QBIO 490: Directed Research – Multi-Omic Analysis

Fall 2024 Review Project

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 is important because it serves a repository for multi-omics data of a wide range of cancers. The data that it provides includes genomic, transcriptomic, epigenomic, methylation, clinical, and more. For this class, we used it to gather data from the TCGA-BRCA (Breast Cancer) cohort to guide us in understanding how to perform multi-omic analysis. Lastly, it's very valuable as the data is publicly available, allowing anyone to access and use the data.
2. **What are some strengths and weaknesses of TCGA?**
   As mentioned in the previous question, one big strength of the TCGA is that the data is publicly available which makes it easy to access and perform analysis on. TCGA houses large-scale data from various omics, including genomic, transcriptomic, epigenomic, and proteomic data across numerous cancer types. Lastly, TCGA has data for 33 different cancers making a comprehensive dataset (NCI Center for Cancer Genomics).

   For the weakness of TCGA, one thing I noticed was that there seems to be a lot of incomplete data in some of the clinical data as well as other datasets. This meant that we had to remove some patients from our study to include only those with valid data. Another weakness of TCGA is the lack of diversity present in the patient cohorts of the cancers. The underrepresentation of ethnic and racial groups could limit the understanding of the relationship between biological factors and environmental factors across various demographics.

Coding Skills

1. **What commands are used to save a file to your GitHub repository?**
   1. **cd** into the local repository where the file is saved (let's say that the file is "test_file.txt")
   2. Assuming that you have already created the file and it's present in your local repository, you do **"git status"** to check which files have local changes that need to be uploaded to Github
   3. git add test_file.txt
   4. git commit -m "[Informative message about file]"
   5. git push

2. **What command(s) must be run in order to use a package in R?**
   Let's say we want to use the package ggplot2 in R. First, we would install the package with "install.packages(ggplot2)". Then, we would load the package to make it usable with "library(ggplot2)".

3. **What command(s) must be run in order to use a Bioconductor package in R?**
   For example, if we want to use the Bioconductor package "TCGAbiolinks" in R, we would first run the command:
   if (!require("BiocManager")) install.packages("BiocManager")
   to install BiocManager as all Bioconductor packages are installed via the BiocManager package. Next, to install the TCGAbiolinks package, the following command is run:
   if (!require("TCGAbiolinks")) BiocManager::install("TCGAbiolinks")

4. **What is boolean indexing? What are some applications of it?**
   Boolean indexing is applying a vector of booleans (T/F) to a column/row in a dataframe. The term "boolean mask" is used to refer to a vector of boolean values. The following are applications of boolean indexing:
   1. Keeping certain data and rewriting into a new dataframe/overwrite existing dataframe
      a. Deleting null data (NAs)
      b. Subsetting data (young/old, male/female)
   2. Selecting certain data points based on given row/column
      a. Getting the patient ids of female patients

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.**
   Here's the sample dataframe (df) that I will be referring to in the example code below.

Patient_ID	Gene	Expression_level	Condition

patient_1	INPP5A	12.5	Healthy
patient_2	CLCN7	5.4	Diseased
patient_3	MAX	8.9	Healthy
patient_4	MLLT1	15.2	Diseased

a. an ifelse() statement

From the above df, I want to create a new column "high_expression_status" to indicate whether the Expression_level is above 10. The following code is used:
df$high_expression_status ← ifelse(df$Expression_Level > 10, "1", "0")

With this code, "df$Expression_Level > 10" evaluates to TRUE or FALSE for each row. The ifelse() command assigns "1" if the condition is TRUE and "0" if FALSE. 1 means yes and 0 mean no in one-hot encoding terms. The result of this command is that a new column, high_expression_status, is added to make the following output df:

Patient_ID	Gene	Expression_level	Condition	high_expression_status
patient_1	INPP5A	12.5	Healthy	1
patient_2	CLCN7	5.4	Diseased	0
patient_3	MAX	8.9	Healthy	0
patient_4	MLLT1	15.2	Diseased	1

b. boolean indexing

If I want to filter rows where the Condition is "Diseased", I would use the following boolean indexing code:
diseased_patients ← df[df$Condition == "Diseased",]
The **$Condition == "Diseased"** part of the code evaluates which rows have "Diseased" in the Condition column. df[**$Condition == "Diseased"**,] returns only the rows that meet this condition. A new dataframe, diseased_genes, is created with only "Diseased" rows:

Patient_ID	Gene	Expression_level	Condition	high_expression_status

patient_2	CLCN7	5.4	Diseased	0
patient_4	MLLT1	15.2	Diseased	1

SKCM Analysis – Results and Interpretations

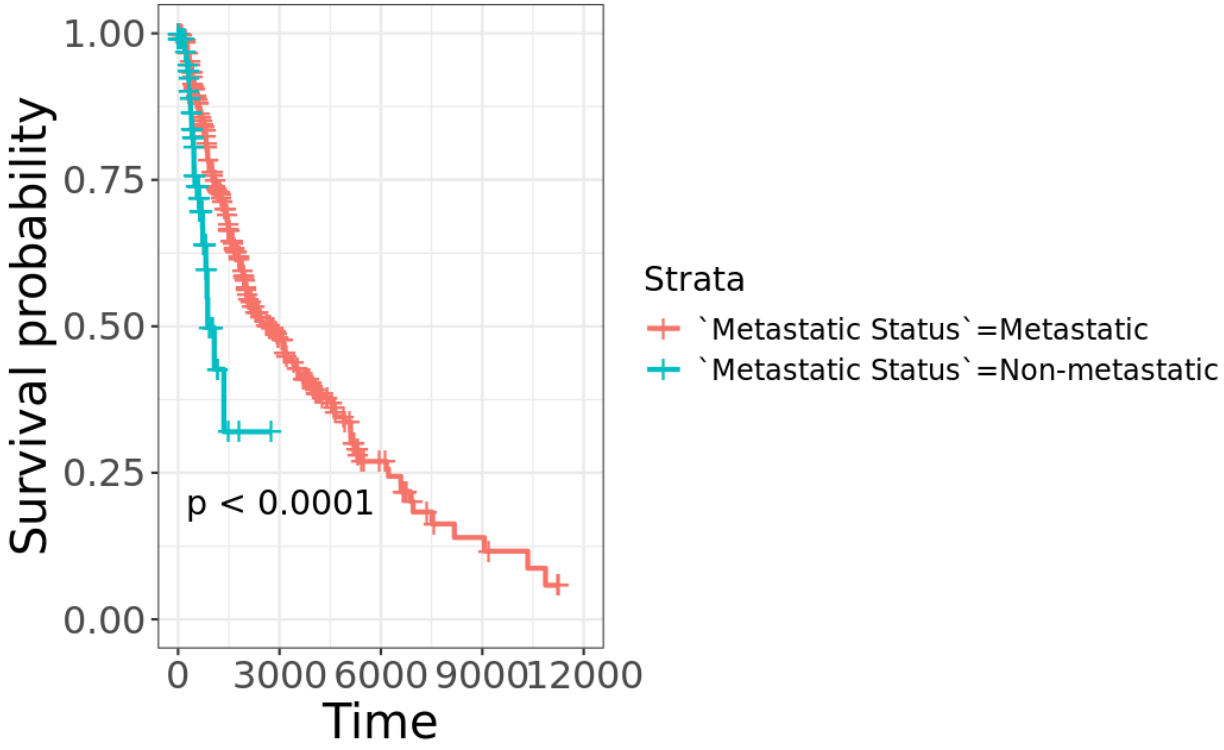
Research question: What are the differences between metastatic and non-metastatic SKCM across the epigenome and do these have any effect on the transcriptome?

Exploration of Methylation Patterns and Effect on Transcription

To do this, you must include at least the following analyses (at least 6 plots):
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?

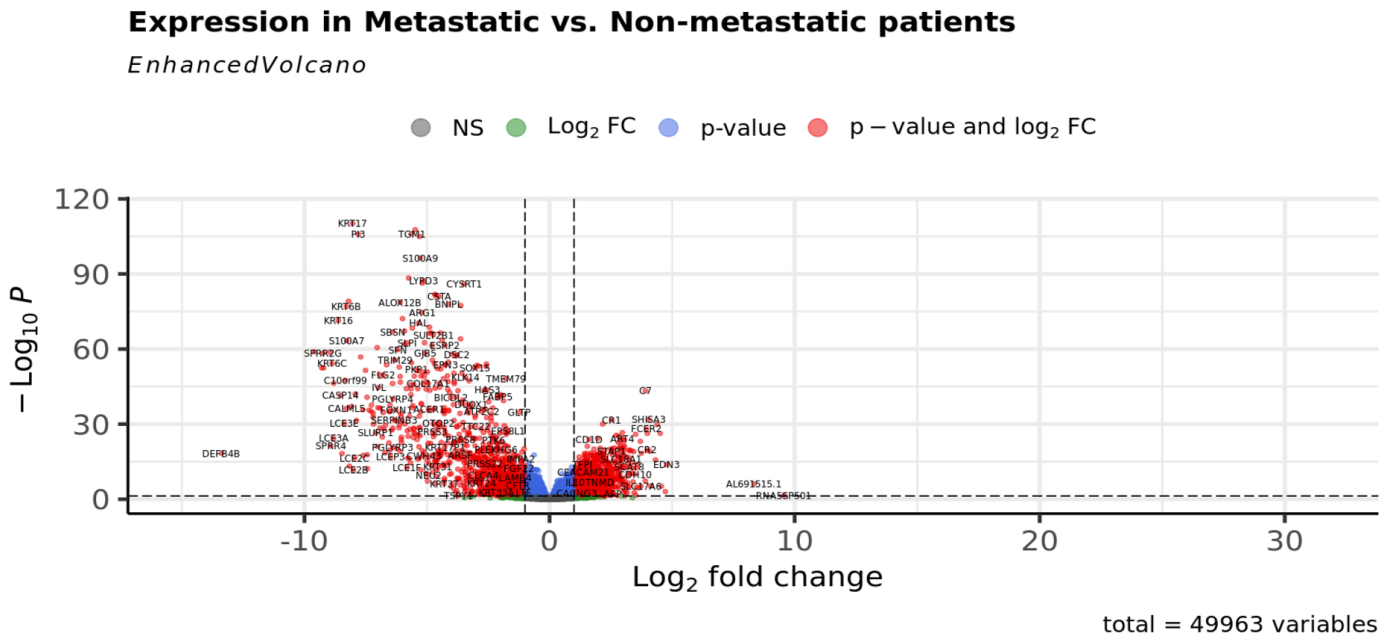
1. **Difference in survival between metastatic and non-metastatic patients (KM plot)**



