Kate Olsen

**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.

**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 serves as the largest public dataset of deidentified patient cancer genomics data. It is important because, being a public dataset, it expanded the number of people that could research cancer genomics, especially given how hard it can be to collect patient samples. Through TCGA data, researchers have been able to make way in the identification of biomarkers to improve diagnosis, prognosis, and therapeutics.

2. What are some strengths and weaknesses of TCGA?

One of the strengths of TCGA being public is that it makes cancer data more accessible to researchers, and it also allows for better checks in terms of reproducibility. A weakness in TCGA and much data in the scientific community would be lack of representation of minority populations and overrepresentation of white populations treated at academic institutions. Another weakness that comes with a public dataset is that it is harder to regulate the quality of the data, so many of the patient barcodes have some form of missing data, mandating quality preprocessing and data cleaning on the side of researchers. The TCGA was also stopped in 2018, so another limitation is lack of more current data.

Coding Skills

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

git add *filename* (saves files to be committed)
git commit -m “commit message”
git push

2. What command(s) must be run in order to use a package in R?
You have to install and load the package.
Ex:
install.packages(“BiocManager”)

BioCManager::install(“TCGAbiolinks”)
library(TCGAbiolinks)

3. What command(s) must be run in order to use a *Bioconductor* package in R?
You have to first make sure Bioconductor is installed, and then, using Bioconductor, you can install and load in the bioconductor packages.

install.packages("BiocManager")
BiocManager::install(version = "3.19")

BiocManager::install("TCGAbiolinks")
library(TCGAbiolinks)

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

Boolean indexing uses boolean values (true and false) in order to select or exclude elements within a dataset, and some applications include data filtering, exclusion, or replacing data.

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.

```
df <- data.frame(
  patient_id = c("1", "2", "3", "4", "5"),
  age = c(25, 62, 54, 81, 42)
)
print(df)
```

| patient_id <chr> | age <dbl> |
|---|---|
| 1 | 25 |
| 2 | 62 |
| 3 | 54 |
| 4 | 81 |
| 5 | 42 |

5 rows

```
```{r}
df$age_category <- ifelse(df$age >= 55, "Old", "Young")

young_adults <- df[df$age > 17 & df$age < 30, ]
```
```

The ifelse() statement is used in this scenario to add a new column called age_category to the data frame, where each row corresponds to a patient. If the patient's age is greater than or equal to 55, that patient is assigned to the "Old" age_category, else they are assigned "Young".

The last line is an example of boolean indexing where conditions like df$age > 17 & df$age < 30 result in TRUE/FALSE vectors. In this line, I am using this boolean vector to filter for rows in the initial data frame where the patient is 18-29 years old and storing them in a new dataframe, young_adults.

# Part 2: SKCM Analysis

SKCM review article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3004577/

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

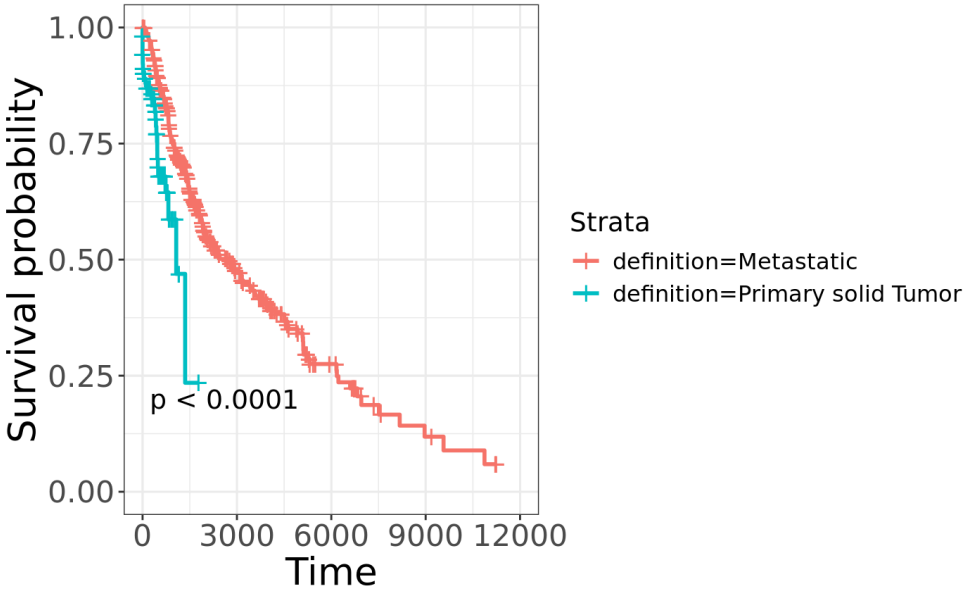
5. Visualization of CpG sites and protein domains for 3 genes for a few genes (use UCSC genome browser)

