Vamsi Chavali

David Wen & Nicole Black

QBIO 490: Public Data Analysis Group

13 Mar 2022

# Mid-Semester Project

## Part 1: Scientific Paper

### Introduction

Colorectal cancer (CRC) is the third most common cancer and the fourth most common cause of cancer-related death. With 1.93 million new cases and 920,000+ deaths and rising per year, the effect of CRC is faced globally in great magnitudes (Xi and Xu, 2021). Despite its gruesome impact, it is evident that it disproportionately affects males and females (Yang et al., 2017). Because of this, scientists have attempted to uncover the molecular mechanisms that cause for this difference. In hopes of curing and treating CRC scientists have found that adenocarcinoma detection rate is a key indicator in patient outcomes, with greater detection rates being associated with lower lifetime risks and mortality (Meester et al., 2015). Realizing that improving detection rates is key to patient outcomes, scientists and researchers continue to study this illness at the molecular level through a multi-omics perspective. Through this approach they have found that gut microbiota and mutations that target oncogenes, tumor suppressor genes and genes related to DNA repair mechanisms are responsible for colorectal carcinogenesis. These effects are most observed as common mutations, chromosomal changes, or translocations. Specifically, through genomics research they have found mutations within the *KRAS, BRAF*, *PIK3CA, SMAD2* and *SMAD4* genes (Marmol et al., 2017) are responsible for CRC. These findings are used predictive markers for detection and indicate patient outcomes. This paper aims

to corroborate these findings by exploring the effects of mutation and survival rates within these genes across male and female patients.

It is hypothesized that gene expression will not differ between males and females and similarly there will be no difference in survival probability between males and females expressing certain mutations. To conduct this analysis, publicly available data from Cancer Genome Atlas (TGCA) will be utilized. Specifically, statistical analyses will be conducted through RStudio and using the DESeq2, maftools, and TCGABiolinks libraries amongst others. The findings show that these genes are not upregulated or downregulated more in either males or females. However, we were able to conclude that KRAS gene mutations do present a difference in survival probability between males and females.

## Methods

Determining the role of sex in colorectal cancer outcomes was analyzed through exploring differences in survival and regulatory effects in specific genes. These analyses were conducted in RStudio using colon cancer clinical data and RNAseq data from TCGA. Using TCGAbiolinks with accession code "COAD" the data was extracted and explored. In RStudio, a plethora of libraries including DESeq2 for differential expression analysis and volcano plots, SummarizedExperiment and maftools for mutation data analysis, and ggplot2 along with survival and survminer were used to create basic plots, survival plots and volcano plots. As a "control", analysis on survival probability between males and females was conducted through constructing survival plots and boxplots to illustrate the findings. Following this, mutation analysis was conducted on to corroborate the initial findings while also being conducted on specific genes. This analysis aimed to explore the relationship between mutations and survival probability to assess the impact of gender on these variables. Finally, to improve the validity of the analyses',

differential expression analysis was performed to determine regulation and expression of specific genes across genders.

## Results (Figures)

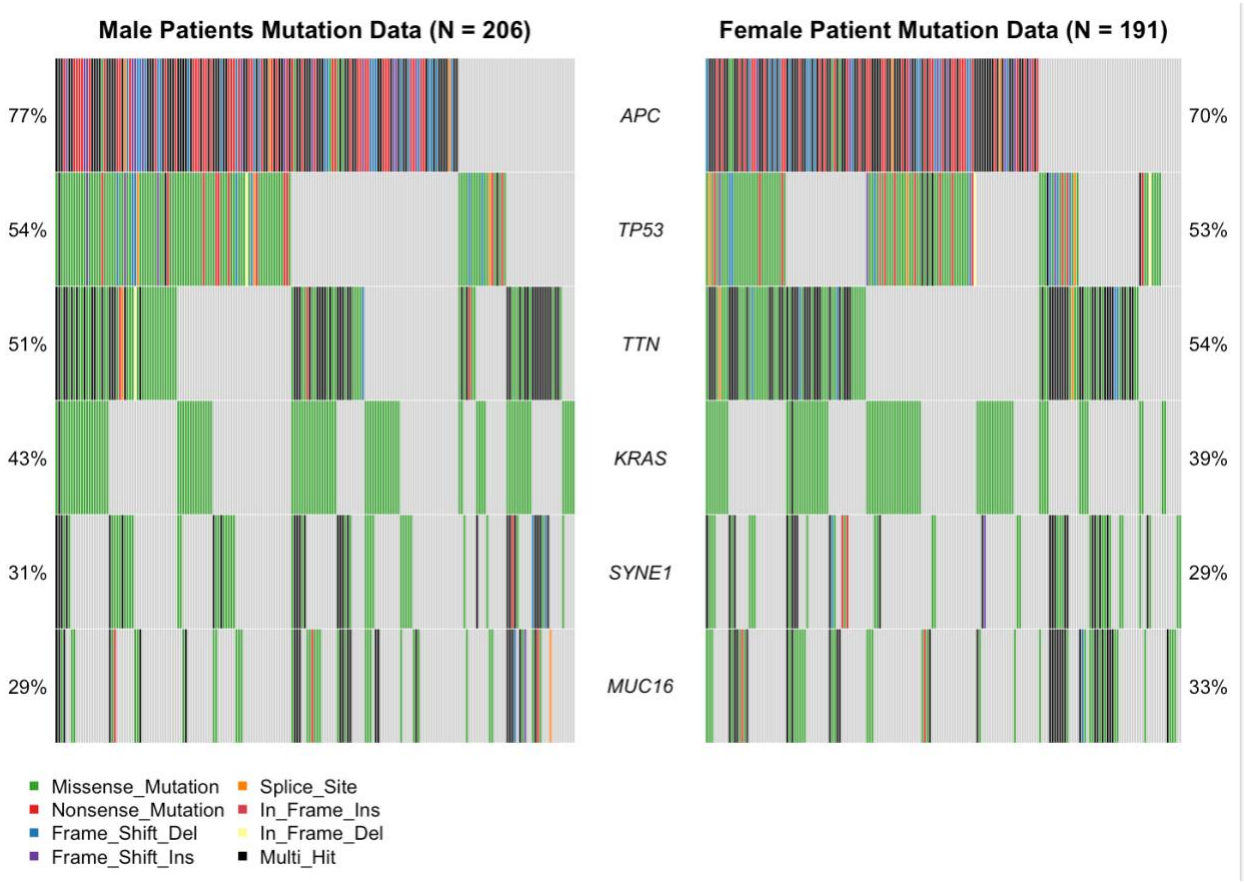


*Figure 1. Plot showing mutation rates of the top 6 mutated genes across males and females.*

