Sabrina Zhong

QBIO490

Mar. 4, 2022

What is the relationship between the gene expression of MSH2 and MLH1, and how does it differ between men and women with colorectal cancer? Does the under-expression of MSH2 and MLH1 affect survival?

Introduction:

Colorectal cancer (CRC) is characterized by an uncontrollable cell growth in the colon or rectum and is one of the most diagnosed cancers making up for around 10% of all cancer diagnosis annually. Specifically, it is the second and third most common diagnosed cancer for women and men respectively. It is not only a commonly diagnosed cancer, but also the fourth deadliest cancer in the world accounting for about 900,000 deaths annually (Dekker et al., 2019).

Abnormal expression levels of the MSH2 and MLH1 genes have been found to have a correlation with the development of CRC. MSH2 and MLH1 are classified as mismatch repair genes (MMR), which are responsible for identifying and repairing mismatched pairs during DNA replication. An under-expression of these genes leads to an increased mismatching of base pairs, causing mutations that can lead to CRC or other cancers (Salem et al., 2020). Moreover, lynch syndrome is a hereditary condition that is characterized by mutations in MMR genes including MSH2 and MLH1. Research has shown that those with Lynch syndrome have an increased risk of developing CRC suggesting that mutations in MSH2 and MLH1 can be important biomarkers for CRC (Duraturo et al., 2019).

Because this cancer is asymptomatic during its early stages, it is imperative to have more accurate ways of detecting risk factors and the early onset of this cancer. Through the analysis of

gene expression amongst colorectal cancer patients, researchers can identify mutations in genes such as, MSH2 and MLH1, that have a direct correlation to the early onset of colorectal cancer. Using the findings from these research efforts, better diagnosis and treatment for this deadly cancer can be developed and implemented.

This research project aims to analyze the count relationship between MSH2 and MLH1 genes amongst men and women and how the under-expression of MSH2 and MLH1 affects a patient's survival rate. This project utilizes data from the Cancer Genome Atlas and R functions to analyze the patient's clinical data. Through these tools, I found that CRC patients tend to have lower counts of both MSH2 and MLH1 genes, but this trend may not be related to their survival rate.

Methods:

The data was from the TCGA database, and the R package TCGAbiolinks was installed to get the clinical and RNAseq data of patients. I used the accession code "COAD" to get data of patients with colon cancer. Of the colorectal cancer patients, 244 of them were female and 280 of them were male. To create the Kaplain-Meier plots, I installed and used the survival and survminer R packages. I used the Kaplain-Meier plots to compare patients with high or low counts of the MSH2 and MLH1 gene. The median of the gene counts for each gene was used to determine if the patient's counts of MSH2 was high or low as the outlier counts affected the median less.

Results:

By comparing the counts of the MSH2 and MLH1 genes, I found that there was a positive correlation between the two counts in the patient sample. Patients with low counts of the MSH2 gene also had around the same counts of the MLH1 gene. The counts seem to be clustered

near the origin of the scatterplot, so most patients had an underexpression of the two genes (Fig. 1). When comparing the counts of each gene between men and women, there was no association between gender and the counts of the gene as both men and women had the same range and median count for each gene. The median count of MSH2 for all patients was 1348 and 948 for MLH1 (Fig. 1 & 2). For the MLH1 gene count, the women patients do have a higher density in the second quartile, but the difference seems negligible (Fig. 2b). For both men and women, the boxplot is more condensed below the median than above. Similar to what was shown in figure 1, the boxplot shows that more patients have an underexpression of MSH2 and MLH1. There are a few patients who have extremely high counts of MSH2 or MLH1 (Fig. 2).

The Kaplain-Meier graphs show the survival probability over time for patients with low and high counts of genes MSH2 and MLH1. For MSH2, there is a different between survivability of patients with low and high counts of the gene. The patients with low counts of MSH2 had a lower survival probability over time. The patients with low counts of MSH2 had a survival probability of about 0.15 after 9 years whereas the patients with high counts had a survival probability of around 0.4 (Fig. 3). For the MLH1, gene the difference between the survival probability of the low counts and high counts groups is very little even though those with low counts have a slightly higher survival probability over time. The survival probability for both categories of counts was about 0.35 after 9 years (Fig. 4). The p-value for the MSH2 counts plot is also lower than the one in the MLH1 counts plot.

