Introduction

The TCGA, or the Cancer Genome Atlas Project, is a joint effort between the National Cancer Institute and the National Genome Research Institute program that characterizes multi-omics data for 33 cancer types. Multi-omics data is analysis that uses the combination of different "omics": specifically, the TCGA has genomic, transcriptomic, epigenomic, and proteomic data. As a landmark database, the TCGA has allowed for breakthroughs in data analysis in research. In this study, we used TCGA breast cancer data to look at the effects of menopause stage. Breast cancer develops in breast tissue cells, and it is the most common cancer among women with more than 2 million new cases in 2020. Around 80% of breast cancer patients are over 50 (Lukaseiwicz et al., 2021). Risk factors include age, gender, genetics (such as mutations in BRCA1 or BRCA 2), and exposure to radiation. It is a prominent health issue worldwide, and researching different risk factors can help develop better preventative healthcare and treatment options. We used clinical, genomic, and transcriptomic data to find out the age distributions between pre and post-menopausal women, whether menopause had an effect on survival, which genes were commonly mutated, and which genes were underexpressed or overexpressed. We did so through using R to conduct SNP analysis and differential expression analysis. We hope that, in looking at the correlation between menopause and breast cancer and also what particular genes may be important, it will be easier for doctors to predict who is more susceptible to breast cancer and also create more targeted treatments. We found that TTN, CDH1, and GATA3 are mutated at different rates between the two groups, and several genes are over or underexpressed.

Methods

To conduct the analysis, we first downloaded the data (accessed through TCGA with the code “TCGA-BRCA”) and relevant packages (maftools, survival, survminer, ggplot2, BiocManager, TCGAbiolinks, EnhancedVolcano, DESeq2) into RStudio. We used the clinical dataset to look at age distributions of those pre and post-menopause by creating a box plot. One represented the ages for those pre-menopause, and one represented the ages for those post-menopause. We also created a Kaplan-Meier plot using ggplot2 to compare survival between the two groups. The SNP mutation data, imported as a MAF file, was used to look at the mutations in those who had breast cancer. This was done by creating a co-oncoplot. We used the transcriptomic data to look at which factors may be relevant to breast cancer. This was done using a differential expression analysis for the RNA counts with Bioconductor. These results were displayed in a volcano plot.

Results

Using the clinical data, we plotted the ages of the patients compared to whether they were pre-menopause or post-menopause. This was important to know because, if there was a strong correlation between age and menopause status, any effect of menopause may be a confounding variable that is actually due to age.

