Alex Zhang

3/10/2023

PIK3CA and Age

Introduction:

TCGA, otherwise known as the Cancer Genome Atlas, is a public dataset that is a collaborative effort between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). TCGA contains multi-omic data from a plethora of cancers. One of these included cancers is breast breast cancer, which we will be taking a look at in our analysis. The multi-omic nature of TCGA will be used to our advantage to analyze biological reasons behind breast cancer development. Breast cancer is the most common form of cancer in women worldwide and is the second leading cause of cancer deaths in women. Some known risk factors for breast cancer include age, family history, genetics, lifestyle, as well as substance use/abuse. Unfortunately, in the US, ~1 in 8 women will develop breast cancer over the course of their lives. The goal of this research was to identify the associations (if any) between breast cancer prognosis and age in the TCGA dataset. We will focus on gene expression and vital status. In order to analyze the data, we will use a combination of techniques, some of which include survival analysis and differential expression analysis. Through this, we discovered that age was, indeed, associated with the prognosis of breast cancer.

Methods:

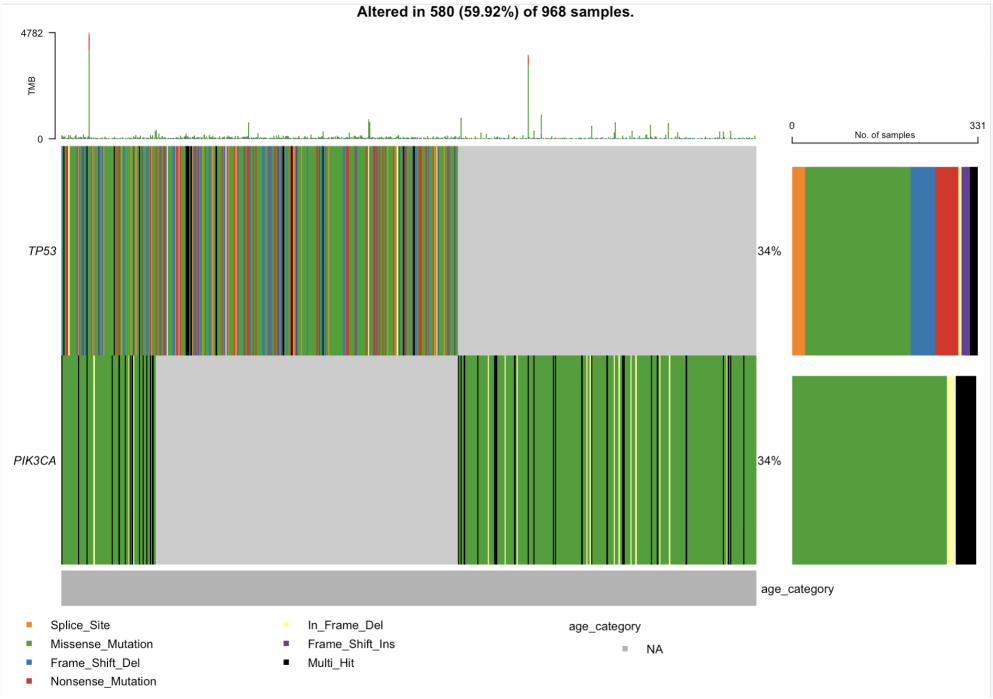
All of our analysis was done using the programming language R. To begin, we first downloaded and loaded all of our required packages and other additional packages for testing. These included BiocManager, dplyr, ggplot2, TCGAbiolinks, maftools, SummarizedExperiment, DESeq2, survival, tidyr, survminer, data.table, Biobase, BiocGenereics, GenomicRanges,

