Question: Are there sex differences in EGFR expression and mutation, and do those differences lead to sex-differentiated survival in CRC?

Introduction

Colorectal cancer (CRC) is the third most common cancer in the world and the fourth most common cause for cancer-caused death with around 700,000 deaths per year (Marmol et al. 2017). In the United States, however, CRC is the second-leading cause of cancer-related death (Ward et al. 2008). However, only around 5-10% of cases of CRC are reported to be related to inherited cancers (Marmol et al. 2017). Therefore, it is very important to study the risk factors associated with colorectal cancer that can be compounded by CRC's molecular biology. Studies have shown that age is a primary risk factor associated with CRC, as those over the age of fifty are more likely to develop CRC (Levin et al. 2008). On the other hand, sex has been found to be a risk factor in liver cancer cases, especially in relation to the growth receptor EGFR; in particular, EGFR was overexpressed in males with little to no mechanisms of mitigation present (Wang et al. 2016). Similarly to liver cancer, overexpression EGFR was sex-differentiated in lung cancer patients where female patients had better regulation and mitigation of EGFR overexpression than male patients. Males faced a lower survival rate than females due to estrogen pathways helping to alleviate the effects of EGFR overexpression (Chang et al. 2015). EGFR is also a gene found to be highly associated in CRC cases, as many CRC patients were found to have overexpression of EGFR (Markowitz and Bertagnolli 2009). Furthermore, EGFR mutations in CRC have led to an absence of cell growth regulating proteins. These are mainly from missense mutations on the furin-like cysteine rich region or deletions on the protein tyrosine kinase on the EGFR gene (Saif and Chu 2010). However, despite the findings about nature of EGFR mutations in CRC and the relationship between sex and EGFR in other cancers, there have been no significant studies that detail the association between sex and EGFR mutations in CRC specifically.

This paper will analyze the relationship between EGFR overexpression and mutation rate and patient sex in CRC. Patient data, RNAseq data, and MAF data were downloaded and extracted from The Cancer Genome Atlas (TCGA) and analyzed through R. RNAseq analysis was first performed comparing gene counts of EGFR between male and female patients. Then, survival analyses were performed comparing male and female patient survival in the presence of overexpression of EGFR and without. Lastly, MAF analysis was performed comparing the incidence of EGFR mutations between male and female patients, as well as examining where such mutations were on the gene. The results of these analyses revealed that there was no significant difference in EGFR gene expression between male and female patients. There was also no statistical difference between male and female patients, both with relative high and relative low EGFR expression. The MAF analysis revealed that although less than 3% of both male and female patients had mutations on the EGFR gene itself, all mutation sites were missense mutations and mutually exclusive between sexes.

Methods

Colon cancer clinical and RNAseq data was accessed from TCGA using the R package TCGAbiolinks with the accession code “COAD.” From the gene counts data extracted from the clinical data, boxplots were generated comparing the gene counts of EGFR between female and male patients. The R packages survival and survminer were used to create Kaplan Meier plots with survival data of patients stratified by gender. In particular, a Kaplan Meier plot was generated comparing survival rates between all male and female patients of relatively low EGFR counts, and a second was generated comparing those with relatively high EGFR gene counts. The patients with relatively high EGFR gene counts were defined as patients with more than the observed mean of gene counts for EGFR across both genders. A p value of 0.05 was decided as the threshold for a result being statistically significant. Lastly, the package maftools was used to generate lollipop plots comparing the positions of EGFR mutations along the gene for male and female patients.

