Differences in Breast Cancer between race groups

Alice Wang

Introduction:

The Cancer Genome Atlas (TCGA), founded in December of 2005, is a cancer genomics program hosted by the National Cancer Institute and the National Human Genome Research Institute. The publicly available data from this project includes genomic, epigenomic, transcriptomic, and proteomic data. This data was collected from 20,000 different samples that span 33 different cancer types. Overtime, TCGA helped establish the importance of cancer genomics, transformed our understanding of cancer, and even begun to change how the disease is treated in the clinic.

Breast cancer, as the most frequently diagnosed cancer in women worldwide with more than 2 million new cases in 2020, is currently one of the most prevalently diagnosed cancers and the 5th cause of cancer-related deaths with an estimated number of 2.3 million new cases worldwide according to the GLOBOCAN 2020 data (Łukasiewicz,2021). What's more, it's also found that differences in breast cancer occur in different race groups. For example, the incidence rate of breast cancer before age 45 is higher among Black women than White women, whereas between the ages of 60 and 84, breast cancer incidence rates are strikingly higher in White women than in Black women (Yedjou,2019). So, figuring out the root cause of the racial differences in breast cancer can help us to use more efficient treatments targeting patients of different races.

In this research, TCGA data is used to examine the differences of gene mutations in different racial groups. In particular, TCGA clinical data, mutation data, and protein data was used to compare the age of diagnosis, survival rate according to days from diagnosis, and gene mutations of different race groups.

According to the results, it is found that differences between race groups exist in the age of diagnosis, survival rate, and different genes in mutation. However, there are no significant differences in gene expression between white and non-white races.

Methods:

In this research, TCGA clinical data, MAF data, and rna data is accessed through "TCGA-BRCA" in data analysis. Firstly, TCGA clinical data was introduced and processed to waive out any NA values which may impact data analysis. It is then used to draw boxplot though "ggplot2" according to four different races exists in the data set: Asian, Black or African American, White and American Indian or Alaska Native.

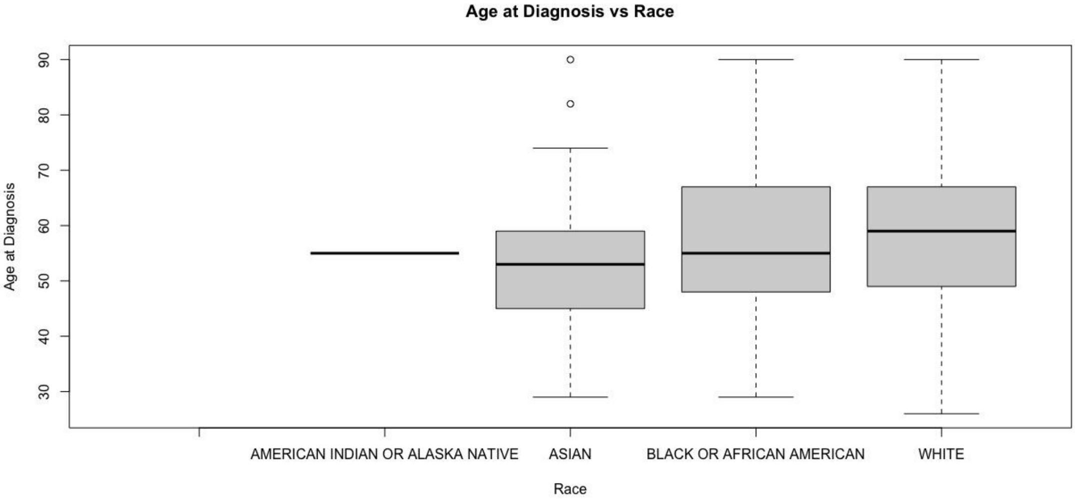


Figure I. Boxplot showing age at diagnosis for different race groups.

From the graph, we can see that the white patients tend to have the highest average age when diagnosis, black patients the second highest, and Asian patients the youngest when diagnosis. Because the American Indian or Alaska Native category only has one patient, so we cannot conclude anything for this race from the data we have.

Then, "survival" and "survminer" package was used to draw Kaplan-Meier Plot for these race groups using clinical data.