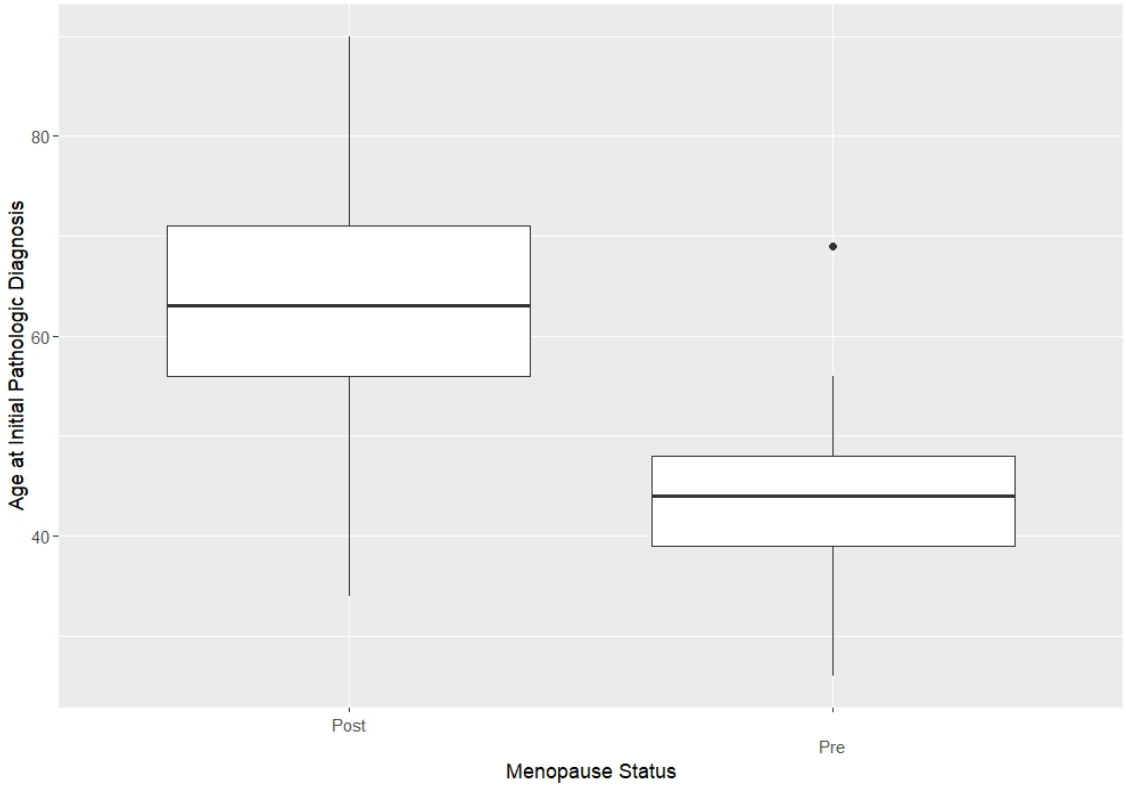

Figure 1: Box plot comparing the age at initial diagnosis of the two groups, with post-menopause on the left and pre-menopause on the right. The pre-menopause group had a significantly younger average and the overall distribution was much lower.

The box plot demonstrated that the average age of those who were pre-menopause was much younger, which made sense (Figure 1). This means that, in analyzing the effects of menopause, we must be careful in differentiating between whether it is related to age or to menopausal status.

We also created a Kaplan-Meier plot to look at the survival outcomes between the two stages.

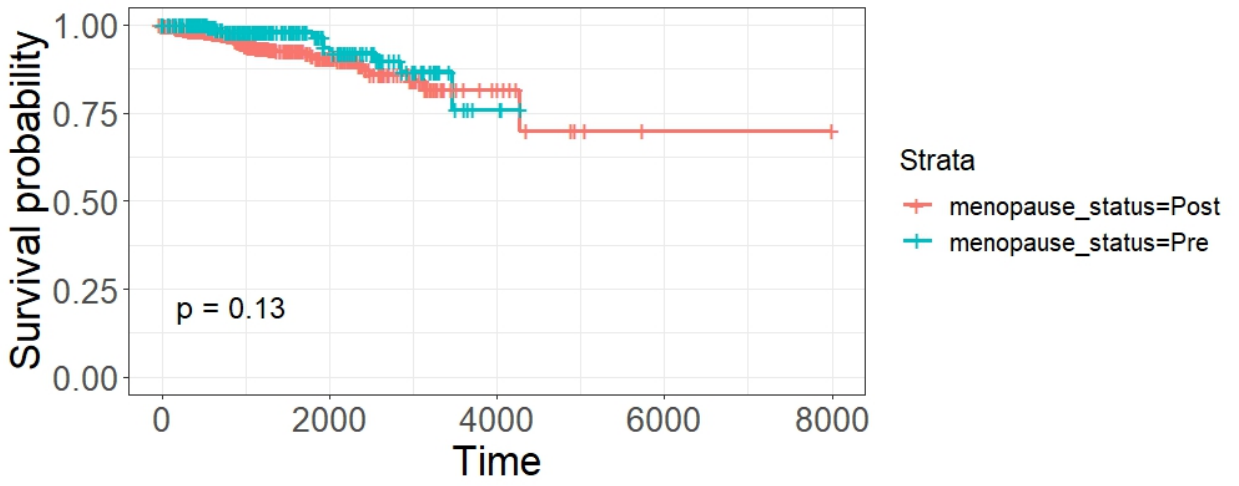

Figure 2: Kaplan-Meier survival plot demonstrating the slightly decreased survival times for the post-menopause patients compared to those who were pre-menopause. P-value of 0.13 denotes that the results were not statistically significant.

The Kaplan-Meier plot was not statistically significant because the p-value of 0.13 was greater than 0.05 (Figure 2), thus no conclusions could be drawn.

We also created a co-oncoplot of the two groups to see which genes were most mutated.

