Title: The Role of TP53 Co-Mutations in Breast Cancer Patient Survival.

Abstract: In this study, the relationships between protein production and RNA expression of TP53, MAP3K1, GATA3, and CDH1 genes were explored in breast cancer patients from The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC) patient datasets. Additionally, survival rates of MAP3K1 mutation-positive (MAP3K+), GATA3+, and CDH1+ breast cancer patients who possessed TP53 mutations were compared to patients in these cohorts without TP53 mutations. Ultimately, it was found that TP53 and CDH1 co-mutations demonstrated a higher length of survival post-diagnosis, but the statistical significance of this finding is questionable ($p > 0.05$). It was also found that a negative relationship exists between TP53 protein and GATA3/MAP3K1 RNA expression, suggesting a potential biochemical relationship between these genes.

Introduction: Multi-omic data analysis is defined as the scientific analysis that examines data that spans areas like the proteome (protein makeup), genome (DNA), and transcriptome. Specifically, this type of data analysis allows researchers to highlight patterns in genetic mutations and disease outcomes, allowing clinicians to make more informed choices based on an individual's biology (Suravajhala et al., 2016). Furthermore, this type of analysis allows scientists to track mutations through the central dogma – from DNA, to RNA, to protein production. As it continues to grow in power, scientists are looking to leverage multi-omic analysis to tackle diseases that were once untreatable, like breast cancer.

Breast cancer is one of the most common cancer types diagnosed in women from developing countries (Donepudi et al., 2014). It represents about 22.9% of all female cancers, with approximately one in eight women developing breast invasive carcinoma in their lifetime – however, 5-year survival of this disease, specifically for white populations, hovers around 90%,

and 10-year survival is about 85% (Ssentongo et al., 2019). Given the advancement of multi-omic analysis, there has been increased advocacy for individualized breast cancer patient care, as multi-omic analysis allows clinicians to begin tailoring treatments based on considerations like an individual's genome or protein makeup.

In this study, trends between certain breast cancer mutation types were observed in public patient datasets from The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC). Specifically, our study looked at a selection of five genes and their co-occurrence, correlations, and potential interactions. Our first choice was TP53, coding for the p53 protein, which is responsible for tumor suppression and cell division regulation. Performing a brief somatic interaction plot analysis, TP53 RNA expression was seen to be mutually exclusive with the following protein-coding genes: MAP3K1 (responsible for cell migration/survival), GATA3 (cell development/lymphoid cell functions), CDH1 (tumor suppression/cell adhesion), and PIK3CA (cell growth). Past literature has investigated the individual purposes of a few of these genes; for example, TP53 and GATA3 mutations (i.e. non co-occurant) have been previously implicated with endocrine resistance in luminal early breast cancer (Grote et al., 2021). Additionally, TP53 and CDH1 mutations have been found to influence polymorphisms within familial breast cancer cases in parts of the world (Rahim et al., 2022). However, this study aims to be a more holistic overview of the above five genes – in essence, the co-occurence and differential expression of these five genes has not been widely explored together.

Using a somatic interaction plot analysis, TP53 mutations were seen to be mutually exclusive with each of the other genes. As such, our study sought to compare TP53 positive and negative survival in each of the other genes; for example, perhaps, patients who carry TP53 and another mutation in a statistically significant way display different patterns of survival.

Ultimately, this study aimed to answer the question: how does TP53 influence the mutation, differential RNA expression, and protein production of other commonly mutated genes in breast cancer patients? Furthermore, does a TP53 mutation in addition to one of these other mutations influence patient survival? By performing a variety of statistical analyses on TCGA and CPTAC breast cancer patient data, we hope to answer this question and open avenues for future exploration in this realm of scientific literature.

Methods: To begin, the datasets that were used for this study were downloaded. Breast Cancer clinical data and MAF/transcriptome data were taken from the TCGA database using accession code “TCGA-BRCA”. The data for transcriptomics and proteomics analysis in python was downloaded from CPTAC using the code “Brca”.

