Part 1: Review Questions

General Concepts

1) TCGA is the The Cancer Genome Atlas project. It is a public dataset that has information on different kinds of omics (mutation data for genomics, RNA count for transcriptomics, methylation data for epigenomics) for many different types of cancers. Thousands of patient samples are stored.

2) Strengths: stores data consisting of thousands of patient samples for many types of cancer, which allows us to explore many genes across a large patient sample; publicly accessible, protects patients, labeled and detailed
Weaknesses: some data is missing, not fully representative of patient population (more people represented than others), follow-up information is very limited

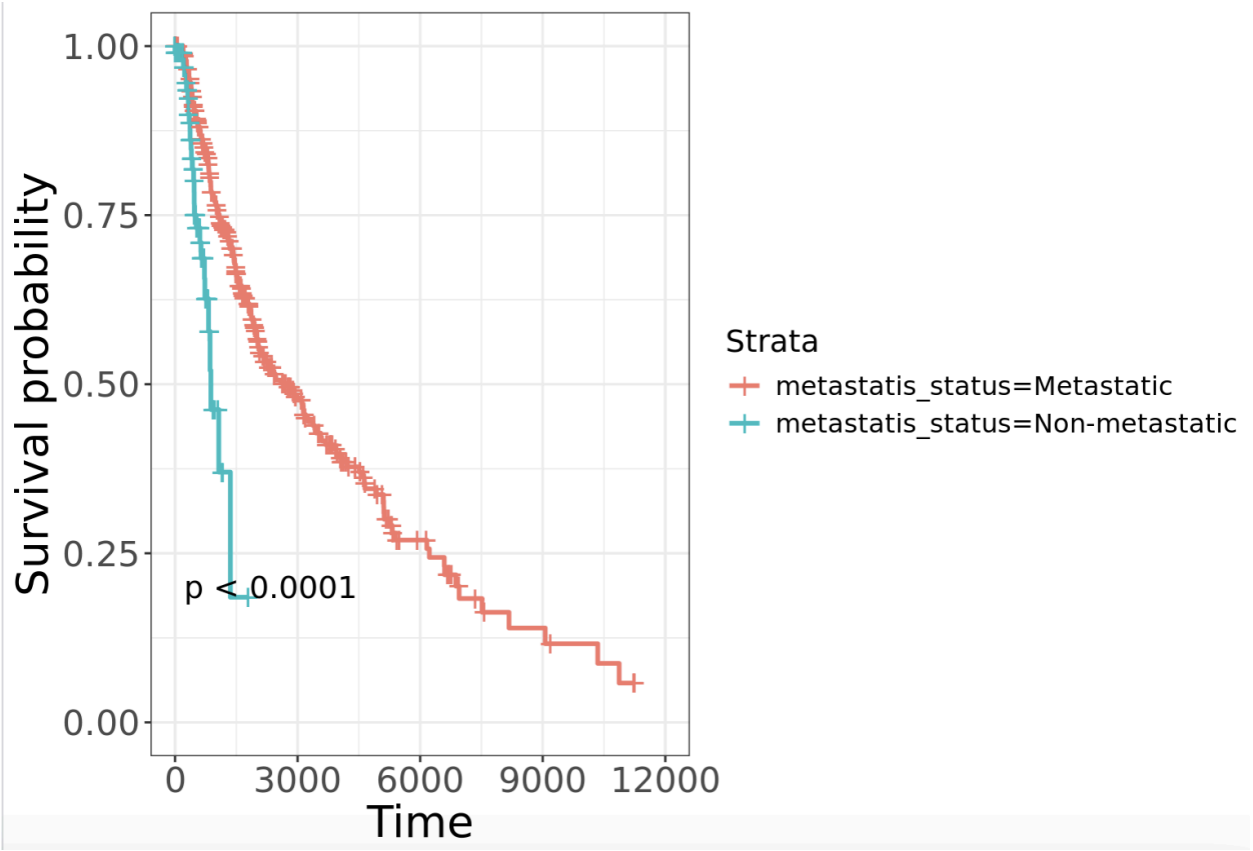
Coding Skills

1) git status, git add filename, git commit -m “message”, git push
2) install.packages() and library()
3) if (!require("BiocManager")) install.packages("BiocManager"), library(BiocManager), BiocManager::install("packagename"), library(packagename)
4) Boolean indexing is when you create a mask/vector of boolean values and apply this to a column or row of a dataframe to subset the data that fit a certain criteria you’re looking for. For example, if you want to filter the data for male patients over a certain age, you would apply a Boolean mask for gender and another for age. You can also use masking to delete data.
5)

Patient	Definition	Gender	Age at Diagnosis
TCGA-1	Primary solid tumor	male	26
TCGA-2	Solid Tissue Normal	female	49
TCGA-3	Solid Tissue Normal	female	68
TCGA-4	Primary solid tumor	female	35

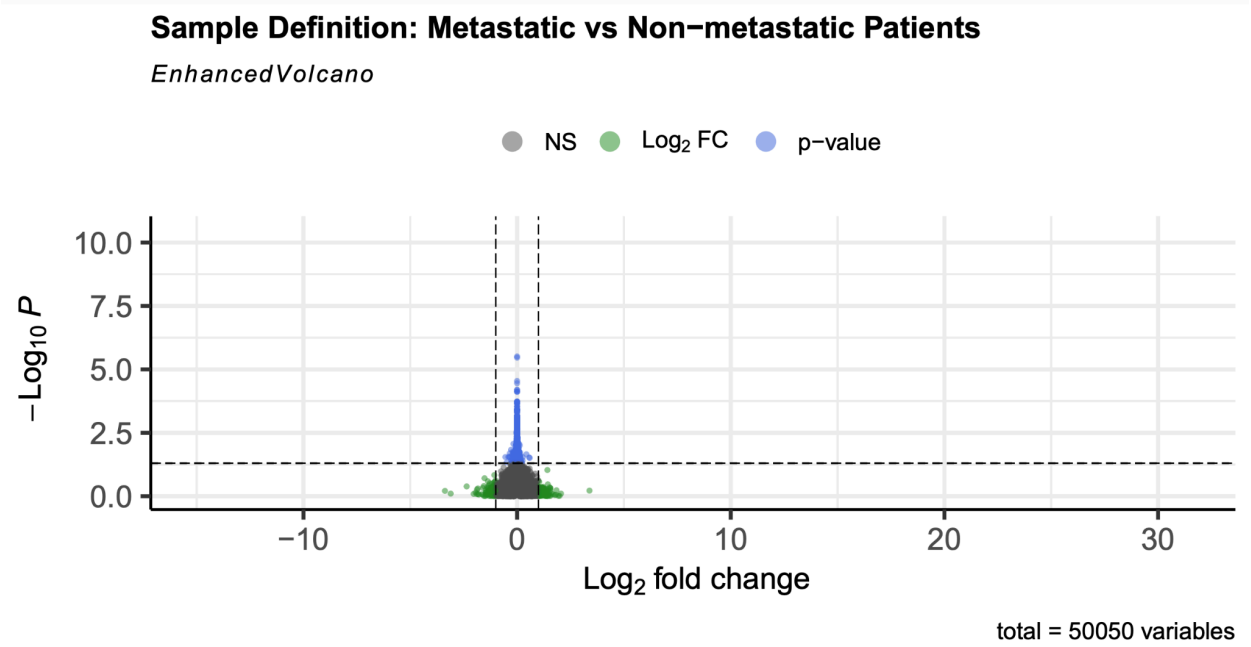
5a) female_mask <- ifelse(clinical$Gender == female, T, F) - filters the data for female patients
5b) female_clinical <- clinical[female_mask,] - applies the mask on the dataframe to extract the only patients who are female and all their associated data

