Rebecca Aviles

QBIO 490: Directed Research

13 October 2023

R Review Project

Part 1: Review Questions

1. What is TCGA and why is it important?

Based on what we have learned in QBIO 490, TCGA is an acronym that stands for "The Cancer Genome Atlas". In this database, there is public information regarding an array of different cancer types in which multi-omics data and future research could learn more about the following cancers. This is important because with the diverse and large amount of patient data we have this can give us insight on patients and their corresponding cancers in a way to understand to a cellular/microscopic level.

2. What are some strengths and weaknesses?

In TCGA, we can depict a lot in multi-omics due to variety of research and data. It is also HIPPA protected, so many of the patient and their identities are concealed. Some of the strength include the abundance of data and the availability to different kinds of data. However, some of the weaknesses can include not being able to study long-term effects of the corresponding cancers and cleaning needed to actually do the analysis that most private institutions do not have problems with as often.

Coding Skills

1. What commands are used to save a file to your Github Repository?

When I normally save a file on my repository. I go to my Terminal to make sure my pathway and cd the file to my own QBIO_490_rebecca folder. Once it is in there, I then go into my GitHub in which I upload the. Rmd File and I upload into my repository.

2. What command(s) must be run to use a package in R?

First, if the package is not installed into your R studio you must implement the install. Packages () function and put your respective package in there. Let that code run and then make sure to use the library function to implement it in your code for the future. Also, once everything is installed, make sure the install package is commented out that way you do not constantly upload it every single time you run your code.

3. What command(s) must be run to use a Bioconductor package in R?

When installing for a Bioconductor Package in R, one must first code in and do an install. Package function for BiocManager. Think of BiocManager as a stream that the code can include. Once this is implemented, the library function can be used to readily use like the question above. Now that you have this function you can actually use it independently BiocManager: install(). Using R studio's columns that give you the options, one can choose what is needed and one can just run what needs to be ran and installed.

4. What is Boolean indexing? What are some applications of it?

Boolean Indexing is a way in which data or other analysis can be filtered out using a Boolean vector. In this case a vector can be seen as a line of code that can make data using specific conditions in which the vector must see if the data adheres to it or not. Some application of it is when we have NA values in data. Specifically, in our breast cancer data, the application when we are reviewing race or gender and there is no useful data, we can use Boolean indexing to make a

na_mask in order to remove it from the set. By doing this, then we can apply it to our new revised data that is cleaner.

5. Draw a mock up (just a few rows and columns) of a sample data frame. Show an example of the following and explain what each line of code does.

Smoothie DATA: [1] bananas kiwis strawberries bananas kiwis kiwis bananas strawberries [9] strawberries bananas bananas

In a data set of 11 features: this can be about finding specific fruit and only wanting bananas and strawberries for the data.

Since we want only bananas, but we have kiwis in this data we have two ways of removing it.

a) An ifelse() statement:

ifelsefruit_data4smoothie_data == 'kiwis', F,T)

by only doing this line of code it should only return the data as if it is true or false; whereas boolean indexing goes the extra mile by removing it since you add the condition to the respective vector.

b) Boolean indexing

Smoothie DATA: [1] bananas kiwis strawberries bananas kiwis kiwis bananas strawberries strawberries [10] bananas bananas

In a data set of 11 features: this can be about finding specific fruit and only wanting bananas and starberries for the data.

Since we want only bananas but we have kiwis in this data we have two ways of removing it.

Boolean Indexing:

kiwi_mask <- ifelse (fruit_data$smoothie_data == 'kiwis' , F, T)

