Peter Chong
Nicole Black, David Wen, and Kate Guion
QBIO 490
9 March 2022

Midterm Project

PART 1

1. What populations according to gender tend to over or underexpress gene CFTR? How does over/underexpression of gene CFTR affect survival?
2. Kaplan-Meier plot, boxplot between male and female, bar graph between male and female, volcano plot (?)
3. https://www.proquest.com/docview/2394731526/fulltextPDF/3651847766E5474FPQ/1, "Cystic Fibrosis, CFTR, and Colorectal Cancer"
    a. This paper links downregulation of the CFTR gene with a decrease in survival. It says that individuals with Cystic Fibrosis have a much higher chance of developing early and aggressive colorectal tumors. CFTR is heavily prominent in the GI tract, as it plays a main role in ion and water homeostasis, as well as intestinal crypt stem cells. There are a variety of mechanisms that the paper speculates is the cause for this increase in colorectal tumors. One such reason is that a lack of CFTR damages physical barrier protections; it causes bacterial dysphoria, damages various homeostasis in the colon, causes abnormal immune responses, and alters cellular oxidative stress.

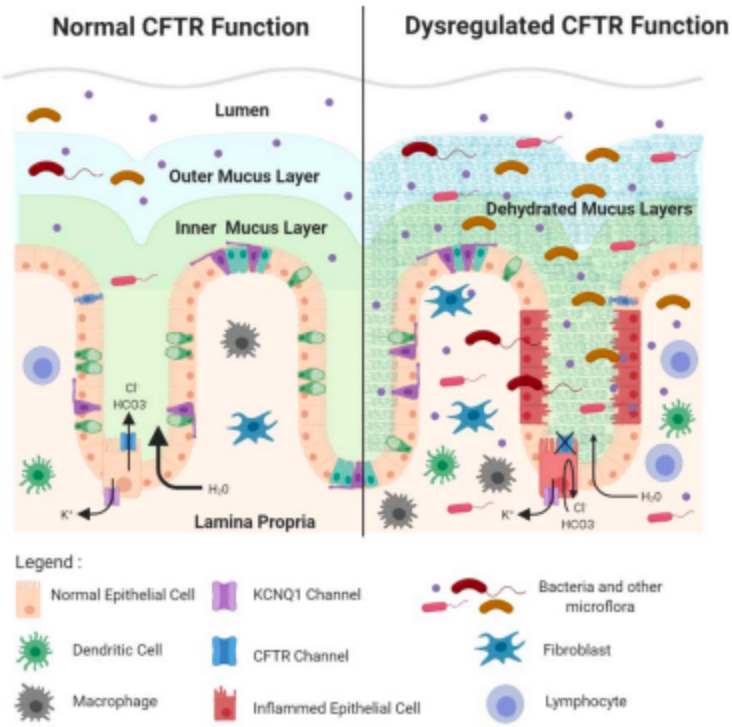
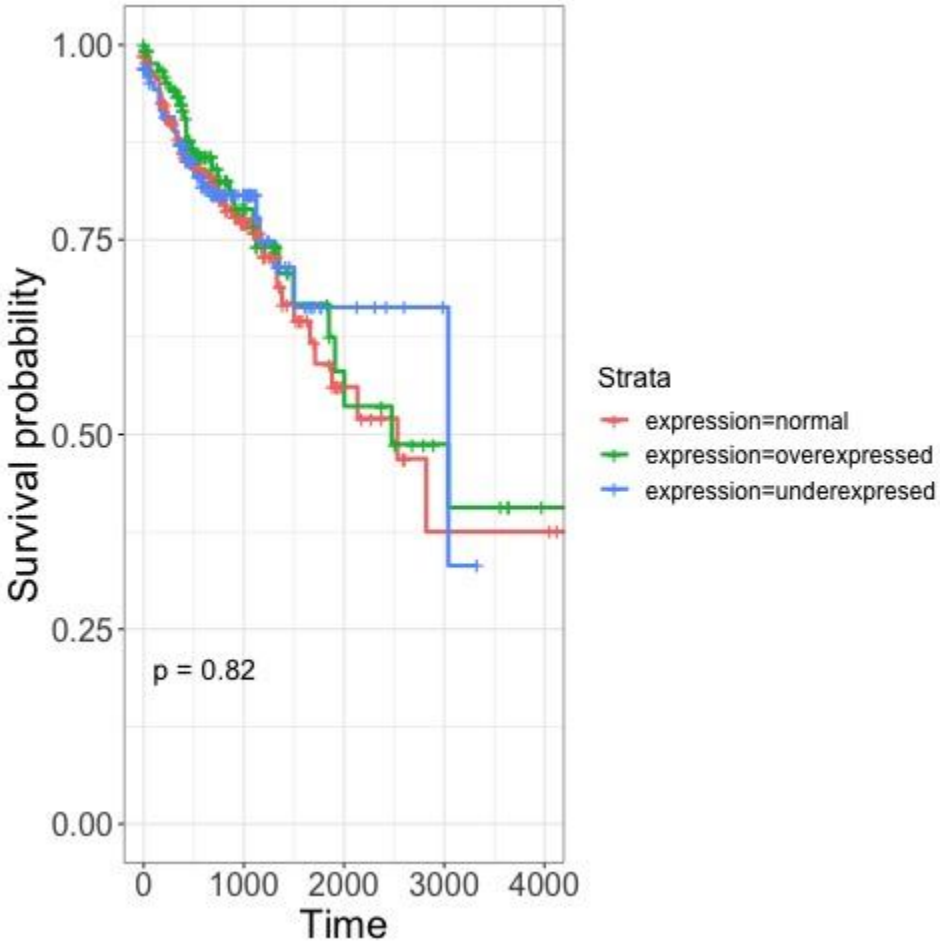


