QBIO 490: Directed Research - Multi-Omic Analysis

Fall 2023 R Review Project

Due: Wednesday, October 11th (11:59 pm). Submit your GitHub link to Blackboard, with all your code and code outputs in a folder called r_review_name within your qbio_490_name repo. The review question answers and final interpretations must also be turned into BlackBoard to check for plagiarism/AI. Please email extension requests (include the reason for your extension and a proposed new due date) to Wade and Kayla by Monday, October 9th 11:59 pm. This is a hard deadline, and no requests will be accepted after this date, except for reasons of emergency or illness.

Purpose:

This review project is meant to recap the analyses we've performed so far in R. It's also intended to rehash various parts of scientific writing and communication. For this project, please do your own work and submit your own written report, but you are more than encouraged to discuss ideas and debug code in groups! Note there are *three parts* to this assignment.

Overview:

So far, we have been working in R with TCGA data for breast cancer. In this assignment, you will be working with Skin Cutaneous Melanoma (SKCM) data. This assignment consists of three parts. In the first part, you will be answering short questions about R and TCGA. In the second part, you will be performing specific analysis of SKCM clinical, genomic, and transcriptomic data to explore a predetermined question about SKCM. In the third and final part, you will briefly write up your interpretations.

Part 1: Review Questions

General Concepts

1. What is TCGA and why is it important?

TCGA is "The Cancer Genome Atlas," a reservoir of 2.5 petabytes of data on biological and clinical markers for 11,000 patients with over 33 different types of tumors. This line of work is important as it elucidates DNA alterations that can better support translational and preventative medicine to address the functional

consequences of a cancerous cell. Ultimately, the large sample size

2. What are some strengths and weaknesses of TCGA?

Some of the strengths include the multifaceted nature of the data (e.g. different stratas of data, transcriptomics, genomics, proteomics, etc.) as well as the large sample size, giving it strong statistical significance. However, these are countered by the static nature of the data points (taken at truncated time intervals), similarly the ethnic and racial composition of the study is predominantly european, therefore skewing the results to more clearly depict european populations (and not those that are of Non-European descent).

Coding Skills

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

One must "git clone [insert github link to repository]" - this typically includes your github username as well as the depository.

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

First, one must make sure that the package is installed through a package such as Biocconductor (e.g. BiocManager::install("EnhancedVolcano")), and then you can invoke the package through the following format, "library ([insert package of interest])."

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

First, you must use "if (!requireNamespace("BiocManager", quietly = TRUE))

install.packages("BiocManager")", then "BiocManager::install("YourPackage")", finally utilize "library([insert package of interest])".

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

Boolean indexing serves to filter data sets via boolean statements (e.g. TRUE and FALSE) and the conditions that serve to isolate a desired variable(s). Furthermore, this logic may be applied to array sets—wherein you apply the same TRUE and FALSE logic to now elements of an array. Lastly, the applications of this can be conditionally assigning elements to an array or DataFrame; on the other hand, one can revise clinical DataFrames to isolate and identify particular elements such as "age" in the TCGA Data Set.

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

```
# This mock-up illustrates the nature of a TCGA data set
data = {
   'Sample_ID': ['TCGA-AB-1234', 'TCGA-CD-5678',
'TCGA-EF-9101'],
   'Patient_ID': ['AB-1234', 'CD-5678', 'EF-9101'],
   'Cancer_Type': ['Breast', 'Colon', 'Lung'],
   'Age': [32, 54, 61], }
```

a. An ifelse() statement

```
#ifelse() statement to pre-process boolean indexing for age, TRUE for 2 values
Data_mask <-ifelse(data$age =< 54, TRUE, FALSE)
```

b. boolean indexing

```
#Boolean indexing (only in rows) for two values
Filtered_data <- data[data_mask$age == 'TRUE', ]



#Displaying filtered data
print(Filtered_data)
```

Part 2: SKCM Analysis

Before starting your analysis, you may find it helpful to read the following review article on SKCM

to get a broad understanding of the cancer pathogenesis and possible treatment options. This may be especially helpful with understanding why each clinical variable was collected and what they mean. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3004577/

In this project, you will conduct multi-omic analyses to answer the following research question:

What are the differences between metastatic and non-metastatic SKCM across the genome and transcriptome?