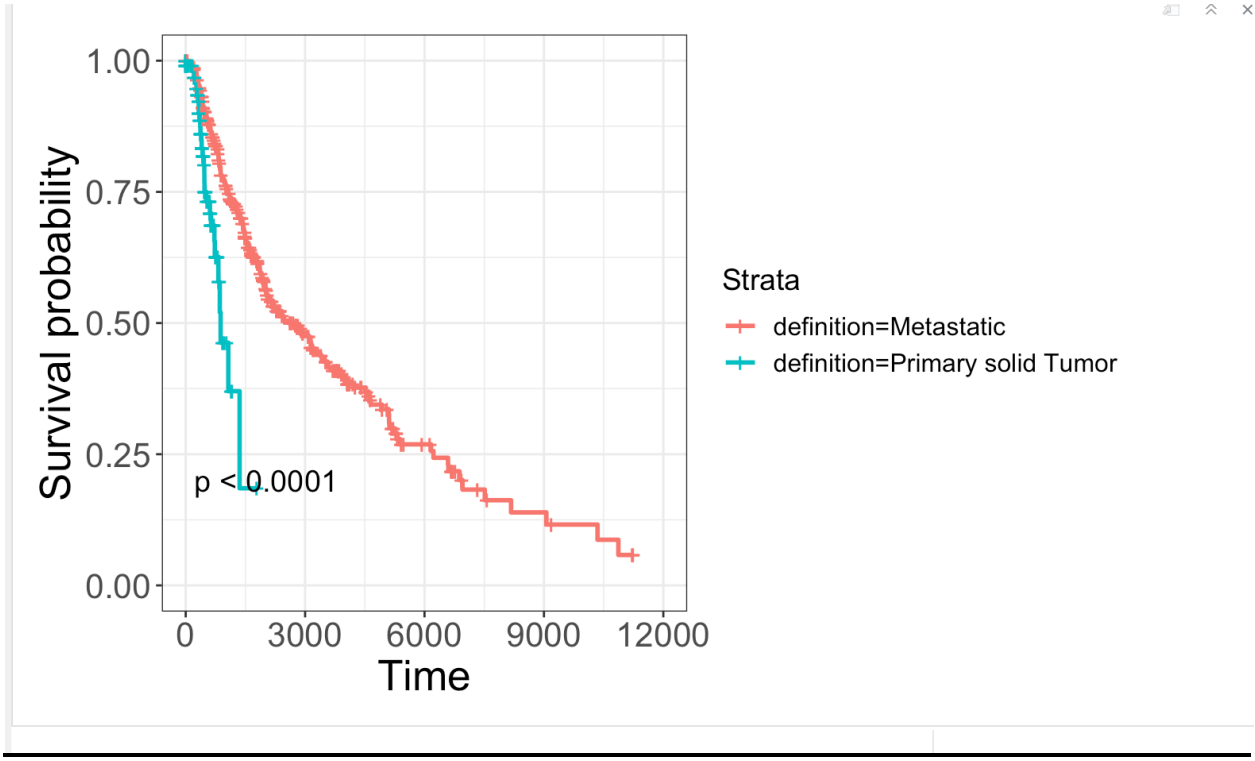
*this code is meant for doing an if else statement for if there is a string with the name ‘kiwis’ that it is false and should be removed from the data set. Anything else is good.

fruit_data <- fruit_data[kiwi_mask,]

this implement the mask using Boolean indexing by sayinf it is a kiwi it considered false and immediately removes it.

*this code is important because it can implement it in the actual database that way when we check it

