added to both the clinical and clinical MAF datasets. The determination of survival time depended on the values of two other variables. If a patient was not NA for the days to death variable, days to death was used for survival time. Otherwise, the days to last followup variable was used as a proxy for days to death. Cleaning the race variable in each dataset simply involved congregating any patients with no race data into an "Other" designation. For analyses regarding clinical MAF data, patients in the dataset were binarily characterized by race; either "White" or "Other" due to constraints revolving around certain plotting functions. To achieve this, "White" and "Other" patient barcodes were extracted from the original dataset and used to subset it into two individual datasets with subsetMaf(), one with only "White" patients, and another with only "Other" patients before any subsequent analysis. Kaplan-Meier plots derived from clinical data were constructed using ggsurvplot(), while Kaplan-Meier plots depicting MAF data utilized the mafSurvival() function. Oncoplots, co-oncoplots, and co-lollipop plots were constructed using the oncoplot(), coOncoplot(), and lollipopPlot2() functions, respectively.

III. Results

Preliminary visualizations depict a heterogenous but skewed distribution of race within the clinical dataset, which can potentially confound any significant results (Figure 1). Kaplan-Meier analysis on the clinical dataset indicates differential survival outcomes for patients based on race, although not statistically significant (Figure 2). With regard to genes of interest TP53 and PIK3CA, preliminary analysis (race-agnostic) depicts a seemingly mutually exclusive relationship between mutations of the two genes within the clinical MAF dataset (Figure 3). On the basis of race, the co-oncoplot constructed from the clinical MAF dataset signifies (a) a higher mutation frequency of TP53 in non-white populations and (b) a higher mutation frequency of PIK3CA in white populations (Figure 4). Together, the two co-lollipop plots (Figures 5 and 6) disprove the hypothesis that the differential mutation frequency observed in the co-oncoplot is detectable on the genome

level. Finally, Kaplan-Meier plots (Figures 7 and 8) from clinical MAF data suggest that TP53 mutations adversely affect survival for BC, while PIK3CA mutations positively affect survival, albeit not to a statistically significant extent, respectively. The distribution of therapy types is similar among different races (Figure 9), pointing to an increasingly apparent need for better personalized BC therapies.

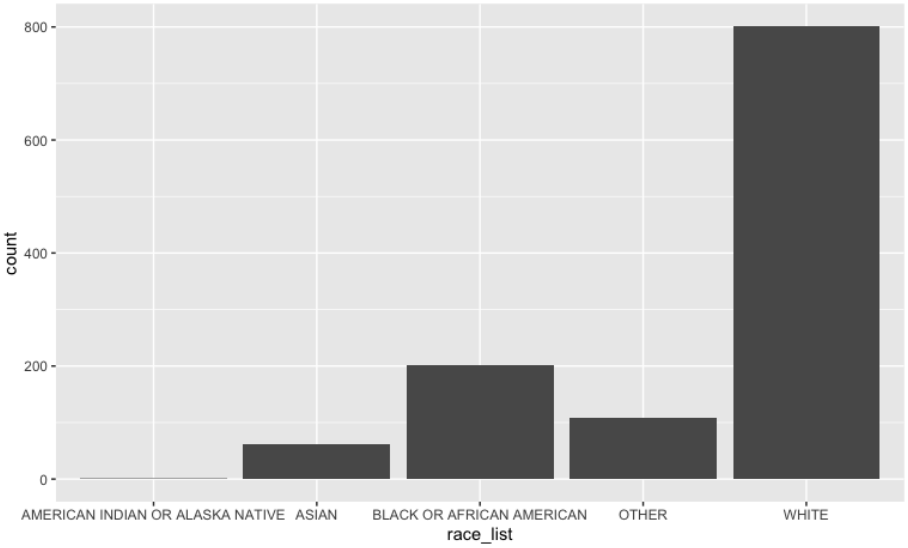

Figure 1: Countplot visualizes highly skewed racial distribution in the clinical dataset. American Indian or Alaska Native (1 patient), Asian (62), Black or African American (201), Other (109), White (801).