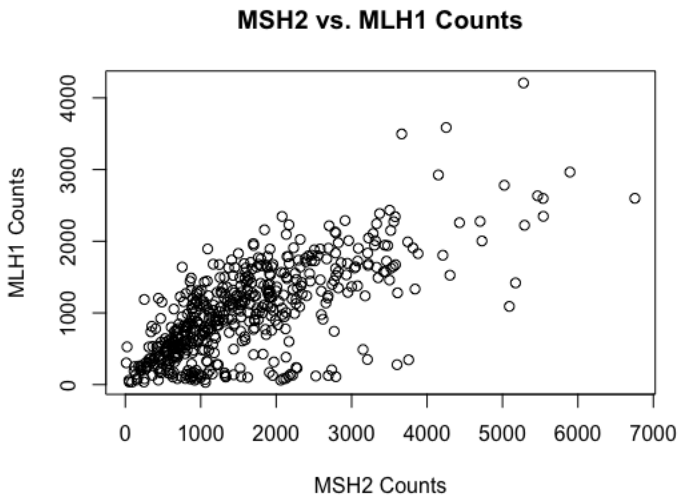

Figure 1: A scatterplot of MSH2 vs. MLH1 gene counts. There is a positive correlation between the two gene counts.

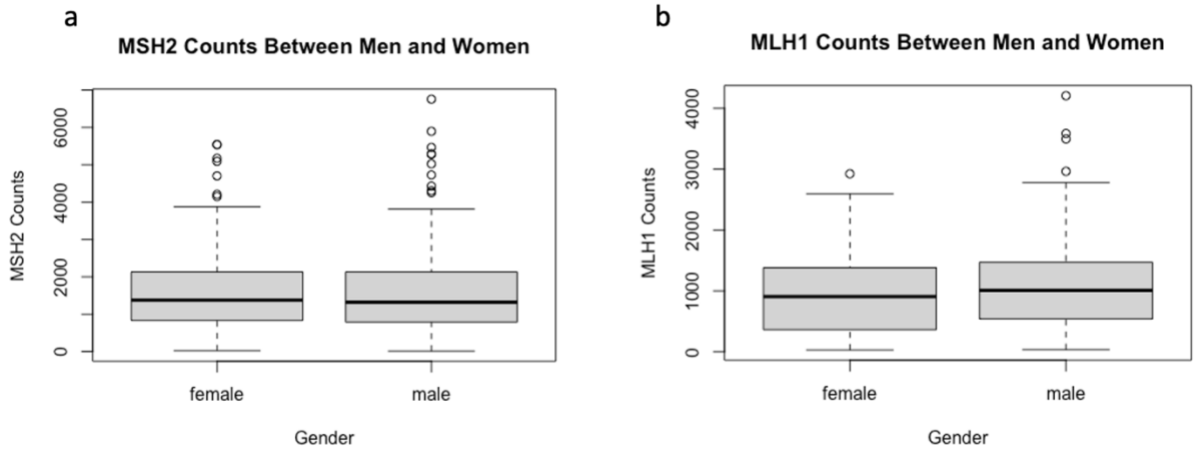

Figure 2: Boxplots of men and women and each subpopulation's counts of MSH2 and MLH1 gene.