Figure 2. CFTR deficiency disrupts protective physical barriers and leads to dysbiosis. CFTR deficiency causes a failure of intestinal cell chloride and bicarbonate ion efflux and accompanying water efflux. This results in dehydration of the mucus layer, making it permissive to bacterial passage, and also causing intestinal obstruction. Disruption of the epithelial barrier leads to infiltration of commensal and pathogenic bacteria, inflammation, epithelial tissue damage, and immune cell infiltration. These alterations in the intestinal landscape (mutations, inflammatory signaling) create favorable conditions for CRC initiation and progression. Figure created using BioRender.com.

b. While this paper jumps towards the biological mechanisms behind how expression of CFTR is linked to CRC, it defies the findings from the expression Kaplan Meier curve. The paper says that underexpression of CFTR in CF patients causes colorectal cancer, thereby leading to a decrease in survival rate. However, the Kaplan Meier curve's underexpressed line maintains a higher survival plot

before sharply dropping below the other lines.



This difference in results could be because the data used to form the Kaplan Meier plot has noise from non-CFTR patients.

Introduction

Colorectal cancer (CRC) is cancer that begins in the colon or rectum, beginning with benign polyps that progress into cancerous polyps, which then cause metastasis tumors (CDC).

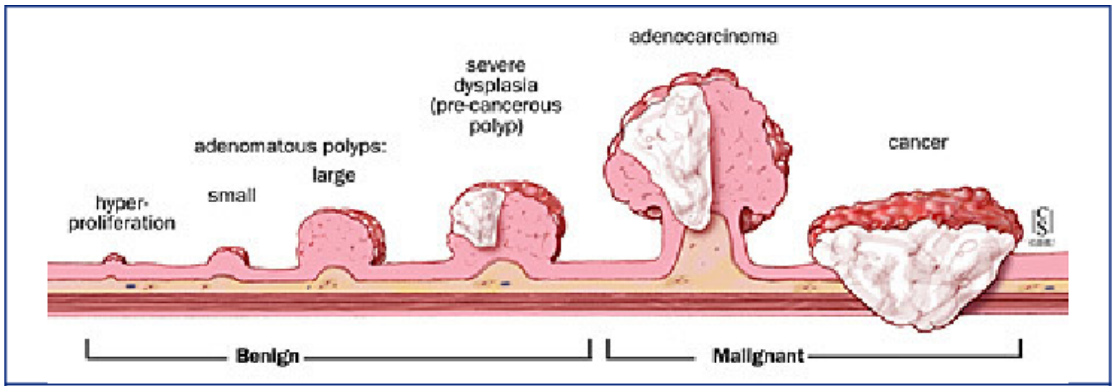


Figure 1. Benign to malignant polyp progression. Credits to "Sporadic (Nonhereditary) Colorectal Cancer: Anatomy."