This Kaplan-Meier plot is conveying that non-metastatic patients have a worse probability of survival than metastatic patients. This does not make sense biologically as in metastatic patients the cancer spreads and leads to a lower chance of survival. The p-value being less than 0.0001 tell us that there is a significant difference in survival probabilities between metastatic and non-metastatic TCGA SKCM patients. However,

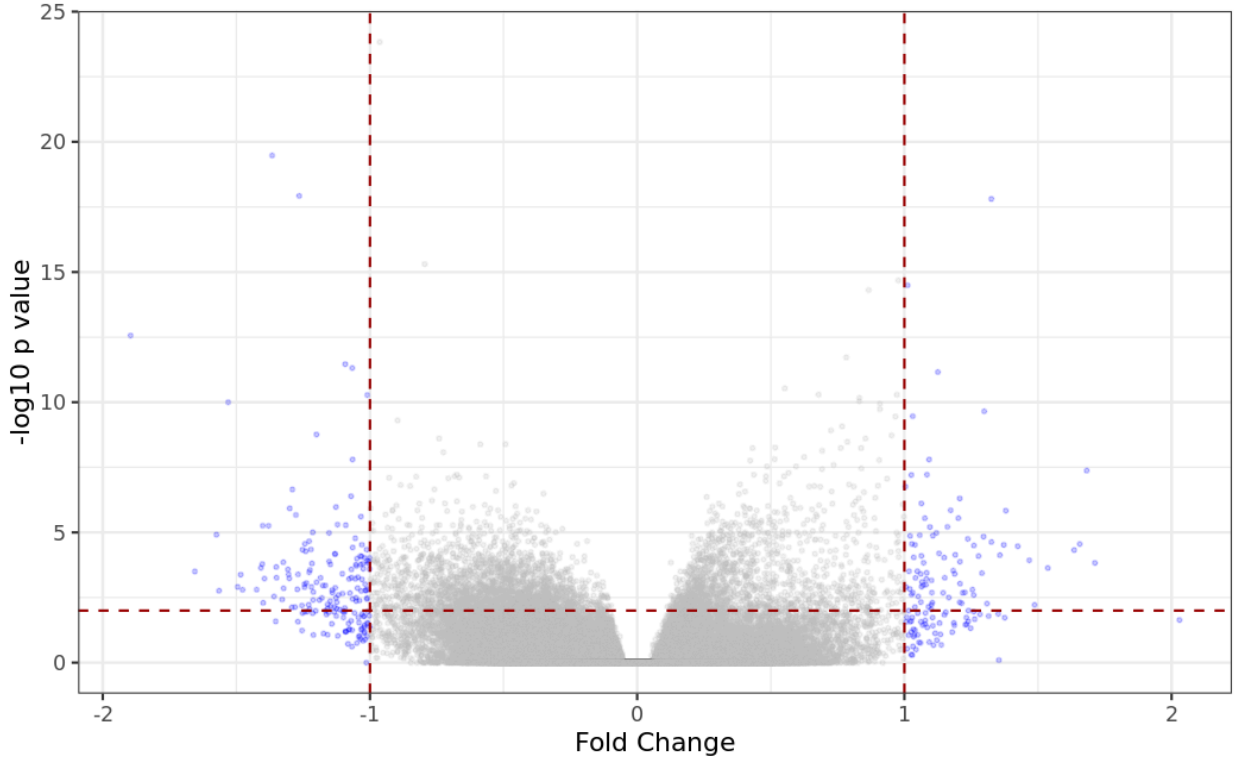
while this suggests a clear distinction in survival outcomes based on metastatic status, the plot does not provide information on potential confounding variables (e.g., treatment type, age, etc.) that might influence these results.

2. **Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status (DESeq2 + Volcano plot)**



This volcano plot highlights the expression in metastatic vs. non-metastatic patients. The plot suggests that specific genes are significantly upregulated or downregulated in metastatic versus non-metastatic patients. The genes found in the top left part of the volcano plot are downregulated in metastatic patients and are significant; KRT17, TGM1. Genes found more to the right are upregulated in metastatic patients; CR1, C7. As you move up the y-axis, that means genes become more significant. We can see that majority of genes are more downregulated in metastatic patients (when comparing expression levels to non-metastatic patients) than upregulated. Perhaps metastatic patients with skin cutaneous melanoma experience more of loss in gene expression than gain in gene expression.

3. **Naive differential methylation between non-metastatic and metastatic patients (Volcano plot)**



This volcano plot highlights the DNA methylation levels in metastatic vs. non-metastatic patients. The x-axis is "fold change" which refers to the relative change in DNA methylation between metastatic and non-metastatic patients. The y-axis tells us the significance level of each point on the plot. The plot seems to be evenly distributed as there is a level of symmetry present. Gray points represent methylation sites with no significant difference in methylation between the metastatic and non-metastatic groups. From the plot, we can conclude that there are several CpG sites with significant differential methylation between metastatic and non-metastatic patients, especially those with large fold changes (both positive and negative). These significantly different sites (blue points) could be of particular interest for further investigation into potential biomarkers for metastasis in this context. The majority of the methylation sites, however, have smaller fold changes or are not statistically significant (gray points), suggesting that most methylation differences are subtle or not robust enough to be considered significant in this analysis.

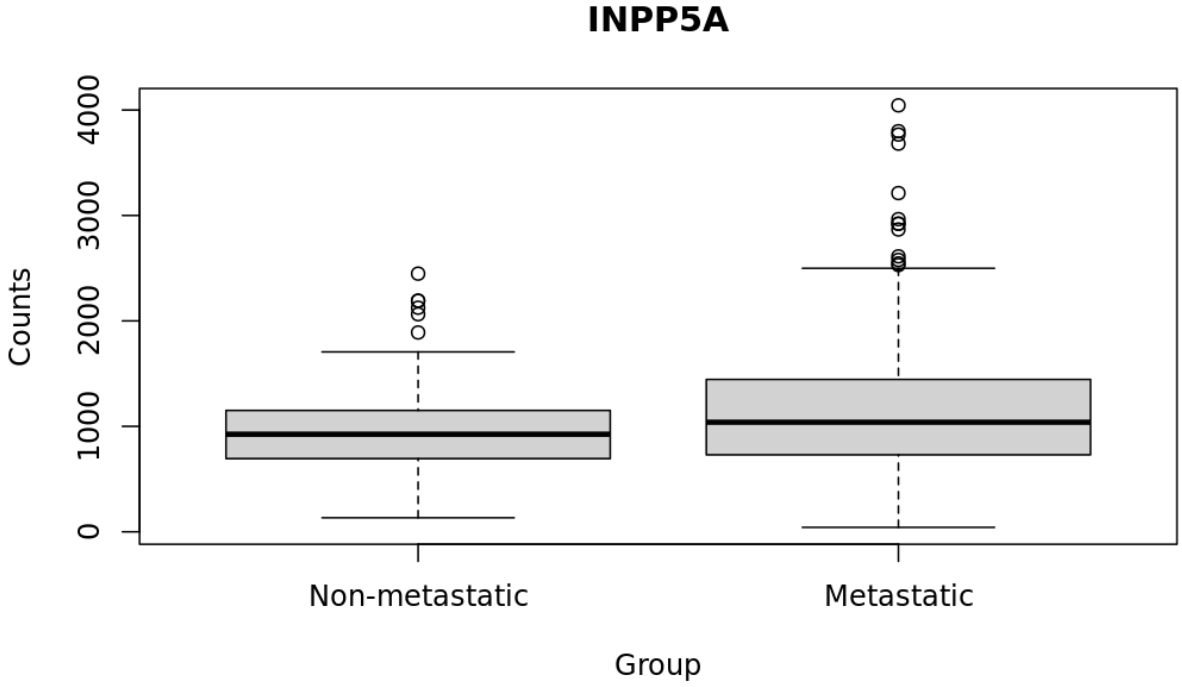
4. **Direct comparison of methylation status to transcriptional activity across non-metastatic vs metastatic patients for 10 genes**

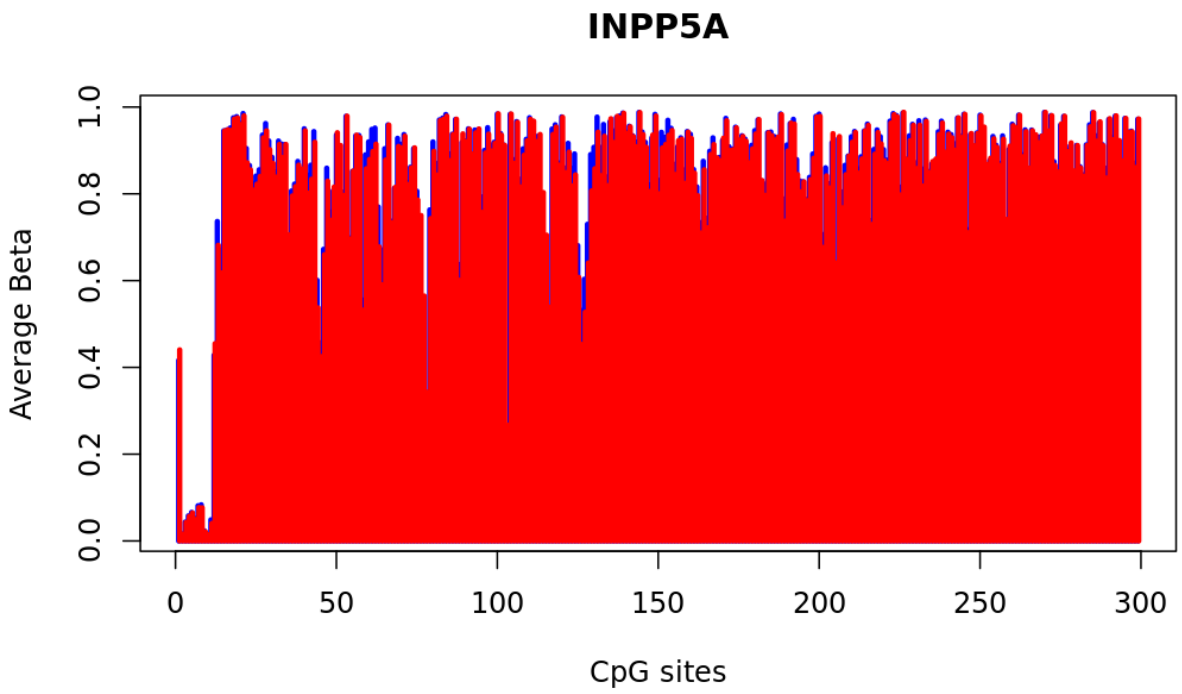
The 10 genes I looked at were: INPP5A, CLCN7, MAX, MLLT1, NPFFR1, TMEM163, TBC1D16, ANKRD11, OTUD3, B3GNTL1.