# 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:

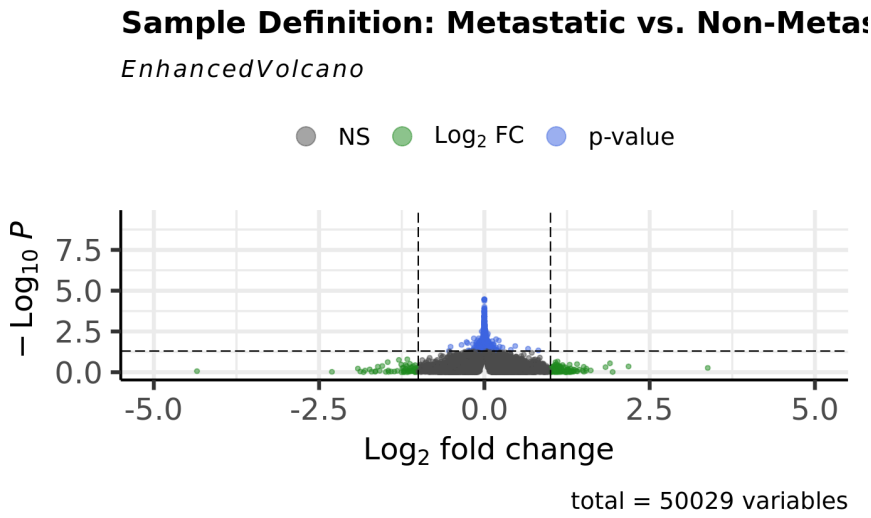
- Analyze the plot. What conclusions can you and can you not draw about differences between metastatic and non-metastatic TCGA SKCM patients? Why?

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



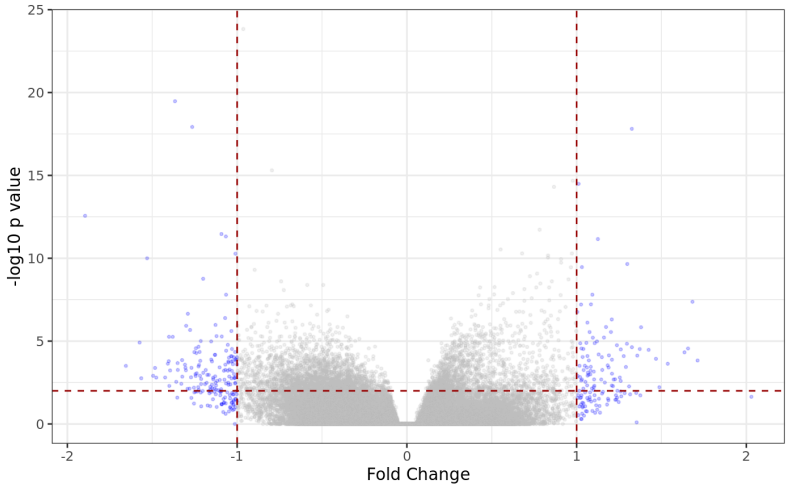

From this plot, it appears that patients with metastatic skin cutaneous melanoma have a better survival probability at all time points than those with non-metastatic cancer. However, this conclusion does not make sense given what we know about metastatic cancer. Instead, looking at the non-metastatic data set (primary solid tumor), the line is rough, indicating fewer data points, likely skewing the data. In this scenario, it is either more common for SKCM to be metastatic or more likely, the patients in TCGA that did have non-metastatic SKCM had null values in their data and were excluded during data cleaning.

