The gene mutation and expression differences in different age categories of patients

Kaining Feng

Introduction

The Cancer Genome Atlas (TCGA) is a publicly available database containing comprehensive genomic and clinical data on over 30 types of cancer. One of the TCGA database's most significant features is its multi-omic data inclusion. This allows the dataset to contain information on multiple levels of biological organization, including DNA sequencing, gene expression profiling, and clinical data. By analyzing these various types of data together, researchers can better understand the molecular processes underlying cancer development and progression.

Multi-omic data analysis has become a powerful tool in cancer research, enabling researchers to identify new cancer biomarkers, develop more accurate prognostic models, and design more effective personalized therapies. With the vast amount of data available in the TCGA database, researchers are able to conduct large-scale studies and ultimately improve patient outcomes.

Breast cancer (BC), the most common cancer in women, with more than 2 million new cases in 2020, is also recorded in the TCGA database. BC incidence and mortality have increased over the past three decades due to changes in risk factor profiles, better cancer registries, and cancer detection. Currently, about 80 percent of people with BC are over the age of 50 (Łukasiewicz,2021). This distinct trend of old patients generates questions about which gene significantly differs in mutation frequency in young and old breast cancer patients. How does mutation of this gene lead to the death of patients? Is it the most upregulated gene?

Based on the clinical data, mutation data frame, and gene information recorded in the TCGA database, the gene expression and mutation frequencies in old patients and young patients could be analyzed and shown in volcano plots and Co-oncoplots, respectively. Besides, the vitality of the gene mutation can be shown

in the Kaplan-Meier plots. After the analysis, the gene GATA3 was mutated most distinctively in old and young patients, gene RNU1 was most upregulated in the old patients, and gene fgf4 was mostly upregulated in the young patient.

Methods

During the analysis, the patients were first separated into two groups based on their ages recorded in the clinical data frame. The patients above the median age (58 years old) are considered “old patients,” and those below 58 are considered “young patients.” Then, the gene mutation data frame was used to find out the top 10 frequently mutated genes respectively in the two groups of people. Two Oncoplots and one Co-Oncoplot were drawn to visualize the discrepancies between the gene mutations. Based on these Oncoplots, I found a gene called GATA3, which has higher mutation frequencies in young patients than in old patients. The influence of GATA3 mutation on survival rate was also studied by drawing a Kaplan-Meier plot. By using the RNA sequencing data frame, I can obtain the barcode and gene names of different genes in patients and also get the counting of the expression of these genes, with gender and ajcc pathologic stage as covariables and age category as the main variable. I first compare the GATA3 gene expression in the young and old patient groups and visualize the differences using two box plots. At last, the most up-regulated gene in old and young patients was also found by ordering the genes based on their log2 fold change. A volcano plot visualizes this. In all the analyses and plots, the significance was set to $p<0.05$.

Results

The top 10 mutated genes in old patients are PIK3CA, TP53, TTN, CDH1, KMT2C, MUC16, MAP3K1, GATA3, FLG, and PTEN (shown in Fig.1). The top 10 genes in young patients are TP53, PIK3CA, GATA3, TTN, CDH1, MUC16, KMT2C, RYR2, HMCN1, and MAP3K1 (Fig.2). I noticed that the GATA3 gene was the No.3 most mutated gene (16%) in young patients, while it was the No.7 (10%) in old patients. This could be more obviously observed in Co-Oncoplots in Fig.3.

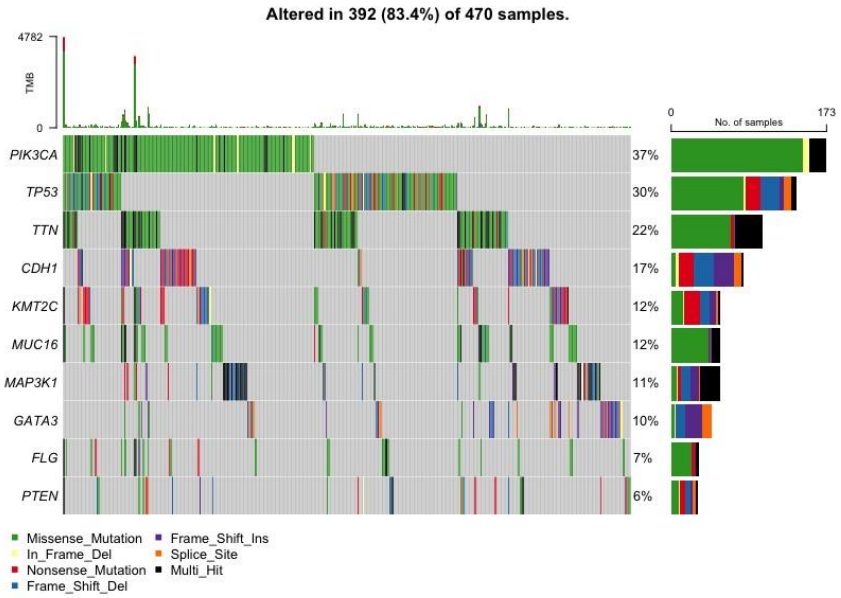

Fig.1 Oncoplot shows the top 10 mutated genes in old patients are PIK3CA, TP53, TTN, CDH1, KMT2C, MUC16, MAP3K1, GATA3, FLG, and PTEN.

