Minwoo Cho

QBIO Student Research Group Midterm

3/5/22

Part 1

Introduction

Colorectal cancer is the cancer related to the colon and rectum. Because of the similarities between the two, they are often grouped together to describe this type of cancer. It typically affects older adults and early stage signs include polyps that form in the colon (Mayo Foundation, 2021). These are clumps of cells which later can metastasize. General causes are from mutations in the DNA. Mutations in certain genes significantly increases susceptibility to cancer such as genes that control cell cycle and proliferation such as APC, TP52 and KRAS to name a few (CDC, 2022). To analyze the effects of gene mutations in relation to gender and general survival based on gender, computational analysis using TCGA data was utilized. Overall, there was a significant difference in survivability based on gender as males had decreased survivability probability which was also in correlation with increased mutations in cell cycle regulatory genes. However, most mutation genes showed no significant correlation to gender, as mutation in those genes induced cancer regardless of gender.

Methods

TCGAbiolinks, SummarizedExperimen and maftools librarys were downloaded and into the R coding platform to conduct analysis specifically on gender. The clinic dataframe was created the clinical.data from maf_object dataframe. Then, Boolean masks were created by accessing the gender column in the clinic dataframe. Using the same method, patient_ids based on the gender masks were also accessed. Then, using the two masks, male_maf factor was created using subsetmaf. Similarly, masks and subsetmaf was created for females. Only then the coOncoplot could be created using male_maf and female_maf which gave the 6 most mutated genes in male and female patients and the mutations associated with them.

Next, the volcano plot was created using the following method. The DESeq2 library was loaded and the counts and patient_data dataframes were isolated from the original sum_exp dataframe. Because there were myriad of NA values present in the data, they were removed from both subset dataframes. The gender column in patient_ data was converted to a factor and genes that showed fewer than 10 mutated occurrences were removed by creating a Boolean mask and

creating a subset of the counts dataframe. Then, DESeqDataSetFromMatrix was run to create dds_obj to get the results. Threshold values were indicated to be 2 for log2foldchange_threshold and 0.05 for padj. Signicant data points would lie outside of these thresholds. Finally, there were graphed on a volcano plot based on gender (male/female).

The Kaplain Meirs curve was created using similar lines of code. The survival and survminer libraries were loaded. To visualize the probability of survival, days_to_death and survival_time columns were created in the clinic dataframe. The NA values in days_to_death were replaced with days_to_last_follow_up and both columns were converted to numeric values with as.numeric. Then, the survplot was created with as survival_time vs. death_event based on gender.

