**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 Mahijaand 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 is a free public database that provides molecular and clinical data from a wide variety of cancer types. It provides researchers with an accessible way to conduct statistical analysis to identify key genetic mutations and pathways involved in cancer, find biomarkers for diagnosis, and spot suitable treatment types.

1. What are some strengths and weaknesses of TCGA?

Strengths: over 11,000 samples for comprehensive modeling, over 30 cancer types for multiple research directions, open access for worldwide collaboration in cancer research.

Weaknesses: lack of longitudinal data, no real-time updates, NAs and none reported information all over the dataset.

Coding Skills

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

# Add the file to staging

git add <filename>

# Commit changes

git commit -m "message\_here"

# Push changes

git push origin <branch-name>

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?

if (!requireNamespace("BiocManager", quietly = TRUE))

install.packages("BiocManager")

BiocManager::install("package\_name")

library(package\_name)

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

Using true or false values to select subsets of data. Possible applications include: filtering rows, identifying values of interest, and so on.

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.

a. an ifelse() statement

b. boolean indexing

df <- data.frame( Name = c("Tony", "Tom", "Toby"), Age = c(18, 20, 21), Score = c (100, 81, 75) )

a.

#identify adult

df$Adult <- ifelse(df$Age >= 18, "Yes", "No")

b.

# Filter rows where Score is greater than 80

high\_scorers <- df[df$Score > 81, ]

**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\_sk cm [-c (1,2),]

drug <- clinical.BCRtab.all$clinical\_drug\_sk cm [-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:

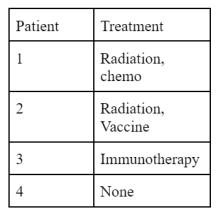
○ Use the rna\_se clinical data (rna\_se@colData).

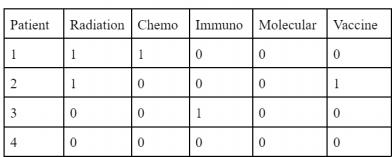
○ Filter out genes with a total expression across all patients of < 20

○ 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:





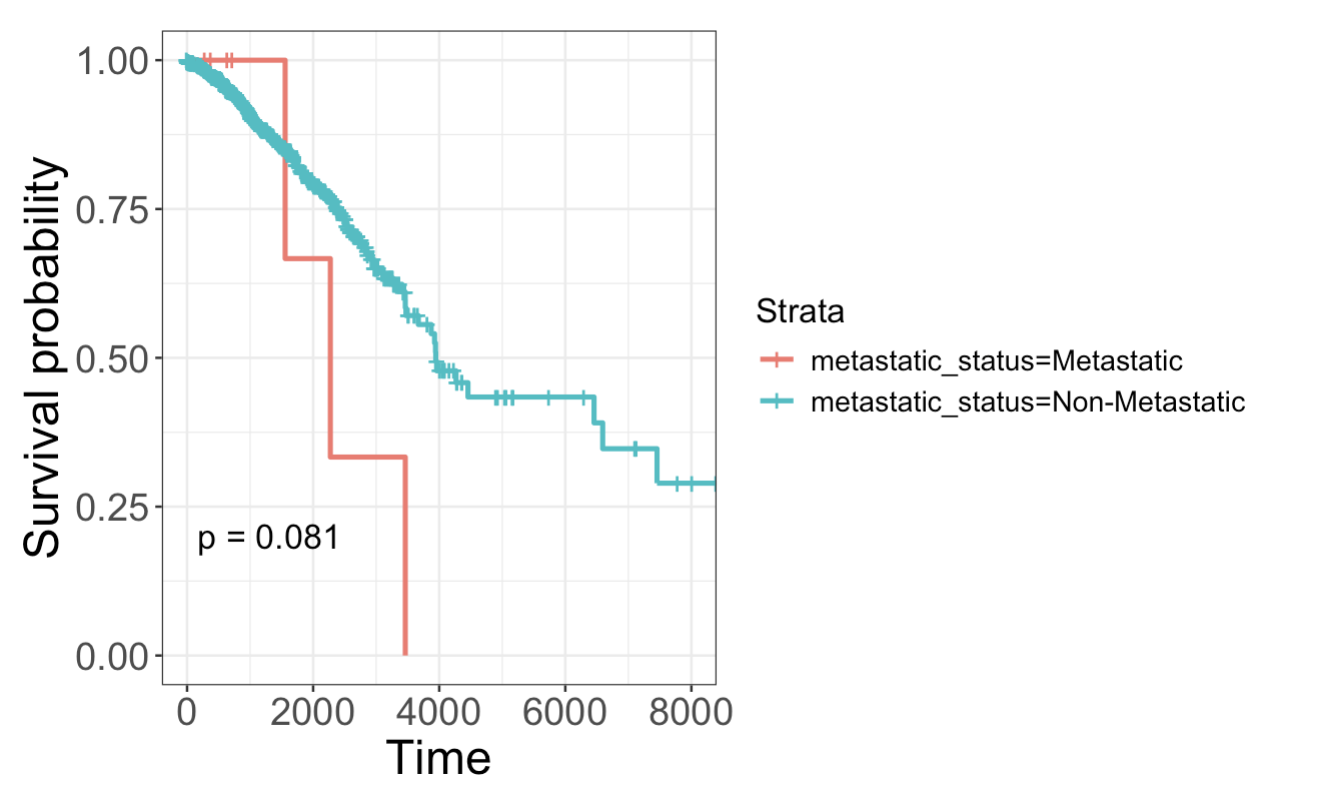
**=**

**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:

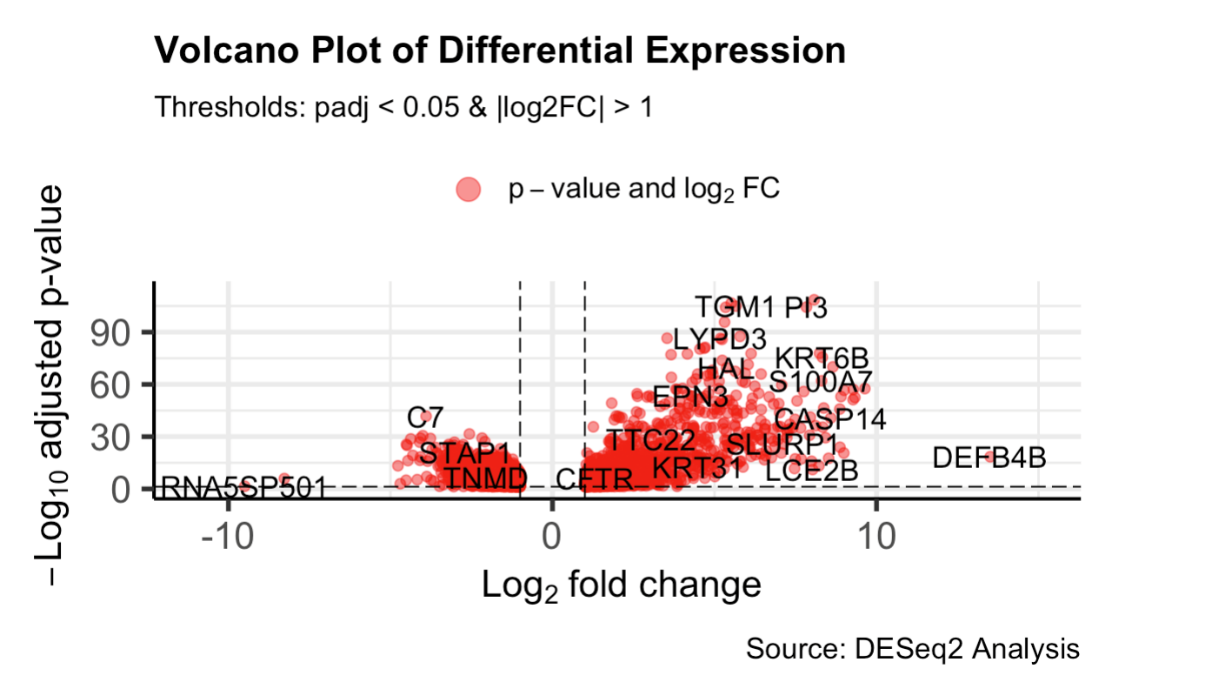
● Analyze the plot. What conclusions can you and can you not draw about differences between metastatic and non-metastatic TCGA SKCM patients? Why?

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



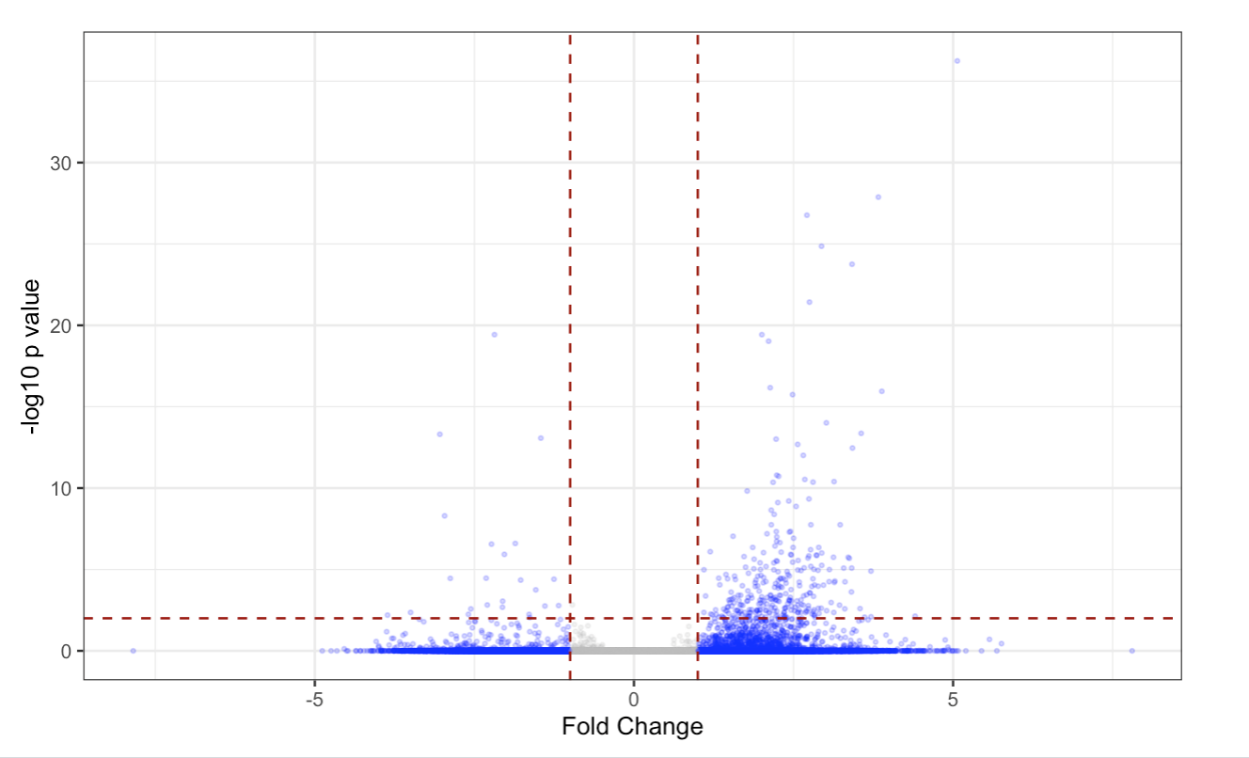
From the KM plot, we can observe a generally better survival probabilities over time for non-metastatic patients in comparison to that for metastatic patients. Also, Metastatic patients show a sharper decline in survival probability earlier in the time course, indicating worse early survival compared to non-metastatic patients. However, the p-value of 0.081 indicates that the difference in survival between these groups is not statistically significant at the commonly used threshold of 0.05. In addition, the plot only establishes correlation but not causality.

