

## **Evaluating the Relation Between *FUT2* Expression and Carcinoembryonic Antigen (CEA) Levels and Survivorship in Colorectal Cancer Patients**

### Introduction

Colorectal cancer (CRC) is one of the most common forms of cancer, and one of the most common causes of morbidity (Kolligs, 2016). Being able to detect CRC early and determine a prognosis is one of the most helpful things doctors and scientists can do to improve patient outcomes and quality of life. As we learn more and more about cancer and even CRC specifically, the role that -omics data play in cancer development is becoming more apparent. Using data analysis, researchers can determine how gene expression or mutation can increase or decrease an individual's likelihood for developing cancer (Bodmer, 2006). One indicator for determining CRC patient prognosis that has been looked into is preoperative Carcinoembryonic Antigen (CEA) levels, which when high can indicate lower likelihood of survival (Becerra et al., 2016). In some studies, CEA levels have been shown to be affected by mutations in the *FUT2* gene on chromosome 19 (Liang et al., 2014). Therefore, in this analysis, the connection between *FUT2* expression, CEA levels, and survivorship will be explored.

Based on the previous knowledge collected about these topics, my hypothesis is that decreased *FUT2* expression causes increased CEA levels, which will negatively impact patient outcomes and survivability.

Data from TCGA was used to perform data analysis. In order to determine if there is a correlation between *FUT2* expression and CEA levels, a scatter plot of patients' CEA levels versus their *FUT2* expression was created as well as a boxplot comparing the *FUT2* expression of patients with normal CEA levels and patients with high CEA levels. In order to assess the

diversity of the data sets, a histogram was made for both *FUT2* expression and CEA levels. Lastly, to determine if there is a correlation between preoperative/pre-treatment CEA levels and survivorship, a Kaplan-Meier plot was created to compare the survivorship of patients with normal CEA levels versus those with high levels.

## Methods

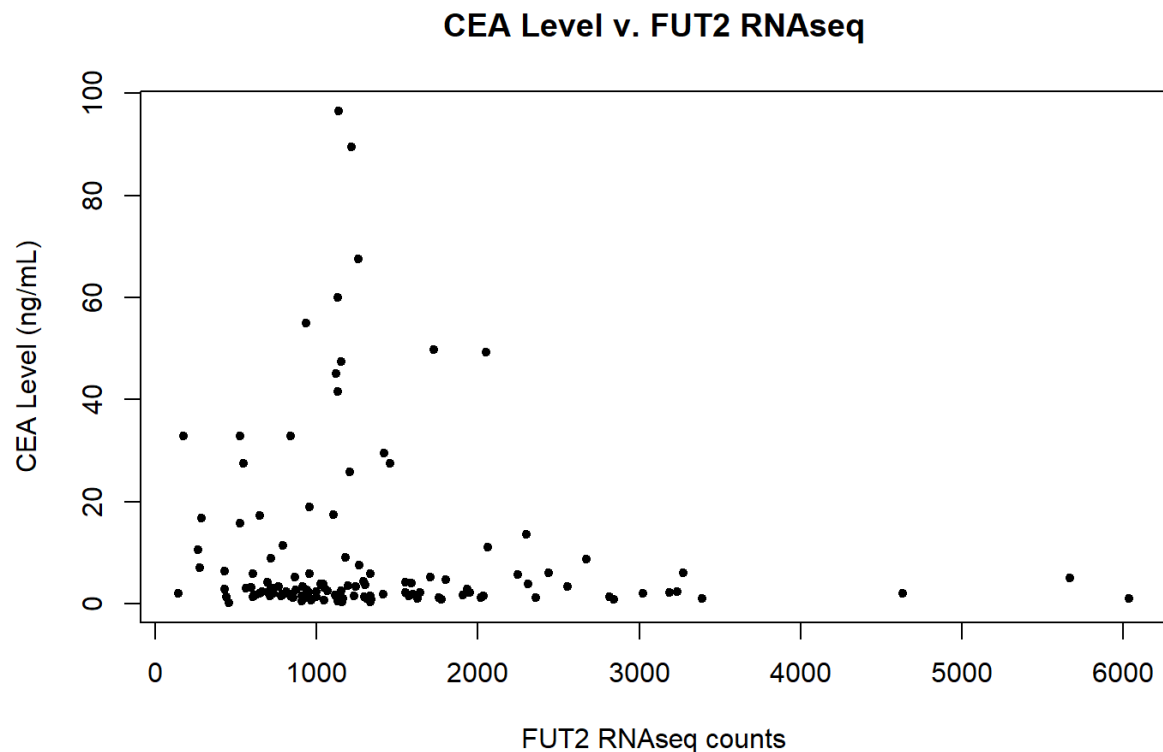
Using the TCGAAbiolinks R package, clinical and transcriptomic data from patients with colorectal cancer (“COAD”) was downloaded from TCGA. In order to investigate CEA level and *FUT2* data, they had to be prepared for analysis. First, the clinical data was placed into dataframe `clinicalData`, and all rows that contained an NA value in the CEA level column were deleted from the entire dataframe. Next, a new column was added to the `clinicalData` dataframe to categorize each patient as either having a normal CEA level ( $\leq 3\text{ng/mL}$ ) or a high ( $> 3\text{ng/mL}$ ) CEA level. To look into the *FUT2* data, first the Ensembl ID of *FUT2* was found via boolean indexing, then that ID was then used in order to create vector `FUT2data` with all the RNAseq counts for each of the patients.