Results

There were no observed statistical differences between the expression of EGFR for male and female patients. The mean of both genders' EGFR gene counts was 2905; however, the range of gene counts for male patients was larger than female patients with one male patient having an EGFR gene count over 25000 (Figure 1). After the comparative analysis of gene counts between female and male patients, a survival analysis was conducted comparing the survival rates of patients with low relative EGFR counts and patients with high relative EGFR counts. Among those with low EGFR counts, male and female patients did not have a significant difference in survival probability for approximately the first one thousand days after diagnosis. However, after one thousand days, female patients' survival probability had a slightly faster rate of decrease than male patients (Figure 2). Similarly to the patients with low EGFR gene counts, both female and male patients with high EGFR gene counts did not see any significant difference in survival probability for the first one thousand days after diagnosis. Furthermore, in both cases, female patients saw a faster decrease in survival probability over time than male patients after one thousand days (Figure 3). Lastly, the mutation sites on EGFR were examined for male and female patients. Although all mutations on EGFR observed were missense mutations, none of the mutation sites were shared between genders. All mutations were unique, each appearing only once (Figure 4).

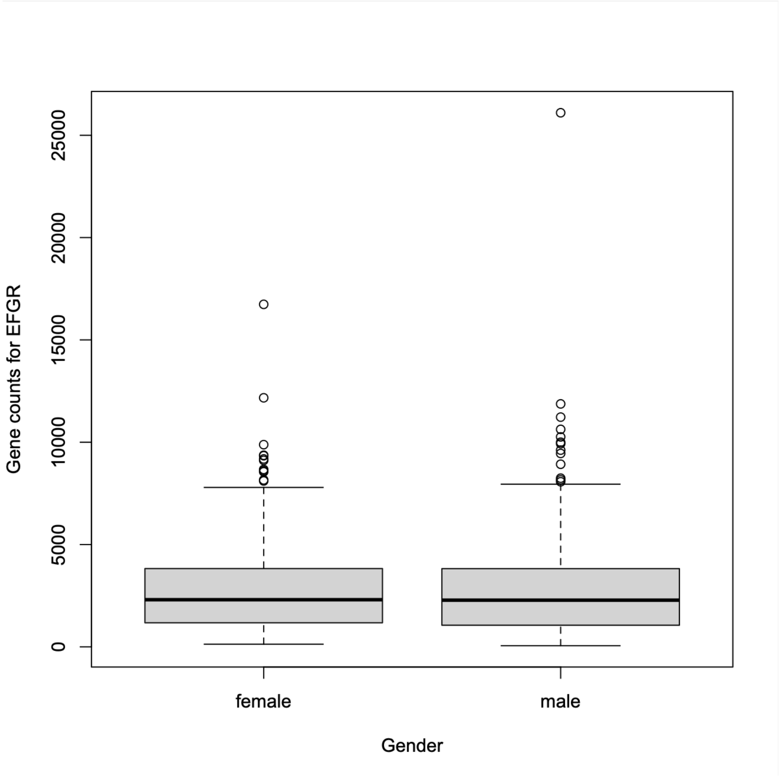

Figure 1: Boxplot comparing gene counts of EGFR between female and male patients. The two groups have no statistical difference in EGFR gene counts, having very similar quartile and median values. However, male patients have a larger range of EGFR counts with one count being over 25000, whereas the maximum EGFR count value for female patients is around 17000. There are also more outliers in the male patient EGFR gene count data compared to the female patient EGFR gene count data.