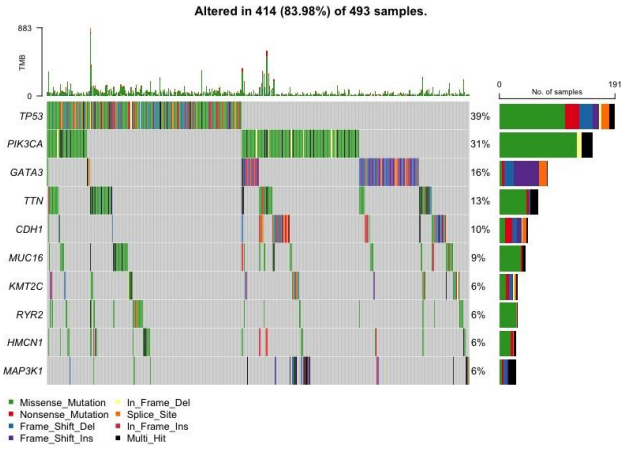

Fig.2 Oncoplot shows the top 10 genes in young patients are TP53, PIK3CA, GATA3, TTN, CDH1, MUC16, KMT2C, RYR2, HMCN1, and MAP3K1

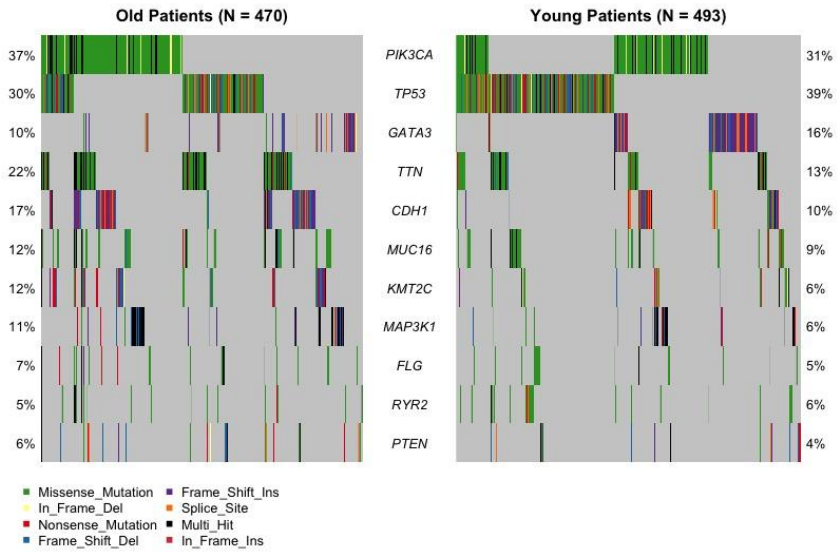

Fig.3 the CO-Oncoplot shows the differences in gene mutation in old and young patients.

The Kaplan-Meier plot (Fig.4) shows the survival probability of patients with mutated GATA3 gene and wild-type GATA3 gene. From the plot, it's obvious that the mutant curve is constantly lower than the wild-type curve, meaning it has a lower survival rate would be resulted from the mutation of the GATA3 gene. However, the p-value was 0.071, which is higher than the significant p-value, which means it cannot support the relationship between GATA3 mutation and patients' survival rates.

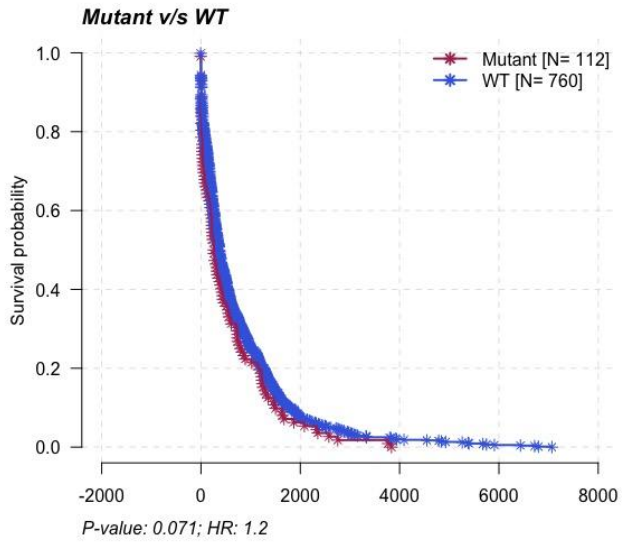

Fig.4 the survival probability of patients with mutated GATA3 gene and wild-type GATA3 gene (p=0.071)

Comparing the GATA3 gene expression in the young and old patient groups based on the RNA sequencing data frame and visualizing the differences using two box plots(Fig.5), I find this gene has a very similar expression in the young and old patient groups, though with different mutation frequencies. Thus, I should find the most expressed genes to find the upregulation genes in these two groups.

