**QBIO 490: Directed Research - Multi-Omic Analysis**

**Fall 2024 Review Project**

**Due: Tuesday, November 19th (11:59 pm).** Submit your GitHub link to Brightspace, with all your code and code outputs in a folder called r\_review\_name within your qbio\_490\_name repo. Please email extension requests (include the reason for your extension and a proposed new due date) to Mahija and Wade by **Thursday, November 21st 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:**

In the first part, you will be answering short questions about R and TCGA. In the second part, you will choose one of two analyses of SKCM clinical, transcriptomic, and epigenomic 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?

The Cancer Genome Atlas is a free, open-access database containing genomic,

transcriptomic, epigenomic, and proteomic data across dozens of cancer types. This vast

set of data allows research to be conducted on numerous aspects of cancer, whether it be

identifying biomarkers or mutations that impact cancer progression.

1. What are some strengths and weaknesses of TCGA?

Strengths: TCGA is that anyone can use it, since it is open access, so its information can be utilized by scientists around the world. Weakness: as a public TCGA, all information is de-identified, which may lead to more NA values compared to private datasets. Private datasets might also have more funding and be cleaned more beforehand compared to TCGA.

## Coding Skills

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

cd, then first commit, then git status, git add, then git commit -m, then git push

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

Install package, then load package.

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

Install it: **if(!requireNamespace(“BiocManager”, quietly = TRUE))**

**Install.packages(“BIOcManager”)**

Then load+install+load: **library(BiocManager)**

**BiocManager::install(“TCGAbiolinks”)**

**library(TCGAbiolinks)**

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

Using logical statements to remove certain elements from a data fragment; can be used to remove na values, filter data on certain conditions such as age group.

1. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Gender | Age | GPA |
| Rudy | Female | 17 | 3.5 |
| Julie | Female | 15 | 2.1 |

* 1. an ifelse() statement

an ifelse() statement will execute a certain function if a certain condition is met. With this data frame, here is an example of checking whether a person is a teenager or not.

df$Age <- ifelse(df$Age > 15, “Not Teenager”, ifelse(df$Age <= 15, “Teenager”)

* 1. boolean indexing

Boolean indexing uses true/false to define a condtion which can be applied to subset a dataframe. For example, in this data you divide the data frame into good/bad students

good\_student <- df[df$GPA > 3.0]

# 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 explore the following research question:

**What are the differences between metastatic and non-metastatic SKCM across the epigenome and do these have any effect on the transcriptome?**

## Exploration of Methylation Patterns and Effect on Transcription

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 (KM plot)
2. Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status (DESeq2 + Volcano plot)
   1. Treatments must include radiation, chemotherapy, immunotherapy, molecular therapy, vaccine
   2. If you run this on CARC, it may take up to 1-2 hours
3. Naive differential methylation between non-metastatic and metastatic patients (Volcano plot)
4. Direct comparison of methylation status to transcriptional activity across non-metastatic vs metastatic patients
5. Visualization of CpG sites and protein domains for 3 genes for a few genes (use UCSC genome browser)

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 Brightspace. As a part of the grading, we will check that your code runs with no errors starting from a clean environment. However, you can assume that any of the csv’s we save in class are present (brca\_clinical\_data, brca\_rna\_clinical, brca\_rna\_genes, brca\_rna\_counts, brca\_methylation\_clinical, brca\_methylation\_betas, and brca\_cpg\_sites). 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.
  1. Only consider “Metastatic” or “Primary solid Tumor” samples
* 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:
  1. Use the rna\_se clinical data (rna\_se@colData).

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

○ 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.
  1. For example:

# =>

## **Part 3: Results and Interpretations**

For each analysis, include an image of the relevant plot you created in Part 2 and a 3-4 sentence description answering the following question:

● Analyze the plot. What conclusions can you and can you not draw about differences between metastatic and non-metastatic TCGA SKCM patients? Why?

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

A graph with a red line

Description automatically generated

The Kaplan-Meier plot shows differences in survival probabilities between metastatic and non-metastatic SKCM patients, with metastatic patients showing higher survival probabilities. The primary solid tumor patients show a sharp decline in survival. The p-value of less than 0.001 shows that these differences are statistically significant. The plot doesn’t reveal causation, because there’s other factors that could may influence what is shown.

2 ) Expression differences between metastatic and non-metastatic patients

A graph with red dots

Description automatically generated

The volcano plot shows differentially expressed genes between metastatic and non-metastatic SKCM patients. It shows that more genes are upregulated in metastatic patients compared to non-metastatic patients; significant genes have a p-value threshold of <0.05 and a log2 fold change >|1|. This data shows some association between gene expression changes and metastatic status, but it doesn’t show causation, because, again, there are other factors that may influence this data.

3 ) Methylation differences between metastatic and non-metastatic patients

A graph showing a diagram

Description automatically generated with medium confidence

The volcano plot compares gene expression differences between metastatic and non-metastatic SKCM patients, again showing significantly upregulated and downregulated genes. Genes in the upper-left and upper-right sides are larger than the thresholds for statistical significance and fold change; this suggests differential expression. More analysis is required to confirm the findings in this plot because again, this plot does not show causation, only correlation.

4 ) Direct comparison of transcriptional activity to methylation status for 10 genes

I tried many times to fix my code to get this to work, but I could not figure it out. I beg you, please go easy me ☹ I promise I tried my best. Mahija knows I was struggling in general too! :’((

**5) Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser) for a few genes. Describe at least one academic article (research or review) that either supports or doesn’t support your final conclusion for one of the genes. If previously published work doesn’t support your analysis, explain why this might be the case.**

SCTR: GDA: A screenshot of a computer

Description automatically generatedEDN3: A screenshot of a computer

Description automatically generated

In a 2021 paper, EDN3 was identified as part of a novel immune gene signature associated with the tumor microenvironment (TME) in SKCM. This gene signature, which included EDN3, was linked to varying prognoses and responses to immunotherapy in SKCM patients. The gene is a potential biomarker for predicting treatment outcomes. EDN3 showed high expression levels in SKCM tissues (Sun, 2021). Though, a larger sample size is needed to confirm its prognostic significance.

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).

References

Zhou, S., Sun, Y., Chen, T., Wang, J., He, J., Lyu, J., Shen, Y., Chen, X., & Yang, R. (2021). The Landscape of the Tumor Microenvironment in Skin Cutaneous Melanoma Reveals a Prognostic and Immunotherapeutically Relevant Gene Signature. *Frontiers in Cell and Developmental Biology*, *9*, 739594–739594. https://doi.org/10.3389/fcell.2021.739594