Part 3: Results and Interpretations
1)



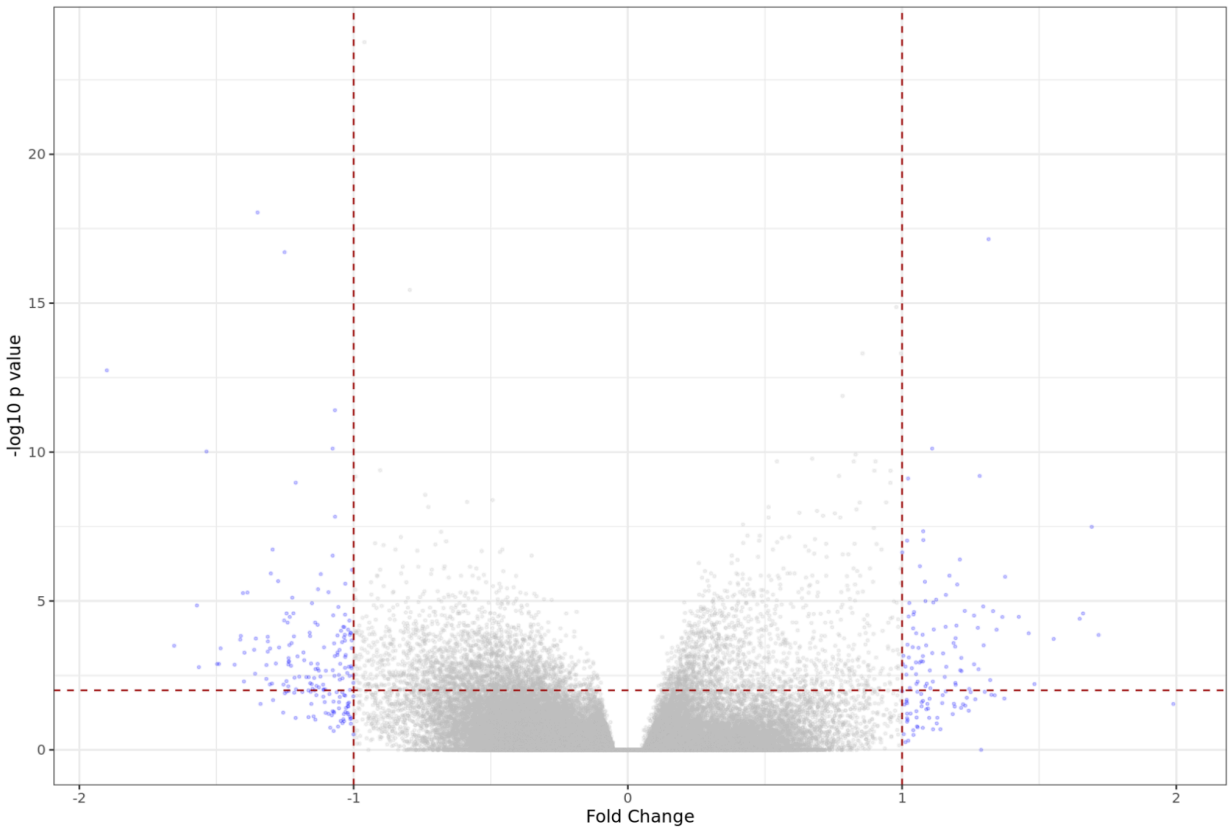
The KM plot was quite similar to the misleading example found in lecture; it appears that non-metastatic patients have worse survival rates compared to metastatic patients. However, this may not necessarily be the case/it goes against intuition. Non-metastatic patients may have gotten their cancer tissue excised or stopped following up, so the data may be skewed in that sense. Therefore, although the p value is significant, nothing significant can be concluded.

2)



From the volcano plot, we see that some genes (blue) are significant in metastatic vs. non-metastatic patients, but neither up nor downregulated. We also see other genes (green) that are upregulated and downregulated in metastatic patients compared to non-metastatic patients, but they are not statistically significant. Since there are no genes that are significantly upregulated/downregulated in metastatic vs. non-metastatic patients, it's difficult to make conclusions from the plot.

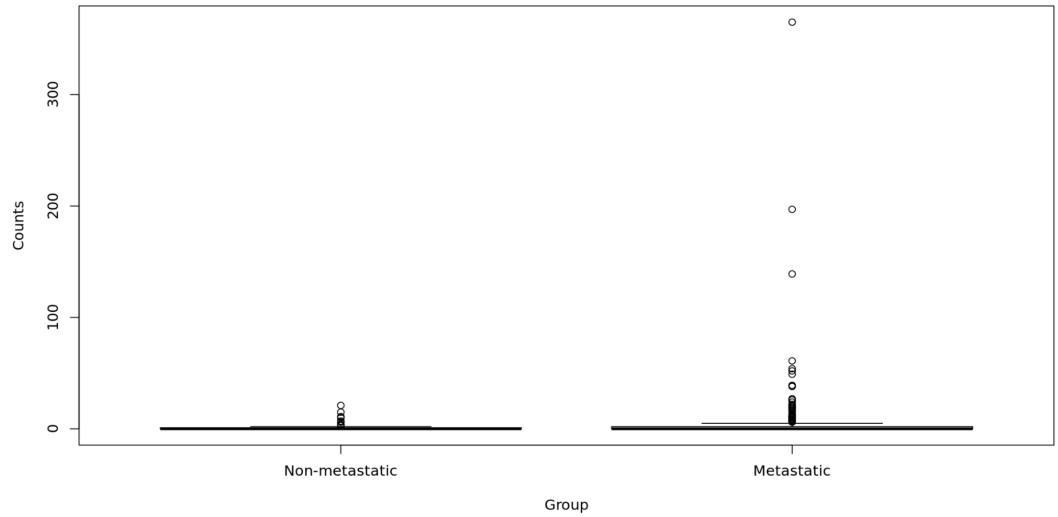
3)