It is the third most prominent cancer in the U.S. In 2022, it is estimated that 151,030 new diagnoses of it will be made. Along with that, an estimated 52,580 people will die of CRC in 2022. A number of genes are associated with CRC, one of which is CFTR, or cystic fibrosis transmembrane regulator. CFTR plays an important role in the ion and water homeostasis within the colon; an upset in the expression of it causes a host of issues, such as damaging physical barrier protections, causing bacterial dysphoria, damaging various homeostasis in the colon, causing abnormal immune responses, and alters cellular oxidative stresses (Scott et al.). At the same time, gender produces numerous biological differences; examining how CFTR and gender are linked through data analysis holds potential deeper insight.

Four primary plots from The Cancer Genome Atlas program were created towards this goal: a comparative boxplot of the gene counts between gender, a volcano plot for expression of

genes between genders, and two Kaplan Meier curves—one for over/underexpression of CFTR and one for gender. Most of the data was not indicative of correlations. The primary deduction is that females experience higher survival rates than males.

Methods

We accessed colon cancer clinical and RNAseq data from TCGA using the R package TCGAbiolinks with the accession code "COAD." We then cleaned the data, removing patients with missing gender information. We then created the boxplot. Subsequently, we then used the R package DESeq2 to create the volcano plots, saving significant results into three tables (the p-adjusted was <0.05, log2FoldChange that were greater than 1, and log2FoldChange that were less than -1). For the expression Kaplan Meier curve, we created a subset of data of counts and the patient data that only contained data associated with CFTR. Then, we estimated days_to_death by using days_to_last_follow_up as an estimate. We then created the CFTR expression Kaplan Meier curve using the built-in R plot. For the gender Kaplan Meier curve, the data was already cleaned from the previous plots; we again used the built-in R plot to create it.