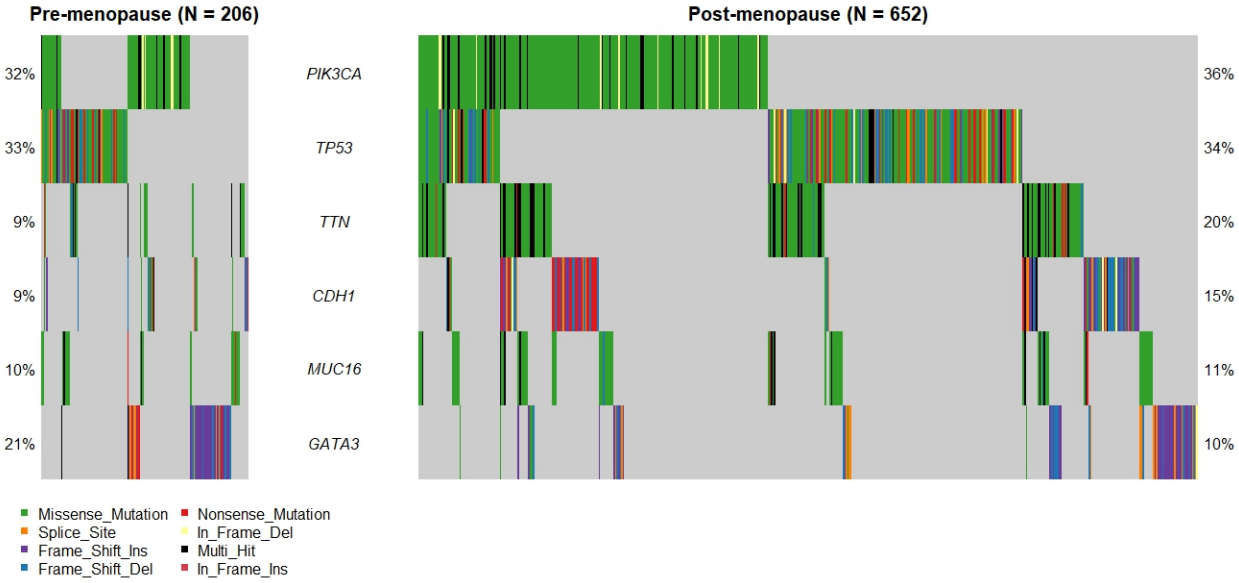

Figure 3: Co-oncoplot of the 6 most commonly mutated genes for patients who were pre-menopause and post-menopause.

For most of the genes (PIC3C!, TP53, CDH1, and MUC16), the mutation rate was similar. Further, the most common type of mutation seemed to be missense mutations for PIK3CA and TTN. However, for CDH1, for example, the most common mutation were frame shifts and multiple hits. For TP53 and GATA3, those who were pre-menopause had 9% and 21% mutated respectively, while those who were post-menopause had 20% and 10% respectively.

We also looked at over and underexpression of RNA using a volcano plot.

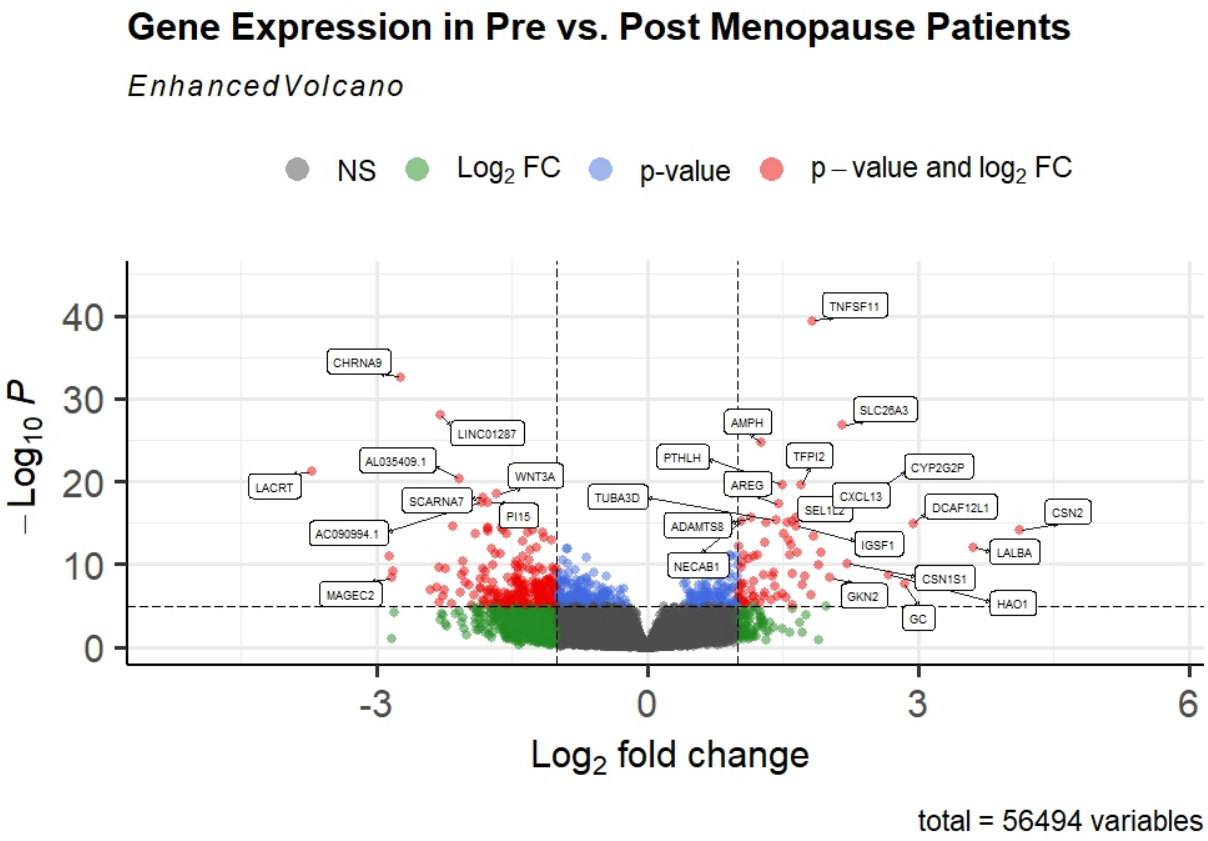

Figure 4: Volcano plot comparing RNA expression for pre versus post menopause patients, acccounting for pathologic stage. The red values are significant according to their p-value.