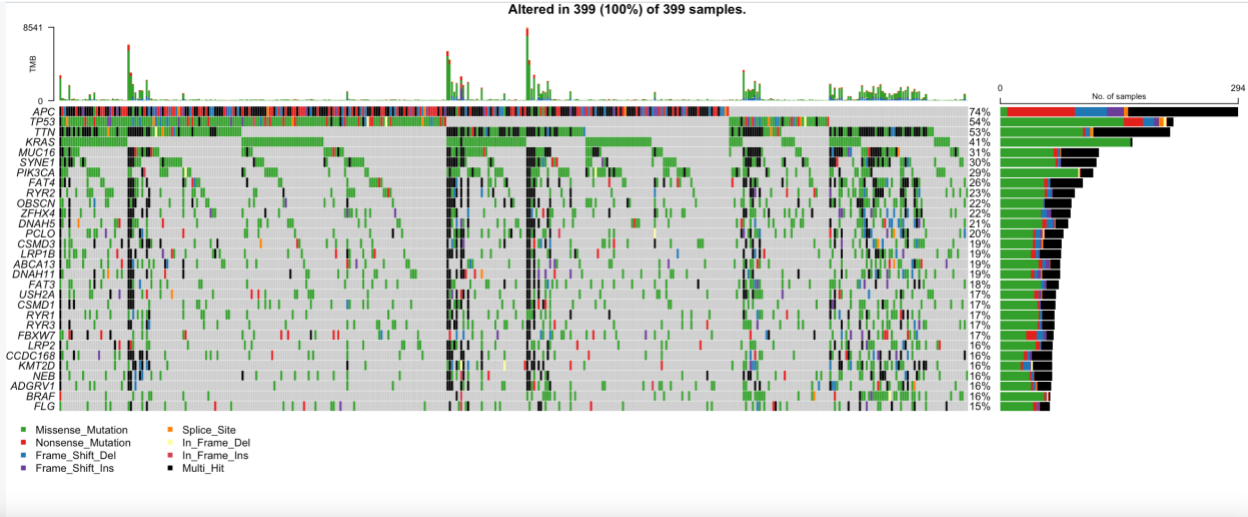


*Figure 2. Plot showing the most mutated genes in CRCs.*

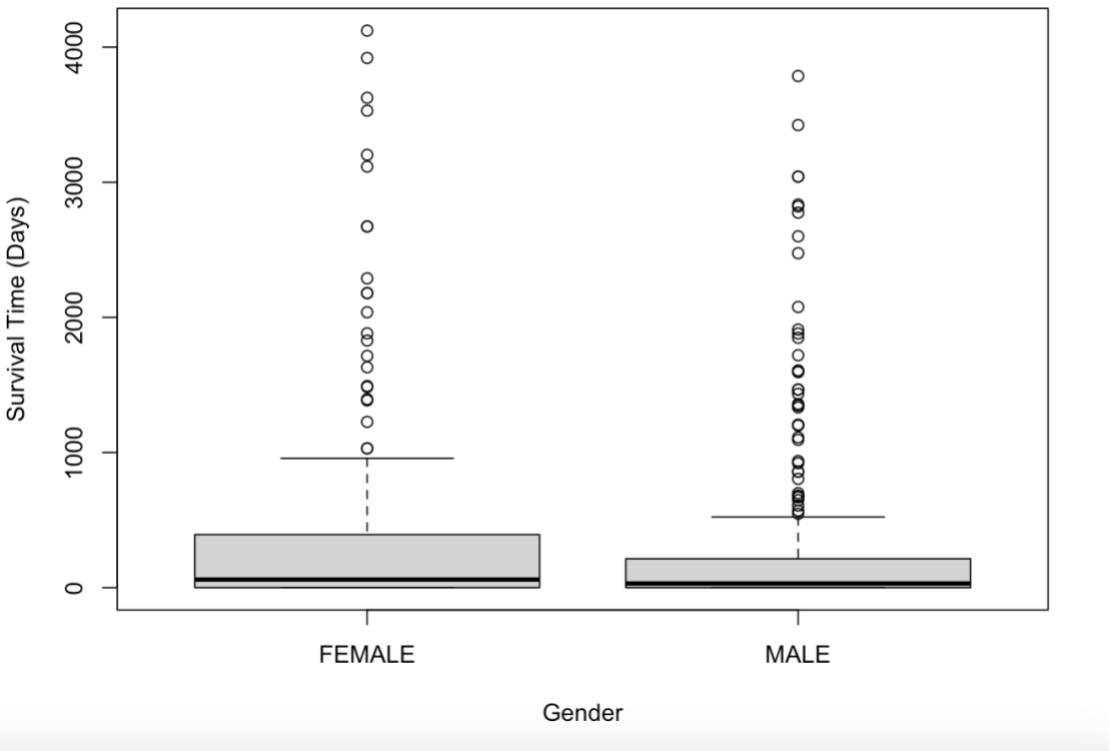


*Figure 3. Boxplot showing survival time of males and females.*

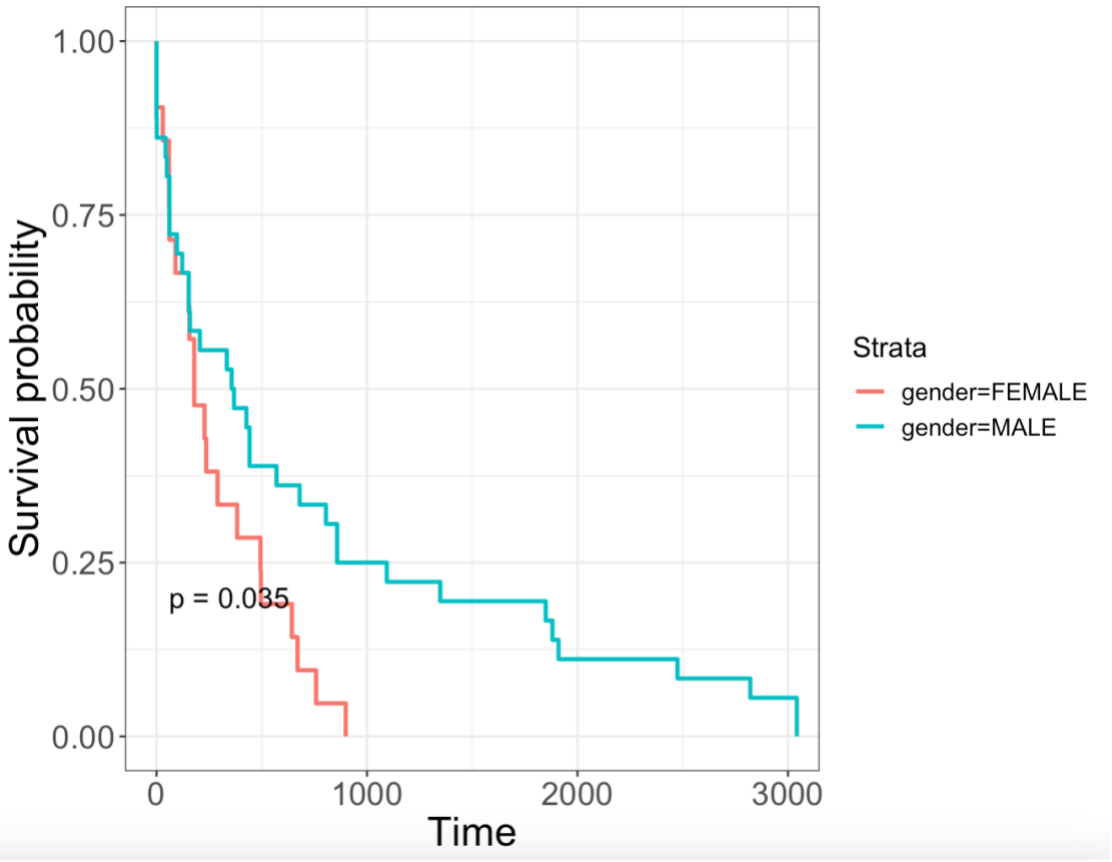


*Figure 4. Survival plot indicating the effect of gender in those with CRCs.*

*Table 1. Table outlining p-adjusted values of gene expression in males and females.*