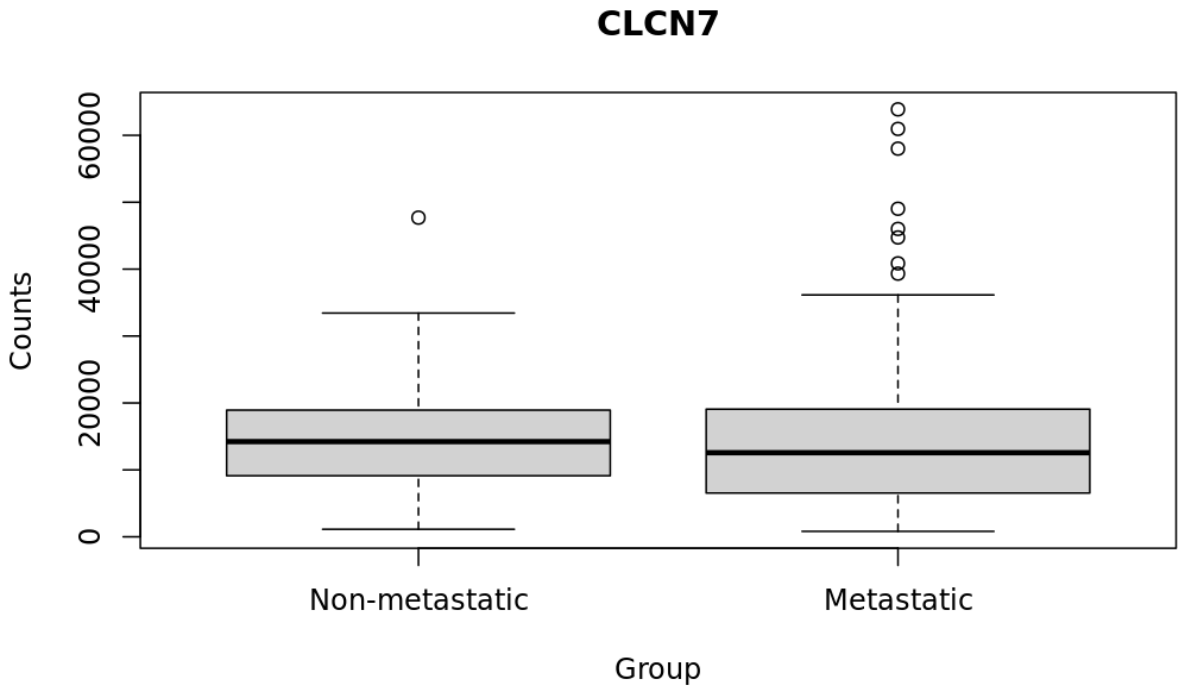
In all genes, when doing a box plot for direct comparison of gene expression levels between metastatic and non-metastatic patients, the metastatic counts is higher on average than the non-metastatic counts.

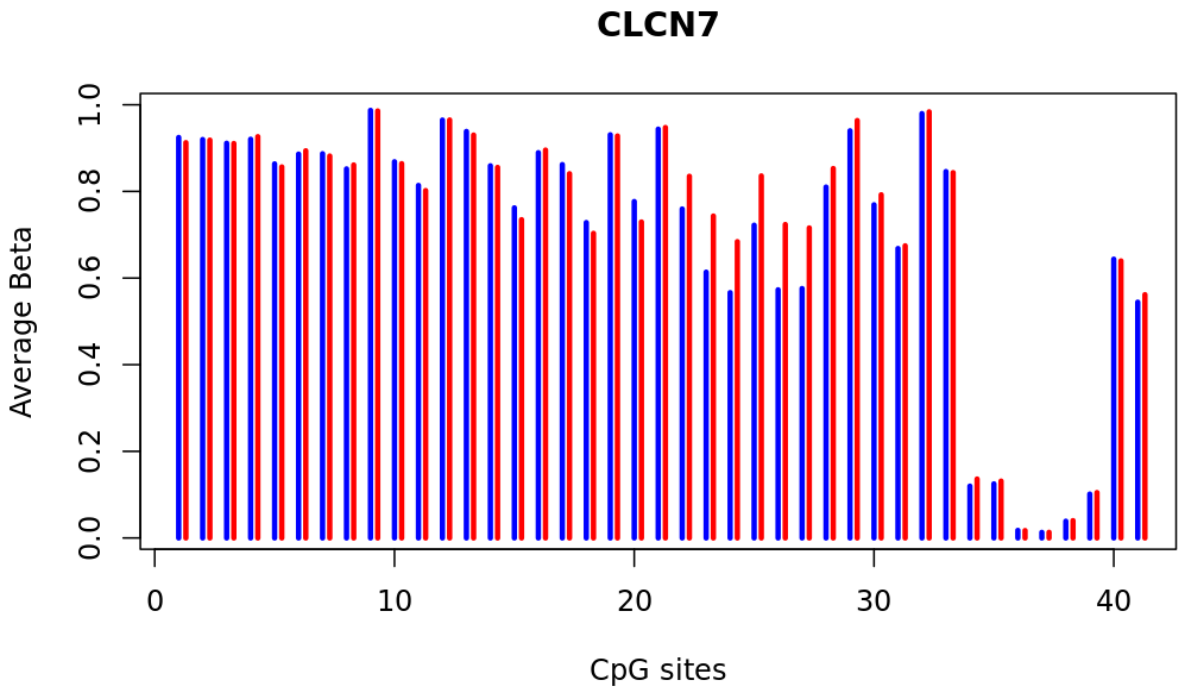
As for the colorful plots of red and blue, for every one of the CpG sites (varying numbers across the genes from ~12 to ~300 sites), we are plotting the methylation beta values of the non-metastatic patients (in blue) vs. metastatic patients (in red). For the most part we can see across all genes that there are instances where methylation levels are higher in metastatic patients compared to that of non-metastatic patients. At some CpG sites, there are also small differences of methylation.

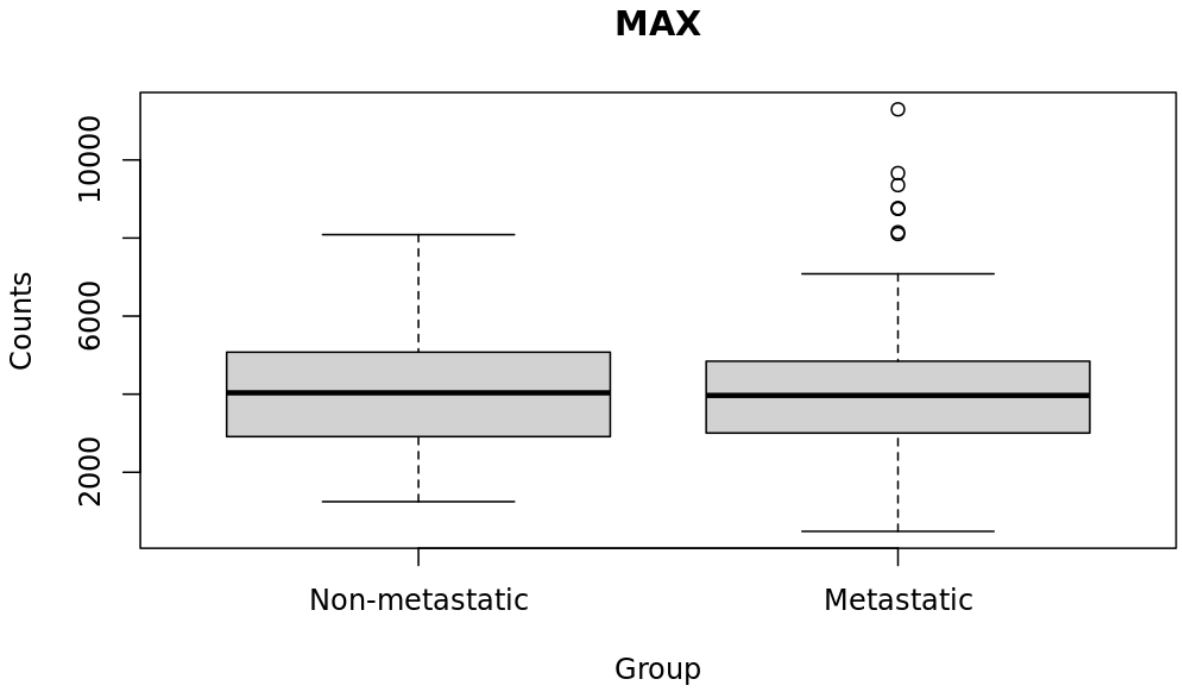
Similar to the conclusion from the volcano plot above, methylation differences may not be robust enough to be considered significant because we don't see significant or eye-opening differences in beta plots across the 10 genes.