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




From the volcano plot, with the base group being non-metastatic patients and the comparison being metastatic patients, this volcano plot shows that there is no significant up or down regulation of genes in the metastatic patients. Any genes to the right of the rightmost vertical line are considered significantly upregulated and the genes to the left of the leftmost line would be the down regulated ones; however, since none of them are above the p-value threshold, none are significant.

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

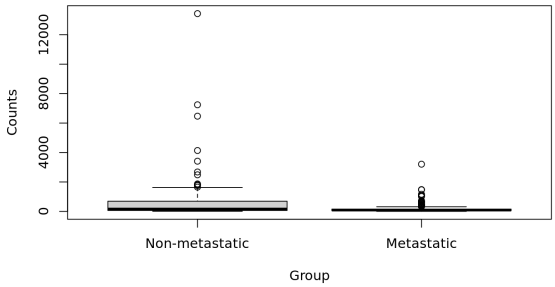


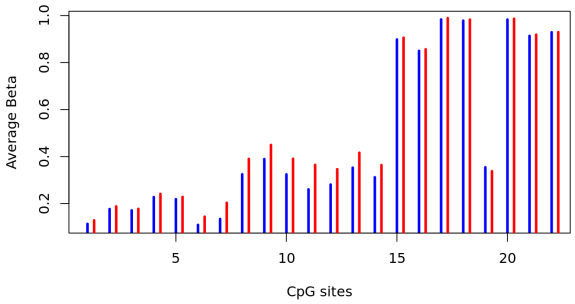

Similar to the prior volcano plot, anything in the top right shows significantly hypermethylated

cpg sites in metastatic cases compared to the non-metastatic samples, and the blue dots in the top left show cpg sites where there is less methylation in metastatic cases compared to the non-metastatic samples.

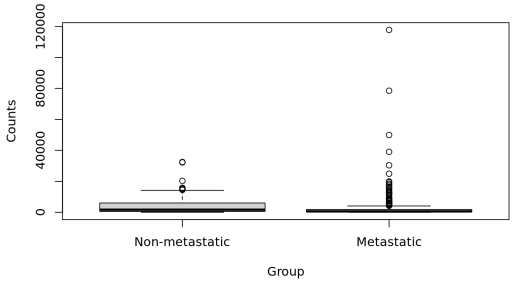
**4 ) Direct comparison of transcriptional activity to methylation status for 10 genes**