Wald test p-value: gender male vs female
DataFrame with 5 rows and 6 columns

| | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj |
|---|---|---|---|---|---|---|
| | <numeric> | <numeric> | <numeric> | <numeric> | <numeric> | <numeric> |
| ENSG00000157764 | 735.977 | 0.04389738 | 0.0362917 | 1.209569 | 0.2264442 | 0.377407 |
| ENSG00000133703 | 1917.660 | -0.05588104 | 0.0359577 | -1.554077 | 0.1201660 | 0.300415 |
| ENSG00000121879 | 595.757 | 0.06286390 | 0.0329058 | 1.910419 | 0.0560793 | 0.280397 |
| ENSG00000175387 | 2196.295 | -0.00628724 | 0.0366872 | -0.171374 | 0.8639296 | 0.863930 |
| ENSG00000141646 | 1626.444 | -0.04852419 | 0.0483051 | -1.004535 | 0.3151206 | 0.393901 |

*Table 2. Table outlining the IDs for specific genes to correlate with Table 1.*

| | |
|---|---|
| ensembl_braf_id | "ENSG00000157764" |
| ensembl_kras_id | "ENSG00000133703" |
| ensembl_pik3ca_id | "ENSG00000121879" |
| ensembl_smad2_id | "ENSG00000175387" |
| ensembl_smad4_id | "ENSG00000141646" |

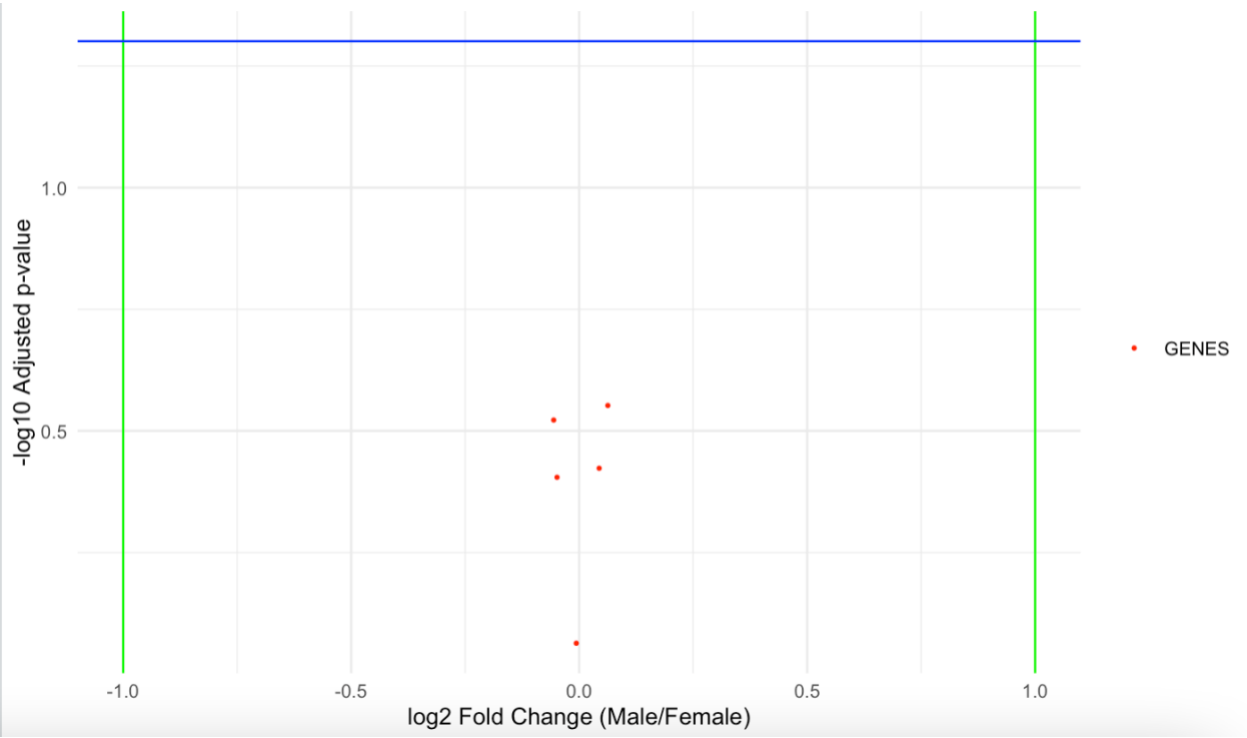


*Figure 5. Volcano plot outlining the expression levels of specific genes across males and females. Must have log2 fold value greater than 1 or p-adj value less than 0.05 to be significant.*