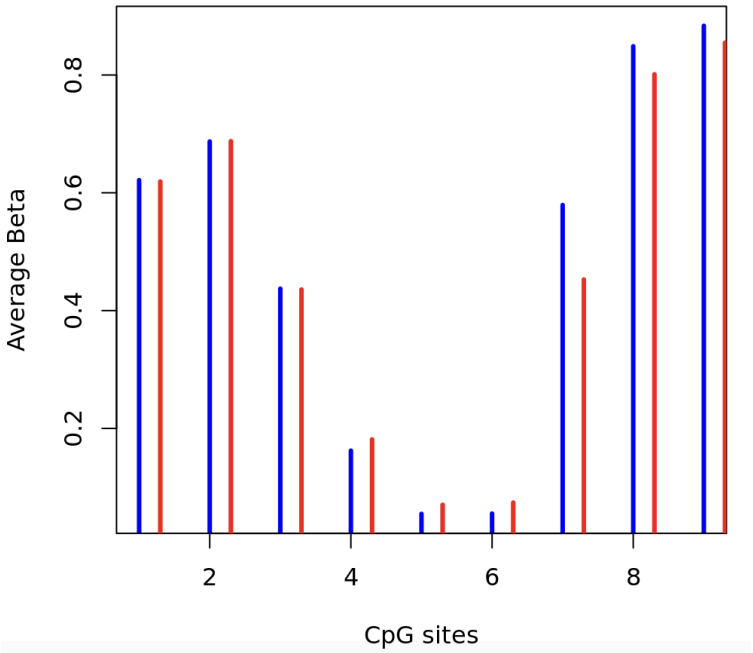
Blue points to the right of the rightmost vertical dashed line represent CPG sites that are hyper-methylated in metastatic samples compared to non-metastatic samples. Blue points to the left of the leftmost vertical dashed line represent CPG sites that are under-methylated in metastatic samples compared to non-metastatic samples. The sites that are significantly under-/hyper-methylated are more sparse compared to sites in the middle section, which suggests that only a select number of sites may be involved in metastasis.

4)
Upregulated + undermethylated:

CISTR:

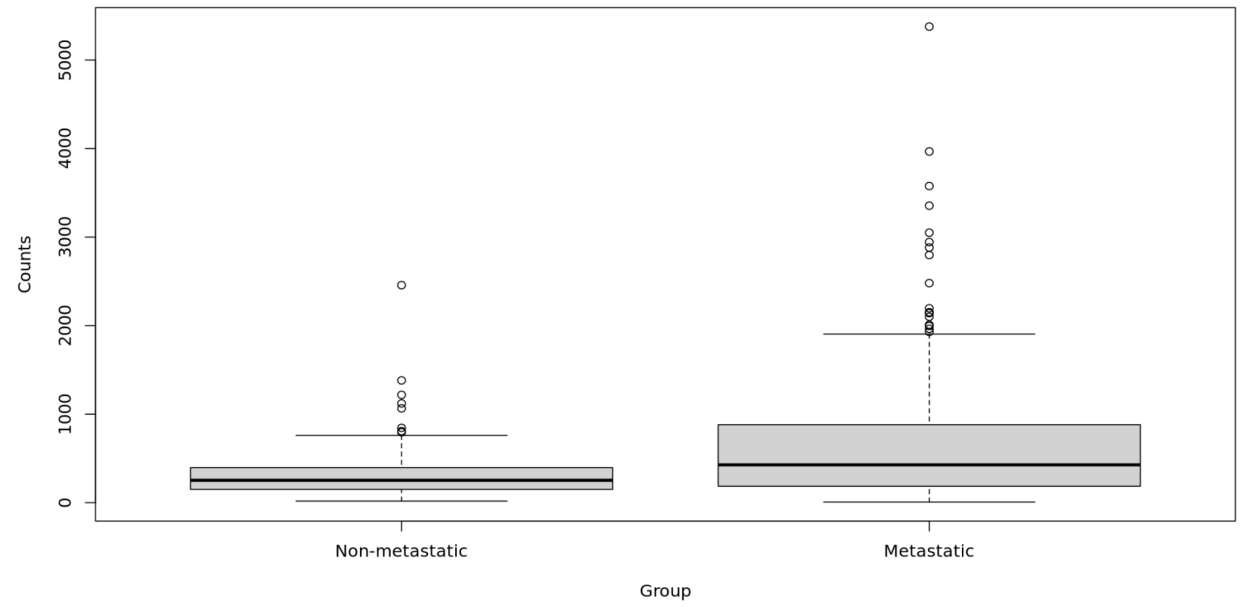


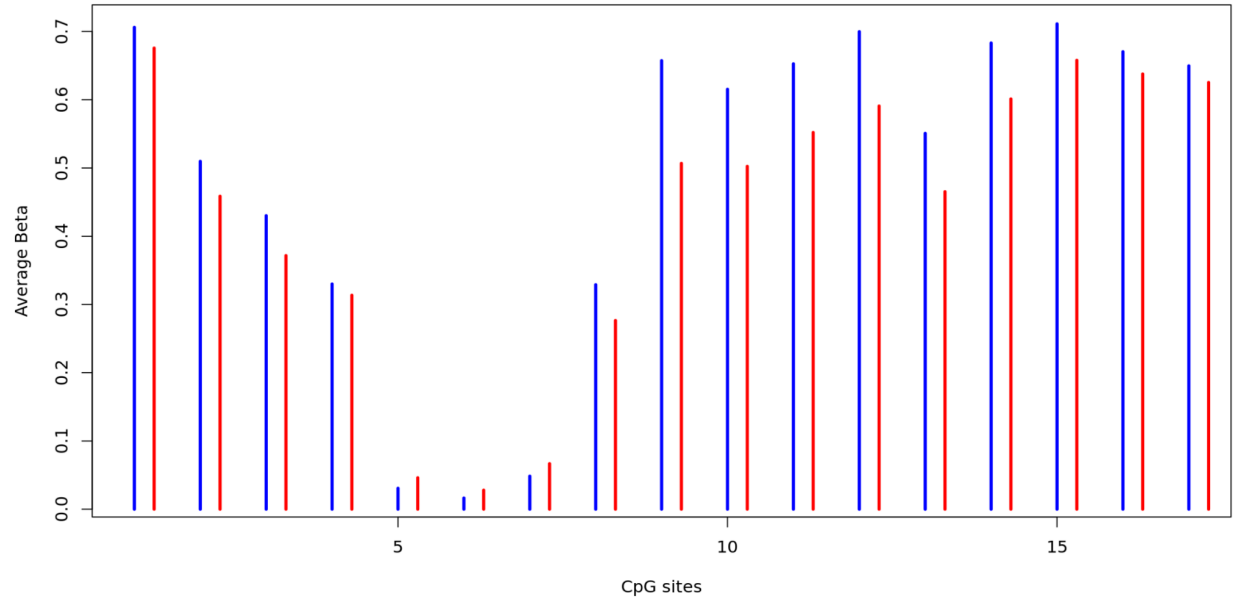
Blue = non-metastatic, red = metastatic



Notes: Generally aligns with what is filtered–the metastatic group is more upregulated as seen by higher counts, and CPG sites after 6 seem to be undermethylated. There are multiple sites where beta is about the same/slightly higher for metastatic sites; it's difficult to conclude anything from those sites.

