R Review Project Part 1

Review Questions

General Concepts

1. What is TCGA & why is it important?
   TCGA stands for "The Cancer Genome Atlas." The TCGA was organized by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) with the goal of creating a public database that displayed a wide range of genes across a vast patient sample. TCGA contains genomic data from over 20,000 samples of thirty three cancer types and presents clinical data, mutation data, and protein data (from the CPTAC). Because TCGA is a large public database that provided both clinical and multi-omic data, it served a crucial role in advancing our understanding of cancer.

2. What are some strengths & weaknesses of TCGA?
   Strengths of TCGA include having a large patient sample for various different cancer types and displaying numerous forms of data, such as clinical, mutation, and protein data. This allows researchers to collaborate with one another and carefully analyze patient and multi-omic data to make new developments in the field of cancer. On the other hand, however, TCGA does not show the progression of cancer in patients over time since they are collected at one point in time. TCGA also only focuses on common cancers and underrepresents non-white demographics.

Coding Skills

1. What commands are used to save a file to your GitHub repository?
   The GitHub commands used to save a file are as follows: use cd to go to your local repository, use git status to check which files have local changes that need to be uploaded to GitHub, use git add add a file, use git commit-m to commit changes, and use git push to push changes into GitHub repository.

2. What commands must be run in order to use a package in R?
   To use a package in R, you must first install the package. Use **install.packages("I_Am_Awesome")**, then **library(I_Am_Awesome)**.

3. What commands must be run in order to use a *Bioconductor* package in R?
   To use a *Bioconductor* package in R, you must first install the package. Use **if (!require("BiocManager", quietly = TRUE)) install.packages("BiocManager") BiocManager::install(version = "3.17")**, then **library(BiocManager)**.

4. What is Boolean Indexing? What are some applications of it?
   Boolean indexing is an effective R technique for data cleaning and subsetting/selection.

Boolean indexing involves creating a vector where selected data is true to filter and create a mask. You can use Boolean indexing to keep certain data, delete null data, subset data (ex: young/old, male/female), select certain data points based on a row or column, finding barcodes of certain patients, etc.

5. Draw a mock-up (just a few rows & columns) of a sample dataframe. Show an example of the following & explain what each line of code does.
    a. An ifelse() statement

```r
data <- data.frame(Name = c("Bartholomew", "Bob", "Charlie", "David"),
                   Age = c(2, 14, 39, 17))
#Using the data.frame function, a new data frame is created, called "data." The two columns are "Name" (which contains "Bartholomew", "Bob", "Charlie", "David") and "Age" (which contains 2, 14, 39, 17).

data$AgeCategory <- ifelse(data$Age >= 18, "Adult", "Minor")
#I am adding a new column called "AgeCategory" to categorize the Names as either "Adult" or "Minor" based on their age. The ifelse is a condition operation that identifies this. If the age is greater than or equal to 18, the name will be categorized as "Adult," and if the age is less than 18, the name will be categorized as "Minor."

print(data)
#Using the print function, the data.frame "data" is now displayed below. It will show the name, age, and age category.
```

Description: df [4 × 3]

Name <chr>	Age <dbl>	AgeCategory <chr>
Bartholomew	2	Minor
Bob	14	Minor
Charlie	39	Adult
David	17	Minor

4 rows

    b. Boolean Indexing

```r
data <- data.frame(Name = c("Bartholomew", "Bob", "Charlie", "David"),
                   Age = c(2, 14, 39, 17))
#Using the data.frame function, a new data frame is created, called "data." The two columns are "Name" (which contains "Bartholomew", "Bob", "Charlie", "David") and "Age" (which contains 2, 14, 39, 17).

data$AgeCategory <- ifelse(data$Age >= 18, "Adult", "Minor")
#I am adding a new column called "AgeCategory" to categorize the Names as either "Adult" or "Minor" based on their age. The ifelse is a condition operation that identifies this. If the age is greater than or equal to 18, the name will be categorized as "Adult," and if the age is less than 18, the name will be categorized as "Minor."

data <- data[data$AgeCategory == "Minor", ]
#This line of code uses Boolean indexing to filter out all of the adults from the data frame "data."

print(data)
#Using the print function, the data.frame "data" is now displayed below. It will show the name, age, and age category. Because adults were filtered out, the data will only show the names and ages of those that are minors.
```

Description: df [3 × 3]

	Name <chr>	Age <dbl>	AgeCategory <chr>
1	Bartholomew	2	Minor
2	Bob	14	Minor
4	David	17	Minor