First, a somatic interaction plot was created in R, to visualize any nonrandom associations between mutation data. The top 25 mutated genes were chosen for analysis in this plot – from this point, we determined that 4 genes demonstrated mutual exclusivity with TP53, our gene of interest: MAP3K1, GATA3, CDH1, and PIK3CA. Following this analysis, we continued our study with just these genes.

To observe the impact that a double-mutation (TP53 with one of the other three) had on patient survival, Kaplan-Meier plots were created that traced the survival rates for each of the three genes, using the functions “Surv”, “surv_fit”, and “ggsurvplot”. These plots drew on clinical data – particularly, survival time post-diagnosis. Each plot contained two groups, one with mutations of TP53 and the target gene, and one with just the target gene mutation (and no TP53 mutation). Boxplots were also created from this data, to visualize the average survival between TP53 +/- groups.

Next, an oncoplot was constructed to visualize potential overlap and clustering between patients who had each of our four genes mutated. This oncoplot was part of our “correlation” analysis, and was followed by the creation of heatmaps between our genes of interest – this was to observe the relationships between RNA expression and protein presence in our data. To create the heatmap, the Spearman correlation coefficient was calculated using the function “spearmanr” in the scipy library with its corresponding p-value. The heatmap was constructed using matplotlib and seaborn after the intersection of transcriptomics and proteomics data was found.

Results:

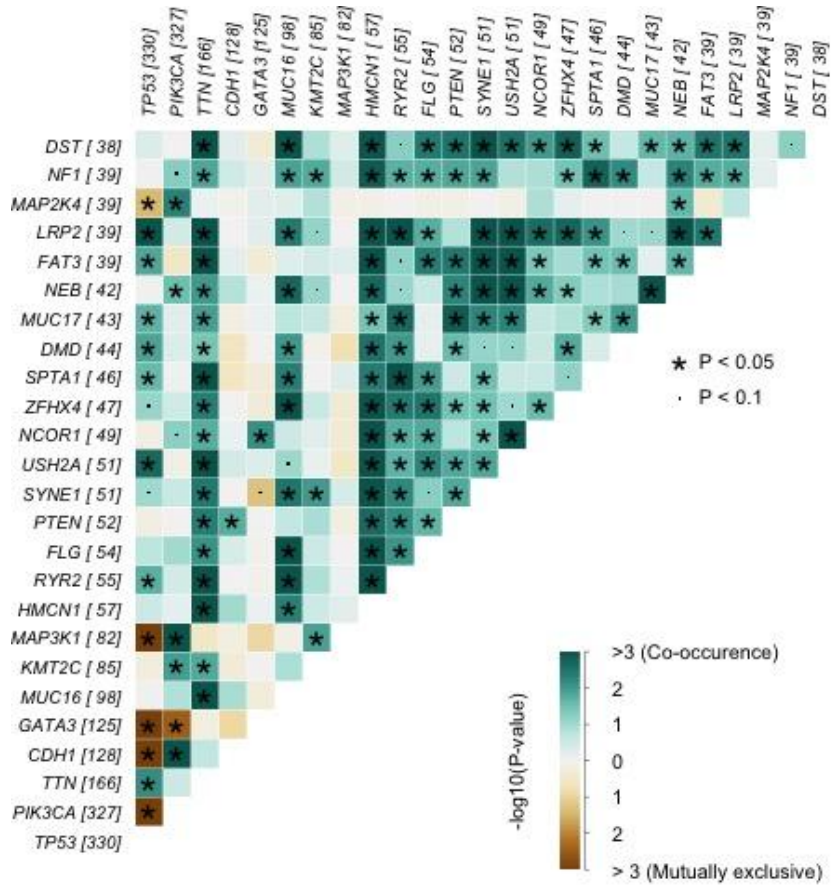


Figure 1. Somatic Interaction Plot between top 25 mutated breast cancer genes shows four genes with mutual exclusive mutations with TP53. Notice in the “TP53” column, there are 4 genes that display extreme mutual exclusivity with a statistically significant p-value ($p < 0.05$).

