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, it was launched in 2005 by the National Cancer Institute (NCI) and it is by far the largest public database of genomic, transcriptomic, epigenomic, and proteomic data for more than 30 cancer types. Most of the deidentified TCGA data is open for access for everyone around the world, therefore allowing students and researchers everywhere to study and gain new insights into the identification of potential biomarkers/mechanisms of the diseases.

2. What are some strengths and weaknesses of TCGA?

Strengths:

a. Data Scale – includes data for 30+ Cancer types across multiple omics
b. Accessibility – free and open access for researchers around the world

Weaknesses:

a. Limited Patient Demographic – mostly North American patients from a social economic status where they can receive good to high quality healthcare. Therefore the findings from TCGA might not be indictive to the greater global patient population.

b. Data quality is not uniform across different cancer type, and portions of clinical/treatment information are often missing from the patients, therefore demanding rigorous data pre-processing from the user/researcher's side.

c. Stopped in 2018 – does include current data.

Coding Skills

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

```
git add . / git add filename
git commit -m "message"
git push
```

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

The package must be installed and loaded into the R session.
Example:

```
install.packages("survival")
library(survival)
```

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

```
install.packages("BiocManager") #install BiocManager from bioconductor
BiocManager::install("DESeq2") #install DESeq2 from bioconductor
library(DESeq2) #load the package
```

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

Boolean indexing involves using arrays of boolean values (true or false) to select elements from data structures such as arrays, data frames, or lists. It's commonly used for filtering data, extracting subsets that meet certain conditions, and performing conditional operations without the need for explicit looping.

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

```
> print(sample)
```

patient_id	age
1	12
2	35
3	76
4	28
5	19
6	64
7	41

a. an `ifelse()` statement

```
sample$age_bracket <- if sample$age > 60, “Old”, “Young”)

#add a new age_bracket column to dataframe sample, if the
patient is over 60-years old, they are categorized as old, and
if they are less than 60-years old, they are catergorized as
young
```

b. boolean indexing

```
middle_age <- sample[sample$age > 30 & sample$age < 60,]


#storing patients from 31 to 59 in a new dataframe called
middle_age
```

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):

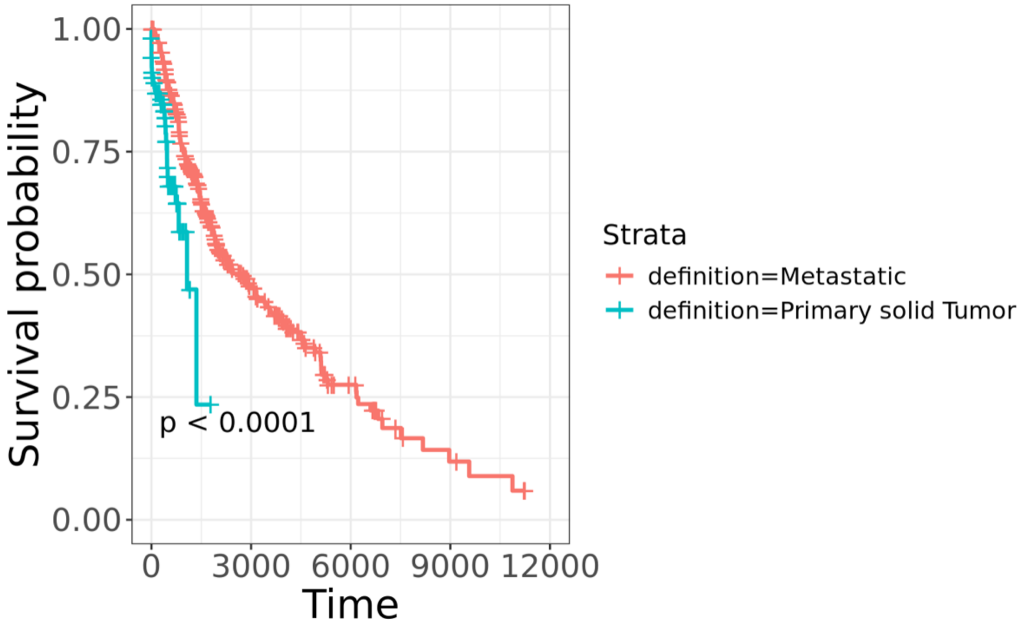
1. Difference in survival between metastatic and non-metastatic patients (KM plot)
2. Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status (DESeq2 + Volcano plot)
    a. Treatments must include radiation, chemotherapy,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

