General Concepts

1. What is TCGA and why is it important?
    a. TCGA stands for The Cancer Genome Atlas Program, which is a large public database that contains multi-omic data of thousands of patients with cancer. This is important since it allows researchers to be able to run statistical tests to notice any patterns and better understand the causes and predictors of cancer.
2. What are some strengths and weaknesses of TCGA?
    a. Some strengths of TCGA are that it has a large sample size (over 20,000) that covers 33 different cancer types. There is also a variety of data: clinical, mutation, RNA count, and methylation. This format and its accessible format allows it to be a very useful resource in comparing all sorts of variables.
    b. One issue with TCGA is that the majority of its patients are identified as White. Due to the small sample size of other races, it can be difficult to determine patterns that differ based on race. The classifications of race also tend to be very broad. This is a common theme amongst other variables, where the description is broad, which allows for less precise analysis. Since no experiments were run to generate this data, it is also not possible to determine causation from our results.

Coding Skills

1. What commands are used to save a file to your GitHub repository?
    a. Git status, git add file_name, git commit -m “message”, git push
2. What command(s) must be run in order to use a package in R?
    a. install.packages(“package_name”)
    b. library(package_name)
3. What command(s) must be run in order to use a Bioconductor package in R?
    a. install.package(“BiocManager”)
    b. library(BiocManager)
    c. Install and call any other needed packages (ex. TCGAbiolinks)
    d. Query, download, and prepare data
4. What is boolean indexing? What are some applications of it?
    a. Boolean indexing involves creating a mask vector with only true and false values and applying that mask to a row or column in a dataframe. Data that has a false value will be removed, while data with a true value will be kept.
    b. This can be useful for data preprocessing and screening for data based on certain criteria.
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.

Patient	Gender	Age
A	Male	54
B	Female	28

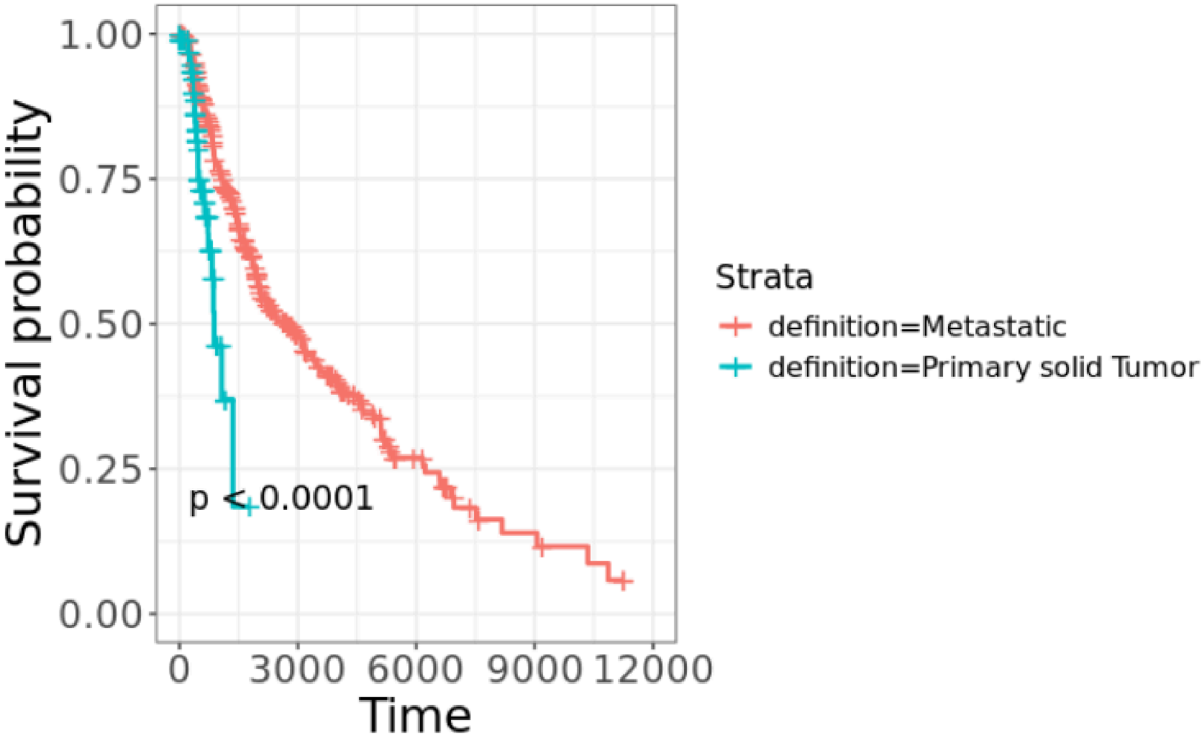
C	Female	72
D	Male	47

b. old_mask <- ifelse(df$Age > 50, True, False)
    i. In age column, applies true is patient age is over 50 but false if below or equal to

c. df [old_mask,]
    i. Applying mask to rows of dataframe removes patients who have false values in mask (age less than 50) leaving only “old” patients

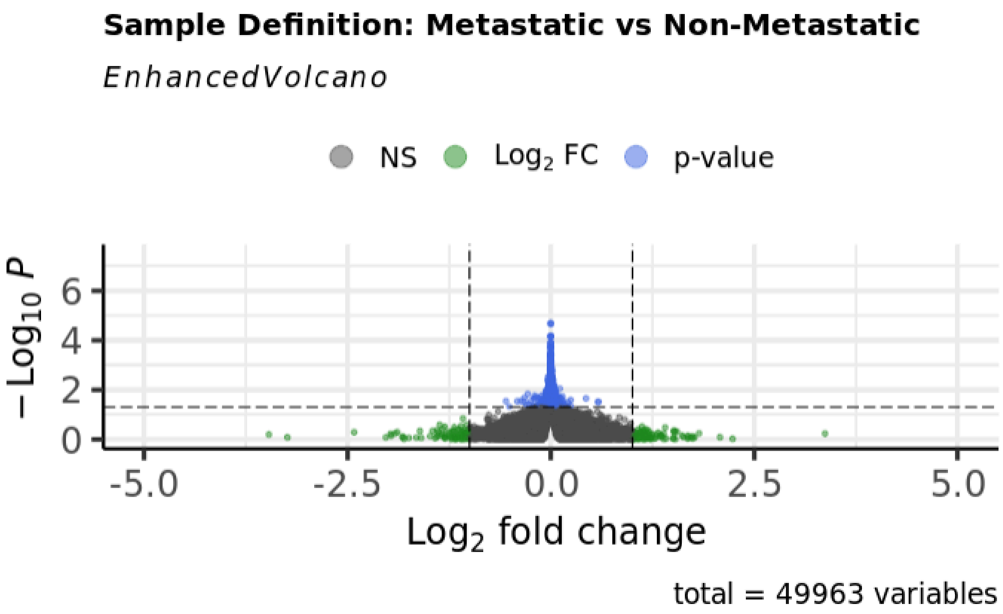
Analyze the plot. What conclusions can you and can you not draw about differences between metastatic and non-metastatic TCGA SKCM patients? Why?

