Aditi Shankar
R-Review Project
QBIO 490

Part 1: Review Questions

1. What is TCGA and why is it important?
    a. TCGA is the Cancer Genome Atlas and it is important because it contains comprehensive, publicly accessible genomic data with over 20,000 samples across 33 cancer types. The dataset is a valuable resource for identifying genetic mutations and pathways for research and treatment methods.
2. What are some strengths and weaknesses of TCGA?
    a. Some strengths of TCGA is that it contains a large amount of data across different omics: genomic, transcriptomic, and epigenomic data. The weakness of TCGA is that it includes limited representation of rare cancers and may contain potential demographic biases which in turn does not encapsulate the genetic diversity of a whole population.
3. What commands are used to save a file to your github repository?
    a. Git status, git add <filename>, git commit -m “done”, git push
4. What commands must be used to run a package in R?
    a. install.packages("package") → library(package)
5. What commands must be used to run the Bioconductor package in R?
    a. BiocManager::install() → library(BiocManager)
6. What is boolean indexing? What are some applications for?
    a. Boolean indexing is a very effective R technique for data cleaning, subsetting or selection, and conditional analysis. Boolean indexing involves applying a vector of booleans (true or false) to a column or row in a dataframe based on a particular condition.
7. 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
    a. students <- data.frame(Name = c("Alice", "Bob", "Charlie", "Diana", "Evan"),

        Score = c(55, 92, 76, 61, 45),

        Age = c(20, 21, 22, 19, 20))

        students

        # this line makes a dataframe with three columns, Name, Score, and Age and creates a vector of values for each column.

students$pass_or_fail <- ifelse(students$Score <= 60, “Fail”, “Pass”)

this if else statement, makes a new columns called pass_or_fail and puts Fail when the score of the student is less than or equal to 60

Passed_student_index <- students[students$pass_or_fail == “Pass”,]

the line filters the students dataframe and creates a new dataframe called passed_student_index, with all the students who passed. It uses the pass_or_fail column to determine whether to be included in the new dataset or not

Part 2: SKCM Analysis

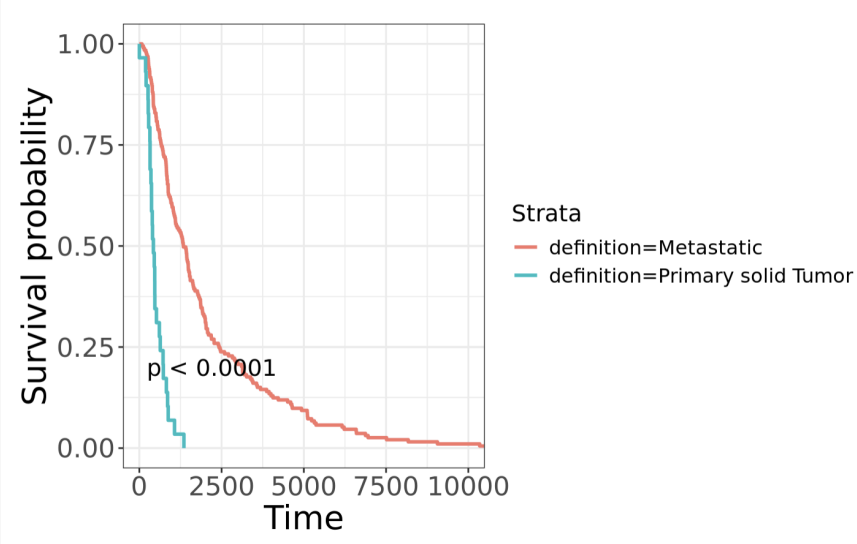
- Full code is pushed to GitHub and this is the repo link: https://github.com/aditiShankar0402/qbio_490_aditishankar.git

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?

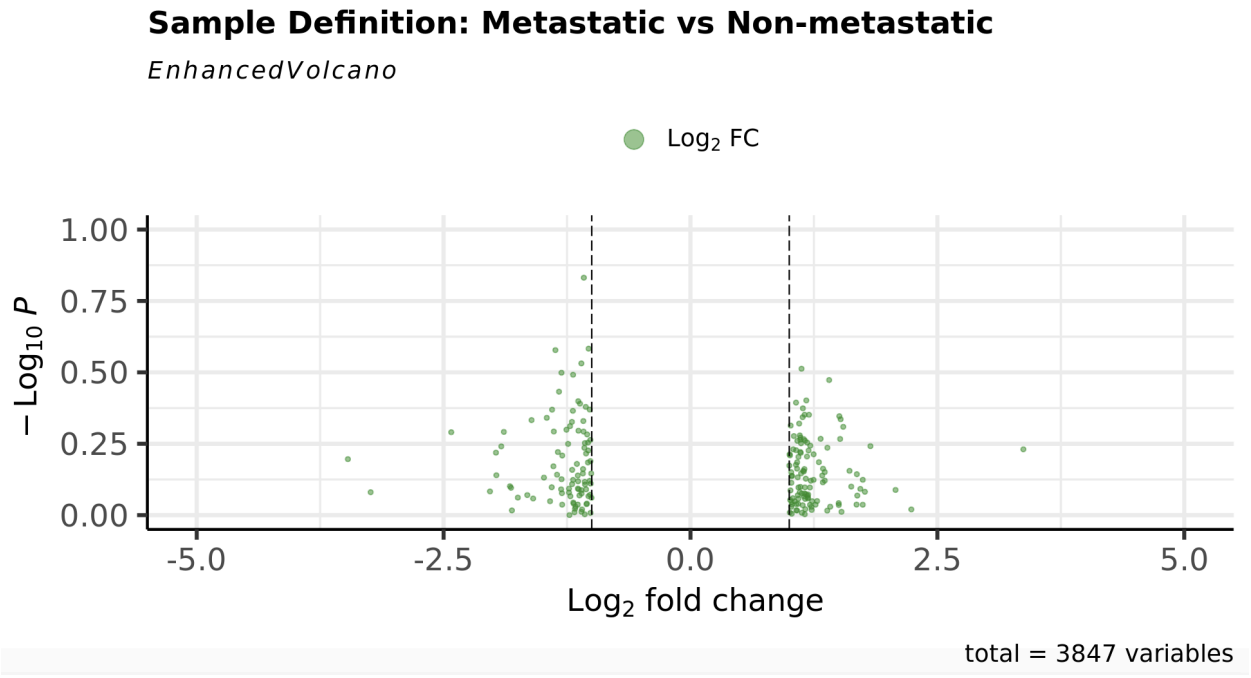
1. Difference in survival between metastatic and non-metastatic patients