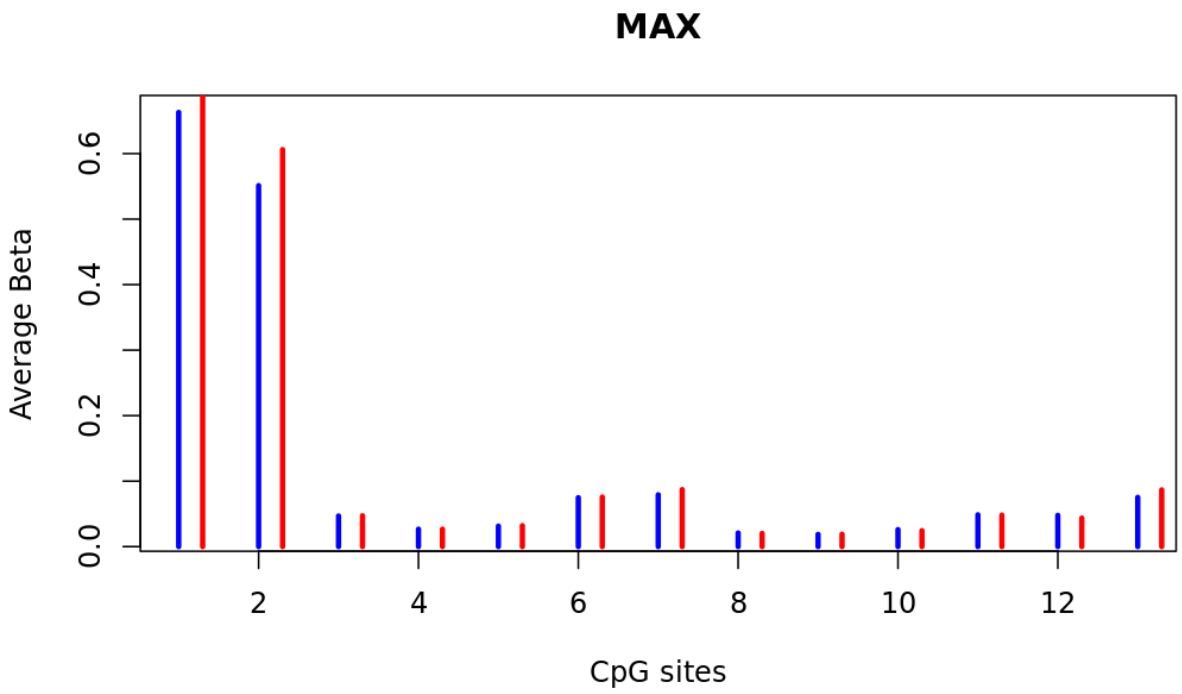


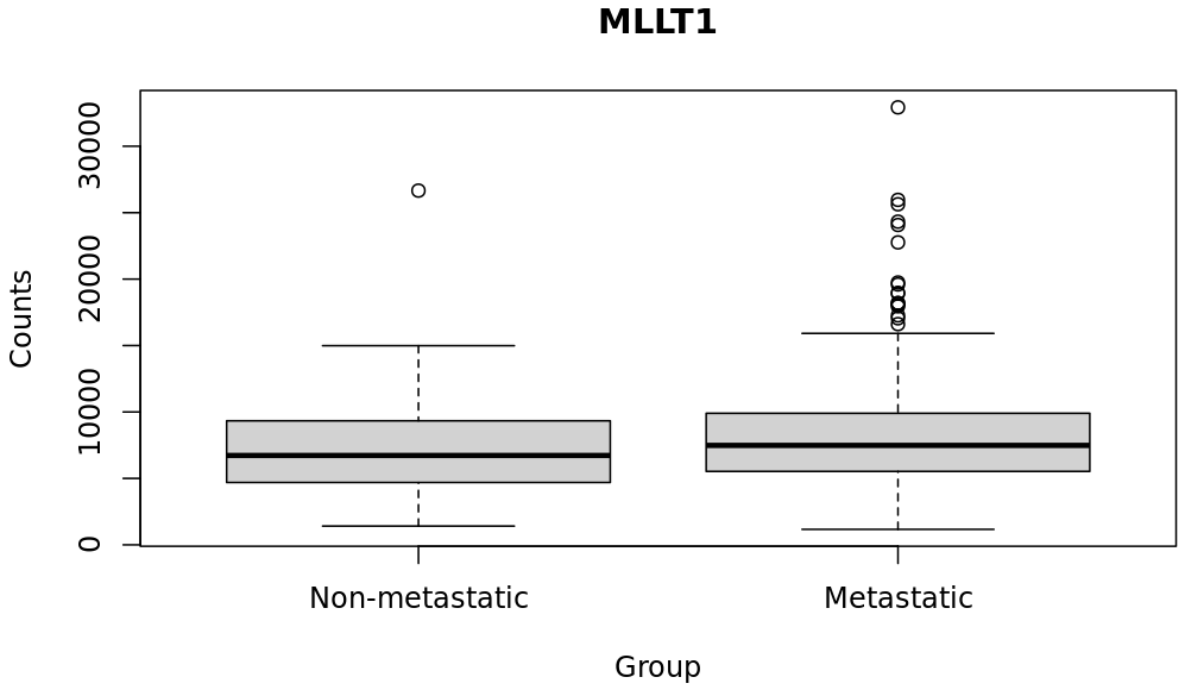

INPP5A
Average Beta
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
CpG sites

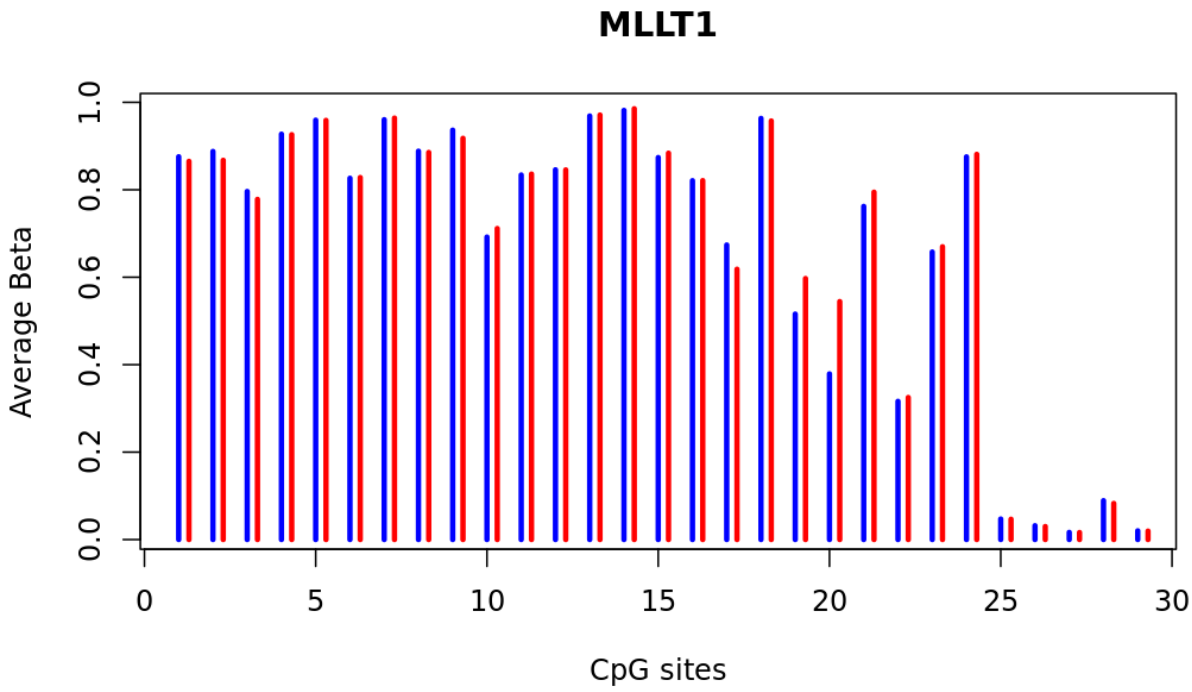

CLCN7
Counts
0
20000
40000
60000
Non-metastatic
Metastatic
Group

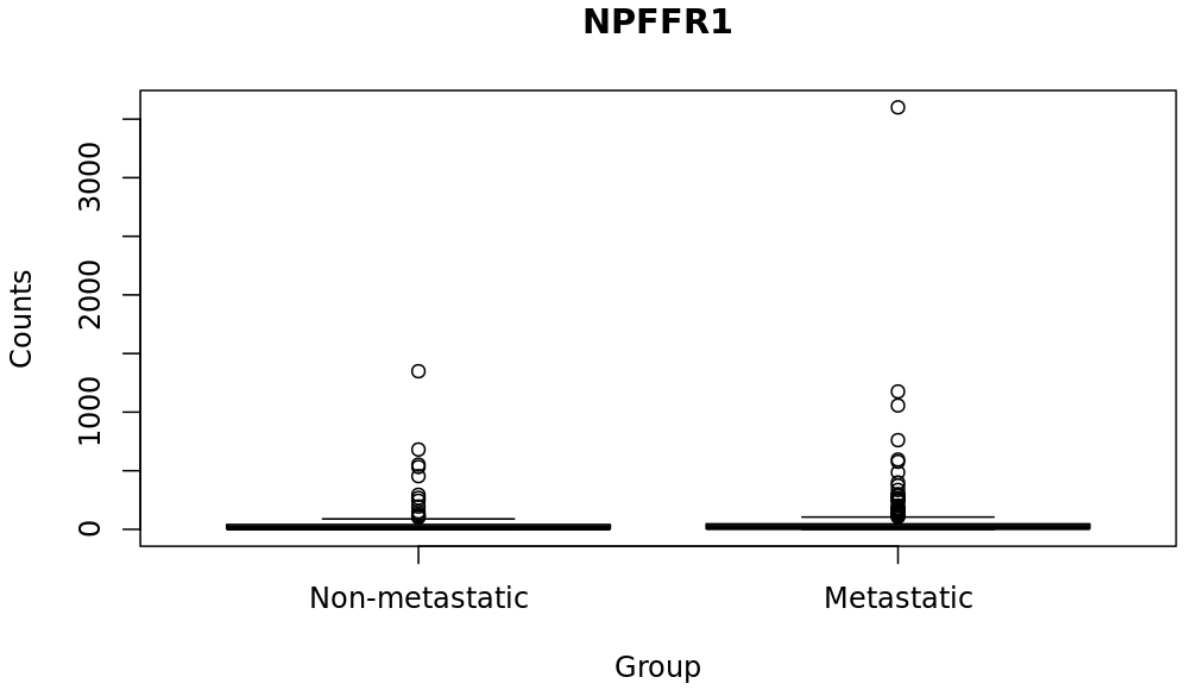

CLCN7
Average Beta
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
CpG sites