Results

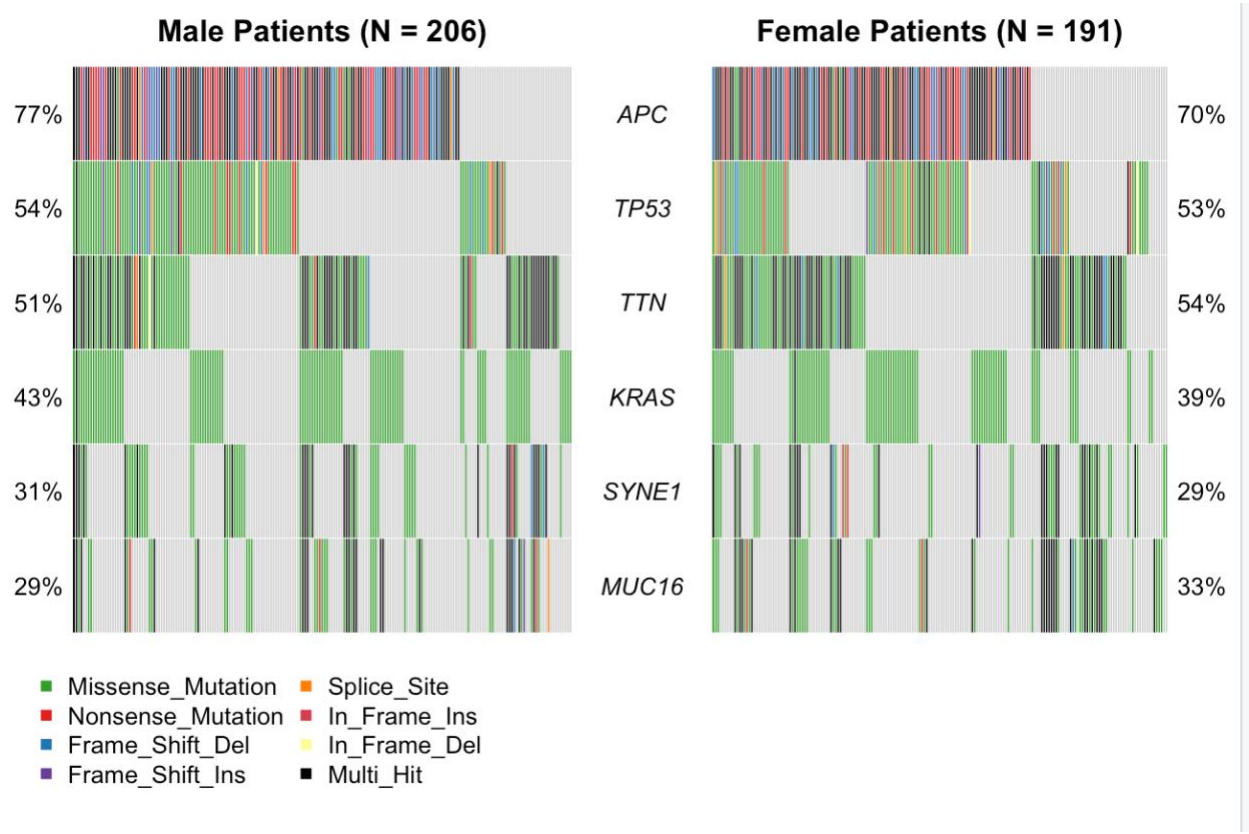

Figure 1: CoOncoplot of Male and Female Patients expressing 6 most mutated genes in colorectal cancer data

In the coOncoplot, 6 common mutated genes, APC, TP53, TTN, KRAS, SYNE1 and MUC16 are shown. The percentage of mutations in 4 of the 6 genes are higher in males then in females. Especially the APC gene is severely mutated in male patients with 77% while it is lower

in females with 70%. The colors represent different types of mutations such and missense mutations seem to be the most common in all genes except APC where all mutations appear to be present.