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



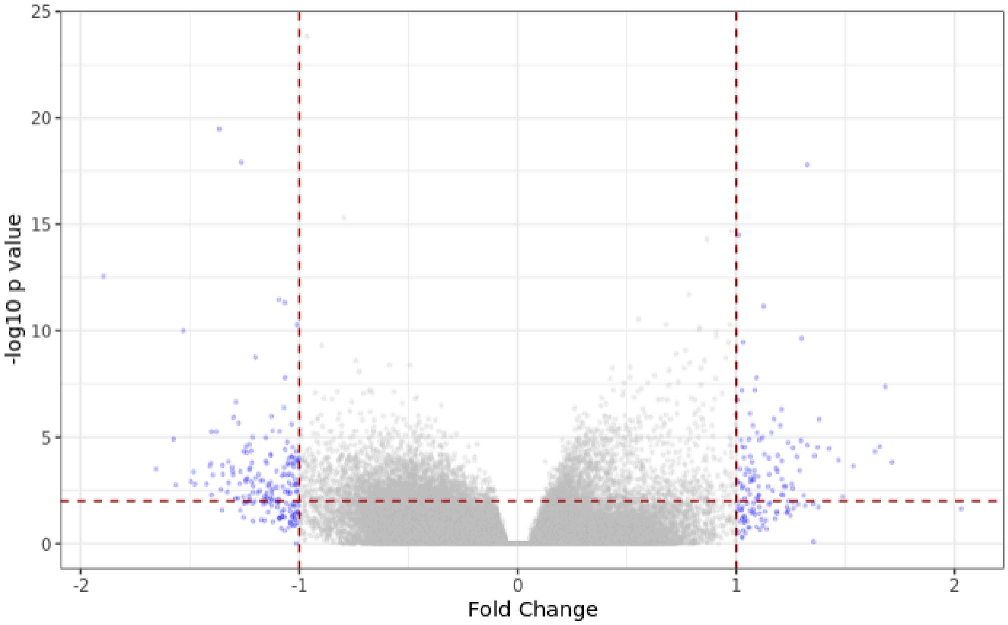
Based on the graph, it appears that patients with metastatic tumors have higher survival probabilities across all time stamps to a significant degree. However, this result does not appear to make sense with what we know about metastatic tumors. One reason why our data may be skewed is that there is a smaller sample size of primary solid tumor patients and those that do follow up are likely to have had more complications after treatment. Those that have no further complications are less likely to follow up, excluding them from our data and making it more biased and misleading.

2. Expression differences between metastatic and non-metastatic patients



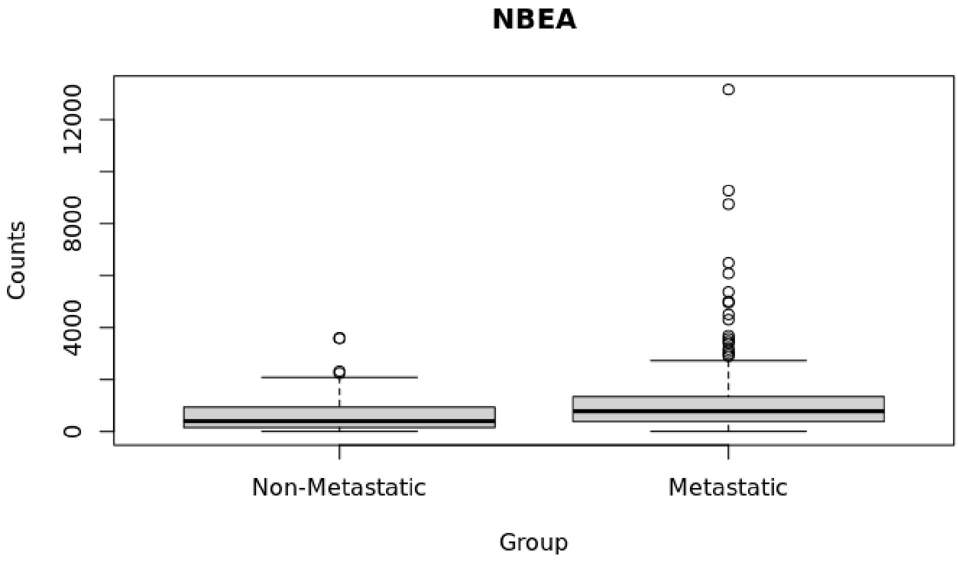
With non-metastatic patients (primary solid tumor) as the baseline, we can see that there are genes in patients with metastatic tumors that are both upregulated and down regulated. However, none of these results seem to be significant based on p-value. There are some genes that are neither upregulated or downregulated with significant results. However, the majority of the genes are neither upregulated or downregulated with insignificant results.

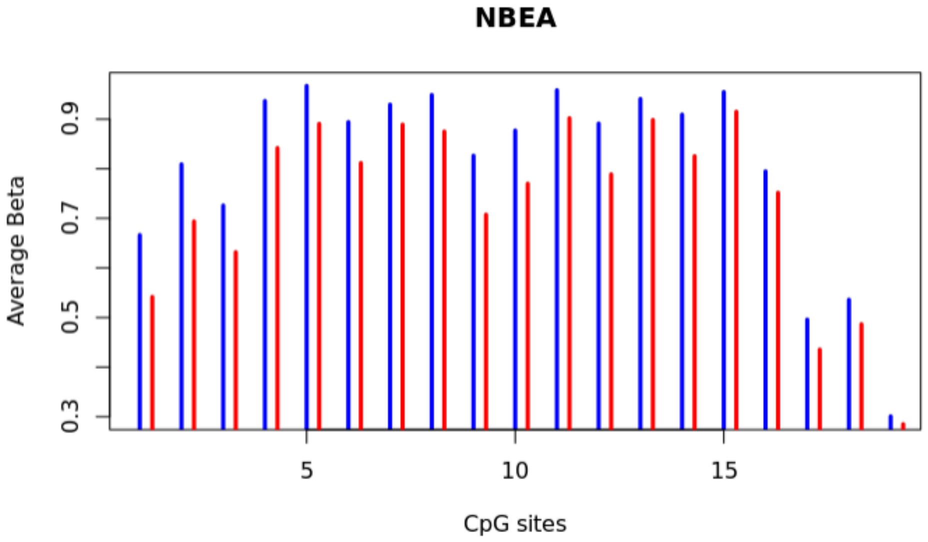
3. Methylation differences between metastatic and non-metastatic patients