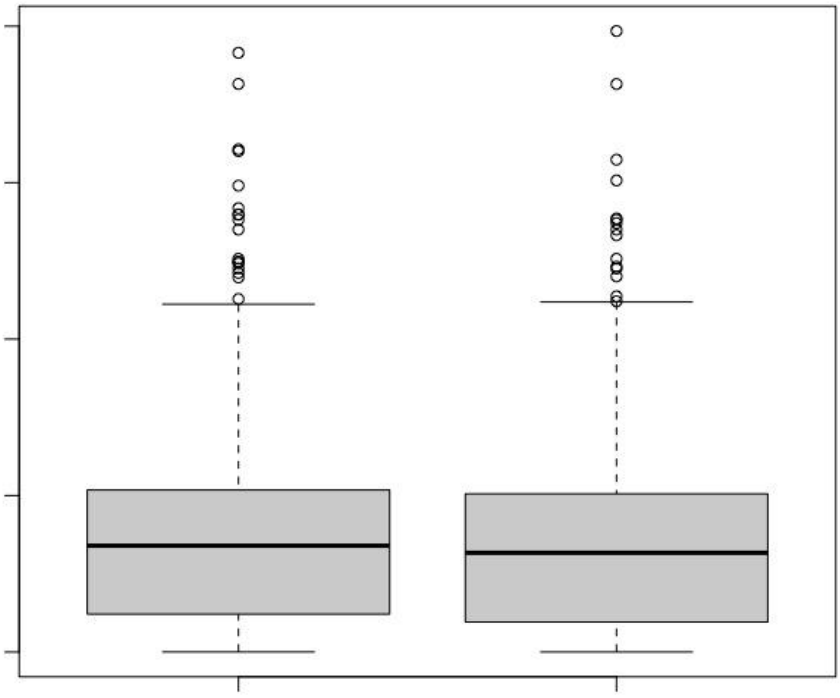

Fig. 5 Boxplots showing the GATA3 gene expression in the young and old patient groups based on the RNA sequencing data frame.

After setting the young patients as the control group, the genes are ordered based on their log2 fold change. The gene with the largest log2 fold change is the most upregulated in old patients, and with the least log2 fold change is the most upregulated gene in young patients. I found the RNU1 genes are the most regulated in old patients and fgf4 to be the most regulated in young patients (shown in Fig.6)

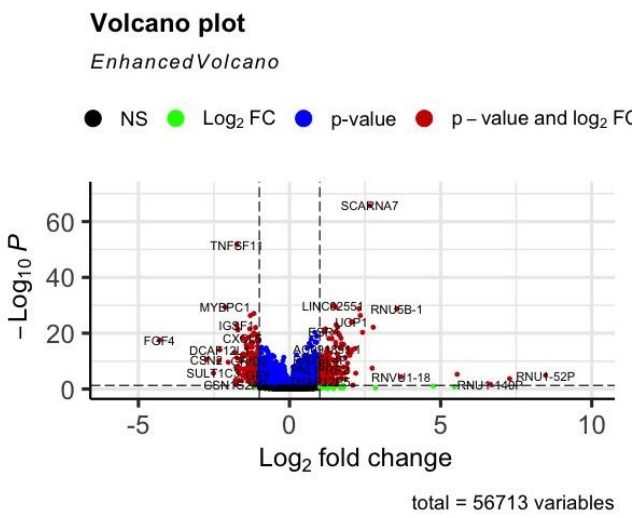

Fig.6 The volcano plot shows the expression frequency of the genes in old patients.

Discussion

Based on my analysis, the GATA3 gene tends to impact patients' survival rates negatively but still needs more research to ensure this relationship. According to Takaku's group (2015) finding, the GATA3 gene is mutated in a high frequency of breast tumors, approximately 10%, with both wild-type and mutant products expressed. Their model shows that GATA3 is a prominent marker of the luminal pattern of gene expression, and loss of GATA3 expression is associated with tumor types with a propensity for invasive growth and poor prognosis. This supports my analysis of the relationship between survival rate and GATA3 mutation. However, their article didn't talk about the didn't mutation rate between old and young patients. It may need more data and further research to confirm.

According to my analysis, the RNU1 genes were the most regulated in old patients, but I didn't find much valuable research on these genes. However, I find research on the CARTPT gene, which is the second upregulated gene in old patients. Kebin's team focused on the Mucinous carcinoma (MC) of the breast, a special histological type of breast cancer. MC patients were older, had lower tumor grade and T and N stage, higher hormone receptor-positive proportions, and were featured by the upregulation of MUC2,

CARTPT, and TFF1. This finding strongly supports my analysis that the CARTPT gene is upregulated in old patients.

References:

Lu, K., Wang, X., Zhang, W., Ye, H., Lao, L., Zhou, X., Yao, S., & Lv, F. (2020). Clinicopathological and genomic features of breast mucinous carcinoma. *The Breast*, *53*, 130–137. https://doi.org/10.1016/j.breast.2020.07.010

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