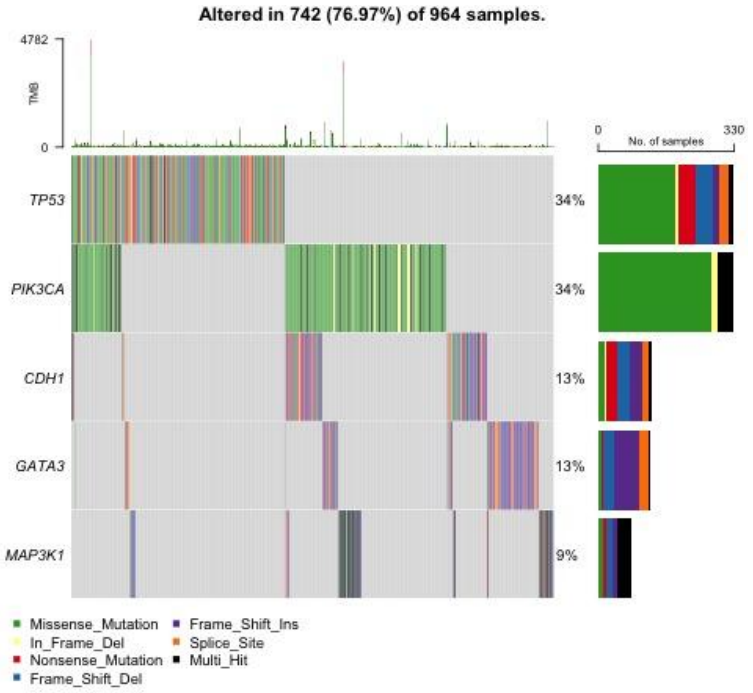
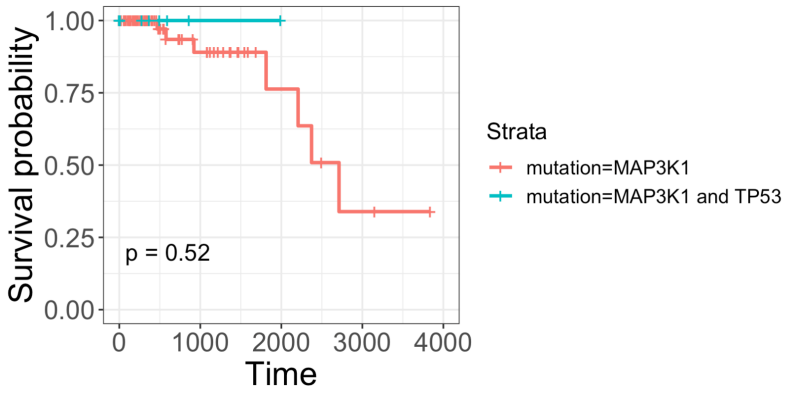
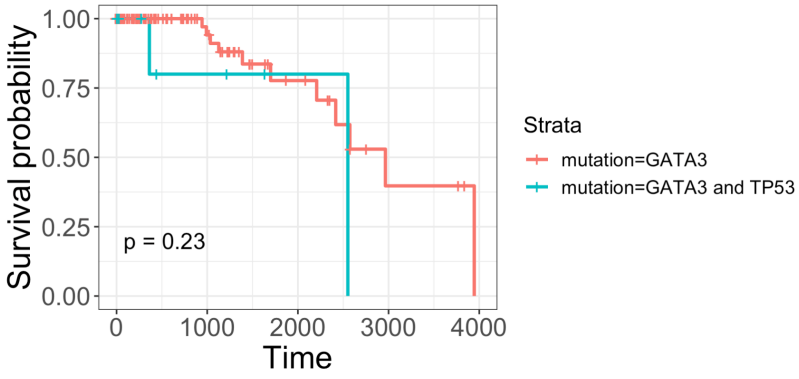