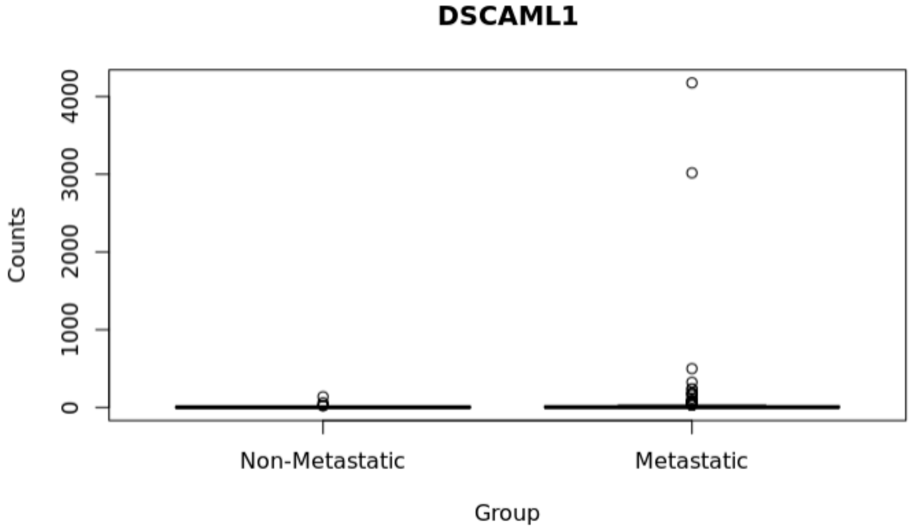


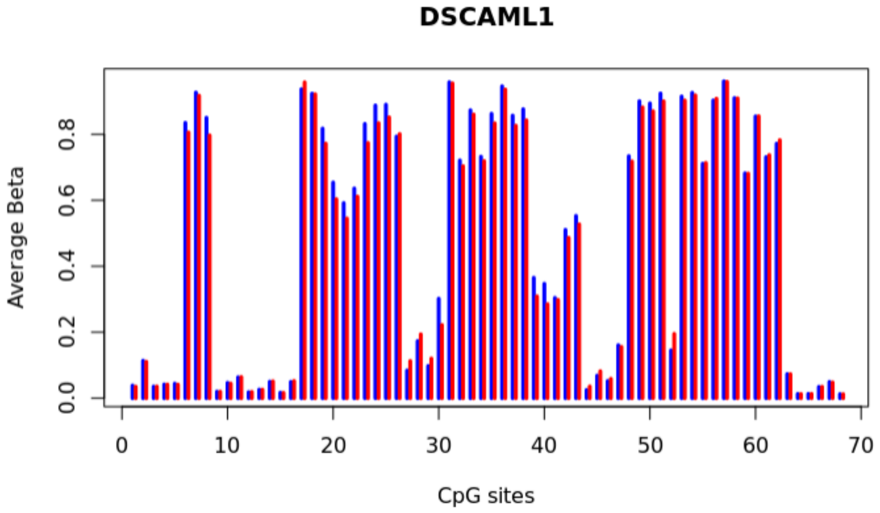
On the right, are cpg sites that are hypermethylated in metastatic patients. Left are undermethylated in regards to non-metastatic samples

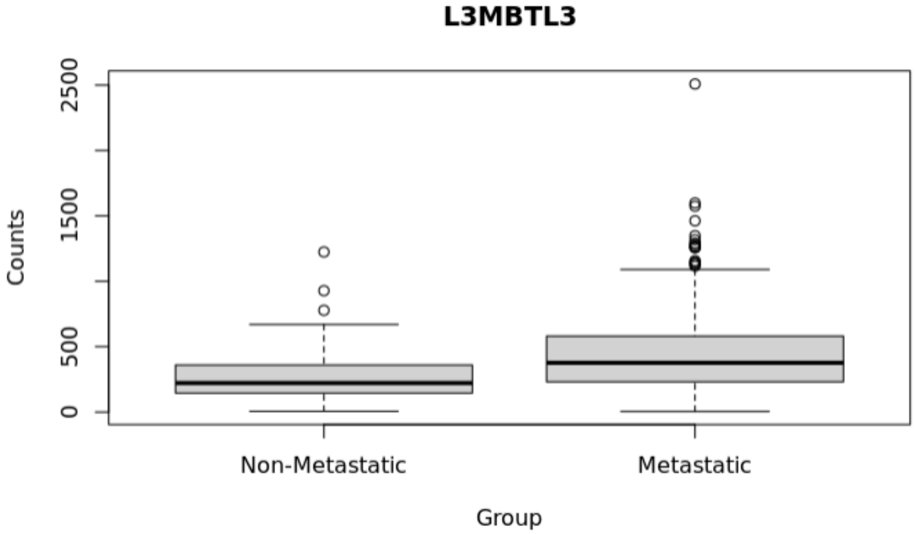
4. Direct comparison of transcriptional activity to methylation status for 10 genes
    a. In general, from the boxplots, the rna counts (expression) do not significantly differ between metastatic and non-metastatic patients. However, in the metastatic patients plot, there does seem to be a greater number of outliers with abnormally high rna counts levels.
       In general, we can see that the average beta value at cpg sites is usually either at similar levels between metastatic and non-metastatic patients. However, the values seem to be slightly higher on average for non-metastatic patients. This is unexpected since this means that non-metastatic patients may tend to have a greater percentage of methylated cells, which may lead to a greater rate of mutations for that gene.

