Pranav Pammidimukkala

Midterm Research Project

10 March, 2022

Introduction

Colorectal cancer is a cancer in which the cells in the colon or rectum experience uncontrolled cell growth. Colon and rectal cancers used to be classified differently, but they are now both called colorectal cancer due to the similarity in disease progression. The colon is another name for the large intestine and the rectum is the pathway that connects the colon to the anus. This is the third most common form of cancer, with a majority of cases in Western countries (Mármol et al.). Known risk factors for colorectal cancer include age, diet, and lifestyle. Like other cancers, colorectal cancer is caused by mutations in genes such as oncogenes, tumor suppressor genes, and other genes that have functions related to DNA replication and cell growth (Màrmol et al.).

Genetic and environmental risk factors both have significant involvement in the diagnosis of this cancer. About 75% of patients with the disease do not have the disease in their family history. Those who do have a genetic history have almost double the risk of getting the disease (Kuipers et al.). They are also more likely to be affected by Lynch syndrome or familial adenomatous polyposis. Lynch syndrome involves mutations in DNA mismatch-repair genes, such as *MLH1*, *MSH2*, and *EPCAMI* (Kuipers et al.). This class of genes encodes for proteins that are responsible for correcting errors in the genome during the DNA replication process. Therefore, a mutation in any of these genes increases the risk for mutations in other parts of the genome. Familial adenomatous polyposis involves mutations in the adenomatous polyposis coli (*APC*) gene (Kuipers et al.). This gene controls the activity of the Wnt signaling pathway, a

group of signal transduction pathways that is involved in cell proliferation and differentiation during development. There are many potential risk factors for the disease, as well. This study will analyze height as a risk factor for the disease. Height was compared with patient survivability and with the number of present mutations in the *IGF1* gene. The data for this experiment was sourced from TCGAbiolinks, and various analyses were run on certain parts of that data to study the stated scenarios. The initial hypothesis was that shorter patients would have a higher survivability than taller patients, and taller patients would have more mutations in all genes. It was determined that patients categorized as short had a higher survivability in both sexes, and also had more mutations in the *IGF1* gene.

Methods

All of the data for this experiment was taken from the TCGA Biolinks library. The data used for the survival analysis was drawn from the clinic dataset, which was split into two different datasets based on the sex of the patient. Each sex dataset was then divided into short and tall categories. This was done by finding the average height for each sex, and categorized everyone greater than or equal to the height as tall. Everyone shorter than the average was categorized as short. The height and survival status of the patients were included in this dataset, and a Kaplan-Meier analysis was done on these two variables using functions from the survival and survminer libraries. The mutation data was sourced from the maf object and an oncoplot and lollipop plot were generated to analyze the relationship between height and the number of mutations.