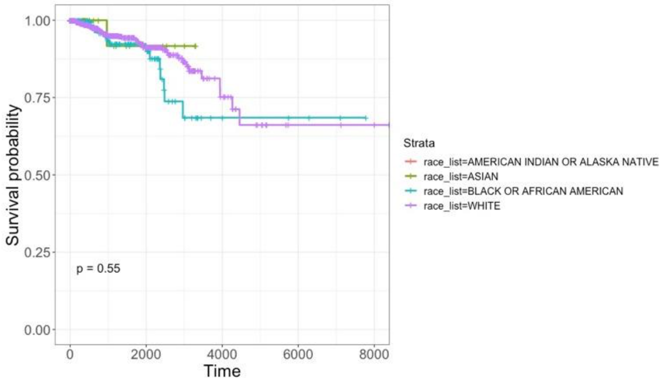


Figure II. Kaplen-Meier Plot showing the survival probability overtime for different race groups.

We can see from the graph that the Asian patients have the highest survival probability between the first 1000 days and after 2000 days. The white patients have a higher survival probability than black patients before 4000 days, after 4000 days, black patients have a higher survival probability. What's more, survival probability for the white and black patients both started to drop after 2000 days.

Moving on to MAF data, "Masked Somatic Mutation" type of MAF data was accessed to data analysis and was processed to waive out the NA values in race info. Then, an oncoplot was made for all the patients showing the top 10 genes mutated.

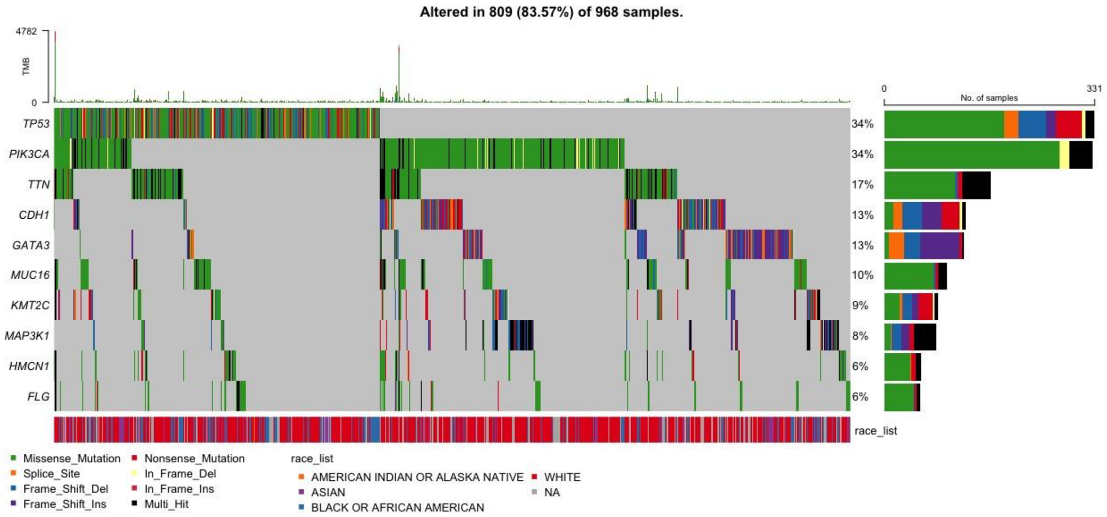


Figure III. Oncoplot showing the top 10 mutated genes for all patients.

As it's hard to compare between race groups, the MAF data was subsetted into four smaller MAF dataframes: Asian MAF, Black MAF, White MAF and Native MAF. And for each small MAF dataframe, oncoplots were made for top 10 mutated genes. As native MAF only contains a single patient, so it is waived from the research.

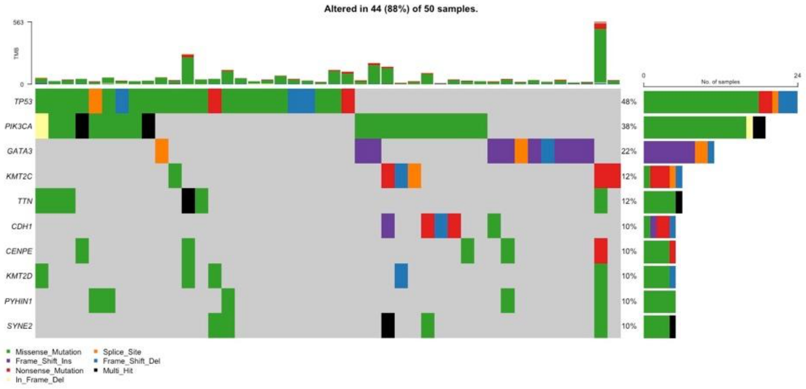


Figure IV. Oncoplot showing the top 10 mutated genes for Asian patients.

