Kiley Huffman
QBIO 490x
Fall 2024
Due: 11/19/2024

QBIO 490: Directed Research - Multi-Omic Analysis

Fall 2024 Review Project

Part 1: Review Questions

General Concepts

1. **What is TCGA and why is it important?**

   TCGA is "The Cancer Genome Atlas." It is a large research project aimed at cataloging and understanding the genetic mutations responsible for cancer. The TCGA project is important because it increases our understanding of cancer at the molecular level, which helps scientists develop targeted, personalized, and effective cancer treatments.

2. **What are some strengths and weaknesses of TCGA?**

   Some strengths of TCGA are that it provides comprehensive, multi-omic views of cancer, it has led to discoveries about the foundations of cancer, and has contributed to the development of targeted and personal therapies. Some weaknesses of TCGA are that it has biases in sample representation, gaps in clinical data, and has limitations in integrating diverse data types.

Coding Skills

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

   In terminal (Bash/Linux):

   cd Users/kileyhuffman/Desktop/qbio_490_kileyhuffman

   git status

   git add (filename)

   git status

   git commit -m 'message'

   git push

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

   install.packages("package_name")

   library(package_name)

3. **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)

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

   Boolean indexing uses an array (or vector) of True and False values (Boolean values) to filter data. Some applications of boolean indexing are to filter rows or columns based on conditions, to handle missing data, to perform time-series filtering, and to perform complex conditional filtering.

5. **Draw a mock up (just a few rows and columns) of a sample dataframe.**

Sample Dataframe:

Name	Age	Score
Anna	21	88
Natalia	22	72
Mario	20	64
Hunter	19	60

Show an example of the following and explain what each line of code does.

Assuming the DataFrame has been loaded as “df” in R:

a. **an ifelse() statement**

   df$Result <- ifelse(df$Score >= 70, "Pass", "Fail")

This ifelse() statement creates a new column in the dataframe labeled “Result” which contains values of “Pass” or “Fail”. The condition being tested is “df$Score >= 70”. If this condition is true for a row, the row’s value in the “Result” column will be “Pass”. If this condition is false for a row, the row’s value in the “Result” column will be “Fail”.

b. boolean indexing

filtered_df <- df[df$Score >= 80,]

The condition being tested is “df$Score >= 80”. This creates a logical vector of True and False values, where each value corresponds to whether the score is greater than or equal to 80. The logical vector is then used to filter the rows in the DataFrame. Only rows where the condition is True (Score >= 80) will be selected and included in filtered_df.

Part 2: SKCM Analysis

Go to **r_review_kiley/part2_code.rmd** on my GitHub, linked below:

https://github.com/kehuffma/qbio_490_kileyhuffman.git

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:

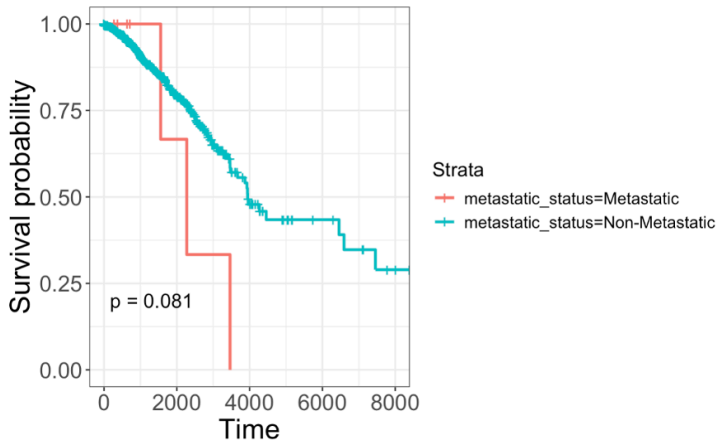

Figure 1. The Kaplan Meier plot of the difference in survival between metastatic and non-metastatic patients.

Based on the plot, there is evidence that non-metastatic patients survive for a longer amount of time than metastatic patients. Near the 0 day time mark, metastatic and non metastatic patients both start at a survival probability near 1. By the time the x axis approaches 4000 days, the survival probability for metastatic patients is less than 0.25, while the non-metastatic patient survival rate is around 0.60. Hence, patients that are non-metastatic are more likely to survive than the metastatic patients. We cannot, however, tell why these differences occur from the plot.