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:

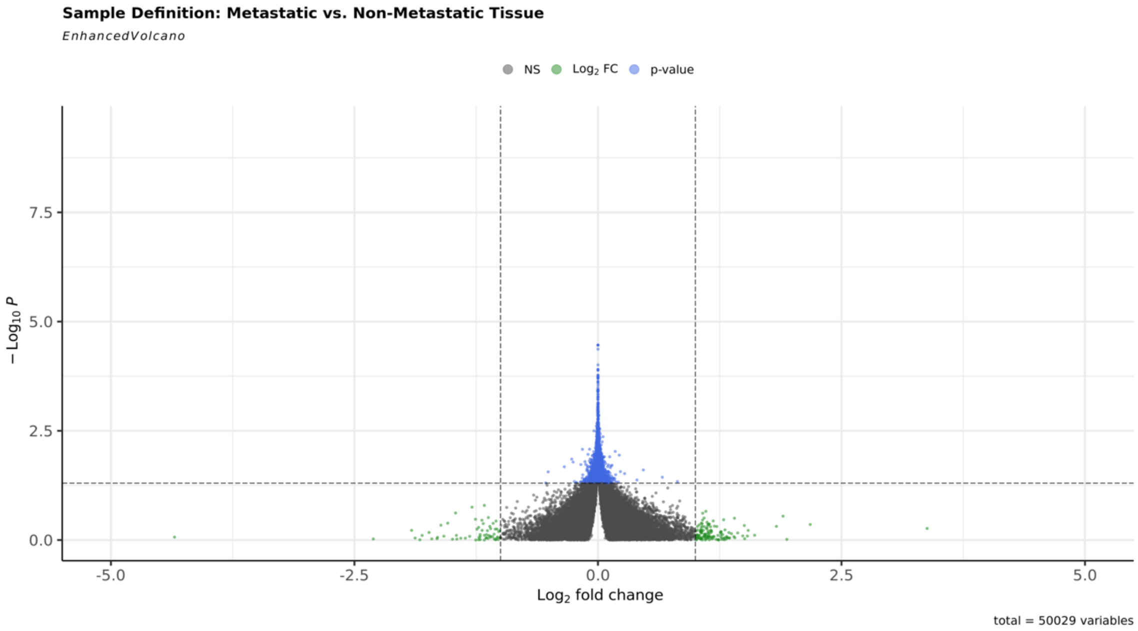
- Analyze the plot. What conclusions can you and can you not draw about differences between metastatic and non-metastatic TCGA SKCM patients? Why?