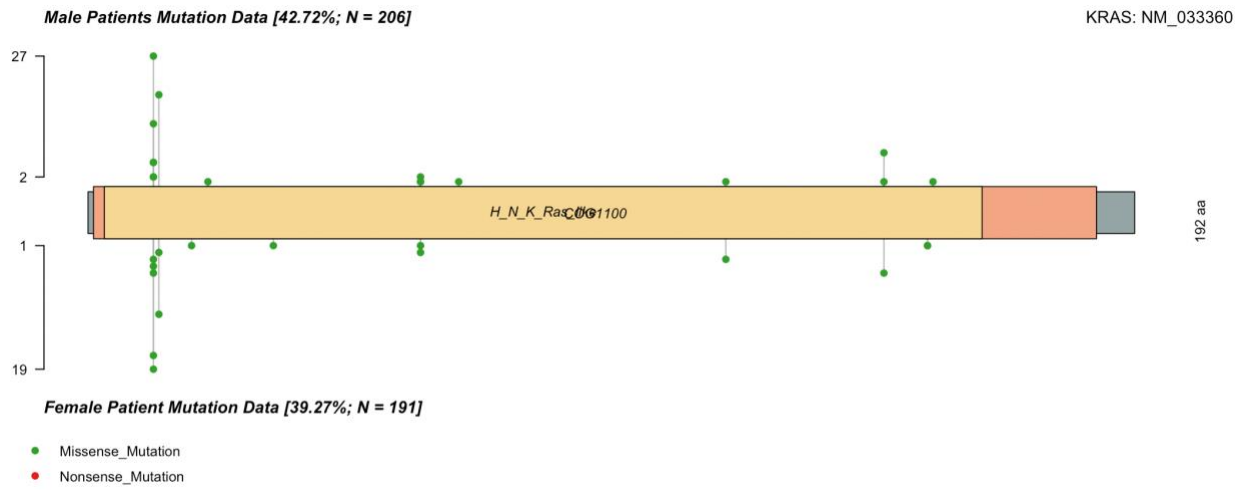


*Figure 6. KRAS gene mutations in  males and females.*

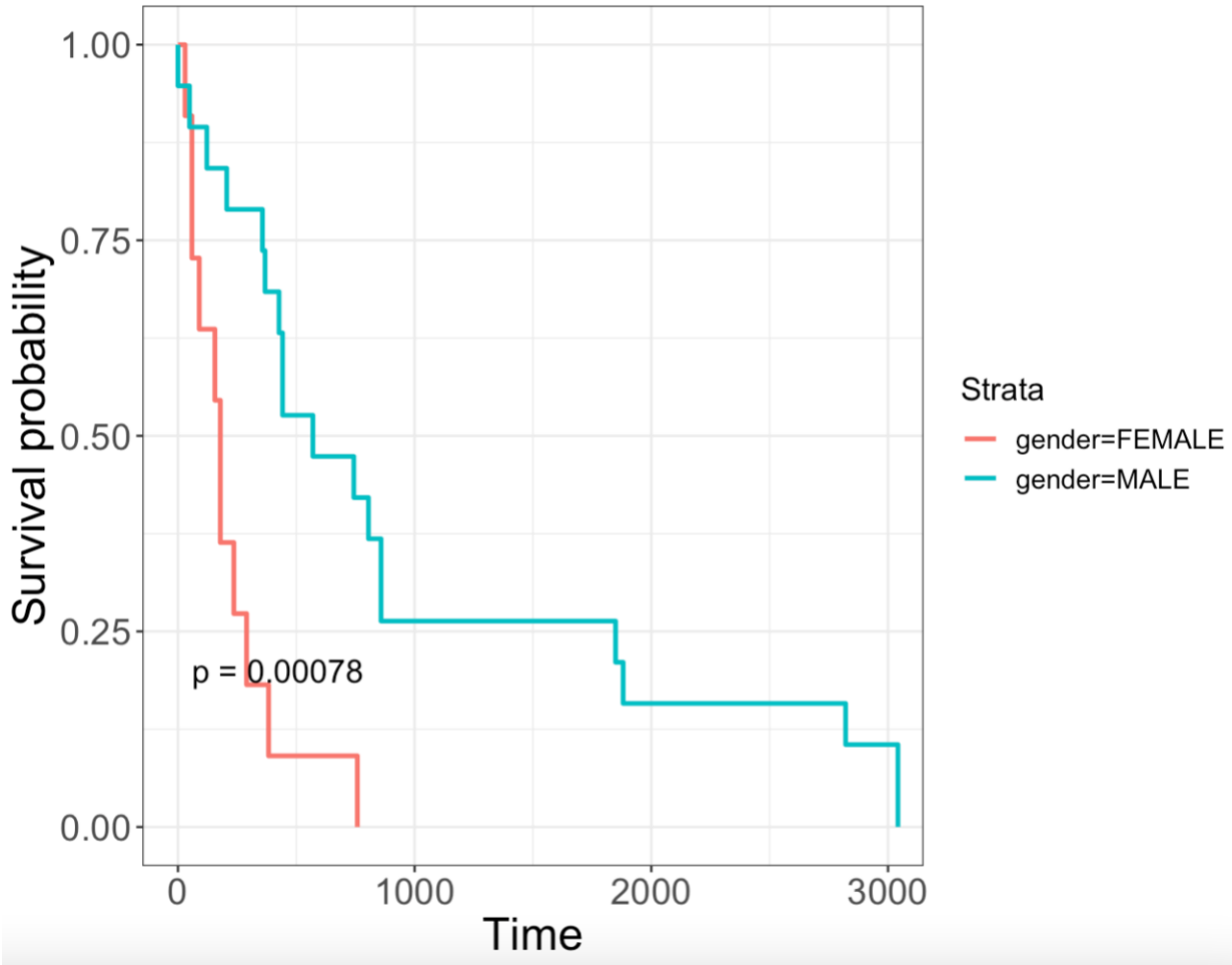


*Figure 7. Survival plot based on KRAS gene mutations in males and females.*