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



The volcano plot suggests that genes like TGM1, KRT6B, and LYPD3 are upregulated, but genes such as TNMD, STAP1, and C7 are downregulated. These top upregulated and downregulated genes may serve as potential biomarker or therapeutic targets. However, the causal relationship is still not established. Therefore, the biological impact of these genes should be further studied.

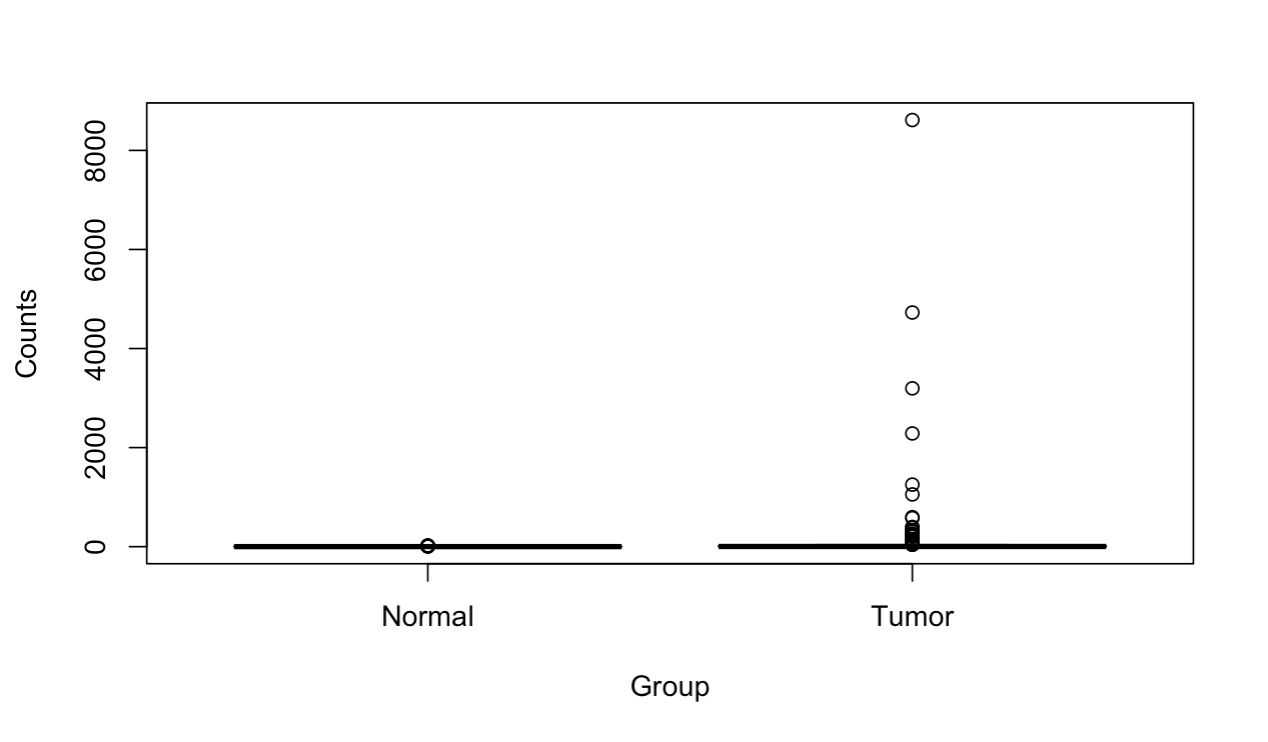
**3 ) Methylation differences between metastatic and non-metastatic patients**

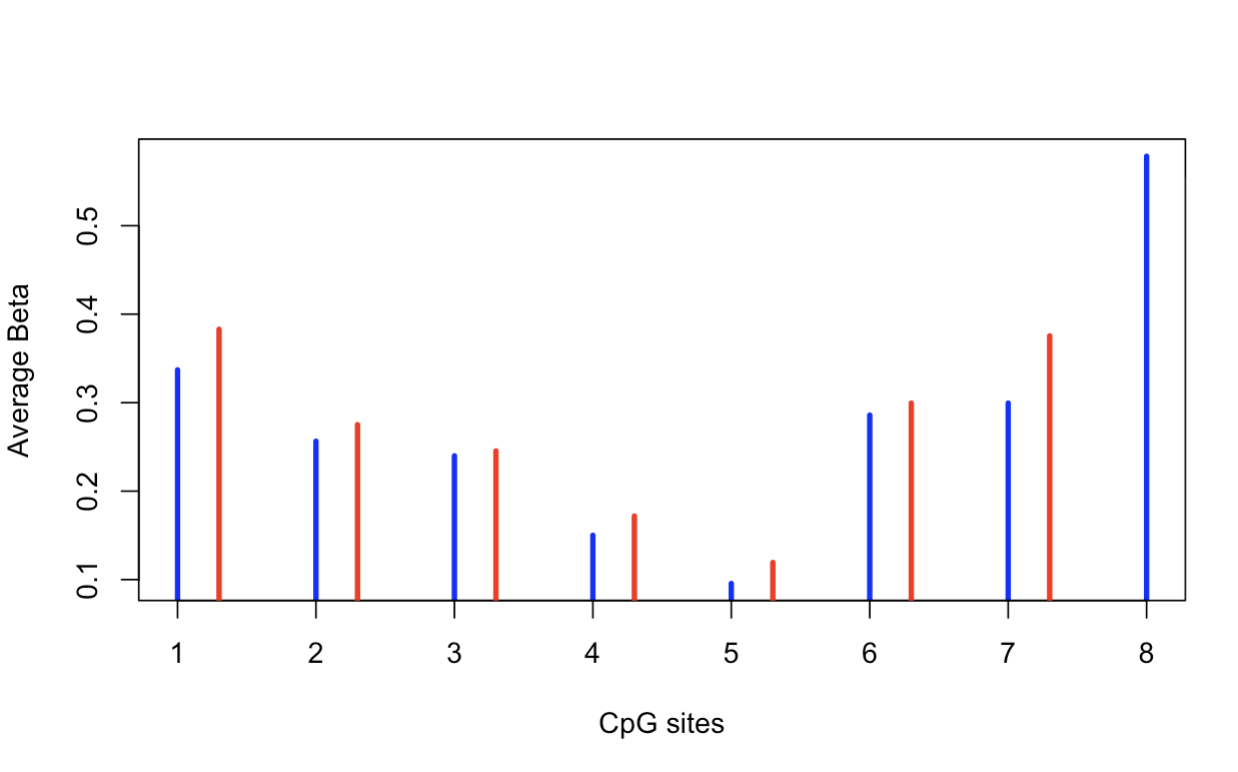


In this plot, the genes on the right are more methylated in metastatic patients and the genes on the left are more methylated in non-metastatic patients. Some outliers may exist as some genes exhibit extreme methylation differences. These outliers show potential in terms of determining metastasis. However, the plot cannot confirm gene functionality or its impact on metastasis. Further experiment is needed.