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



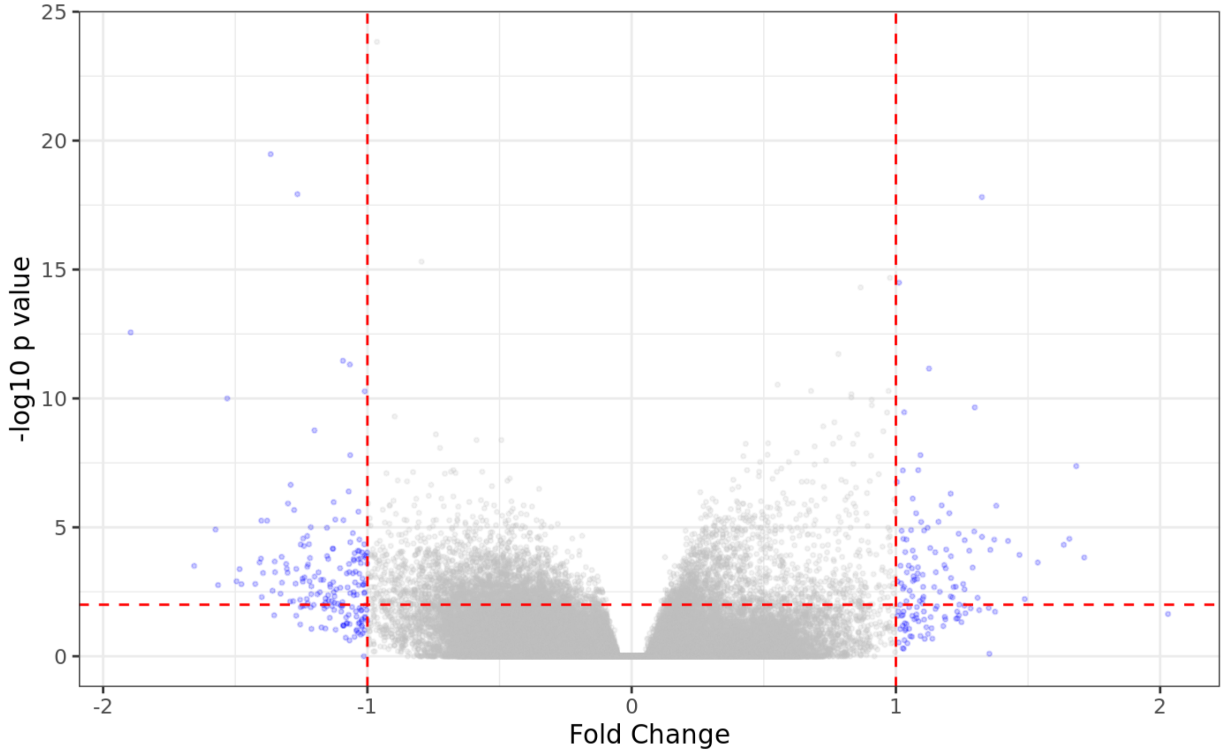
From the KM plot shown above, it seems that metastatic SKCM patients have better survival probability compared to those with non-metastatic SKCM (primary solid tumor group). These counterintuitive results could be the result of TCGA SKCM dataset not having enough non-metastatic patient data, as seen from the rougher primary solid tumor survival probability line, with fewer data point, and more drastic changes when each patient dies. This ununiform data could reflect the metastatic nature of Skin Cutaneous Melanoma, or simply a sampling issue when building the patient cohort to collect data.

2) Expression differences between metastatic and non-metastatic patients



No significant over-expressed or under-expressed DEGs were identified in metastatic patients from the above volcano plot. A few DEGs were shown in green and are either over- or under-expressed, but none passed the p value threshold and therefore none are considered significant. There are no meaningful expression differences between metastatic and non-metastatic skin cutaneous melanoma patients.

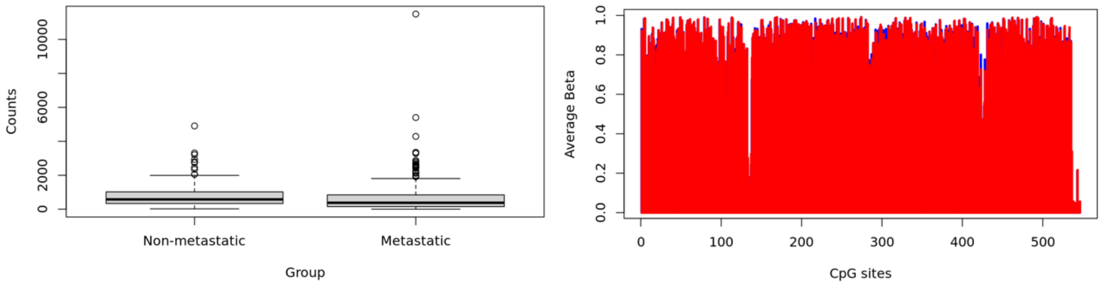
3) Methylation differences between metastatic and non-metastatic patients



