QBIO 490 Mid Semester Project

Research Question: What populations (young or old) tend to over/underexpress the EGFL6 gene? How does over/underexpression of EGFL6 affect survival?

Introduction

Colorectal cancer (CRC) is a type of cancer that occurs in the colon and rectum, which are parts of the large intestine. It is the third most common cancer type, responsible for the fourth highest number of cancer related deaths (Mármol et al., 2017). Given the prevalence and potency of this disease, research on CRC is of considerable importance. This project, in line with the QBIO 490 course, takes a genomic approach to CRC research. Through preliminary research on CRC, it was found that the EGFL6 gene was universally upregulated in the incorporated sample (n = 2233) which was expressed in extremely low frequency in normal colorectal cells (Song et al., 2019). With this knowledge, this project looked into firstly, if there was a difference in the level of expression of EGFL6 based on age, and secondly, how levels of expression affect survival. The methodology of this project, which will be described in detail in the Method section, was largely carried out on R.

Method

The data used in this project was obtained from The Cancer Genome Atlas program. The relevant data – TCGA–COAD – was loaded onto a script. As mentioned in the Introduction section, there were two main parts in this project. Since the first part involved a comparison of EGFL6 expression levels based on age, patients were categorized whether they were over or under the age of 50, after which boxplots were generated comparing the frequency of EGFL6 expression between “Young” (patients younger than 50) and “Old” (patients older than 50) patients. The second part regarded survival rates, a Kaplan-Meier plot was created comparing the survival rates of "Young" and "Old" patients.

Results

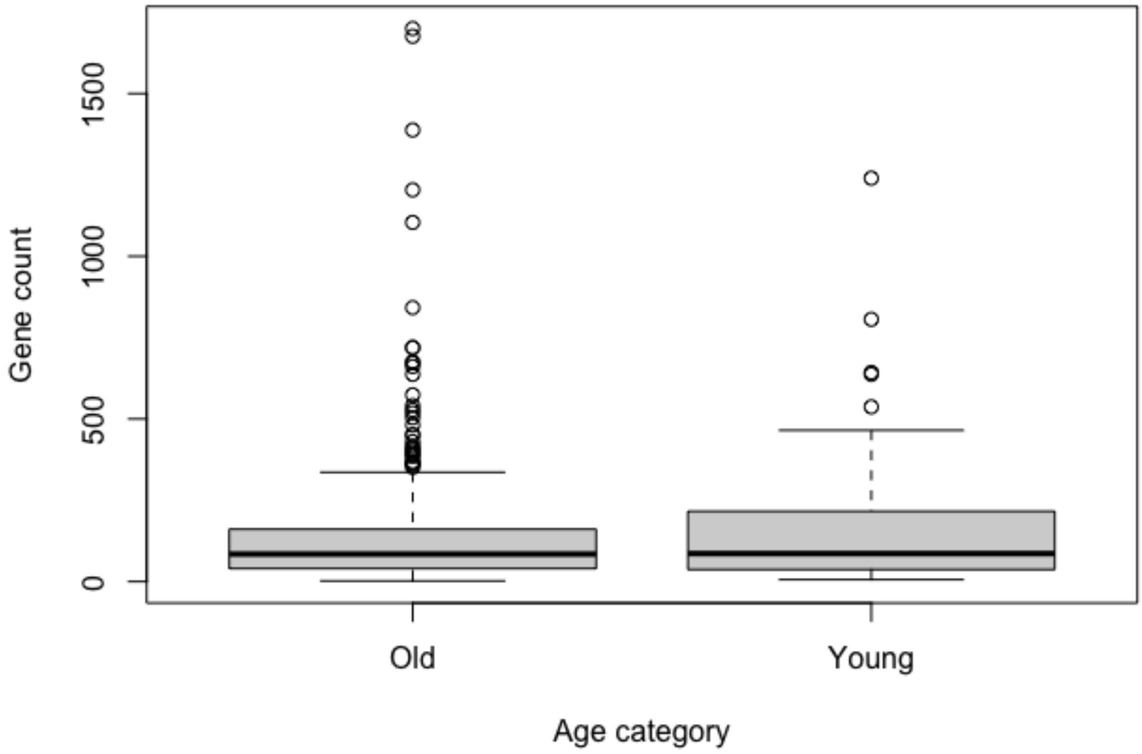

Figure 1: Boxplot of EGFL6 gene expression (gene count) in "Young" (patients under the age of 50) and "Old" (patients 50 years old or older) patients

As can be seen in Figure 1, the minimum, upper quartile, and maximum gene counts of "Young" patients are higher than those of the "Old" patients. However, the difference is very small and most importantly, there is a significant overlap between the two boxes (i.e. the whiskers of one box do not begin or end above or below the other). Further observations can be made. Since the minimum/maximum gene counts and the length of the box of "Young" patients are both greater than those of “Old” patients, there is a greater variance of EGFL6 expression in “Young” patients which spans a range slightly above that of “Old” patients. This signifies that although there is no significant difference in the expression of EGFL6 between “Young” and “Old” patients, there is a small likelihood that any given “Young” patient has a higher EGFL6 expression, but not the opposite case – in other words, there will be no “Young” patient that has a lower EGFL6 expression than any “Old” patient.

