The Effect of Age on the CDH1 Gene Mutation and Survival Rate of Breast Cancer

I. Introduction

Breast cancer is caused by the uncontrolled growth of cells in the breast. The cancer can occur in lobules (the glands that produce milk), ducts (tubes that carry milk to the nipples) and surrounded connective tissues. The most common types of breast cancer are invasive ductal carcinoma and invasive lobular carcinoma. The risk factors include getting older, gene mutations, family history of breast cancer, and reproductive history. According to Breastcancer.org, about 1 in 8 U.S. women (about 13%) will develop invasive breast cancer over the course of her lifetime ("Breast Cancer Facts and Statistics", 2022). It is the second highest cancer among women, just after lung cancer. Age, as a risk factor of breast cancer, was studied in this research to figure out its effect on CDH1 gene mutation and survival rate. CDH1 gene codes for E-cadherin protein, which is a strong invasion suppressor in tumor cell systems, whose mutation would result in lobular breast cancer. It was shown that family history of breast cancer can lead to increased risk of CDH1 mutation (Corso et al., 2020). This research was based on TCGA database, the Cancer Genome Atlas program that focuses on genomic, transcriptomic, proteomic and clinical data. It provides vast amount of data about patient information, gene type and mutation type. Specifically, the clinical data and mutation data were imported into R to construct oncoplots and lollipop plot for the comparison between the effect of young and old age groups on CDH1 gene mutation, the Kaplan-Meier plot for the survival analysis of the mutant and wild type gene, as well as the boxplot to see the relationship between age and days of radiation therapy. The result showed significant increase in gene mutation among old

population and decrease in survival probability. However, there is no strong correlation between days of radiation and age.

II. Method

The breast cancer clinical data and the mutation data (MAF) were accessed from TCGA with the accession code “TCGA-BRCA”. Then, various masks were constructed to filter out the NA values and categorize continuous variables into groups. Patients with age greater than or equal to 50 years old at the time of diagnose were considered old, while patients less than 50 years old were young. The MAF data of young and old patients were then subsetted out to construct a co-oncoplot using the R package “maftools”, which visualized the mutational landscape of various mutant genes among different age groups. A co-lollipop plot were also generated by maftools to compare the number and position of mutated CDH1 gene between young and old population. After that, R packages “survival” and “survminer” were loaded to construct the MAF survival plot (Kaplan-Meier plot) for survival analysis of patients with wild type and mutant CDH1 gene, and a boxplot in basic plot function was used to compare the days of radiation therapy accepted by young and old patients.

III. Result

The co-lollipop plot provided us an overview of the mutated gene among different age groups (Figure 1). Figure 1 showed the top 6 mutated genes. The total number of mutation was young population is 263, while that in old population was 700, indicating that old population had an overall higher mutation rate in genes. The CDH1 gene was the fourth commonly mutated

gene, accounting for 8% of mutants in young and 15% in old population, which was twice as much as young's, showing that both the number of mutation and the percentage of CDH1 gene mutation are higher in the old population.In addition, unlike other genes, such as PIK3CA and TP53, lots of CDH1 mutations were not missense mutation, but nonsense mutation and frame shift deletion (shown as red and blue bars in Figure 1).