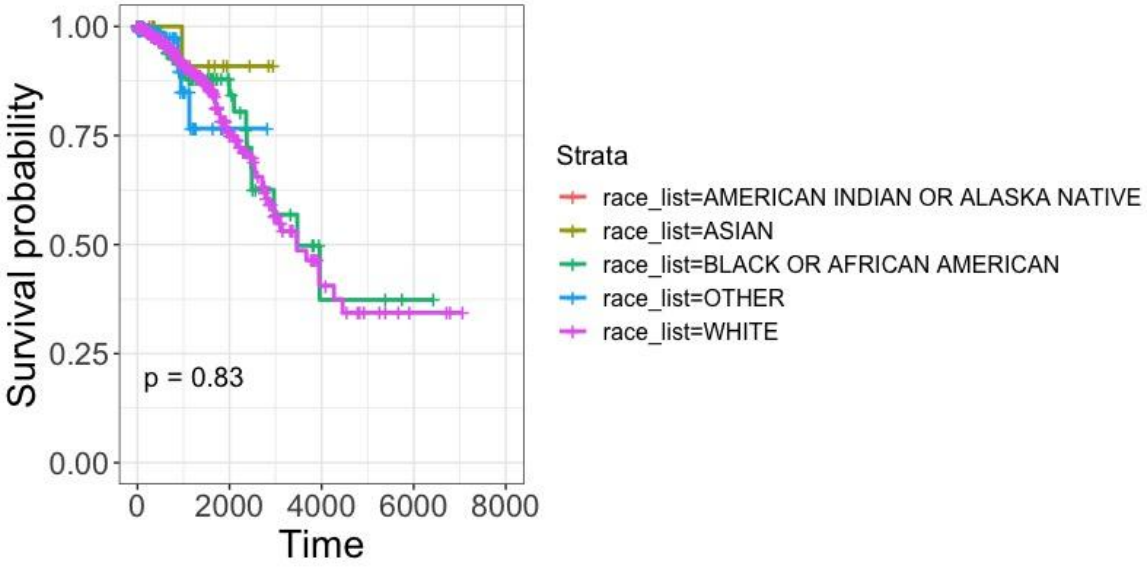

Figure 2: Kaplan-Meier plot suggests differential survival outcomes by race. Whites present linear survival probability by time, while minority survival probability is more variable. *p*-value of 0.83 suggests that results are insignificant, contradicting existing literature.