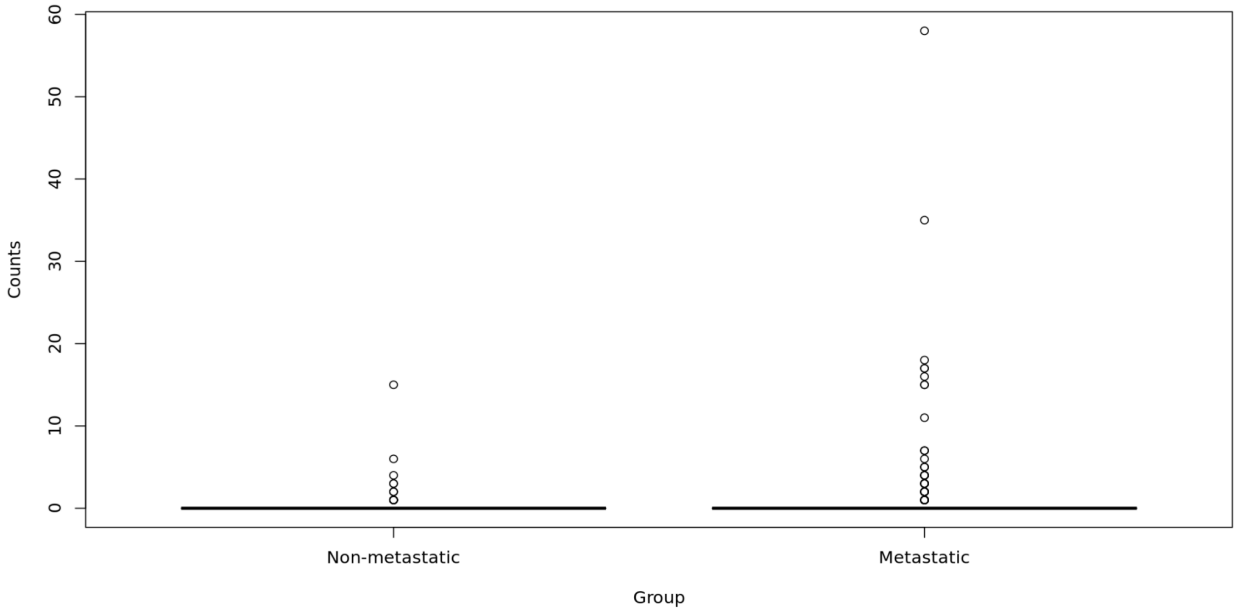
HHEX:

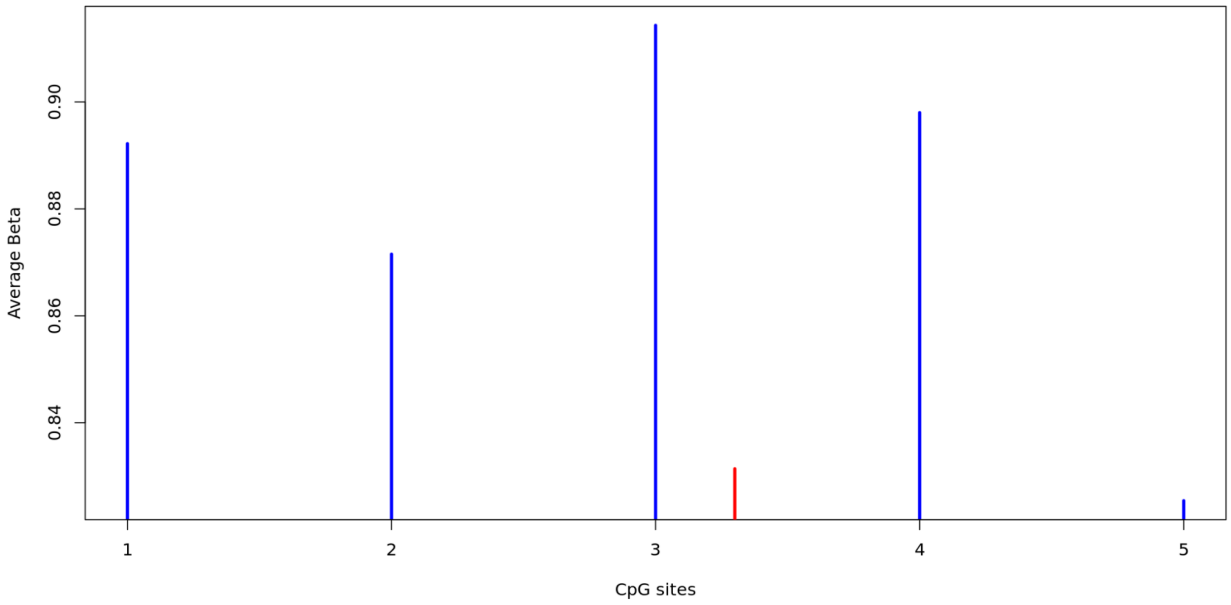




Notes: This gene gives a more definitive confirmation of what was filtered–the metastatic group has higher counts, and at the majority of CpG sites, there is a significant difference in methylation levels–the metastatic group is undermethylated. Like the previous gene, there are still a few sites where the metastatic group is more methylated compared to the non-metastatic group.

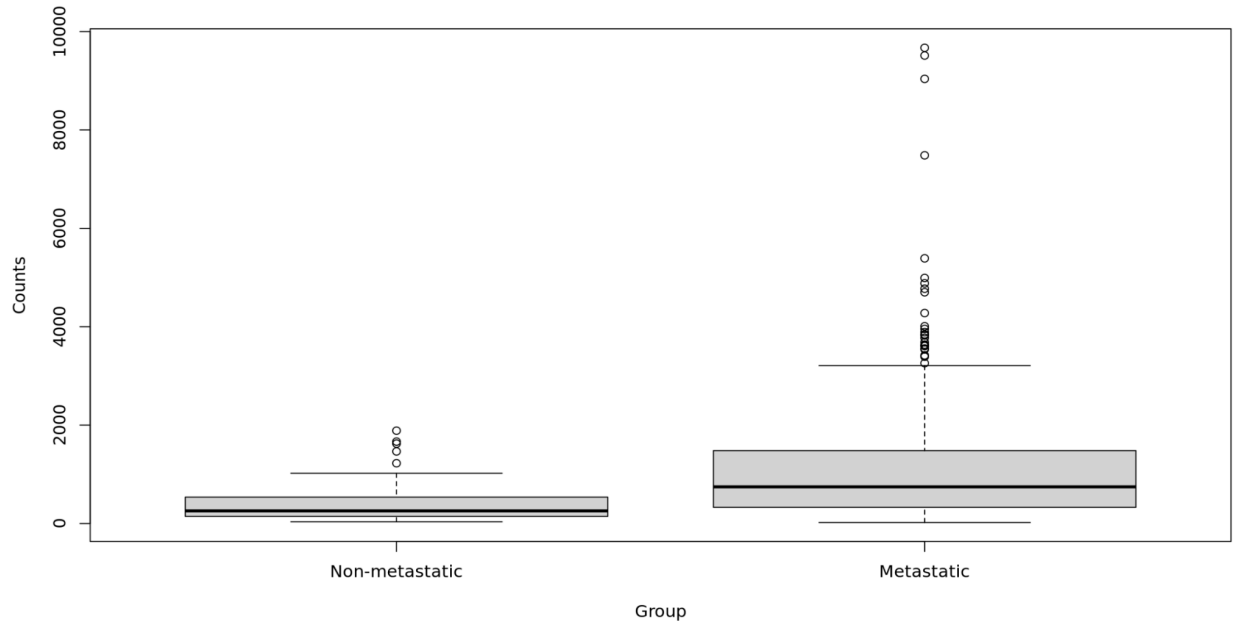
C7orf33:

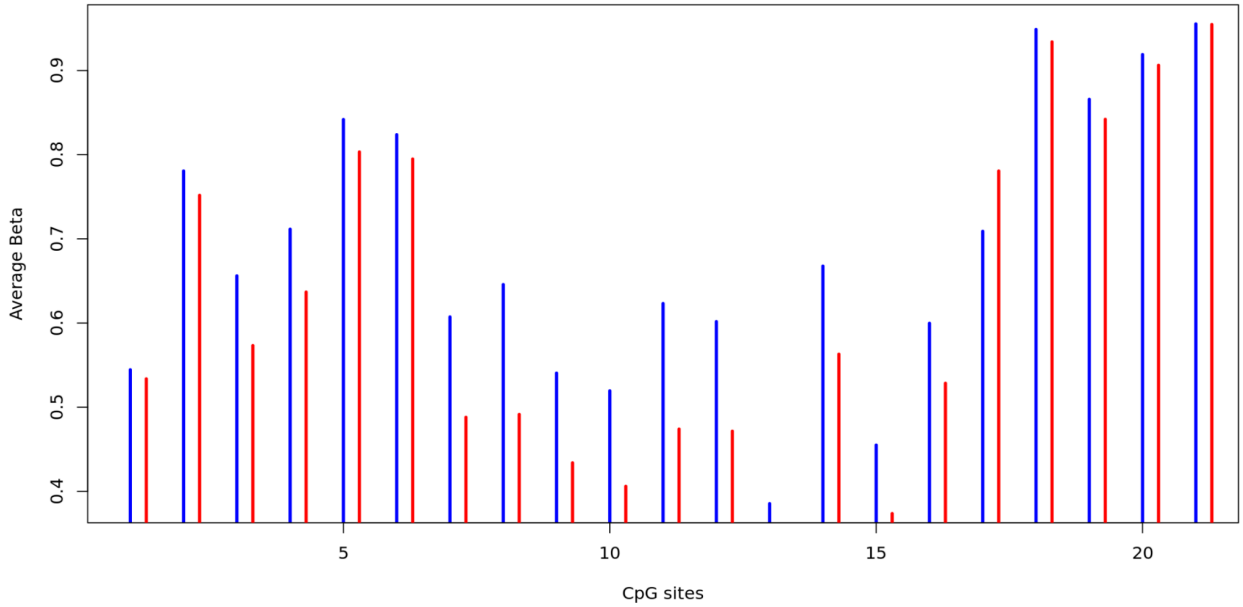




Notes: This gene gives the clearest evidence of methylation differences and undermethylation in metastatic patients; in 4 out of 5 sites, there seems to be no methylation present. It is also the gene with the least number of counts.

ARHGAP25:

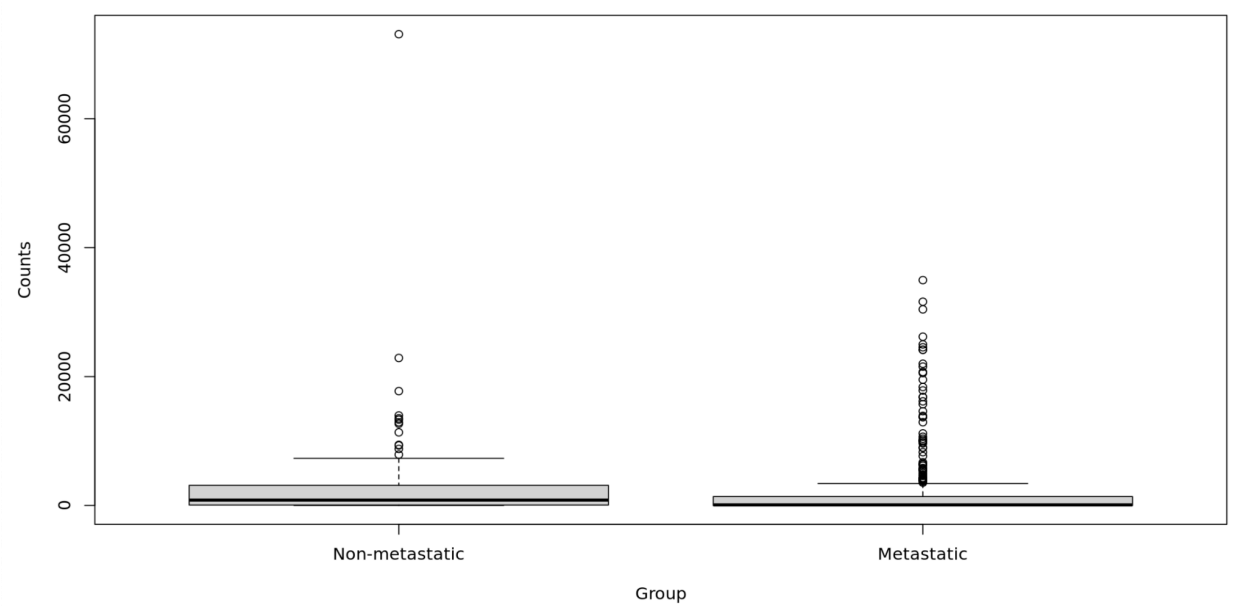


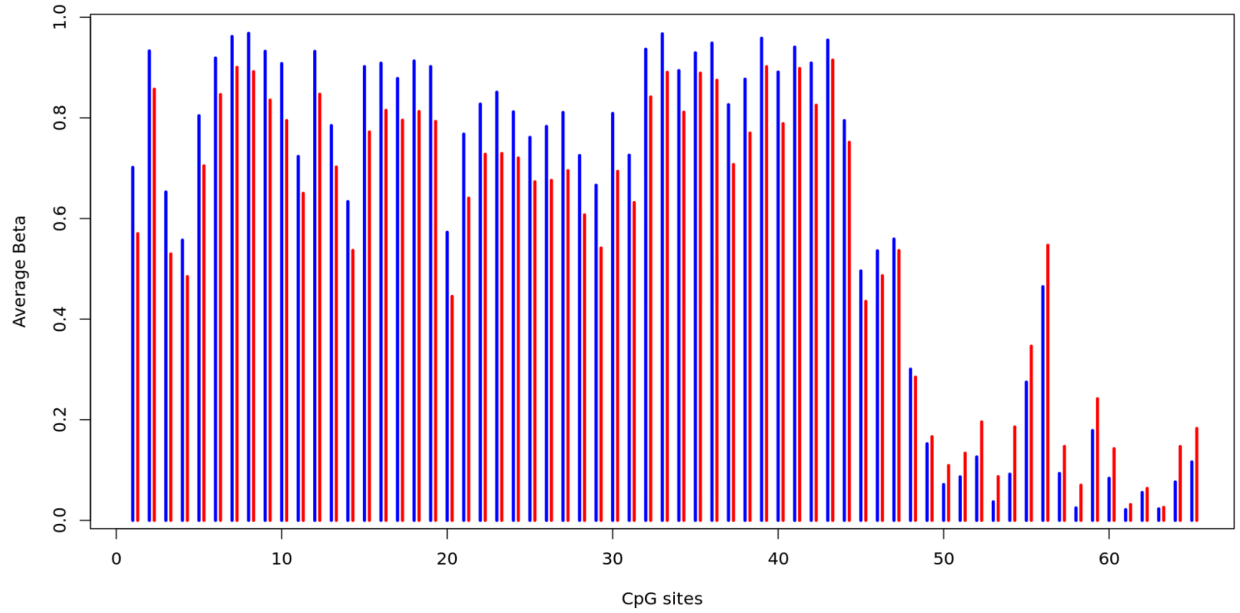


Notes: This gene gives the highest number of counts, clearly showing the upregulation of metastatic patients compared to non-metastatic patients. Methylation data also shows a difference in methylation levels, and it can be seen that sites for metastatic patients are less methylated.

