Jonathan Le
2/28/22
QBIO-490

Part 1: Scientific Paper

Introduction

Colorectal cancer is the second leading cause of death from cancer among adults in the U.S., with more than 57,000 deaths from colorectal cancer every year (Markowitz & Bertagnolli, 2009). Thus, analyzing the molecular characterization of colorectal cancer to drive personalized medicine and improve survival rates remains an important area of biomedical research. According to Armaghany et al. (2012), APC, TP53, and KRAS are three of the most commonly mutated genes in colorectal cancer associated with tumorigenesis. Both APC and TP53 are tumor-suppressor genes that are involved in the control of the cell cycle, while KRAS is a gene that produces a protein that is involved in cell signaling pathways that control cell growth.. As such, mutations in these genes result in a transition towards uncontrolled cell growth and tumorigenesis. By understanding the role that mutations in these genes have in the colorectal cancer development process, these genes and their protein products can become biomarkers or potential drug targets that can improve patient outcomes.

In this research paper, we aim to validate the most commonly mutated genes in colorectal cancer using publicly available data from The Cancer Genome Atlas (TCGA). TCGA is a cancer genomics program that molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types and that provides a vast database of genomic, epigenomic, transcriptomic, and proteomic data for various cancer types. While our analysis confirmed the most commonly mutated genes in colorectal cancer that have been reported by previous authors, our results found that there is no difference in the expression of these genes between old and young patients. Moreover, the survival rate for old and young patients with mutations in these genes was also similar.

Methods

The clinical and genomic data for colorectal cancer for this project were accessed from the Genomic Data Commons and TCGA. The analysis for this project was done in R using several Bioconductor packages. The clinical data was accessed using the R package "TCGABiolinks". The gene mutation data was acquired via the "maftools" package, while the differential gene expression analysis was done using the "DESeq2" package. Lastly, the survival plots were generated with functions from the "survival" and "survminer" packages.

The first step in this project was to determine the genes that most commonly mutated in colorectal cancer. After the top 4 most mutated genes were identified, a differential gene expression analysis was done to determine if these genes were over- or under-expressed in older or younger CRC patients. Patients that were age 50 or older were defined as "old", while those younger than 50 were defined as "young". The last step in my analysis was segmenting the clinical data into patients that had each of the 4 most mutated genes into four different data frames and conducting a survival analysis for each subset based on age category.