Part 3 : Results and Interpretations

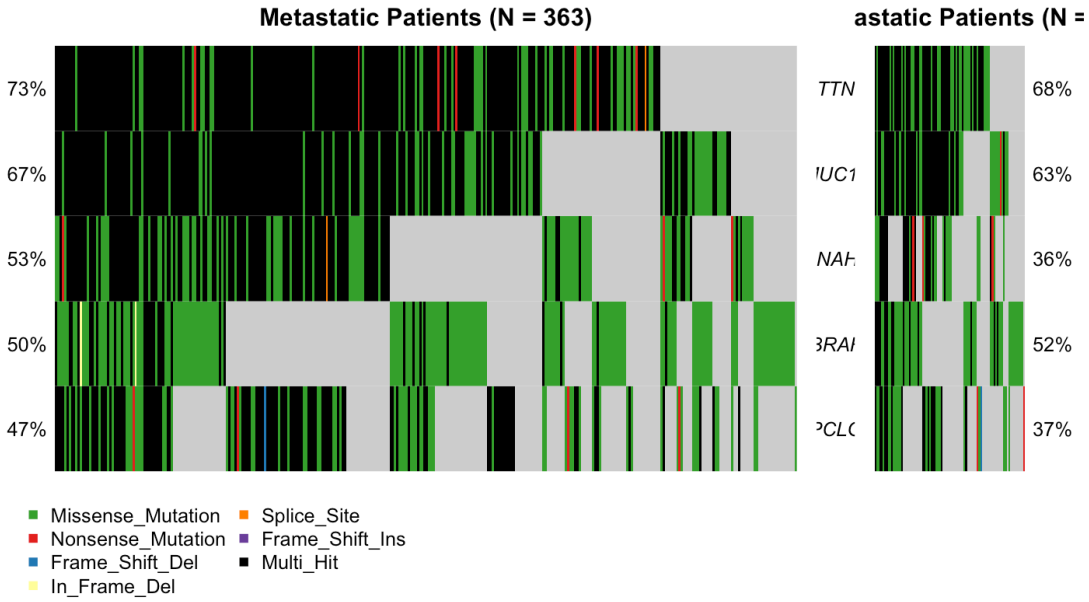


1. **Difference in survival between metastatic and non-metastatic patients**

In this analysis, a Kaplan Meier plot is what I got to distinguish between metastatic and non-metastatic patients. In this plot, the purpose is to show the two opposing patients and their

probability in terms of their survival. In this case, my metastatic patients are in red, and they are displaying signs of a prolonged time frame in comparison to the blue line of non-metastatic patients. Both patients have lower probabilities as the time gets longer; however, the non-metastatic patients have a much shorter probability at a shorter amount of time. While the graph is showing that metastatic patients have a longer survival rate (albeit at a low probability), I cannot make a distinction regarding the general Populus of all cancer data nor this specific data since the data is very short-term and it cannot show the general trend towards metastatic and non-metastatic. This is said similarly to the article, "The nucleosome remodeling and deacetylase complex has prognostic significance and associates with immune microenvironment in skin cutaneous melanoma" written by Xinhua Liu and Ju Wang. In this they did a Kaplan-Meier of the different gene expression showing similar rests in which metastatic status had a longer survival probability over time. It states, "Kaplan-Meier survival analysis demonstrated significant differences in overall survival (OS) among SKCM patients within different clusters (p value = 0.023, log-rank test) as shown in Fig. 1B, and samples in cluster2 exhibited the most superior OS." (Liu and Wang, 88) . This shows similar conclusions to my own plot in which clusters that were involved in metastatic tumor cells were more significant than those that were not.

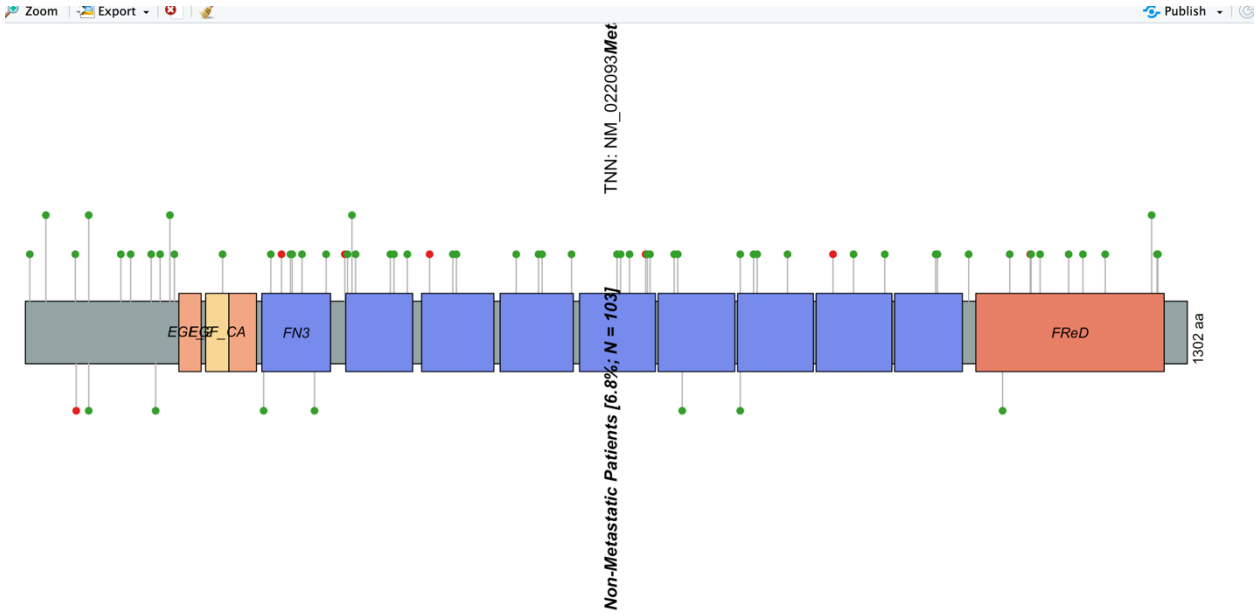
2. **Mutation Differences between metastatic and non-metastatic patients for multiple genes**