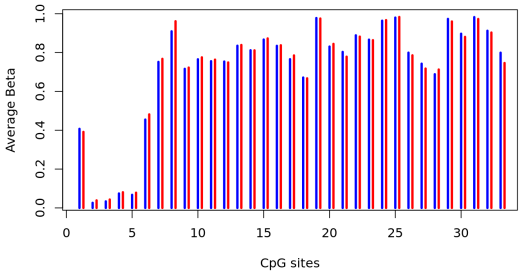
1. **HAS3**



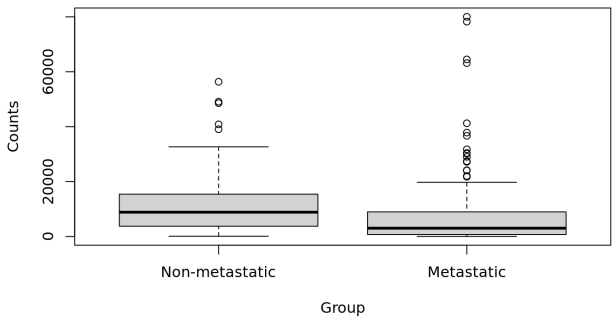



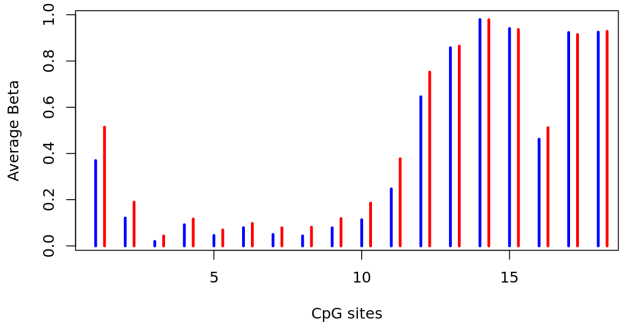

2. **ITGB4**



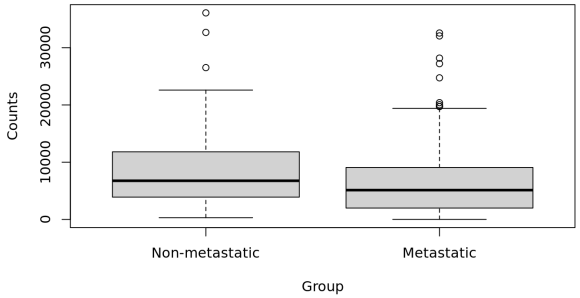



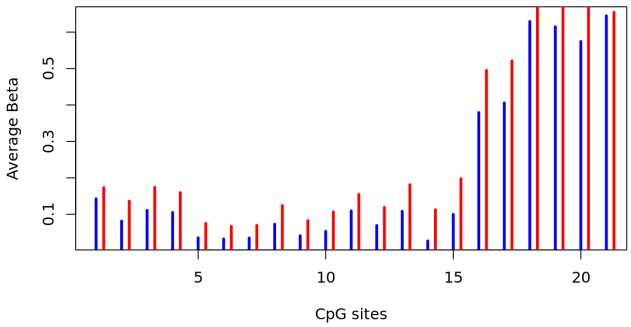

3. **GMPR**



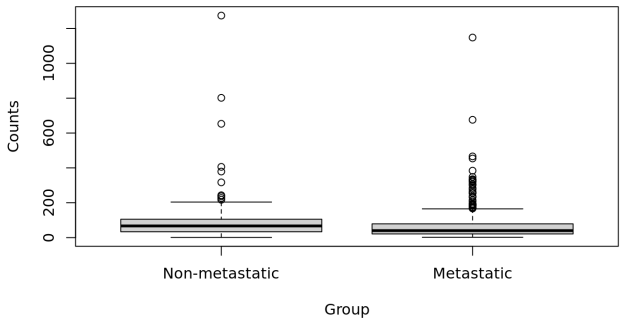



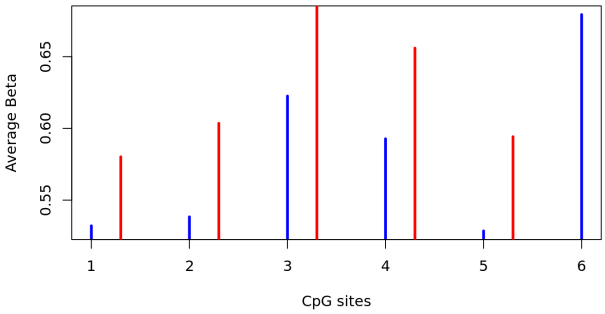
## 4. **IRF4**



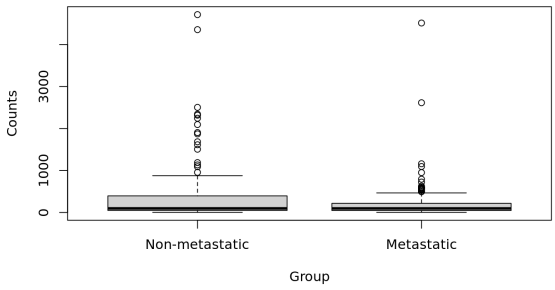



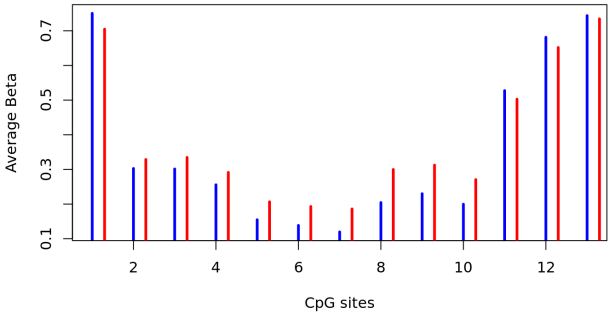

## 5. **KRTCAP3**



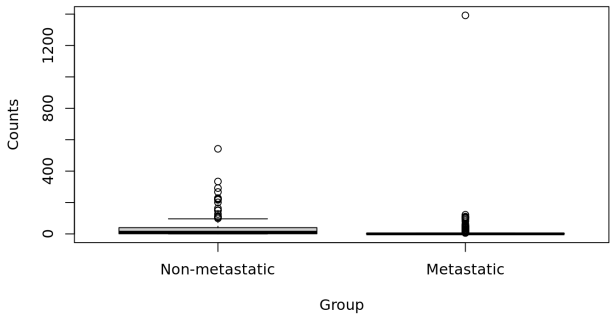



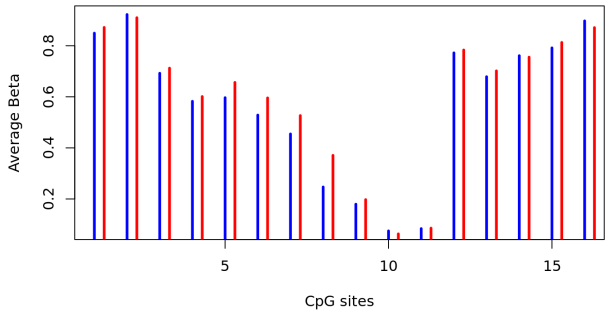