2) Expression differences between metastatic and non-metastatic patients:

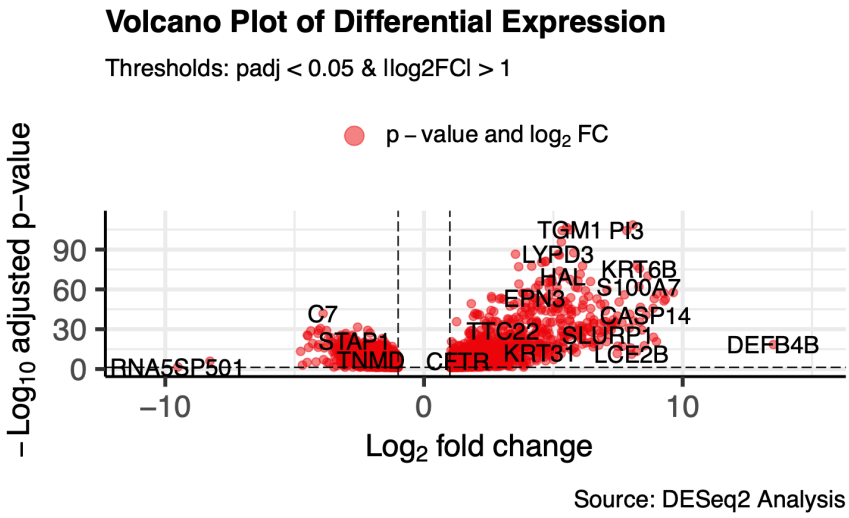

Figure 2. The Volcano Plot of the Differential Expression Analysis between Metastatic and Non-Metastatic patients.

This plot suggests that genes such as TGM1, KRT6B, and LYPD3 are upregulated. It also suggests that genes such as TNMD, STAP1, and C7 are downregulated. The position of these genes on the plots illustrates a trend in patients who have melanoma, thus they may serve as potential biomarkers for the cancer. The plot, however, does not establish any casual relationship, thus further study must be done to link these genes to melanoma.

3) Methylation differences between metastatic and non-metastatic patients:

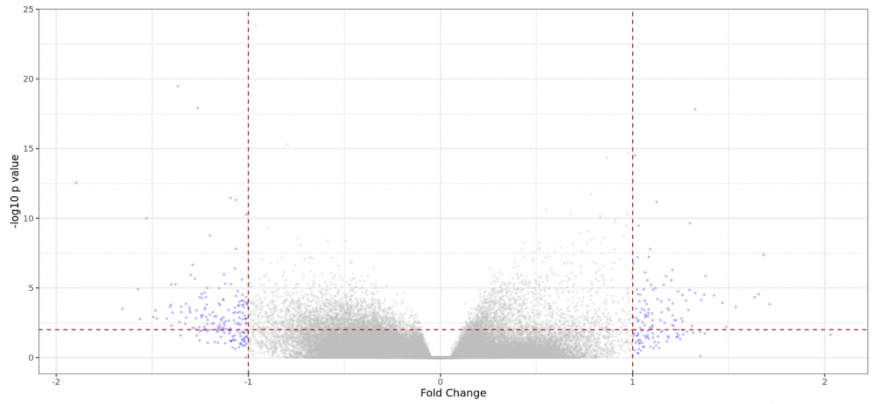

Figure 3. The volcano plot for the methylation differences between metastatic and non-metastatic patients.

Based on the plot, there is evidence that the genes on the right side are more methylated in metastatic patients than in non-metastatic patients. In contrast, the genes on the left are more methylated in non-metastatic patients. There are a few outliers on the plot, indicating more experiments may be needed to examine how those specific genes differ in metastatic versus non-metastatic patients.