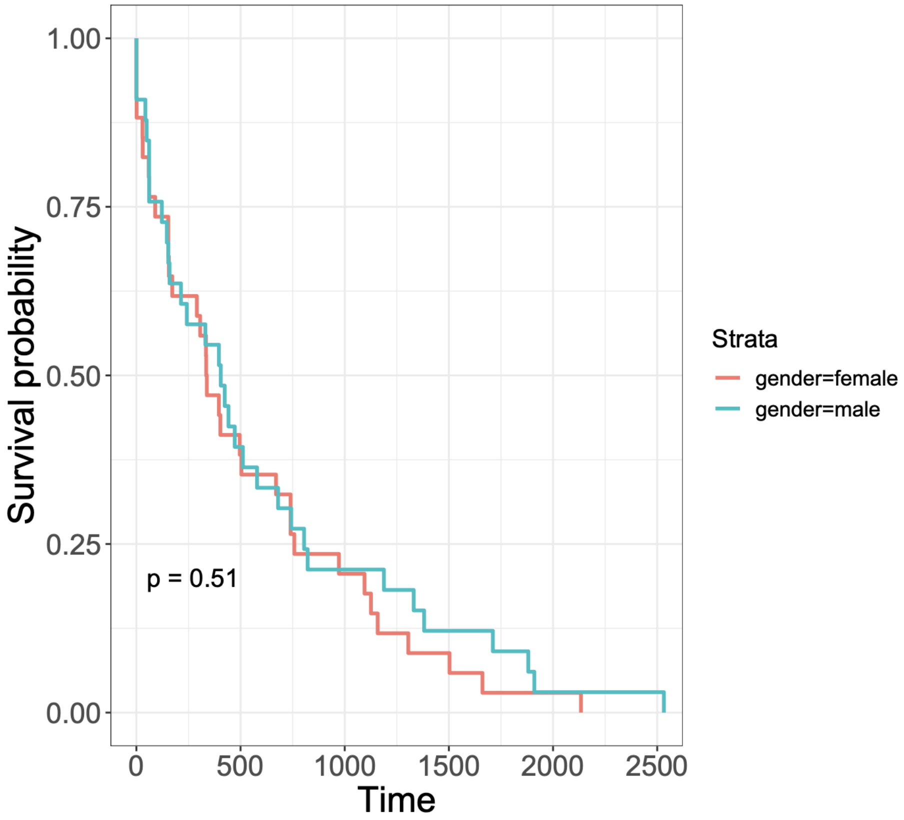

Figure 2: Kaplan-Meier plot of survival probability over time for male and female patients of relative low EGFR expression. For those with relatively low gene counts of EGFR, female and male patients had no significant differences in survival probability for the first one thousand days after diagnosis. After one thousand days, female patients' survival probability had a slightly faster rate of decrease over time than male patients with survival probability reaching zero at around 2100 days. In contrast, male patients' survival probability reached zero at a little over 2500 days. In this comparison, the p value was 0.51.