EnhancedVolcano, and ggrepel. After setting the working directory to the folder with all of our data, we then read in our data. Using GDCquery and GDCprepare, we read in the TCGA clinical dataset, rewriting it as “brca_clinical_data.csv” to make it more readily accessible. Using GDCprepare_clinic, we also defined datasets for clinical drug and radiation treatment data. After that, we loaded in a mutation frequency dataset (maf). Using GDCquery, we created datasets maf_query and maf_object. With maf-related data loaded, we then moved onto loading data from the dataset rna_se. This dataset was directly uploaded to R from a local file. After defining variables for rna_counts and rna_genes, we then loaded DESeq (differential expression analysis) for the variables vital_status and age_at_diagnosis. With this, all of our necessary variables and data have been loaded. After this, we started working on creating plots. Using the maf data, we were able to mask the age variable to split it. Using this, we created oncoplots, cooncoplots, and lollipop plots. After this, we are able to use BRCA clinical data in order to create Kaplan-Meier survival curves. Following this, we can also make regular boxplots in R using the same dataframes derived from the BRCA clinical data. Finally, we are able to create volcano plots using the prepared DESeq data and the rna_se gene data.

Results:

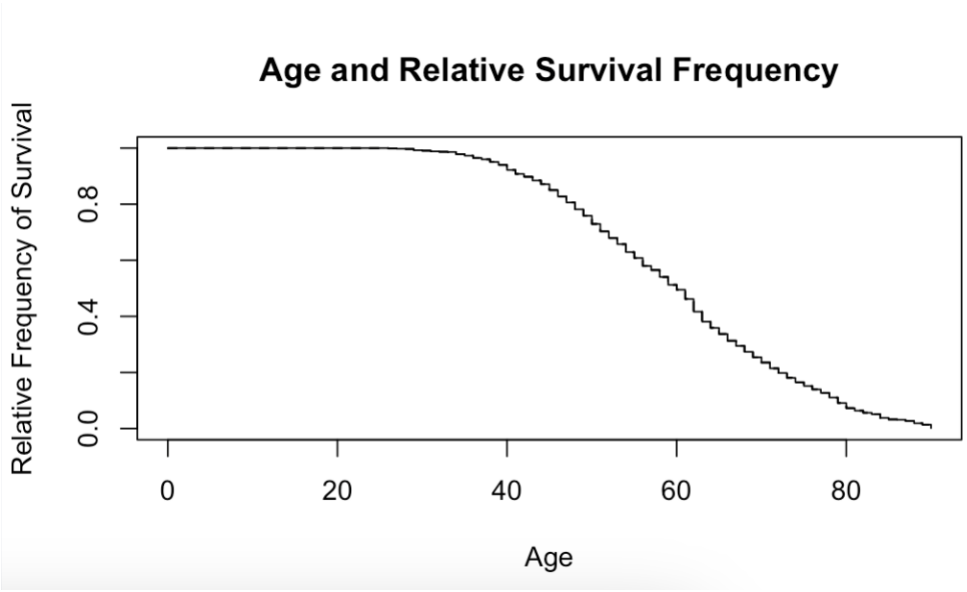
Throughout our experiment, we’ve created many visualizations of data to help us come to conclusions. To begin, a lollipop plot was made to find a gene suitable for analysis. For the sake of convenience, the lollipop plot as shown below is limited to only show the top 2 most mutated

genes in the dataset.