To create a scatter plot, the `FUT2data` vector and the column of `clinicalData` with the CEA levels were passed to the plot function, with both of them masked to remove data points with a CEA level over  $250\text{ng/mL}$  (to get rid of extreme outliers). To create the box plot, the `FUT2data` vector and the column of `clinicalData` with the normal/high CEA categories were passed to the plot function, again masked to remove outliers. To create two histograms for the *FUT2* RNAseq data, the `FUT2data` and the CEA level data was passed to the histogram function each, making two plots showing each distribution. Finally to create the Kaplan-Meier plot, first the NA values in the `days_to_death` column were replaced with values from the

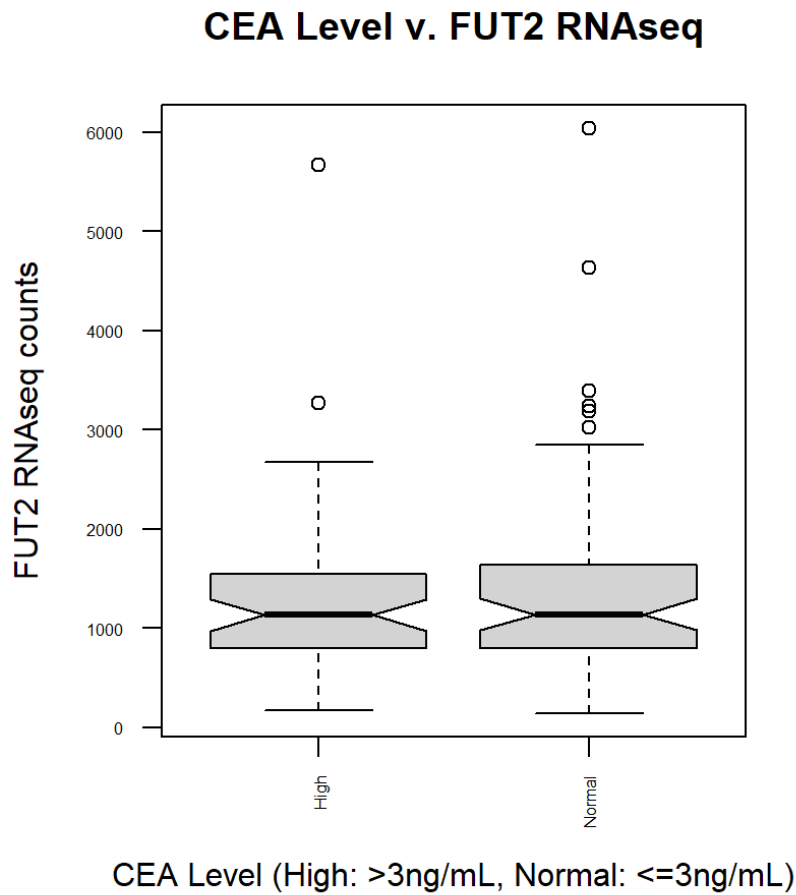
days\_to\_last\_follow\_up column, then a boolean death\_event column was created with TRUE values if the patient was deceased, and those columns were used to initialize the survival object. Then the fit object was created using the CEA categories column, and it was all fed to the survplot function.