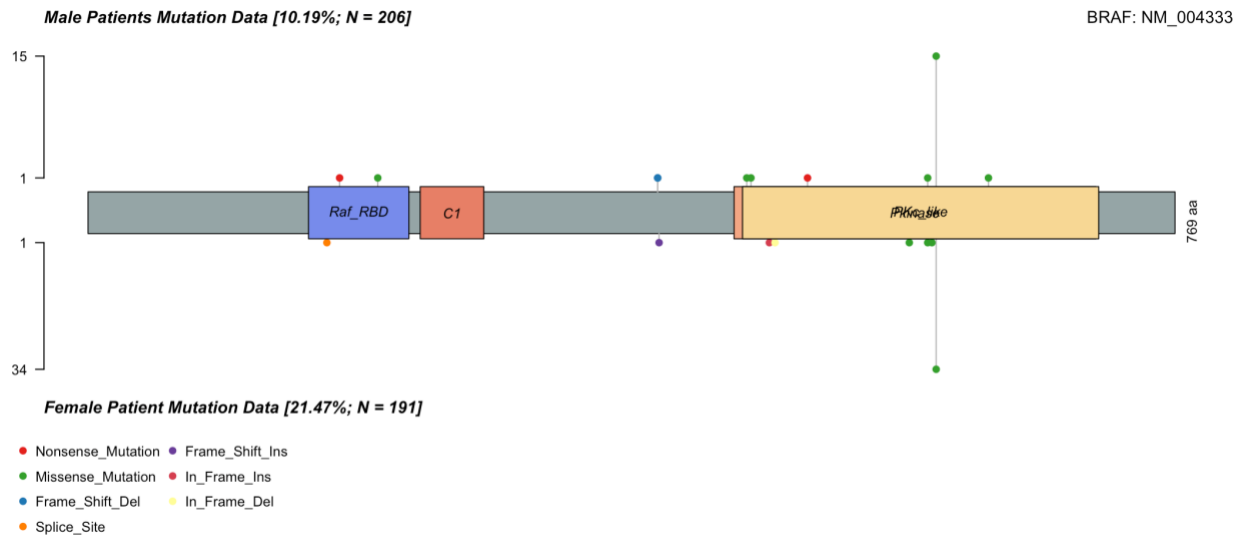


*Figure 8. BRAF gene mutations in males and females.*

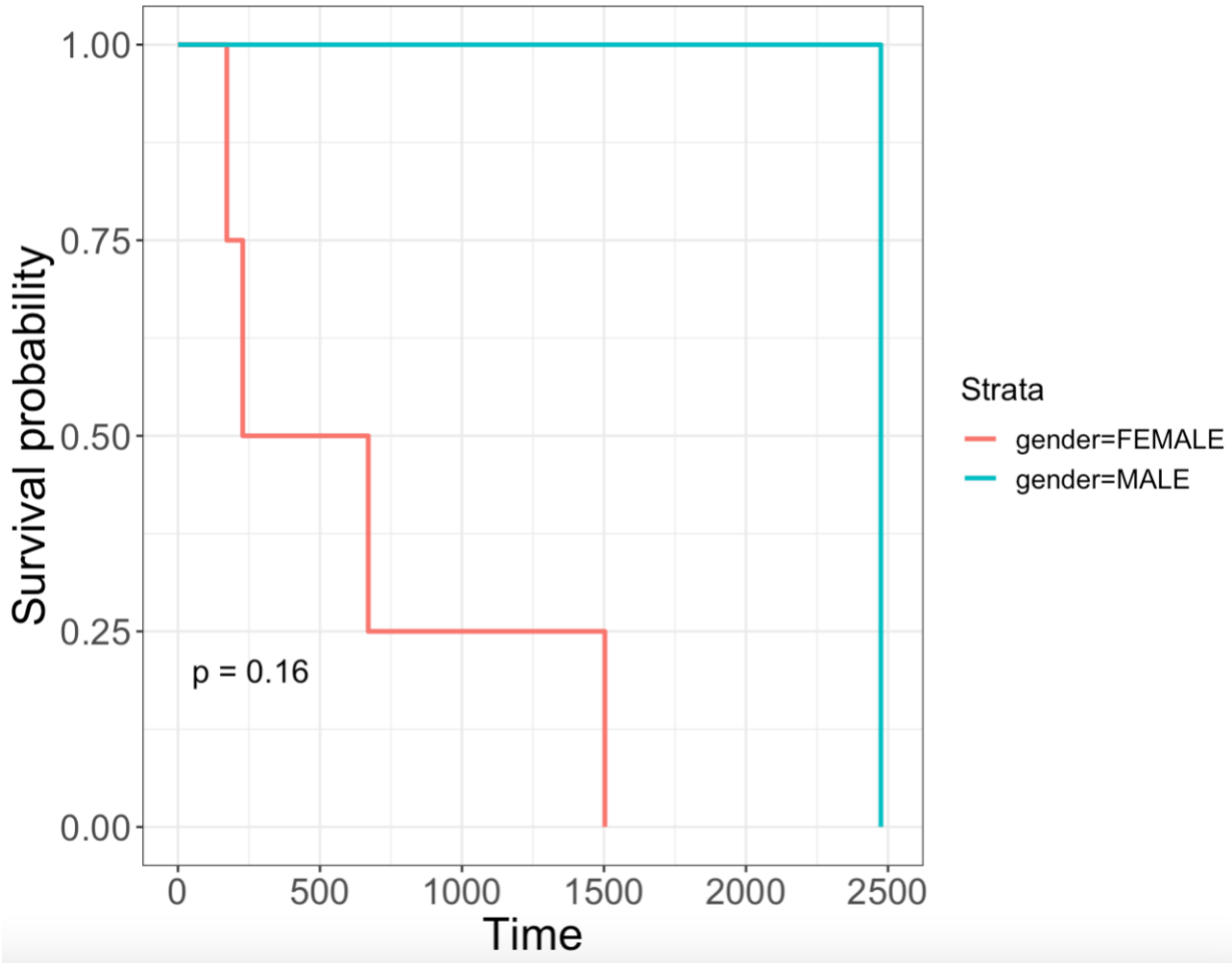


*Figure 9. Survival probability plot of BRAF gene mutations in males and females.*