In this analysis, we were able to use maf to make a co-on coplot for both metastatic and non-metastatic patients. In the following graph, I found the top 5 genes that were the most prominent in metastatic patients and the comparison to its control of non-metastatic patients. There is a prominent number of missense mutations, in BRAF genes at about 50%. With a sample of 363 Metastatic Patients, there is a lot of Mutli-Hit mutations displaying a combination of mutations that is also noticeable. In comparison for both types of patients, it is said that there

is a different form of mutation differences in different multiple because there are different percentages for each corresponding oncoplot. There is a unanimous decision that TTN and MUC16 are the most expressed genes in the SKCM. A conclusion I feel I cannot draw using a co-oncoplot is also based on the amount of data, and if this a general trend between the data. This cannot be said because we only have about 300 patients and about 100 patients in our control. With this, it can be difficult to tell if the actual data matches up or because of the other lack of samples show a form of bias. In the article, “Construction and Identification of an NLR-Associated Prognostic Signature Revealing the Heterogenous Immune Reponses in Skin Cutaneous Melanoma”, Yi Geng, Yu-Jie Sun, et al. show different results to mine. There graphs show “the highest gene is NLRP1 and NLRP3” (Geng et al.). This is shown in the oncoplot that is shown with the most prominent genes in which there is only about 11% of the genes holding these missense mutations in these corresponding genes. This could be partially attributed to the lack of sample size or improper coding on my part.

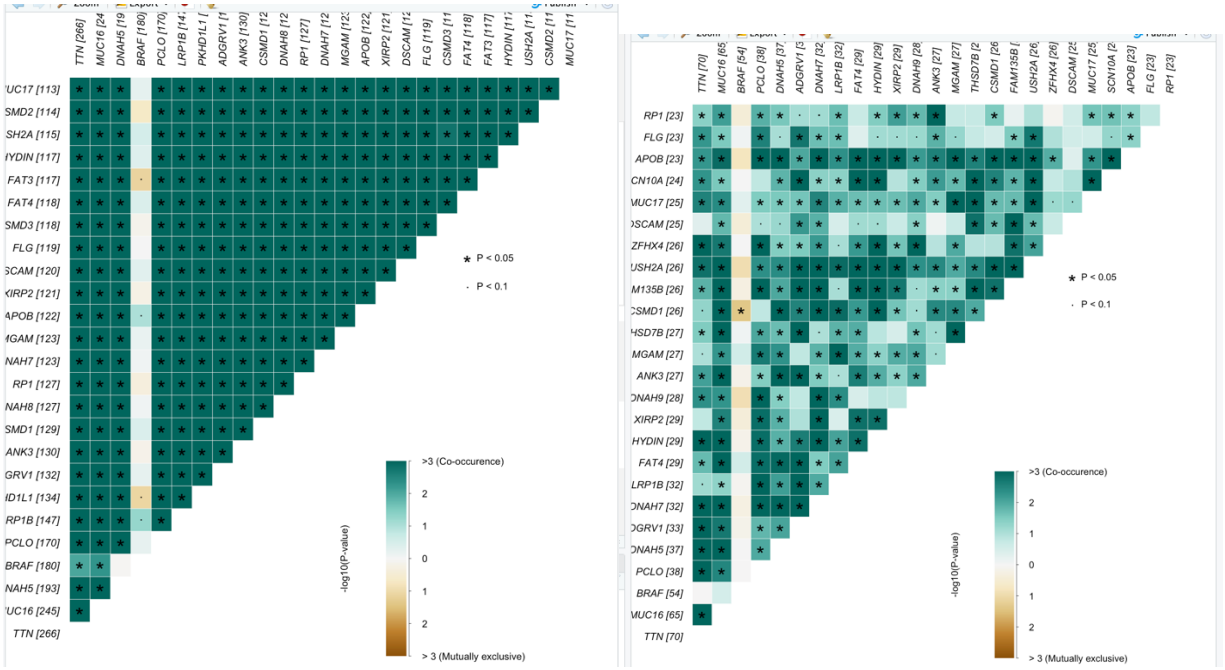
3)Mutation Differences for specific genes of interest



