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 publicly available multi-omic data set organized by the National Cancer Institute and the National Human Genome Research Institute. It allows for the exploration of a wide range of genes across a large patient sample.

2. What are some strengths and weaknesses of TCGA?

TCGA has comprehensive and high-quality clinical data. There are over 20,000 samples of 33 cancer types that are all easily accessible. The diversity of patients is not great, but it is better than other data sets.

Coding Skills

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

    git add (file) -> git commit -m (message) -> git push

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

    install.packages("Package")
    library(Package)

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

    Example for clinical data:

    clin_query <- GDCquery(project = <ACCESSION CODE>, data.category = " Clinical", data.type = " Clinical Supplement", data.format = 'BCR Biotab')

    GDCdownload(clin_query)

    clinical.BCRtab.all <- GDCprepare(clin_query)

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

    Boolean indexing applies a vector of booleans (T/F) to a column/row in a dataframe. A boolean mask can be used to select for certain data, rewrite it into a new dataframe, or overwrite the existing dataframe. Boolean indexing can be used to delete null data or subset the data with a clinical variable (young/old, male/female).

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.

patient_barcodes	age
0001	10
0002	30
0003	50
0004	60
0005	90

a. an ifelse() statement

df$age_category <- ifelse(df$age < 30, ‘young’, ifelse(df$age > 60, ‘old’, ‘middle’))

This nested ifelse statement creates 3 separate age categories based on the age column in the dataframe. The output is set to a new column in the dataframe.

b. boolean indexing

age_mask <- ifelse(df$age_category == “young”, T, F)
df$patient_barcodes[age_mask]

The boolean vector mask will list TRUE for all patients in the young age category. The mask is then applied to overwrite the existing data frame to only include young patients.

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)
    a. Treatments must include radiation, chemotherapy, immunotherapy, molecular therapy, vaccine
    b. 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:

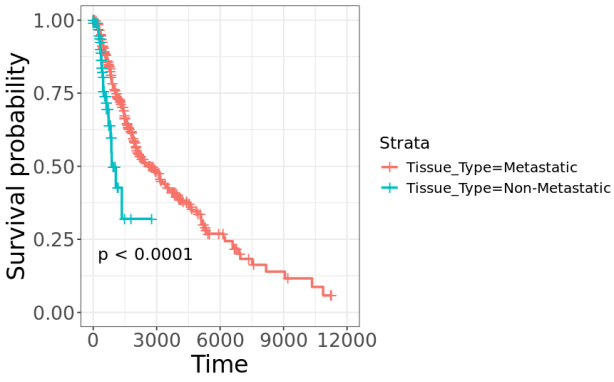
Patient	Treatment
1	Radiation, chemo
2	Radiation, Vaccine
3	Immunotherapy
4	None

Patient	Radiation	Chemo	Immuno	Molecular	Vaccine
1	1	1	0	0	0
2	1	0	0	0	1
3	0	0	1	0	0
4	0	0	0	0	0

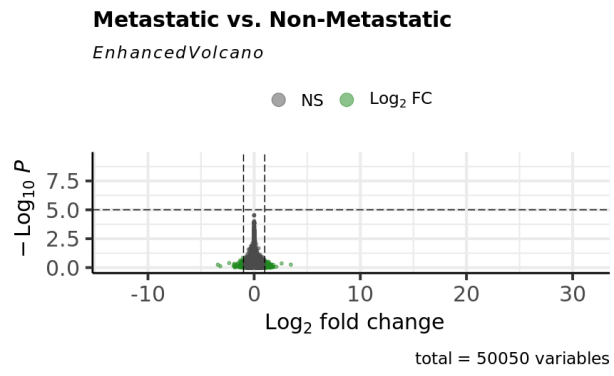
Part 3: Results and Interpretations

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?

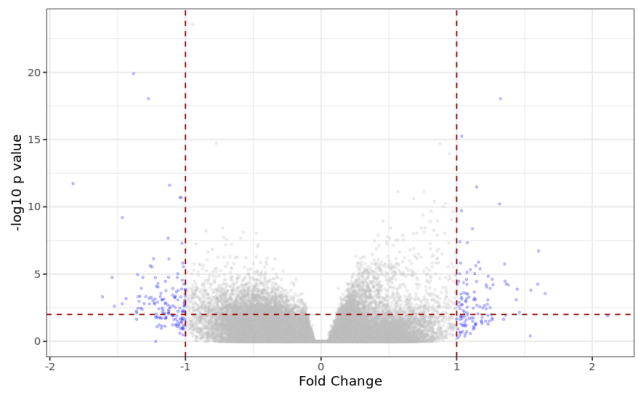
1. Difference in survival between metastatic and non-metastatic patients

	With a high degree of probability, this plot seems to indicate that non-metastatic skin cancer patients have lower survivability than metastatic skin cancer patients. However, this does not align with our current understanding of metastasis. This may be a misleading KM plot where non-metastatic patients are less likely to follow-up unless they experience complications.