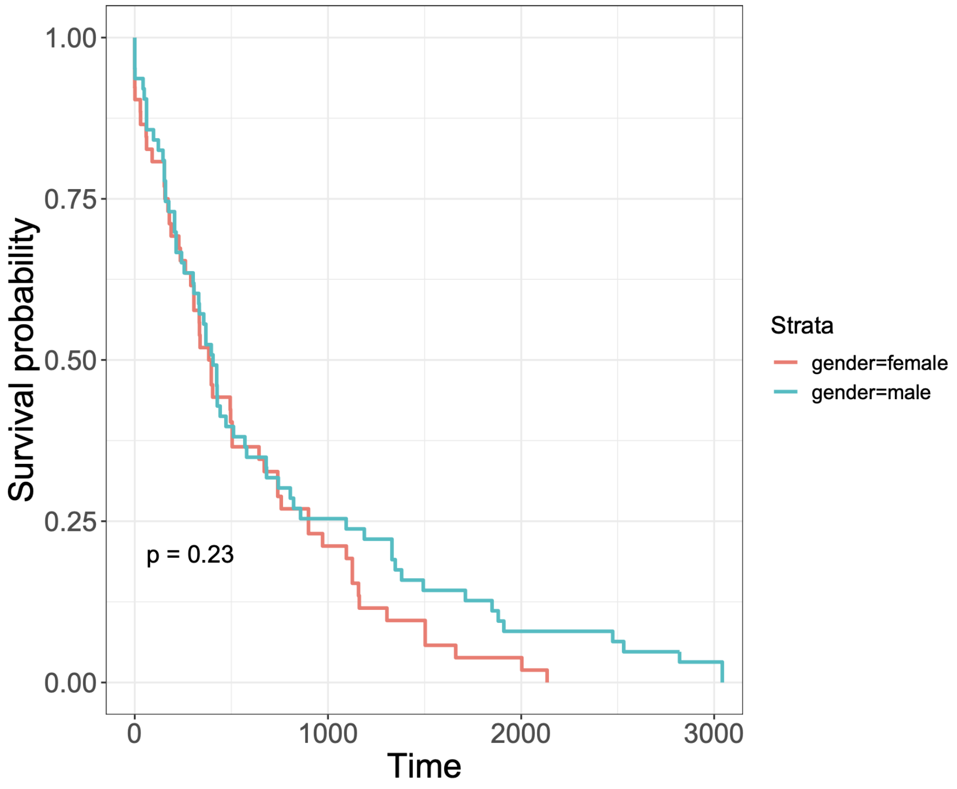

Figure 3: Kaplan-Meier plot of survival probability over time for male and female patients of relatively high EGFR expression. Similarly to the survival curves for patients with relatively low gene counts, the survival probability between female and male patients had no significant difference for approximately the first 1000 days after diagnosis. Furthermore, female patients after the first 1000 days saw a faster rate of decrease in survival probability compared to male patients like their counterparts with lower EGFR gene counts. The survival probability of female patients with high EGFR counts reached zero at around 2100 days like those with lower EGFR counts; male patients' survival probability reached zero later than those with lower counts at around 3000 days compared to 2500 days. The p value in this comparison was 0.23.

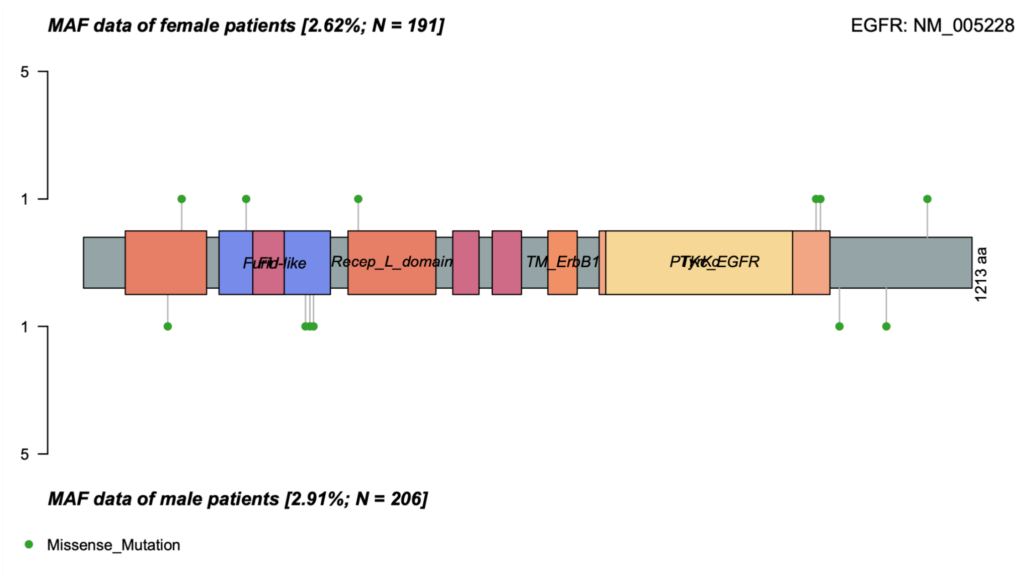

Figure 4: Lollipop plot comparing mutation location and mutation type on the EGFR gene between female and male patients. Though many patients had high expression of EGFR, less than 3% of male and female patients had mutations on the gene itself. Of the patients with mutations on the gene, all mutations present were missense mutations. Despite the homogeneity of mutation type across patients, all mutation sites were mutually exclusive between genders. However, there were no sites that appeared more frequently than others, with each mutation site only being observed once along the EGFR gene.

Discussion

There were no significant differences in the EGFR gene counts between female and male patients from this dataset. This result complicates findings by both Wang et al. and Chang et al., which reported higher frequency of EGFR overexpression in males for both lung and liver cancers. The lack of sex-differentiated significant difference in EGFR gene counts in CRC suggests that there is no association between EGFR gene counts and sex in CRC, and sex-related molecular pathways are not the underlying cause of high gene counts in CRC. Given the departure from studies of EGFR in other cancers, this further emphasizes that genes will exhibit different behaviors depending on the cancer.