In this KM plot the y axis represents the survival probability and the x-axis represents the time in days. When analyzing which group has the greatest survival rate, patients with primary solid tumors have better survival rates overall compared to metastatic patients, but, primary solid

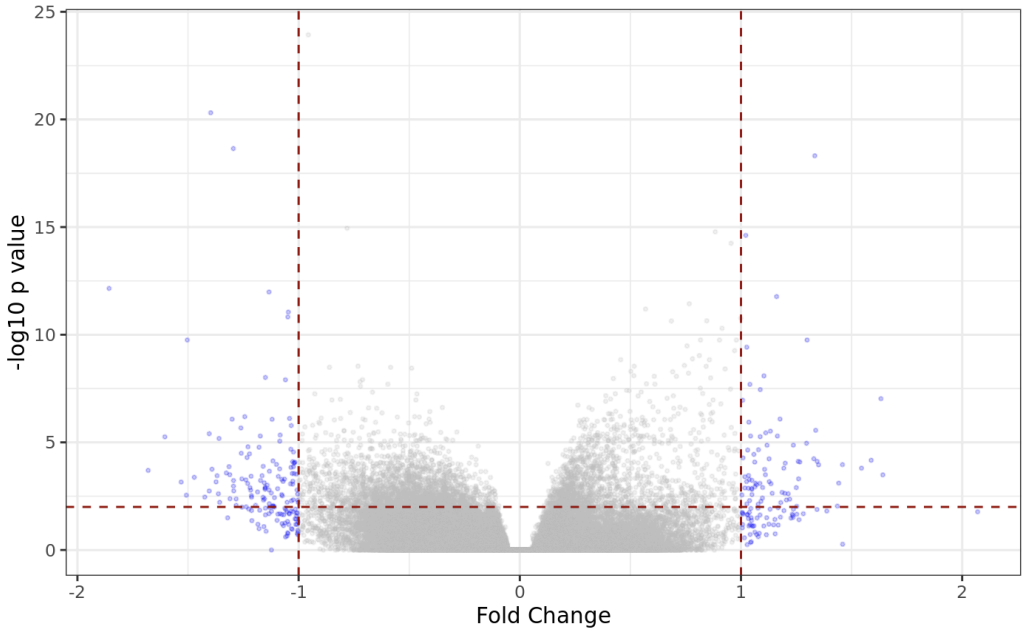
tumor patients die faster initially, while metastatic patients have a more gradual decline in survival over time. The p value is 0.0001. Since the value is less than 0.05, this is statistically significant and means that there is enough evidence to conclude that there is a significant difference in survival rates among the metastatic patients and primary solid tumor patients.

2. Expression differences between metastatic and non-metastatic patients



This volcano plot shows the expression difference between metastatic and non-metastatic patients. The genes in the top right are genes that are upregulated in metastatic patients, while the top left genes are genes that are downregulated in metastatic patients. The regions near 0 have low P values and are neither significantly upregulated or downregulated. In this plot there are no points in the middle region, meaning the data is very significant.

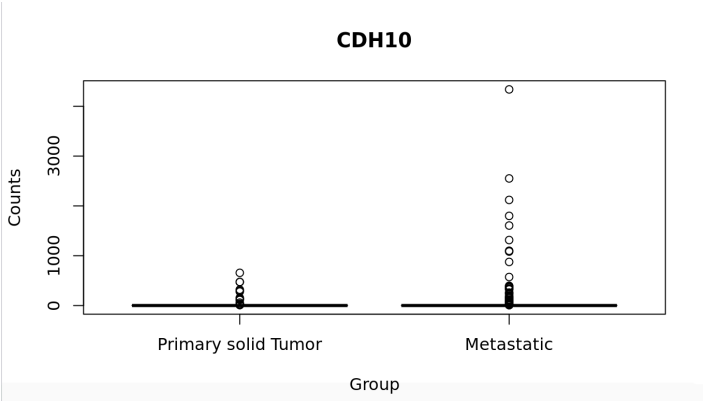
3. Methylation differences between metastatic and non-metastatic patients