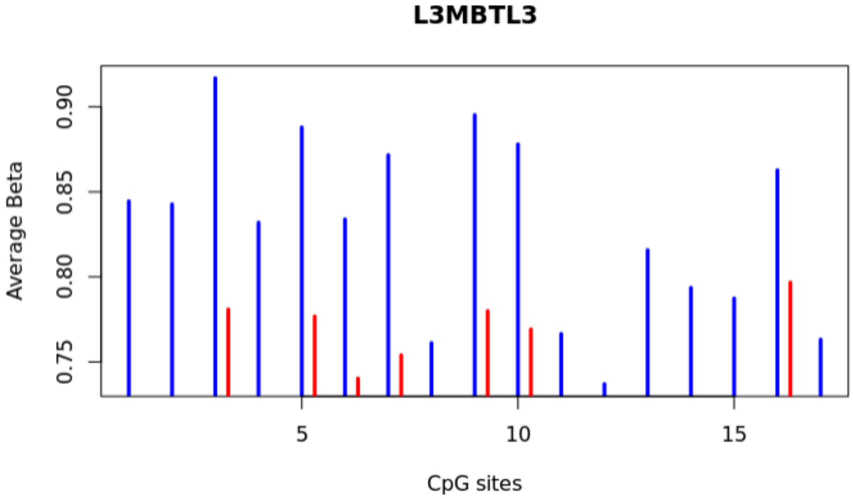


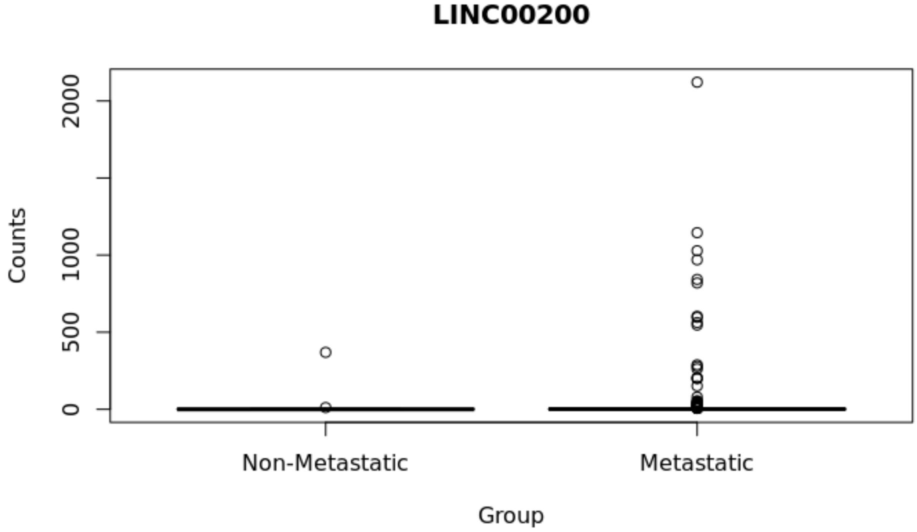

NBEA

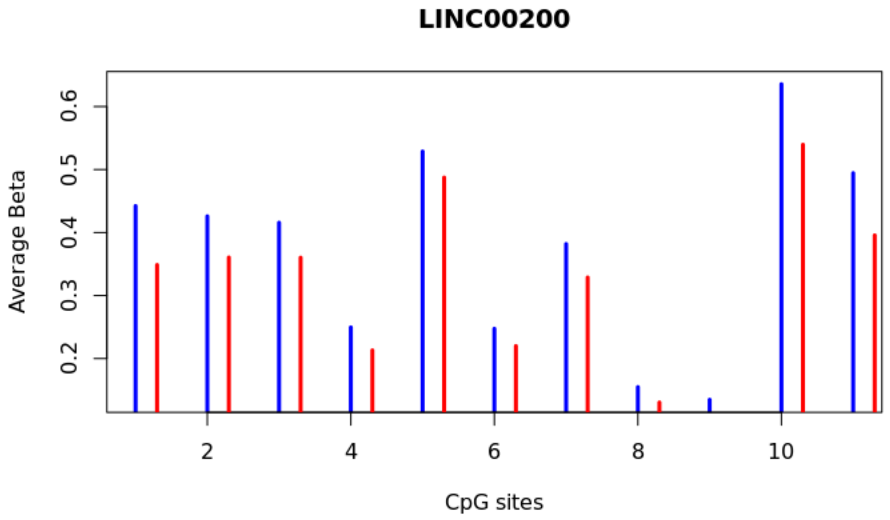

DSCAML1

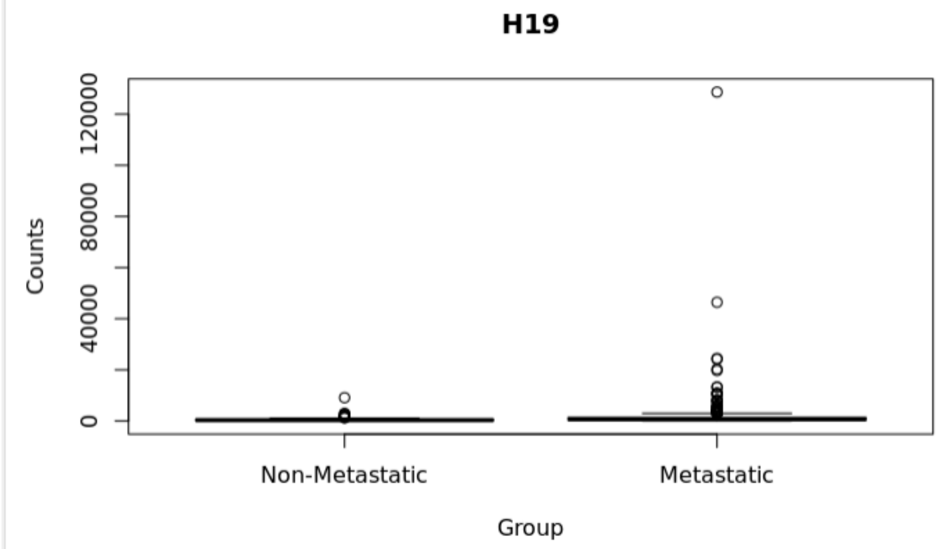

DSCAML1

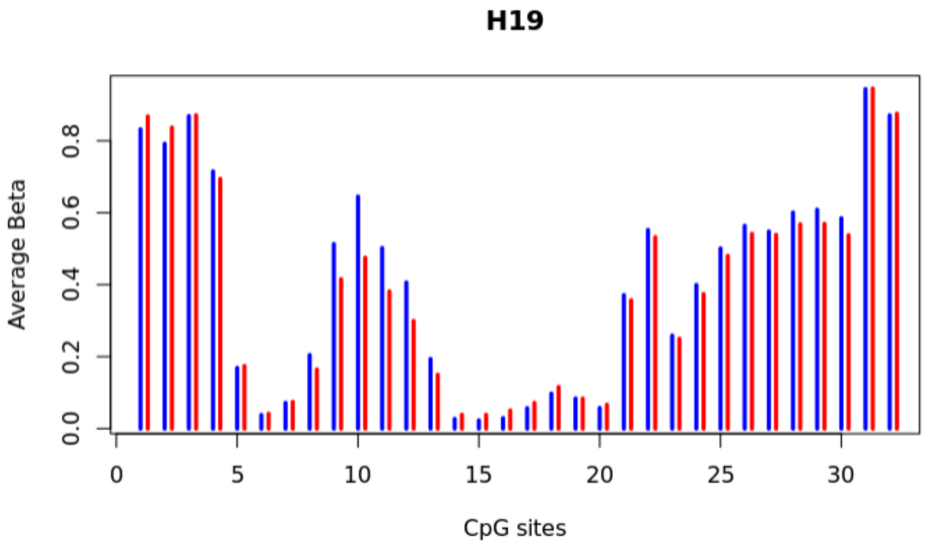

L3MBTL3
Counts
0
500
1500
2500
Non-Metastatic
Metastatic
Group

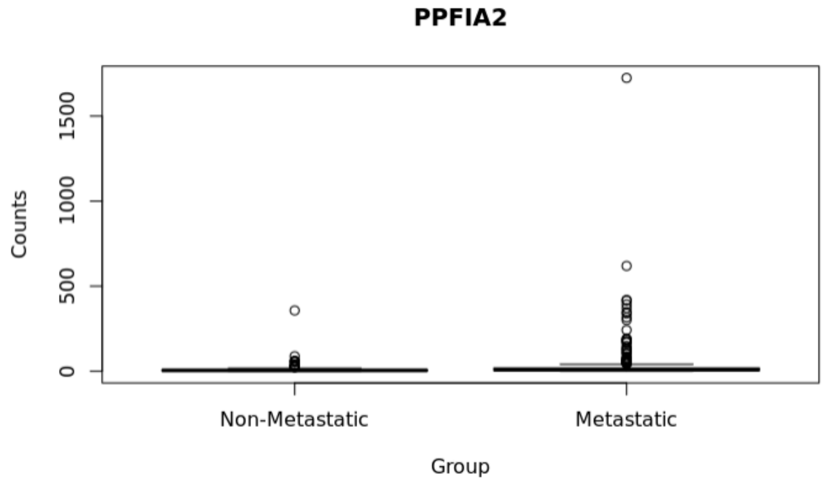

L3MBTL3
Average Beta
0.75
0.80
0.85
0.90
5
10
15
CpG sites

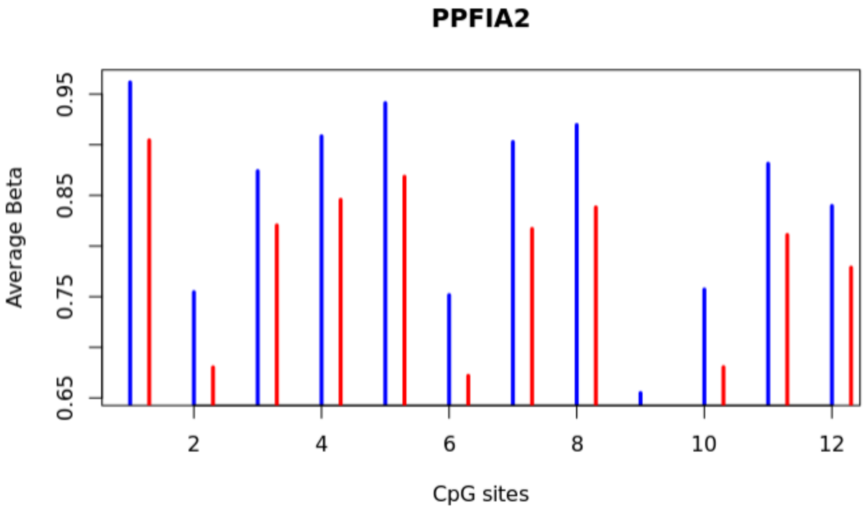