To do this, you must include at least the following analyses (at least 6 plots):

1. Difference in survival between metastatic and non-metastatic patients
2. Mutation differences between metastatic and non-metastatic patients for multiple genes
3. Mutation differences for specific gene of interest (one of most mutated genes)
4. Coocurrence or mutual exclusion of common gene mutations: one for metastatic patients, one for non-metastatic patients
5. Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status
    a. Treatments must include radiation, chemotherapy, immunotherapy, molecular therapy, vaccine
    b. If you run this on CARC, it may take up to 4-5 hours, but you should not have to if your own personal machine has at least 8 GB available RAM

Hint: Each analysis lines up with a plotting method/analysis we've learned in class.

All of your code can be in a R Notebook or R script, which you will push to GitHub and provide a repo link to BlackBoard. As a part of the grading, we will check that your code runs with no errors starting from a clean environment. This means that while you can work with csv's during your drafting process, your final code should install and load all libraries and pull all dataframes from the TCGA data download. Remember to comment your code so other people can follow along.

Technical Tips:

The accession code for SKCM is TCGA-SKCM

The following commands can be used to access the drug and radiation dataframes once SKCM clinical data has been downloaded from TCGA:

rad <- clinical.BCRtab.all$clinical_radiation_skcm[-c(1,2),]

drug <- clinical.BCRtab.all$clinical_drug_skcm[-c(1,2),]

Metastasis status should be based on the rna_se@colData$definition column.

- Only consider "Metastatic" or "Primary solid Tumor" samples

When creating your maf_object using the read.maf command, it will make your life easier if you use rna_clinical (rna_se@colData) for the clinical information. To do this, all you need to do is run this command beforehand

rna_clinical$Tumor_Sample_Barcode <- rna_clinical$patient

Be careful about what "barcode" columns you use! The patient id, sample id, and sample barcode columns are all named slightly differently across the different dataframes. Double check that the columns you are using to match index values are correct!

- For DESeq2 data preprocessing:
  - Use the rna_se clinical data (rna_se@colData).

    Filter out genes with a total expression across all patients of <20

  - O Threshold padj values at 0.05 and log2FoldChange at |1|

Since there are 5 different treatments and each individual may have multiple treatments, you must use a technique called one-hot encoding where you create a column for each treatment and give a 1/0 value for whether each patient underwent that treatment.

- o For example:

Patient	Treatment	Patient	Radiation	Chemo	Immuno	Molecular

						Vaccine
1	Radiation, chemo	1	1	1	0	0	0
2	Radiation,	2	1	0	0	0	1
	Vaccine	3	0	0	1	0	0
3	Immunotherapy	4	0	0	0	0	0
4							

Non
e

Part 3: Results and Interpretations

For each analysis, include an image of the relevant plot you created in Part 2 and a 5-6 sentence description answering the following questions:

1. Describe the plot(s). What kind of plot is it? What is it showing?

The plots are KM, Co-Oncoplot, Co-Lollipop, Somatic Interaction, and Volcano. They are all included below—minus the volcano plot because my code wasn't able to output this.

2. What conclusions can you draw about differences between metastatic and non-metastatic TCGA SKCM patients? Why?

Deriving from the KM plot, I see that immediately there is a significant disparity between the survivorship probabilities with Metastatic patients and primary solid Tumor (non-metastatic) where the former demonstrated that there is a more conserved survival rate relative to the latter (non-metastatic) that indicated that there was a significant event that lead to a sharp decrease in survivorship. Moving forwards, there is a significance in the correlation noted in the graph because the p-value is less than 0.05. Moving forward, the co-oncoplot fails to accurately have axes labels, so I will avoid this analysis; similarly, the volcano plot failed to print, therefore, I will not include this analysis. For the co-lollipop graph, I see that for the metastatic patients there were 310 data points available relative to the 2 that were available for non-metastatic patients, prompting me

to assess that the comparison appears to be quite lopsided. However, evaluating solely the metastatic portion of the co-lollipop graph, there were a myriad of missense mutations in the MUC16 gene, which is fairly interesting. And lastly, the somatic interaction plot illustrated a co-occurrence between the genes, MUC16 and BRAF.