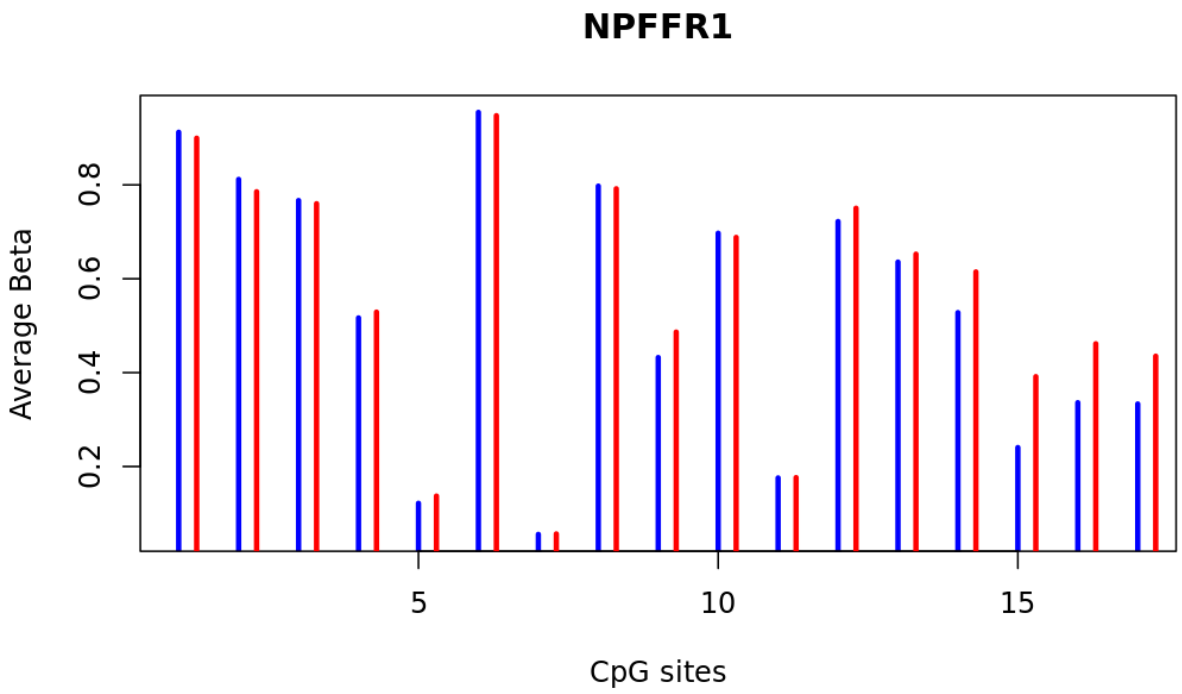

MAX
Counts
2000
6000
10000
Non-metastatic
Metastatic
Group

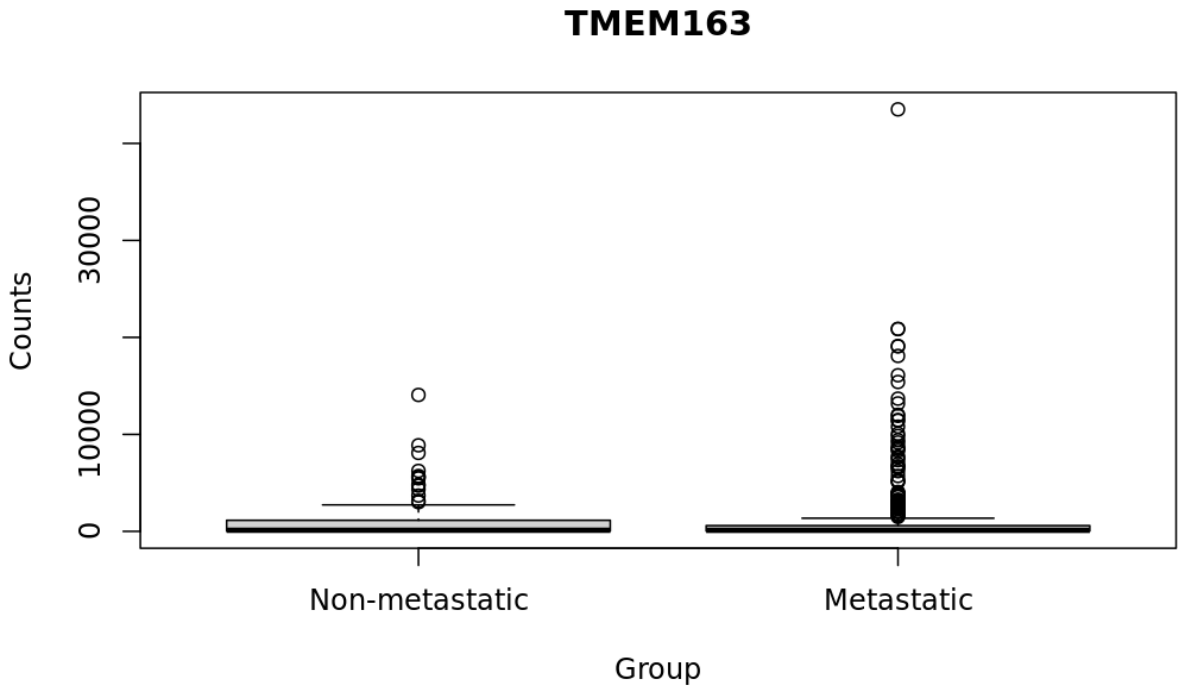

MAX
Average Beta
0.0
0.2
0.4
0.6
2
4
6
8
10
12
CpG sites

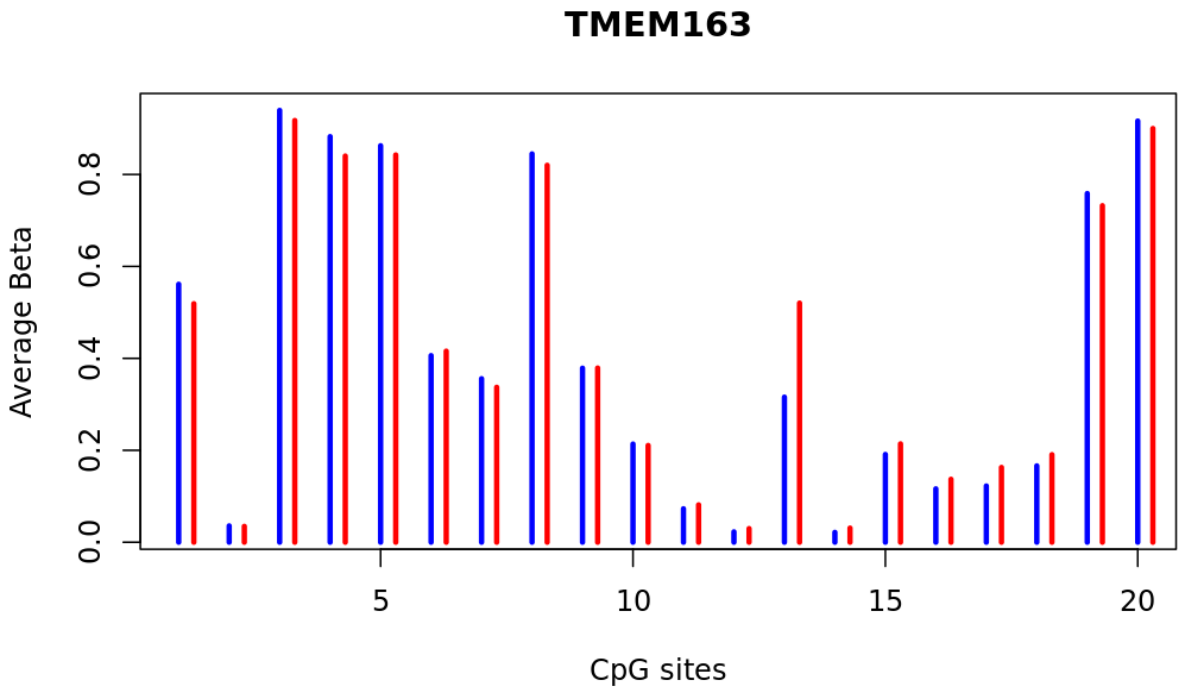

MLLT1
Counts
0
10000
20000
30000
Non-metastatic
Metastatic
Group

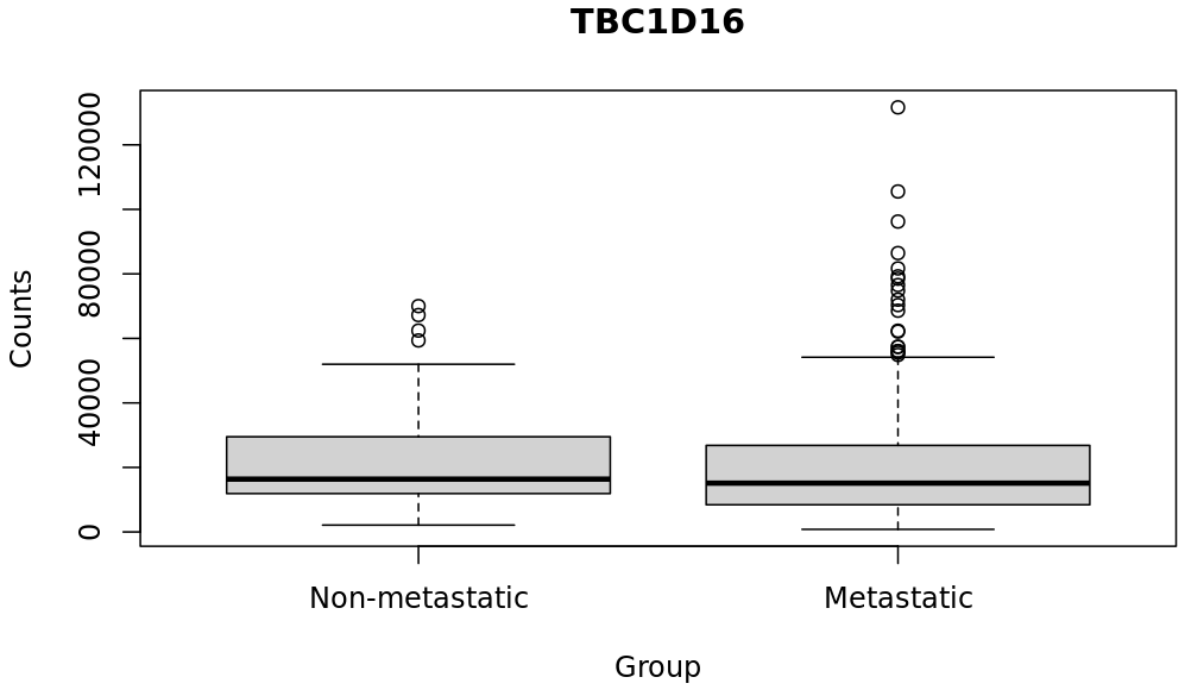

MLLT1
Average Beta
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
CpG sites