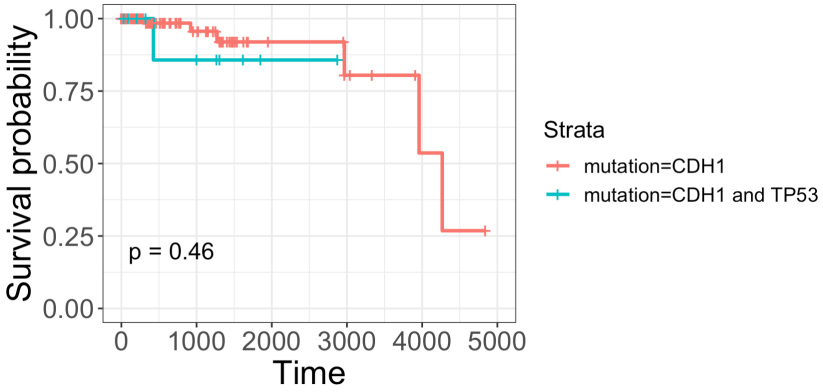
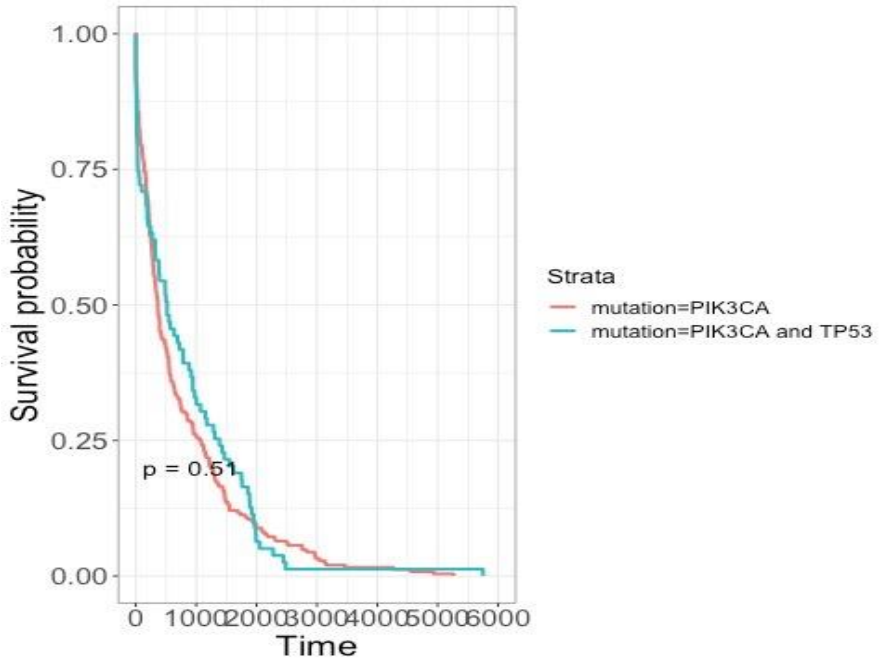