The red genes lying at a high enough log2 fold change and -log(p-value) were significant. The genes on the right are upregultated with the genes on the left are downregulated.

Discussion

Starting with the box plot, we found there was a strong correlation between age and menopause status (Figure 1). The average age at menopause is about 50 years (Key et al., 2001), thus it would make sense that those who are pre-menopause are younger and those who are post-menopause are older. Looking at the Kaplan-Meier survival plot, there is no significant impact on survival correlated with menopause status (Figure 2). Other studies have also looked at breast cancer risk correlated with menopause. One study concluded that women who go through menopause at an older age, such as 55, have twice the risk for breast cancer compared to those who go through menopause before 45 (Trichopoulos et al., 1972). Another research team found that premenopausal women are at a higher risk of breast cancer than postmenopausal women of the same age (Key et. al, 2001). However, these studies were similarly unable to conclude if menopausal status would affect survival, only that it would affect risk.

Looking at the most mutated genes for each group, they shared similar mutation rates for PIK3CA, TP53, and MUC16. However, the rates for TTN, CDH1, and GATA3 differed (Figure 3). TP53, PIK3CA, TTN, and GATA3 have been commonly identified as often mutated in breast cancer patients (Rajendran & Deng, 2017). TTN encodes titin, the largest known protein. It plays a role in cardiac and skeletal muscles. (Chaeveau et al., 2014). The higher mutation rate of TTN in post-menopausal women may make sense considering the studied decrease in muscle mass and strength after menopause (Matlais et al., 2009). Further, the 21% mutation rate of GATA3 in premenopausal women compared to the 10% in postmenopausal women is interesting. GATA3 is a transcription factor that helps development of the mammary gland. It acts either as a tumor

suppressor or oncogene, and is often mutated in those with breast cancer (Takaku et al., 2015). This is corroborated by findings by Hosoda et al., who found that GATA3 was significantly associated with improved disease-free survival in premenopausal women but not postmenopausal women. Although this does not address mutations, it indicates why GATA3 may be important. The same study also postulates that the estrogen-dependent growth and clinical role of GATA3 might differ with differing menopausal status, and different levels of GATA3 expression may affect endocrine responsiveness (Hosoda et al., 2014).

Lastly, looking at the under or overexpression of genes in pre and postmenopausal patients (Figure 4), we see that LINC01287 was underexpressed. This is logical as other studies have concluded that LINC01287 plays an oncogenic role in hepatocellular carcinoma cells, a form of liver cancer (Mo et al., 2018). The impact of LINC01287 could be similar for breast cancer.

It would be interesting to study how the biological impacts of menopause are correlated with the differences in mutation and gene expression. This could better our look at menopause and how we can see it as a marker for breast cancer.

References

Chauveau, C., Rowell, J., & Ferreiro, A. (2014). A rising titan: TTN Review and Mutation Update. *Human Mutation*, *35*(9), 1046–1059. https://doi.org/10.1002/humu.22611

Hosoda, M., Yamamoto, M., Nakano, K., Hatanaka, K. C., Takakuwa, E., Hatanaka, Y., Matsuno, Y., & Yamashita, H. (2014). Differential expression of progesterone receptor, FOXA1, GATA3, and p53 between pre- and postmenopausal women with estrogen receptor-positive breast cancer. *Breast Cancer Research and Treatment*, *144*(2), 249–261. https://doi.org/10.1007/s10549-014-2867-0

Key, T. J., Verkasalo, P. K., & Banks, E. (2001). Epidemiology of Breast Cancer. *The Lancet Oncology*, *2*(3), 133–140. https://doi.org/10.1016/s1470-2045(00)00254-0

Key, T. J., Verkasalo, P. K., & Banks, E. (2001). Epidemiology of Breast Cancer. *The Lancet Oncology*, *2*(3), 133–140. https://doi.org/10.1016/s1470-2045(00)00254-0

Menopause and breast cancer risk. (1972). *JNCI: Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/48.3.605

Takaku, M., Grimm, S. A., & Wade, P. A. (2015). GATA3 in breast cancer: Tumor suppressor or oncogene? *Gene Expression*, *16*(4), 163–168. https://doi.org/10.3727/105221615x14399878166113

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