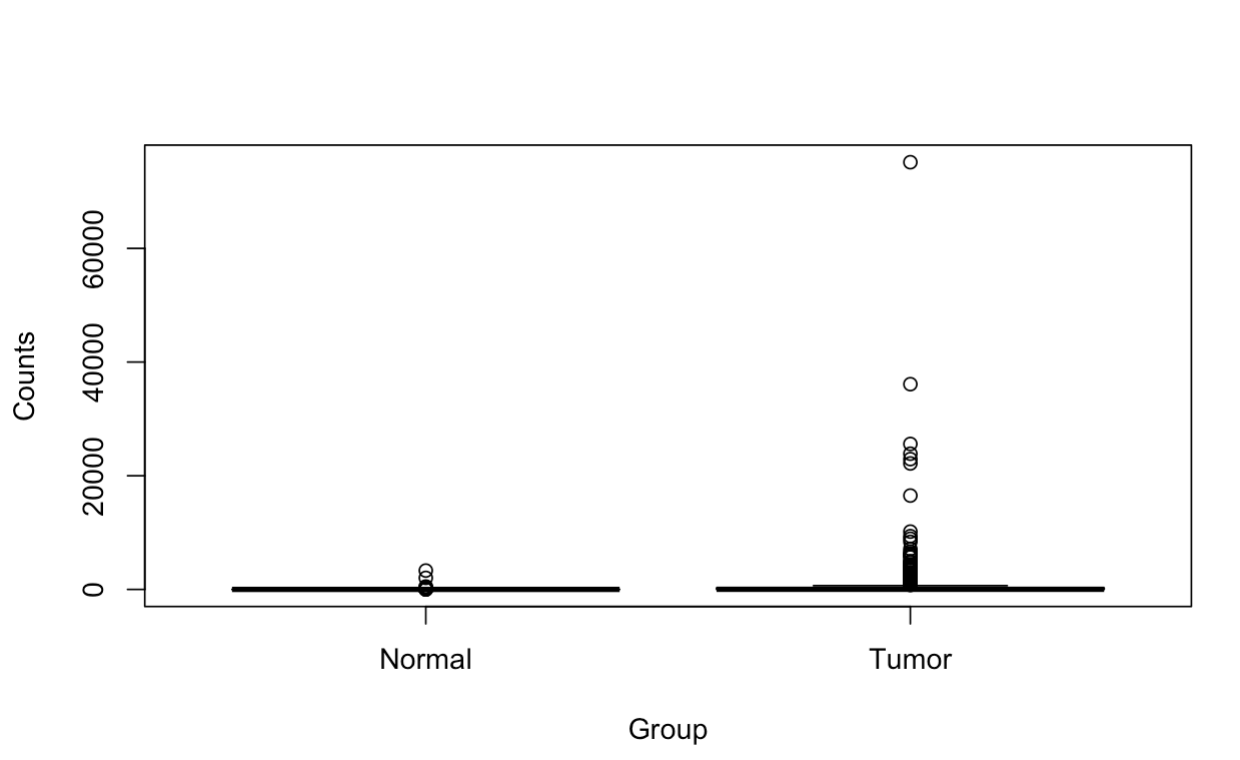
**4 ) Direct comparison of transcriptional activity to methylation status for 10 genes**