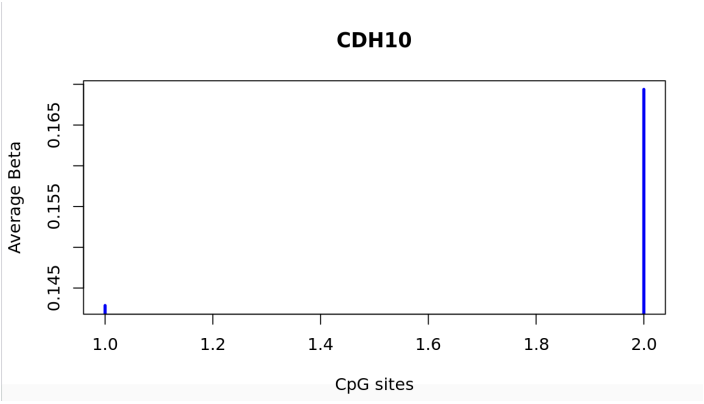


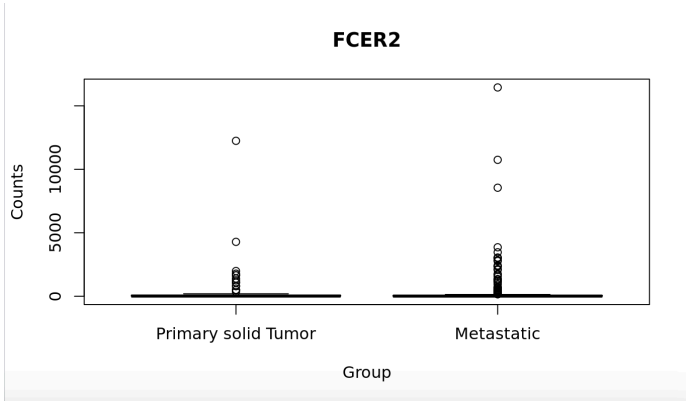
This plot shows the methylation differences between metastatic and non-metastatic patients. All the points on the top right represent cpg sites that are hyper-methylated in metastatic patients. The ones on the left represent cpg sites that are under-methylated in metastatic patients relative to patients with primary tumors or normal tissue. Points near the center or below the significance threshold line are not statistically or biologically significant. There are many points toward the center of the plot.

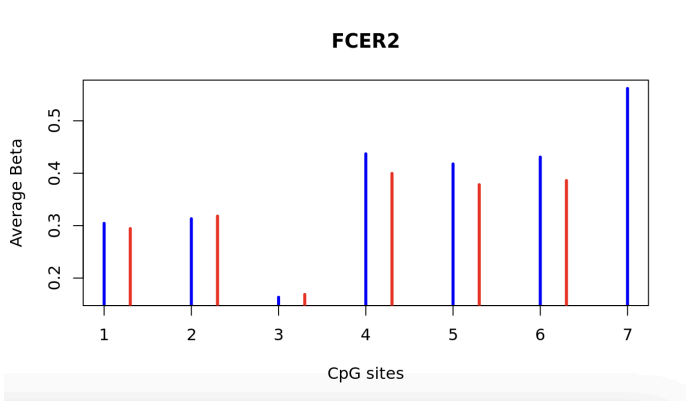
4. Direct comparison of transcriptional activity to methylation status for 10 genes

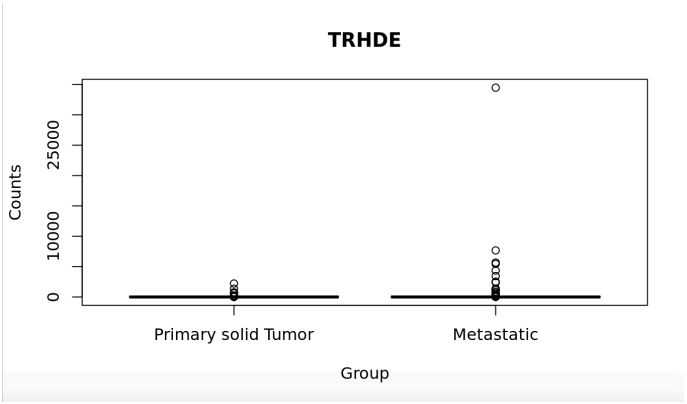
This comparison was done with any genes that were hypermethylated and upregulated.