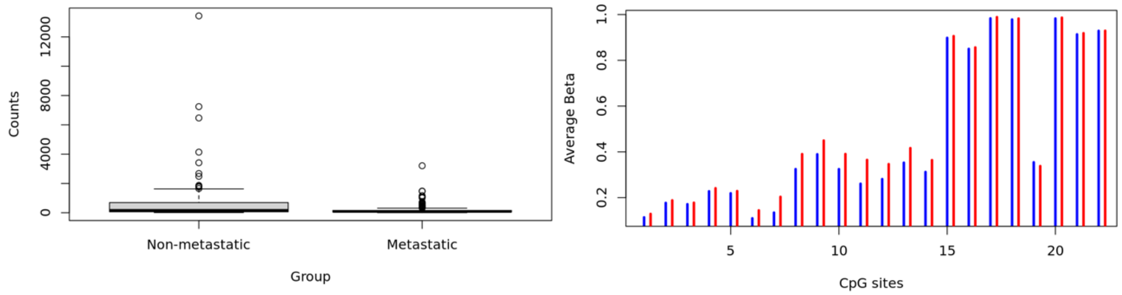
As seen from the above plot, there are significantly hypermethylated (upper right corner, blue dots) and hypomethylated (upper left corner, blue dots) CpG sites identified in metastatic skin cutaneous melanoma patients as compared to the non-metastatic patients. Dots shown in grey do not have change in methylation levels, and dots below the horizontal red line are not considered statistically significant.