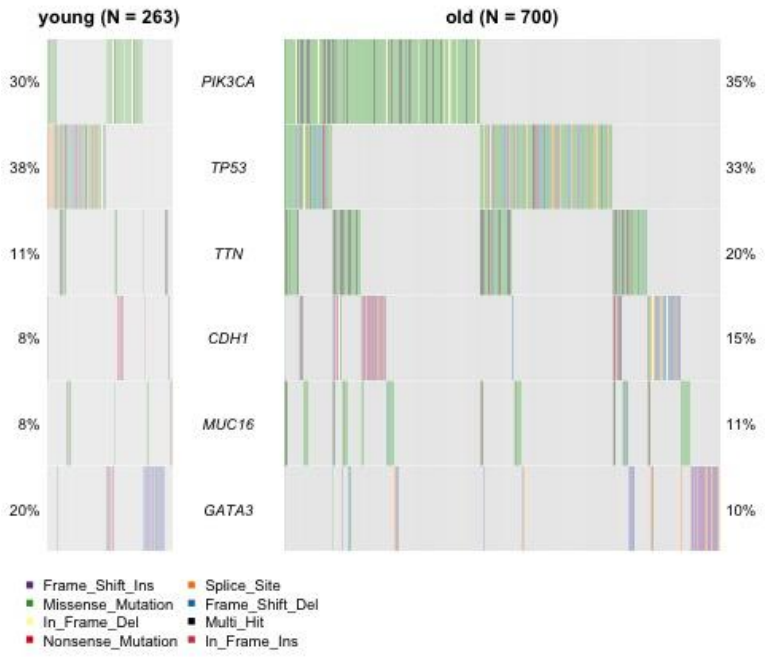

Figure 1. The co-oncoplot plotted using R package "maftools" with the top 6 mutated genes among two age categories (young population <50 y/o and old population >=50 y/o), showing that the old population have a higher mutation rate in genes, including CDH1 gene.

Then, a co-lolliplot was constructed to visualize the number of CDH1 mutations along the position of gene in each domain. The CDH1 gene contains a Cadherin_pro region, a CA_like region, a Cadherin_repeat region, a Cadherin region and a Cadherin_C region. The mutations were evenly scattered across every region with 1 at each position and not necessarily within the

functional region. There was a noticeable mutation before the Cadherin_pro region, which had 8 nonsense mutations at the same position among the old population. The nonsense mutation terminates the translation process before any useful protein can be made, which may be the reason why no tumor suppressor protein is made and cancer occurs.

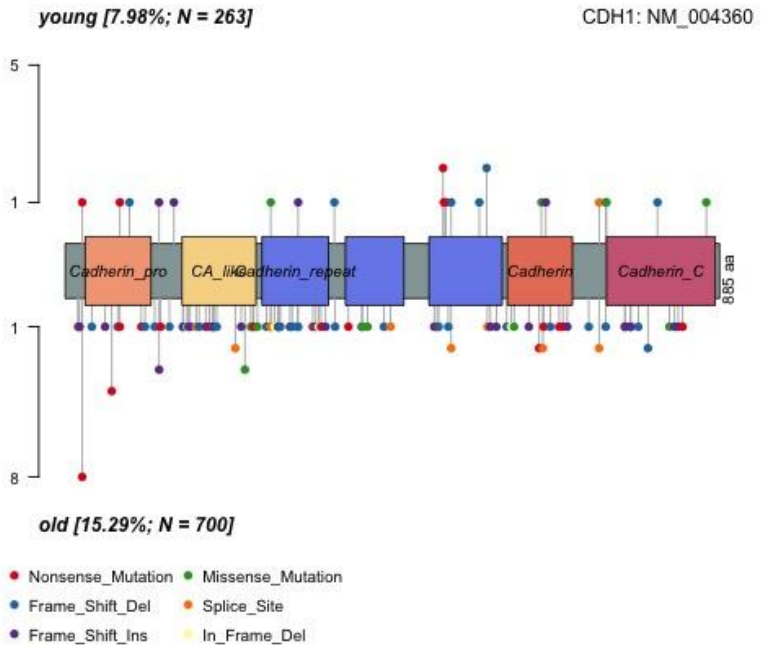

Figure 2. The co-lollipop plot by R package "maftools" with lollipopPlot2 function, showing that the old population had a higher mutation rate in CDH1 gene, especially the nonsense mutation before the Cadherin_pro region (8 mutations) compared with that in the young population (only 1 mutation).

Both the co-oncoplot and the co-lollipop plot presented a larger number of mutations of CDH1 gene among the old population. The next question is whether a higher mutation rate would lead to a lower survival rate when diagnosed of breast cancer. A maf survival plot of the old population and young population were plotted separately to see the survival probability of

patients with wild type and mutant CDH1 gene. The number of mutant gene was much lower than that of the wild type and the graph showed that the survival probability of patients with normal CDH1 genes decreased more rapidly than those with mutant CDH1 gene, to around 0.4. The same result was found in the maf survival plot of the young population, with the survival probability of patients with mutant CDH1 gene remained at 1, while that of the wild type was as low as 0.4. The patients with no CDH1 mutation have a lower survival rate because they would have other mutations that may be more lethal. Comparing solely the patients with mutant CDH1 gene, the survival probability of the old population falls more drastically than that of the young population, meaning that the old population is more susceptible to the CDH1 gene mutation.