Results

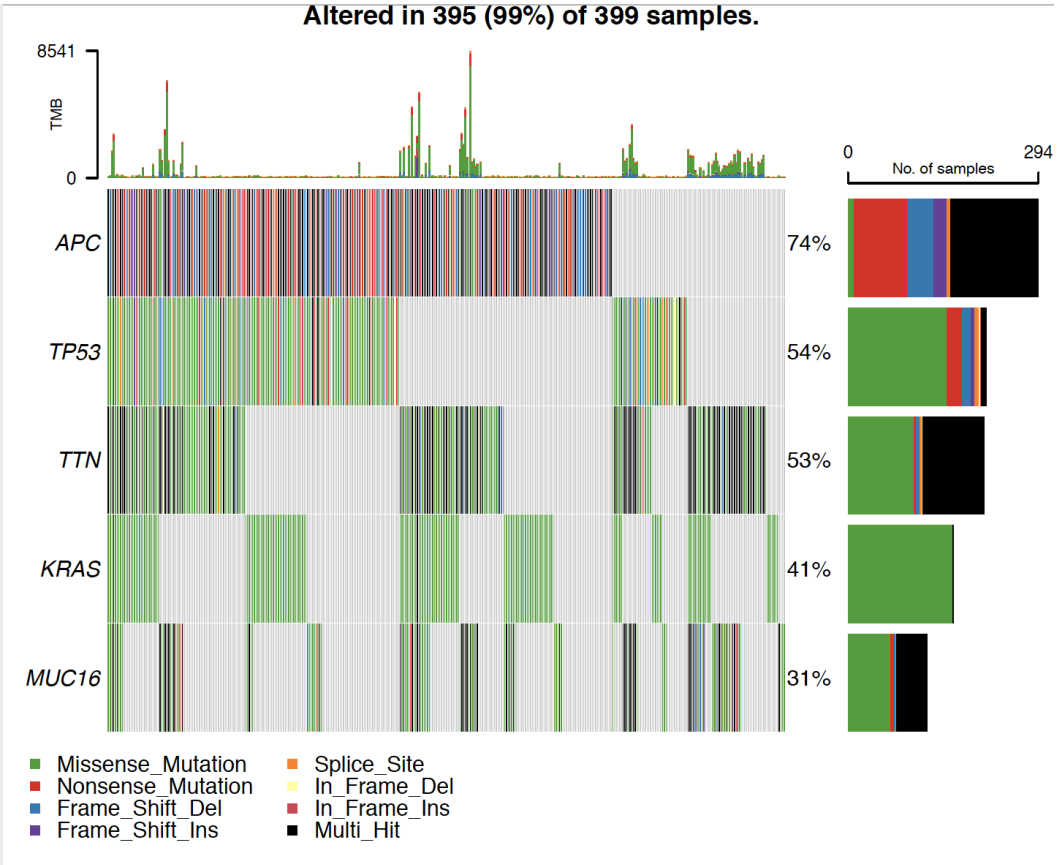

Figure 1. Oncoplot of the top 5 most mutated genes in colorectal cancer.

According to Figure 1, the top 5 most mutated genes across all colorectal cancer patients were APC, TP53, TTN, KRAS, and MUC16. APC was mutated in the most patients (74%), while TP53, TTN, and KRAS were mutated in 54%, 53%, and 41% of patients, respectively.

```
log2 fold change (MLE): age_category young vs old
Wald test p-value: age_category young vs old
DataFrame with 4 rows and 6 columns
                 baseMean log2FoldChange     lfcSE      stat    pvalue      padj
                <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
ENSG00000155657   209.125     -0.2754623 0.1868999 -1.473849  0.140522  0.562089
ENSG00000133703  1960.663     -0.0780488 0.0785549 -0.993557  0.320439  0.640877
ENSG00000134982   740.651      0.0182190 0.0757570  0.240493  0.809948  0.809948
ENSG00000141510  4235.871     -0.0537149 0.1521193 -0.353111  0.724006  0.809948
```

Table 1. DESeq2 Differential Expression Analysis for APC (ENSG00000134982), TP53 (ENSG00000141510), TTN (ENSG00000155657), and KRAS (ENSG00000133703).

Table 1 shows that although APC, TP53, TTN, and KRAS were the most commonly mutated genes in colorectal cancer patients, there was not a significant difference in expression of these genes between young and old patients. The largest difference in expression was TTN, with a log2FoldChange of -0.275, which indicates that TTN is slightly over expressed in old patients compared to young. The other genes had log fold changes near 0, which indicates that these genes are expressed at similar levels in young and old patients.

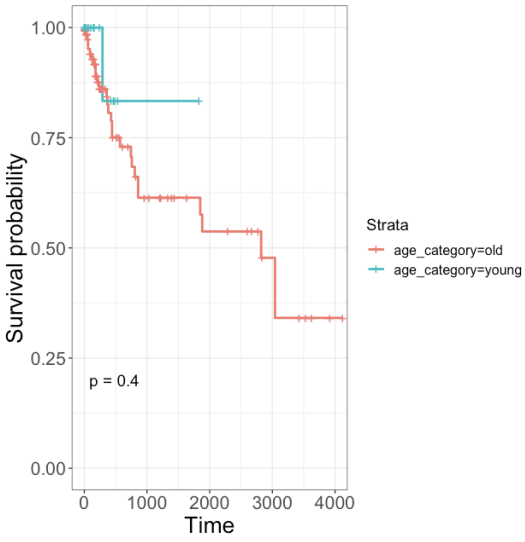

Figure 2. Kaplan-Meier survival plot by age category for patients with KRAS mutations.