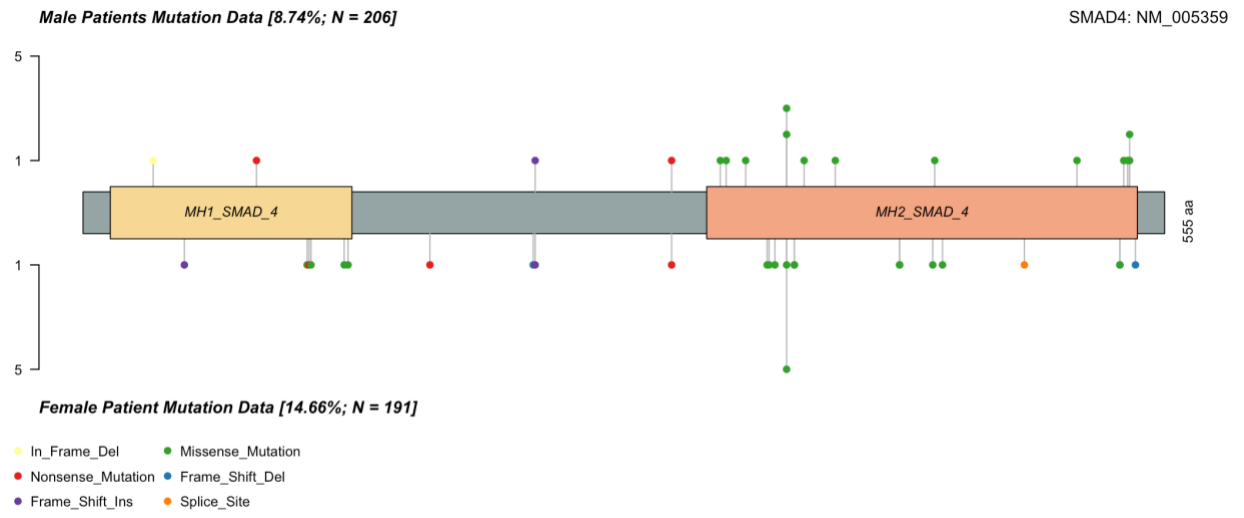


*Figure 10. SMAD4 mutations in males and females.*

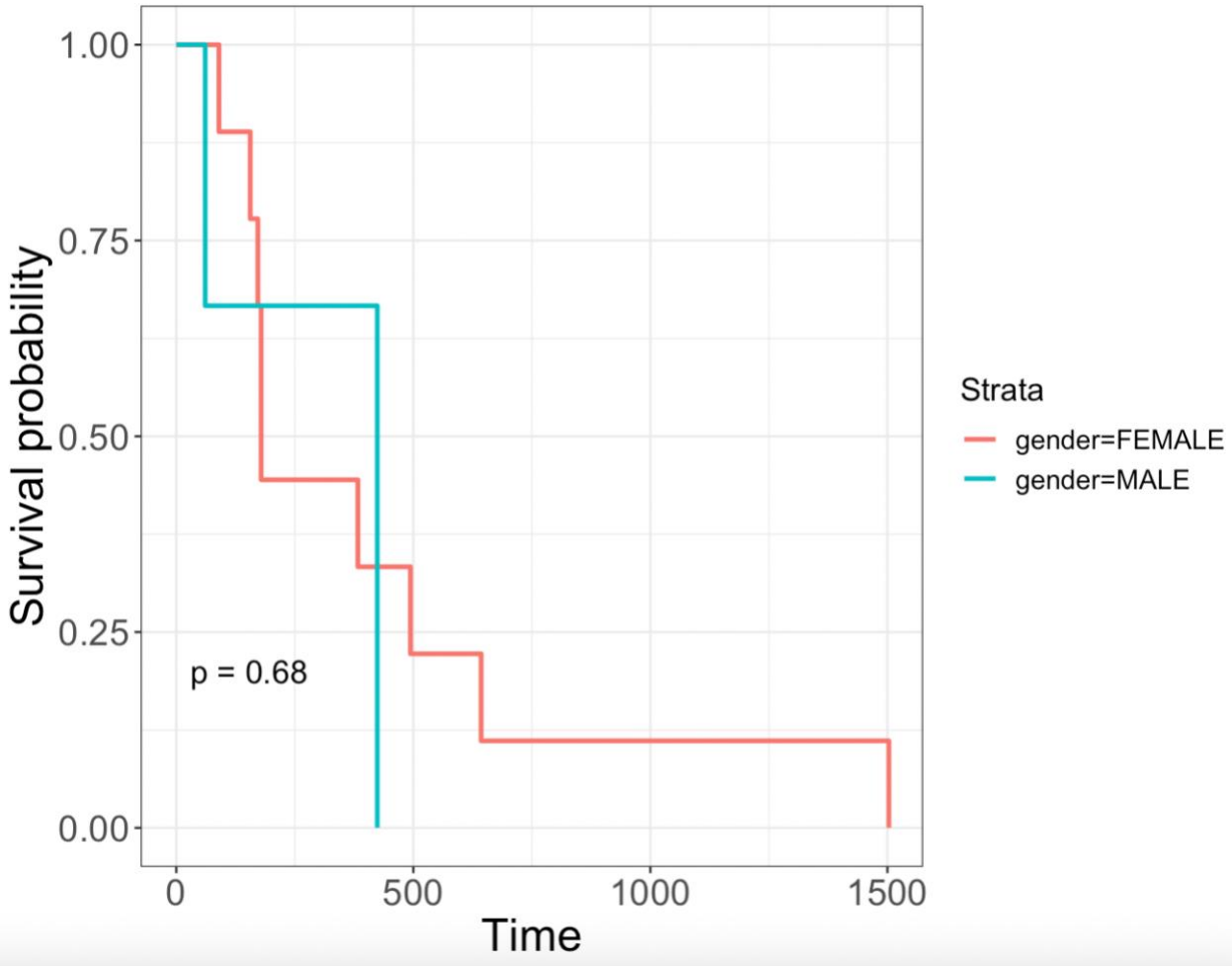


*Figure 11. Survival probability plot of SMAD4 mutations in males and females.*

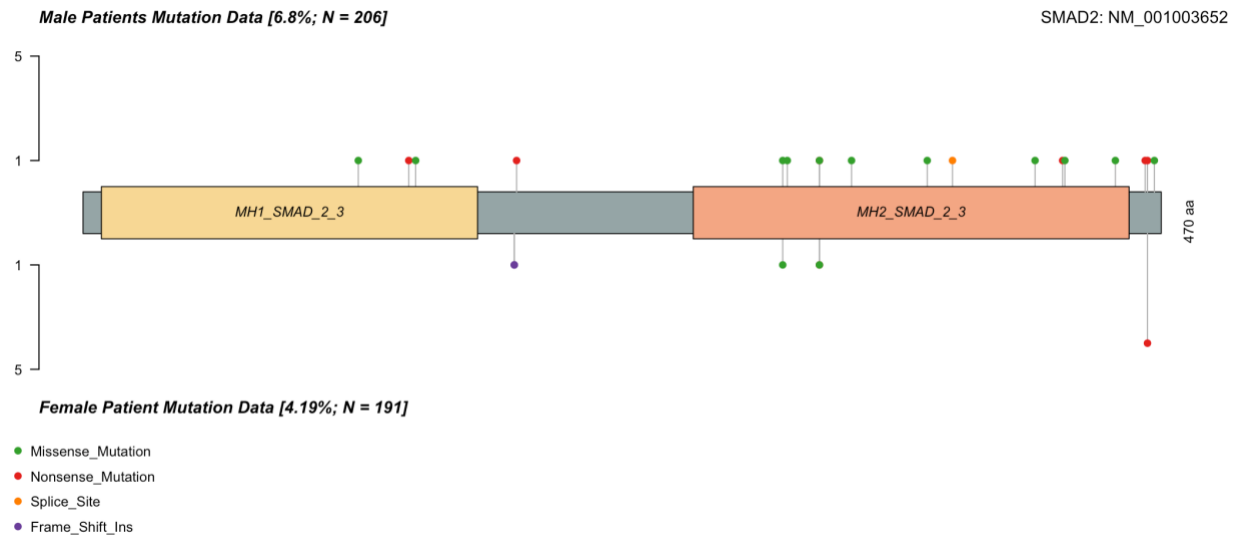


*Figure 12. SMAD2 mutations in males and females.*

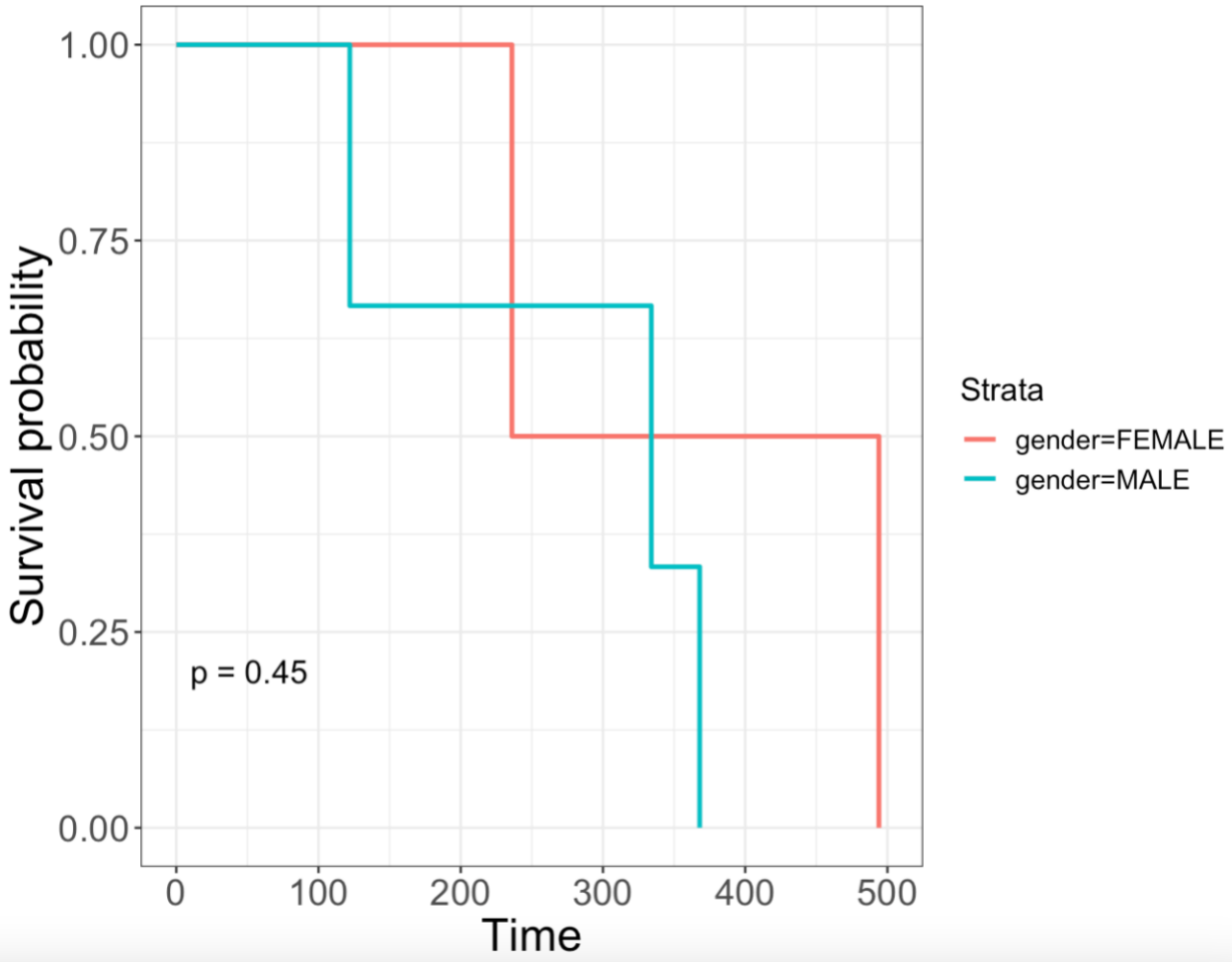


*Figure 13. Survival probability plot of SMAD2 mutations in males and females.*