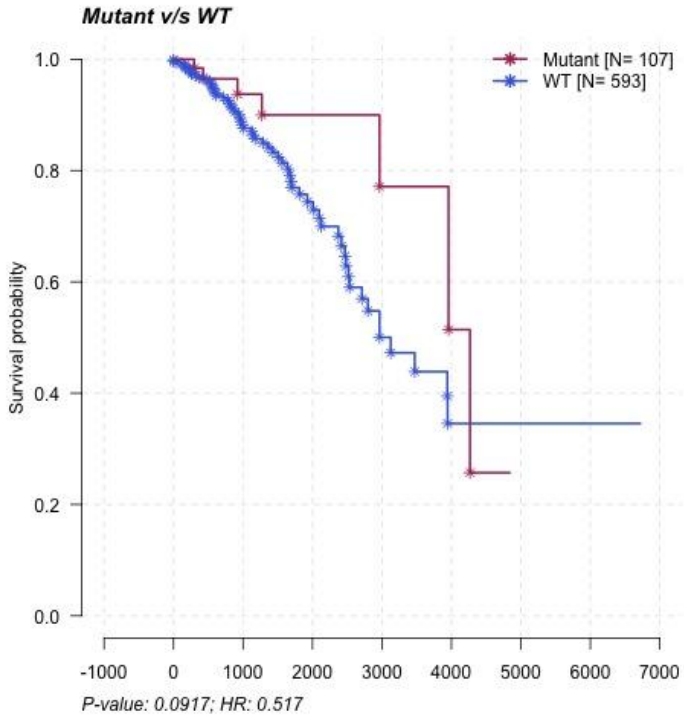

Figure 3. The MAF survival plot by R package “maftools” with mafSurvival function, visualizing the survival rate by mutational status of the CDH1 gene among old patients. The dataset of the old population was subsetted out from the whole maf data using the mask.

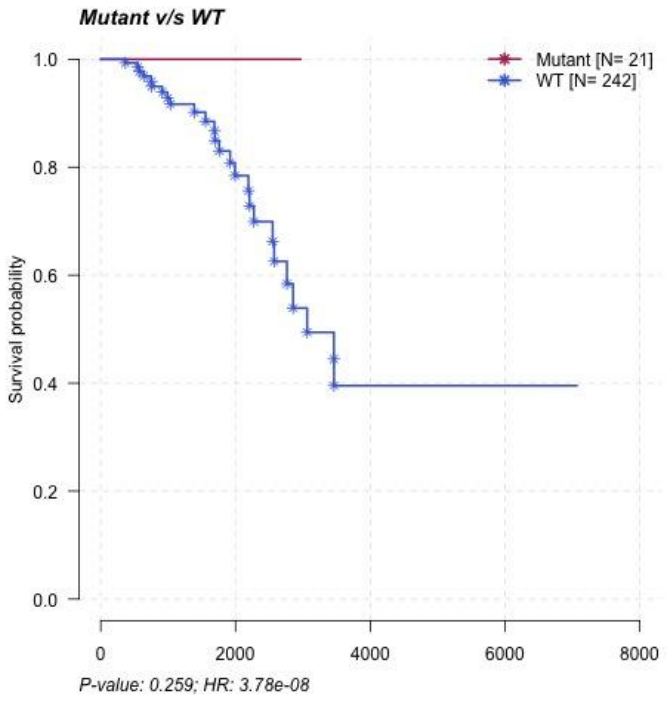

Figure 4. The MAF survival plot by R package “maftools” with mafSurvival function, visualizing the survival rate by mutational status of the CDH1 gene among young patients. The dataset was subsetted out from the whole maf data by inverting the mask set for the old population.

Finally, we looked at the relationship between age and days of radiation therapy. The boxplots of the days of radiation were plotted separately for the young and old population (Figure 5). The mean days of radiation therapy taken by the young population were longer than

the old population. The summary function was then called to quantified the boxplot. The mean and median values were low among the old population, meaning that they accepted shorter days of radiation on average. However, there was no significant difference between the first and third quantile, indicating that the variation was not significant and the days of radiation therapy does not strongly relate to age categories.