Downregulated + undermethylated:

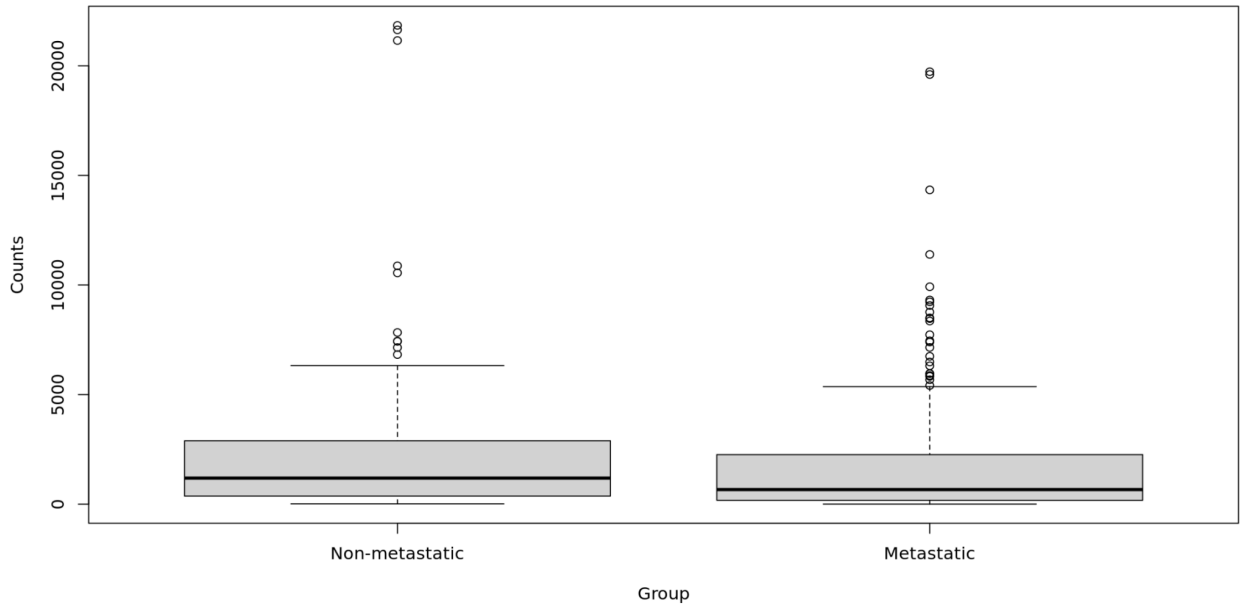
OCA2:

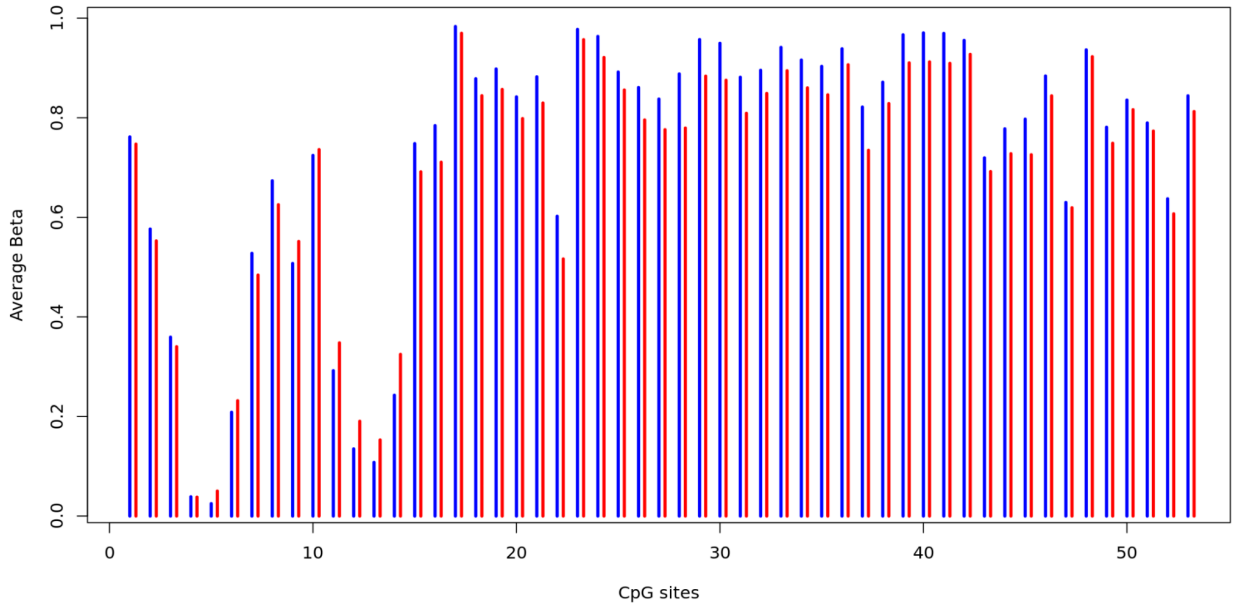




Notes: There is some conflicting evidence seen here; there is an outlier in the non-metastatic RNA counts data, which makes it appear that there is more upregulation in non-metastatic patients, but other than that, metastatic patients show more evidence of upregulation. CpG sites show clear undermethylation in metastatic patients, but after that, there seems to be greater methylation in metastatic patients.