3. What's one conclusion you cannot draw? Why?

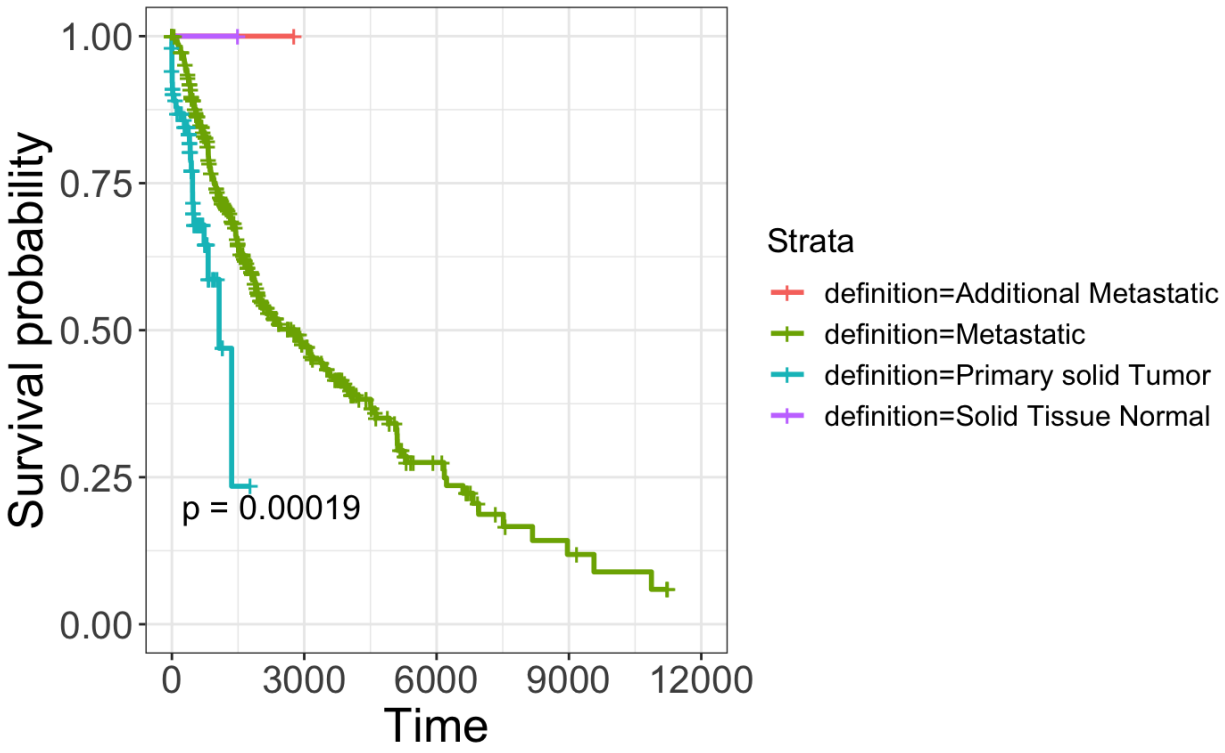
One conclusion I cannot draw is that non-metastatic patients on their MUC16 gene can only possibly have missense mutation(s). This can be attributed to the small sample size (n = 2), which lends a rather flimsy basis for critical genetic analysis.

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.

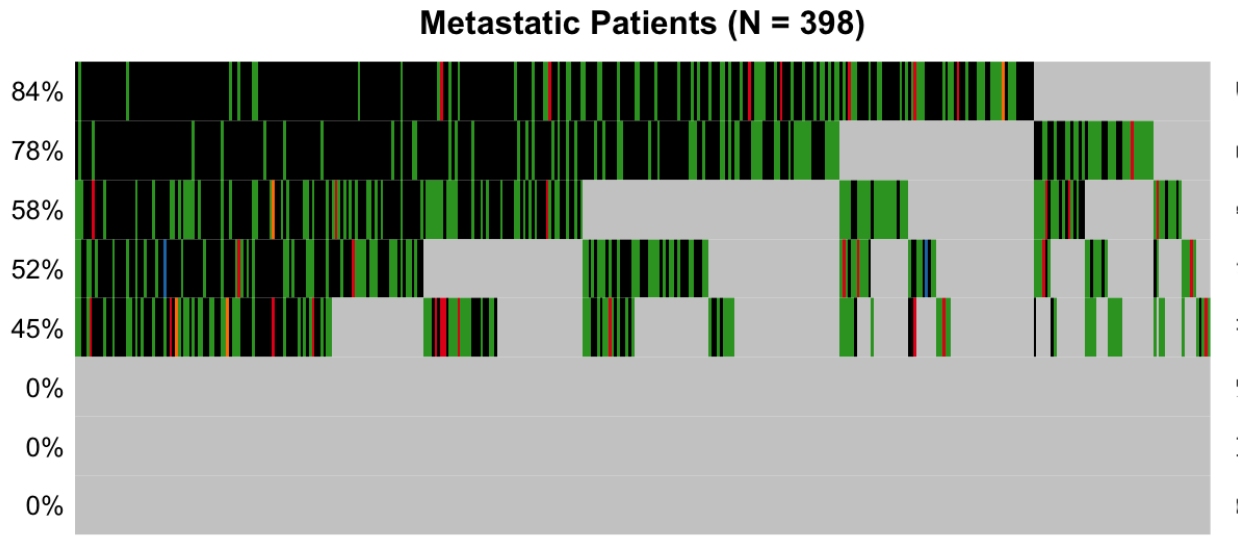
This article, https://pubmed.ncbi.nlm.nih.gov/36436416/, observes—and substantiates the claim above—that there are more than just missense mutations at the MUC16 gene, specifically there are "in-frame or frameshift insertions/deletions and splice-site or nonsense mutations" that are also observed in the MUC16 gene.