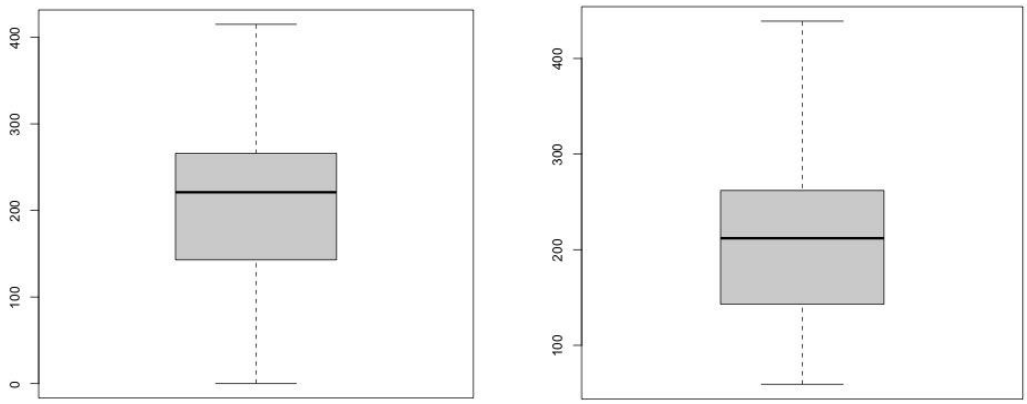

Figure 5. The boxplot of radiation days constructed by the boxplot function in the basic functions. The days of radiation was derived from substracting the days when radiation ends from the days when radiation starts. The mean of young population was higher than that of old population, but the 1st and 3rd quantile remained at a similar level.

Table 1. The statistic data of the days of radiation for young and old population by using summary function in R.

Age group	**Min**	**1st quantile**	**Median**	**Mean**	**3rd quantile**	**Max**
Young	0.0	141.0	218.0	214.7	267.0	472.0

Old	59.0	143.0	212.0	207.9	262.0	439.0

IV. Discussion

In conclusion, old patients generally have more CDH1 gene mutations than young patients, both in terms of numbers and percentages. Nonsense mutaiton and frameshift deletion are the common types of mutation for CDH1 gene. The position before the Cadherin_pro region on CDH1 is a common nonsense mutation point for old patients. Although CDH1 gene mutation is the fourth highest mutation among all the breast cancer gene muation, it is not as lethal as other gene mutations, since the result showed that patients with no CDH1 gene mutation (i.e. with other types of mutations) have lower rate of survival than patients with mutant CDH1 gene. Still, the old patients have a much lower survival probability than the young patients with CDH1 mutant genes. In terms of days of radiation therapy accepted, there is no significant difference between age categories, with an average of days between 200 and 250 days.

The research done on the “Incidence of Gastric and Breast Cancer in CDH1 Mutation Carriers From Hereditary Diffuse Gastric Cancer Families” showed that the risk of women getting breast cancer at age 40 is approximately 3%, while that at age 80 is 39% among the CDH1 mutation carriers, verifying the conclusion that there are more old patients carring CDH1 gene mutations than the young patients (Pharoah, Guilford & Caldas, 2001). Another research on genetic alterations of APC and CDH1 genes in lobular breast cancer also looked at the genetic variations in breast cancer patients, showing that most of the mutations are in-frame deletion

that lead to a stop codon, which is nonsense mutation, corresponding to the observation of the oncoplot constructed in our research (Sarrió et al., 2003). The nonsense mutation before the Cadherin_pro region prevents the coding gene to be translated, leading to the lack of tumor suppressor protein and a higher chance of getting cancer. However, the sample is still too small for us to have a general pattern of common mutation point on the CDH1 gene. Especially for young patients, the number of mutation at each position was only one, make it hard to analyze.