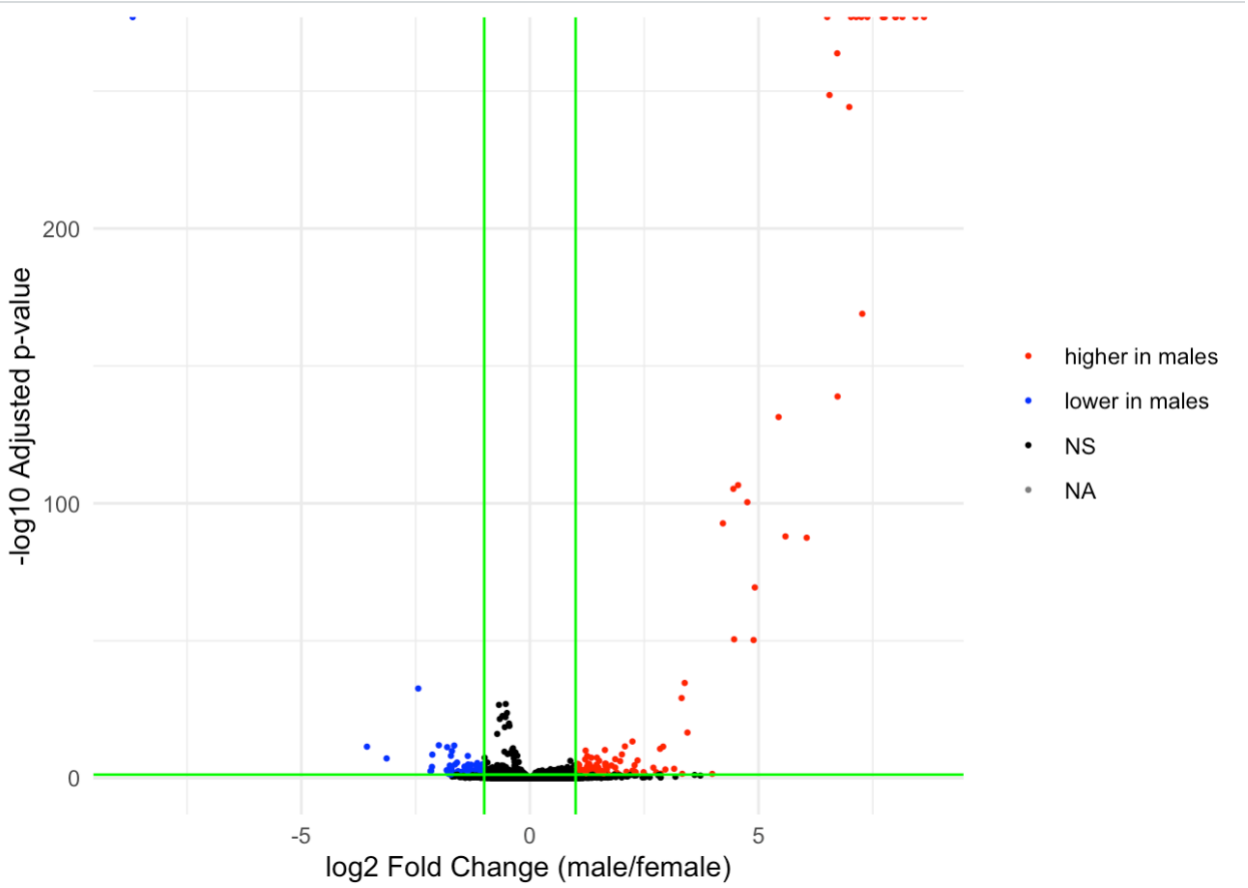

Figure 2: Volcano plot of Significance of Gene Mutations in Male and Females

This plot shows the p-values of gene mutations in relation to gender, the red representing genes showing higher mutations in males and blue representing genes mutated lower in males. Significant p-values lie outside the threshold indicated by the green lines and higher occurrence of the red can be observed. This indicates that there are more genes mutated in males then females although there are few exceptions. However, p-values of the mutated genes higher in the red dataset which gives more statistical evidence regarding these mutated genes expressed higher in males.

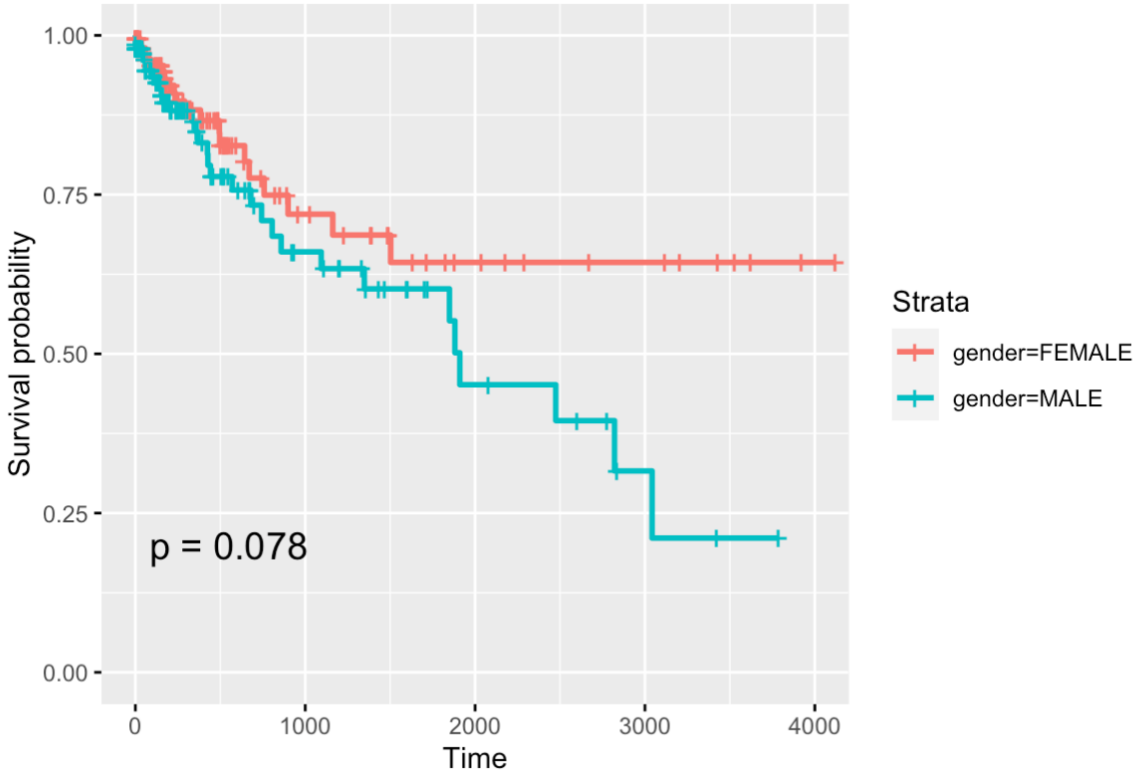

Figure 3: Kaplain Meirs Curve of Survival Probability vs. Time regard to Gender

The KM curve depicts the decrease in survival probability with time after diagnosis in males and females. In females, the line of interest plateaus around after 2000 days while the slope of the line remains negative in males. With time, survival probability decreases for both genders but the rate of decline is significantly faster in males, especially after the 2000 day mark.