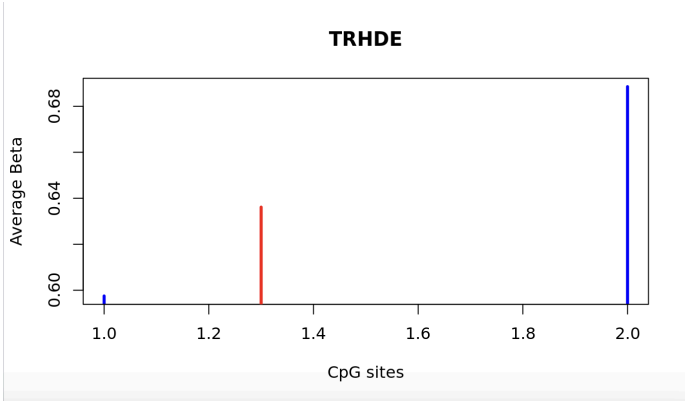


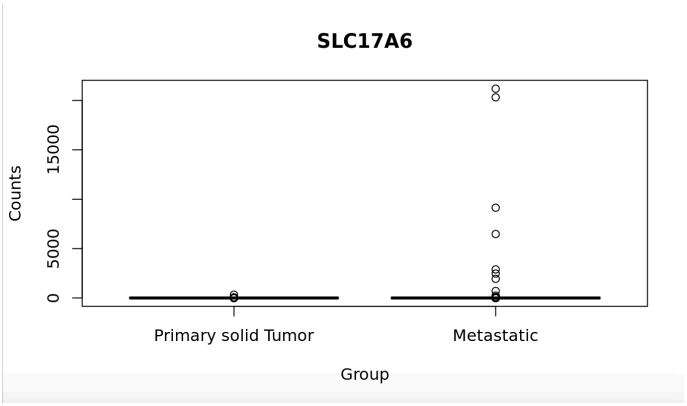


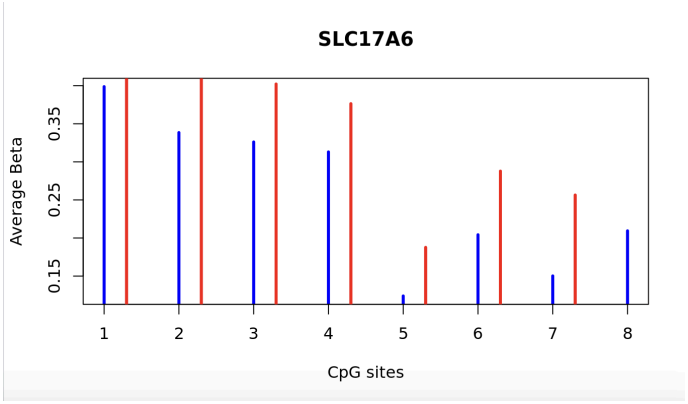

FCER2
Counts
0
5000
10000
Primary solid Tumor
Metastatic
Group

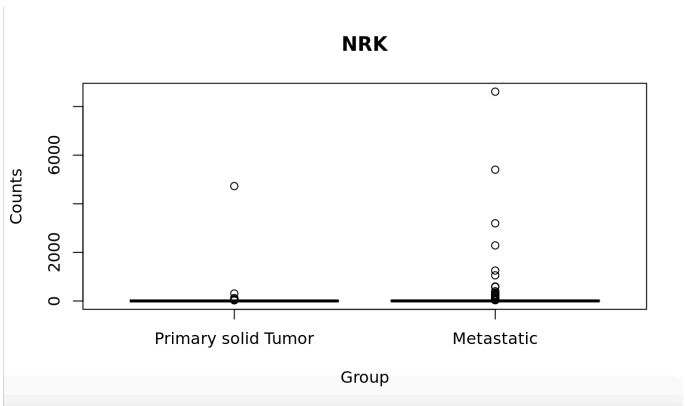

FCER2
Average Beta
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
CpG sites