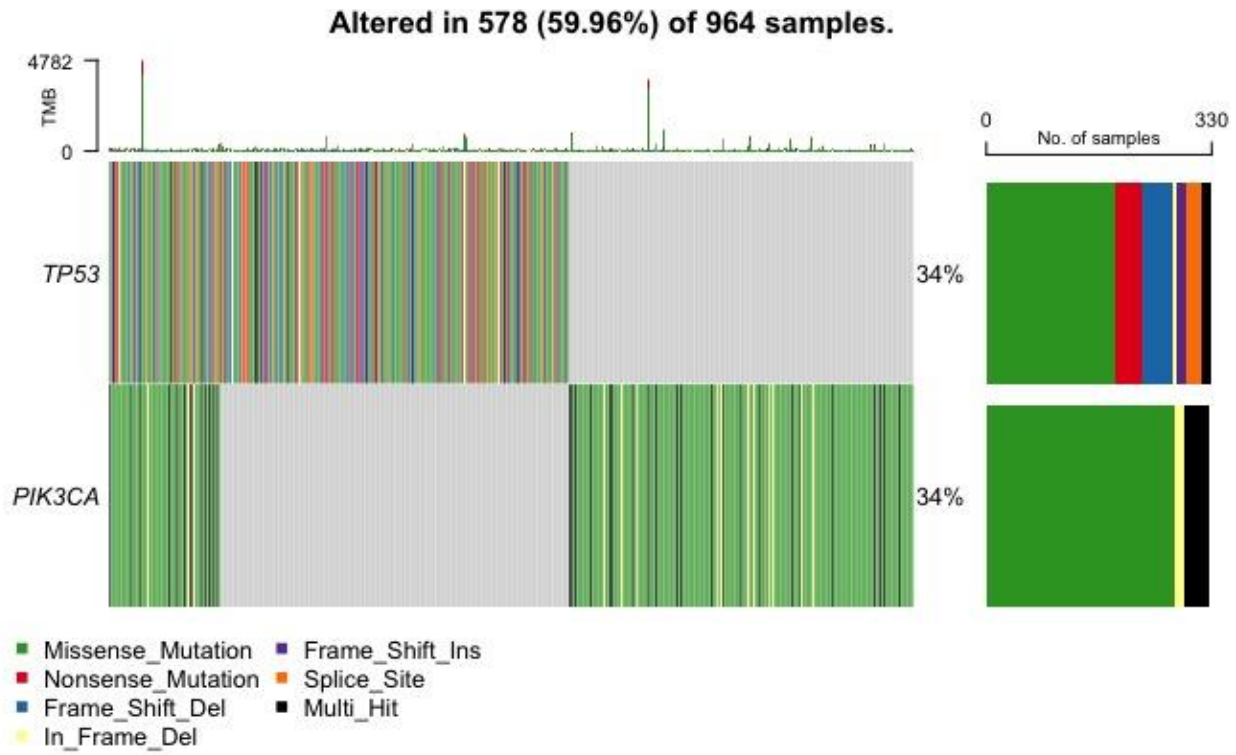

Figure 3: Oncoplot displaying differential mutations in individual patients in the mutation data dataframe. Mutual exclusivity of TP53 and PIK3CA mutations is suggested by the low overlap between the two. Mutations for both TP53 and PIK3CA are primarily missense, although TP53 displays higher heterogeneity in mutation type.

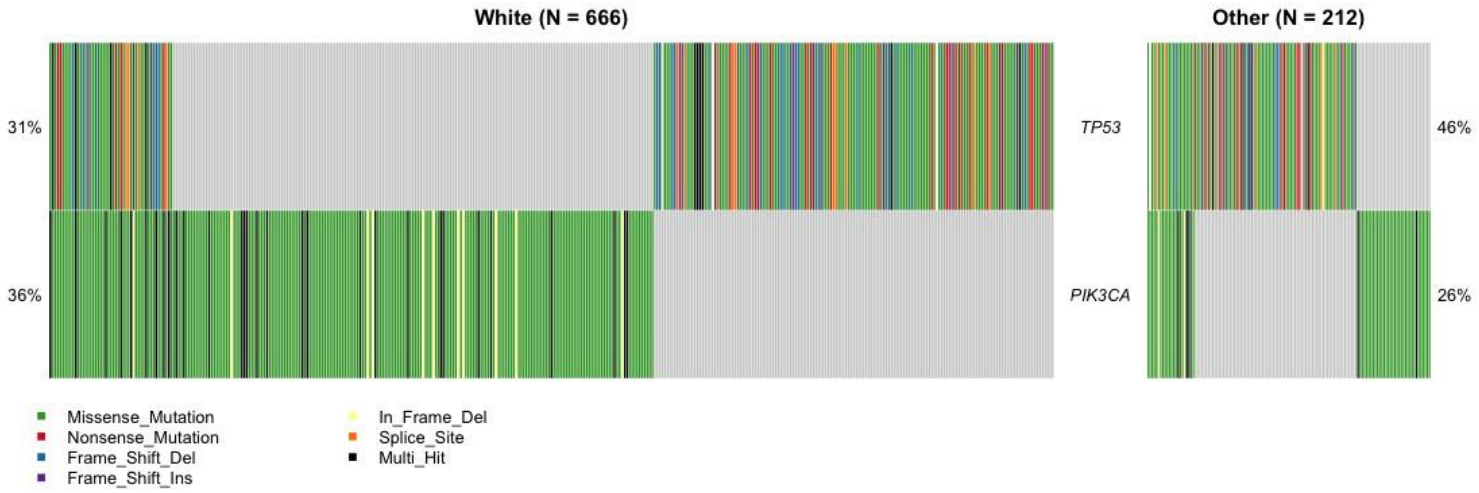

Figure 4: coOncoplot reveals differential mutation frequencies of TP53 versus PIK3CA in white versus minority populations. The White subpopulation has a decreased TP53 and increased PIK3CA mutation frequency in comparison to its minority counterpart.