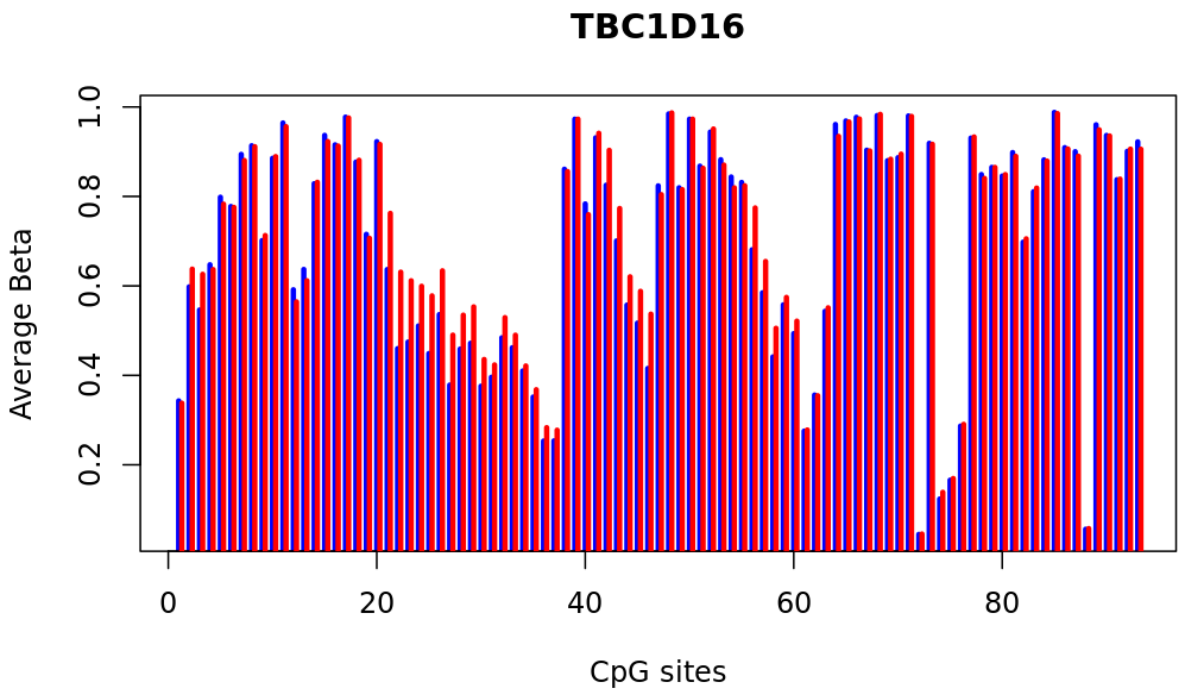

NPFFR1
Counts
0
1000
2000
3000
Non-metastatic
Metastatic
Group

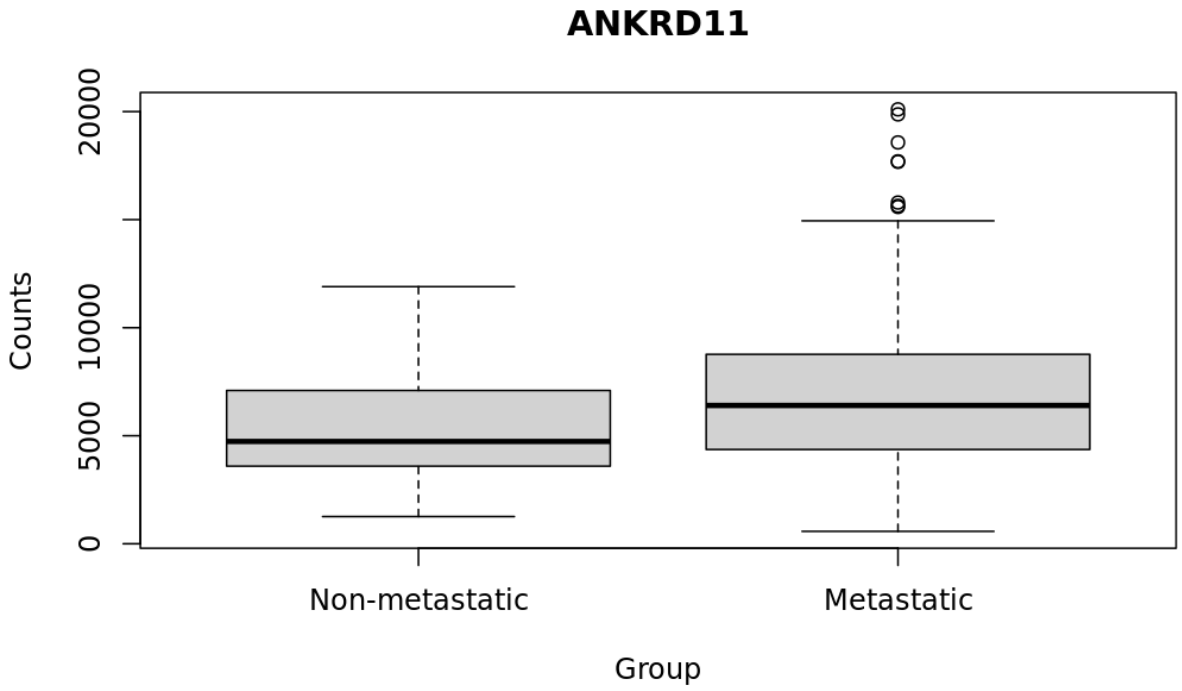

NPFFR1
Average Beta
0.2
0.4
0.6
0.8
5
10
15
CpG sites

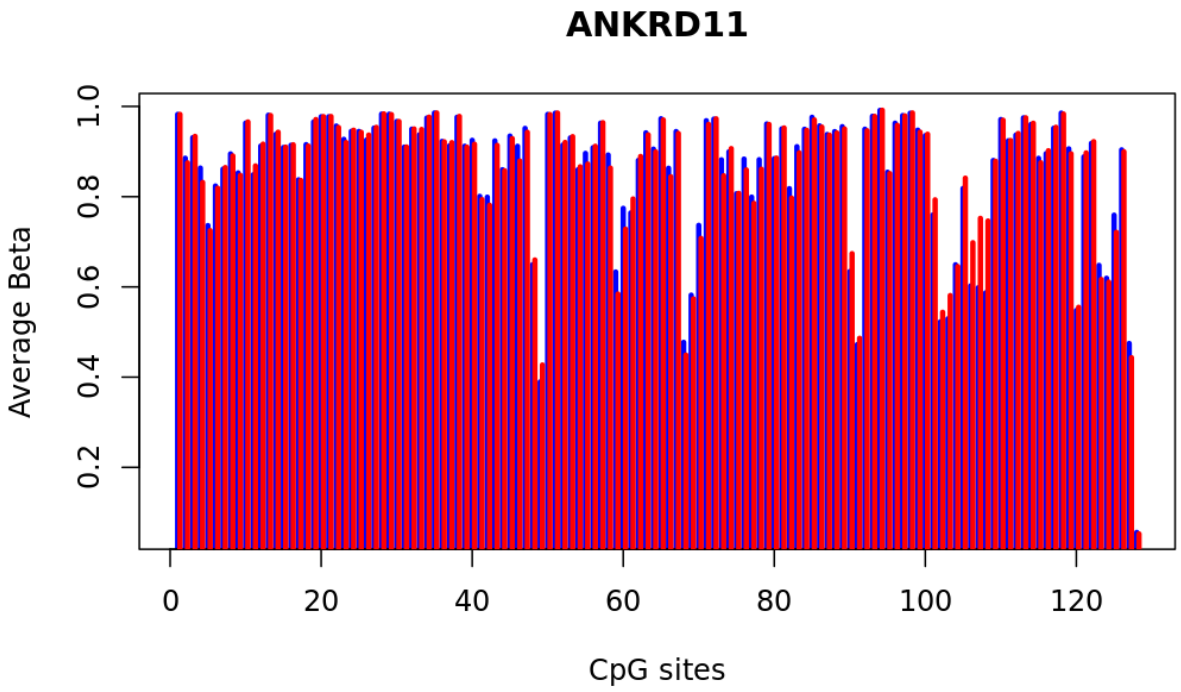

TMEM163
Counts
0
10000
30000
Non-metastatic
Metastatic
Group

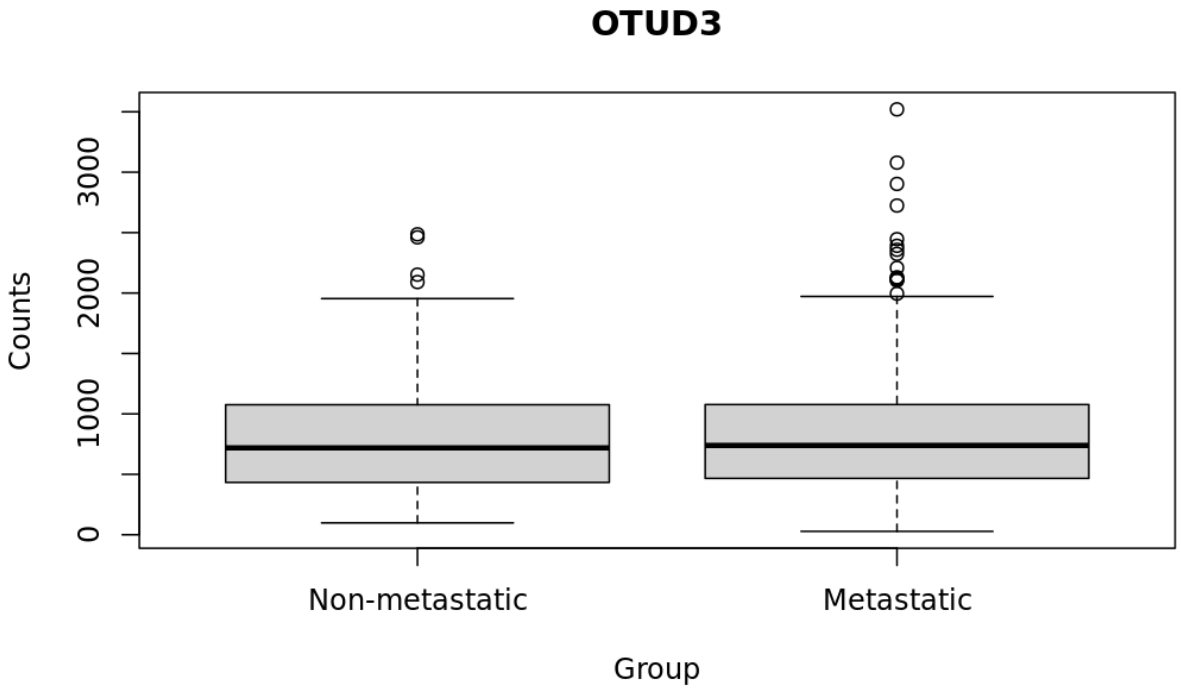

TMEM163
Average Beta
0.0
0.2
0.4
0.6
0.8
5
10
15
20
CpG sites