2. Expression differences between metastatic and non-metastatic patients

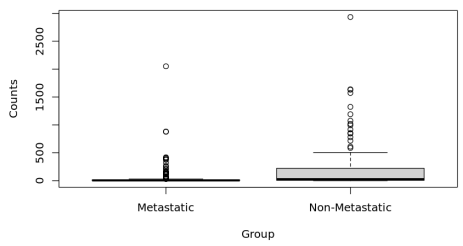
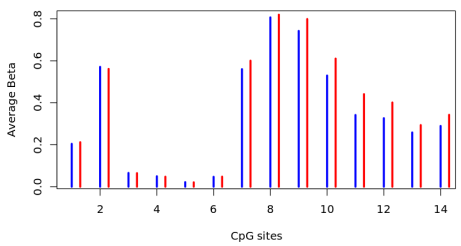
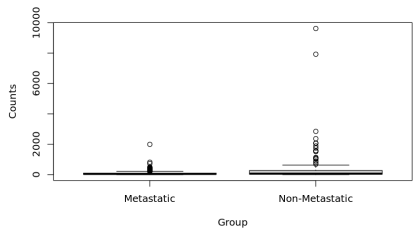
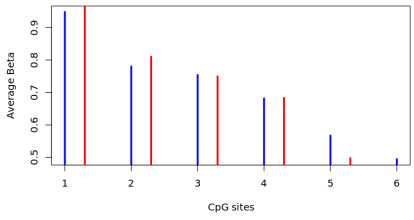
	There are some genes that are up and downregulated in metastatic patients, but none with a probability great enough to be statistically significant. This would imply that metastasis is controlled on an epigenomic, rather than a transcriptomic, level.

3. Methylation differences between metastatic and non-metastatic patients

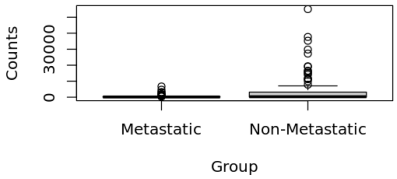
	All CpG sites hypermethylated in metastatic patients appear to the right of the dotted line and all those undermethylated in metastatic patients appear to the left of the dotted line (in comparison to non-metastatic patients). There are relatively few genes that are hypermethylated and undermethylated in metastatic patients, indicating that methylation does not significantly differ between the two groups.

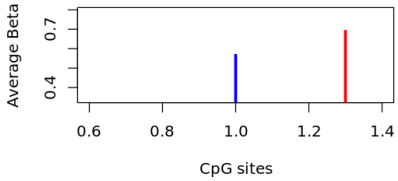
4. Direct comparison of transcriptional activity to methylation status for 10 genes

The following 10 genes are downregulated and hypermethylated (top 1% of methylated genes) in metastatic patients when compared to non-metastatic patients.

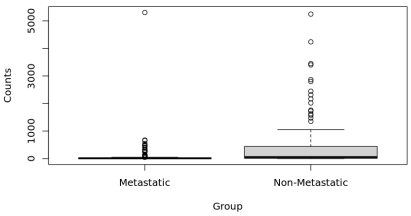
AP1M2 - upregulated in non-metastatic patients, around half of the CpG sites are more methylated in metastatic patients	
	
CLIC3 - neither upregulated in metastatic or non-metastatic patients, few CpG sites with slightly varying methylation	
	

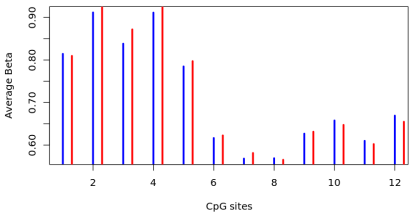
CSTA - slightly upregulated in non-metastatic patients, only 1 CpG site in both patient types



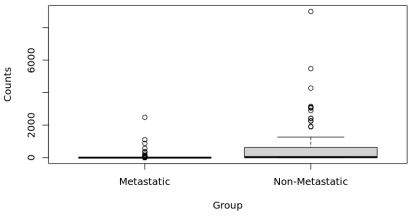


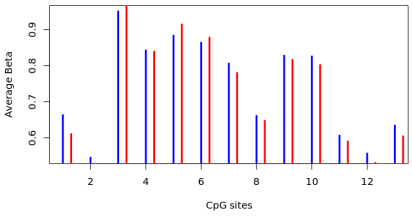
ESRP2 - upregulated in non-metastatic patients, CpG sites with slightly varying methylation



