Brian Tinsley

Qbio Student Group

Midsemester Project

*How does the presence of lymphatic invasion relate to age, survivability, gene expression, and mutation frequency?*

**Introduction:**

Colorectal cancer is a major cause of death and tragedy worldwide. The National Cancer Institute estimates that there were 149,500 new cases of colorectal cancer in 2021, with 52,980 (35%) resulting in death; the American Cancer Society predicts over 52,000 people to die from colorectal cancer in 2022. In this study, I will be examining lymphatic invasion, which is involved in the process of lymphatic metastasis. It is important to investigate lymphatic invasion because cancers that spread to areas of the body other than the affected area are often more deadly and untreatable. One study strongly associates lymphatic invasion with poorer survival and prognosis (Harris). In this paper, I will examine how age, gene expression, mutation frequency, and survivability relate to the presence of lymphatic invasion. Using data from the TCGA data portal, R packages and scripts will be used to analyze this data. Clinical data of patients in the study contains information about lymphatic invasion being present (YES/NO), as well as other information about the lymph nodes such as the number of nodes examined, and the number of nodes tested positive by multiple tests. Using the clinical data, RNA seq data, and mutation allele frequency (MAF) data, I have generated plots that support the thesis that there are significant differences between patients with lymphatic invasion and those without.

**Methods:**

I accessed colorectal cancer clinical and RNAseq data from TCGA using the R package TCGAbiolinks with the accession code “COAD.” Looking at the clinical data, some of the relevant column names were *lymphatic\_invasion*, *lymph\_node\_examined\_count*, *number\_of\_lymphnodes\_positive,* and *primary\_lymph\_node\_presentation\_assessment.* The first step in analysis of lymphatic invasion was to create a new column called *age\_category* which would turn the continuous age variable into a discrete categorical variable, labeling patients as young or old with a threshold of 50 years as the cutoff. Next, I subset the data to remove patients who had NA data for the lymphatic invasion column, removing 65 of the initial 524 patients in the study. Of the remaining 459 patients, 188 had lymphatic invasion, while 271 did not. The barplot in Figure 1 shows the distribution of young and old patients with lymphatic invasion.

Using the R packages “survival” and “survminer,” I was able to create a the necessary items for a survival analysis plot. After loading in the packages, I created a days to death column, as well as a death event column using the vital status column. Then, I created a survival object with these columns, followed by the creation of a survival fit object using lymphatic invasion. Finally, I created the Kaplan Meier plot shown in Figure 2.

For RNAseq analysis, I installed the package “SummarizedExperiment” and queried the transcriptome profiling HTSeq - counts data. After creating a summarized experiment object, I looked at the column names of the colData and found the same columns from the previous methods. The column *paper\_lymphatic\_invasion\_present* had many <NA> values, as well as “NA” values. I cleaned these values up separately using boolean indexing, leaving 191 patients in the data. I also used boolean indexing to locate the rows with two genes I selected, TP53 and APC, and extract their ensemble gene id. Figures 3 and 4 show the gene counts for TP53 gene and APC gene, respectively, for patients with and without lymphatic invasion.

For DESeq analysis, I loaded the R package “DESeq2.” Using the same data fram as the RNAseq data, I extracted the clinical data from colData and the counts from assays(sum\_exp). After converting the counts data to a data frame, I created another boolean mask to extract rows without NA values in the lymphatic invasion column. I subset the counts and patient data to match this mask, and then calculated the rowSums of the counts data frame to get the absolute number of counts for each gene. I filtered out counts that were less than 10 using another boolean mask, which filtered out 8752 genes that would have probably made the data much messier for graphing. After looking at the results and listing them in descending order, we set the log2 fold change threshold to -1 and the padj threshold to 0.05. Using the “ggplot2” library and some of the DESeq2 functions, I created the volcano plot found in Figure 5.

For MAF analysis, I loaded the R package “maftools” and renamed the colnames of the clinic data to Tumor\_Sample\_Barcode. After querying the data and creating a MAF object, we access the clinical data and mutation data using the “@” symbol. Using the oncoplot function, I created the plot shown in Figure 6. Since we changed the colnames in the data, both the clinical data and MAF data have tumor sample barcode columns. I used boolean indexing again to find the patient id’s that had lymphatic invasion and the patients that did not. Using the subsetMaf function, I subset the MAF data using the “YES” patient ids and the “NO” patient ids. This allowed me to create the co-oncoplot shown in Figure 7. Using the lollipopPlot function, I was able to plot the mutation data for TP53 and APC genes, and I used lollipopPlot2 to plot “YES” vs. “NO” for each gene as shown in Figures 8 and 9.

**Results and Figures:**

In Figure 1, we can see that the overwhelming majority of patients with lymphatic invasion are older than 50.

Chart, bar chart

Description automatically generated

Chart, line chart

Description automatically generatedIn Figure 2, we can see that the overall survival time of patients with lymphatic invasion is much lower than patients without.

Chart, box and whisker chart

Description automatically generatedIn Figure 3, we can see there is a wider range of counts for the TP53 gene in patients with lymphatic invasion, but patients without lymphatic invasion have more upward outliers.

In Figure 4, we can see that the counts for the APC gene for patients with and without lymphatic invasion are very similar.

**Chart, box and whisker chart

Description automatically generated**

Figure 5 shows us that patients with no lymphatic invasion have many more genes that are significantly upregulated.

**Chart, scatter chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**Figure 6 shows the mutation data for the top 10 most mutated genes in patients with colorectal cancer.