Figure 2. Oncoplot showing the clustering of patients for the four selected genes. Notice that there aren’t any large clusters of patients in any one category, but that TP53 shows slight overlap with each of the four genes, and each of the chosen genes overlaps with each other, as well.





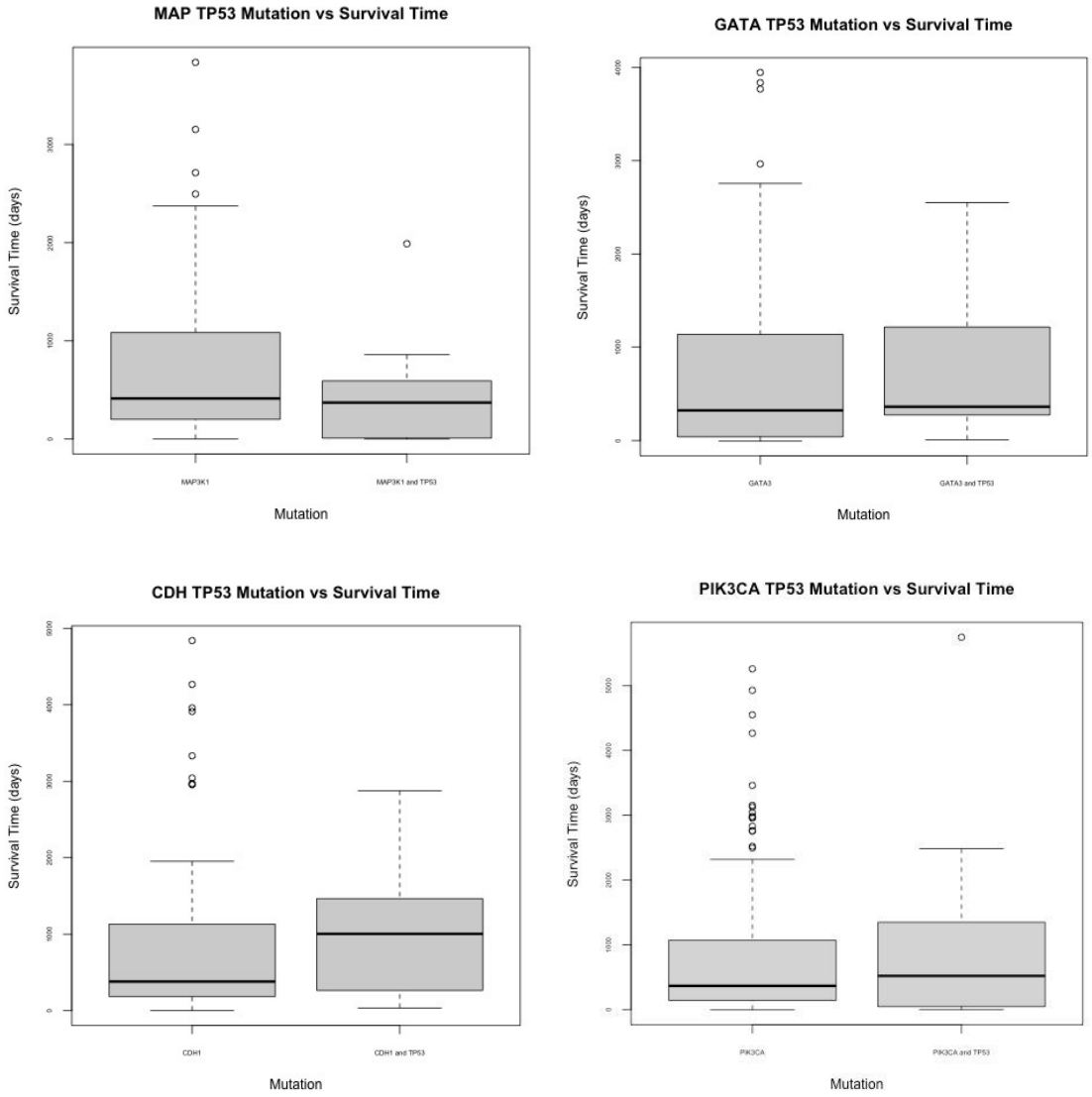




Figures 3a-d. Kaplan Meier survival plots for each of the four chosen genes, with and without TP53 mutations. Note that there seem to be differential survival trends within each figure, but

that the p-values are much higher than the statistical significance cutoff of $p < 0.05$ for all of these figures.

From Figure 3a, it can be observed that MAP3K1 mutations without a TP53 mutation lead to slightly better survival. From Figure 3b, the same can be said about GATA3 mutations (i.e. GATA3 + TP53 mutations had worse early survival than GATA3 on its own). Figure 3d shows very little difference between the survival of PIK3CA-only and PIK3CA-TP53 co-mutation patients. However, interestingly, it was observed in Figure 3c that having CDH1 and TP53 co-mutations led to slightly better survival than a CDH1 mutation on its own.



Figures 4a-d. Boxplots that display the average survival time of individuals who are TP53 mutation + or -, for the three selected genes (MAP3K1, GATA3, CDH1). These plots were

constructed using clinical survival data. Note that these boxplots are not necessarily statistically significant, as their margins of error display overlap.

The findings of Figures 3a-d are reflected in Figures 4a-d, as both MAP3K1, GATA3, and PIK3CA showed comparable survival times, but the CDH1 + TP53 group showed a better average survival time than CDH1 by itself.