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

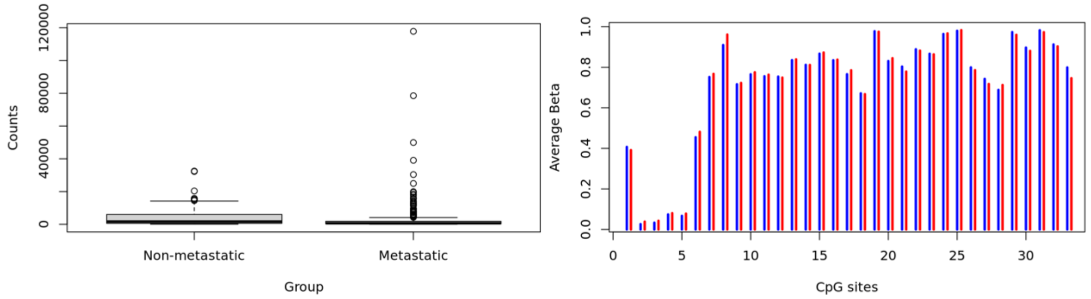
MAD1L1



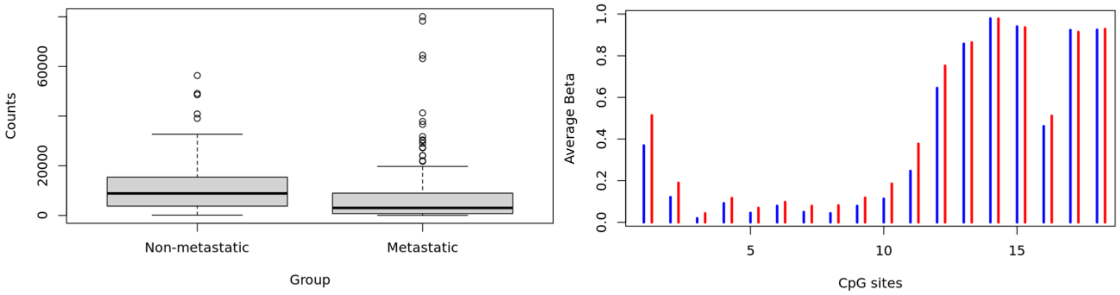
HAS3



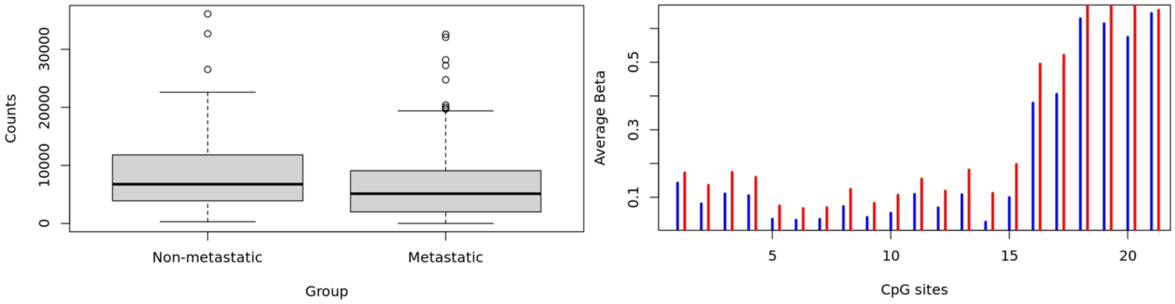
ITGB4



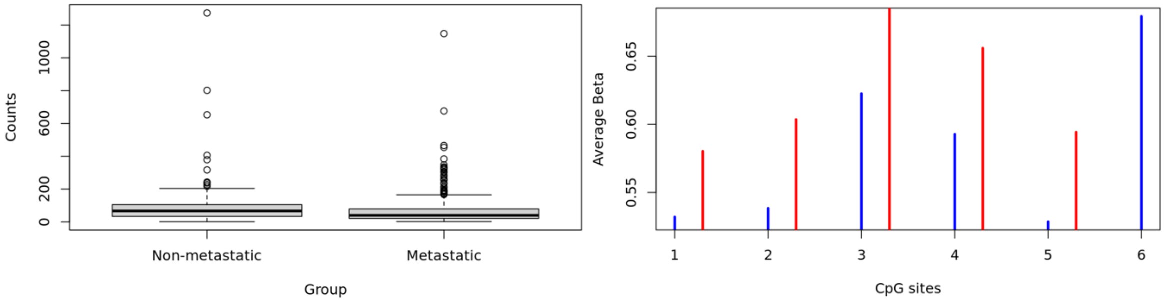
GMPR



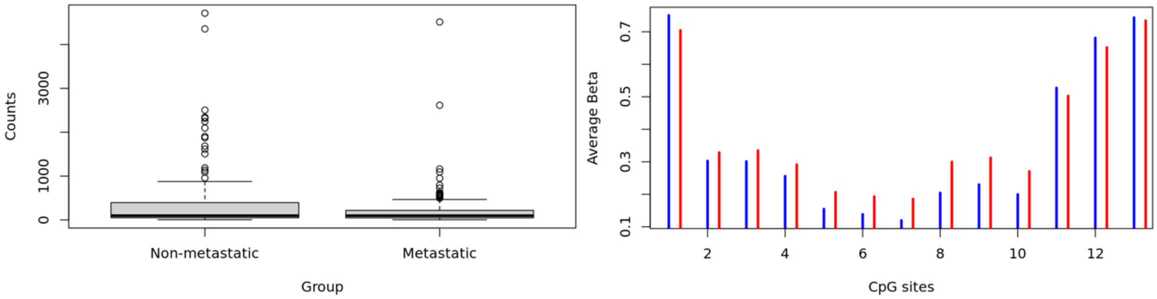
IRF4



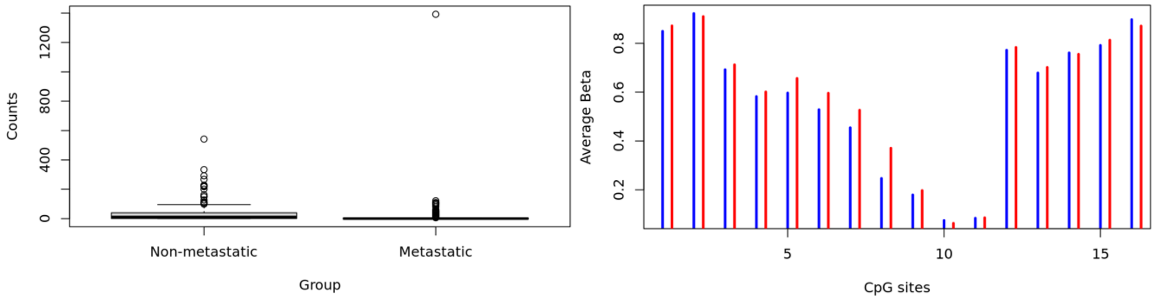
KRTCAP3



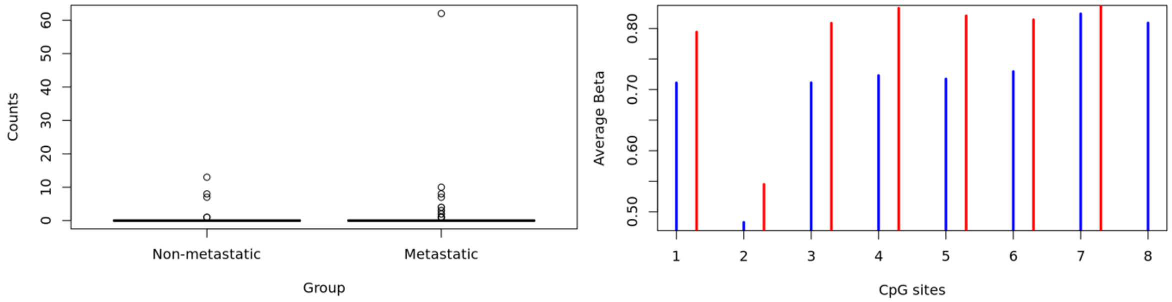
SPINT2



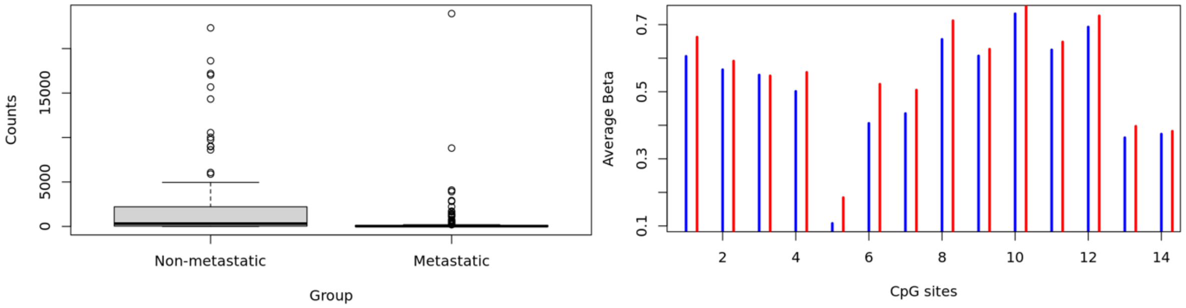
CD164L2



CETN1