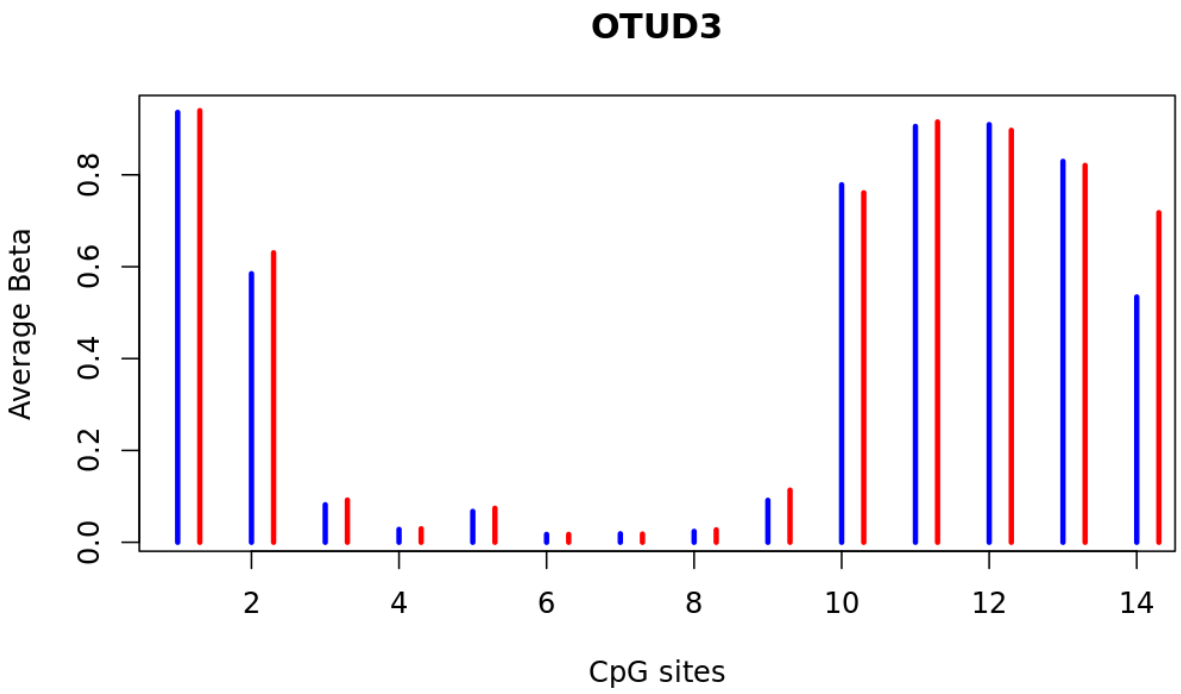

TBC1D16
Counts
0
40000
80000
120000
Non-metastatic
Metastatic
Group

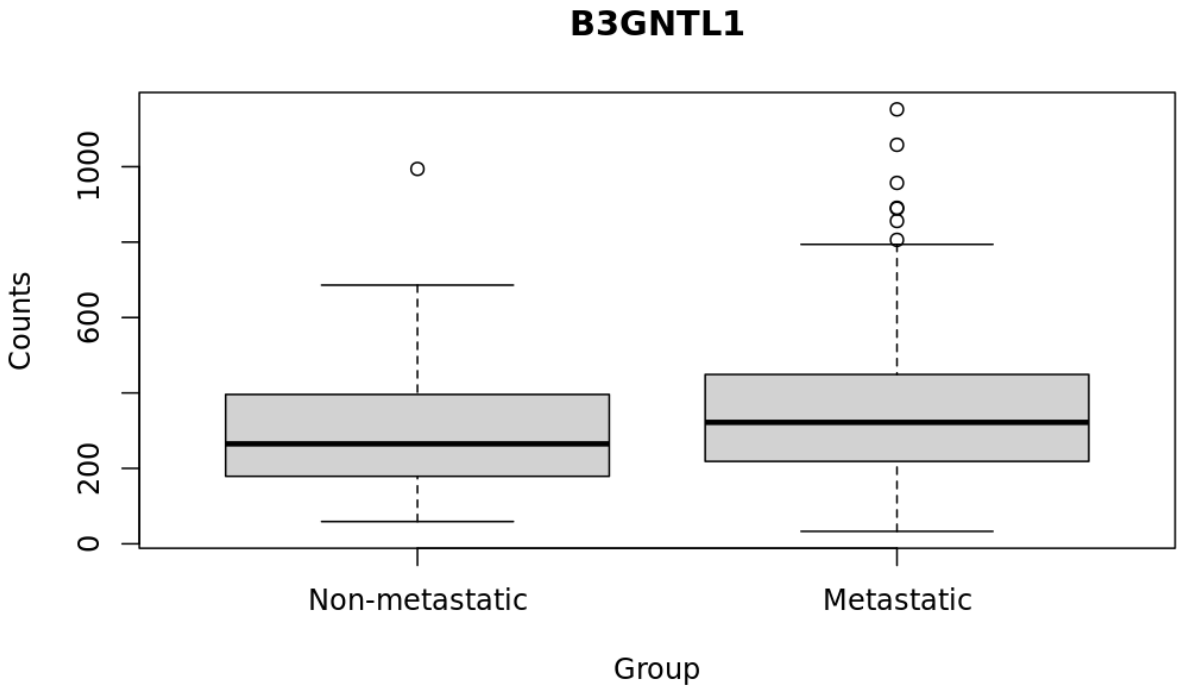

TBC1D16
Average Beta
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
CpG sites

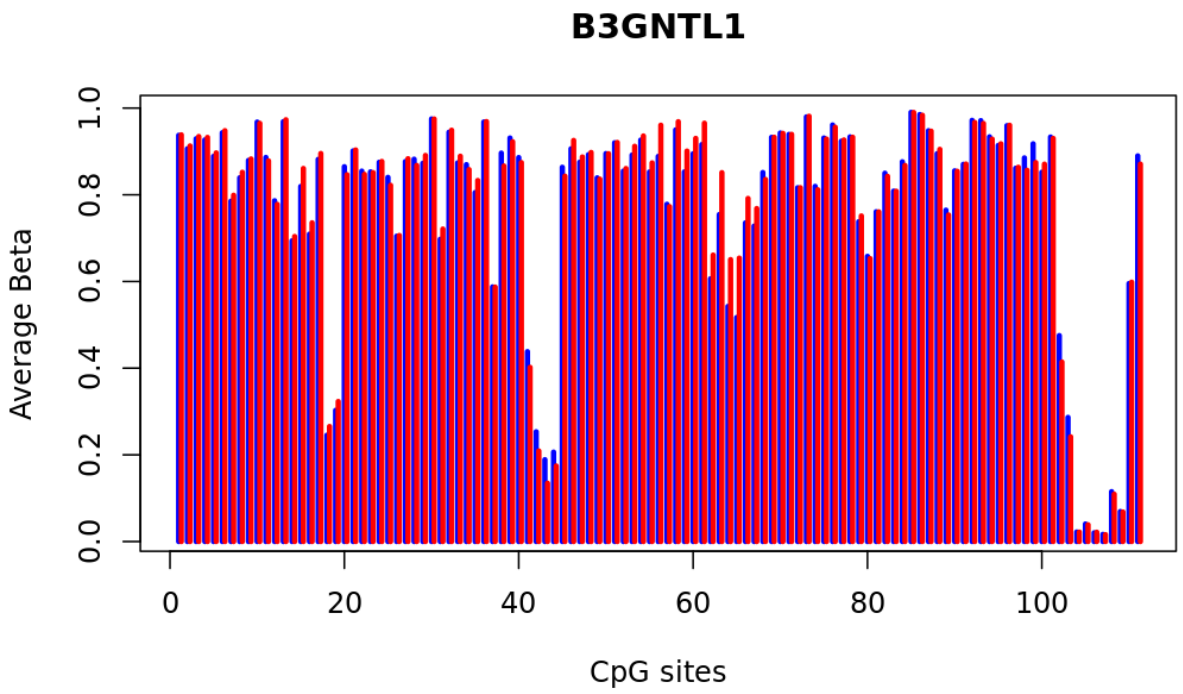

ANKRD11
Counts
0
5000
10000
20000
Non-metastatic
Metastatic
Group


ANKRD11
Average Beta
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
CpG sites


