Ethnicity and Breast Cancer

Introduction:

The Cancer Genome Atlas (TCGA) is a collection of multi-platform molecular profiles of thousands of tumours from many types of cancers. TCGA was created to share accessible, usable data with researchers worldwide. Since 2006 it has grown and helped advance cancer biology (Liu et al., 2018). Multi-omic data analysis interprets data at the genomic, epigenomic, transcriptomic, proteomic and metabolomic levels. The use of multi-omic data has improved insights into cellular functions and aided medical research(Subramanian et al., 2020).

Breast cancer is a leading cause of death in women globally. It is a metastatic cancer, meaning it transfers to other organs. Many genes have been identified as causes of breast cancer. These include but are not limited to Breast Cancer Associated genes 1 and 2 (BRCA1/2), Human Epidermal Growth Factor Receptor 2 (HER2), and many more. Other risk factors for breast cancer include lifestyle, oestrogen levels, family history and ageing (Sun et al., 2017).

A disparity between the survival of different populations is seen within the healthcare and medical research fields in general(Sachdev et al., 2010). It is therefore imperative to ensure that ethnicity is considered when investigating the survival rates of any disease, including breast cancer. With different gene mutations having varying frequencies within different ethnicities it is important to use this information to better understand why there is a difference in survival rates.

This investigation looks into the survival outcomes and mutation rates of Hispanic and Latino patients, in comparison with not Hispanic or Latino patients. Data was sourced

from the TCGA public data set and was analysed in R. Survival probability and mutation rates in Hispanic and Latino patients are significantly lower than that of patients of other ethnic backgrounds.

Methods:

To investigate the relationship between race and gene mutation and survival in breast cancer, breast cancer clinical data and MAF data were accessed from TCGA, using "TCGA-BRCA" as the project. Two oncoplots and lollipop plots of gene mutation data, one from the Hispanic and Latino patients, and the other for other ethnicities were created using the R package maftools. A Kaplan Meier Survival plot was created using survminer package.

Results:

We saw a strong difference in survival between Hispanic and Latino patients when compared to patients of other ethnicities. Hispanic and Latino patients were associated with significantly greater survival probabilities than other patients.

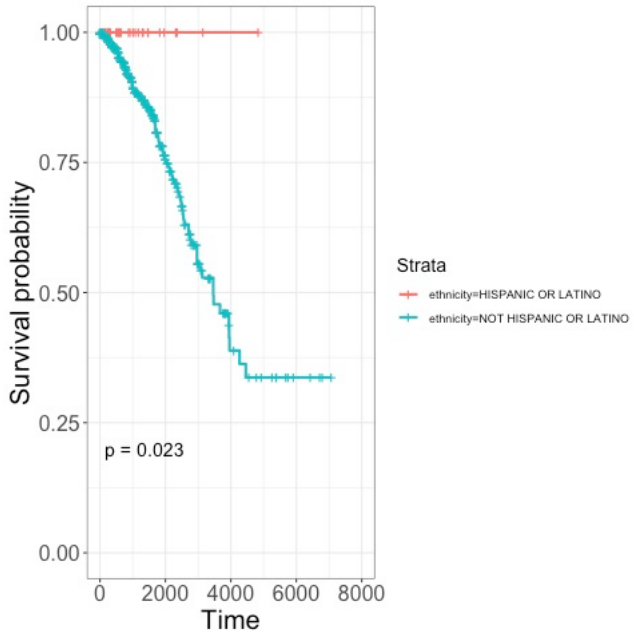

Figure 1. Kaplan-Meier survival plot showing that Hispanic or Latino patients experience significantly increased survival times in comparison to patients of other ethnicities. P-value of 0.023 denotes the statistical significance of the found results.

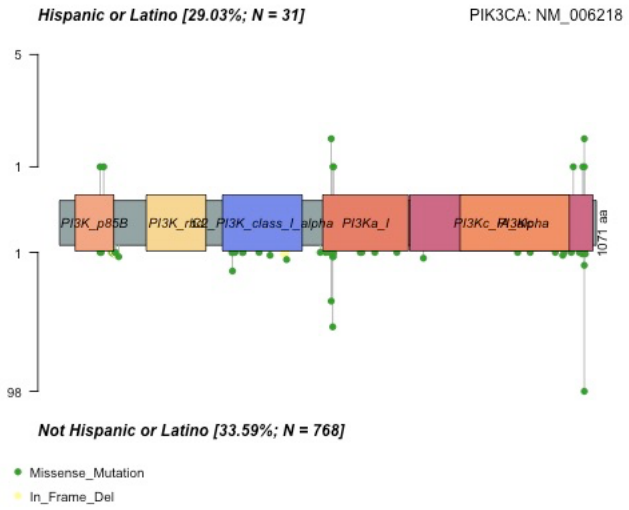

Figure 2.1. Lollipop plot showing that patients of ethnicities other than Hispanic or Latino have significantly more mutations in the PIK3CA gene than Hispanic or Latino patients.