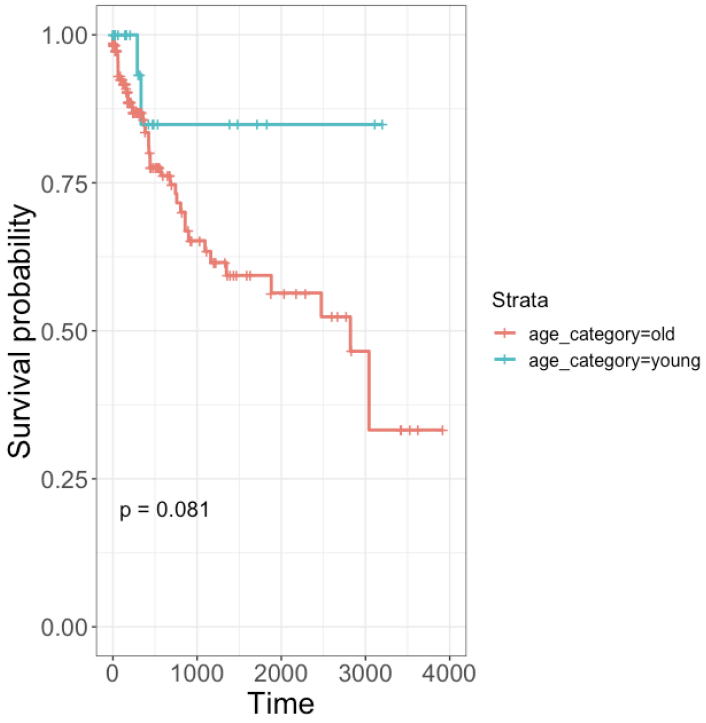

Figure 3. Kaplan-Meier survival plot by age category for patients with APC mutations.

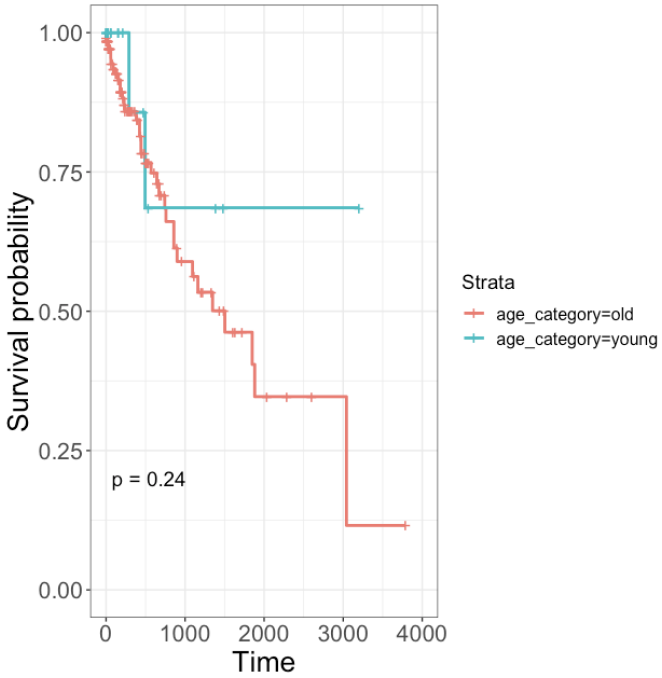

Figure 4. Kaplan-Meier survival plot by age category for patients with TTN mutations.