Results

Counts of CFTR Between Gender

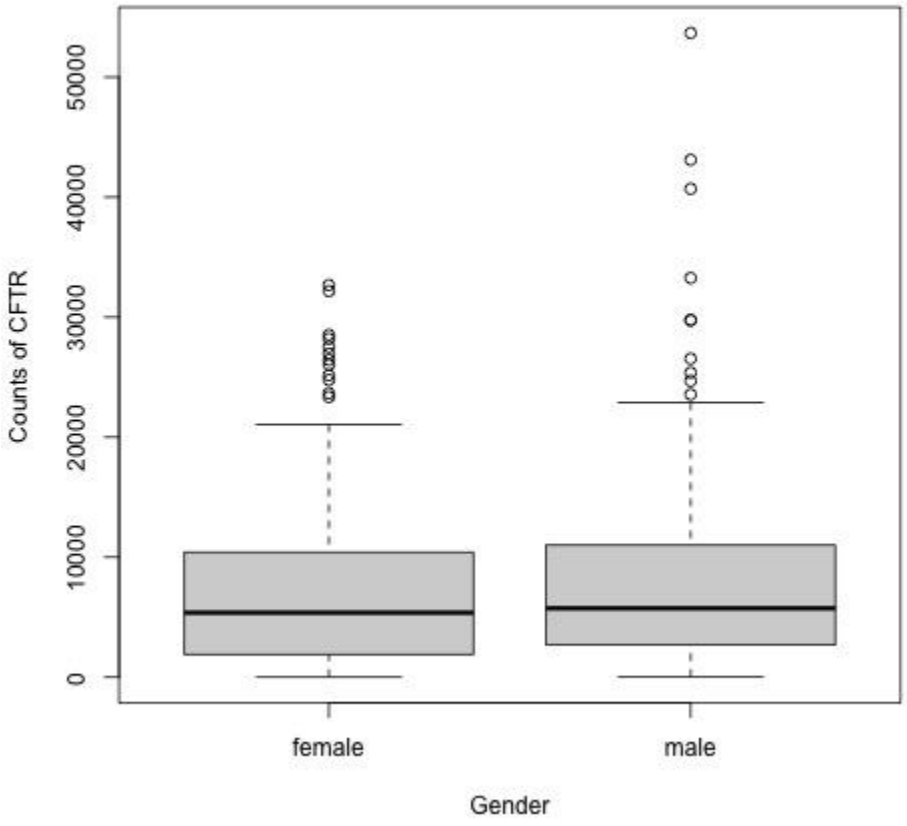


Figure 2. The median value for females was 5376, while for males it was 5742. Although males have a wider spread, the uppermost values seem to be outliers. The spread therefore of the values, excluding those outliers, is very similar between females and males. Additionally, the range of the first quartile to median is more condensed than that of median to third quartile, indicating that the data is skewed to less counts.

Based on the boxplot comparing counts of CFTR between males and females, females had a lower median than males. However, this is not necessarily indicative of a huge skew

towards the bottom in females. The difference between them is marginal enough that significance tests should be further conducted to examine this difference.

Volcano Plot of Expression in Gender

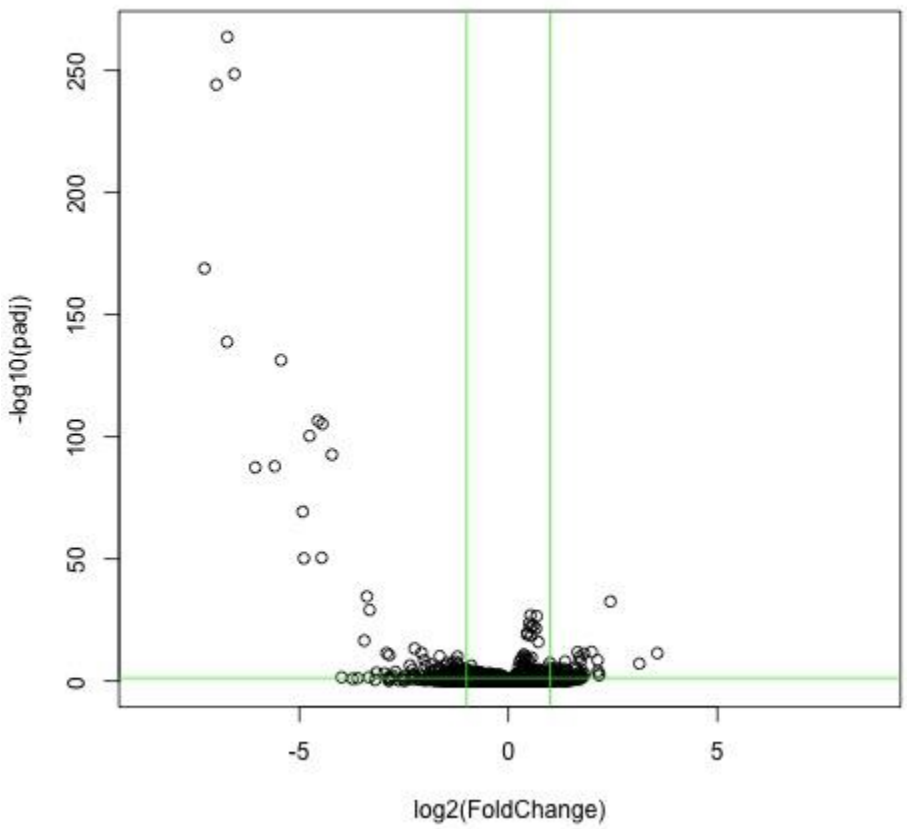


Figure 3. The left side's dots are underexpressed genes in females; alternatively, they are the overexpressed genes in males. On the right, there are the overexpressed genes in females/the underexpressed genes in males. The further they are from zero, the more change is indicated, while the higher on the y-axis it is, the more statistically significant it is according to the p-value. Most genes are in the middle (although there is a skew towards the underexpressed side), and

relatively insignificant. Of those genes that are more extreme, as shown by the dots in the upper left, there are many more relatively underexpressed genes in females than in males.

Genes overall tend to be more underexpressed within females relative to males, as indicated by the slight skew towards the left. Further evidence for this is the relative abundance of extreme dots in the upper left compared to the upper right dots.