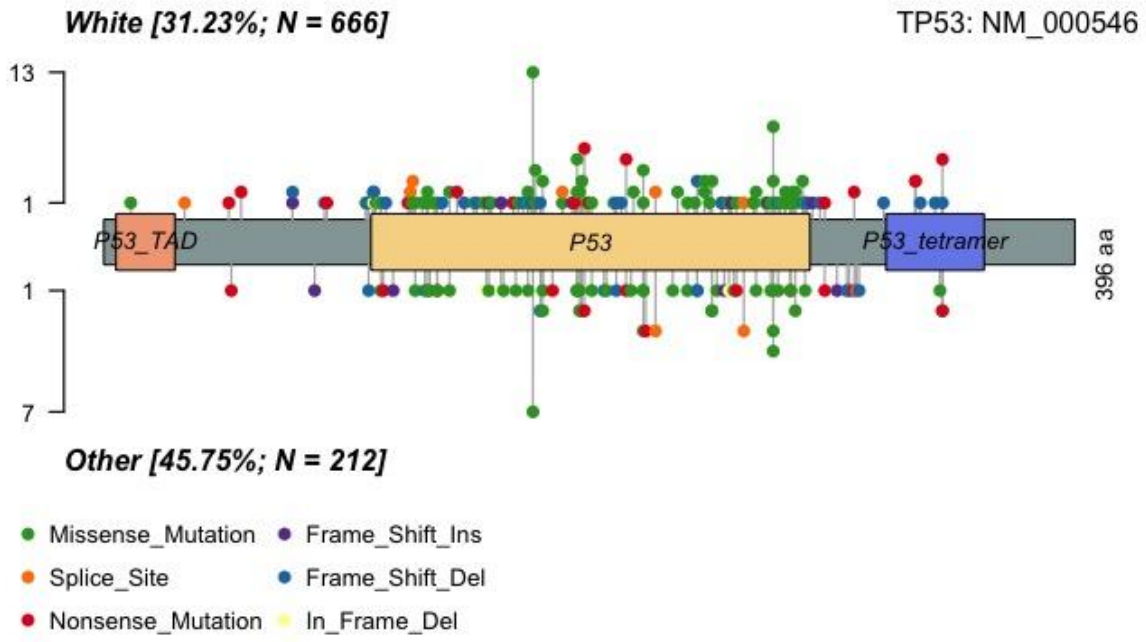

Figure 5: Cololipop plot for the TP53 gene displays little mutational diversity in white versus minority patients on the genome level. Mutations for both subpopulations occur most frequently in the P53 domain of the gene.