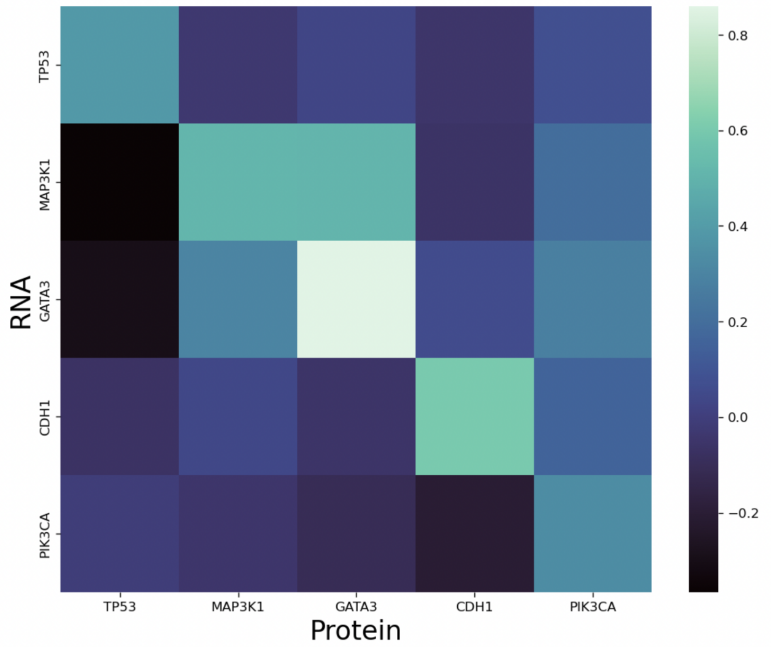


Figure 5. Heat map showing the relationship between the RNA and protein expression of “TP53”, “GATA3”, “CDH1” and “MAP3K1” genes. The correlation coefficient indicates that a low presence of TP53 protein is associated with a high expression of GATA3 and MAP3K1 gene, while there is no significant relationship between the expression of TP53 gene with the protein presence of other genes.

Discussion: One of the most notable trends found in the study lies within Figures 7-10 – specifically that the average survival time of individuals with mutations in both TP53 and CDH1 genes survived longer, on average, post-diagnosis. This is an interesting finding, as TP53 and CDH1 are both responsible for tumor suppression, and one would expect that having both of these genes mutated, rather than just one, is detrimental to an individual’s survival (Rosenthal et al., 2017). In fact, both TP53 and CDH1 mutations have been found to be correlated with increased breast cancer risk globally (Shabnaz et al., 2015). In comparison, having a TP53

mutation did not seem to influence survival if you also possessed a MAP3K1, GATA3, or PIK3CA mutation.

While these findings do demonstrate a notable result, it is important to recognize that these results may not be statistically significant. From Figures 3-6, which draw on the same data as Figures 7-10, the observed p-values are quite high (0.52, 0.23, 0.46, and 0.51, which don't meet the significance threshold of < 0.05) – presumably, this is due in part to the small sample size that was used for this study. For this reason, and because there is a lack of literature that suggests that CDH1 and TP53 co-mutations improve survival, our finding might warrant further study before any conclusions may be drawn about CDH1-TP53 relationships.

In addition to the results from our survival data, another point of discussion lies within Figure 11. In particular, it seems that a low presence of TP53 protein is associated with high expression of GATA3 and MAP3K1 genes/RNA. Essentially, this suggests that the expression of TP53 protein might influence the production of GATA3 and MAP3K1 RNA in some way, or vice versa (i.e. low production of GATA3 might cause a high presence of TP53 proteins). Looking at current literature, a recent article studying ovarian cancer proposed that GATA3 may be responsible for apoptosis (i.e. cell death) of cells that produce wild-type TP53 protein (El-Arabey et al., 2022). Put simply, without the presence of GATA3, which potentially induces cell death and halts TP53 protein production, TP53 protein levels are allowed to increase and reach high concentrations. This type of relationship is one that our heatmap demonstrates, and though the study was focused on ovarian cancer, the biology behind these mechanisms may be comparable.