Discussion

From the graphical analyses, there was clear indication that gender was a significant factor in survivability and progression of colorectal cancer. Just looking at figure 1, it is unclear whether there is a difference between genders in terms of frequency of mutations. The most mutated genes are closely mutated between both genders and both correlate with low probability of survival. Figure 2 supports this claim because for example, the APC gene has a p-value of 0.39 which is below the threshold and is therefore insignificant between genders. This indicates that the APC gene will be as likely mutated in males as females with colorectal cancer. However, further analysis of figure 2 reveals that there are much more genes mutated in males that show statistically significance as illustrated by the red dataset. This directly correlates with figure 3 which depicts probability of survival in females and males. Male survivability severely decreases

while females have a higher chance of survival. This can be explained by the more frequent mutations observed in males and defective genes that control cell cycle have a higher chance to create tumors and cause cancer.

Current studies have opposing results, however. A study indicated that females over 65 show higher morbidity and a lower chance of survival in 5 years in comparison to males. Also, they explain that women have a higher chance of developing right-sided colon cancer than men, which is more aggressive than cancer formation on the left (Kim et al., 2015). The contradiction in results may be due to unclear guidelines between men and women because sex specificity is not a study design in colorectal cancer. However, genes commonly associated with colorectal cancer match the data. The APC gene is the most commonly mutated and in absence, the Wnt gene is expressed and thousands of polyps develop (Munteaunu & Mastalier, 2014). The TP53 gene is also heavily associated as it acts as the "guardian of the genome" and a facilitator of carcinogenesis. Future treatment methods are focusing on immunotherapy. In a clinical study, Nivolumab and Pembrolizumab agents were used in patients with colorectal cancer and showed positive results. The stable disease rate was at 69% and encouraged researchers to potentially combine immunotherapeutic agents (Florescu-Tenea et al., 2019). However, these methods are still experimental.

References

Centers for Disease Control and Prevention. (2022, February 17). What is colorectal cancer? Centers for Disease Control and Prevention. Retrieved March 5, 2022, from https://www.cdc.gov/cancer/colorectal/basic_info/what-is-colorectal-cancer.htm#:~:text=Colorectal%20cancer%20is%20a%20disease,the%20colon%20to%20the%20anus.

Florescu-Ţenea, R. M., Kamal, A. M., Mitruţ, P., Mitruţ, R., Ilie, D. S., Nicolaescu, A. C., & Mogoantă, L. (2019). Colorectal Cancer: An Update on Treatment Options and Future Perspectives. Current health sciences journal, 45(2), 134–141. https://doi.org/10.12865/CHSJ.45.02.02

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, 21(17), 5167–5175. https://doi.org/10.3748/wjg.v21.i17.5167

Mayo Foundation for Medical Education and Research. (2021, June 11). Colon cancer. Mayo Clinic. Retrieved March 5, 2022, from https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669

Munteanu, I., & Mastalier, B. (2014). Genetics of colorectal cancer. Journal of medicine and life, 7(4), 507–511.

Part 2

General Concepts

1. TCGA is the cancer genome atlas and is a public dataset consisting of a myriad of samples form 33 different cancer types. It allows study across wide range of genes from a large patient sample which would be difficult from independent studies.
2. A major strength is the range of cancer data that it covers.
3. We are analyzing the genes related to colorectal cancer which is part of the patients' DNA. These genes are translated into RNA which synthesizes a dysfunctional protein which contributes to proliferation of tumors and development of cancer.

Coding Skills

1. Git add .
   Git commit -m "upload message"
   Git push
2. Downloading TCGAbiolinks data
   BiocManager::install("TCGAbiolinks")
   library(TCGAbiolinks)
   query <- GDCquery(project = "TCGA-COAD",
   data.category = "Transcriptome Profiling"
   data.type = "Gene Expression Quantification"
   workflow.type = "HTSeq - Counts")
   GDCdownload(query)
   sum_exp <- GDCprepare(query)
3. Boolean indexing is filtering data based on a certain test. For example, the ifelse statement uses a test to differentiate data into true or false values. We've used Boolean indexing on age. We wrote a test such as if the patient age was greater than 50. Then, we wrote that if this test was true for a patient, the patient would be assigned "old." If false, they were assigned "young." Therefore, this classified patient age data into young and old.
4.
   a. Ifelse(sum_exp$age_at_index > 50, "Young", "Old")

This is a test where if the patient ages in the age at index column in the sum_exp data set is greater than 50, they will be assigned “Young.” If the test is false, meaning the age is less than or equal to 50, they are assigned “Old.”

b. geneA_id_mask <- rowData(sum_exp)$external_gene_name == “KRAS”
ensemble_geneA <- rowData(sum_exp)$ensemble_gene_id[geneA_id_mask]
The first line creates a variable geneA_id_mask which is represented by the KRAS gene in the external_gene_name column of the sum_exp dataset. Then, the second line of code assigns the KRAS gene to its corresponding ensemble_gene_id based on the mask we created from the first line. The output is saved to ensemble_geneA.