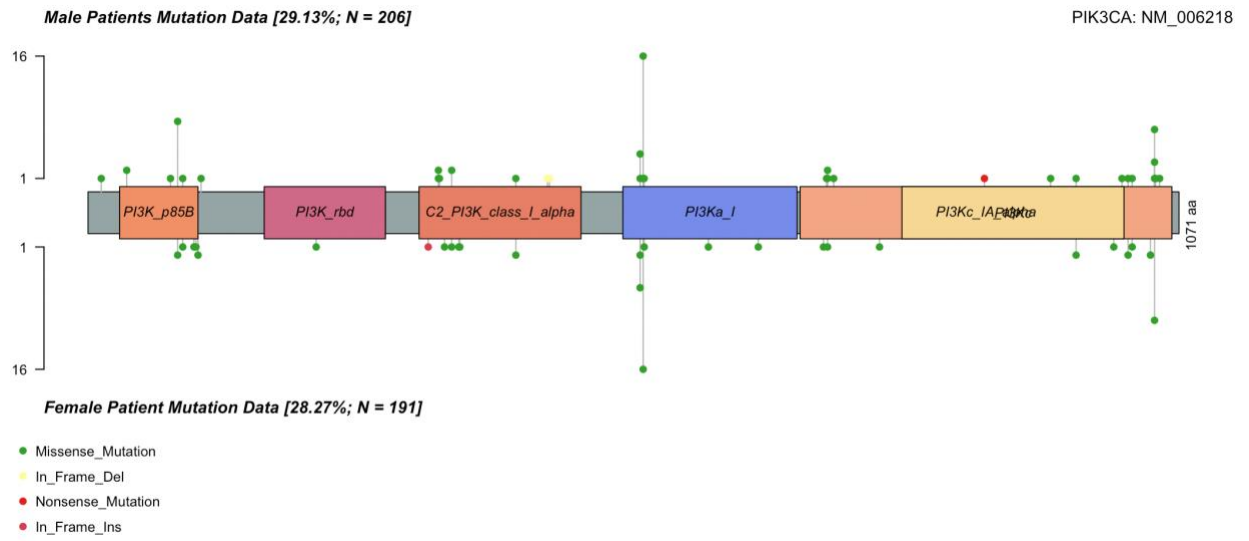


*Figure 14. PIK3CA mutations in males and females.*

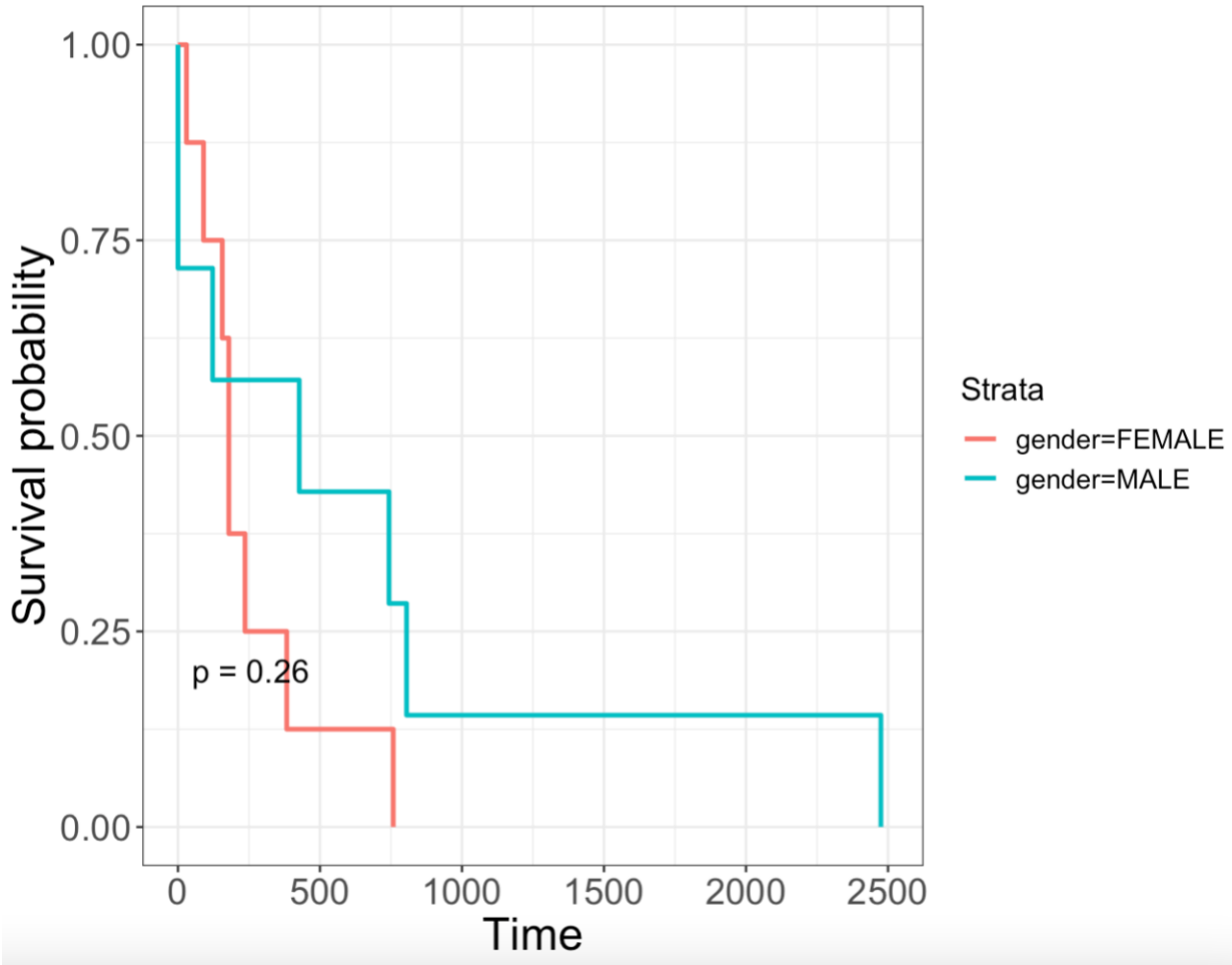


*Figure 15. Survival plot of PIK3CA mutations in males and females.*

## Results

Analysis of survival rates across different genders, indicates that gender plays a role in survival from CRC. As seen through a p-value of 0.035, the null hypothesis must be rejected, and it is observed that gender impacts survival (Figure 4).