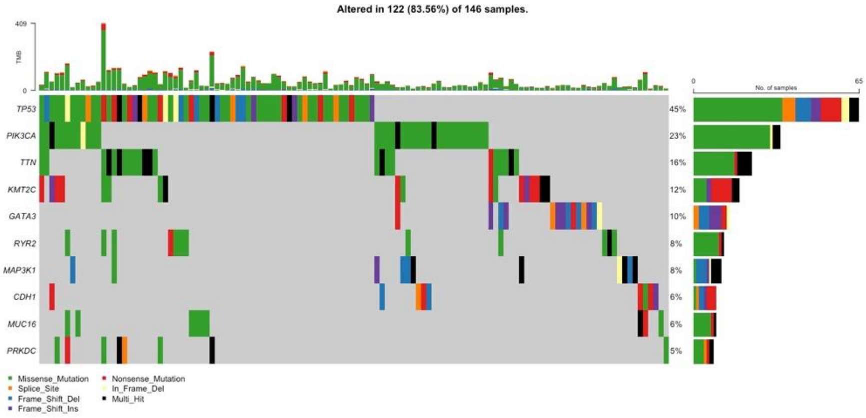


Figure V. Oncoplot showing the top 10 mutated genes for Black patients.

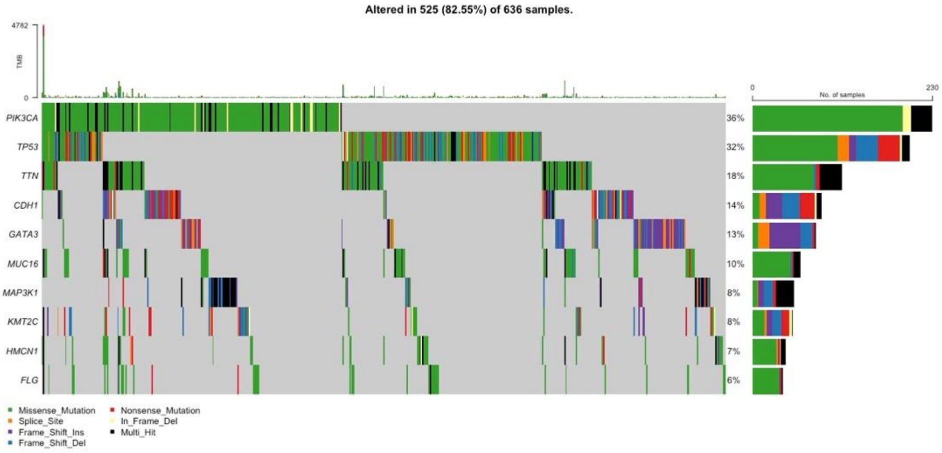


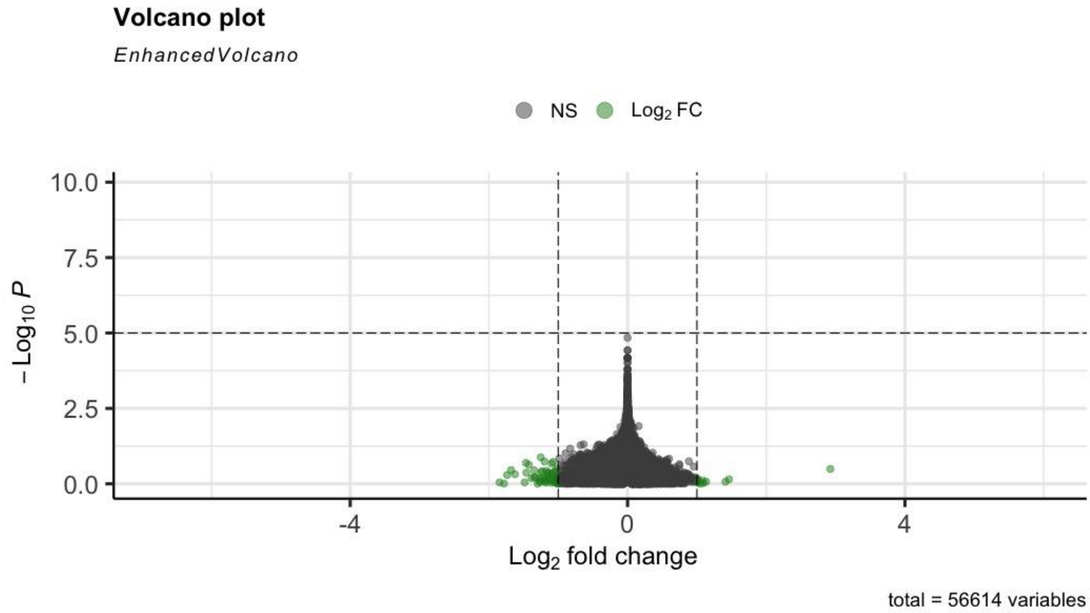
Figure VI. Oncoplot showing the top 10 mutated genes for White patients.