LINC00200
Counts
0
500
1000
2000
Non-Metastatic
Metastatic
Group

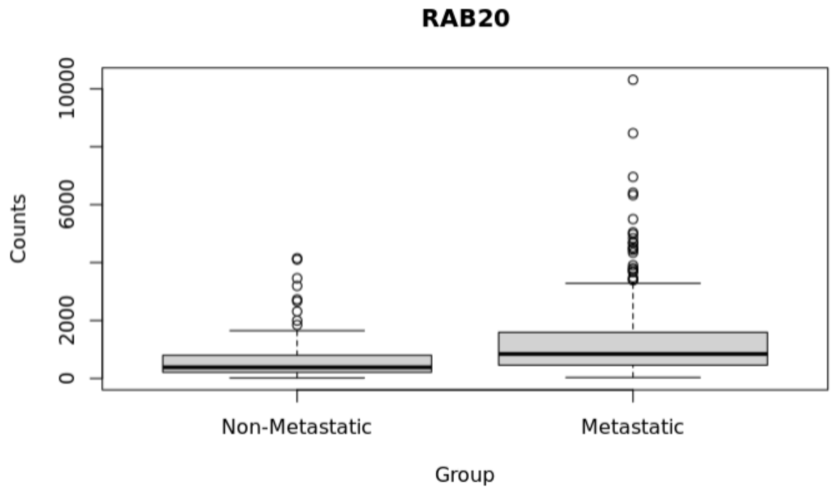

LINC00200
Average Beta
0.2
0.3
0.4
0.5
0.6
2
4
6
8
10
CpG sites

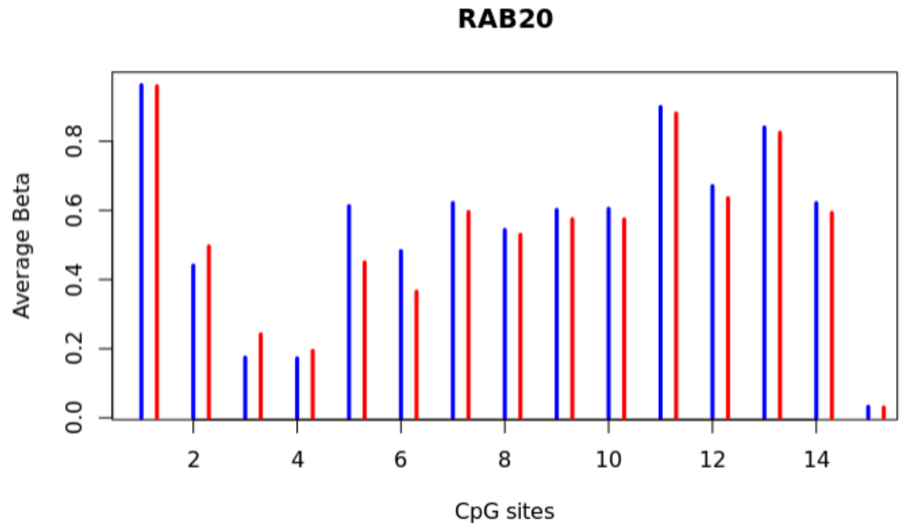

H19
Counts
0
40000
80000
120000
Non-Metastatic
Metastatic
Group

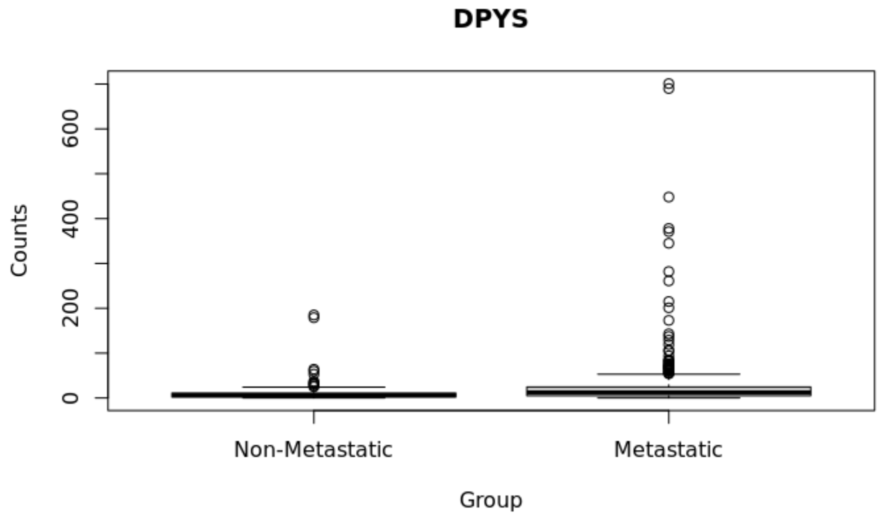

H19
Average Beta
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
30
CpG sites