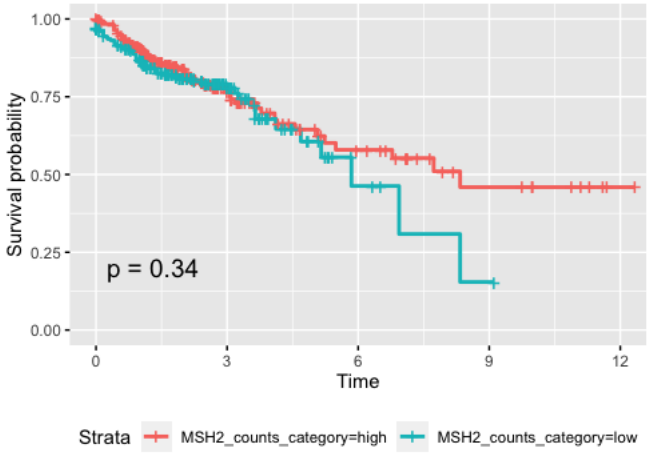

Figure 3. A Kaplain-Meier plot depicting the survival rates of patients with high or low counts of MSH2. The x axis is the time in years and the y axis is the survival probability.

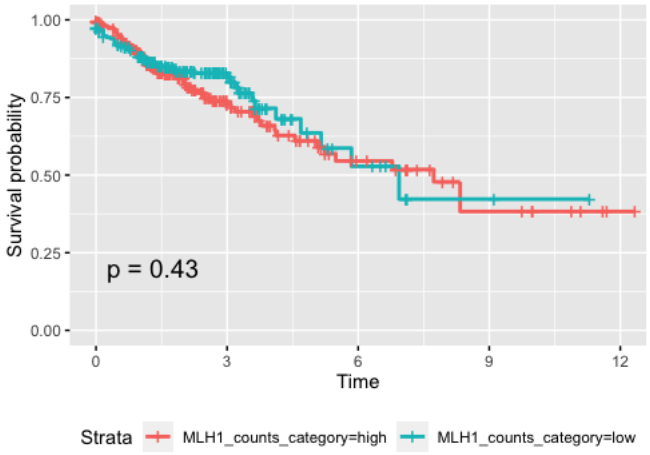

Figure 4. A Kaplain-Meier plot depicting the survival rates of patients with high or low counts of MLH1. The survival rates seem to be about the same.

Discussion:

The trend seen in figures 1 and 2 where most patients have lower counts of MSH2 and MLH1 can also be seen in previous research efforts. In Moufid F.Z., et al.'s research on MLH1 and MSH2 gene expression among Moroccan patients, they found that among the 214 patients, 24 of the tumors were identified as MMR deficient. This shows how a mutation in MSH2 and MLH1 genes leading to the underexpression of those genes can lead to an increased risk of developing CRC. As mismatch repair genes, MSH2 and MLH1 are responsible for recognizing and repairing mistakes during DNA replication. Without an ample amount of these genes, errors during DNA replication can lead to uncontrollable cell growth, a major characteristic of cancer. In addition to this, the research team also found that of the MMR deficient tumors, 66.7% of them had a complete loss of MLH1 and MSH2 gene expression and were associated with the patients who were considered younger at CRC diagnosis (Moufid et al., 2018).

In my analysis, patients with a lower count of the MSH2 gene had a slightly lower survival probability (Fig. 3). However, Wang S.H. et al.'s research concluded that mutations in MLH1 or MSH2 gene was independent of the patient's survival outcome. By studying 681 patients where 131 of them had a mutation in either MLH1 or MSH2 or both, those with at least one mutation had an 86.9% chance of survival after 5 years compared to 59.1% for patients without any mutations. This shows how mutations in these genes may not be related to or decrease the survival of the patient as those without mutations had a significant decrease in survivability. Wang's research concluded that other factors such as the patient's family history of CRC, anatomy of the intestine, age, and much more are more imperative to the patient's survival outcome. In addition to this, Wang's research also found that patients in stage III CRC with

deficient MMR genes had a much higher chance of survival with chemotherapy treatment than those with proficient MMR genes (Wang et al., 2019).

The results from this study provide insight on the relationship between the gene expression of MSH2 and MLH1 and the risk of developing CRC. This suggests that a deficiency in MSH2 and MLH1 could be a potential biomarker for CRC. Further research can be done to understand how mutations and which type of mutation leads to a deficiency of these MMR genes so that risk factors could be detected earlier. The other result that the gene expression of MSH2 and MLH1 may not be indicative of a patient's survival rate. This suggests a focus on other factors such as expression of other genes, family history of CRC, and/or the specific anatomy of the patient's colon should be made in future research. Identifying the factor or factors that significantly decreases a patient's survival rate would be beneficial to developing CRC treatment tailored towards it to improve survival outcomes.