In this analysis, this is a co-lollipop plot with the the focus on gene “TNN”. In co-lollipop plots in which it shows both metastatic patients and non-metastatic patients. In this case, we can see that they have the highest mutation count in our maf_object is TNN for both metastatic and non-metastatic patients. This can bring me to my conclusion that is like my co-oncoplot that TNN. Based on the plot as RNA is being translated further out in Metastatic patients, we see more missense mutation—which when there is a DNA change that affects future protein making. There are more found in metastatic patients. A conclusion that would be harder to explain is what is causing the mutations what types of mutations are affecting the non-metastatic patients. This proves more difficult because of the readily available data and the lack of marks in this graph shown here indicating that the gene that could be affecting proteins is another one entirely or if there is something on a transcriptomic level that is leaving scarce results. This can be supported by another article, “Differential mutation frequencies in metastatic cutaneous

squamous cell carcinomas versus primary tumors", written by Ayse Selen Yilmaz and others. In the following article, they also discuss skin cutaneous melanoma in length in which they also found similarities to my conclusion that these genes did not have any missense mutations in common with non-metastatic tumors. They state, "Two primary tumors (289199T and 387481T) for which we also sequenced a metastatic sample from the same individual did not have any mutations in common with the metastatic tumor, suggesting that they were not the cSCC giving rise to the metastasis" (Yilmaz et. al) . In this case, it shows that the two groups could hold different mutations at the same gene or there is a potential change in the following frequencies in other genes.

4) Cooccurence or mutual exclusion of common gene mutations : one for metastatic patients, one of non-metastatic patients



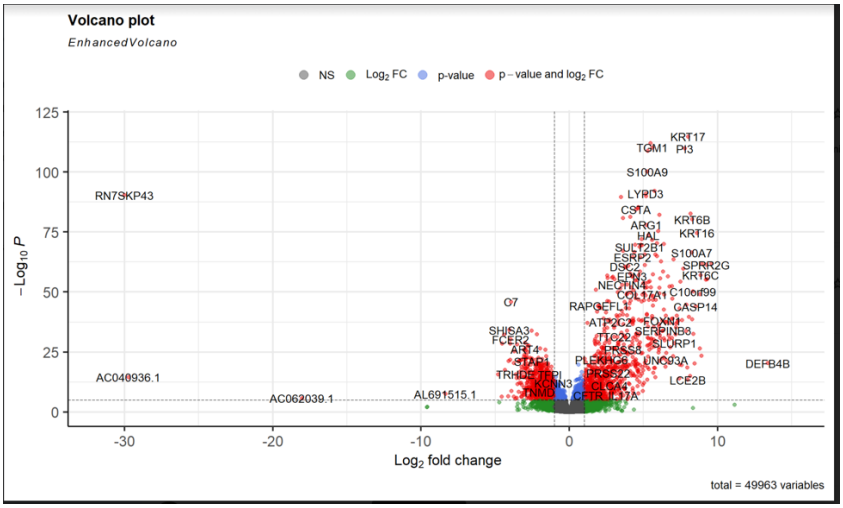
Left: Metastatic Patients

Right: Non-Metastatic Patients

In this -omic analysis, Somatic Interaction plots were made for both groups to see if genes when put against themselves of the top 20 if there would be intersection or if it would be mutually exclusive. This plot is showing that in Metastatic Patients most genes excluding BRAF are all co-occurrent with one another with a P-value of less than 0.05 meaning it is significant. There is an exception with BRAF in which there is actually mutually exclusive occurrences at SMDH2 and FAT3, and I/H?D1L1 . These are also different as they have a significant p value that is less than 0.01; this is a very high association. This could potentially tell us the type of mutations that this impacts the most, and in this case, we see on the left the genes impact one