Kaplan Meier Curve of Gender

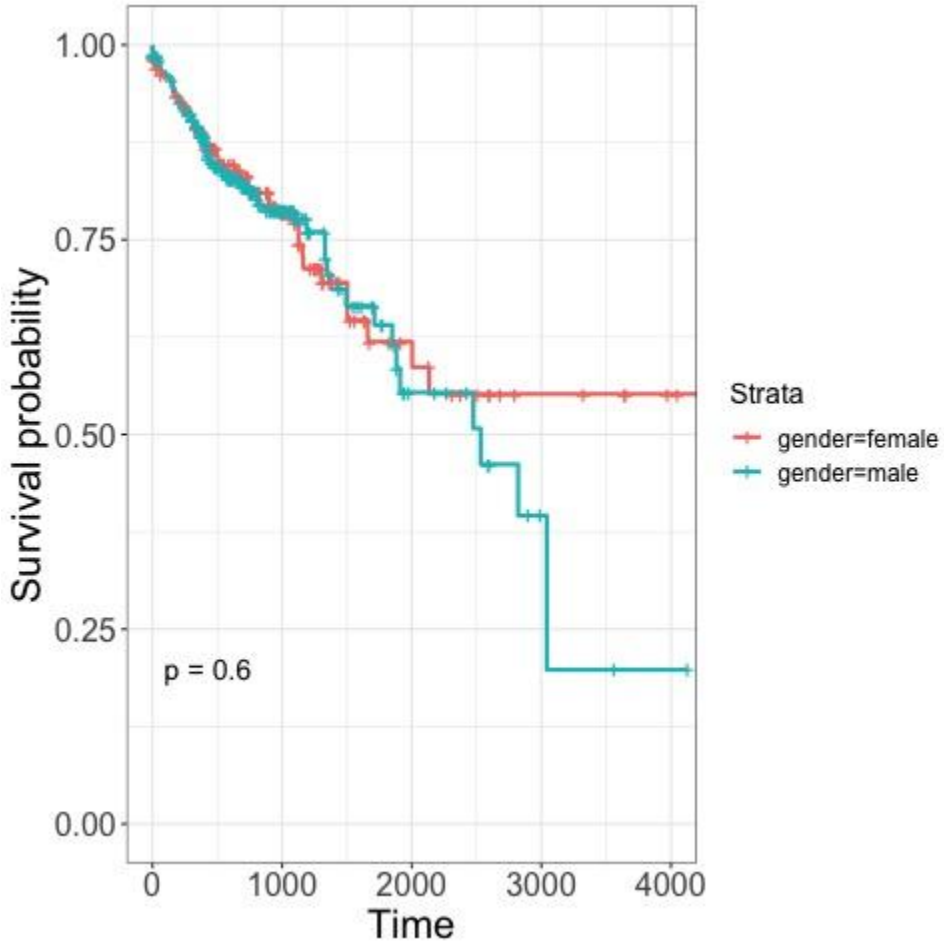


Figure 4. A survival plot between genders. In the initial years, as time progresses, survival for both females and males follow each other closely, declining at a steady rate. After the initial years, females maintain survival probability, while male survival continues to decrease.