Results

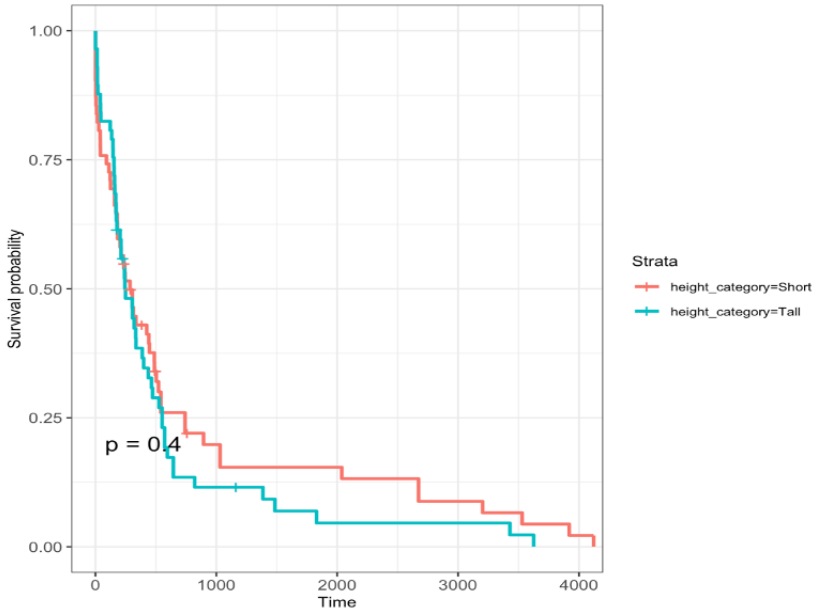

Figure 1A. Survival Analysis between short and tall female patients

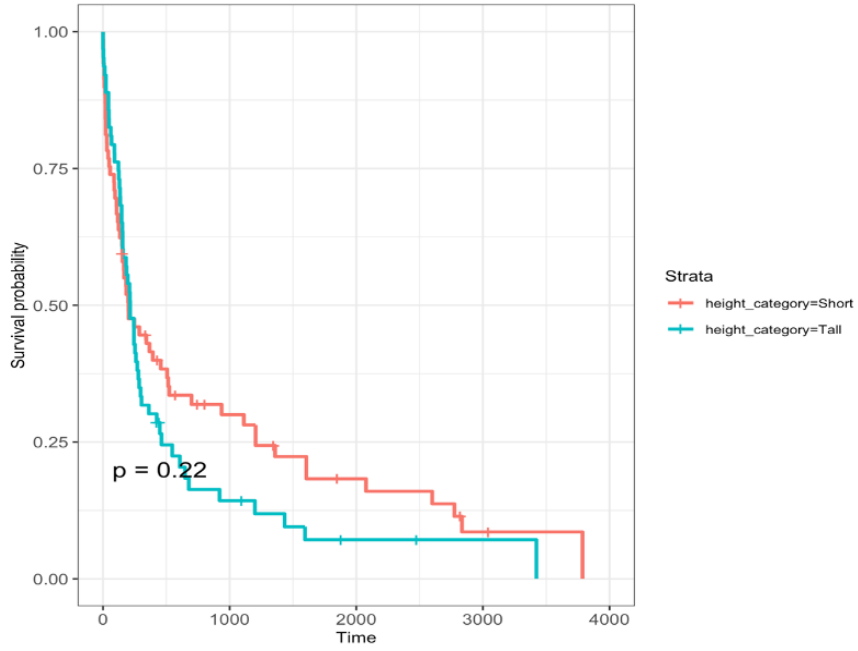

Figure 1B. Survival Analysis between short and tall male patients

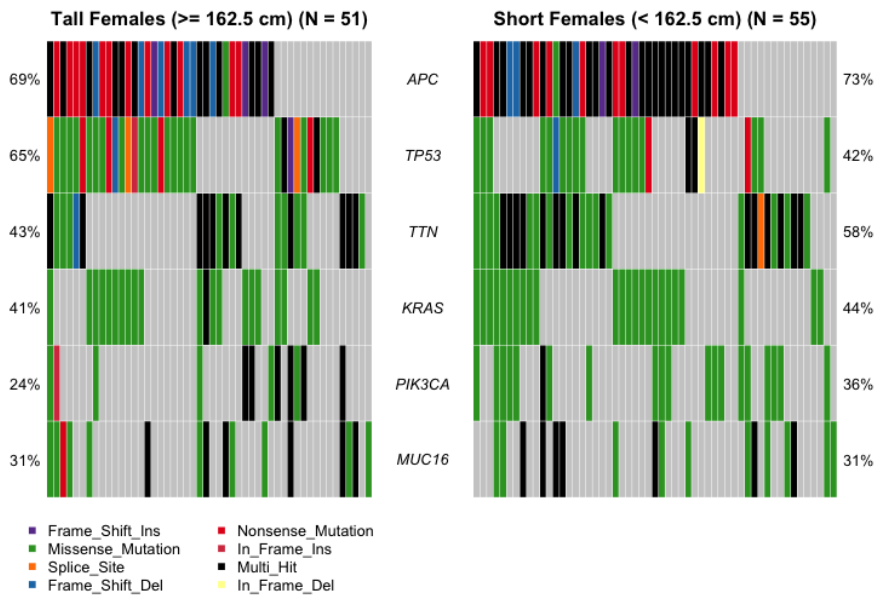

Figure 2A. Comparison of mutations in listed genes between tall and short female patients