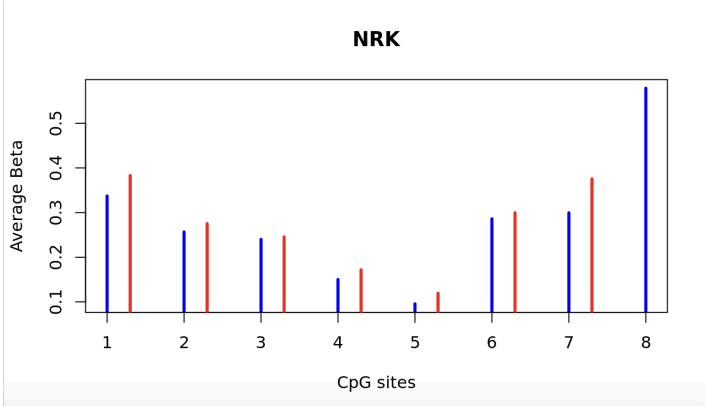

TRHDE
Counts
0
10000
25000
Primary solid Tumor
Metastatic
Group

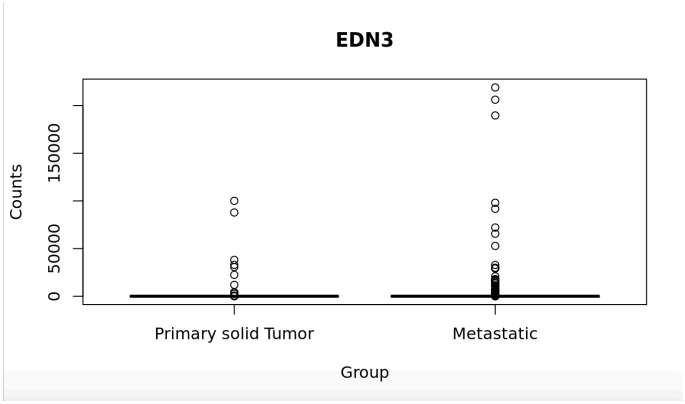

TRHDE
Average Beta
0.60
0.64
0.68
1.0
1.2
1.4
1.6
1.8
2.0
CpG sites

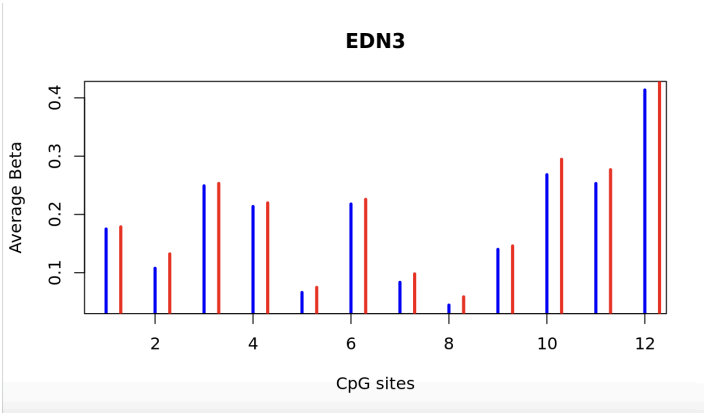

SLC17A6
Counts
0
5000
15000
Primary solid Tumor
Metastatic
Group

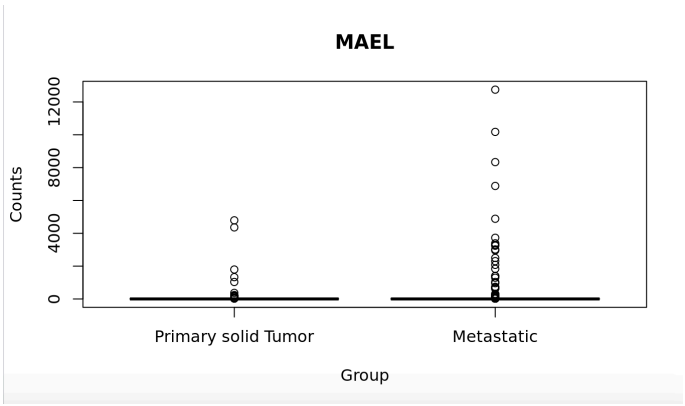

SLC17A6
Average Beta
0.15
0.25
0.35
1
2
3
4
5
6
7
8
CpG sites

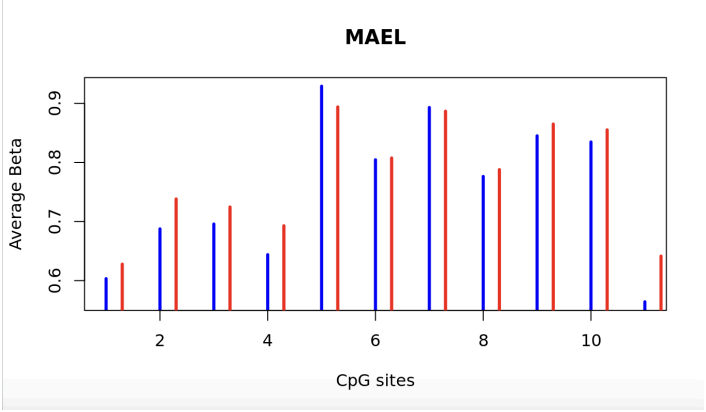

NRK
Counts
0
2000
6000
Primary solid Tumor
Metastatic
Group

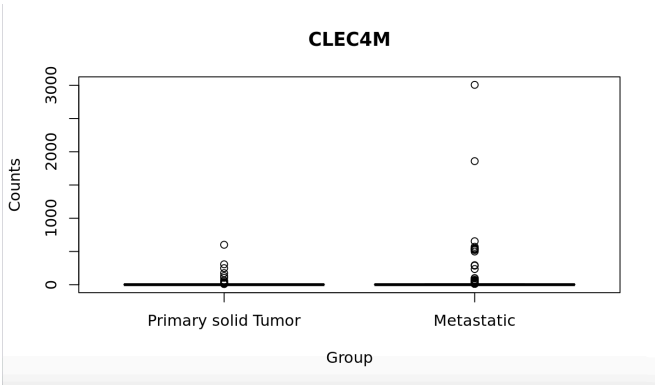

NRK
Average Beta
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
CpG sites

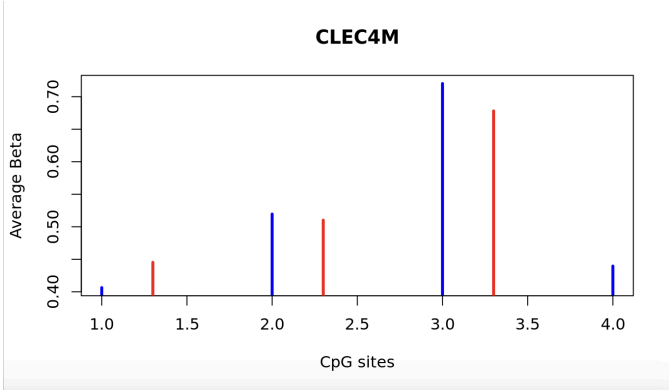

EDN3
Counts
0
50000
150000
Primary solid Tumor
Metastatic
Group

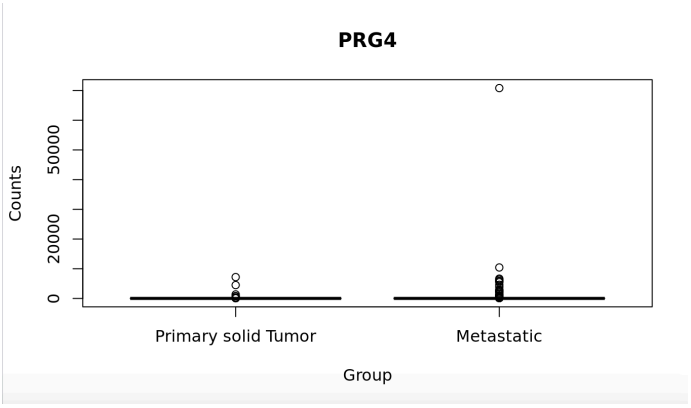

EDN3
Average Beta
0.1
0.2
0.3
0.4
2
4
6
8
10
12
CpG sites

