Leon Zha
QBIO 490
Kate, Nicole, David
March 9, 2022

Examining Relative TP53 and KRAS Gene Expressions Between Races

INTRODUCTION:

Colorectal cancer (CRC) is a disease characterized by unchecked division of cells in the colon or rectum. CRC is one of the most common and deadly types of cancer, being the third most common cancer worldwide and the fourth most common cause of death as of 2009 (Haggar and Boushey 2009). A plethora of sources agree that there are clear racial/ethnic disparities for incidence, diagnosis, and survival when it comes to CRC (Ollberding et al. 2011, Lee et al. 2021). For example, after adjusting for age, studies have found that incidence and mortality is generally highest for African Americans and lowest for Hispanics and Asians/Pacific Islanders (Ollberding et al. 2011, Lee et al. 2021).

In this assignment, I took transcriptomic data of patients afflicted by CRC from The Cancer Genome Atlas (TCGA) and compared the relative expression of two genes – TP53 and KRAS – across four racial groupings – White, Asian, Black or African American, and American Indian or Alaska Native. TP53, sometimes called the "the guardian of the genome," codes for a tumor suppressor protein, and its mutation is associated with many types of cancer (GeneCards Human Gene Database). KRAS, which encodes for proteins that are involved in cell signaling pathways that control cell growth, maturation, and apoptosis. KRAS is frequently found mutated in CRC, with about 35%-45% of cases having a mutation (Dinu, D et al. 2014). Additionally, I also compared mortality rates between the four racial groupings. I found that although the mean expression for TP53 and KRAS was about the same across the races, Asians tended to have better survival rates than the White population, who tended to have better survival rates than Black population.

METHODS:

To perform the analysis comparing TP53 and KRAS expression, I first loaded in transcriptomic data from TCGA. I extracted the rows corresponding to the genes for TP53 and KRAS, plotting them against each other to see if they have any correlation. Next, I created a boolean mask to extract data from counts where there was corresponding valid race data for that patient. Finally, I created two boxplots, one for each gene, and plotted the distribution of counts for each race.

Next, to investigate how race impacts survival, I created Kaplan-Meier plots using the clinical data of patients afflicted by CRC from TCGA. As before, the first step was to clean up the data. I created boolean masks to filter out patients for which there was NA or "No Data." After, I created a fit object, which was used as the main input for the creation of the plot itself.

RESULTS:

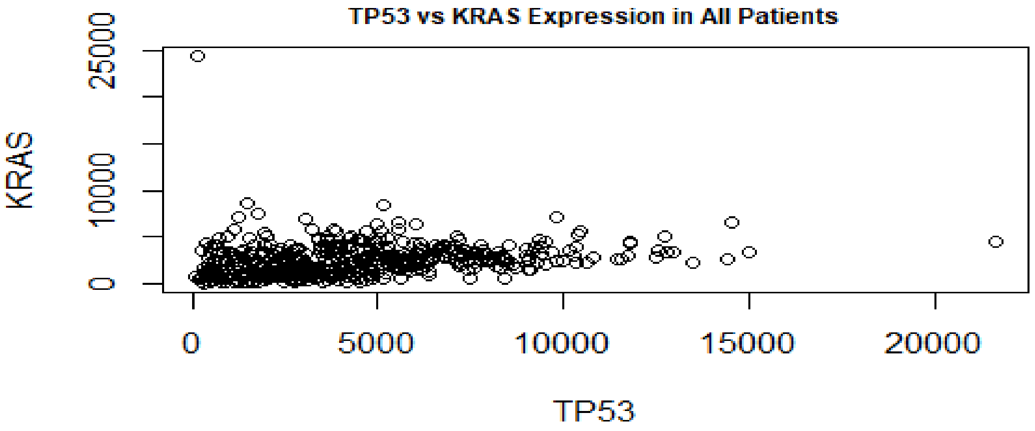

Fig. 1. Scatterplot of RNA counts for TP53 vs KRAS counts

The above scatter plot shows that KRAS and TP53 have a weak positive correlation - as TP53 counts increase, so do KRAS counts. This indicates that TP53 and KRAS may be connected in some way. Interesting to note is that there are at least two major outliers, one along each axis, which should probably be excluded should more in depth statistical analyses be done.