**Chart, bar chart

Description automatically generated**Figure 7 shows the mutation data for the top 6 most mutated genes for patients with and without lymphatic invasion.

Figure 8 shows us the number of mutations of the TP53 at certain locations in the data set for patients with and without lymphatic invasion.

**Timeline

Description automatically generated**

Figure 9 shows the number of mutations of the APC gene at certain locations in the data set for patients with and without lymphatic invasion.

Chart

Description automatically generated

**Discussion:**

After analysis, we can see that presence of lymphatic invasion is affiliated with older patients, and it does in fact lead to poorer survival. It should not come as a surprise since lymphatic invasion is a part of lymphatic metastasis, a more untreatable type of cancer; once cancer spreads from the colon to the lymphatic system and the rest of the body, prognosis is typically much worse. Figures 3 and 4 show us no significant findings about the difference in expression between patients with and without lymphatic invasion in the TP53 and APC genes. One potential area for future studies is looking at the outliers in the counts for TP53 for patients without lymphatic invasion, as well as the wider range of data for patients with lymphatic invasion. Figure 5 is one of the most interesting charts of the entire results section; it shows a significant difference in the relative expression for patients with and without lymphatic invasion. Here, we have more genes that have more significance in the upregulated side for the patients with no lymphatic invasion. One area to look into in the future would be changing the thresholds around for the volcano plot and seeing if the results are drastically different. Figure 6 is less informative as to the differences in mutations between patients with and without lymphatic invasion, but it gives a good general focus for which genes would be interesting to look into for this study. Figure 7 yields interesting results; we can see that TP53 and APC mutation rates were actually not as different as we would have expected, while KRAS and PIK3CA had significant differences in mutation rates between the two groups. In future studies, more close examination into those two genes could yield more informative results, potentially in the RNAseq data. In Figure 8, we can see a few differences between patients with lymphatic invasion and those without. First, we can see that the patients with lymphatic invasion have a 5% higher mutation frequency of this gene. Furthermore, these patients have mutations at the beginning of the DNA segment whereas the other patients do not. This could be a very important finding; maybe the mutations at the beginning of the sequence are a potential cause for lymphatic invasion. More tests would have to be done in future studies to examine this. Interestingly, these mutations at the beginning are all insertions and deletions, which are more impactful mutations than missense and nonsense mutations. Figure 9 shows us that the APC gene is actually mutated less often in patients with lymphatic invasion, so perhaps another gene like KRAS or PIK3CA would be a better area of focus. Interestingly, and similarly to in figure 8, there is a frame shift insertion at the beginning of patients with lymphatic invasion. Both of these genes are almost equally expressed (from figures 3 and 4) and have fairly similar mutation frequencies, but in both genes, the patients with lymphatic invasion have a frame shift mutation at the beginning. Lymphatic invasion is associated with poorer survival and is therefore an important area for research, and this study has opened the door to future studies.

**References:**

Harris, Elizabeth I. MD\*; Lewin, David N. MD†; Wang, Hanlin L. MD, PhD‡; Lauwers, Gregory Y. MD§; Srivastava, Amitabh MD∥; Shyr, Yu PhD¶; Shakhtour, Bashar PhD¶; Revetta, Frank BS, HT, QIHC\*; Washington, Mary K. MD, PhD\* Lymphovascular Invasion in Colorectal Cancer, The American Journal of Surgical Pathology: December 2008 - Volume 32 - Issue 12 - p 1816-1821 doi: 10.1097/PAS.0b013e3181816083

Yuan, Hang, et al. "Lymphovascular invasion is a high risk factor for stage I/II colorectal cancer: a systematic review and meta-analysis." Oncotarget [Online], 8.28 (2017): 46565-46579. Web. 8 Mar. 2022

*These websites were just for quick statistics for the introduction section:*

<https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html>

<https://www.cancer.gov/types/common-cancers>

**Part 2: Review Questions**

Attach the answers to the following questions after your References.

General Concepts

1. What is TCGA and why is it important?

The Cancer Genome Atlas is an organization that collects data from cancer patients and allows them to be publicly analyzed by people like us to help make an impact in the fight against cancer.

1. What are some strengths and weaknesses of TCGA?

A strength of the TCGA is that many different organizations are helping contribute to these data sets, so it is getting bigger and bigger. One weakness is that some data, such as hereditary mutations, cannot be publicly available because these patients could be identified, so there is a loss of valuable data in the public datasets.

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

We are looking at mutations, which can happen in both transcription and translation. We use MAF and RNA seq to see where mutations are happening, and how often mutated strands of genetic information are being expressed in hopes of discovering new findings.

Coding Skills

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

Git add, git commit, git push

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

Library()

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

Boolean indexing is saving the row/column indexes where a certain value is found, such as the rows containing NA, and then passing those indexes in to subset the data either including or excluding those specific rows.

1. Draw out a dataframe of your choice. Show an example of the following and explain what each line of code does.
   1. an ifelse() statement

an ifelse statement can be used to create a new column. The first parameter is an expression, and the second parameter is what should go in the column if the expression is true, and the third parameter is what to fill in if the expression is false. For example, for the age category, to create a categorical variable rather than a continuous one, we could write…

dataframe$age\_category = ifelse(dataframe$age > 50, “old”, “young”)

* 1. boolean indexing

na\_mask <- dataframe[is.na(dataframe$age),]

#this saves the indexes of rows with NA in age column

Newdatafram <- dataframe[!na\_mask,]

#this subsets the dataframe where there are no NA’s in the age column using our mask above