Once the data was all collected, the results were analyzed to draw final conclusions, and determine if the hypothesis is supported.

## Results

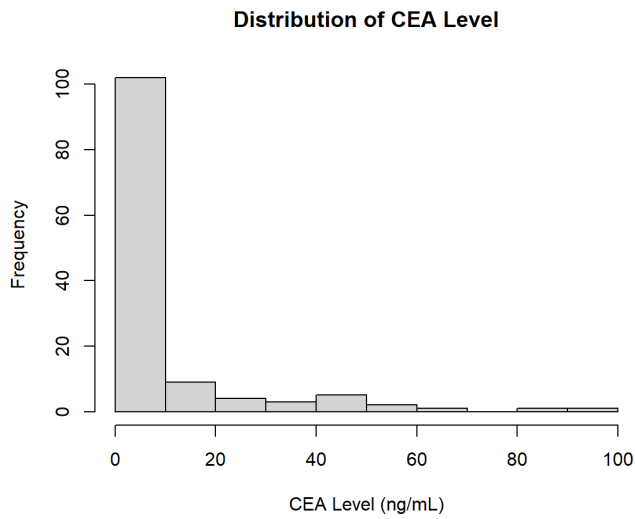


**Figure 1.** Scatter plot comparing the *FUT2* expression data with the raw CEA level data (non-categorized). The x-axis is measured in RNA transcript counts and the y-axis is measured in ng/mL

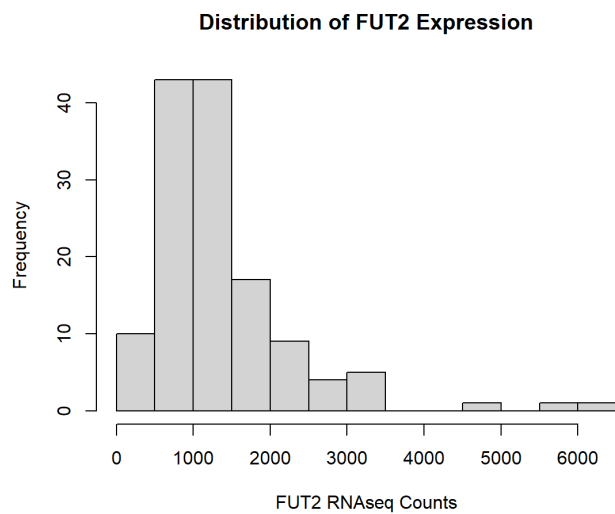


**Figure 2.** Box plot comparing the *FUT2* expression of patients with normal CEA levels and patients with high CEA levels. Normal CEA levels are considered  $\leq 3\text{ng/mL}$  and high levels are considered  $> 3\text{ng/mL}$ .

Figure 1 and Figure 2 show that there is not significant correlation between *FUT2* expression and CEA level. In Figure 1, there is no obvious best fit line or pattern between the two axes. In Figure 2, there seems to be a very slightly higher expression of *FUT2* in patients with normal CEA levels, however it is very small and not nearly enough to be statistically significant.