another, so if there is a mutation it tends to impact the next one over. In the right side this is not really the case; however, we do see some co-occurrences with a p-value of 0.05; however, both BRAF are considered to show the least number of co-occurrences and it is more neutral or mutually exclusive with the other top 20 genes on both sides. This is supported in a review article discussing somatic interaction plots. In the article, "Genomic Landscape of the immunogenicity regulation in skin melanomas with diverse tumor mutation burden "written by George Georgoulias and Apostolos Zaravinos also did somatic mutation sin which they also saw mutual exclusivity in their BRAF genes (Georgoulias and Zaravinos). There was also contrasting conclusions that showed specific genes that were only co-occurrent, whereas most of mine for metastatic patients was co-occurrent.

5) Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status.



*This is not my graph. This is property of Wade Boohar and QBIO 490 class as my graph did not appear in my R Studio

For the final plot, this needed to be done using Differential Analysis and a Volcano Plot. Volcano Plots show a statistical significance and the magnitude of change as their axes. In the following graph, it shows that most of the data shown in our rna_counts data is significantly up regulated in metastatic patients compared to non-metastatic patients. There is more RNA for more of the following genes in metastatic than there is in non-metastatic patients. I can draw these conclusions because if we compare gene KRT17 in comparison the volcano plot it is significant upregulated when being compared to the volcano plot. A conclusion I cannot draw is if there is a consensus. Yes, there is a lot of clustering going upwards; however, it is possible to be unable to determine all the genes in the data set and prove if they are upregulated and it will prove hard to show which one. In an article, "Integrated Analysis to reveal potential therapeutic targets and prognostic biomarkers of skin cutaneous melanoma". This article is also studying SKCM data, they did general trends of IncRNA, miRNA, and mRNA. Both their IncRNA and mRNA plots shows a similar trend upward in which "there were 1457 DEIncRNAs (779 upregulated and 678 downregulated" (Zhou, et al). This conclusion supports mine because we

are also focusing on RNA and genetic transcriptomic data that shows a general trend with an skew of upregulated genes for metastatic versus non-metastatic patients.

Work Cited

Geng, Yi, et al. “Construction and Identification of an NLR-Associated Prognostic Signature Revealing the Heterogeneous Immune Response in Skin Cutaneous Melanoma.” Clinical, Cosmetic and Investigational Dermatology, vol. 16, 2023, pp. 1623–39, https://doi.org/10.2147/CCID.S410723.

Georgoulias, George, and Apostolos Zaravinos. “Genomic Landscape of the Immunogenicity Regulation in Skin Melanomas with Diverse Tumor Mutation Burden.” Frontiers in Immunology, vol. 13, 2022, pp. 1006665–1006665, https://doi.org/10.3389/fimmu.2022.1006665.

Liu, Xinhua, and Ju Wang. “The Nucleosome Remodeling and Deacetylase Complex Has Prognostic Significance and Associates with Immune Microenvironment in Skin Cutaneous Melanoma.” International Immunopharmacology, vol. 88, 2020, pp. 106887–106887, https://doi.org/10.1016/j.intimp.2020.106887.

Yilmaz, Ayse Selen, et al. “Differential Mutation Frequencies in Metastatic Cutaneous Squamous Cell Carcinomas Versus Primary Tumors.” *Cancer*, vol. 123, no. 7, 2017, pp. 1184–93, https://doi.org/10.1002/cncr.30459.

Zhou, Sitong, et al. “The Landscape of the Tumor Microenvironment in Skin Cutaneous Melanoma Reveals a Prognostic and Immunotherapeutically Relevant Gene Signature.” *Frontiers in Cell and Developmental Biology*, vol. 9, 2021, pp. 739594–739594, https://doi.org/10.3389/fcell.2021.739594.