## 6. **SPINT2**



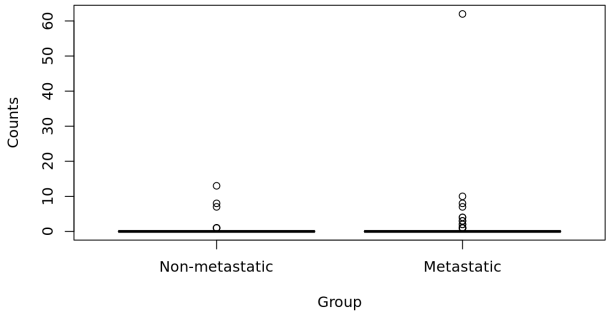



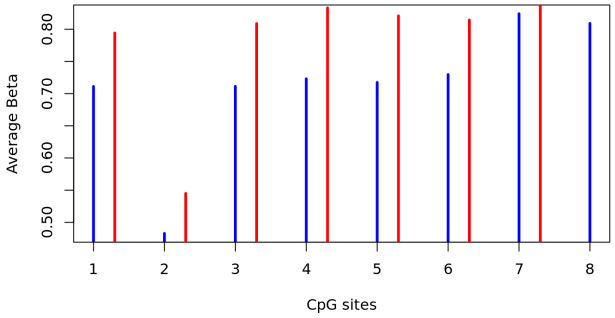
## 7. **CD164L2**



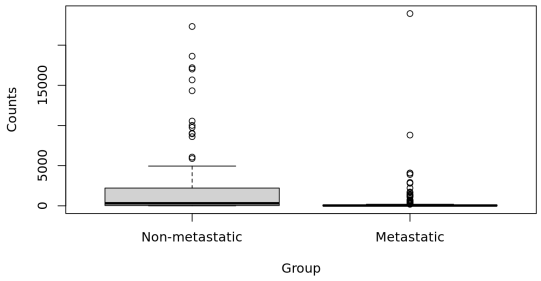



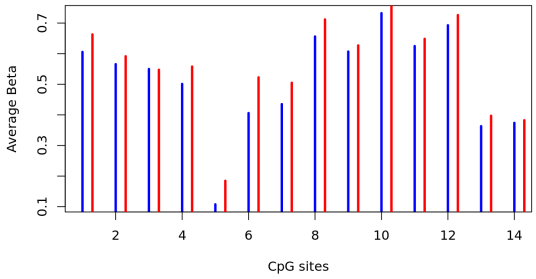

## 8. **CETN1**



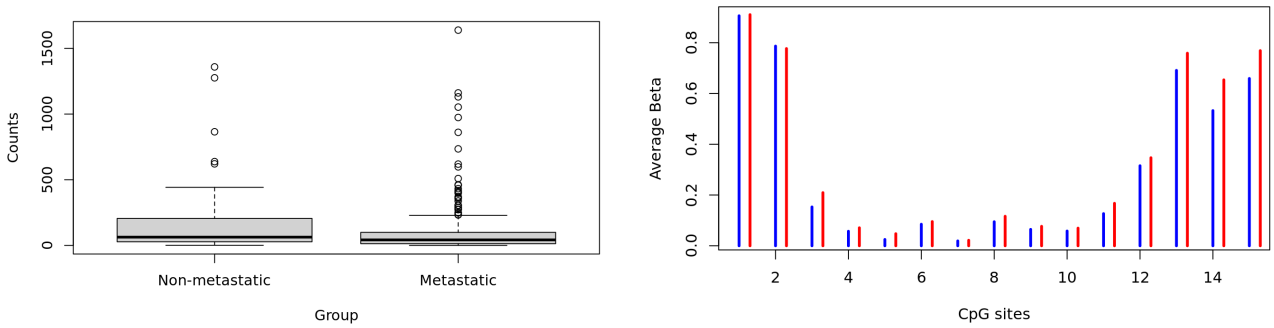




## 9. **TACSTD2**



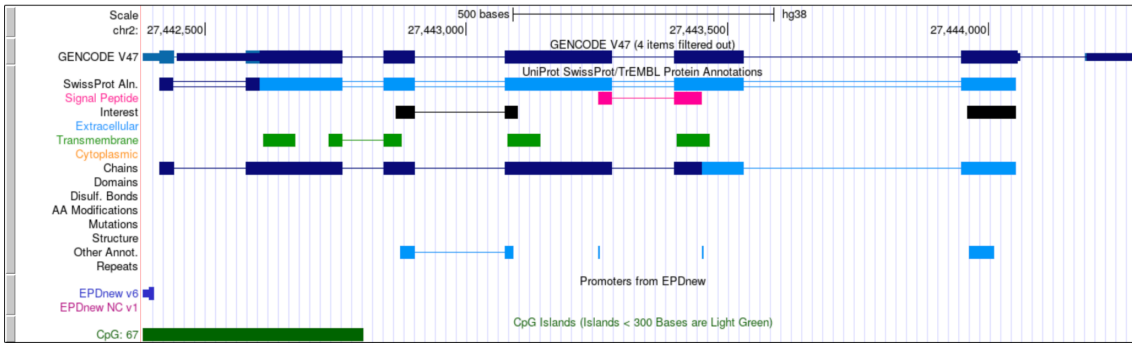
10. **LINC00482**




Based on these comparisons, the difference in rna_counts does not look statistically significant for most of the genes. If there is a difference, it seems like non-metastatic patients