3 rows

R Review Project Part 3

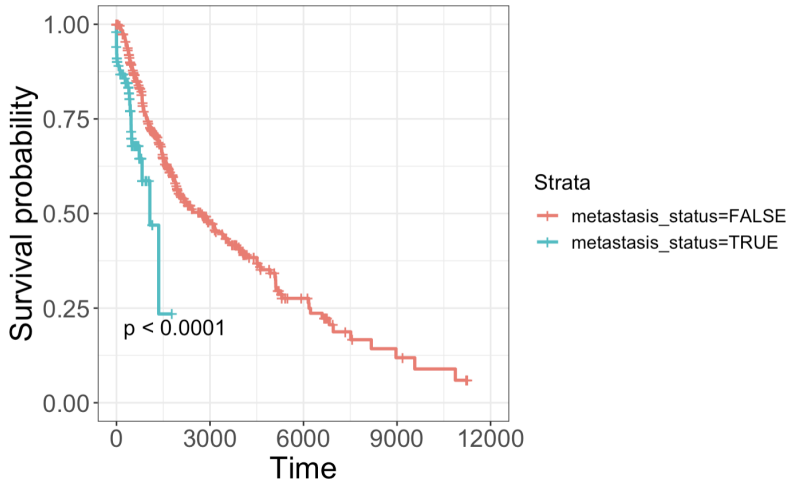
Results & Interpretations

For each analysis, include an image of the relevant plot, as well as a description.

1. Describe the plot(s). What kind of plot is it? What is it showing?
2. What conclusions can you draw about differences b/n metastatic & non-metastatic TCGA SKCM patients? Why?
3. What's one conclusion you cannot draw? Why?
4. Describe at least one academic article (research or review) that either supports or doesn't support your conclusion. If previously published work doesn't support your analysis, explain why this might be the case.

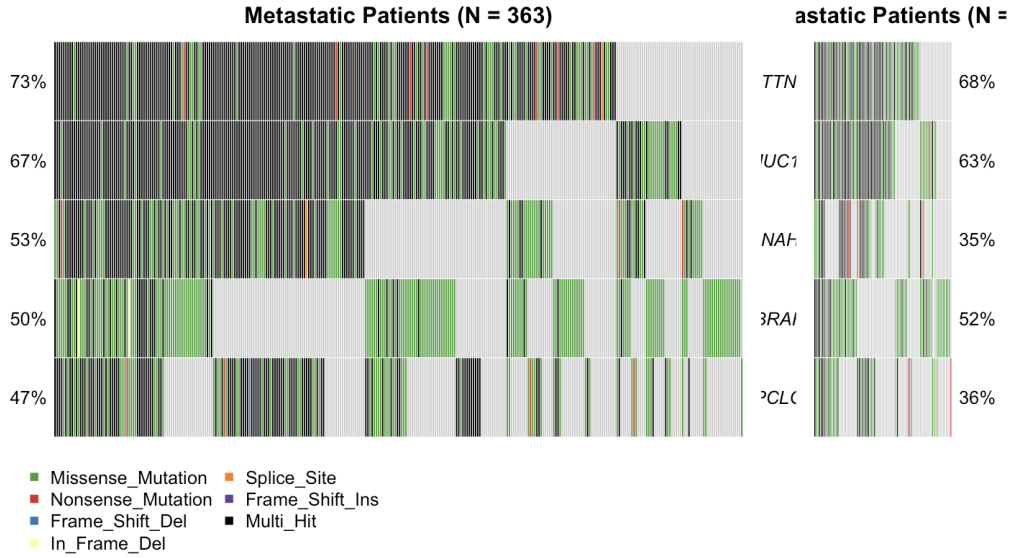
---

1. Differences in survival b/n metastatic & non-metastatic patients.



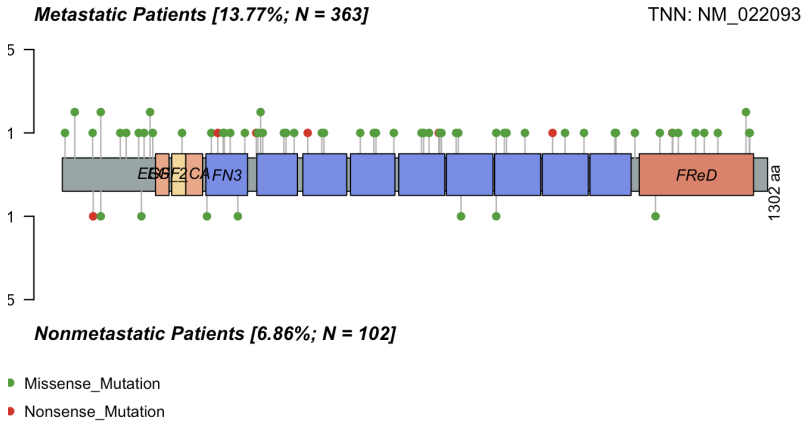
This plot is a Kaplan-Meier survival plot, which shows survival probability over time (days). There are two lines on the plot: the blue line represents patients with metastatic cancer, while the red line represents patients without metastatic cancer. Starting at 100%, the survival probability for each group of patients decreases as time goes on; in this case, the blue line (patients with metastatic cancer) is less steep than the red line (patients without metastatic cancer), meaning that the KM plot indicates that patients with metastatic cancer have a lower probability of surviving over time than patients with non-metastatic cancer. A conclusion I cannot draw is that a certain treatment is the reason behind the higher/lower mortality rate, since the KM plot merely presents the differences in survival and not the possible reasons/factors behind the displayed survival rates. The conclusions made by this KM plot is supported by an article called "Analysis of prognostic factors for melanoma patients" by Letautienė, Simona, et al. The article discusses how metastatic melanoma patients generally have lower survival rates than non-metastatic melanoma patients by citing several prognostic factors that influence their likelihood of surviving.