For the survival analysis, research done by Beretta et. al stated that patients with hereditary syndrome carrying BRCA1/2 CDH1 germline mutations have a worse breast cancer-specific survival compared with BRCA-negative/sporadic cases (Corso, Veronesi, Sacchini & Galimberti, 2018). This is inconsistent with the result by the maf survival analysis, which showed that the survival probability of patients with CDH1 mutant gene is higher than that of wild type. However, we could not determine whether the low survival probability of the wild type is due to the normal CDH1 gene or due to other gene mutations, which limited our ability to preform the single CDH1 gene survival analysis. Another research on the effect of age on breast cancer survival rate showed that the relative survival rate is the highest among the 46-50 years old age group, around 70%, while that of young patients (30 years old) and old patients (75 years old) are around 60% (Holli & Isola, 1997). This is also inconsistent with our survival plot, which presented a lower survival rate among the old population. This may be due to the threshold of age category, which is 50 years old, which only divide the population into two groups and was not able to characterize the specific trend of survival rate.

In future research, there are mainly two ways to improve and dig deeper into the topic. Firstly, more sample needs to be gathered for the analysis of survival and mutation, so that we

can find the common mutation position on the gene and the type of mutation among different age categories. More samples also allows us to have a more solid understanding of the age on the survival rate. In current analysis, the number of young patients with CHD1 mutant gene was only 21, which is far from enough to obtain a general trend. Moreover, we will divide the age into more than two categories, such as 10 years as a group, to explore different patterns among these groups, since women in the middle age go through a series of hormonal changes that affect the development of breast cancer.

V. Reference

Breast Cancer Facts and Statistics. (2022). Retrieved 14 October 2022, from https://www.breastcancer.org/facts-statistics

Corso, G., Montagna, G., Figueiredo, J., La Vecchia, C., Fumagalli Romario, U., & Fernandes, M. et al. (2020). Hereditary Gastric and Breast Cancer Syndromes Related to CDH1 Germline Mutation: A Multidisciplinary Clinical Review. Cancers, 12(6), 1598. doi: 10.3390/cancers12061598

Corso, G., Veronesi, P., Sacchini, V., & Galimberti, V. (2018). Prognosis and outcome in CDH1-mutant lobular breast cancer. European Journal Of Cancer Prevention, 27(3), 237-238. doi: 10.1097/cej.0000000000000405

Holli, K., & Isola, J. (1997). Effect of age on the survival of breast cancer patients. European Journal Of Cancer, 33(3), 425-428. doi: 10.1016/s0959-8049(97)89017-x

Pharoah, P., Guilford, P., & Caldas, C. (2001). Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology, 121(6), 1348-1353. doi: 10.1053/gast.2001.29611

Sarrió, D., Moreno-Bueno, G., Hardisson, D., Sánchez-Estévez, C., Guo, M., & Herman, J. et al. (2003). Epigenetic and genetic alterations ofAPCandCDH1genes in lobular breast cancer: Relationships with abnormal E-cadherin and catenin expression and microsatellite instability. International Journal Of Cancer, 106(2), 208-215. doi: 10.1002/ijc.11197

VI. Review questions

1. TCGA is a public dataset including genomic, transcriptomic, proteomic and clinical data of various cancer types. It is important because it helps us analyze the similarity and alterations of tumor samples with large amount of data.
2. TCGA has a clear classification of the statistics of different cancer subtypes, and provide us with data from different -omics aspects. But TCGA lacks the normal sequencing data that can be compared with the tumor data. We need to search for other datasets such as GTEx for more data.
3. Git add; git commit -m; git push
4. install.packages()
5. BiocManager::install()
6. Boolean indexing returns true and false at a specific index with respect to the condition. By applying the boolean mask, we can get the data we want (with boolean index TRUE) and ignore the data we do not want (with boolean index FALSE). We can use it for filtration of NA, subsetting out a data frame and categorizing continuous variables.
7.

	Age	Race	Gender
Patient 1	77	Hispanic	Female
Patient 2	39	African American	Female
Patient 3	56	Asian American	Male

- age_mask <- ifelse (df$Age > 50, T, F): if the age is greater than 50, return TRUE at that index, otherwise return FALSE

- df <- df[age_mask,]: subset out the data frame with patient age greater than 50 (keep all columns and ignore the rows with patient age less than 50).