Through the differential expression gene analysis, it was found that none of the *KRAS*, *BRAF*, *PIK3CA*, *SMAD2* and *SMAD4* genes are differentially expressed across males and females (Table 1 & Figure 5). The p-values (p-adj for Table 1) are all greater than 0.05 for those figures and tables analyzing expression and regulation, thus the null-hypothesis is failed to be rejected, indicating that gender differences most likely do not attribute to regulation patterns of the *KRAS, BRAF*, *PIK3CA*, *SMAD2* and *SMAD4* genes.

These findings are confirmed by the data that shows mutations in *BRAF*, *PIK3CA*, *SMAD2* and *SMAD4* genes does not lead to statistically significant differences in survival across genders (Figures 9, 11, 13, and 15). This is because the p-values are greater than 0.05, we fail to reject the null hypothesis, indicating that mutations of these genes across males and females most likely does not impact survival. Despite this finding, the mutation rates between males and females are vastly different in BRAF genes (Figure 8) with roughly 10% of the male population having BRAF mutations while nearly 20% of the female population shares similar mutations.

Moreover, the data does show that mutations in KRAS gene within females does affect survival probability as shown by the p-value of 0.00078 (Figure 7). Additionally, this occurrence may the explanation for the findings (Figures 3&4) that outline that gender does impact survival with CRC. Despite this, the mutation rate data within male and female populations for the KRAS gene is nearly equivalent with both males and females facing 40% occurrence of these mutations (Figure 6).

## Discussion

This study aimed to explore the effects of gender on mutation regulation and survival of colorectal cancer. It built upon previous studies that found specific genes responsible for colorectal carcinogenesis. Despite the aims, many of the genes explored are not within the top 30 mutated genes responsible for CRC (Figure 2). More statistically relevant findings would probably be found through further exploration of the most mutated genes as opposed to the genes selected. Additionally, the statistical findings and lack thereof of this study must be contextualized before being taken as true. Because of difficulty in data collection efforts, much of the data that determined survival rates and longevity was approximated to the most recent checkup from initial diagnosis of CRC. Although this may not heavily skew data, it must be considered before evaluating or building upon the findings of this study.

Despite these discrepancies, this study was able to conclude that gender impacts survival from CRC and specifically that KRAS mutations within can affect survival probability between genders. The impact of gender on survival from CRC has been found in other studies as well (Yang et al., 2017). In a recent study published in the World Journal of Gastroenterology, scientists found that women over 65 show higher mortality and lower survival rates and that this finding can be attributed to a higher portion of women presenting with right-sided colon cancer (Kim et al., 2015). To further this study in conjunction, it may be interesting to study if cells in the right-side of the colon express the KRAS differently, thus affecting survival.

Additionally, because this study inconclusively determined the role of gender in colorectal cancer outcomes, it is essential that future research be directed towards understanding this the mechanisms and pathways that protect one sex more than the other, to improve overall outcomes. Future studies should aim to correlate survival rates and probability with upregulation

to further tangible solutions that be solved through regulated genes and their expression. Hopefully, through such an approach, the impact of the second most deadly cancer will be reduced.

## References


Kim, S. E., Paik, H. Y., Yoon, H., Lee, J. E., Kim, N., & Sung, M. K. (2015). Sex-and gender-specific disparities in colorectal cancer risk. *World journal of gastroenterology: WJG*, *21*(17), 5167.

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., & Rodriguez Yoldi, M. J. (2017). Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *International journal of molecular sciences*, *18*(1), 197.

Meester, R. G., Doubeni, C. A., Lansdorp-Vogelaar, I., Jensen, C. D., van der Meulen, M. P., Levin, T. R., ... & van Ballegooijen, M. (2015). Variation in adenoma detection rate and the lifetime benefits and cost of colorectal cancer screening: a microsimulation model. *Jama*, *313*(23), 2349-2358.

Xi, Y., & Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Translational Oncology*, *14*(10), 101174.

Yang, Y., Wang, G., He, J., Ren, S., Wu, F., Zhang, J., & Wang, F. (2017). Gender differences in colorectal cancer survival: a meta-analysis. *International journal of cancer*, *141*(10), 1942-1949.


## Part 2: Review Questions

### General Concepts

1) The Cancer Genome Atlas (TCGA) is a cancer genomics program that has characterized 20,000+ primary cancer types at the molecular level. Through a joint effort between the NCI and the National Human Genome Research Institute, TCGA has compiled over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. This data is publicly available and has been vital for diagnosing and treating cancer.

2) TCGA is incredibly powerful with over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data is easily accessible and free to conduct

analyses upon, thus making collaboration more possible and furthering the field of computational biology. Through TCGA, scientists and researchers have used this tool to treat and prevent cancers.

TCGA has tangible weaknesses that have been observed within instructional settings as well. Primarily, to truly garner the full value of TCGA requires an incredible amount of training. With researchers not having this training, accessing, and using TCGA can be difficult. Another weakness includes the fact that clinical data can be missing or spotty as some samples are untreated or do not receive follow-up. Additionally, TCGA does not have immune-oncology data which is especially useful when studying cancer.

3) The central dogma of biology dictates that DNA is transcribed into mRNA (RNA) and then translated into protein. Using the framework, we have delved into exploring colorectal cancer, by analyzing different datasets to understand various aspects of tumorigenesis and mutagenesis. Moreover, using a multi-omics perspective incorporates this underlying framework as genomics is the study of DNA, transcriptomics is the study of RNA, epigenomics is the study of methylation patterns affections gene expression, and proteomics is the study of proteins. Each "omic" presents a different viewpoint of the same process, thus the central dogma is essential to our exploration.

Coding Skills

1) To save a file to GitHub, you must first initialize the GitHub repository you are attempting to save it to. First navigate to the specific folder and then "git init." Then you must "git add" along with the relative or absolute file path (cd/Documents/qbio_folder/filename). Next, you must "git commit -m" and enter a message about the file for tracking purposes. Finally, you must "git push" to ensure that the file is pushed to your repository.

Commands:

cd /filepath/filename
git init
git add /filename/
git commit -m “label”
git push

2) To use any package in R, the package must first be downloaded and called to be run.

Commands:

BiocManager::install(“Package”)
library(Package)

3) Boolean indexing is a mechanism to filter or sort data by creating true/false values for specific data frames and its rows or columns. This method creates a Boolean vector that can be used as a mask to select specific data from larger data frames. It is especially useful for removing missing values or selecting values based upon a threshold.

4) Data Frame:

cancer_data

| Patient | Name | Age | Height | Weight | Cancer Status |
|---|---|---|---|---|---|
| 701 | Robert Sunshine | 71 | 68 | 154 | II |
| 702 | Irene Moon | 56 | 64 | 128 | I |
| 703 | Phillip Gold | 61 | 71 | 173 | IV |

5) a) cancer_data$age_category = ifelse(cancer_data$Age > 50, TRUE, FALSE)
This line of code creates a new row wherein “TRUE” or “FALSE” is added to new row labelled “age_category” based upon whether their age exceeds 50 or not.

b) young_patients = cancer_data[(cancer_data$Age < 40),]

This line of code filters out the patients that are older than or equal to 40. This is done by assigning true or false values based on the condition described (age < 40), thus resulting in the vector that has the necessary separated data.