1) Difference in survival between metastatic and non-metastatic patients

KM plot



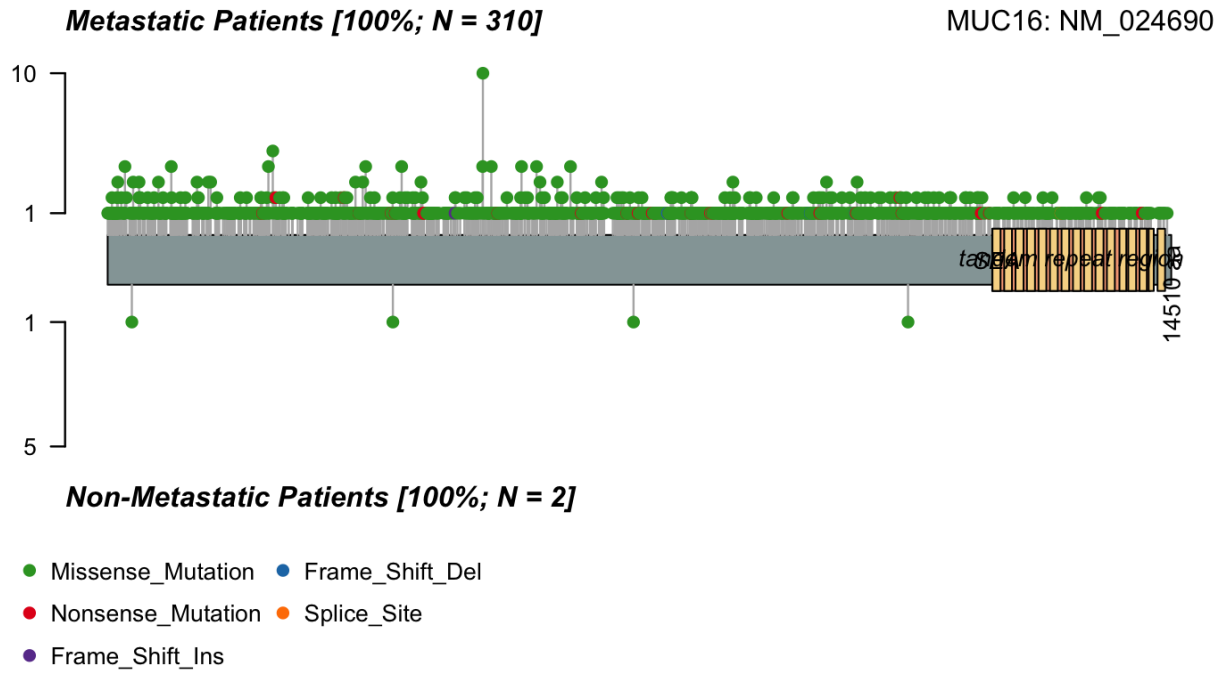
2) Mutation differences between metastatic and non-metastatic patients for multiple genes