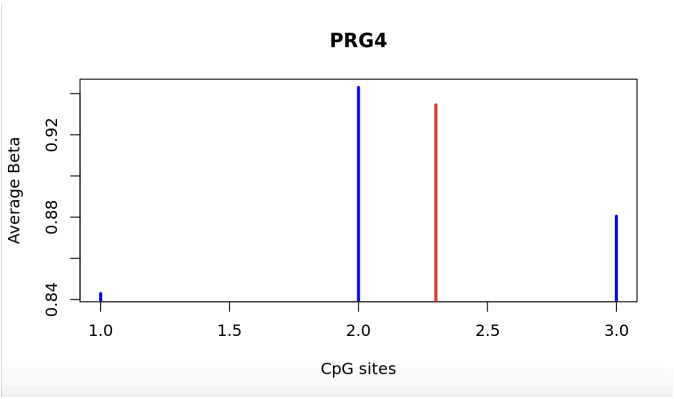

MAEL
Counts
0
4000
8000
12000
Primary solid Tumor
Metastatic
Group

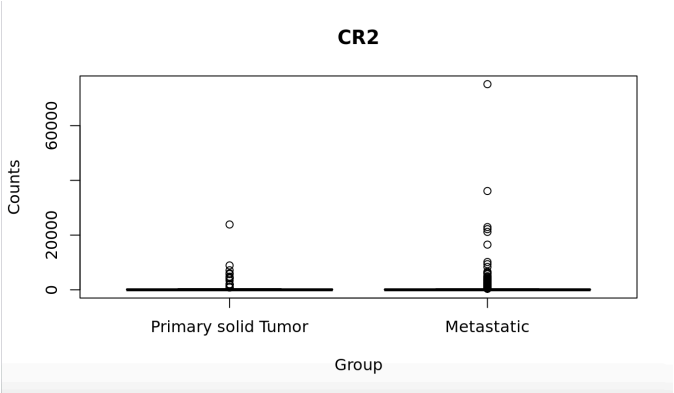

MAEL
Average Beta
0.6
0.7
0.8
0.9
2
4
6
8
10
CpG sites

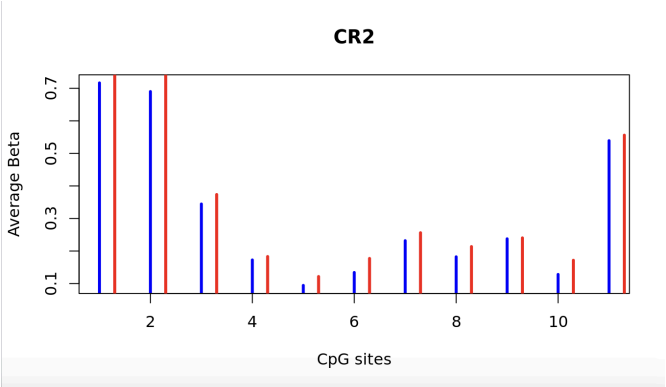

CLEC4M
Counts
0
1000
2000
3000
Primary solid Tumor
Metastatic
Group


CLEC4M
Average Beta
0.40
0.50
0.60
0.70
1.0
1.5
2.0
2.5
3.0
3.5
4.0
CpG sites


PRG4
Counts
0
20000
50000
Primary solid Tumor
Metastatic
Group


PRG4
Average Beta
0.84
0.88
0.92
1.0
1.5
2.0
2.5
3.0
CpG sites

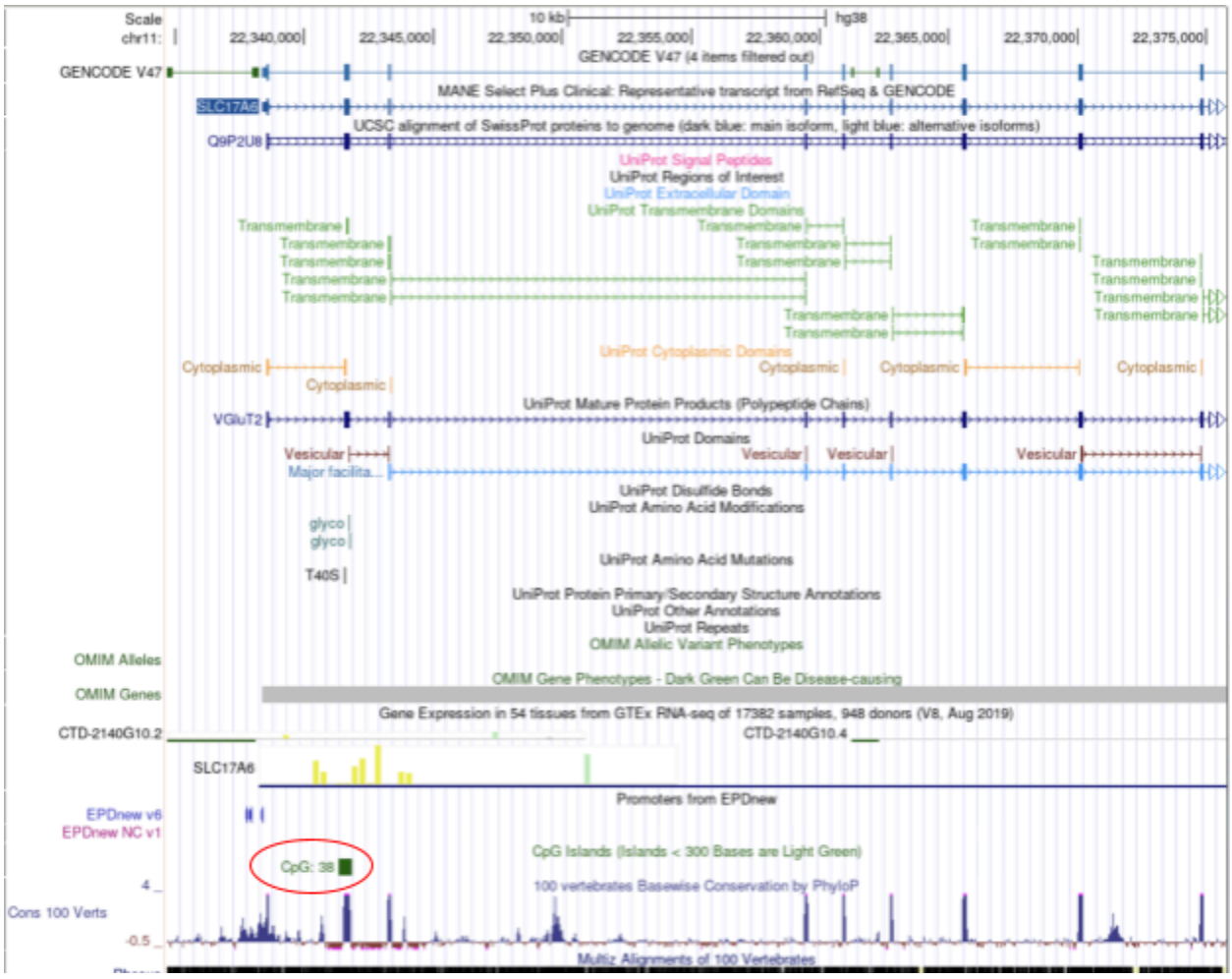




The ten intersection genes were CDH10, TRHDE, SLC17A6, FCER2, CLEC4M, PRG4, CR2, NRK, EDN3, and MAEL. For all of genes' box plots representing RNA transcription expression between Metastatic and primary solid tumor. In general, it can be seen that the Metastatic count is higher than the primary solid tumor for all genes. While the median expression levels of all genes are not different between both groups, the outliers in the metastatic group can mean that the gene is upregulated in those patients. The horizontal line plot indicates the methylation differences of the genes in certain CpG sites. Blue indicates primary solid tumor, while red represents metastatic. For some methylation sites, for example in SLC17A6, there is a difference in that the metastatic tumor is more methylated than the primary solid tumor. In other cases, the levels are relatively similar, indicating no significant methylation differences between the two groups. Additionally, there are also instances where one group (primary solid tumor or metastatic) appears to completely dominate the methylation.

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.