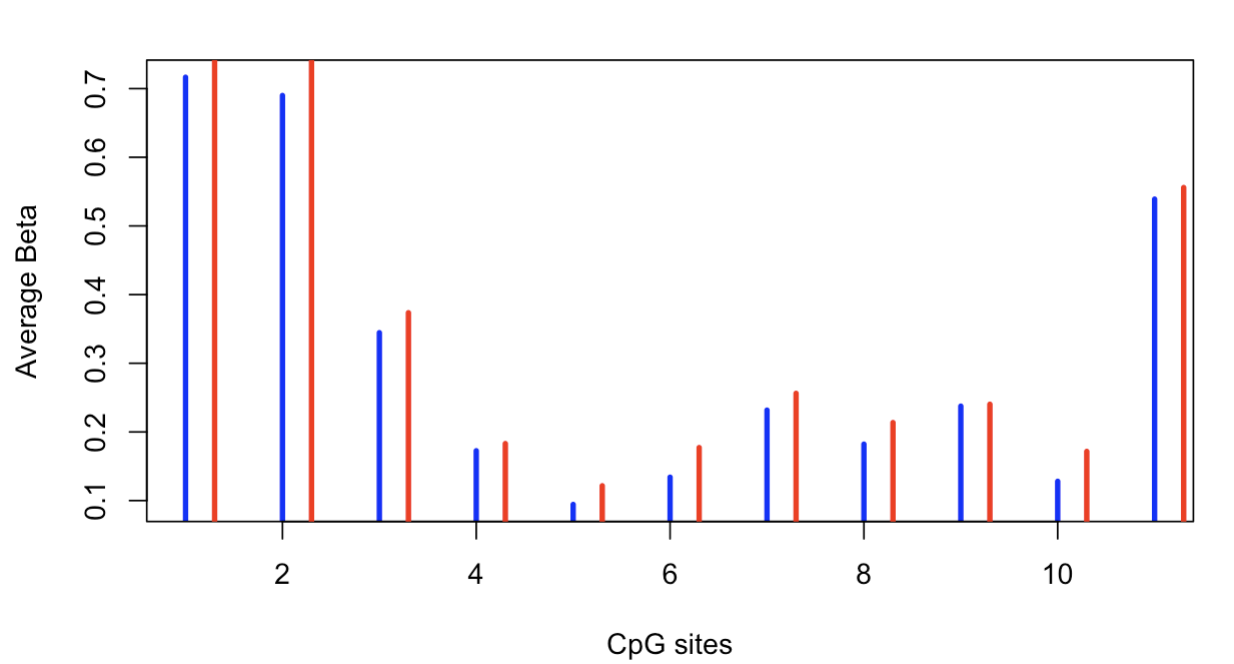
NRK

****

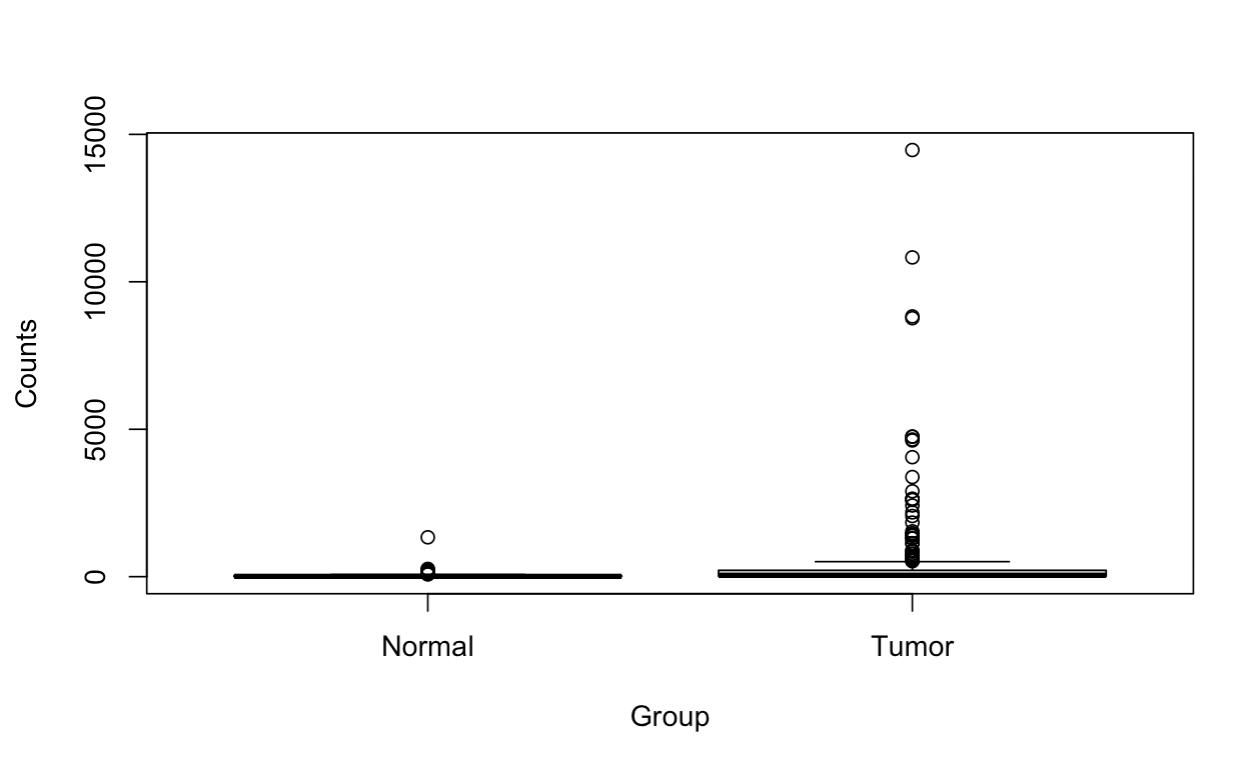
****

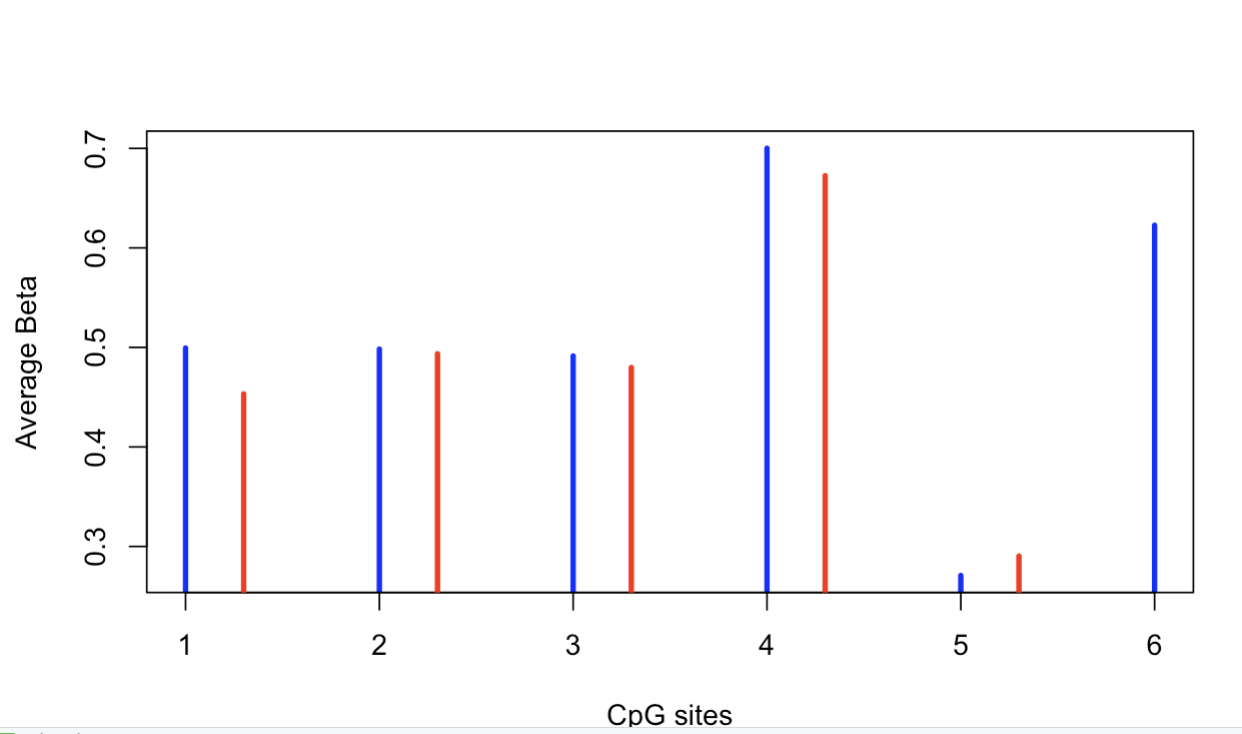
CR2



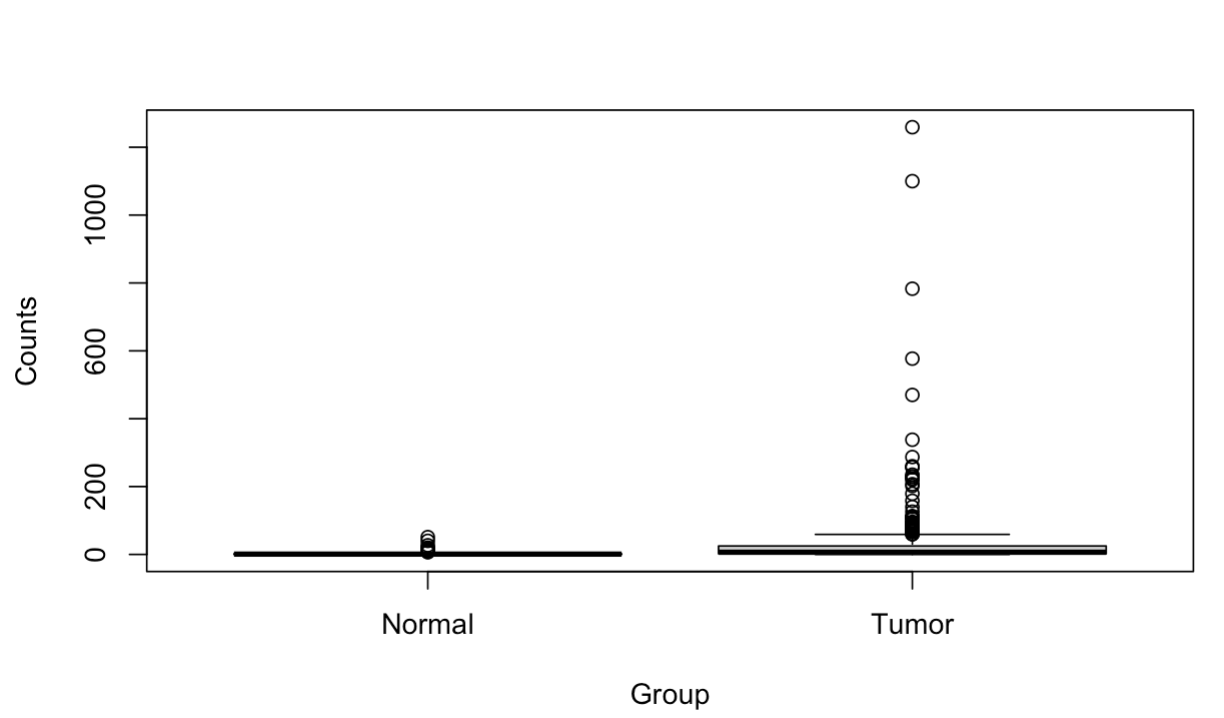
****

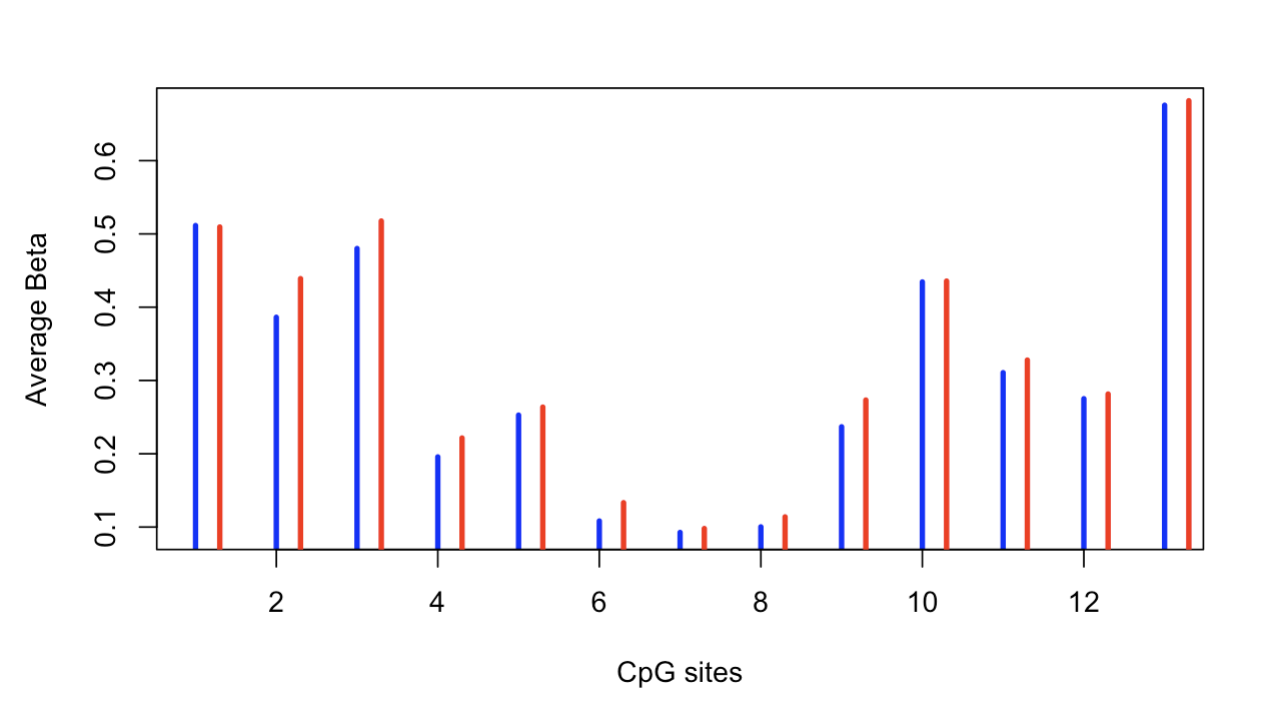
BANK1



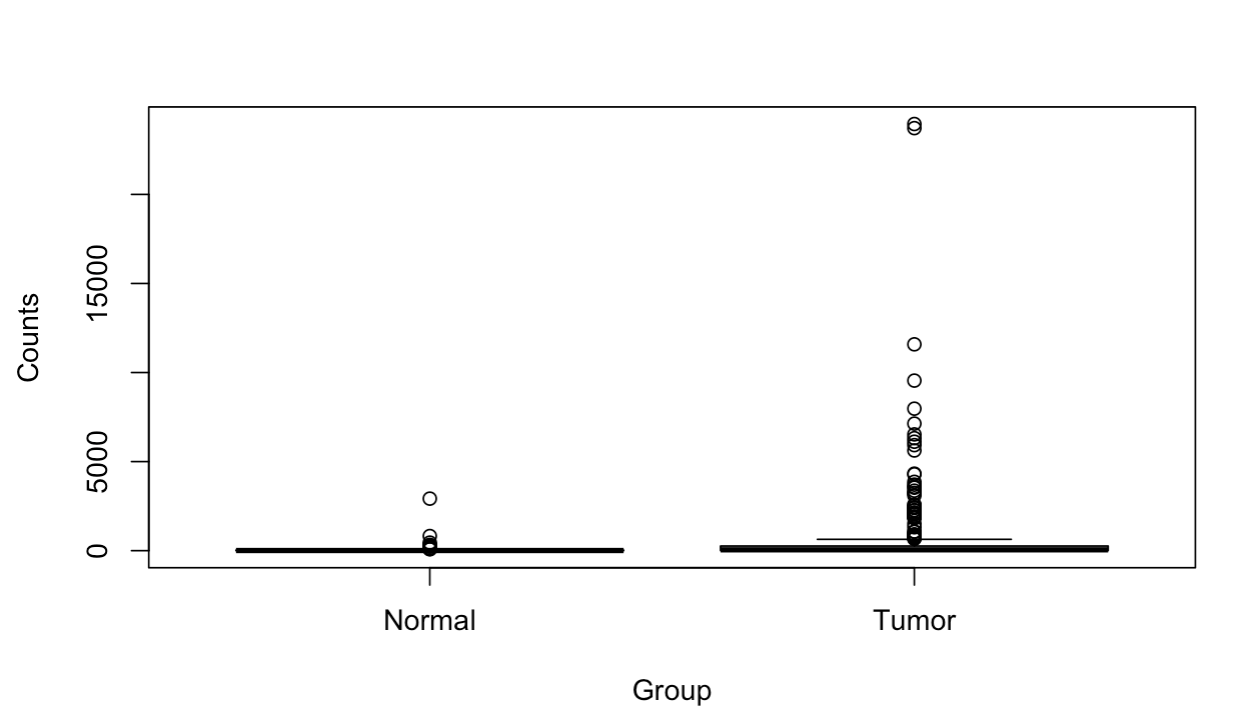
****

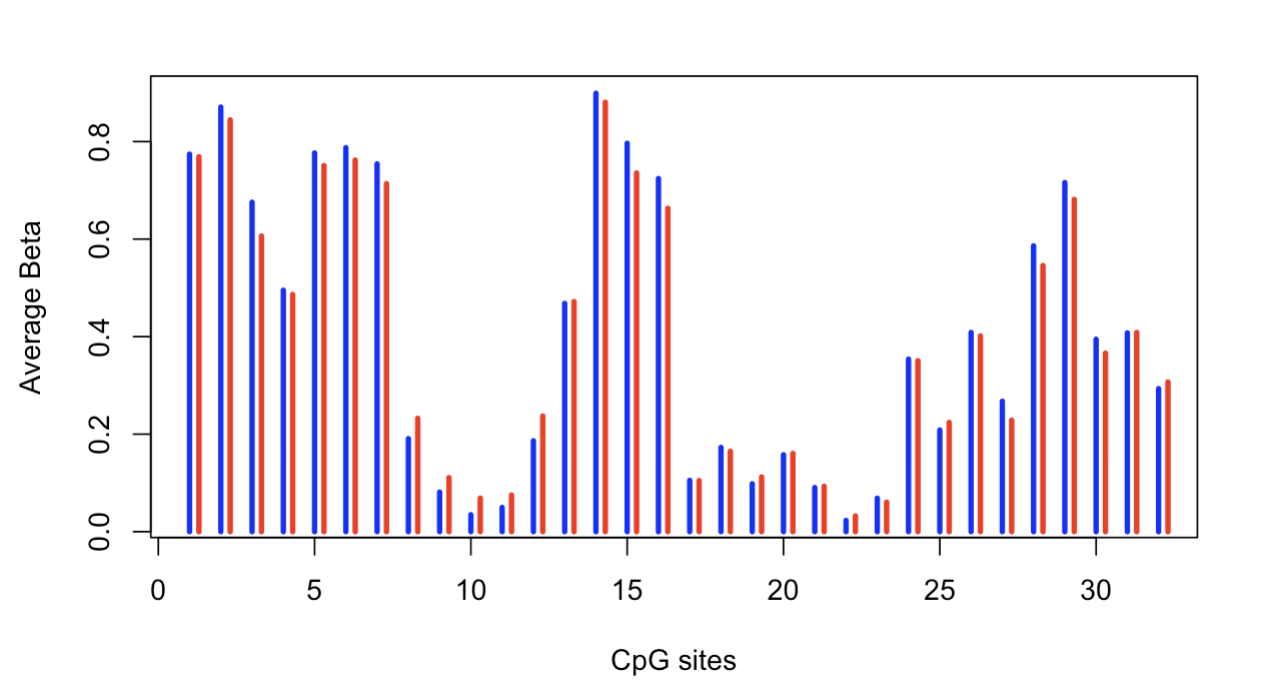
BEND4



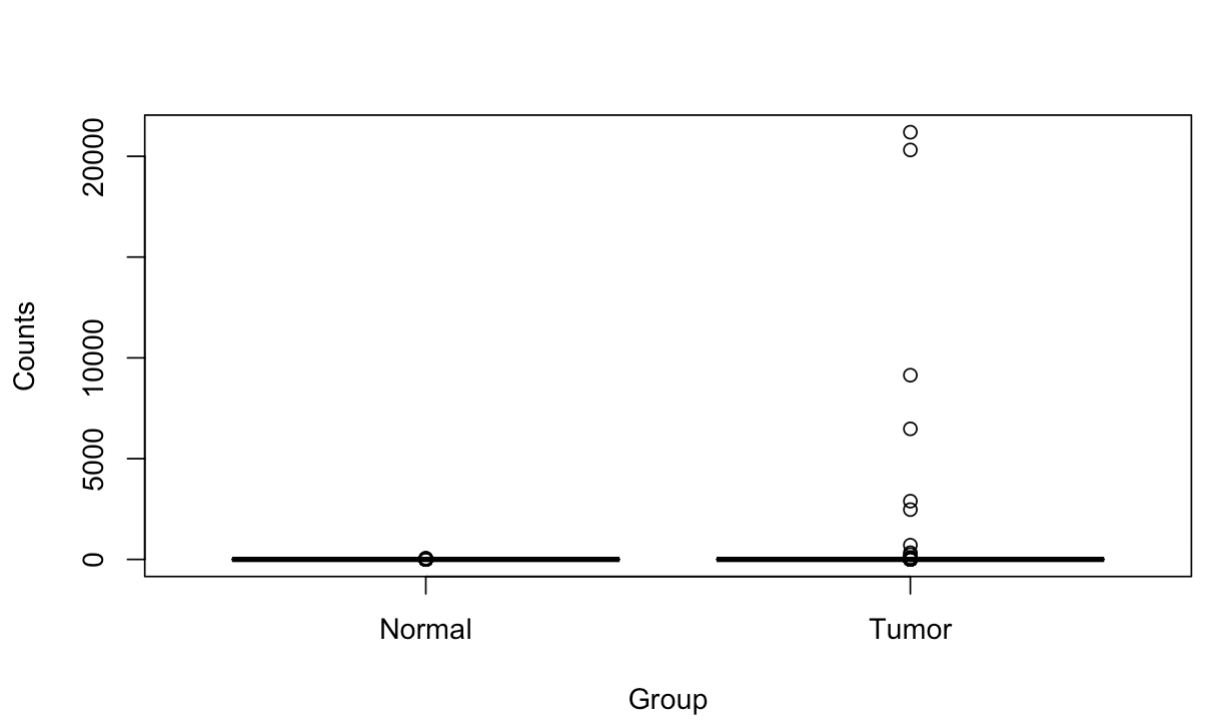
****

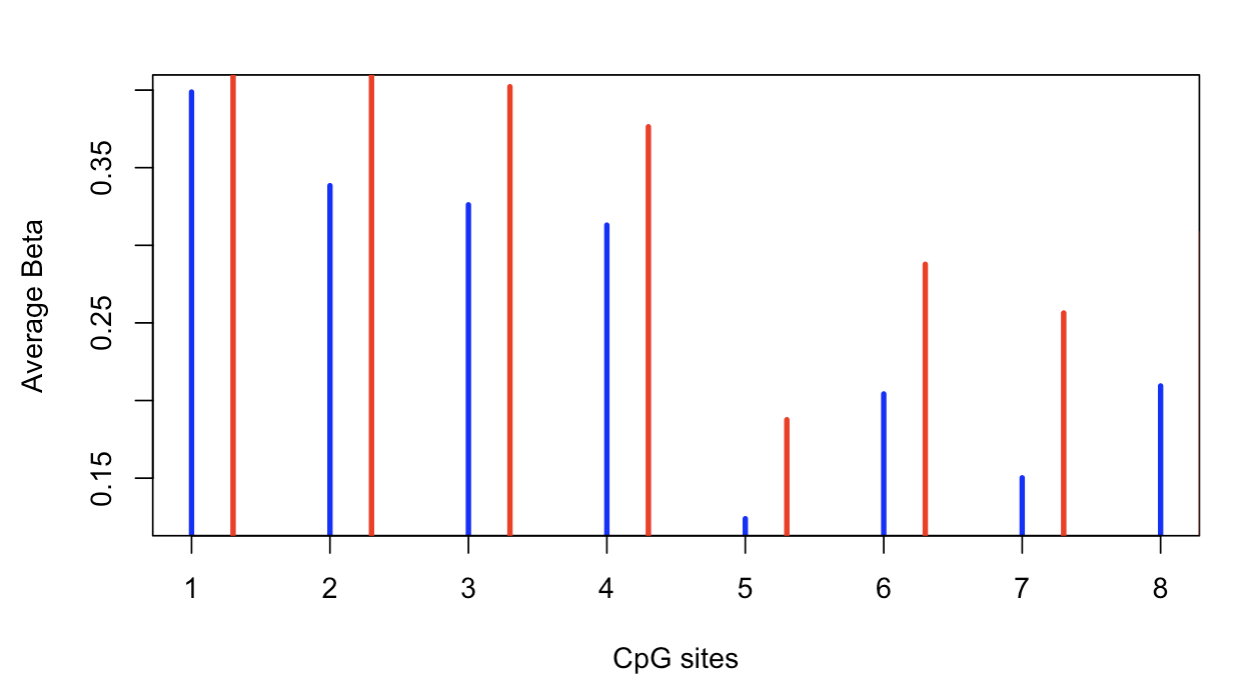
PAX5



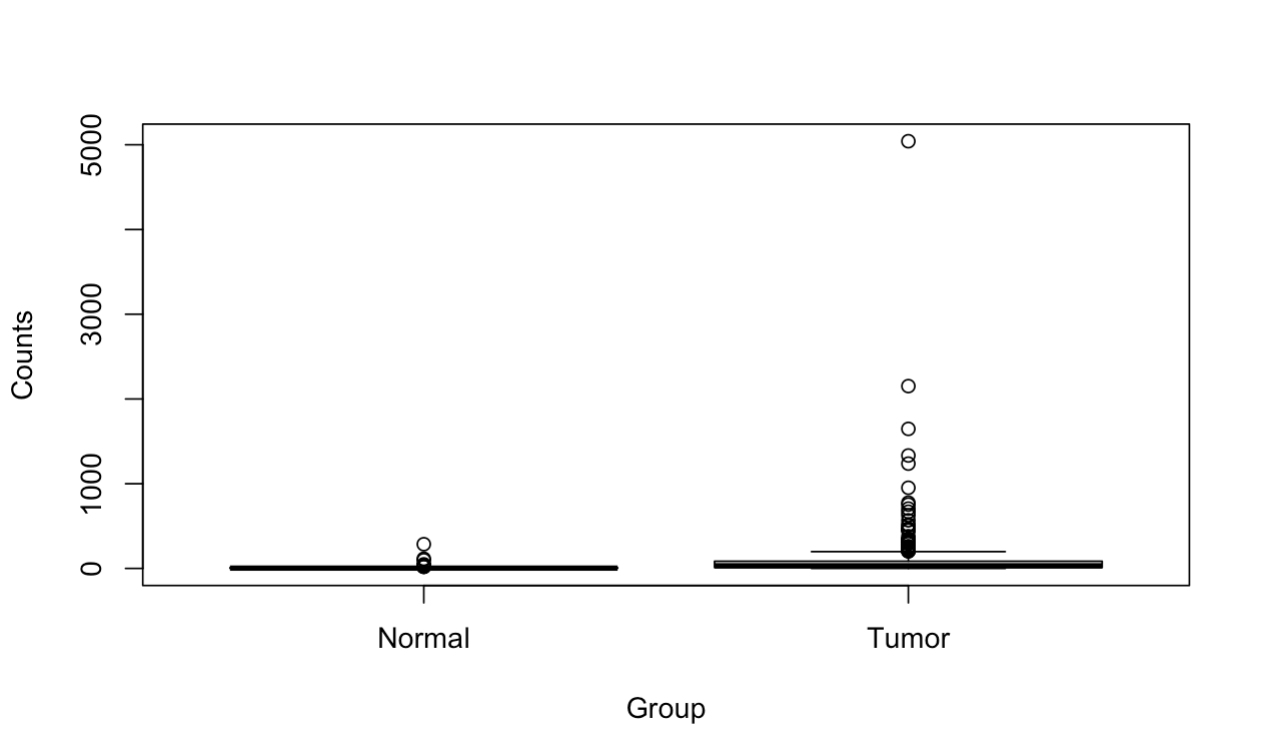
****

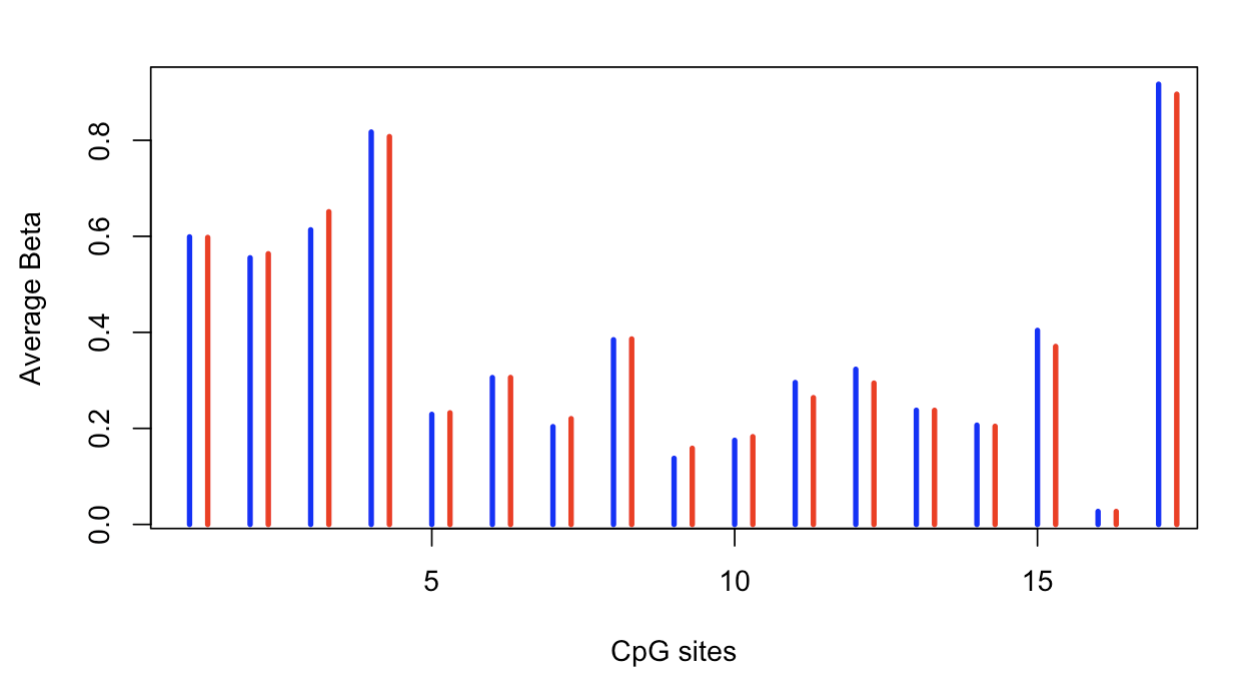
SLC17A6



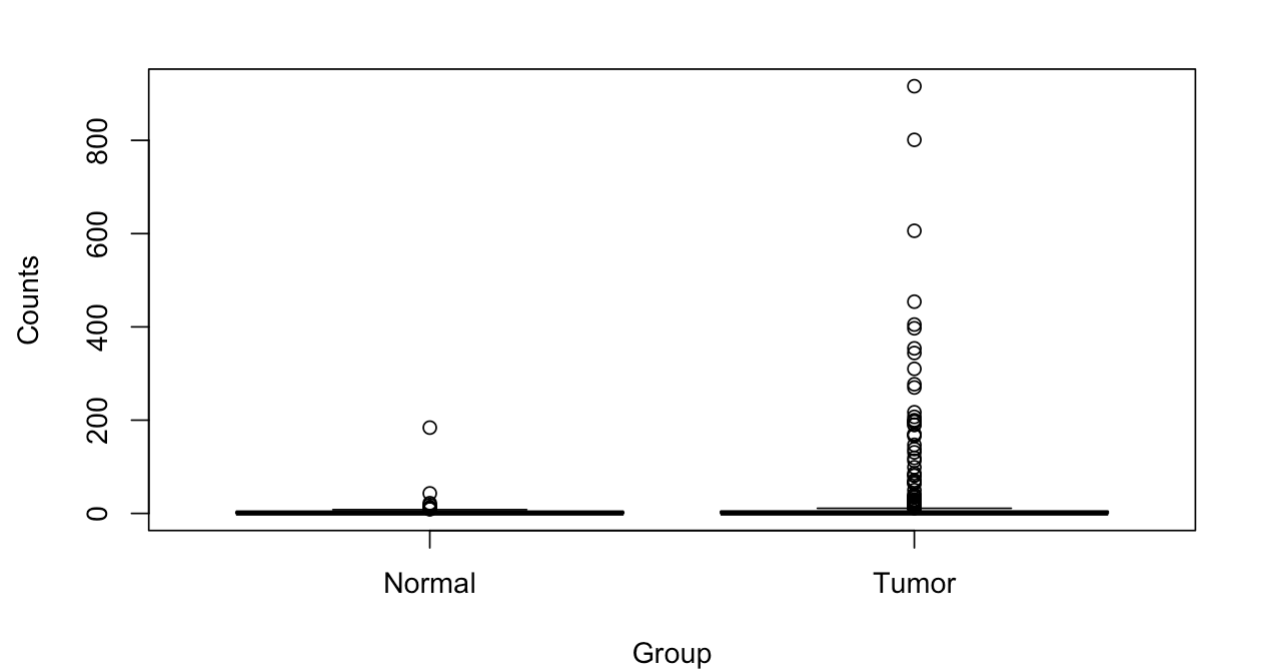


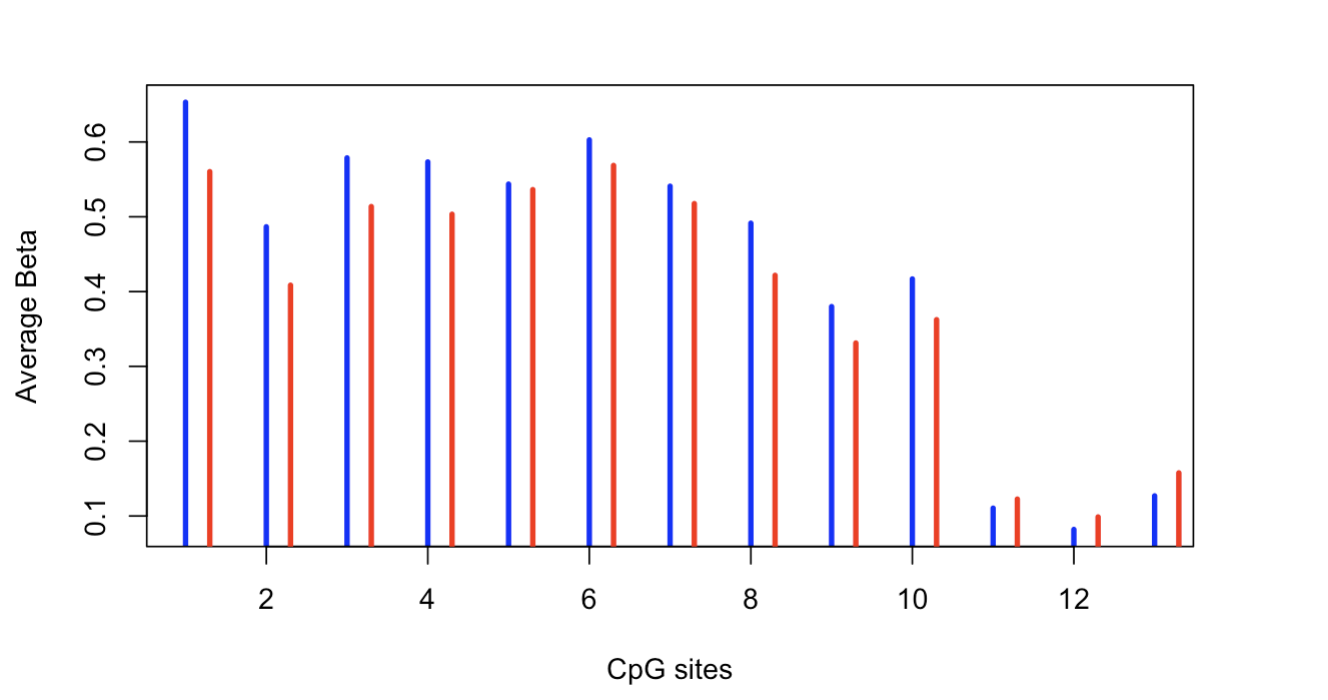
TNFSF11



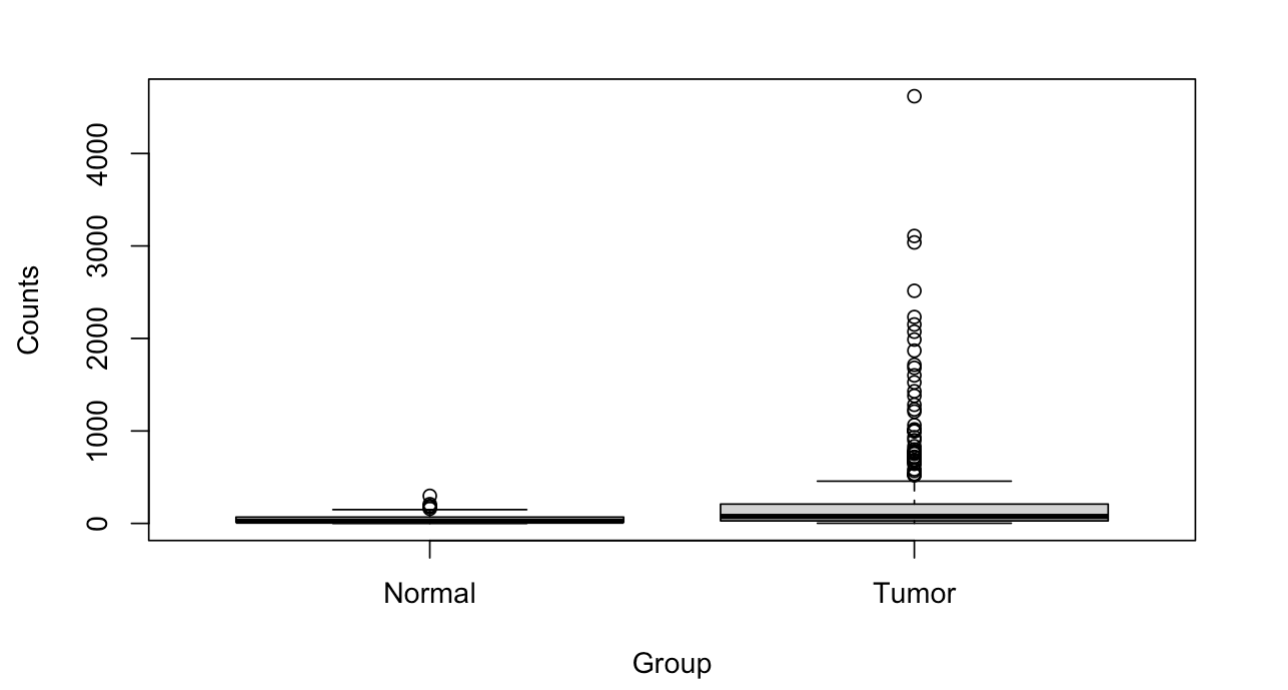