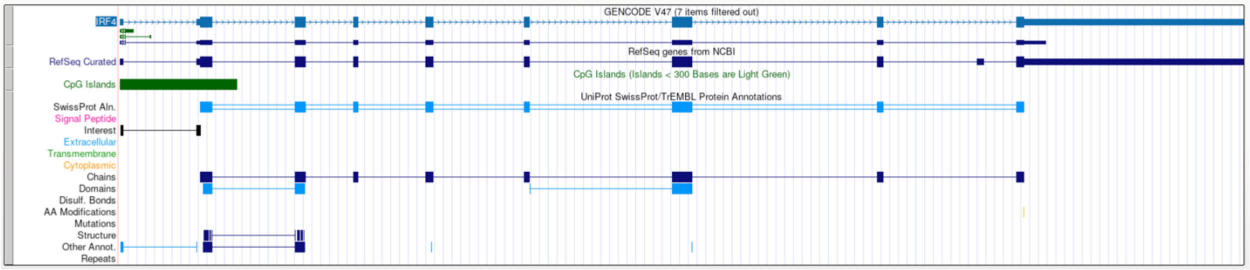
TACSTD2



As seen from the above plots, across the ten analyzed genes, transcriptional activity and methylation status show varying levels of overlap between metastatic and non-metastatic groups, with no consistent or significant differences observed in most cases. While certain genes display slight differences in expression or methylation patterns, these are not pronounced enough to draw definitive conclusions.