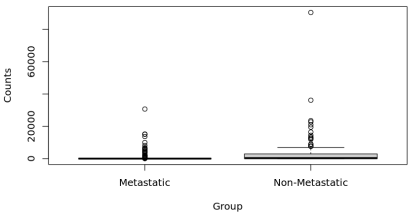


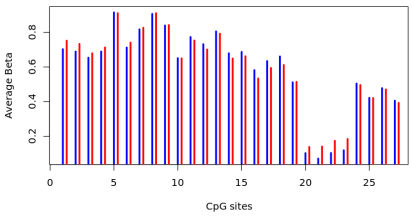
FAM83C - upregulated in non-metastatic patients, CpG sites with slightly varying methylation





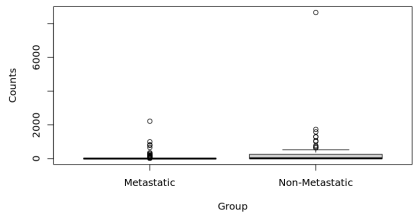
GJB6 - slightly upregulated in non-metastatic patients, many CpG sites with slightly varying methylation

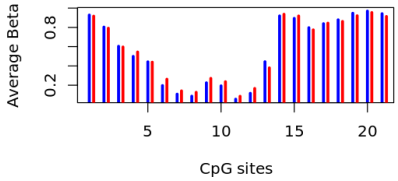




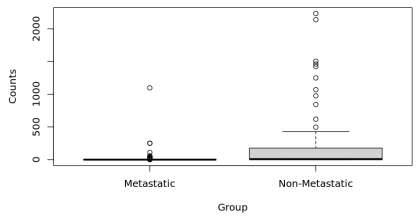
OVOL1 - neither upregulated in metastatic or non-metastatic patients, many CpG sites with

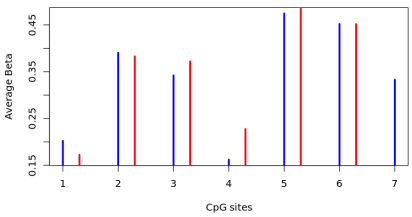
slightly varying methylation



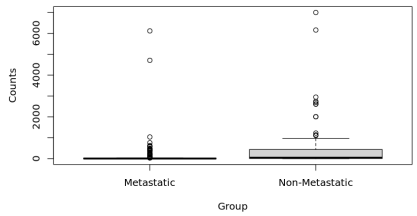


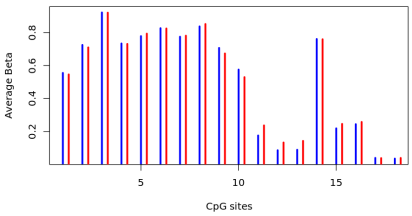
PGLYRP4 - significantly upregulated in metastatic patients, CpG site 4 is significantly more methylated in metastatic patients while CpG site 7 has no methylation (or does not exist)



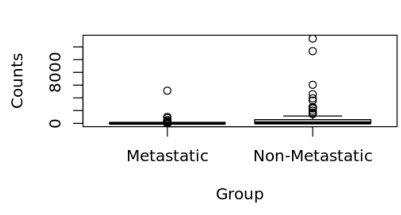


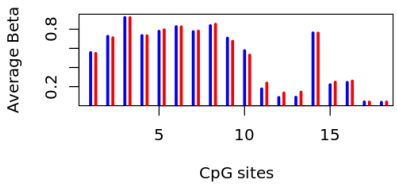
PRSS8 - significantly upregulated in non-metastatic patients, many CpG sites with slightly varying methylation





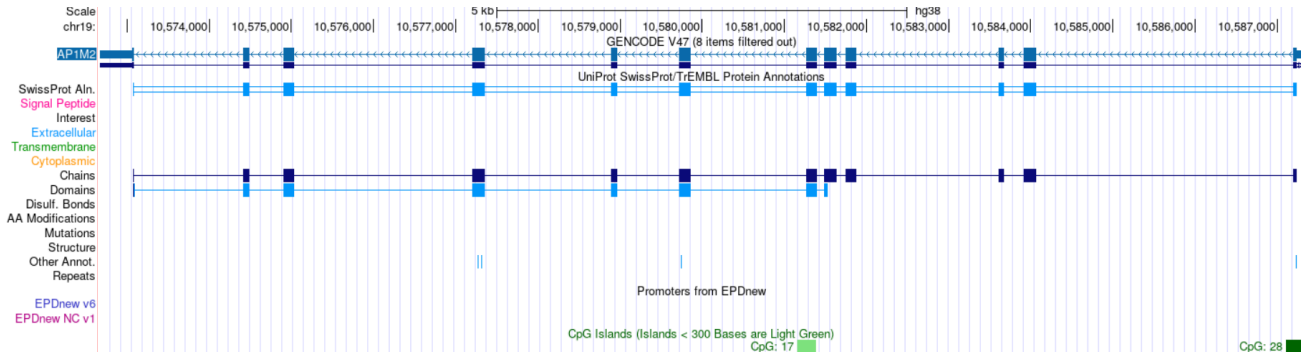
SULT2B1 - slightly upregulated in non-metastatic patients, many CpG sites with slightly varying methylation