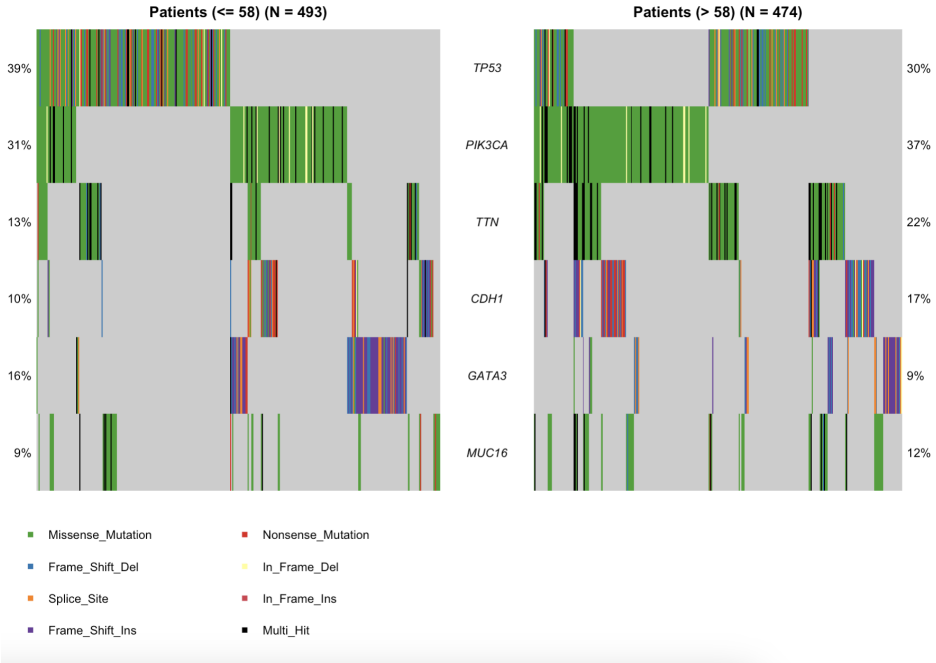
Graph 1.1, Lollipop Plot of Most Mutated Genes:

Using this, we were able to determine our gene of interest/analysis, PIK3CA. After taking our Kaplan-Meier survival curve, we receive this data:



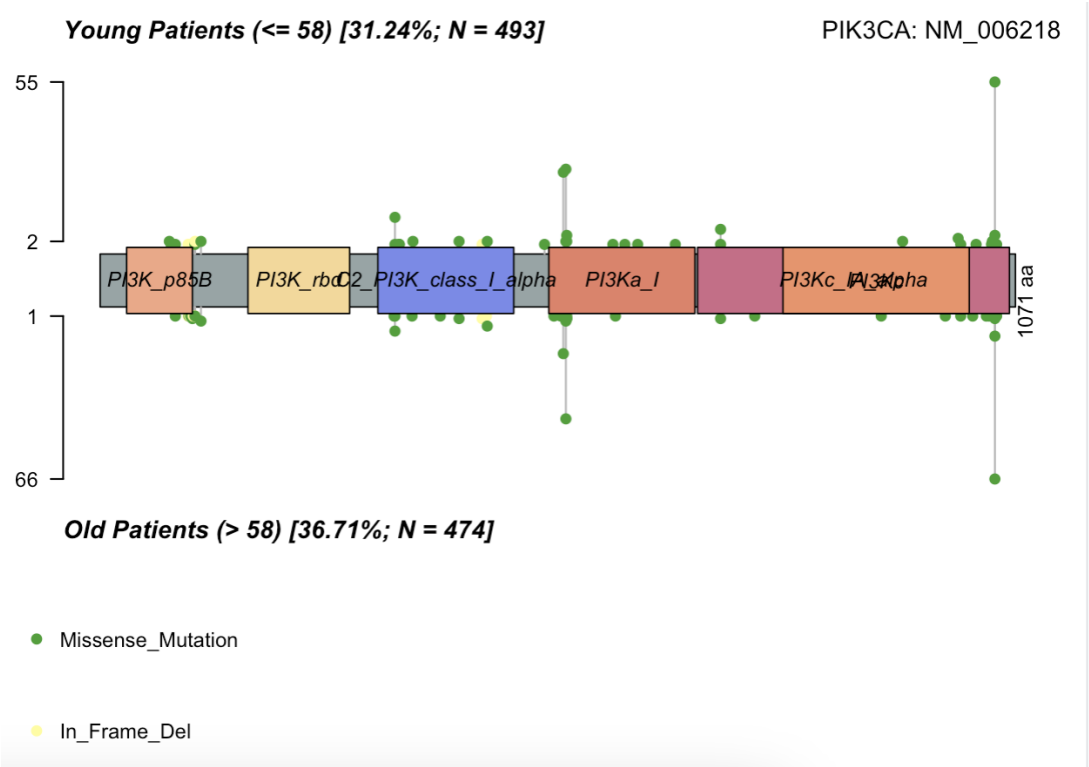
Graph 1.2, Kaplan-Meier Survival Curve of Age and Survival:

From this, we can conclude that in general (as this doesn't take into account other factors), it seems as if there is a correlation between age and survival: the older the patient, the less likely they are to survive. After this, we can pull another graph to compare the mutation patterns of gene PIK3CA in between old and young patients:



Graph 1.3, Co-Oncoplot of top 5 Most Mutated Genes and Age:

From this graph, there isn't much we can conclude about gene mutation status and age. The data for PIK3CA seems to only differ very slightly between the 2 different age classes. However, there is a slight difference between the overall mutation ratios between old and young patient classes. Old patients see a ~37% mutation rate overall, while young patients see a ~31% mutation rate overall. However, it would be hard to call this change significant, as the values are still somewhat close to each other. Finally, for our last graph, we have a co-lollipop plot:



Graph 1.4, Co-Lollipop Plot of PIK3CA and Age

From this graph, there isn't much we can conclude about gene mutation status and age. The data for PIK3CA seems to differ very very slightly between the 2 different age classes. However we can see where exactly these mutations are taking place for both young and old patients. In this case, it seems as if the majority of the mutations are clustered around C2_PI3K_class_I_alpha, PI3Ka_I, and PI3Kc_P_alpha for both younger and older patient data.

Discussion

Through our experiment, it can be determined that age seems to affect (to some degree) the mutation status of PIK3CA. As shown in graph 1.3, we can determine that age is correlated with a higher likelihood of PIK3CA mutation. In addition to this, we can determine that the majority of all of these mutations, regardless of age, occur around C2_PI3K_class_I_alpha, PI3Ka_I, and PI3Kc_P_alpha. Finally, we can also conclude that age is negatively correlated to survival rate;

as in, as age increases, relative rate of survival decreases. Compared with the results of other studies such as the one done by Cizkova et. al., the results for this study are fairly similar. Cizkova et. al. also found that PIK3CA mutation was highly associated with longer metastasis-free survival in the overall population. The p-value of this relationship is 0.0056, making the correlation significant and essentially confirming the importance of PIK3CA mutations in breast cancer. In another study done by Cho et. al. titled "PIK3CA Mutation as Potential Poor Prognostic Marker in Asian Female Breast Cancer Patients Who Received Adjuvant Chemotherapy", the PIK3CA mutation was found in 78 (49.4%) of 158 samples. This study also used the TCGA database, and essentially concluded the PIK3CA (as well as a few related genes) might be relevant to poor prognosis in BC subsets, especially in Asian women. This also slightly aligns with the findings of our own experiment. The paper "Analysis of PIK3CA mutations in breast cancer subtypes" by Arsenic et. al. may shed some light onto the biological properties of PIK3CA and why it is so important to cancers. PIK3CA is a central element of the cell signaling pathway which is involved in a variety of processes such as cell proliferation, survival, and growth. The study essentially found that, in specific exons (such as 9 or 20), mutations can have drastic, cancer-causing effects. To summarize, there are many other academic studies whose conclusions mirror ours. In addition, the gene PIK3CA is biologically important to the development of cancers because it is integral to many regulatory processes that prevent cancer and keep the cell in normal functioning order.

References

Cizkova, M., et. al, 2012 Feb 13, "PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups," Breast Cancer Res. 2012; 14(1): R28, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3496146/

Arsenic, R. et. al., 2014 Jan 22, “Analysis of PIK3CA mutations in breast cancer subtypes,” Appl Immunohistochem Mol Morphol. 2014 Jan;22(1):50-6, https://pubmed.ncbi.nlm.nih.gov/24471188/

Cho, Y. A., et. al., 19 April 2022, “PIK3CA Mutation as Potential Poor Prognostic Marker in Asian Female Breast Cancer Patients Who Received Adjuvant Chemotherapy,” Curr. Oncol. 2022, 29, 2895–2908, https://doi.org/10.3390/curroncol29050236