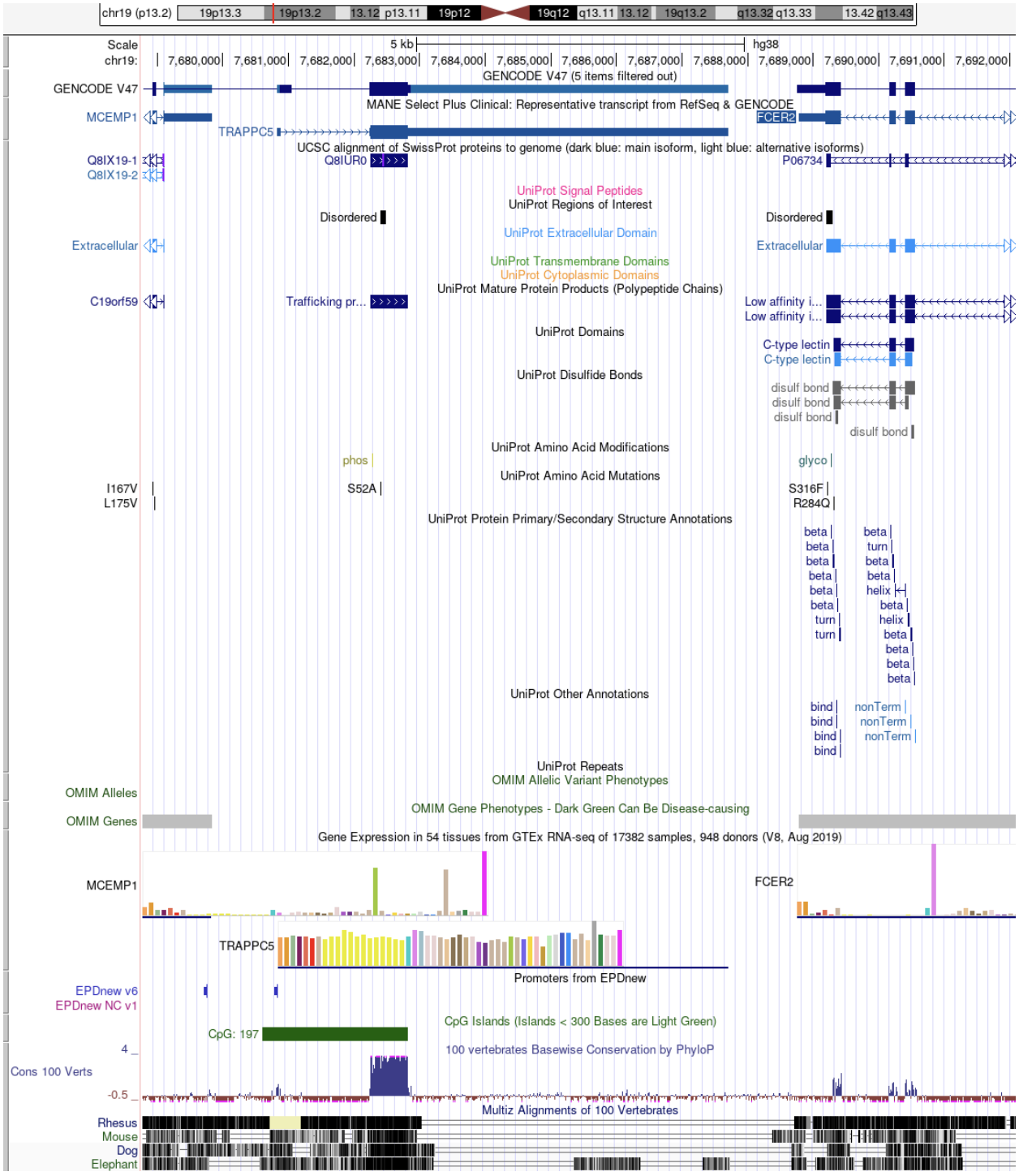
SLC17A6



MAEL


Scale
chr1:
10 kb
hg38
166,990,000
166,995,000
167,000,000
167,005,000
167,010,000
167,015,000
GENCODE V47
GENCODE V47 (4 items filtered out)
MANE Select Plus Clinical: Representative transcript from RefSeq & GENCODE
MAEL
UCSC alignment of SwissProt proteins to genome (dark blue: main isoform, light blue: alternative isoforms)
Q96JY0-1
Q96JY0-2
Q96JY0-3
UniProt Signal Peptides
UniProt Regions of Interest
Disordered
UniProt Extracellular Domain
UniProt Transmembrane Domains
UniProt Cytoplasmic Domains
UniProt Mature Protein Products (Polypeptide Chains)
1
UniProt Domains
HMG box
HMG box
Maelstrom
UniProt Disulfide Bonds
UniProt Amino Acid Modifications
UniProt Amino Acid Mutations
S41A
UniProt Protein Primary/Secondary Structure Annotations
helix
helix
helix
helix
helix
helix
UniProt Other Annotations
DNA-binding
UniProt Repeats
OMIM Allelic Variant Phenotypes
OMIM Alleles
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
MAEL
RNA5SP65
Promoters from EPDnew
EPDnew v6
EPDnew NC v1
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 43
100 vertebrates Basewise Conservation by PhyloP
4
Cons 100 Verts
-0.5

FCER2



A study done in 2021, examined miRNA and mRNA expression profiling to find potential biomarkers for metastatic cutaneous melanoma. The data and clinical information, similar to our project, was from the TCGA database for skin cancer. The Deseq analysis indicated that C7, VAT1L, LIFR, CR1, and FCER2 (Fc fragment of IgE receptor II), were the top 5 deMRNAs in

patients with metastatic cutaneous melanoma in comparison to non-metastatic patients. More clustering analysis by the study showed that these five DEmRNAs could help to identify between patients with metastatic or non-metastatic CM. This relates to our results as we observed that FCER2 was upregulated in metastatic patients.

References

Human hg38 Chr19:7,678,776-7,692,131 UCSC genome browser V473. (n.d.). https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr19%3A7678776-7692131&hgsid=2385272617_7DmK5w9EeNYA1PbFau08nUEvBxk9

Vizkeleti, L., Kiss, T., Koroknai, V., Ecsedi, S., Papp, O., Szasz, I., Adany, R., & Balazs, M. (2017). Altered integrin expression patterns shown by microarray in human cutaneous melanoma. *Melanoma Research*, *27*(3), 180–188. https://doi.org/10.1097/cmr.0000000000000322

Wang, J., Tao, Y., & Bian, Q. (2021). miRNA and mRNA expression profiling reveals potential biomarkers for metastatic cutaneous melanoma. *Expert Review of Anticancer Therapy*, *21*(5), 557–567. https://doi.org/10.1080/14737140.2021.1882860