have a greater number of RNA counts that are downregulated. In terms of methylation, the metastatic cases seem to consistently be more methylated, especially in the KRTCAP3 gene.

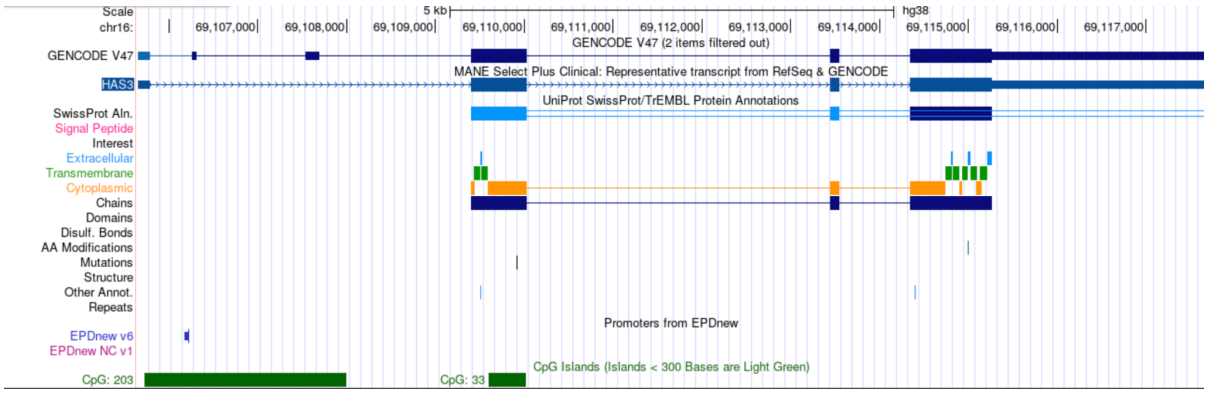
**5 ) Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser) for a few genes. Describe at least one academic article (research or review) that either supports or doesn't support your final conclusion for one of the genes. If previously published work doesn't support your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