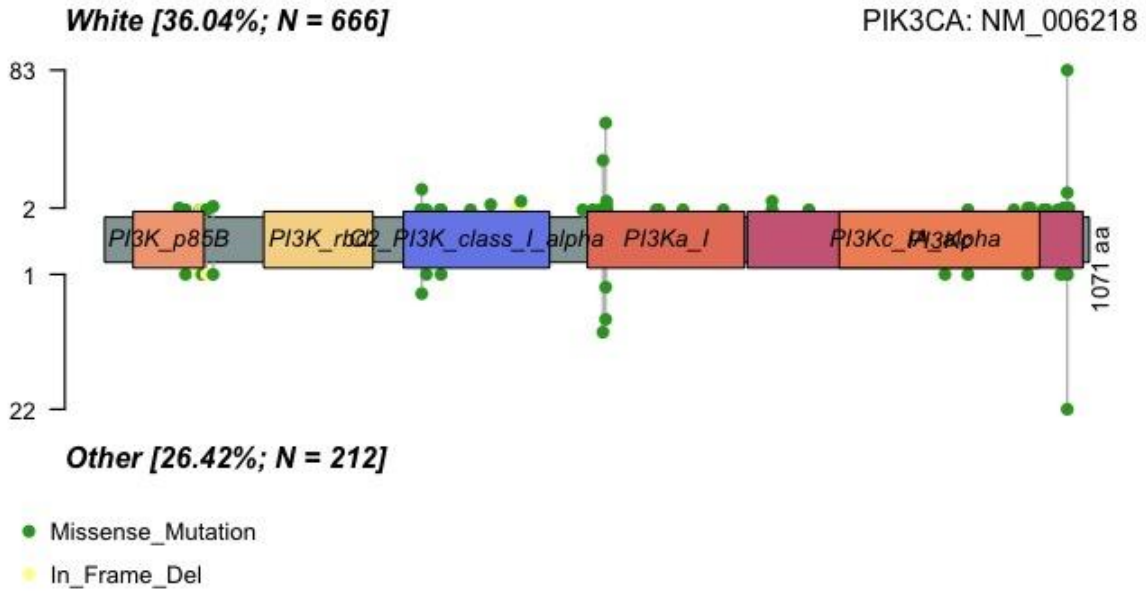

Figure 6: Cololipop plot for the PIK3CA gene displays little mutational diversity in white versus minority patients on the genome level. Mutations for both subpopulations occur primarily in the PI3Ka_I and PI3_PI4_kinase domains.

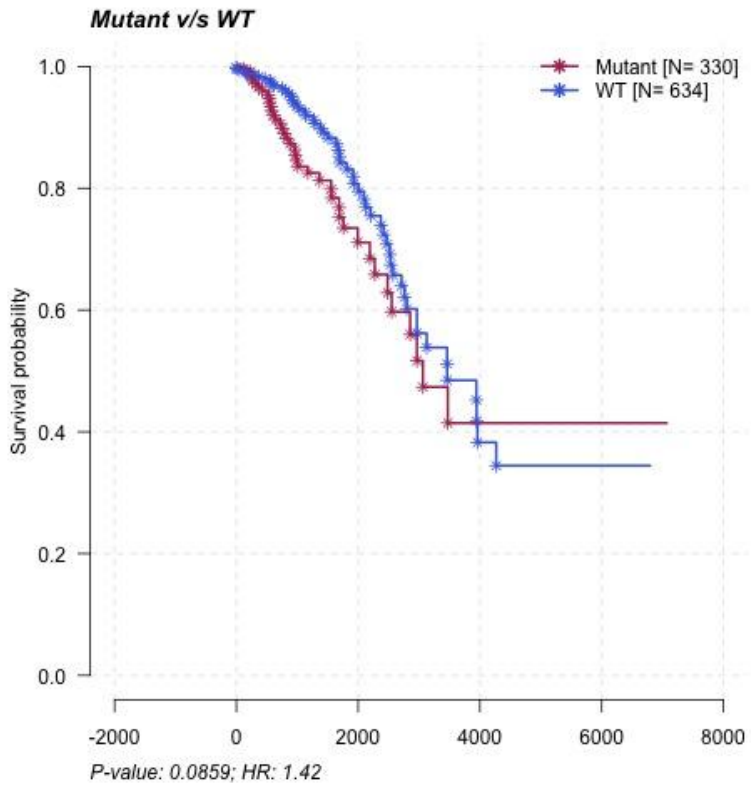

Figure 7: Kaplan-Meier plot for the TP53 gene suggests adverse survival outcomes for mutant subpopulations. *p*-value of 0.0859 indicates statistically significant results.

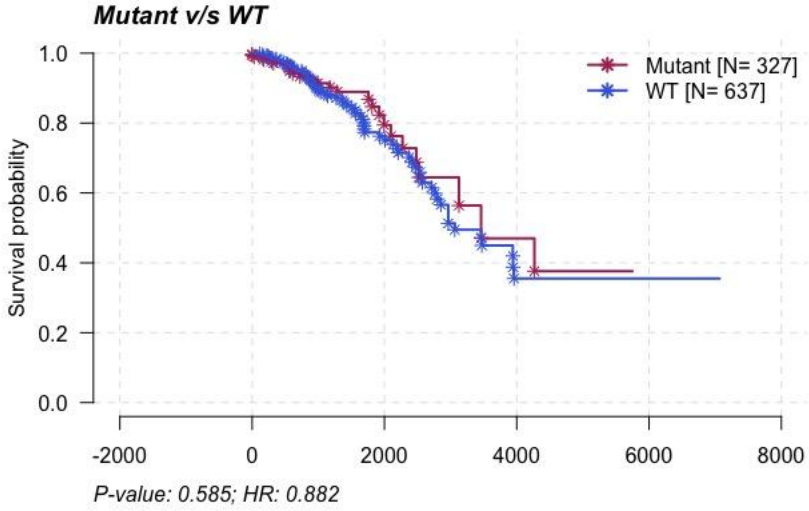

Figure 8: Kaplan-Meier plot for the PIK3CA suggests adverse survival outcomes for wildtype subpopulations. *p*-value of 0.585 indicates statistically insignificant results.