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

N/A…I wrote all of the code for this part of the assignment but I was unable to get R to process it to produce a visual. If R had been able to process the code, I believe these are the conclusions I would be able to draw:

The visual would likely show that the 10 genes have a wider range and more outliers in patients with metastatic versus non-metastatic status. This may or may not be related to metastasis or cancer progression. Moreover, it is likely that methylation patterns vary a lot from each gene. A more comprehensive study would likely examine each gene individually to determine the difference between the two groups.

5) Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser):

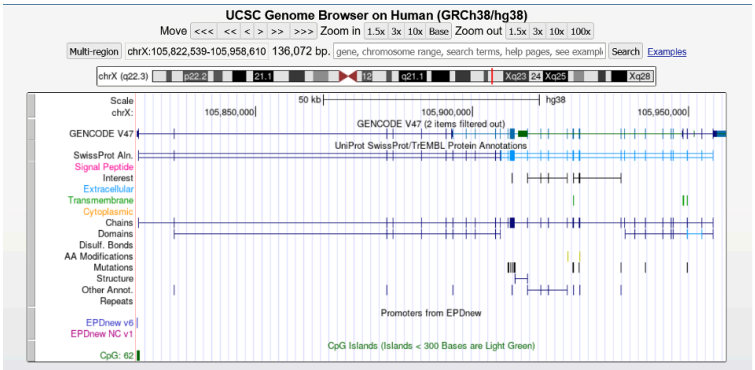

Figure 5.1 The visualization of CpG sites and protein domains for the NRK gene.

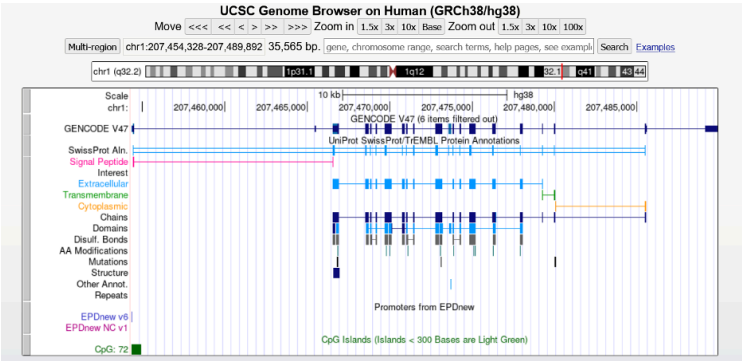

Figure 5.2 The visualization of CpG sites and protein domains for the CRK gene.

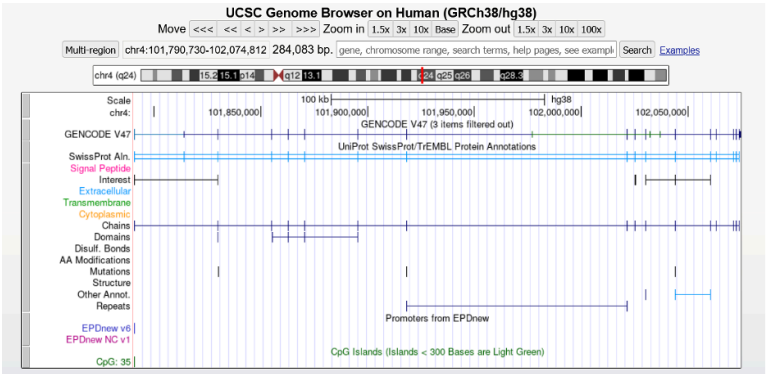

Figure 5.3 The visualization of CpG sites and protein domains for the BANK1 gene.

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:

Gene: NRK

The article, "The role of nucleoside diphosphate kinases in tumorigenesis and their therapeutic potential", corroborates the idea that NRK genes are often overexpressed in melanoma cells, contributing to tumor growth and metastasis. The visualization above for NRK also shows that NRK genes are present in melanoma cells. The article comes to the conclusion that the overexpression of NRK may contribute to the survival of melanoma cells, making it a potential therapeutic target. This supports my final conclusion for the NRK gene.

References

Liu, Y., & Zhang, X. (2017). The role of nucleoside diphosphate kinases in tumorigenesis and their therapeutic potential. *Frontiers in Oncology, 7*, 92. https://doi.org/10.3389/fonc.2017.00092