For those with relatively low gene counts of EGFR, there is no statistically significant difference between survival rates given the high p-value of 0.51. Though with a smaller p-value of 0.23, those with relatively high gene counts of EGFR displayed similar patterns in survival probability. Despite the smaller p-value in the comparison of patients with high EGFR counts, the p-value is still not sufficiently small enough to deduce that there is a statistical difference between male and female survival rates. This is in opposition to the findings of Chang et al. where female estrogen pathways had a shielding effect on high EGFR gene counts, leading to better mitigation of unchecked cell growth. The similarities between survival rates of both males and females in both cases of EGFR gene counts suggest that the estrogen pathway that could protect female patients from unchecked cell growth in lung cancer did not offer female patients the same protection in CRC. This implies that any differences in survival rates of CRC are most likely not determined by patient sex. It is important to note that the definition of low EGFR gene counts in this analysis is in relation to only the gene counts observed in the sample; there was only one patient with gene counts lower than 100 and none with gene counts under 50. Further investigation may include more samples that have lower gene counts to deduce if there is a correlation between EGFR counts and survival .

Lastly, although EGFR mutations were not present in many patients, the MAF analysis revealed a variety of locations where EGFR mutations occurred. All mutation sites observed were missense mutations. Though each site only appeared once, four out of the twelve observed mutation sites were on the furin-like cysteine region of the EGFR gene; this region accounted for half of all male mutations on the gene. This supports the conclusions of Saif and Chu, which reported that one of the most common EGFR mutations in CRC appeared on the furin-like cysteine region as a missense mutation. There were more male patients who displayed mutations at that region than female patients, suggesting that there many also be a sex-based difference on EGFR mutation sites.

The findings from this analysis indicate that gender is not a primary determining factor of EGFR expression and subsequent survival rates. Further research directions include exploring other risk factors such as age, race, and family history and their associations with EGFR expression and mutation. Other research can also delve into sex-based differences in EGFR mutations for CRC, as the MAF analysis revealed that all mutation sites were mutually exclusive to male and female patients. However, given that not many patients had EGFR mutations in this dataset, redoing the MAF analysis with a larger sample size may also be of interest.

References

Chang, C.-H., Lee, C.-H., Ho, C.-C., Wang, J.-Y., &amp; Yu, C.-J. (2015). Gender-based impact of epidermal growth factor receptor mutation in patients with nonsmall cell lung cancer and previous tuberculosis. *Medicine*, 94(4), 44–62. https://doi.org/10.1097/md.0000000000000444

Levin, B., Lieberman, D. A., McFarland, B., Smith, R. A., Brooks, D., Andrews, K. S., Dash, C., Giardiello, F. M., Glick, S., Levin, T. R., Pickhardt, P., Rex, D. K., Thorson, A., & Winawer, S. J. (2008). Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: A joint guideline from the American Cancer Society, the US multi-society task force on colorectal cancer, and the American College of Radiology. *CA: A Cancer Journal for Clinicians*, 58(3), 130–160.

https://doi.org/10.3322/ca.2007.0018

Markowitz, S., & Bertagnolli, M. (2010). Molecular basis of colorectal cancer. *New England Journal of Medicine*, 362(13), 1245–1247. https://doi.org/10.1056/nejmc1000949

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., & Rodriguez Yoldi, M. (2017). Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *International Journal of Molecular Sciences*, 18(1), 197–201. https://doi.org/10.3390/ijms18010197

Saif, M., &amp; Chu, E. (2010). Biology of Colorectal Cancer. *The Cancer Journal: The Journal of Principles & The Practice of Oncology*, 16(3), 196–201. https://doi.org/https://doi.org/10.1097/PPO.0b013e3181e076af