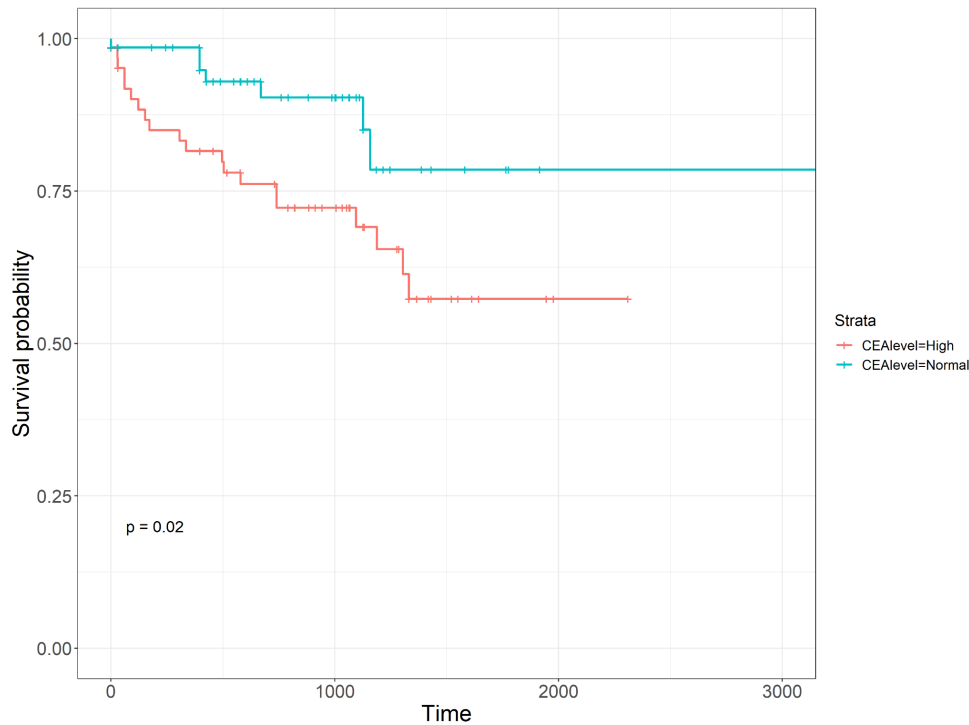
a)



b)

**Figure 3.** Histograms showing the frequency distribution of a) CEA level (ng/mL) and b) *FUT2* RNAseq counts.

Figure 3 shows the distribution of both CEA level and *FUT2* expression level. The CEA level seems to have the highest frequency between 0 and 10 ng/mL, with a very large drop in frequency in the next bin. The *FUT2* expression level seems to have the highest frequency between 500 and 1500 counts, with the distribution centered around 1000 counts. Both histograms show far outliers.



**Figure 4.** Kaplan-Meier plot showing the survival probability of two groups, patients with normal CEA levels ( $\leq 3\text{ng/mL}$ ) and patients with high CEA levels ( $>3\text{ng/mL}$ ). Key is labeled “strata” with the high CEA group labeled red and the normal CEA group labeled blue. The p-value is shown on the empty space of the graph.

Figure 4 shows a statistically significant ( $p < 0.05$ ) difference between the survival probability of patients with normal versus high CEA levels. Patients who have normal CEA levels are more likely to live longer, and those who have high CEA levels are less likely.

## Discussion

Based on the results collected from the analysis of the TCGA colorectal cancer data, the initial hypothesis can only be partially supported. Based on the data from Figures 1 and 2, there is no obvious correlation between FUT2 expression and CEA levels in CRC patients. This result is interesting, because it contradicts the findings of Liang et al. which determined that SNPs on

the *FUT2* gene (which would lower expression) are associated with elevated CEA levels. However this study was conducted from data collected from a CRC patient population in southern China, so it is possible that the mutation is a localized phenomenon. The narrow scope of our data set could also have something to do with our unsupported results, as Figure 3 shows that the distribution of data is very concentrated near a relatively smaller range.

However, the data does support the idea that elevated CEA levels are associated with lower survival rates in CRC patients. In Figure 4, it's clearly shown that the high CEA level group has lower survivorship, with a statistically significant p-value of  $p = 0.02$ . These findings can be used to improve prognostic methods for CRC patients by referencing the CEA levels in their body.

Overall based on our results, it cannot be concluded that there is a correlation between the expression of *FUT2* and CEA levels in CRC patients, either direct or inverse. However, there is statistically significant evidence that shows that elevated CEA levels are associated with lower survival in CRC patients than that of CRC patients with normal CEA levels.