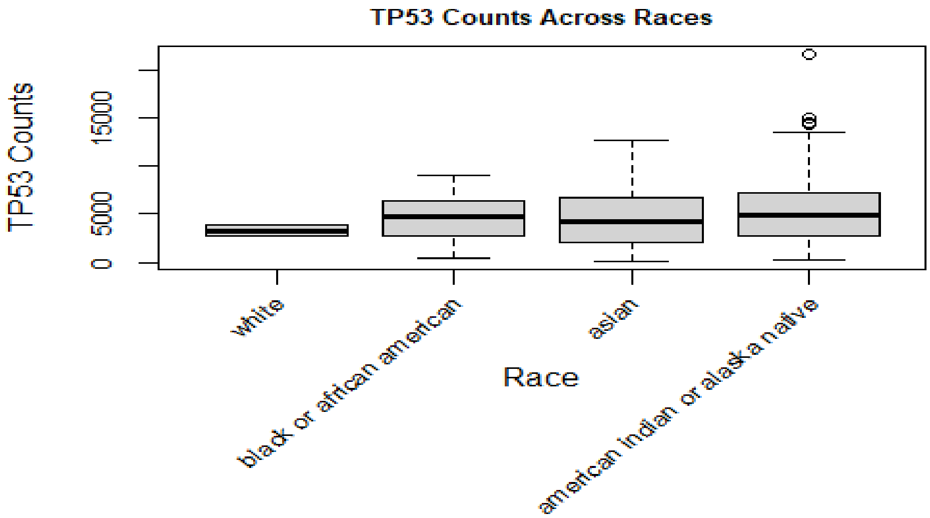

Fig 2. RNA Counts of TP53 stratified across 4 racial groupings

Next, the box plots above illustrate the counts for TP53 when stratified by race. Looking at these boxplots, it becomes clear that the White population has the lowest mean, and the American Indian or Alaska Native population the highest mean. Generally, however, the means of the populations are all roughly the same. In terms of variation, the White population has the lowest variation, while the interquartile range of the other three demographics are all roughly similar.

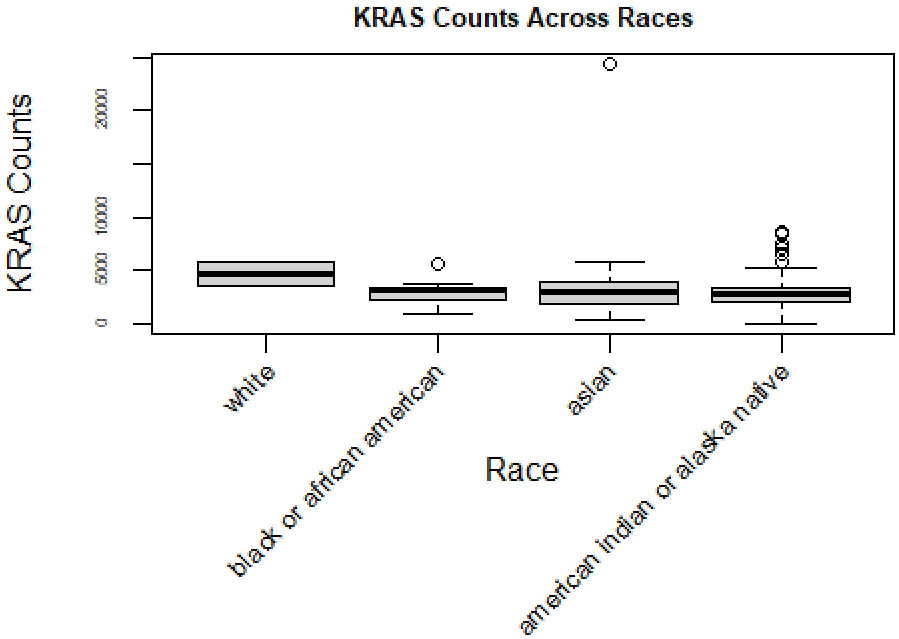

Fig. 3. RNA counts of KRAS stratified across 4 racial groupings

This second group of box plots show the KRAS counts, which have considerably less variability than the TP53 counts. Interestingly, the extremes of the mean count is reversed compared to TP53, with the highest mean count found in the White population, while the lowest is in the American Indian or Alaska Native population.