KRTCAP3



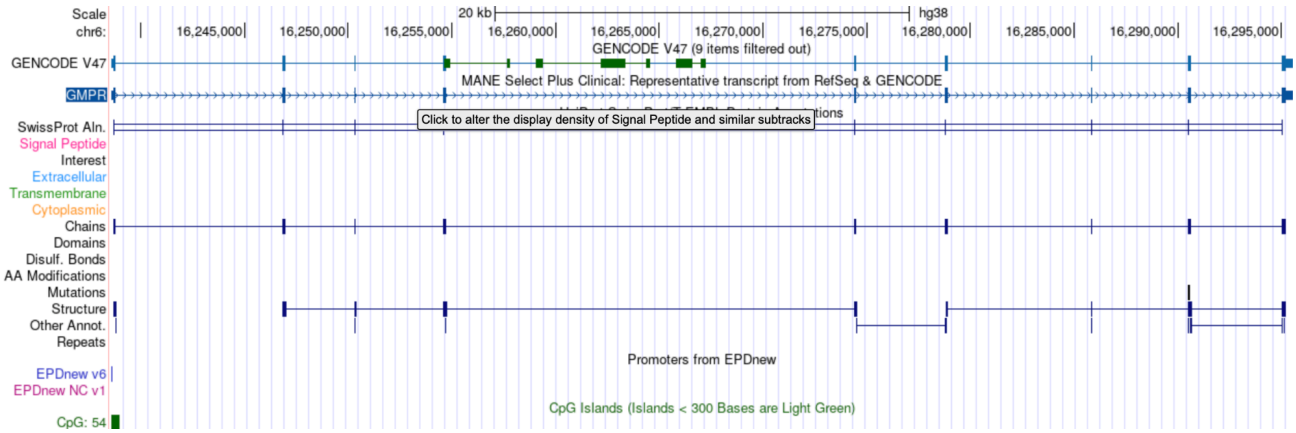

The green CpG islands are linked with promoter regions. This gene has a large CpG island which overlaps with chains, transmembrane, GENCODE V47, and SwissProt Aln.

HAS3




This gene has 2 CpG islands, one significantly larger towards the start and overlapping with just GENCODE V47 and the second one overlapping with more characters, including GENCODE V47, HAS3, SwissProt Aln, TransMembrane, Cytoplasmic, and Chains.

GMPR




This gene has the smallest CpG Island but overlaps with 7 factors.

Generally, CpG Islands that are near gene promoter regions suggest that these genes are regulated by DNA methylation. KRTCAP3 was the gene that was shown to show the most variety in methylation pattern across metastatic vs. non metastatic states of these three while GMPR showed the least, and this corresponds to the degree of overlap that their corresponding CpG islands have with other promoter regions.

From Dobre et. al's research in "Interrogating Epigenome toward Personalized Approach in Cutaneous Melanoma," they mention that KRTCAP3 is hypermethylated in melanomas compared to melanocytes. Melanocytes are healthy skin cells, so from this, KRTCAP3 hypermethylation is certainly prevelant in

cancerous cells compared to normal ones. Although this article does not specifically break down hypermethylation into metastatic and nonmetastatic patients, it seems plausible that KRTCAP3 could be linked to a more aggressive form of skin cutaneous melanoma in general, explaining the higher methylation rates in the metastatic cancer cohort.

References

Dobre, E. G., Constantin, C., Costache, M., & Neagu, M. (2021). Interrogating Epigenome toward Personalized Approach in Cutaneous Melanoma. Journal of personalized medicine, 11(9), 901. https://doi.org/10.3390/jpm11090901