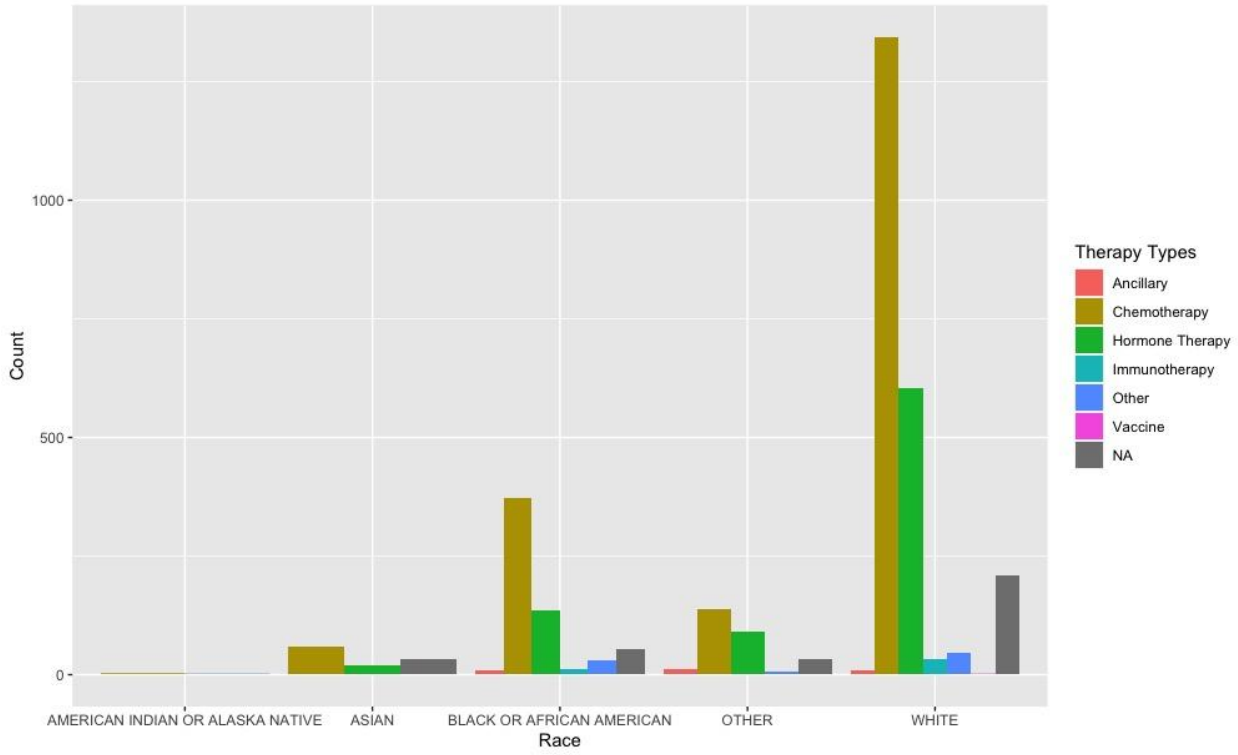

Figure 9: Grouped countplot suggests that chemotherapy is the most commonly received therapy among BC patients regardless of race. Furthermore, the distribution of therapy types across different races is notably similar except in "Other", where hormone therapy frequency nears that of chemotherapy.

IV. Discussion

One of the fundamental premises of this study–the fact that BC patient race is associated with survival outcomes in a statistically significant manner–is unable to be verified by the clinical and demographic data contained in the TCGA BC database (*p*-value of 0.83, as shown in Figure 1). However, it is widely accepted in current literature that certain racial disparities exist in both BC incidence and survival. For example, the incidence rate of BC before age 45 is higher among Black than White women, and importantly, Black women are found to be more likely to die from BC at virtually every age (Yedjou, C. G. et al., 2019). Hispanic women have a lower incidence of BC than non-Hispanic white women, but their cancers are often diagnosed at later stages and present with large tumors; BC incidence for Asian and White women is similar (Yedjou, C. G. et al., 2019).