2. Mutation differences b/n metastatic & non-metastatic patients for multiple genes.



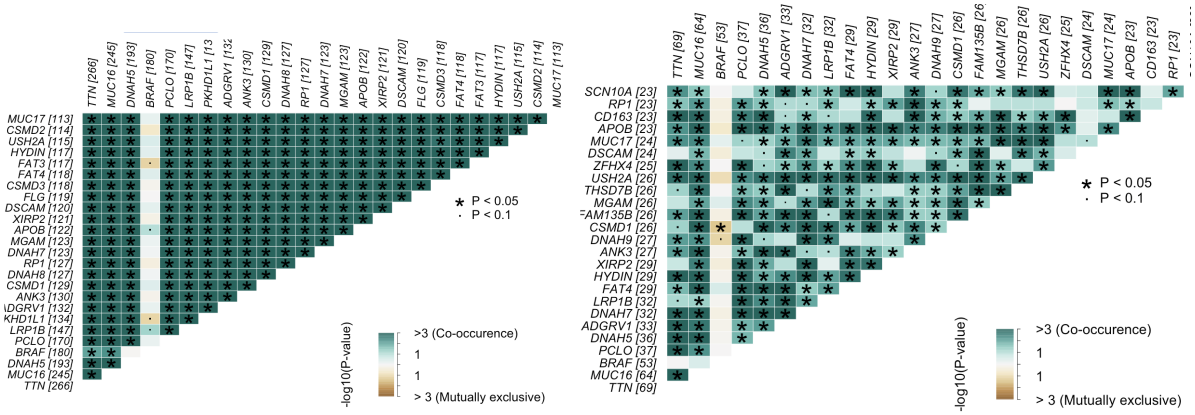
This plot is a Mutation Co-Oncoplot, which compares the mutation distribution of two subsets of patients— in this case, the subsets are metastatic patients (the left one) & non-metastatic patients (the right one). Each bar represents a gene, and the bars are color-coded based on the type of mutation it is. I can conclude that most commonly mutated genes by order are TTN, MUC16, DNAH5, BRAF, and PCLO, and the most common types of mutations are missense and multi-hit mutations. A conclusion I cannot draw is that these specific mutations will cause or prevent metastasis, since the plot only presents data and does not establish a cause-and-effect relationship between the mutations and metastasis status. The conclusions made by this Mutation Co-Oncoplot is supported by an article called "Genetics and genomics of melanoma" by Chin, Lynda, et al. The article discusses genetic mutations that are associated with various types of melanoma and emphasizes differences between metastatic and non-metastatic melanoma patients.

3. Mutation differences for specific genes of interest.



This plot is a Co-Lollipop Plot, which, like a standard Lollipop Plot, shows mutation count by location on the gene; except, unlike a standard Lollipop Plot, it compares two sets of patients rather than showing just one set of patients. In this case, the Co-Lollipop shows a side-by-side comparison of the mutations found on TNN for metastatic patients and non-metastatic patients. Looking at the Co-Lollipop Plot, one can conclude that metastatic patients have a lot more missense mutations— and mutations overall— compared to non-metastatic patients. A conclusion I cannot draw is that having a mutation on specific locations of the TNN gene causes or is a result of metastasis, since the Co-Lollipop Plot only aids in comparing two subsets of patients and does not demonstrate or prove causality. The conclusions made by this Co-Lollipop Plot is supported by numerous articles. “TTN mutations predict a poor prognosis in patients with thyroid cancer” by Chen, Jianrong, et al.; “Mutations in the TTN Gene are a Prognostic Factor for Patients with Lung Squamous Cell Carcinomas” by Hu, Sheng, et al.; and “Spontaneous mutations in the single TTN gene represent high tumor mutation burden” by Cho, Eun J., et al. all discuss how TNN has been observed as a common mutation in metastasized cancers.