OTUD3
Counts
0
1000
2000
3000
Non-metastatic
Metastatic
Group

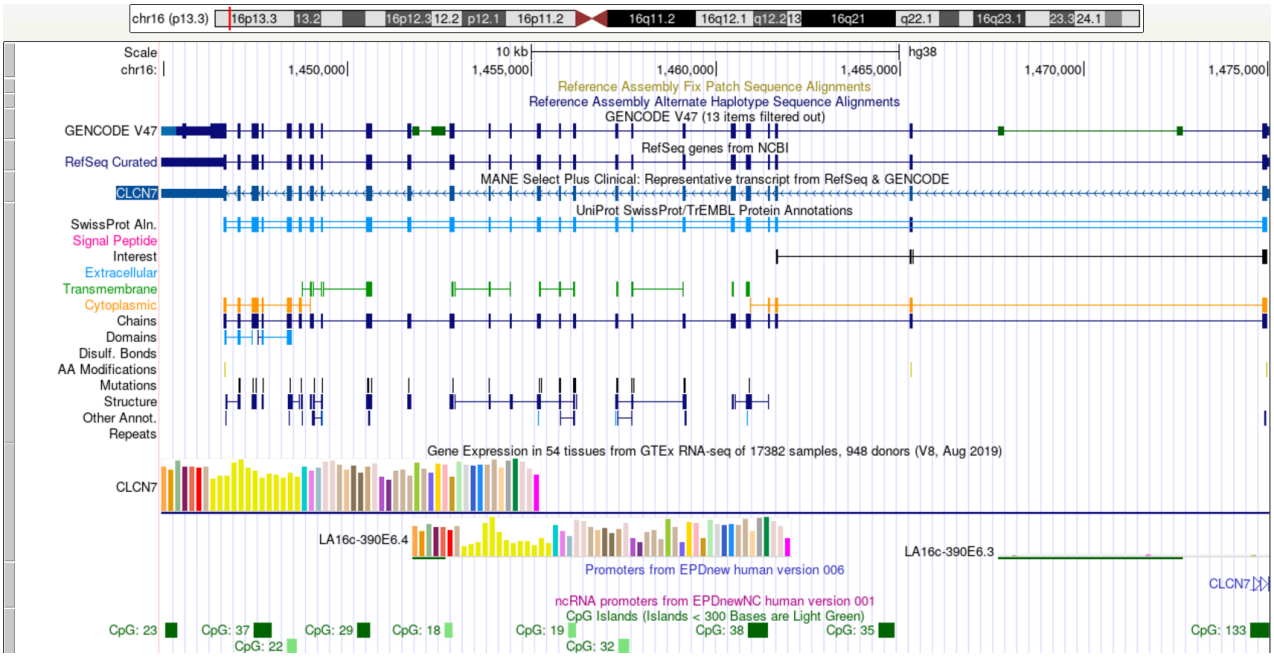

OTUD3
Average Beta
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10
12
14
CpG sites


B3GNTL1
Counts
0
200
600
1000
Non-metastatic
Metastatic
Group



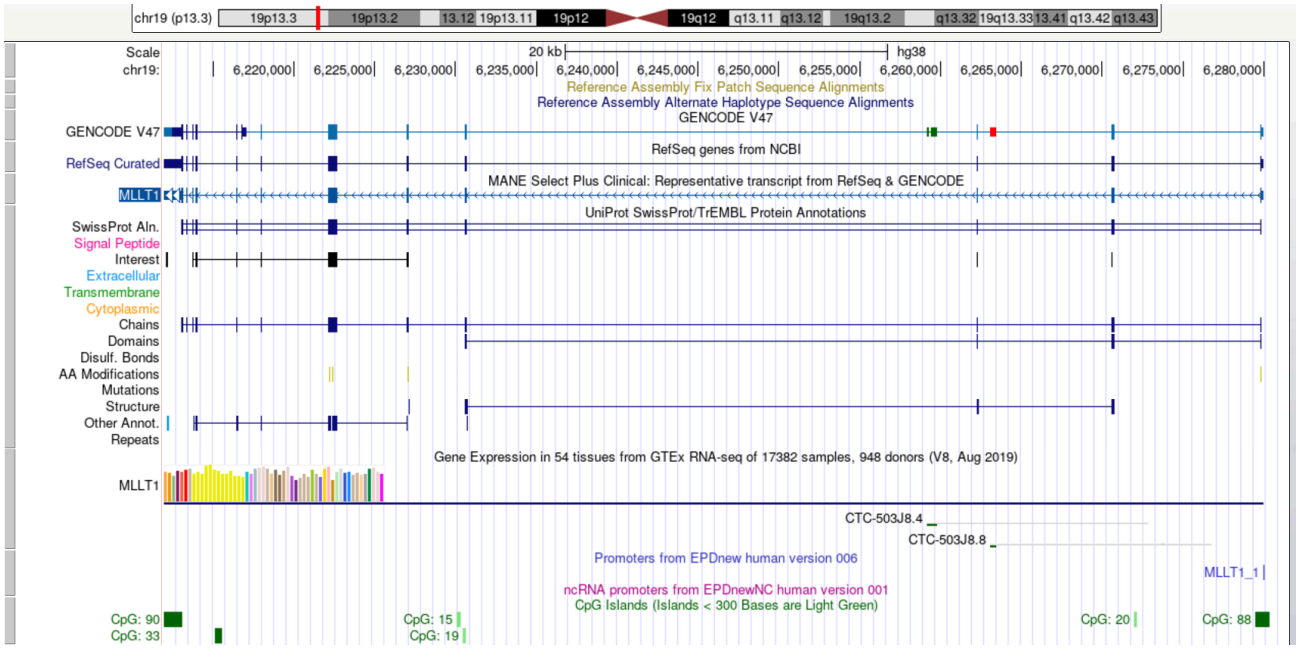
5. **Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser) for a few genes. Describe at least one academic article (research or review) that either supports or doesn't support your final conclusion for one of the genes. If previously published work doesn't support your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

Visualization of CpG sites and protein domains of **CLCN7**



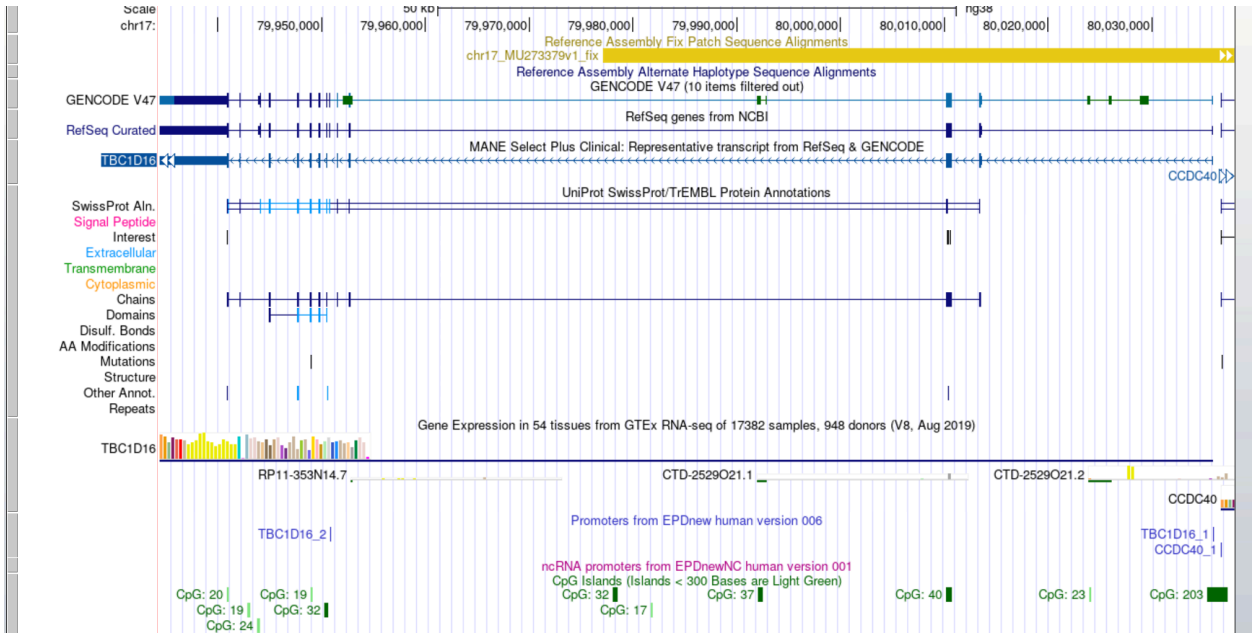
For CLCN7, there is a promotor region present and also a CpG island located right at the same location as well. There does not appear to be any signal peptide protein domains present where they code for parts of the protein which will be translated in peptide chains. There are protein domains present for transmembrane and cytoplasmic which refer to parts of the protein which will be extracellular and cytoplasmic.

Visualization of CpG sites and protein domains of **MLLT1**



For MLLT1, there is a promotor region present and also a CpG island located right at the same location as well. There does not appear to be any signal peptide protein domains present where they code for parts of the protein which will be translated in peptide chains. The absence of transmembrane regions or extracellular domains further supports the gene's intracellular localization.

Visualization of CpG sites and protein domains of **TBC1D16**



For TBC1D16, there is a promotor region present and also a CpG island located right at the same location as well. There does not appear to be any signal peptide protein domains present where they code for parts of the protein which will be translated in peptide chains. There are some “interest” protein domains which are regions that have been experimentally defined, such as the role of a region in mediating protein-protein interactions or some other biological process.

The paper, “Epigenetic activation of a cryptic TBC1D16 transcript enhances melanoma progression by targeting EGFR”, confirms that TBC1D16 is involved in skin cutaneous melanoma as it progresses melanoma by EGFR. In the paper, it states that they “found a hypomethylation event that reactivates a cryptic transcript of the Rab GTPase activating protein TBC1D16 to be a characteristic feature of the metastatic cascade. This short isoform of TBC1D16 exacerbates melanoma growth and metastasis both in vitro and in vivo” (Vizoso et al.).

References

"The Cancer Genome Atlas Program (TCGA)." NCI, www.cancer.gov/ccg/research/genome-sequencing/tcga. Accessed 24 Nov. 2024.

Vizoso, Miguel et al. "Epigenetic activation of a cryptic TBC1D16 transcript enhances melanoma progression by targeting EGFR." Nature medicine vol. 21,7 (2015): 741-50. doi:10.1038/nm.3863