**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?

TCGA stands for The Cancer Genome Atlas, and it’s a database of multi-omic data collected from over 20,000 samples of 33 different types of cancer. It’s important because it has led us to get a better understanding of different types of cancer on a molecular level. It helps lead us into the prevention and treatment of cancer.

1. What are some strengths and weaknesses of TCGA?

Strengths: publicly accessible on the internet, collaborative

Weaknesses: only contains data samples of 33 different types of cancer when there are over 200 types of cancer

Coding Skills

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

git add

git status

git commit -m “”

git push

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

Install.packages(“name\_of\_package”)

Library(name\_of\_package)

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

Install.packages(“BiocManager”)

BiocManager::install(“package\_name”)

Library(package\_name)

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

It’s a technique used to select or filter data from a vector or dataframe using a Boolean value. It’s used to create masks which these masks help clean out data.

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.
   1. an ifelse() statement
   2. boolean indexing

students dataframe

|  |  |  |
| --- | --- | --- |
| STUDENT ID # | Age | Major |
| 1 | 19 | QBIO |
| 2 | 19 | BUSINESS |
| 3 | 20 | FILM |
| 4 | 21 | FILM |

Age\_mask <- ifelse(students$age < 20, T, F)

Ag mask = [TRUE TRUE FALSE FALSE]

Students <- students[, Age\_mask]

students dataframe

|  |  |  |
| --- | --- | --- |
| STUDENT ID # | Age | Major |
| 1 | 19 | QBIO |
| 2 | 19 | BUSINESS |

# 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):

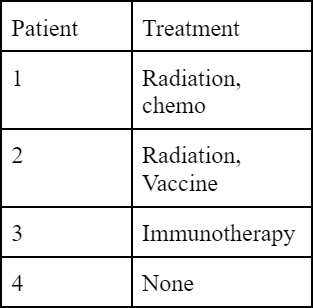
1. Difference in survival between metastatic and non-metastatic patients (KM plot)
2. Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status (DESeq2 + Volcano plot)
   1. Treatments must include radiation, chemotherapy, 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)

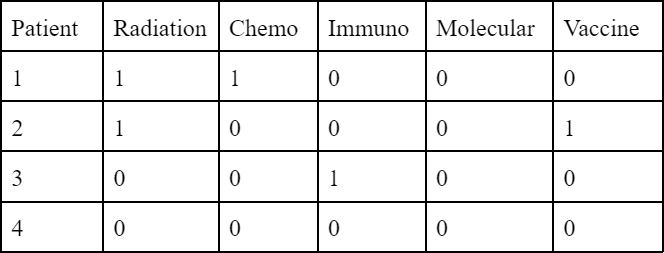
All of your code can be in a R Notebook or R script, which you will push to GitHub and provide a repo link to Brightspace. As a part of the grading, we will check that your code runs with no errors starting from a clean environment. However, you can assume that any of the csv’s we save in class are present (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.
  + 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:
  + 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.
  + 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 survival plot illustrates a significant difference in survival probabilities between metastatic and non-metastatic SKCM patients (p < 0.0001). Patients with metastatic melanoma show lower survival probabilities over time compared to non-metastatic patients. This highlights the poorer prognosis associated with metastatic disease. However, this analysis does not account for treatment or patient age, which could influence survival outcomes.

1. **) Expression differences between metastatic and non-metastatic patients**

**A graph with numbers and a blue line

Description automatically generated with medium confidence**

The volcano plot has limited differences in gene expression between metastatic and non-metastatic SKCM patients, with only a small subset of genes showing significant fold changes and p-values. Most genes cluster around the origin, suggesting overall similarity in gene expression. While the highlighted genes may serve as potential biomarkers for metastasis, the plot does not establish causation or provide pathway-level insights.

1. **) Methylation differences between metastatic and non-metastatic patients**

**A graph with blue dots and red lines

Description automatically generated**

This volcano plot displays the differential methylation analysis between metastatic and non-metastatic SKCM patients. The x-axis shows the fold change in methylation levels, and the y-axis indicates the -log₁₀(p-value), representing statistical significance. Points outside of the red dashed thresholds suggest significant changes in methylation, with higher significance in those farther from the origin. The symmetric distribution of points suggests a mix of hypermethylated and hypomethylated sites, though the dots past the thresholds are few compared to the others (gray). These significant sites may be key in understanding epigenetic differences between the groups but require further analysis.

1. **) Direct comparison of transcriptional activity to methylation status for 10 genes**

**A graph with red and blue lines

Description automatically generated** A graph of a group and metastatic

Description automatically generated

Bar Plot: The bar plot compares the average beta values of CpG sites for metastatic and non-metastatic SKCM patients with ALX1 gene. Blue and red bars represent the two patient groups, respectively. From this plot, it can be inferred that certain CpG sites exhibit markedly higher methylation levels in one group compared to the other. However, this analysis does not confirm whether these differences are statistically significant.

Box Plot: The box plot compares transcriptomics count distributions for metastatic and non-metastatic SKCM patients. Both groups show similar medians suggesting there is not really a significant difference in overall gene expression counts between them.

1. **) 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.**

**A screenshot of a computer

Description automatically generated**

I’m not sure if I did this part right, the genome browser is not user friendly ToT.

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:

“The Cancer Genome Atlas Program (TCGA).” *NCI*, www.cancer.gov/ccg/research/genome-sequencing/tcga#:~:text=The%20Cancer%20Genome%20Atlas%20(TCGA,samples%20spanning%2033%20cancer%20types. Accessed 3 Dec. 2024.