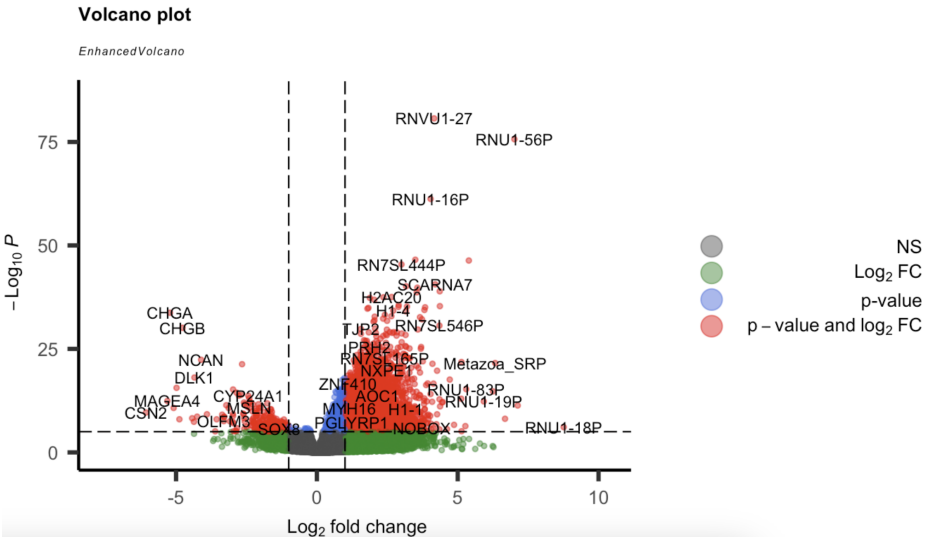
4. Co-occurrence or mutual exclusion of common gene mutations (one for metastatic & one for non-metastatic patients).



These plots are both Somatic Interaction Plots, with the left one being for metastatic patients and the right one being for non-metastatic patients. Somatic Interaction Plots show interactions between somatic mutations in genes, with the color corresponding to the interaction either being co-occurring or mutually-exclusive and the opacity corresponding to the strength of the association. The significance is labeled by . or *. For non-metastatic patients, it can be seen that mutations in CSMD1 and BRAF are significantly mutually exclusive, meaning that they almost never occur together in non-metastatic patients. This is a contrast to metastatic patients, since the Somatic Interaction Plot for metastatic patients shows that there is a slight association between the two genes and that they are more likely to be co-occurring (the box is light blue). A conclusion that I cannot draw is that metastatic patients have more mutations than non-metastatic patients, since the Somatic Interaction Plots only show whether genes are likely to be mutated together or not. The conclusions made by the Somatic Interaction Plots is supported by an article

called “BRAF mutations in metastatic melanoma: a possible association with clinical outcome” by Angelini, Sabrina, et al. This article discusses how BRAF gene mutations are associated with metastatic melanoma. The study found that patients with metastatic melanoma and BRAF mutations saw shorter durations of response to treatment, meaning that such mutations potentially played a role in disease outcomes.

5. Differential expression b/n non-metastatic & metastatic patients controlling for treatment effects, race, gender, & vital status.



This plot is a Volcano Plot, which helps visualize results made by differential gene expression analysis for metastatic and non-metastatic patients, taking into consideration treatments, gender, race, and vital status. Each circle on the plot represents a gene, the x-axis represents log2 fold change, and the y axis represents -log10 p-adjusted value. The red circles on the upper right corner of the plot are significantly up-regulated, the red circles on the upper left corner are significantly down- regulated, the green circles on the bottom right corner are insignificantly up-regulated, the green circles on the bottom left corner are insignificantly down-regulated, and the blue circles in the middle show approximately equal expression. I can conclude that RNVU 1-27 is significantly up-regulated for metastatic patients, but I cannot conclude that RNU 1-16P has more influence than RNVU 1-27 or vice versa. This is because the Volcano Plot only displays significance and it cannot be used to establish a cause-and-effect relationship; therefore, relative impact or influence cannot be concluded based on the Volcano Plot alone. The conclusions made by this Volcano Plot is dissimilar to the Volcano Plot shown by an article called “The Integrative Analysis Identifies Three Cancer Subtypes and Stemness Features in Cutaneous Melanoma” by Cheng, Yaqi, et al. This article also presents a Volcano Plot to visualize their differential gene expression analysis, but theirs had much more significantly down-regulated genes (319) than significantly up-regulated (45). The difference is most likely as a result of our differences in covariates.

Works Cited

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467960/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771951/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9310696/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8742622/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7424531/

https://pubmed.ncbi.nlm.nih.gov/12960123/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7921163/