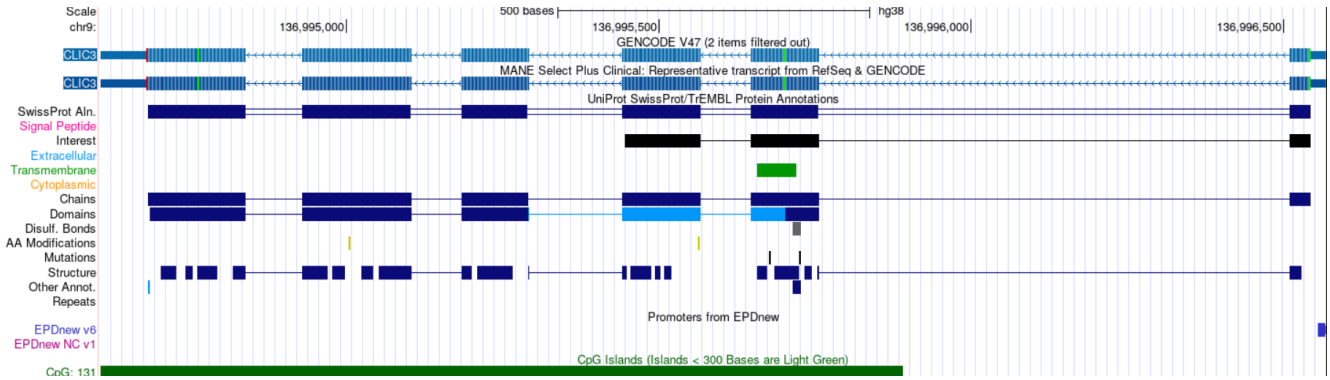
5. Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser) for a few genes. Describe at least one academic article (research or review) that either supports or doesn't support your final conclusion for one of the genes. If previously published work doesn't support your analysis,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

AP1M2 - 2 CpG islands with no promoter

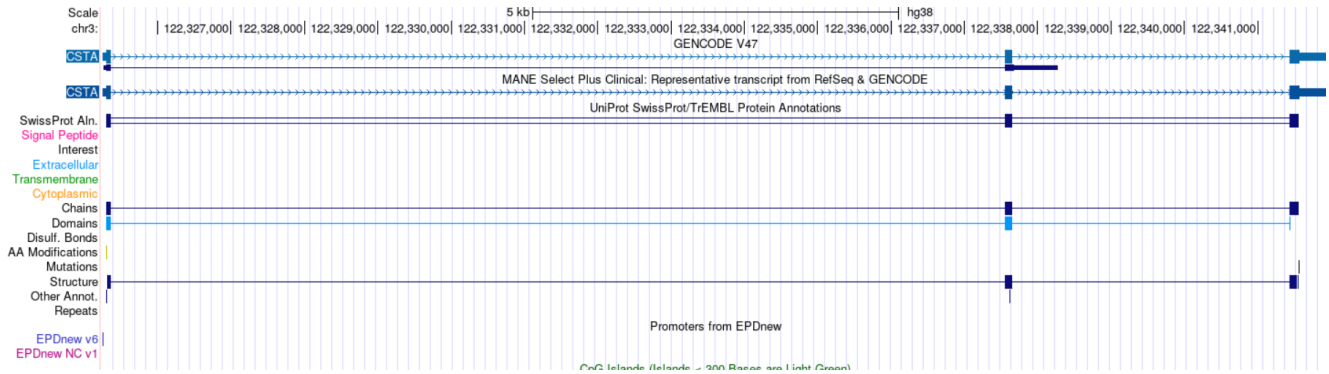


A 2022 pan-cancer system analysis on AP1M2 in malignant tumors found that "AP1M2 expression levels in most tumors influenced the activation of tumor-associated pathways and immune-associated pathways." Although mainly focusing on breast cancer, the study found upregulation of AP1M2 was tied to immune infiltration in most cancer types, including T cells, macrophages, neutrophils, and dendritic cells (Yi et al., 2022).

CLIC3 - 1 large CpG island with a non-CpG site promoter



CSTA - No CpG islands or promoter



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).

Yi, Y., Zhang, Q., Shen, Y., Gao, Y., Fan, X., Chen, S., Ye, X., Xu, J., Wang, F., & Fu Wang. (2022). System Analysis of Adaptor-Related Protein Complex 1 Subunit Mu 2 (AP1M2) on Malignant Tumors: A Pan-Cancer Analysis. *Journal of Oncology*, *2022*, 7945077–17. https://doi.org/10.1155/2022/7945077