Operating under the assumption that patient race is indeed associated with survival outcomes in BC as existing literature suggests, this study provides valuable insights into two key genes (TP53 and PIK3CA) that influence BC progression and presentation in the

context of race. We find that the preliminary MAF analysis presented in Figure 4 is largely corroborated by existing findings regarding differential mutation rates in certain landmark BC genes in minority versus white populations. Previous studies find that African Americans had more TP53 mutations than Whites (42.9% versus 27.6%, respectively) and fewer PIK3CA mutations (20% versus 33.9%) (Keenan, T. et al., 2015). While our analyses also incorporate minority populations beyond African Americans such as Asians, the disparities remain evident (46% “Other” versus 31% “White” in TP53 and 36% “White” versus 26% “Other” in PIK3CA).

Analysis of the potential differences in mutation distribution on the genome level (Figures 5 and 6) provides novel insights. Here, we demonstrate that the observed differential mutation frequency among Whites and minority populations in TP53 and PIK3CA is not immediately detectable on the genome level. In other words, larger-scale biological forces are at play when it comes to determining the mutation status for these two genes among different ethnic populations. However, further research is needed to determine whether this result is specific to BC or rather a generalizable observation across multiple different cancer indications. Currently, there seems to be no such research that is both publicly accessible and well-cited.

Finally, Kaplan-Meier analysis (Figures 7 and 8) support much of what is already known about the implications of TP53 and PIK3CA survival outcomes. As a critical tumor suppressor gene whose product, tumor protein p53, functions as a checkpoint control following DNA damage, somatic mutations in TP53 are highly associated with early onset, poor prognosis, and ultimately poorer survival outcomes (Schon, K. et al., 2018). BC patients mutant for PIK3CA, a gene that encodes the p110α subunit of a lipid kinase family responsible for mediating cell survival, differentiation, and proliferation, present with more favorable clinicopathological features, including positive expression of estrogen receptors, smaller tumor size, and low histological grade (Zardavas, D. et al., 2014). Together with differential mutation data by race as presented in Figure 4, the survival analysis conducted here corroborates the important finding that patients of minority descent face harsher outcomes compared to their White counterparts, in part influenced by the highly mutated TP53 and PIK3CA genes. Importantly though, the notable similarity in therapy type by race (Figure 9) despite differing mutation frequencies and survival outcomes points to a need for a shift in the current clinical paradigm of BC care. Ultimately, new strategies, primarily in the form of improved personalized therapies, are needed to promote BC prevention, improve survival outcomes, and reduce mortality, especially for populations of ethnic minority.

Consideration for future direction includes addressing apparent class imbalance as well as increasing the number of genes ultimately analyzed. Expanding beyond TP53 and

PIK3CA a landscape of the top *X* most mutated genes among BC patients and subsequent analysis of how they correlate with race and ultimately survival would perhaps be more useful and even potentially clinically relevant. Additionally, considering that the TCGA BC data is skewed heavily toward White patients in number, a more evenly-distributed dataset by race could provide novel, more accurate insights. Addressing this described class imbalance will potentially be useful moving forward.

V. References

1. Kalinsky, K., Jacks, L. M., Heguy, A., Patil, S., Drobnjak, M., Bhanot, U. K., Hedvat, C. V., Traina, T. A., Solit, D., Gerald, W., & Moynahan, M. E. (2009). PIK3CA Mutation Associates with Improved Outcome in Breast Cancer. *Clinical Cancer Research*, *15*(16), 5049–5059. https://doi.org/10.1158/1078-0432ccr-09-0632

2. Keenan, T., Moy, B., Mroz, E. A., Ross, K., Niemierko, A., Rocco, J. W., Isakoff, S., Ellisen, L. W., & Bardia, A. (2015). Comparison of the genomic landscape between primary breast cancer in African American versus White Women and the association of racial differences with tumor recurrence. *Journal of Clinical Oncology*, *33*(31), 3621–3627. https://doi.org/10.1200/jco.2015.62.2126

3. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A), A68–A77. https://doi.org/10.5114/wo.2014.47136

4. Meric-Bernstam, F., Zheng, X., Shariati, M., Damodaran, S., Wathoo, C., Brusco, L., Demirhan, M. E., Tapia, C., Eterovic, A. K., Basho, R. K., Ueno, N. T., Janku, F., Sahin, A., Rodon, J., Broaddus, R., Kim, T.-B., Mendelsohn, J., Mills Shaw, K. R., Tripathy, D., Mills, G. B., Chen, K. (2018). Survival Outcomes by TP53 Mutation Status in Metastatic Breast Cancer. *JCO Precision Oncology*, (2), 1–15. https://doi.org/10.1200/po.17.00245

5. Schon, K., & Tischkowitz, M. (2018). Clinical implications of germline mutations in breast cancer: TP53. *Breast cancer research and treatment*, *167*(2), 417–423. https://doi.org/10.1007/s10549-017-4531-y

6. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

7. van Beek, E., Hernandez, J. M., Goldman, D. A., Davis, J. L., McLaughlin, K., Ripley, R. T., Kim, T. S., Tang, L. H., Hechtman, J. F., Zheng, J., Capanu, M., Schultz, N., Hyman, D. M., Ladanyi, M., Berger, M. F., Solit, D. B., Janjigian, Y. Y., & Strong, V. E. (2018). Rates of TP53 Mutation are Significantly Elevated in African American Patients with Gastric Cancer. *Annals of surgical oncology*, *25*(7), 2027–2033. https://doi.org/10.1245/s10434-018-6502-x

8. Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., Alo, R. A., Payton, M., & Tchounwou, P. B. (2019). Health and Racial Disparity in Breast Cancer. *Advances in experimental medicine and biology*, 1152, 31–49. https://doi.org/10.1007/978-3-030-20301-6_3

9. Zardavas, D., Phillips, W. A., & Loi, S. (2014). PIK3CA mutations in breast cancer: Reconciling findings from Preclinical and Clinical Data. *Breast Cancer Research*, *16*(1). https://doi.org/10.1186/bcr3605

VI. Coding Skills

1. What commands are used to save a file to your GitHub repository?
    a. git add –all
    b. git commit -m “message”
    c. git push

2. What command(s) must be run in order to use a standard package in R?
    a. install.packages(“Package”)
    b. library(“Package”)

3. What command(s) must be run in order to use a Bioconductor package in R?
    a. if(!requireNamespace(“BiocManager”, quietly = TRUE))
        install.packages(“BiocManager”)
        BiocManager::install(version = “3.15”)
    b. library(BiocManager)
    c. if(!require(“Package”))
        BiocManager::install(“Package”)
    d. library(“Package”)

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

    Boolean indexing refers to a method of subsetting data by value using booleans. It can be especially useful for categorizing otherwise continuous datatypes (for example, we can define an “old” patient as someone older than 50, while a “young” patient as someone 50 or younger). Boolean indexing is also useful for certain types of data analysis, as it allows for otherwise complicated variables to be subsetted into more manageable sub-portions.

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.

df

name	gender	age	fav_color
Bob	Male	15	Blue
Sally	Female	26	Purple
Joe	Male	12	Green

a. an ifelse() statement

   If we wanted to generate a column to describe whether or not the person is a minor, we could write something like:

   df$is_minor <- ifelse(df$age < 18, T, F)

   The ifelse() statement checks the age column in each row. If the age is less than 18, then they are a minor, otherwise, they are not.

b. boolean indexing

   If we wanted to subset out all the females from this dataset, we could use boolean indexing to create a mask as follows:

   female_mask <- ifelse(df$gender == “Female”, T, F)
   df[!female_mask,]

   The first line of code uses ifelse() logic (explained above) to create a mask containing all people in our dataset that are female. Then, in the second line, we apply our mask to all rows of the dataframe, using boolean indexing to subset out all female patients. Note that we use “!” when doing this, as subsetting the dataframe in such a way will only include “true” occurrences, and we are looking to remove all females.