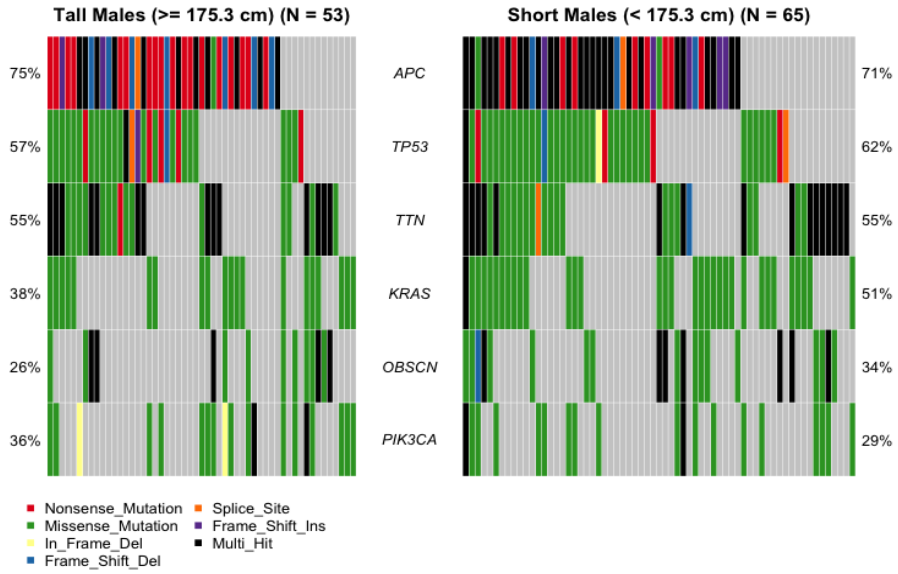

Figure 2B. Comparison of mutations in listed genes between tall and short male patients

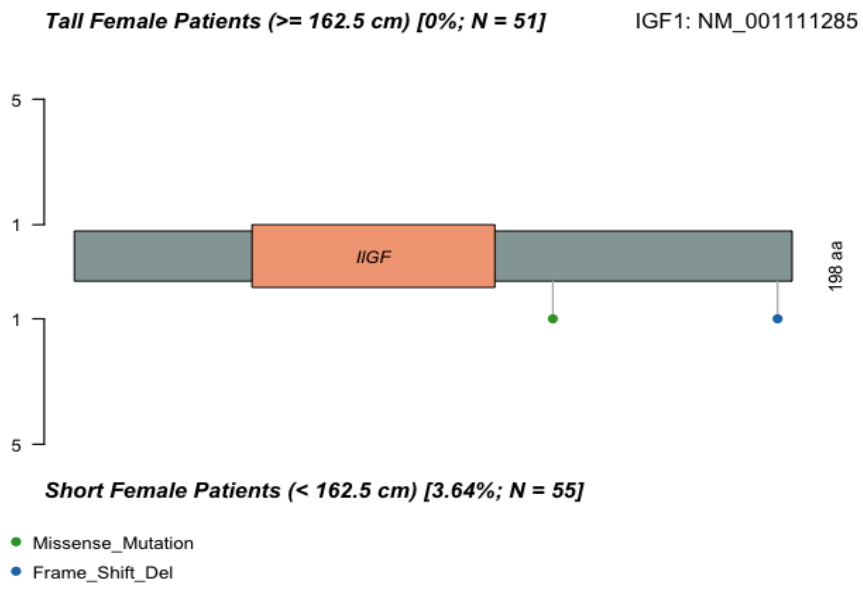

Figure 3A. Comparison of number of mutations of *IGF1* gene in tall and short female patients