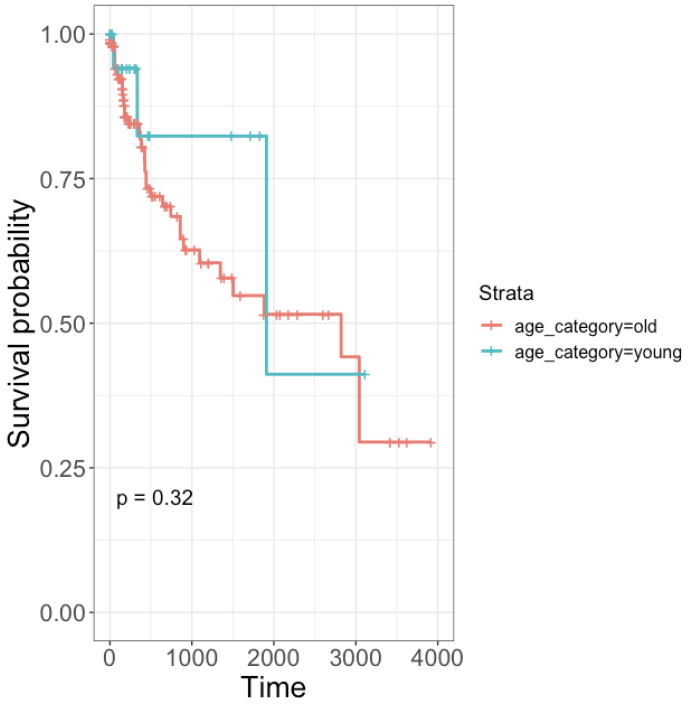

Figure 5. Kaplan-Meier survival plot by age category for patients with TP53 mutations.

The last step of this project consisted of a survival analysis of the groups of patients with each of the top 4 most mutated genes (APC, KRAS, TTN, and TP53). Based on Figures 2-5, there does not appear to be a significant difference in survival between young and old patients that have these mutations. For each of the figures, both young and old patients have similar survival rates at each of the time points. However, Figures 2 and 3 (patients with KRAS and APC mutations, respectively) may show a difference due to a difference in sample sizes of young and old patients.

Discussion

The top 4 most commonly mutated genes were found to be APC, TP53, TTN, and KRAS, respectively. This result is consistent with the findings reported previously by Armaghany et al.

(2012). However, while these genes were the most commonly mutated, our differential expression analysis showed that there was not a significant difference in the expression of these genes between young and old patients (Table 2). Thus, these genes are not over- or under-expressed in either population, which suggests that the expression of these genes may not be dependent on age.

Moreover, our survival analysis also revealed that the survival rates between young and old patients with each of these mutations was similar, suggesting that mutations of these genes have similar impacts in both young and old patients. This finding is consistent with reports from Calvert & Frucht (2002) and others that state that APC, TP53, TTN, and KRAS all play essential roles in the tumorigenesis process when mutated. Thus, the presence of these mutations are common indicators of CRC development, but may not be useful indicators for personalized prognoses for young and older patients, as the survival rates are similar for both populations.

One limitation of this study is that the sample sizes between old and young patients was the vast difference in sample size between the populations, as there were 348 old patient samples for the mutation data, while there were only 43 young patient samples. Thus, the comparison between the two populations may not be accurate, as the sample sizes are quite small. This difference in sample sizes for the mutation data is likely due to the fact that mutations are more common in older patients, as cancer is a disease that is associated with aging, which allows for more mutations to accumulate in older patients over time.

To build off of this project, future research directions could be to identify the genes that are more differentially expressed between young and old patients and the associations they may have with survival. By focusing on these genes, they may potentially be used as biomarkers and indicators of prognosis in specific patient populations, which can help guide treatment decisions.

Another potential direction of study is the repeat this study but with comparisons between men and women to identify if there are differences in gene expression between those populations. Having an understanding of the impact of the expression and mutation of specific genes in specific populations is a key step in advancing personalized medicine and is the future of biomedical cancer research.

References

Calvert, P., & Frucht, H. (2002). The Genetics of Colorectal Cancer. *Annals of Internal Medicine*, *87*(19).

Cho, K. R., & Vogelstein, B. (1990). Genetic alterations in colorectal tumors. *Hereditary Colorectal Cancer*, 477–482. https://doi.org/10.1007/978-4-431-68337-7_71

J, S., & JA, W. (2013). Cancer/testis antigens and colorectal cancer, gene expression. *Journal of Genetic Syndromes & Gene Therapy*, *04*(05). https://doi.org/10.4172/2157-7412.1000149

Markowitz, S., & Bertagnolli, M. (2009). Molecular Basis of Colorectal Cancer. *New England Journal of Medicine*, *361*, 2449–2460. https://doi.org/10.1056/NEJMra0804588

Sameer, A. S. (2013). Colorectal cancer: Molecular mutations and polymorphisms. *Frontiers in Oncology*, *3*. https://doi.org/10.3389/fonc.2013.00114

Part 2: Review Questions

General Concepts

1. What is TCGA and why is it important?

   TCGA is a cancer genomics program that molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. TCGA provides a publicly available database of genomic, epigenomic, transcriptomic, and proteomic data for various cancer types. The development and release of this database has enabled many advances in the understanding of cancer and driven forth discoveries in precision medicine, including biomarkers and drug discovery that have transformed care for cancer patients.