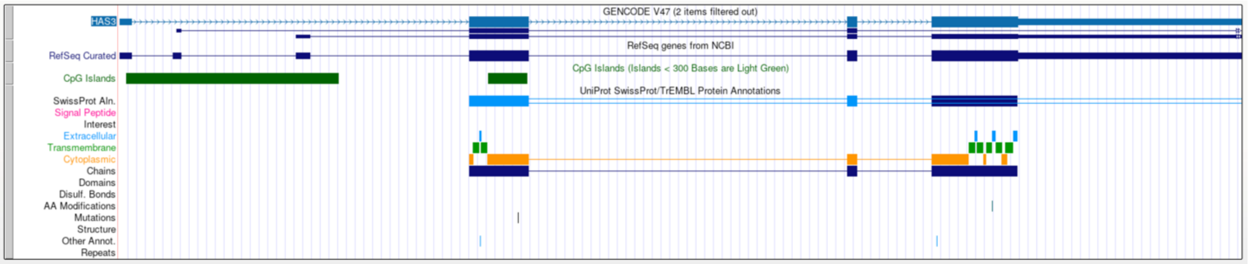
5) Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser). Describe at least one academic article (research or review) that either supports or doesn't support your final conclusion for one of the genes. If previously published work doesn't support your analysis, explain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

IRF4



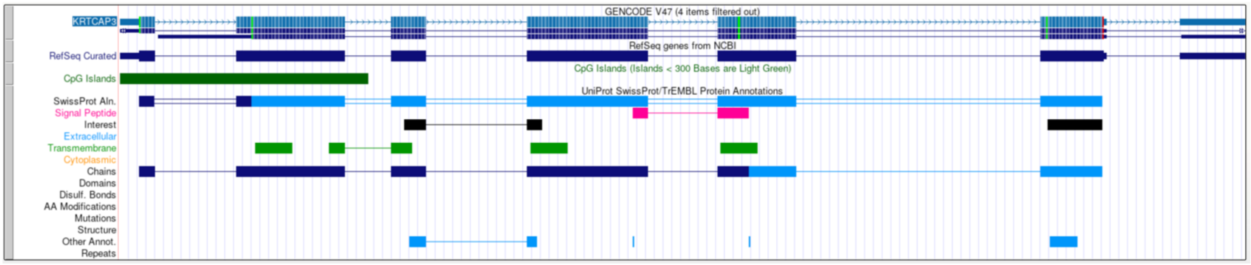
The IRF4 gene shows dense CpG islands in its promoter region and spans multiple exons, with annotated protein domains suggesting regulatory roles in transcription.

HAS3



The HAS3 gene includes CpG islands near its coding region and several protein domains involved in enzymatic functions, which may influence its transcriptional regulation.

KRTCAP3



The KRTCAP3 gene contains CpG islands overlapping its promoter and coding regions, with annotations for protein structural domains, indicating potential involvement in cellular structural integrity.

A past study by Takabe et al. from 2015, *Hyaluronan synthase 3 (HAS3) overexpression downregulates MV3 melanoma cell proliferation, migration and adhesion*, provides evidence that increased HAS3 expression reduces the aggressive properties of MV3 melanoma cells. It was observed that HAS3 overexpression leads to the formation of a hyaluronan-rich extracellular matrix, which subsequently decreases cell proliferation, adhesion, and migration by downregulating key signaling pathways, such as ERK1/2.

This finding contrasts with the results seen in part 4, where HAS3 showed no significant transcriptional or methylation differences between metastatic and non-metastatic groups in SKCM. The discrepancy might stem from the experimental context, as Takabe et al.'s study focuses on **in vitro** manipulation of HAS3 expression in a melanoma cell line, while the TCGA dataset reflects **patient-derived** samples with heterogeneous regulatory mechanisms.

Works Cited

Takabe, Piia, et al. “Hyaluronan Synthase 3 (HAS3) Overexpression Downregulates MV3 Melanoma Cell Proliferation, Migration and Adhesion.” *Experimental Cell Research*, vol. 337, no. 1, Sept. 2015, pp. 1–15, https://doi.org/10.1016/j.yexcr.2015.07.026.