Co-Oncoplot

3) Mutation differences for specific genes of interest

co-lollipop

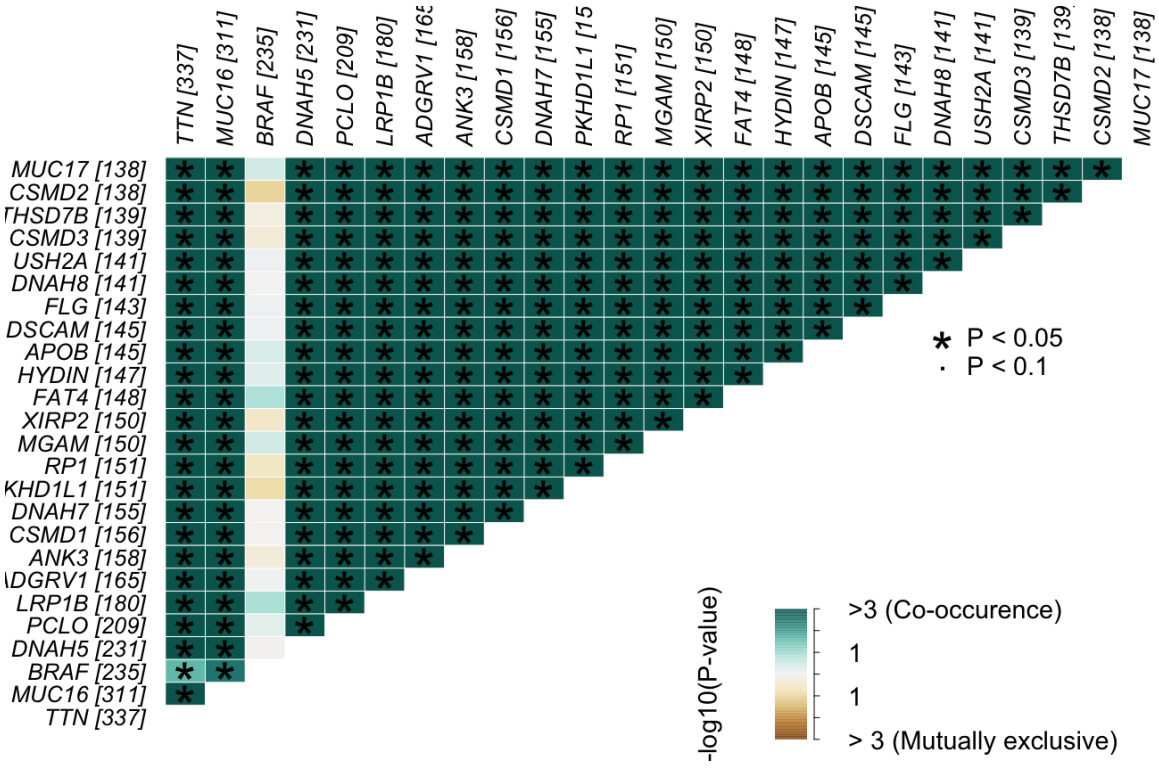


4) Coocurrence or mutual exclusion of common gene mutations: one for metastatic patients, one for non-metastatic patients

gene1 <chr>	**gene2** <chr>	**pValue** <dbl>	**oddsRatio** <dbl>	**00** <int>	**11** <int>	**01** <dbl>	**10** <dbl>	**pAdj** <dbl>	**Event** <chr>
MUC16	TTN	1.448571e-21	8.240837 2	88	269	68	42	3.353174e-20	Co_Occurence
DNAH7	CSMD1	4.172239e-18	6.184481 7	250	94	62	61	8.991895e-17	Co_Occurence
DNAH5	MUC16	7.946476e-18	6.174716 8	122	197	114	34	1.602112e-16	Co_Occurence
LRP1B	PCLO	1.388649e-17	5.469866 1	203	125	84	55	2.630017e-16	Co_Occurence

DNAH5	CSMD1	1.857201e-17	5.9810759	200	120	36	111	3.316430e-16	Co_Occurence
PCLO	MUC16	2.902391e-17	6.3389375	128	181	130	28	4.902688e-16	Co_Occurence
RP1	MUC16	9.923754e-17	8.7410724	143	138	173	13	1.590345e-15	Co_Occurence

Somatic Interaction plot (need 2 to show metastatic and the other)



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

Volcano plot (for DESeq2)

At the end of your report, include a References page of all the articles you used. Any citation format works, as long as you are consistent (all MLA, APA, etc.). Reminder: we are permitting the use of properly attributed AI work on the coding portion of this assignment (ie part 2), but not on any written portions (parts 1 and 3).