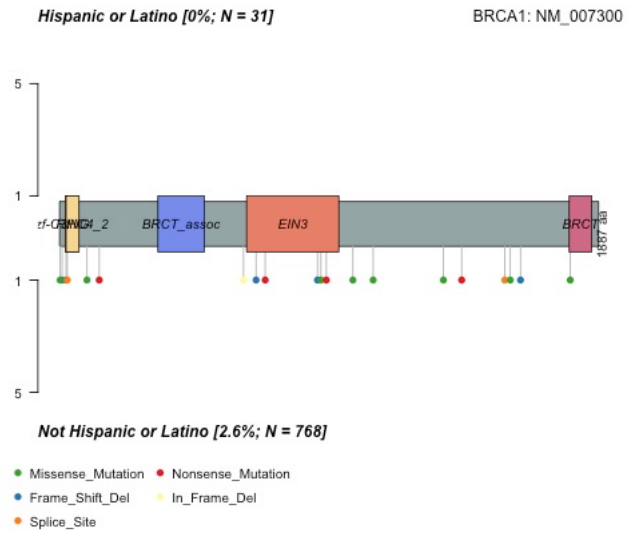

Figure 2.2. Lollipop plot showing that patients of ethnicities other than Hispanic or Latino have significantly more mutations in the BRCA1 gene than Hispanic or Latino patients.

The research also resulted that the frequency of mutation in the two investigated genes:

BRCA1 and PIK3CA is much greater in patients that are not Hispanic or Latino (Figure 2.1,

Figure 2.2).

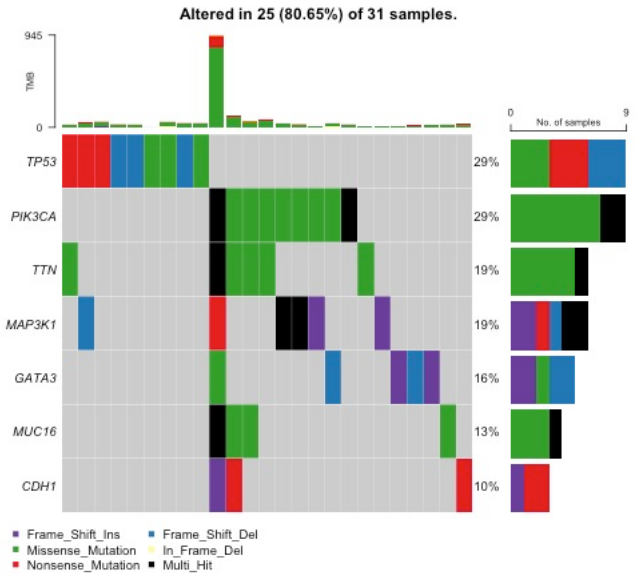

Figure 3.1 Oncoplot showing that the frequency of gene mutation of the genes TP53, PIK3CA, MAP3K1, GATA3, MUC16 and CDH1 is greatest in TP53 and PIK3CA in Hispanic and Latino patients.

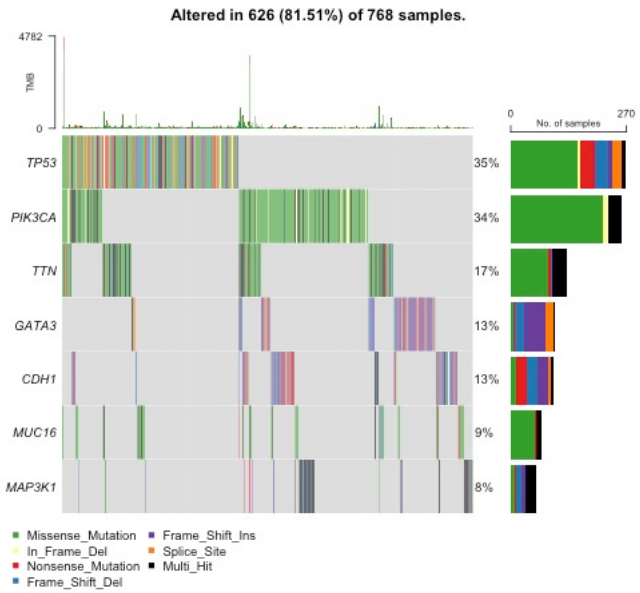

Figure 3.2 Oncoplot showing that the frequency of gene mutation of the genes TP53, PIK3CA, MAP3K1, GATA3, MUC16 and CDH1 is greatest in TP53 and PIK3CA in patients that are not Hispanic or Latino.