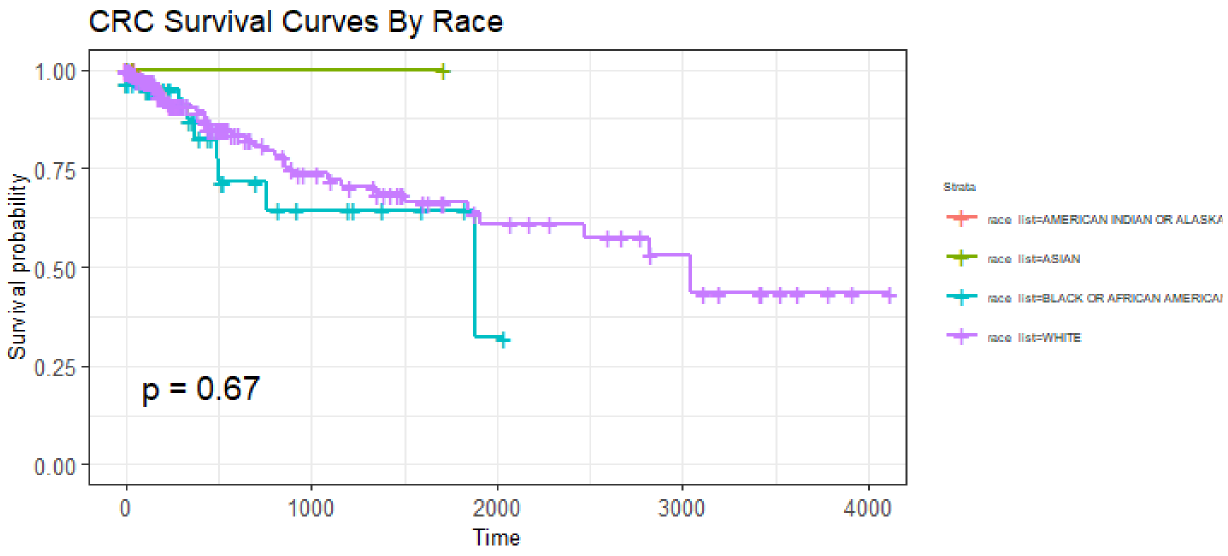

Fig. 4. Survival Probability of CRC Patients Over Time

Finally, the Kaplan-Meier plot above shows the survival probability over time for the different races. Unfortunately, since the dataset used to build this graph is slightly different from the ones used for the plots above, there was significantly less data for American Indian or Alaska Native populations, so much so that they don't even show up on the graph. However, looking at the data we do have, it becomes clear that Asians have the best survival over time, the White population has the second best survival rate over time, and the Black population the worst.

DISCUSSION:

It's unsurprising to see that there's a correlation between TP53 and KRAS counts - this is supported by existing literature (Lee, Chul Seung, et al. 2021). In regards to the boxplots, having the lowest TP53 counts for the White population is interesting because one would expect more of it in a healthy system. I was unable to locate any studies directly measuring RNA counts for either TP53 or KRAS among races for comparison, although I found evidence to suggest that certain populations, especially Asian ones, are more likely to have TP53 mutations in general

(Guttery et al. 2018). Shifting focus to the survival curves, the findings in this paper support the notion that there are severe gaps in racial disparity for CRC survival. However, to note is that further analysis should use a dataset that is both broader and more fine-grained. A broader dataset could include other prominent racial groups such as the Latinx demographic, while a more detailed dataset could break down the existing groups into smaller sections to address issues like the fact that Native Hawaiins have a particularly high CRC mortality rate, but are generally grouped in with Asians, which overall have low CRC mortality rates (Ollberding et al. 2011).