For future research, it would be beneficial to use a data set with more variety of both CEA levels and *FUT2* expression, because the fact that many of the patients had relatively low CEA levels may have influenced the accuracy of the data analysis. It would also be interesting to look into how the expression and/or mutation of the gene that encodes CEA can affect CEA levels and survivorship. Lastly, one study concluded that *postoperative* CEA levels were more indicative of prognosis than *preoperative* CEA levels, and the data used for this analysis was the latter (Lin et al., 2011). Therefore it would be pertinent to reconduct this data analysis using *postoperative* CEA levels in the future, to see if there is a stronger correlation.

## References

- Becerra, A. Z., Probst, C. P., Tejani, M. A., Aquina, C. T., González, M. G., Hensley, B. J., ... & Fleming, F. J. (2016). Evaluating the prognostic role of elevated preoperative carcinoembryonic antigen levels in colon cancer patients: results from the national cancer database. *Annals of surgical oncology*, 23(5), 1554-1561.
- Bodmer, W. F. (2006). Cancer genetics: colorectal cancer as a model. *Journal of human genetics*, 51(5), 391-396.
- Kolligs, F. T. (2016). Diagnostics and epidemiology of colorectal cancer. *Visceral medicine*, 32(3), 158-164.
- Liang, Y., Tang, W., Huang, T., Gao, Y., Tan, A., Yang, X., ... & Peng, T. (2014). Genetic variations affecting serum carcinoembryonic antigen levels and status of regional lymph nodes in patients with sporadic colorectal cancer from Southern China. *PloS one*, 9(6), e97923.
- Lin, J. K., Lin, C. C., Yang, S. H., Wang, H. S., Jiang, J. K., Lan, Y. T., ... & Chang, S. C. (2011). Early postoperative CEA level is a better prognostic indicator than is preoperative CEA level in predicting prognosis of patients with curable colorectal cancer. *International journal of colorectal disease*, 26(9), 1135-1141.



## Review Questions

### General Concepts

1. What is TCGA and why is it important?

*The TCGA is The Cancer Genome Atlas, and it is important because it is a publicly available data bank containing relevant genomic, transcriptomic, proteomic, and clinical information about cancer patients that can be used to conduct cancer research more effectively.*

2. What are some strengths and weaknesses of TCGA?

*One strength of the TCGA is that it allows the sharing of data to allow researchers to have larger data sets to work with when conducting research. One weakness of the TCGA is that privacy issues are a problem that can prevent patients from wanting to participate in the project.*

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

*The data we are exploring are literally the quantitative components of the central dogma of biology. Using the TCGA we can look directly at genomic, transcriptomic, and proteomic data.*

## Coding Skills

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

*cd \*FILEPATH\*/qbio\_data\_analysis\_tessa*

*git status*

*git add file.R*

*git commit -m "commit"*

*git push*

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

*BiocManager::install("Package") # only if the package has not been installed yet*


*library(Package) # must be run every time you open a new work*

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

*Boolean indexing is creating a vector of TRUE/FALSE values in order to classify specific rows or columns from a dataframe. This can be used for applications such as filtering out data or putting data in specific categories.*

4. Draw out a dataframe of your choice.

ensembl gene ID	external gene name	original ensembl gene ID
ENSG00...#	GENE1	ENSG00...##
ENSG00...#	GENE2	ENSG00...##
ENSG00...#	GENE3	ENSG00...##
etc....	etc....	etc....



*rowData(sum\_exp)*

5. Show an example of the following and explain what each line of code does.

- a. an ifelse() statement

```
dataframe$column2 = ifelse(dataframe$column >= 5, "Grp1", "Grp2")
```

*# if the value of 'column' is >=5, 'column2' reads "Grp1"*

*# else (value is <5), it reads "Grp2"*

- b. boolean indexing

```
dataframe = dataframe[dataframe$column <= 10, ]
```

*# filters out all rows of 'dataframe' in which the value of 'column' is >10*