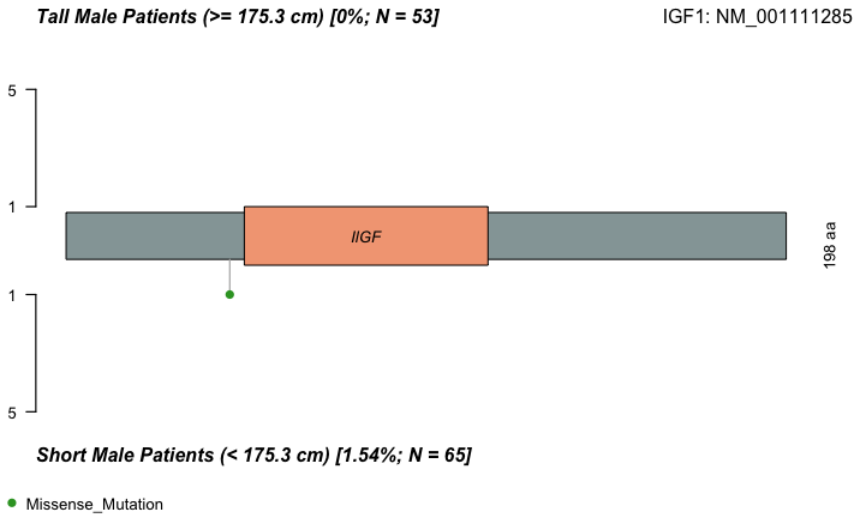

Figure 3B. Comparison of number of mutations in *IGF1* gene in tall and short male patients

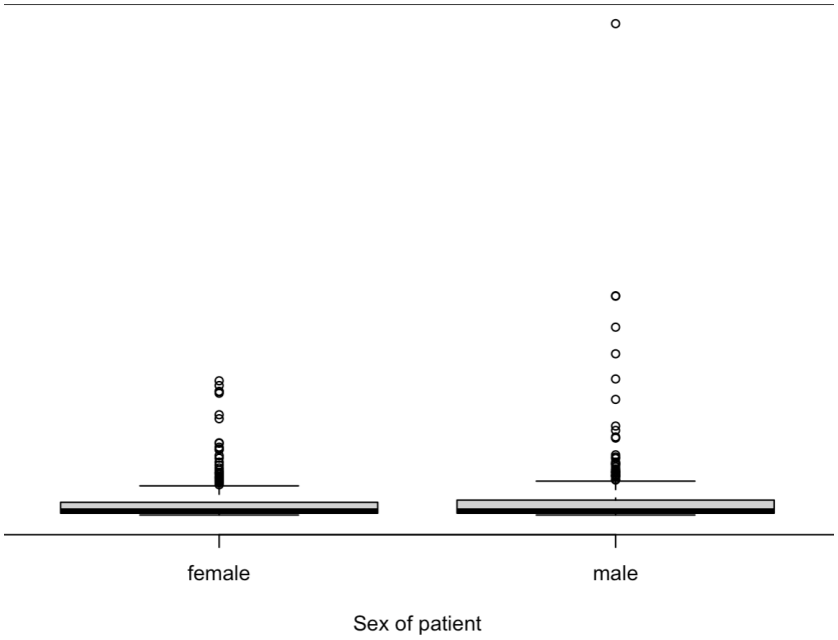

Figure 4. Comparison of expression of *IGF1* gene between male and female patients

An analysis of the height and survival data revealed that the patients categorized as short in the data sample had a higher survival probability in both sex groups. A visual representation of both these trends can be seen in Figures 1A and 1B. However, there was no clear trend in the amount of mutations in the 6 genes listed in Figures 2A and 2B. Some genes were expressed more in short patients, some were expressed more in tall patients, and some were expressed equally. There were a greater number of genes with more mutations in shorter patients than expected, with different genes varying in the category that was expressed more between sex groups.