Kaplan Meier Curve of Expression of CFTR

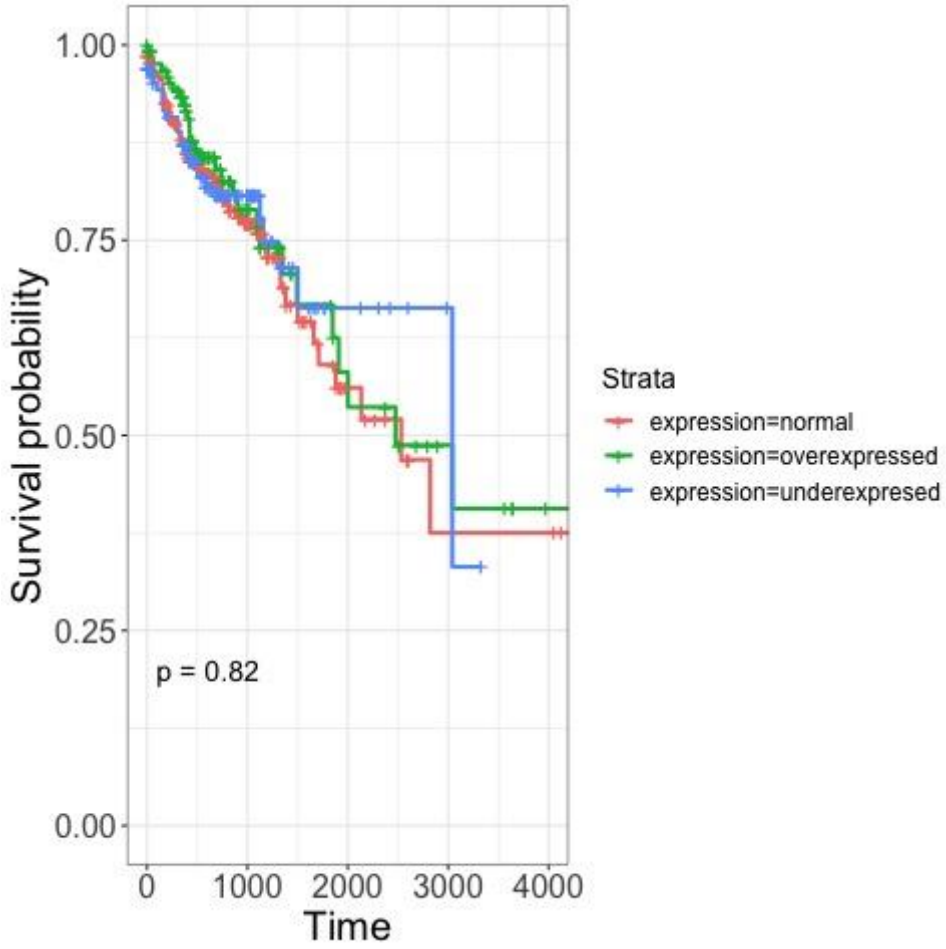


Figure 5. In the initial years, survival of all three categories follows each other at a steady decline. However, the underexpressed survival then maintains at a constant survival probability, while the other two categories continue to decline. Then, the underexpressed line drops down to meet the other two curves.

For gender, it is apparent that females' genes tend to be underexpressed compared to males. This is most evidenced by Figure 2. This is also supported by Figure 1's; females typically had a lower count of CFTR than males. Females also have a higher survival rate after the first initial years than males. As for how expression of CFTR affects survival, the data is

nebulous. Although the underexpressed survival seems to be higher, it is sandwiched between survival rates that are equal to the other two categories.

Discussion

According to the data, specifically in Figure 3, females have a higher survival rate than males. In fact, CRC occurs more frequently in males than females (Abancens et al.). The main biological reason underpinning this is estrogen. Progression of CRC tumors is dependent on pathways that exhibit sexual dimorphism. This is primarily due to the influence estrogen exerts on genes and cell signaling. One such example of sexual dimorphism in CRC is how females have a better, more robust response to inflammatory infections. This difference may have protective effects in how the immune system responds to CRC early on.

According to the data, there is not a significant correlation between differential expression of the gene CFTR and survival. However, other research has found that patients with cystic fibrosis have a dramatically increased chance of CRC (Scott et al.). Although there is yet to be definitive reasons for how CFTR acts as a tumor suppressor, there are numerous speculations for this link. Underexpression of CFTR affects numerous mechanisms. The primary potential cause is that CFTR may heavily influence intestinal crypt stem cells, which are the primary cell source of CRC.

The reason for this difference in results between the created graph and the other paper may be that all of the patients examined have been diagnosed with CRC already. It may be that cystic fibrosis increases the chances of getting CRC, but does not actually change survival rates within CRC patients.