References

Dekker, Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., & Wallace, M. B. (2019). Colorectal cancer. The Lancet (British Edition), 394(10207), 1467–1480.

Duraturo, F., Liccardo, R., De Rosa, M., & Izzo, P. (2019). Genetics, diagnosis and treatment of Lynch syndrome: Old lessons and current challenges. *Oncology letters*, *17*(3), 3048–3054.

Moufid, Bouguenouch, L., El Bouchikhi, I., Chbani, L., Iraqui Houssaini, M., Sekal, M., Belhassan, K., Bennani, B., & Ouldim, K. (2018). The First Molecular Screening of MLH1 and MSH2 Genes in Moroccan Colorectal Cancer Patients Shows a Relatively High Mutational Prevalence. *Genetic Testing and Molecular Biomarkers*, *22*(8), 492–497.

Salem, Bodor, J. N., Puccini, A., Xiu, J., Goldberg, R. M., Grothey, A., Korn, W. M., Shields, A. F., Worrilow, W. M., Kim, E. S., Lenz, H., Marshall, J. L., & Hall, M. J. (2020). Relationship between MLH1, PMS2, MSH2 and MSH6 gene-specific alterations and tumor mutational burden in 1057 microsatellite instability-high solid tumors. *International Journal of Cancer*, *147*(10), 2948–2956.

Wang, S. M., Jiang, B., Deng, Y., Huang, S. L., Fang, M. Z., & Wang, Y. (2019). Clinical significance of *MLH1*/*MSH2* for stage II/III sporadic colorectal cancer. *World journal of gastrointestinal oncology*, *11*(11), 1065–1080.

General Concepts

1. What is TCGA and why is it important?
   TCGA stands for "The Cancer Genome Atlas", and it is a database of genomic, epigenomic, transcriptomic, and proteomic data of cancer patients. It is important because it provides data that research groups can analyze to help better understand, diagnose, and treat specific cancers.
2. What are some strengths and weaknesses of TCGA?
   A strength of TCGA is that it is the largest database of genotypic information on cancer patients. With a large set of data, there can be more accurate findings for the specific cancer.
   A weakness of TCGA is that it takes time to download and analyze the data because there is so much data.
3. How does the central dogma of biology (DNA → RNA → protein) relate to the data we are exploring?
   Our gene expression is produced through the central dogma of biology, and our gene expression is a huge part of how certain cancers come about.

Coding Skills

1. What commands are used to save a file to your GitHub repository?
   Git add (filename) or git add .
   Git commit -m "informative message"
   Git push
2. What command must be run in order to use a package in R
   Install.packages("name of package")
   Library(name of package)
3. What is boolean indexing? What are some applications of it?
   Boolean indexing is the use of a Boolean mask to index into a column of a dataset. We use Boolean indexing to get a subset of the data that has the specific category we need. For example, if we want to get the females in a dataset, we can use Boolean indexing to pick out all of the females in the gender column and save this into a Boolean mask. We then use the Boolean mask as an index of the data frame to extract the data that makes the mask true or false.

4. Draw out a dataframe of your choice. Show an example of the following and explain what each line of code does.
   1. an `ifelse()` statement
      vector = c(1, 5, NA, 3, 9, NA, NA)
      vector = ifelse(is.na(vector), -1, vector)
      The first line is creating a vector of integers and NA. The ifelse statement is converting the values of vector that are NA into -1. The line is basically saying, if the value in vector is NA, change it to -1, if it is not, then keep the same value.
   2. boolean indexing
      vector = c(1, 5, NA, 3, 9, NA, NA)
      vector[is.na(vector)] = -1
      The first line is creating the vector of integers and NA values. The second line is converting the values at each index in vector that have NA into -1.