Figures 3A and 3b revealed that shorter patients had more mutations in the *IGF1* than in taller patients. It was also revealed that the male patients had a slightly higher expression of the *IGF1* gene than in female patients, according to Figure 4.

Discussion

The hypothesis of the experiment was that shorter patients would have a higher survival probability than taller patients, and that taller patients would have a higher number of mutations in all the studied genes. This hypothesis was partially supported by the results of this experiment. According to Figures 1A and 1B, patients categorized as short in both sex groups had a higher survival probability as time increased. One reason for this trend is that taller people tend to have larger colons. This increased amount of cells increases the chance of some of those cells turning into cancer cells (Khankari et al.). Another reason could be the increased amount of growth factors that taller people are exposed to. These growth factors increase the risk of mutation, which could result in malignant cells (Abar et al.). The increased survival probability in shorter patients is further supported by results of studies by Trift et al. and Khankari et al.

It was initially hypothesized that taller people would have a higher expression of the insulin-like growth factor 1, or *IGF1*. This growth factor is known to have carcinogenic effects, such as increasing cell proliferation and inhibiting control processes such as apoptosis (Weroha et al.). The relationship between the number of mutations in *IGF1* and the height of the patient was explored in Figures 3A and 3B. In both sexes, shorter patients had more mutations in *IGF1*, with shorter male patients having one more mutation than taller male patients and shorter female patients having two more mutations than taller female patients. These results were different than initially hypothesized, but could indicate that the mutations in shorter patients result in a less active *IGF1* gene. This could explain their shorter height. A similar study found that the *IGF1* gene increases the risk of colorectal cancer (Murphy et al.). According to Figure 4, male patients had a higher expression of the *IGF1* gene than female patients. Male patients also had a larger difference in survival rate between height categories than female patients. These results could

indicate that taller patients have a higher expression of *IGF1*, which would result in a lower survival probability due to an increased risk of colorectal cancer.

A source of error in this study is the standard used to measure height. Tall was defined as patients above or equal to the average value, while short was defined as below the average value. If the clinical definitions for these categories differ, it could result in different trends in survival and mutations. A future experiment could explore the effect of an increased expression of the *IGF1* gene and the survival probability of a patient. Another future experiment could study the effects of the expression of other growth factor genes such as EGF on survivability and the number of present mutations in a patient.

Works Cited

Abar, L., Vieira, A. R., Aune, D., Sobiecki, J. G., Vingeliene, S., Polemiti, E., Stevens, C., Greenwood, D. C., Chan, D., Schlesinger, S., & Norat, T. (2018). Height and body fatness and colorectal cancer risk: an update of the WCRF-AICR systematic review of published prospective studies. *European journal of nutrition*, *57*(5), 1701–1720. https://doi.org/10.1007/s00394-017-1557-1

Khankari, N. K., Shu, X. O., Wen, W., Kraft, P., Lindström, S., Peters, U., Schildkraut, J., Schumacher, F., Bofetta, P., Risch, A., Bickeböller, H., Amos, C. I., Easton, D., Eeles, R. A., Gruber, S. B., Haiman, C. A., Hunter, D. J., Chanock, S. J., Pierce, B. L., Zheng, W., … Transdisciplinary Research in Cancer of the Lung (TRICL) (2016). Association between Adult Height and Risk of Colorectal, Lung, and Prostate Cancer: Results from Meta-analyses of Prospective Studies and Mendelian Randomization Analyses. *PLoS medicine*, *13*(9), e1002118. https://doi.org/10.1371/journal.pmed.1002118

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., van de Velde, C. J., & Watanabe, T. (2015). Colorectal cancer. *Nature reviews. Disease primers*, *1*, 15065. https://doi.org/10.1038/nrdp.2015.65