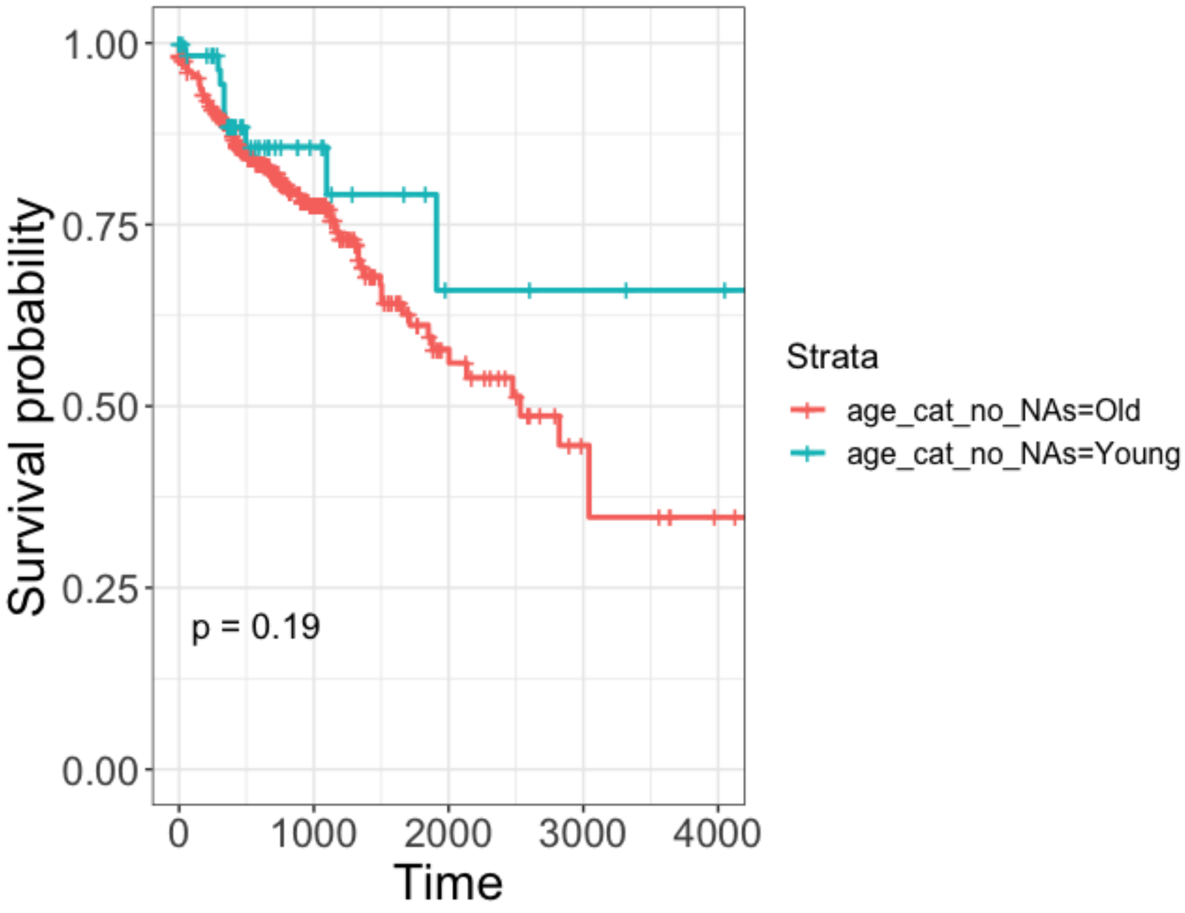

Figure 2: Kaplan-Meier plot of the survival probability of “Young” (patients under the age of 50) and “Old” (patients 50 years old or older) patients

Figure 2 shows that at any given time, “Young” patients have a greater probability of survival than “Old” patients. It can also be seen that while the survival probability of all patients decrease over time, the survival probability of “Young” patients plateaus at approximately 0.65, while “Old” patients do at approximately 0.35, almost a two-fold smaller survival probability.