From the oncoplots of different races, it's noticeable that although the top 10 mutated genes for different races share similarity, there are differences in their frequency. For example, TP53 gene is the most mutated gene for Asian and Black patients, however, PIK3CA was the most mutated gene in white patients. What's more, mutations in FLG genes are more frequent in white patients, SYNE2 gene mutations are more likely to happen to Asian patients, and we can find uniqueness of the genes mutated in each race group, suggesting the possibility of different mechanisms in race groups can be further explored.

Finally, "Gene Expression Quantification" RNA data was imported. Three dataframes, clinical data, gene data, and gene expression data were extracted for further analysis. The clinical data was processed to waive out the patient with NA information, and the corresponding data was also eliminated from gene expression dataframe. As the white patients are the most abundant group in the dataframe, in order to simplify analysis, we created a category consisting of white and non-white to examine which gene has significance in white patients than non-white patients. Then, a differential expression was conducted on white/non-white with 2 co-variables pathologic stage and definition types of tumors. A volcano plot was drawn using the result.

Figure VI. Volcano Plot showing the gene expression differences between white and non-white patients.

From the plot, we can see that no gene has significant differences in expression among white patients and non-white patients, suggesting that there is still need of more strong evidence in different gene expression between races.



Results:

From this research, we found that there are some noticeable differences in breast cancer between race groups which include Asian, White and Black. The differences are shown in multiple aspects, including age of diagnosis, survival probability overtime, genes that are mutated. White patients tend to have the highest average age when diagnosis, black patients the second highest, and Asian patients the youngest when diagnosis, while Asian patients have the highest survival probability between the first 1000 days and after 2000 days. The white patients have a higher survival probability than black patients before 4000 days, after

4000 days, black patients have a higher survival probability. What's more, survival probability for the white and black patients both started to drop after 2000 days.

In gene mutation, races groups are to be found to have different tendencies. For example, mutations in FLG genes are more frequent in white patients, and SYNE2 gene mutations are more likely to happen to Asian patients.

However, when it comes to gene expression, there is no significant difference to be found between white and non-white patients, suggesting more research on this topic needed.

Discussion:

The result of this research aligns with the existing study showing that the mortality rate of breast cancer remains significantly higher among Black compared to White women and other ethnic groups, and black women tend to be diagnosed with breast cancer at a younger age than White women (Yedjou,2019). According to existing research, it is likely due to the fact that minority women, especially African Americans, Hispanic whites, and American Indians, are more likely to be diagnosed at more advanced stages of breast cancer (1–3), less likely to receive recommended treatment regimens (1, 3, 4), and more likely to have worse survival outcomes (1–3, 5–7) compared with non-Hispanic white patients (Lu,2015).

However, there is no significant difference in gene expression discovered, suggesting more research on this field in the future in order to find the root cause of the differences. Of course, cultural and lifestyle differences should also be taken into consideration.

What's more, the study has its limitations as the data was not abundant enough to reach conclusions. In the data set used, there is much more data for white patients than Asian and Black patients, and there is only one sample for the American Indian and Alaska Native race, which is too few to draw conclusions. If we can conduct data analysis using more abundant data, we can draw more reliable conclusions in the future.

References:

Lu Chen, Christopher I. Li; Racial Disparities in Breast Cancer Diagnosis and Treatment by Hormone Receptor and HER2 Status. *Cancer Epidemiol Biomarkers Prev 1* November 2015; 24 (11): 1666–1672. https://doi.org/10.1158/1055-9965.EPI-15-0293

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

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