****

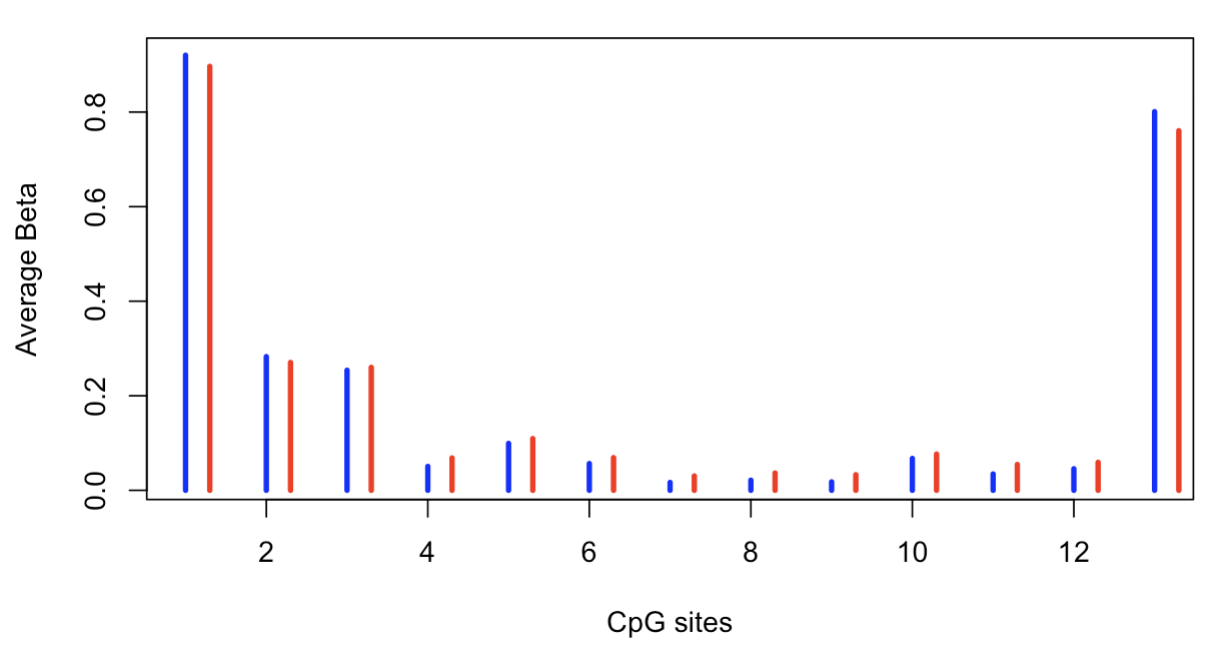
HOXD11

****

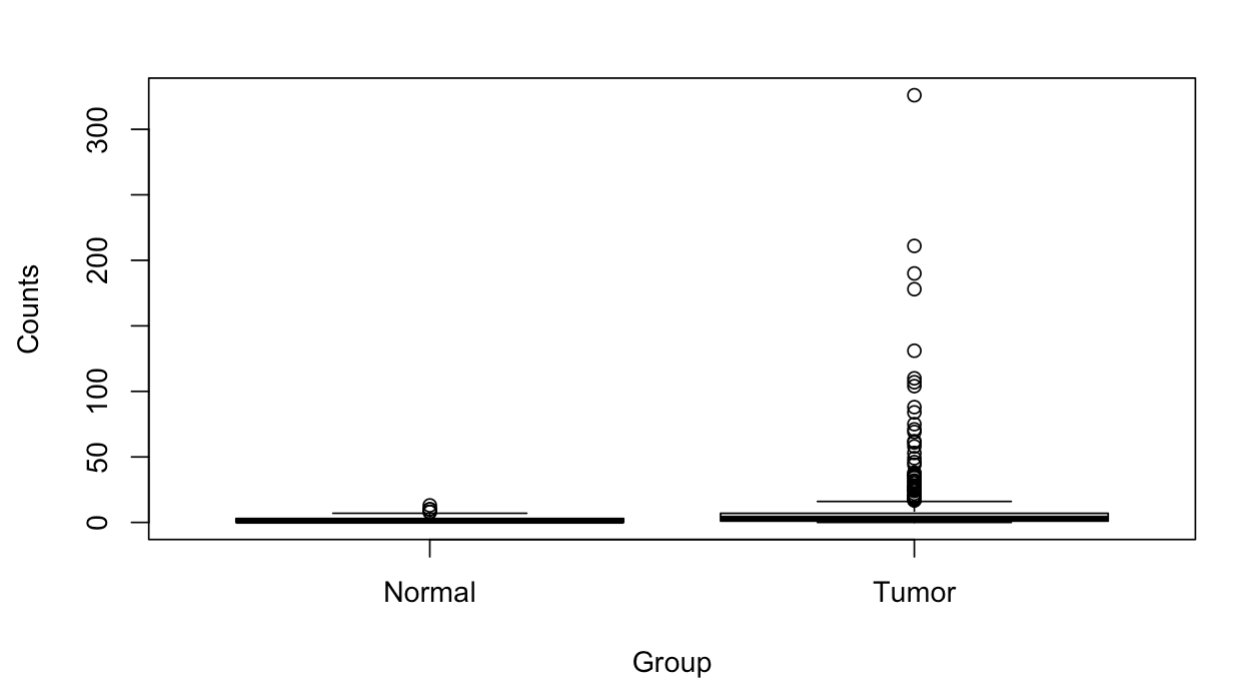
****

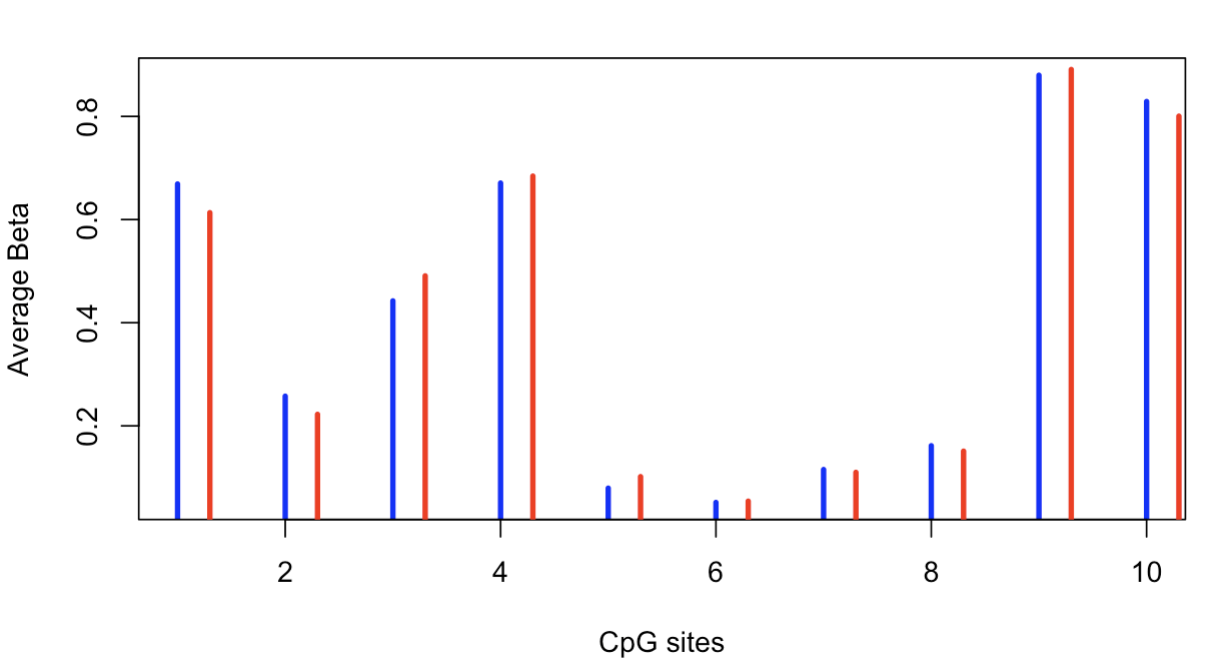
PLAC8

****

****

MDS2

****

****

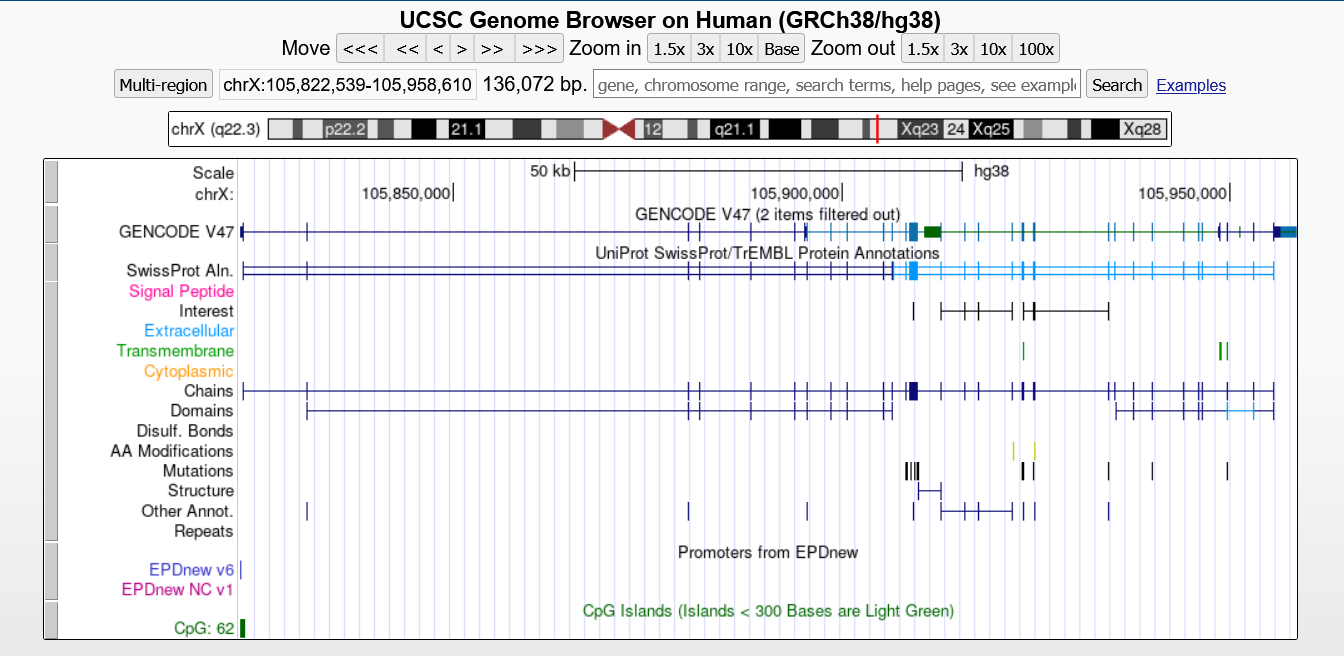
All 10 genes show a wider range of counts and more outliers in tumor samples in comparison to the normal group. They have potential to be related to metastasis or cancer progression. Their methylation patterns also vary a lot from site to site according to the bar graph. These sites with clear difference in beta values are worth for further study as they may correspond to genes showing distinct distributions between normal and tumor groups.