Wang, L., Xiao, J., Gu, W., & Chen, H. (2016). Sex difference of EGFR expression and molecular pathway in the liver: Impact on drug design and cancer treatments? *Journal of Cancer*, 7(6), 671–680. https://doi.org/10.7150/jca.13684

Ward, E., DeSantis, C., Robbins, A., Kohler, B., & Jemal, A. (2014). Childhood and adolescent cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, 64(2), 83–103. https://doi.org/10.3322/caac.21219

General Concepts

1. What is TCGA and why is it important?
    a. TCGA is a cancer genomics program hosted by the National Cancer Institute and the National Human Genome Research Institute. The publicly available data from this project includes genomic, epigenomic, transcriptomic, and proteomic data. This data was collected from 20,000 different samples that span 33 different cancer types. The data in TCGA is important because the data can then be used to improve diagnosis, treatment, and prevention of cancer. TCGA is important for answering questions like which genes have associated effects in cancer, which groups are more likely to have mutations in cancer-associated genes, etc.
2. What are some strengths and weaknesses of TCGA?
    a. Some strengths of TCGA is its diversity of cancer types and its availability. TCGA has data on 33 cancer types, emphasizing that the types of cancer the data covers is quite comprehensive. Furthermore, TCGA is publicly available, making it much more financially accessible than other datasets. This availability allows multiple researchers from different institutions to collaborate on data collection and analysis. Some weaknesses are its size and lack of transparency with data acquisition and collection. The size can make it difficult to download and process, and oftentimes, a strong base knowledge in coding and software will be needed in order to analyze the data. Transparency is also a weakness for large publicly available datasets like TCGA; oftentimes, the data collection is not documented well or at all. This can be a problem from minoritized groups, which have historically faced medical exploitation in the past. There are some steps that

TCGA has taken to address ethical concerns with documentation of informed consent, but as a large publicly available database, there might be incomplete information on how data was collected and acquired.

3. How does the central dogma of biology (DNA → RNA → protein) relate to the data we are exploring?
    a. TCGA has extensive proteomic, genomic, and transcriptomic data for many cancers. Proteomic data is data on protein expression and translation, genomic data is on the physical and chemical structure of DNA, and transcriptomic data is on gene expression as quantified through RNA. All three data types are connected as seen with the central dogma, so keeping that central dogma in mind is important in understanding what each kind of data represents and communicates.

Coding Skills

1. What commands are used to save a file to your GitHub repository?
    a. git add, git commit -m "Message describing commit," git push
2. What command must be run in order to use a package in R?
    a. library(package)
3. What is boolean indexing? What are some applications of it?
    a. Boolean indexing checks to see which members of a dataframe or vector satisfy a certain condition, then creating a vector of only boolean values that say whether or not the original value meets that condition. An application of it would be removing NA's from a dataset; by using boolean indexing, the NA values can be obtained by index, and consequently removed by index without ever actually

having to go into the vector or dataframe and changing values. Another application would be categorizing data; for example, if there was a dataframe holding information about number of mutations, boolean indexing would be useful in determining which members of the dataset have more than 1 mutation.

4. Draw out a dataframe of your choice. Show an example of the following and explain what each line of code does.

Dataframe called covid_cases

day_of_week	num_cases
Sun	10
Mon	8
Tues	4
Wed	2
Thurs	7
Fri	12
Sat	15

a) an ifelse() statement

i) covid_cases$case_amount = ifelse(covid_cases$num_cases > 10, “high”, “low”)

1) This code creates a new column called case_amount in the covid_cases dataframe. It categorizes the cases as “high” or “low,” where num_cases is labeled “high” if the daily case number is more than 10, and “low” if otherwise.

b) boolean indexing

i) low_cases = covid_cases$num_cases(covid_cases$num_cases < 10)

ii) covid_cases = covid_cases[, !low_cases]

1) This code updates the data frame to only include the days of the week (rows) where the number of cases is greater than or equal to 10. It creates a vector of only true and false values; days with under 10 cases return true and those with 10 or more return false. The covid_cases dataframe is updated to only include cases that are not considered low.