Works Cited

Dinu, D et al. “Prognostic significance of KRAS gene mutations in colorectal cancer--preliminary study.” Journal of medicine and life vol. 7,4 (2014): 581-7.

GeneCards Human Gene Database. “TP53.” *GeneCards Is a Searchable, Integrative Database That Provides Comprehensive, User-Friendly Information on All Annotated and Predicted Human Genes*, 2017, https://www.genecards.org/cgi-bin/carddisp.pl?gene=TP53.

Guttery, David S et al. “Racial differences in endometrial cancer molecular portraits in The Cancer Genome Atlas.” Oncotarget vol. 9,24 17093-17103. 30 Mar. 2018, doi:10.18632/oncotarget.24907

Haggar, Fatima A, and Robin P Boushey. “Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors.” Clinics in colon and rectal surgery vol. 22,4 (2009): 191-7. doi:10.1055/s-0029-1242458

Lee, Chul Seung, et al. “Enhancing the Landscape of Colorectal Cancer Using Targeted Deep Sequencing.” *Scientific Reports*, vol. 11, no. 1, 2021, https://doi.org/10.1038/s41598-021-87486-3.

Lee, Seohyuk, et al. “Race, Income, and Survival in Stage III Colon Cancer: Calgb 89803 (Alliance).” *JNCI Cancer Spectrum*, vol. 5, no. 3, 2021, https://doi.org/10.1093/jncics/pkab034.

Ollberding, Nicholas J et al. “Racial/ethnic differences in colorectal cancer risk: the multiethnic cohort study.” International journal of cancer vol. 129,8 (2011): 1899-906. doi:10.1002/ijc.25822

General Concepts

1. What is TCGA and why is it important?
    a. TCGA is a landmark cancer genomics program that molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. All the data is publicly available, which has led to improvements in our ability to diagnose, treat, and prevent cancer
2. What are some strengths and weaknesses of TCGA?
    a. A strength of TCGA is that it's free to access for all, providing easy access to those that might not have the resources to get samples from patients directly, leading to more experiments and analyses. Additionally, findings are easily replicable because everyone uses the same data. A weakness of TCGA is that any potential error in the dataset is propagated through a multitude of studies because everyone shares the data. Another weakness is that TCGA was very labor intensive to get started, factoring in many issues such as privacy.
3. How does the central dogma of biology (DNA → RNA → protein) relate to the data we are exploring?
    a. The central dogma explains the flow of genetic information, from DNA to RNA to proteins. This relates to the data we explore because we have access to genomic, transcriptomic, and proteomic data through TCGA, so we can see how changes in DNA and RNA propagate down into symptoms, deadliness, effectiveness of specific cures, etc.

Coding Skills

1. What commands are used to save a file to your GitHub repository?
    a. git add [FILENAME].[FILEEXTENSION]
    b. git commit -m"COMMIT MESSAGE"
    c. git push
2. What command must be run in order to use a package in R?
    a. install::("PACKAGENAME")
    b. library("PACKAGENAME")
3. What is boolean indexing? What are some applications of it?
    a. Boolean indexing is a type of indexing which uses actual values of the data in the DataFrame. The main application is speed - it's faster than looping when used for certain data structures, particularly in R.
4. Draw out a dataframe of your choice. Show an example of the following and explain what each line of code does.
    a. Consider the dataframe "students" below

ID	firstName	lastName	Sex	Year	GPA
1	Notdavid	Wen	M	Senior	3.9
2	Notnicole	Black	F	Sophomore	3.9
3	Notleon	Zha	M	Junior	1.9

    b. an `ifelse()` statement
        i. Is_failing_mask <- ifelse(students$GPA < 2.0, TRUE, FALSE)

ii. // creates and fills Is_failing_mask, which is a vector filled with either TRUE if the student is failing or FALSE if not, based on the student's GPA. In this case [FALSE, FALSE, TRUE]

c. boolean indexing

i. Student_id_failing <- students$ID[Is_failing_mask]

ii. // creates and fills Student_id_failing, a vector filled with the student IDs of those that are failing. In this case [3]