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

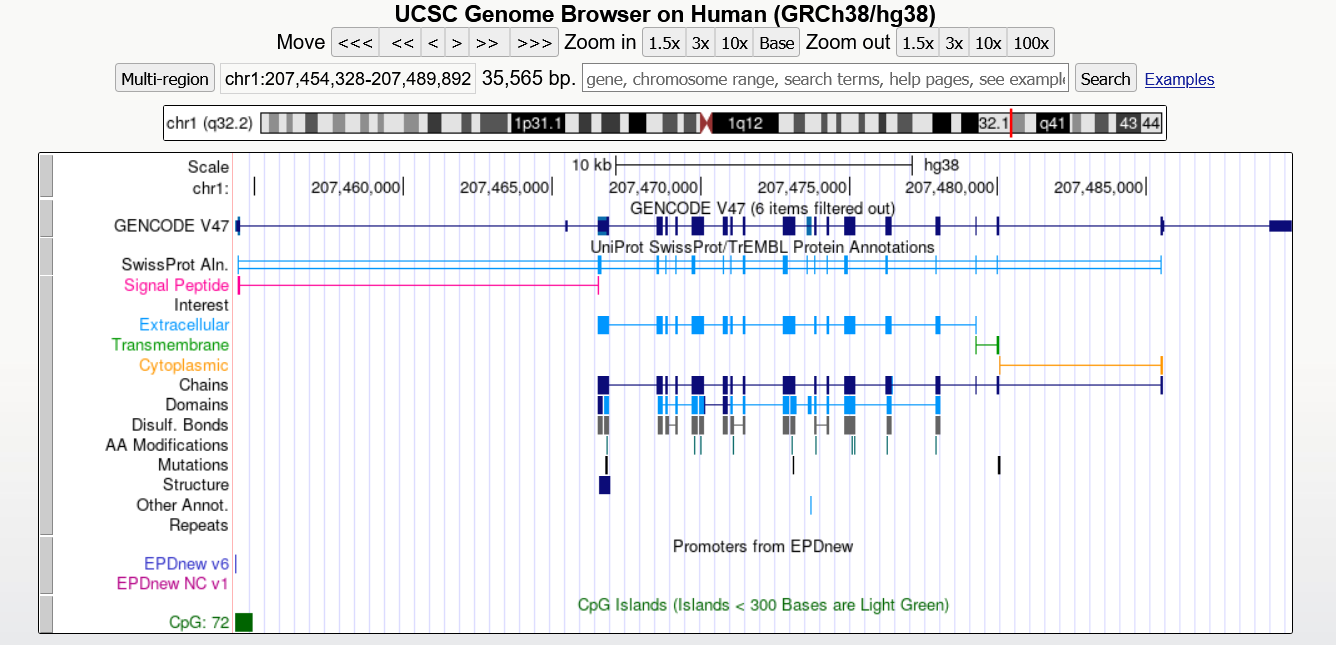
**for a few genes. Describe at least one academic article (research or review) that either supports or doesn’t support your final conclusion for one of the genes. If previously**

**published work doesn’t support your analysis, explain why this might be the case.**

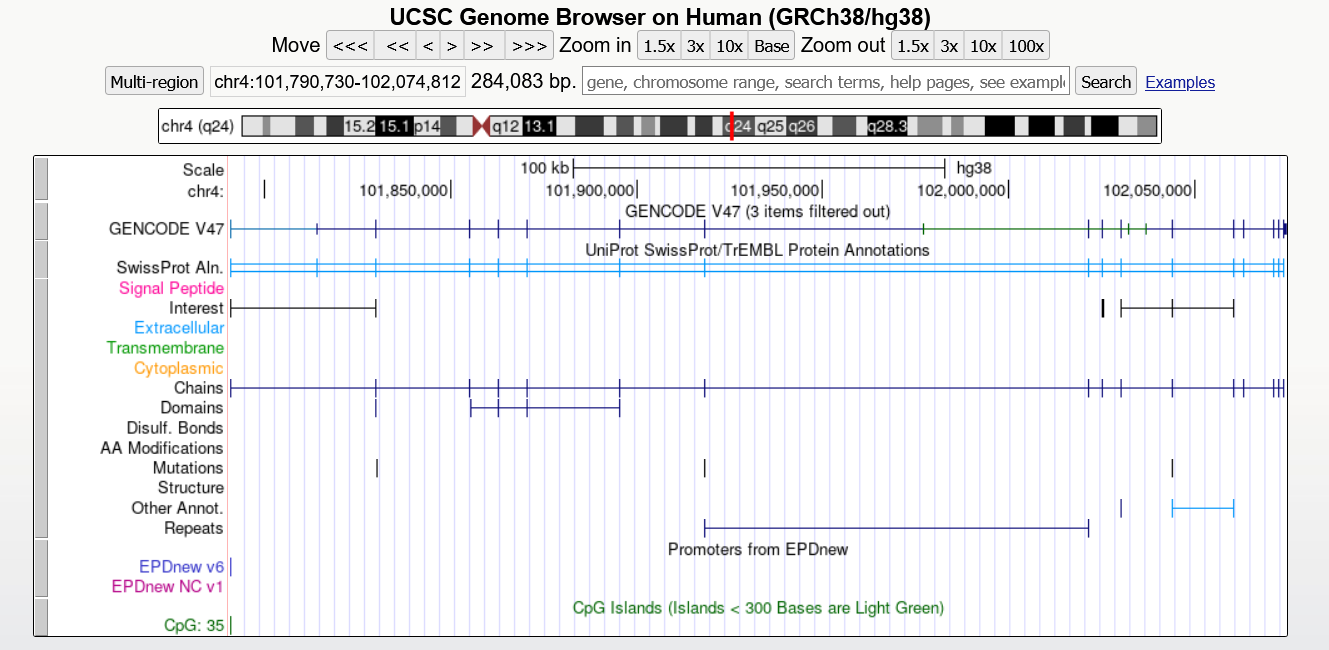
NRK



CR2



BANK1



Gulliver, G A et al. “Both conserved region 1 (CR1) and CR2 of the human papillomavirus type 16 E7 oncogene are required for induction of epidermal hyperplasia and tumor formation in transgenic mice.” *Journal of virology* vol. 71,8 (1997): 5905-14. doi:10.1128/JVI.71.8.5905-5914.1997

Even though there is no article directly linking CR2 with skin cancer, Gulliver et al., discovered that CR2 is essential for inducing epidermal hyperplasia and tumor formation in transgenic mice. The article demonstrates that CR2 is indispensable for tumor initiation and progression. Therefore, CR2 may influence metastasis by similarly regulating genes involved in cell proliferation, adhesion, and invasion. Nevertheless, further research should be conducted on the biological pathways of CR2 and its role in skin cancer progression to confirm this hypothesis.

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).

Reference:

Gulliver, G A et al. “Both conserved region 1 (CR1) and CR2 of the human papillomavirus type 16 E7 oncogene are required for induction of epidermal hyperplasia and tumor formation in transgenic mice.” *Journal of virology* vol. 71,8 (1997): 5905-14. doi:10.1128/JVI.71.8.5905-5914.1997

Tomczak, Katarzyna et al. "ReviewThe Cancer Genome Atlas (TCGA): an immeasurable source of knowledge." Contemporary Oncology/Współczesna Onkologia, 2015, pp. 68-77. doi:10.5114/wo.2014.47136.

Zhang, Zhuo et al. “A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data.” *Briefings in bioinformatics* vol. 20,4 (2019): 1524-1541. doi:10.1093/bib/bby023