2. What are some strengths and weaknesses of TCGA?

   The vast amounts of data provided by TCGA have given cancer researchers an invaluable source of information about cancer genetic and epigenetic profiles, potential biomarkers, and drug targets. These data have enabled advances in the field of computational biology and deepened our understanding of the molecular characterization of cancer. The data from TCGA is publicly available and free for anyone to download, which gives anyone with the proper training to access and mine this data.

   One weakness of TCGA data is that the tumor samples come from untreated patients, so the data only provide information about baseline tumors, but no insights about genomic changes due to therapy. Another limitation of TCGA data is that it does not allow you to to map genetic and protein changes to the single cells or distinct cell populations within tumors. Lastly, a weakness of TCGA is that although the data is publicly available and free to access, it can be challenging and difficult to actually obtain the data and conduct analysis without training or guidance through an educational program like QBIO.

3. How does the central dogma of biology (DNA → RNA → protein) relate to the data we are exploring?

   Because we are studying colorectal cancer from a multi-omics perspective, we are studying the different levels of the central dogma of biology (genomics = DNA, transcriptomics = RNA, and proteomics = protein) and trying to identify relationships between these levels. For instance, we may be interested in whether mutations at the genomic level may impact expression or transcription levels of that gene, as well as what

impact that may have on protein expression. Moreover, we may also be interested in exploring differences in gene expression in different age groups or cancer samples and how that impacts protein expression or survival, which ultimately ties back to the central dogma of biology.

Coding Skills

1. What commands are used to save a file to your GitHub repository?

   To save a file to your GitHub repository, we must first do git add /PATHTO/FILE, then git commit -m “clear message about file”, then git push.

2. What command must be run in order to use a package in R?

   To use a package in R, we must first download it and call library(package name).

3. What is boolean indexing? What are some applications of it?

   Boolean indexing is a method of filtering and accessing certain subsets of a dataframe. Because R is a vectorized language, we can use this characteristic to create a boolean vector that will allow us to access the specific rows or columns that we would like to access. This boolean vector, known as the mask, will enable us to quickly select certain data from large data sets and filter the data we would like to analyze.

   Some applications of boolean indexing are to delete all NAs in a dataframe or to subset the dataframe based on a characteristic such as young vs. old or male vs. female.

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

   Basketball Player Stats Dataframe called ballers
   Stats represent each player’s average per game

Player	Points	Assists	Rebounds	Steals
Lebron James	27.8	7.3	7.4	1.1
Steph Curry	25.3	6.4	5.3	2.3
Luka Doncic	28.5	8.4	7.9	0.8
Dwight Howard	12.4	2.1	5.3	0.4

5.

a. an ifelse() statement

ballers$all_star = ifelse(ballers$points > 22 & ballers$assists > 5 & ballers$rebounds > 5, TRUE, FALSE)

This line of code would create a new column called all_star that would hold boolean values for whether a basketball player was an all star player. The if else statement checks if a player has an average points per game of greater than 22, assists per game of greater than 5, and rebounds per game greater than 5. If a player satisfies all of these conditions, then the all_star column for that player would be TRUE, while those that do not satisfy all of these conditions would have FALSE.

The new dataframe would look like this:

Player	Points	Assists	Rebounds	Steals	All_star
Lebron James	27.8	7.3	7.4	1.1	TRUE
Steph Curry	25.3	6.4	5.3	2.3	TRUE
Luka Doncic	28.5	8.4	7.9	0.8	TRUE
Dwight Howard	12.4	2.1	5.3	0.4	FALSE

b. boolean indexing

elite_scorers_mask = ballers$points > 20

This code is an example of boolean indexing and creates a mask called “elite_scorers_mask” that contains boolean values for whether a player averages more than 20 points per game. In this example, the mask would hold [TRUE TRUE TRUE FALSE]. This mask can then be used to filter or select a subset of the ballers dataframe of only players that average more than 20 points per game.