Through the oncoplots (Figure 3.1, Figure 3.2) it can be seen that the two ethnic groups both have the most frequent mutations in the same genes, TP53 and PIK3CA. However, Hispanic and Latino patients have the fewest mutations in CDH1 and other patients have the fewest mutations in MAP3K1.

Discussion:

The results support the hypothesis of this investigation, suggesting that Hispanic and Latino patients do not have lower survival probabilities and gene mutations in relation to breast cancer. This agrees with existing literature that found that there was a significant difference in the survival probabilities of different ethnic groups, more specifically one concluded that there is a correlation between African ancestry and developing triple-negative tumours (Sachdev et al., 2010).

Other existing literature directly disagrees with the findings of this investigation. A study done in New Mexico noticed that Hispanic women are less likely to survive breast cancer than non-Hispanic women, however, it was speculated that this was due to adverse tumour prognostic characteristics(Hill et al., 2010). Our findings may have opposed this due to a lack of investigation into at what stage the cancer was diagnosed, and other factors.

In the future, a larger sample of Hispanic and Latino patients could be investigated to ensure the reliability of the results. Also, investigations into other ethnic groups would enable the identification of the ethnic groups with the lowest survival probabilities of breast cancer. This information could be used to tailor the treatment of breast cancer, or used as a preventative measure. A further area of future research could be to investigate if the stage of cancer at diagnosis is impacting the results of this investigation.

References:

Hill, D. A., Nibbe, A., Royce, M. E., Wallace, A. M., Kang, H., Wiggins, C. L., & Rosenberg, R. D. (2010). Method of detection and breast cancer survival disparities in Hispanic women. *Cancer Epidemiology, Biomarkers & Prevention*, *19*(10), 2453–2460. https://doi.org/10.1158/1055-9965.epi-10-0164

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Hu, H., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., … Mariamidze, A. (2018). An integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, *173*(2). https://doi.org/10.1016/j.cell.2018.02.052

Sachdev, J. C., Ahmed, S., Mirza, M. M., Farooq, A., Kronish, L., & Jahanzeb, M. (2010). Does race affect outcomes in triple negative breast cancer? *Breast Cancer: Basic and Clinical Research*, *4*, 117822341000400. https://doi.org/10.1177/117822341000400003

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, *14*, 117793221989905. https://doi.org/10.1177/1177932219899051

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., & Zhu, H.-P. (2017). Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*, *13*(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

General Concepts

1. What is TCGA and why is it important?

TCGA is the Cancer Genome Atlas, it consists of clinical information, molecular analyte metadata and molecular characterisation data. It is important as it has increased the availability and access of data, enabling many advancements in many cancer fields.

2. What are some strengths and weaknesses of TCGA?

Strengths: The large accessible data reduces costs of research as data is already collected, it also increases the sample size for research as it provides access to more data than what an individual researcher would be able to generate.

Weaknesses: There is a high error rate in the data, there are missing genes in the data so it may not be applicable to all cancer research or may result in missed correlations between genes. Also there are issues with data privacy (most of which have been overcome by TCGA)

Coding Skills

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

Git status

Git add .

Git commit -m “”

Git push

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

Library(package)

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

if(!require(package)) {BiocManager::install("package")

}

4. library(survival) What is boolean indexing? What are some applications of it?

It converts data to true or false based off of information given, for example it can be used to categorise data such as age data into young and old. Or it can be used to remove invalid patients, for example if they are above a specific age, or if they have an NA value in a column.

5. Draw a mock up (just a few rows and columns) of a sample data frame. Show an example following and explain what each line of code does.
    a. an `ifelse()` statement
    b. boolean indexing

Patient #	Age	vital_status	death_event	
1	47	Alive	False	
2	83	Dead	True	
3	12	Alive	False	
4	NA	Alive	False	

ifelse(vital_status == "Alive", death_event <- FALSE, death_event <-TRUE)

This ifelse statement looks at the data in the vital_status column and determines whether the contents of the death event column, and fills that column

na_mask <- ifelse(is.na(age), F, T)

This Boolean mask stores the age values of data set as true and false, if there is a non-integer value present it is false.

data <- data[,na_mask]

This applies the mask to the columns of our data set