In contrast, a recent case study found that MAP3K1 signaling increased TP53 stability and enhanced the ability of TP53 to induce apoptosis in cells; essentially, the presence of MAP3K1 led to an increased presence of TP53 (Fuchs et al., 1998). This contradicts our finding that low

MAP3K1 RNA is associated with high TP53 protein in some way – however, it is important to note that this study focused on a single patient, not a set of data, like our study did.
These results, as a whole, are interesting from a clinical perspective – as different patients with various combinations of mutated genes are treated, perhaps clinicians may look to push for more individualized treatments based on a patient's genomic makeup. Furthermore, understanding the way that co-mutations influence survival may provide families with a more holistic understanding of their loved one's prognosis in a clinical setting. Given that the analysis in this study showed potential relationships between co-mutated genes and survival, but that its results didn't necessarily hold statistical significance, one potential path would be to redo this study, but on a bigger dataset with more patient data. The TCGA dataset that was used in this study only contained about 1,000 patients – however, a larger group might yield a more fruitful study. Furthermore, it might be worthwhile to explore the same types of relationships as were explored in this study, but with other highly mutated genes for breast cancer patients.

Works Cited:

Donepudi, M., Kondapalli, K., Amos, S. & Venkanteshan, P. (2014). Breast Cancer Statistics and Markers. *J Cancer Res Ther.* 10(3): 506-511. DOI: 10.4103/0973-1482.137927.

El-Arabey A., Denizli, M., Kanlikilicer, P., Bayraktar, R., Ivan, C., Rashed, M., Kabil, N., Ozpolat, B., Calin, G., Salama, S., Abd-Allah, A., Sood, A. & Lopez-Berestein, G. (2020). GATA3 as a master regulator for interactions of tumor-associated macrophages with high-grade serous ovarian carcinoma. *Cellular Signaling* 68: 109539.

Fuchs, S., Adler, V., Pincus, M. & Ronai, Z. (1998). MEKK1/JNK signaling stabilizes and activates

p53. *Proc. Natl. Acad. Sci. USA* 95: 10541-10546.

Grote, I., Bartels, S., Kandt, L., Bollmann, L., Christgen, H., Gronewold, M., Raap, M., Lehann, U., Gluz, O., Nitz, U., Kuemmel, S., Eulenburg, C., Braun, M., Aktas, B., Grischke, E., Schumacher, C., Luedtke-Heckenkamp, K., Kates, R., Wuerstlein, R., …Kreipe, H. (2021). TP53 mutations are associated with primary endocrine resistance in luminal early breast cancer. *Cancer Medicine* 10(23): 8581-8594. DOI: 10.1002/cam4.4376.

Rahim, A., Jan, A., Ali, J., Khuda, F., Muhammad, B., Khan, H., Shah, J. & Akbar, R. (2022). Association of ATM, CDH1 and TP53 genes polymorphisms with familial breast cancer in patients of Khyber Pakhtunkhwa, Pakistan. *African Health Sciences* 22(3). DOI: 10.4314/ahs.v22i3.17

Rosenthal, E., Evans, B., Kidd, J., Brown, K., Gorringe, H., Orman, M. & Manley, S. (2017). Increased Identification of Candidates for High-Risk Breast Cancer Screening Through Expanded Genetic Testing. *Journal of the American College of Radiology* 14(4): 561-568.

Shabnaz, S., Ahmed, M., Islam, Islam, S., Islam, R., Al-Mamun, A., Islam, M. & Hasnat, A. (2016). Breast cancer risk in relation to TP53 codon 72 and CDH1 gene polymorphisms in the Bangladeshi women. *Tumor Biology* 37: 7229-7237. DOI: 10.1007/s13277-015-4612-7.

Ssentongo, P., Lewcun, J., Candela, X., Ssentongo, A., Kwon, E., Ba, B., Oh, J., Amponsah-Manu,

F., McDonald, A., Chinchilli, V., Soybel, D. & Dodge, D. (2019). Regional, racial, gender, and tumor biology disparities in breast cancer survival rates in Africa: A systematic review and meta-analysis. *PLOS.* DOI: 10.1371/journal.pone.0225039.

Suravajhala, P., Kogelman, L. & Kadarmideen, H. (2016). Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genetics Selection Evolution* 48(38).