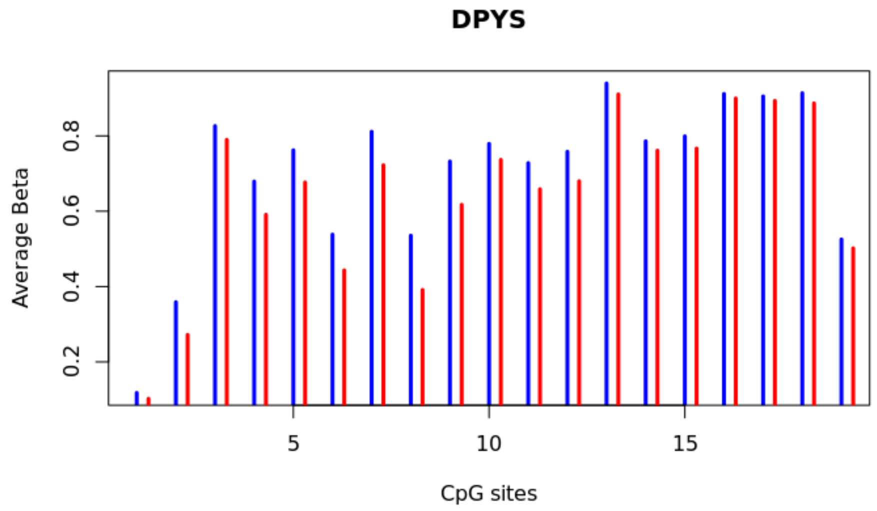

PPFIA2
Counts
0
500
1000
1500
Non-Metastatic
Metastatic
Group

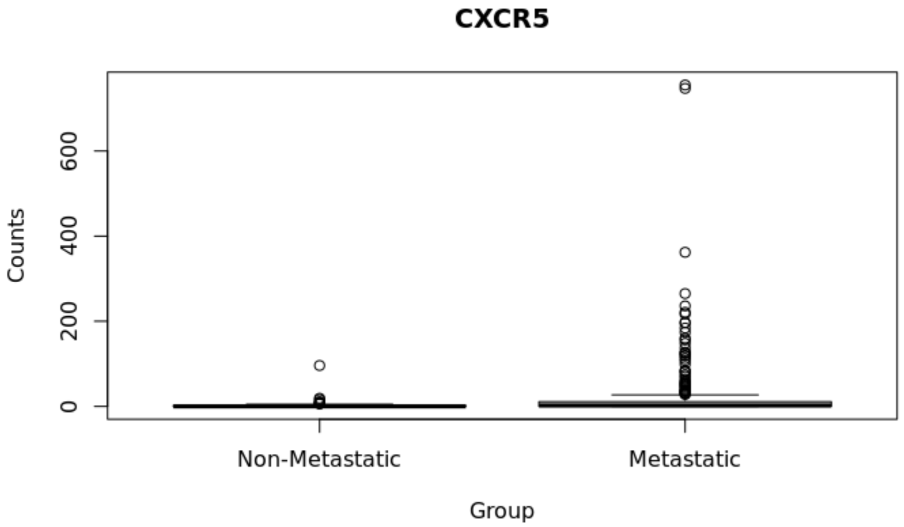

PPFIA2
Average Beta
0.65
0.75
0.85
0.95
2
4
6
8
10
12
CpG sites

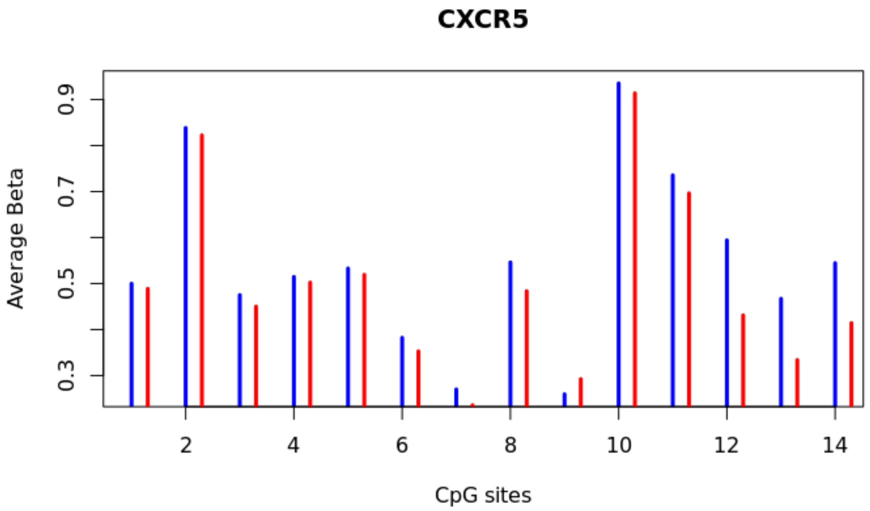

RAB20
Counts
0
2000
6000
10000
Non-Metastatic
Metastatic
Group

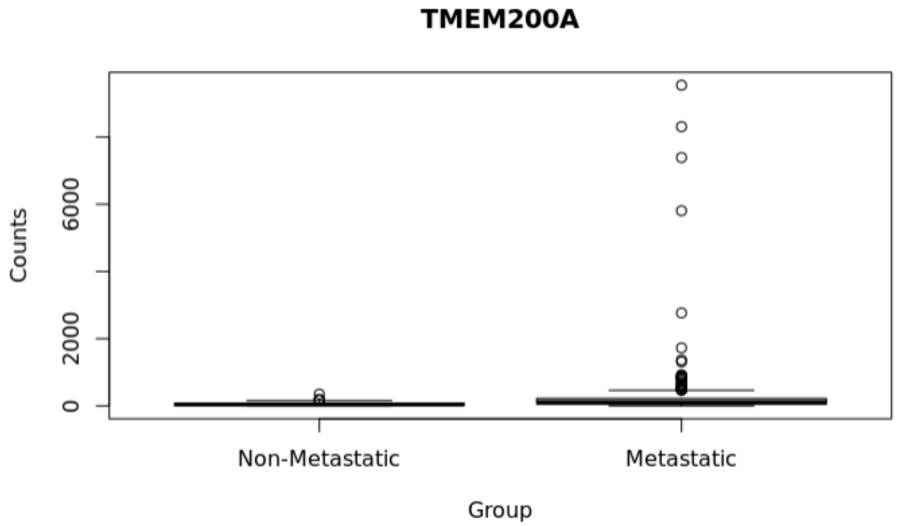

RAB20
Average Beta
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10
12
14
CpG sites