References

Abancens, M., Bustos, V., Harvey, H., McBryan, J., & Harvey, B. J. (2020, December 9). *Sexual dimorphism in colon cancer*. Frontiers in Oncology. Retrieved March 9, 2022, from https://www.frontiersin.org/articles/10.3389/fonc.2020.607909/full

American Cancer Society. (2022, January 12). *Colorectal cancer statistics: How common is colorectal cancer?* American Cancer Society. Retrieved March 9, 2022, from https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html

Centers for Disease Control and Prevention. (2022, February 17). *What is colorectal cancer?* Centers for Disease Control and Prevention. Retrieved March 9, 2022, from https://www.cdc.gov/cancer/colorectal/basic_info/what-is-colorectal-cancer.htm

Hopkins. (2001). *Sporadic (Nonhereditary) Colorectal Cancer: Introduction*. Retrieved March 9, 2022, from https://www.hopkinsmedicine.org/gastroenterology_hepatology/_pdfs/small_large_intestine/sporadic_nonhereditary_colorectal_cancer.pdf

Scott, P., Anderson, K., Singhania, M., & Cormier, R. (2020, April 21). *Cystic fibrosis, CFTR, and colorectal cancer*. International journal of molecular sciences. Retrieved March 9, 2022, from https://pubmed.ncbi.nlm.nih.gov/32326161/

PART 2

General

1. TCGA is The Cancer Genome Atlas, and it is a cancer genomics data bank. It is the data bank that we perform analysis on in class.
2. TCGA is a relatively large bank of genetic information, coupled with clinical metadata. This metadata is important to make conclusions. On the other hand, TCGA does have holes in this data which can make it hard to form conclusions for those categories of data which have many holes. For instance, the time of death for many of the patients is missing, so we have to approximate by using their last check-up. And we also assume that people who have relapses return to the same cancer care facility, which makes it difficult to record post-cancer treatment clinical data,
3. The data that we explore is gene information. One such example is the frequency/count of genes or gene mutations that occur in cancer patients. Such genes or gene mutations that are correlated with cancer therefore fit within the central dogma of biology. These genes are transcribed into RNA and then translated into proteins. These cancer-associated proteins therefore likely have some cancer-related function, such as signaling to proceed through a cell cycle checkpoint despite DNA damage. Thus, we can examine genetic data that directly corresponds to cancer-causing proteins.

Coding

1. The command that we use to save a file to your GitHub repository is first to cd into the local repository (cd qbio_data_analysis_peterc). Then we use git status in order to check which files have local changes that can be uploaded onto GitHub. Here, we should see

that the file we want to upload onto GitHub pops up in red. Then we use git add name_of_file. Then we commit the file using git commit -m “text that describes files”. Finally, we git push it into GitHub.

2. What command must be run in order to use a package in R?
3. What is boolean indexing? What are some applications of it?
4. Draw out a dataframe of your choice. Show an example of the following and explain what each line of code does.
    a. an `ifelse()` statement
    b. boolean indexing

2. In order to use a package in R we must first download it using the install(“name_of_package”) function. Then we must run the library(name_of_package) function.
3. Boolean indexing is a technique of turning vectors into boolean vectors, also known as masks, through setting the mask equal to a logical operator of the vector. We then use these use myvector[mask] in order to create a new vector that only contains information from the original vector that matches our logical operator. Boolean indexing is important so that we can isolate specific information from long vectors. Typically we use loops; however, since R is very clunky with loops, boolean indexing is how we do this in R. Through boolean indexing, we can do things such as clean data.
4.

element_name	atomic_number	atomic_mass	electronegativity	boiling_point	melting_point
“Hydrogen”	1	1.008	2.20	-252.9	-259.1
“Helium”	2	4.0026	NA	-269	NA
“Lithium”	3	6.94	0.98	1342	180.54
“Beryllium”	4	9.0122	1.57	2470	1287
“Boron”	5	10.81	2.04	4000	2075
“Mercury	80	200.59	2.0	356.73	-38.830

a)
```
# creates a new column gas_at_room_temp in the dataframe that tells you what
# the state the element is at room temperature
# if the boiling point is below 20, then it is a gas; if 20 is between the melting and
# boiling point, then it is a liquid; if the melting point is above 20 then it is a solid
df$state_at_room_temp=ifelse(df$boiling_point<20, “gas”,
        ifelse(df$melting_point<20, “liquid”, “solid”))
```

b)
```
# create a mask for if the boiling point is less than 20 degrees C, meaning it is a
# gas at room temp
mask=df$boiling_point<20
# creates a new dataframe solely made up of the elements that are gaseous at
# room temperature
gas_df=df[mask,]
```