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., & Rodriguez Yoldi, M. (2017). Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *International Journal of Molecular Sciences*, *18*(1), 197. https://doi.org/10.3390/ijms18010197

Murphy, N., Carreras-Torres, R., Song, M., Chan, A. T., Martin, R. M., Papadimitriou, N., Dimou, N., Tsilidis, K. K., Banbury, B., Bradbury, K. E., Besevic, J., Rinaldi, S., Riboli, E., Cross, A. J., Travis, R. C., Agnoli, C., Albanes, D., Berndt, S. I., Bézieau, S., …

Gunter, M. J. (2020). Circulating levels of insulin-like growth factor 1 and insulin-like growth factor binding protein 3 associate with risk of colorectal cancer based on serologic and Mendelian randomization analyses. *Gastroenterology*, *158*(5). https://doi.org/10.1053/j.gastro.2019.12.020

Weroha SJ, Haluska P. The insulin-like growth factor system in cancer. *Endocrinol Metab Clin North Am*. 2012;41(2):335-vi. doi:10.1016/j.ecl.2012.04.014

General Concepts

1. What is TCGA and why is it important?
    a. The Cancer Genome Atlas is a program that collected the data and analysis of a large amount of cancer patient data. Most of this data is available to the public in a way that maintains the confidentiality and privacy of the patient. This is important because it solves one of the biggest difficulties with studies - a lack of a large data sample. People all over the world can use TCGA to study various aspects of different cancers, making the process much more efficient and cheaper.
2. What are some strengths and weaknesses of TCGA?
    a. Some strengths of TCGA are that it is publicly available and maintains the privacy of the patients. This allows people to study the data with minimal resources. A big issue with statistical studies is maintaining the privacy of the participants. Different variables can be used to identify an individual, and TCGA takes this hassle out of the equation by only providing public data that is ensured to respect the patients. A weakness of TCGA is that it can be difficult to access the data without the proper training. A decent grasp of a computer language, along with the ways to call and analyze different aspects of the dataset is required in order to meaningfully study the data. As a result, there is some friction that people have to overcome before accessing the data.
3. How does the central dogma of biology (DNA → RNA → protein) relate to the data we are exploring?
    a. The central dogma of biology is fundamental to the data we are exploring. Much of the data, such as mutation analysis, sequence analysis, and expression are all various stages of the central dogma. Errors in various steps of the central dogma, such as transcription or translation, are fundamental to cancer. The different data that we are exploring is a quantification of these errors or processes.

Coding Skills

1. What commands are used to save a file to your GitHub repository?
    a. “git status” identifies files that are not saved to GitHub. “git add file” allows for the selection files to be uploaded to github. “git commit -m “ “ allows for a commit message to be uploaded with the file. “git push” pushes the files to GitHub, saving them to the repository.
2. What command must be run in order to use a package in R?
    a. “install.packages()” installs the package and “library()” allows R to access the contents of that package.
3. What is boolean indexing? What are some applications of it?
    a. Boolean indexing is creating a vector of True False values that results from applying a conditional statement to part of a dataframe. This is a very useful technique, as it allows for an efficient method of extracting data from a dataframe. For example, boolean indexing can be used to remove NA values from a dataframe. It can be used to divide data based on a condition such as age.
4. Draw out a dataframe of your choice. Show an example of the following and explain what each line of code does.
    a. install.packages("palmerpenguins")
    library(palmerpenguins)
    b. an `ifelse()` statement
        i. penguins$height_category <- ifelse(penguins$height > 50, “tall”, “short”)
        ii. Creates a new category in the dataframe called height_category that says tall if the penguins height value is greater than or equal to 50, and short if it is less.
    c. boolean indexing
        i. penguins$height
        is.na(penguins$height)
        boolean_mask <- is.na(penguins$height)
        penguins_new <- penguins[boolean_mask = ‘FALSE’,]