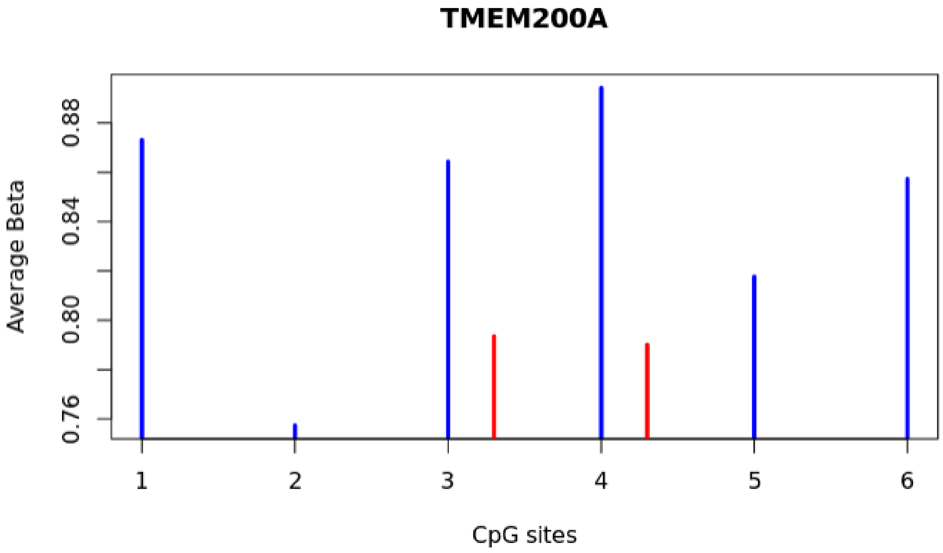

DPYS
Counts
0
200
400
600
Non-Metastatic
Metastatic
Group


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

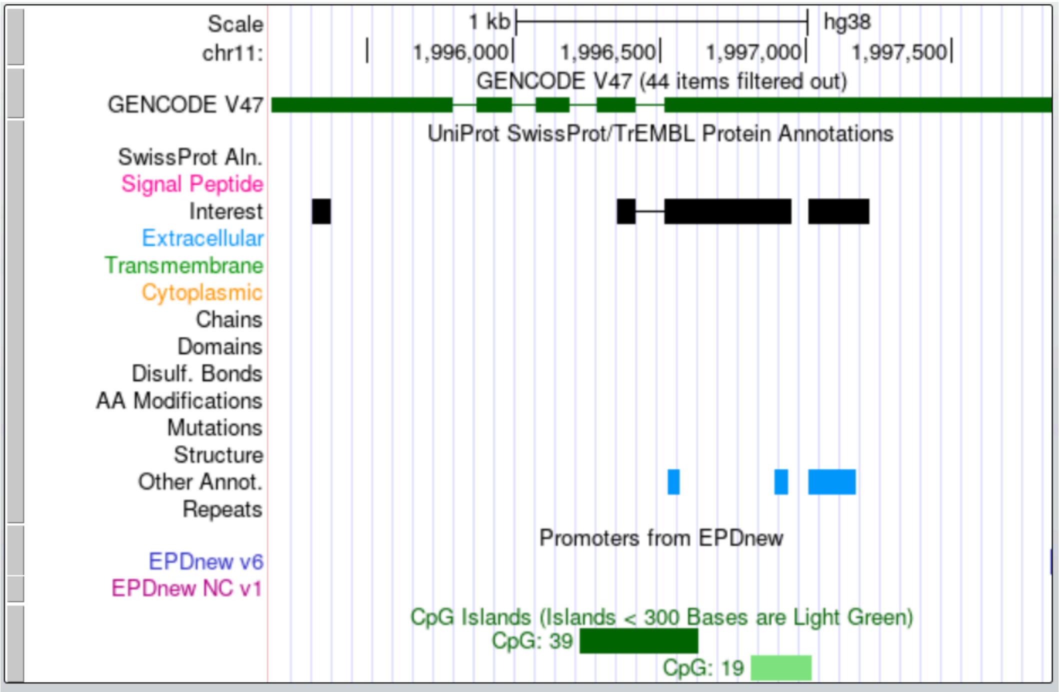

CXCR5


CXCR5


TMEM200A

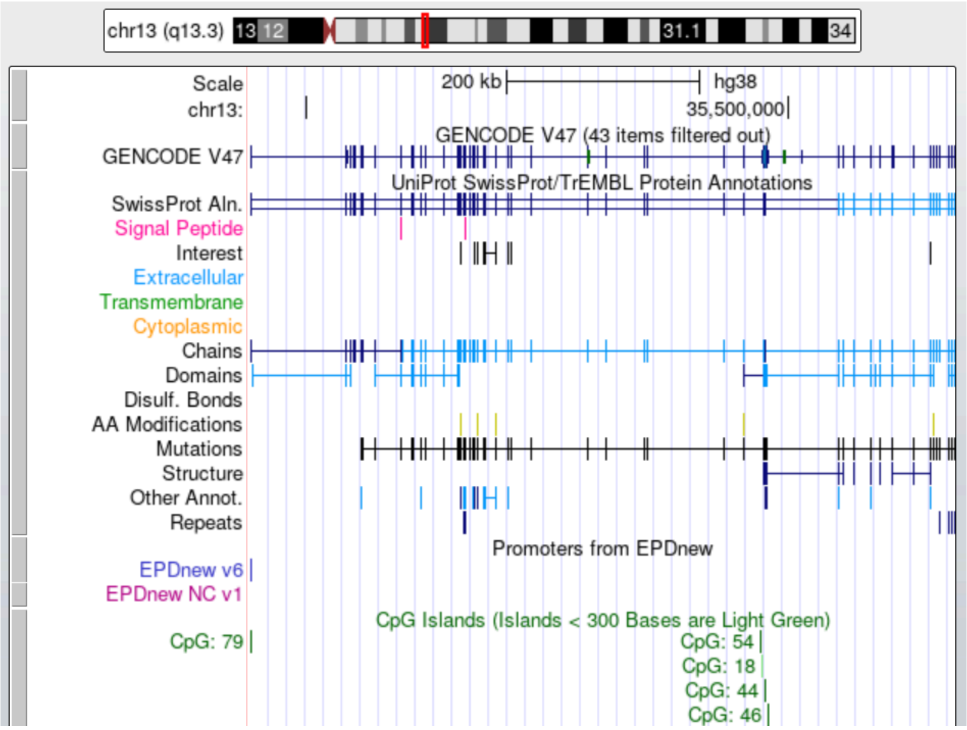


5. Visualization of CpG sites and protein domains for 3 genes (use UCSC genome browser) for a few genes. Describe at least one academic article (research or review) that either supports or doesn't support your final conclusion for one of the genes. If previously published work doesn't support your analysis, explain why this might be the case.
   a. H19:



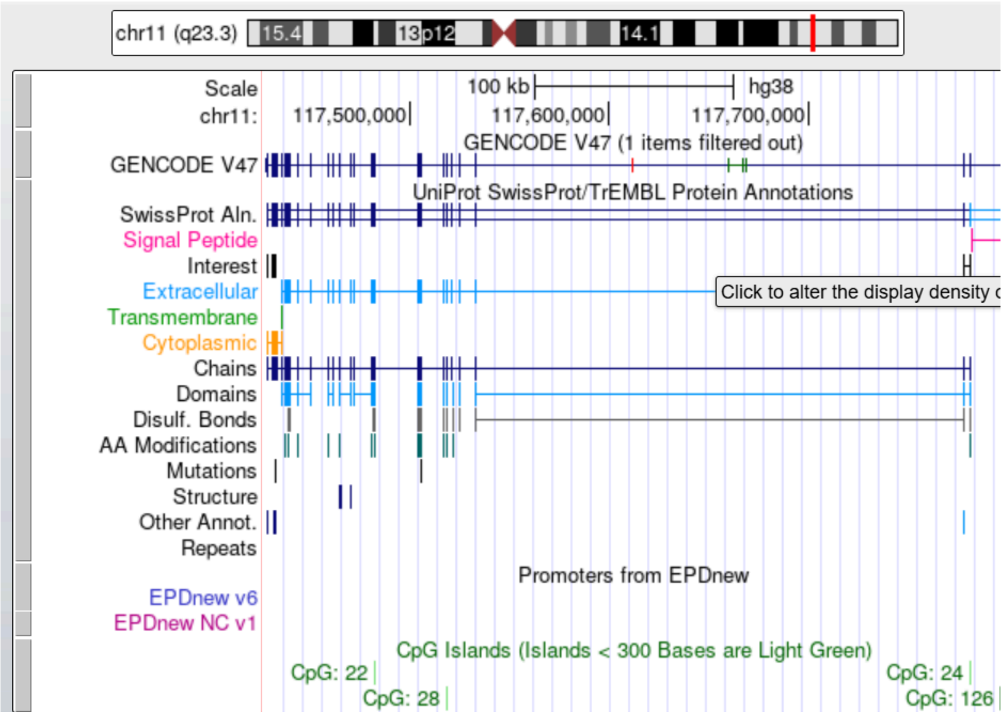
      i. There are two distinct CpG islands, with one having many more bases. However, both do not appear to be in close proximity to the promoter region, which is at the very end. The CpG sites seem to overlap with the GENCODE V47, interest, and other sites.

b. NBEA:



i. This gene seems to have many different CpG sites, 5 to be exact. The first CpG site seems to coincide with the promoter region while the last 4 all seem to be around the same position.

c. DSCAML1:



        i. This gene also has many CpG sites with a total of 4. Interestingly, no promoter region is shown. Also, the CpG site at the very end of the gene also seems to have the most sites.

    d. Article: “Downregulation of lncRNA H19 inhibits the migration and invasion of melanoma cells by inactivating the NF-κB and PI3K/Akt signaling pathways”

        i. This article finds that higher expression levels of H19 were found in tumor tissue samples compared to healthy samples. They also found that the H19 gene tends to be upregulated in tumor tissue.

        ii. These findings match what we found as in our volcano plots H19 was upregulated in metastatic tumor patients vs non-metastatic patients. On the other hand, there does not seem to be a significant difference in the RNA expression counts between metastatic and non-metastatic patients in our graphs, which contradicts the findings. However, the metastatic side did have more outliers with high RNA counts. This discrepancy could be due to a sampling size error.

References:

Liao, Zhichao et al. “Downregulation of lncRNA H19 inhibits the migration and invasion of melanoma cells by inactivating the NF-κB and PI3K/Akt signaling pathways.” *Molecular medicine reports* vol. 17,5 (2018): 7313-7318. doi:10.3892/mmr.2018.8782