Discussion

To discuss the findings of the first part of the project, it was surprising to see a minimal difference in the expression levels of EGFL6 gene between “Young" and “Old” patients, as it was hypothesized that as older individuals are likely to abnormally over/underexpress genes due to diminished cellular function (Pedro de Magalhães et al., 2009). However, upon further research, it was found that the findings of the project were consistent with literature, which stated that “... expression of EGFL6 ... was not correlated with patient’s age (Cao et al., 2018).”

The second part of the research question – the effect of EGFL6 expression on survival – was created with the assumption that aging affects the expression of EGFL6. Since Figure 2 was generated based on the “Young" and “Old” age groups, which were shown later not having a significant difference in EGFL6 expression, the results of this project cannot be utilized to answer this question. The fact that survival probabilities of either age group decrease over time is likely due to other factors. However, it cannot be concluded that the EGFL6 gene does not play a role in decreasing survival probabilities, as it has been shown that its overexpression in colorectal cells of CRC patients “affects the proliferation of colorectal cancer cells, regulates cell cycle, and inhibits apoptosis (Kang et al., 2020).” An improvement that can be made to this project is to study the survival rates of CRC patients directly based on stratified groups of EGFL6 gene counts, rather than comparing them to clinical factors.

References

Cao, Y. Q., Li, Z., Wang, L. F., Li, N., & Chang, H. (2018). High EGFL6 expression is associated with clinicopathological characteristics in colorectal cancer. *International journal of clinical and experimental pathology*, *11*(12), 5893–5900.

Kang, J., Wang, J., Tian, J., Shi, R., Jia, H., & Wang, Y. (2020). The emerging role of EGFL6 in angiogenesis and tumor progression. *International journal of medical sciences*, *17*(10), 1320–1326. https://doi.org/10.7150/ijms.45129

Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., Rodriguez Yoldi, MJ. (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

Pedro de Magalhães, J., Curado, J., & Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, *25*(7), 875–881, https://doi.org/10.1093/bioinformatics/btp073

Song, K., Su, W., Liu, Y., Zhang, J., Liang, Q., Li, N., Guan, Q., He, J., Bai, X., Zhao, W., & Guo, Z. (2019). Identification of genes with universally upregulated or downregulated expressions in colorectal cancer. Journal of gastroenterology and hepatology, 34(5), 880–889. https://doi.org/10.1111/jgh.14529

Review Questions

General Concepts

1. What is TCGA and why is it important?

The Cancer Genome Atlas program is an initiative under the National Cancer Institute, with an aim of freely providing a vast multi omic database of cancer patients. The potential in using such data for cancer research, treatment, and prevention is the importance of the TCGA.

2. What are some strengths and weaknesses of TCGA?

One strength of the TCGA is the fact that it is publicly and freely available. Easy access to a broad range of data can catalyze cancer research. A weakness of TCGA is the lack of data for certain cancers.

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

The data used in this project is the frequency of RNA expression. This is useful data as according to the central dogma of biology, RNA lies in the middle of the protein production process: because it is a product of transcription of the DNA, and because proteins are made through the translation of RNA, RNA expression data can be used to deduce the frequency of gene expression, as well as its visible effects on the human body (as facilitated by proteins).

Coding Skills

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

git add

git commit -m “”

git push

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

library()

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

Boolean indexing involves sorting data to TRUEs and FALSEs. One useful application of boolean indexing is extracting data that are within a certain parameter. For example, the frequently done Young/Old differentiation based on whether a patient was older than 50 used boolean indexing, denominating, either Young or Old as TRUEs and FALSEs.

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

colData(sum_exp)

patient_id	year_of_diagnosis	vital_status	gender
TCGA-D5-6530-01A-11R-1723-07	2010	Alive	male
TCGA-G4-6320-01A-11R-1723-07	NA	Alive	male
TCGA-AD-6888-01A-11R-1928-07	2010	Dead	male
TCGA-CK-6747-01A-11R-1839-07	2008	Alive	female
TCGA-AA-3975-01A-01R-1022-07	NA	Alive	male

TCGA-A6-6780-01A-11R-1839-07	2011	Alive	male

a. an `ifelse()` statement

ifelse(test, TRUE, FALSE); a parameter, which is a true or false question, converts datasets in chosen vector to a label of choice.

Input:

```
ifelse(colData(sum_exp)$vital_status == Dead, 1, 0)
```

Output:

```
0  0  1  0  0  0
```

b. boolean indexing

Converts dataset into TRUEs and FALSEs, then TRUE or FALSE are removed/kept.

Input:

```
year_no_NA =
colData(sum_exp)$year_of_diagnosis[!is.na(colData(sum_exp)$year_of_diagnosis)]
```

Output:

```
2010  2010  2008  2011
```