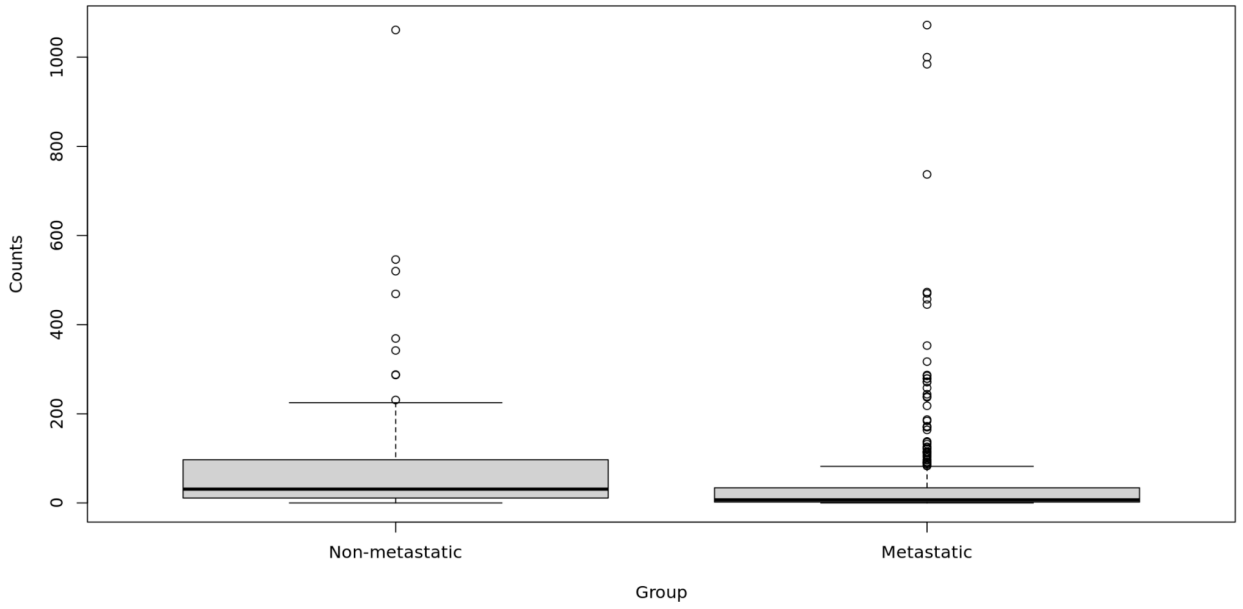
MGAT5B:

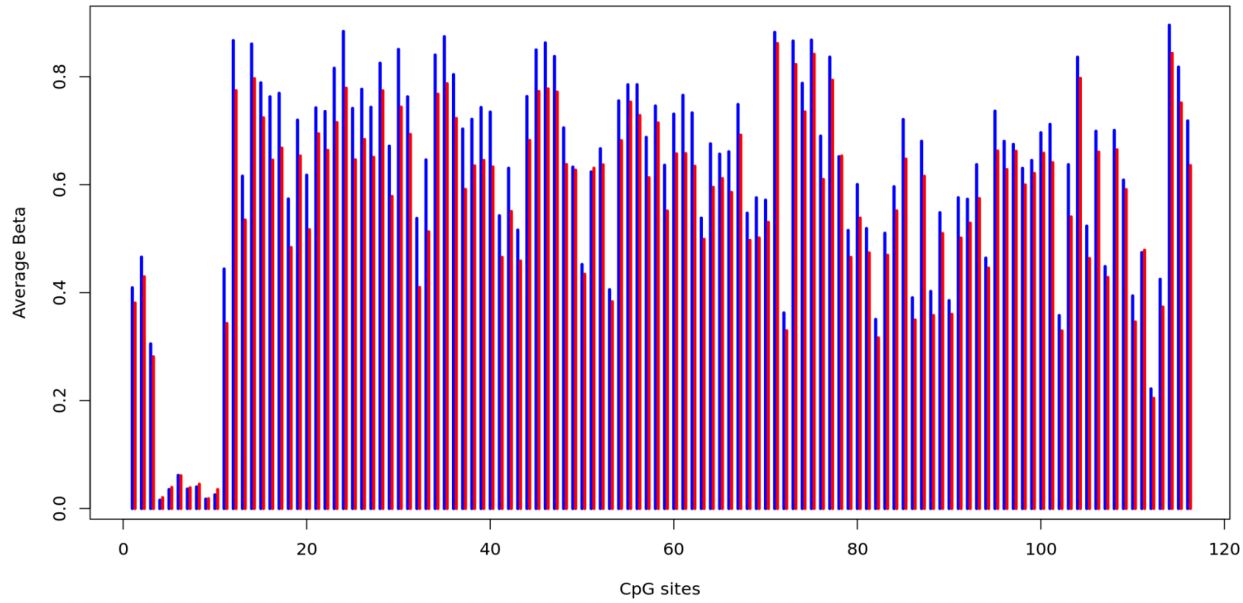




Notes: Like the previous gene, regulation levels seem close, though there is more evidence of upregulation in non-metastatic patients. The methylation graph shows better support for undermethylation in metastatic patients.

KNDC1:

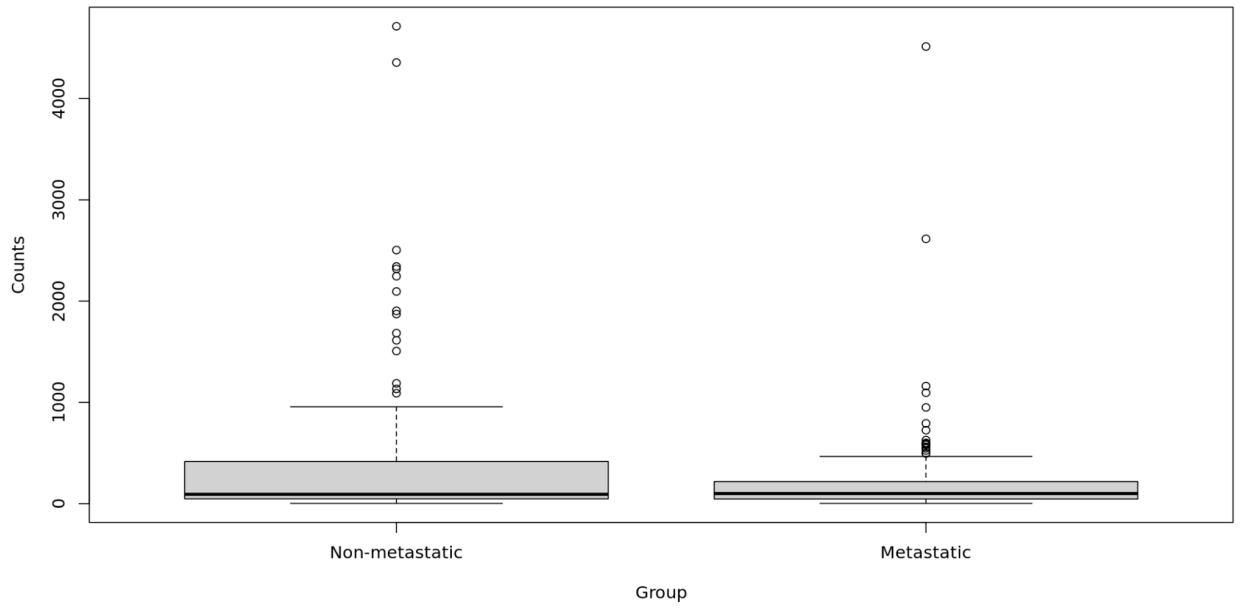


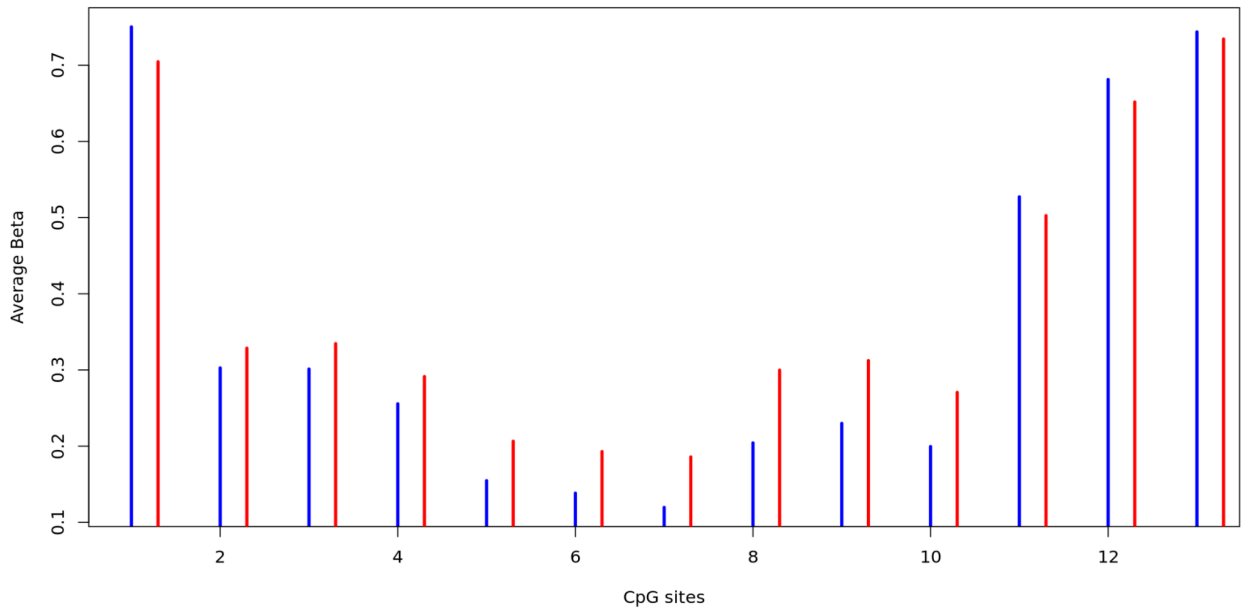


Notes: The counts boxplot is similar to the first gene's, where one outlier is dragging the non-metastatic group's counts up; however, the box is wider than the first gene's. The methylation graph shows the strongest evidence of undermethylation in metastatic patients, and it has the most CpG sites.

Downregulated + hypermethylated:

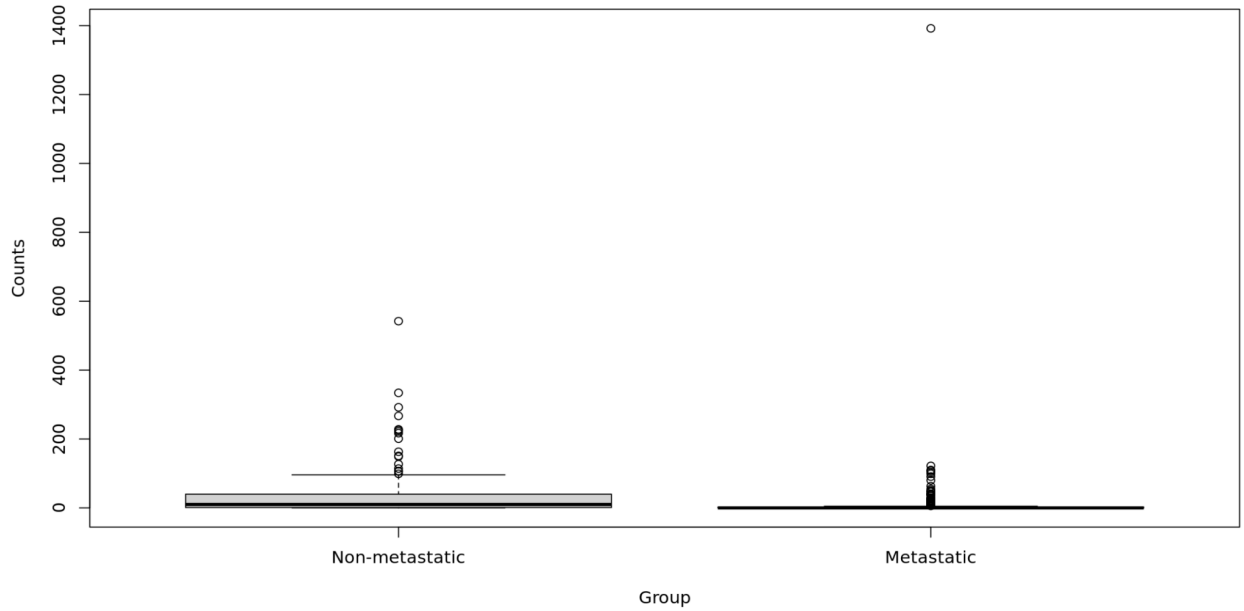
SPINT2:

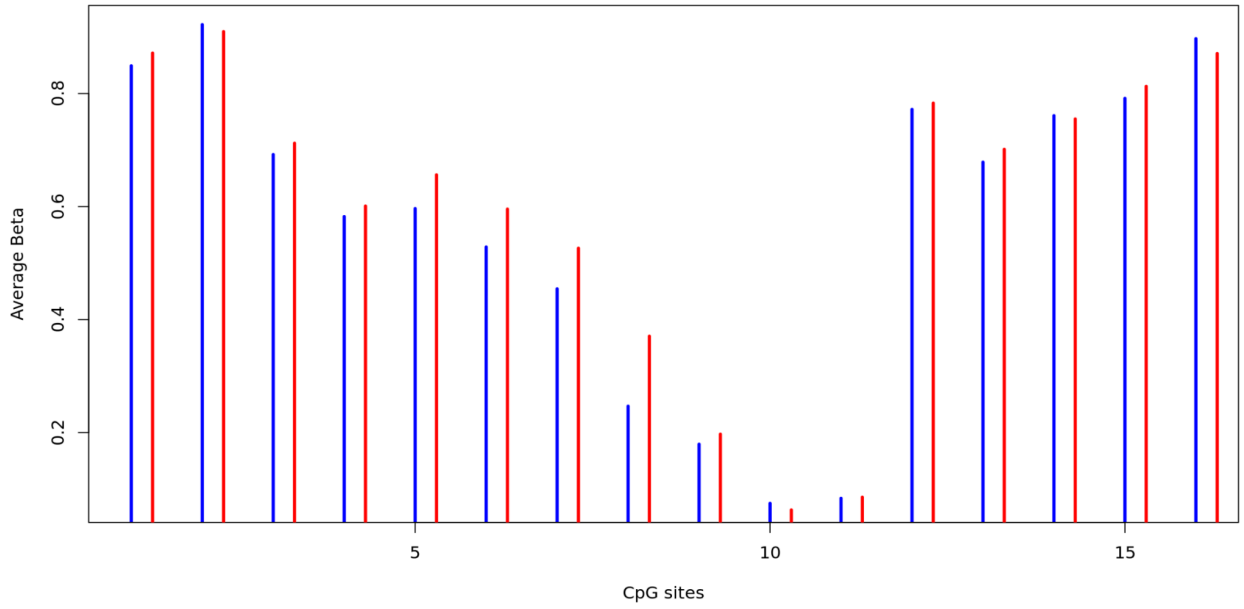




Notes: The counts plot aligns with downregulation in metastatic patients, and the methylation graph also generally aligns with hypermethylation in metastatic patients. CpG sites near the more extreme values seem to stray from this, as there is less methylation in metastatic patients at those sites.

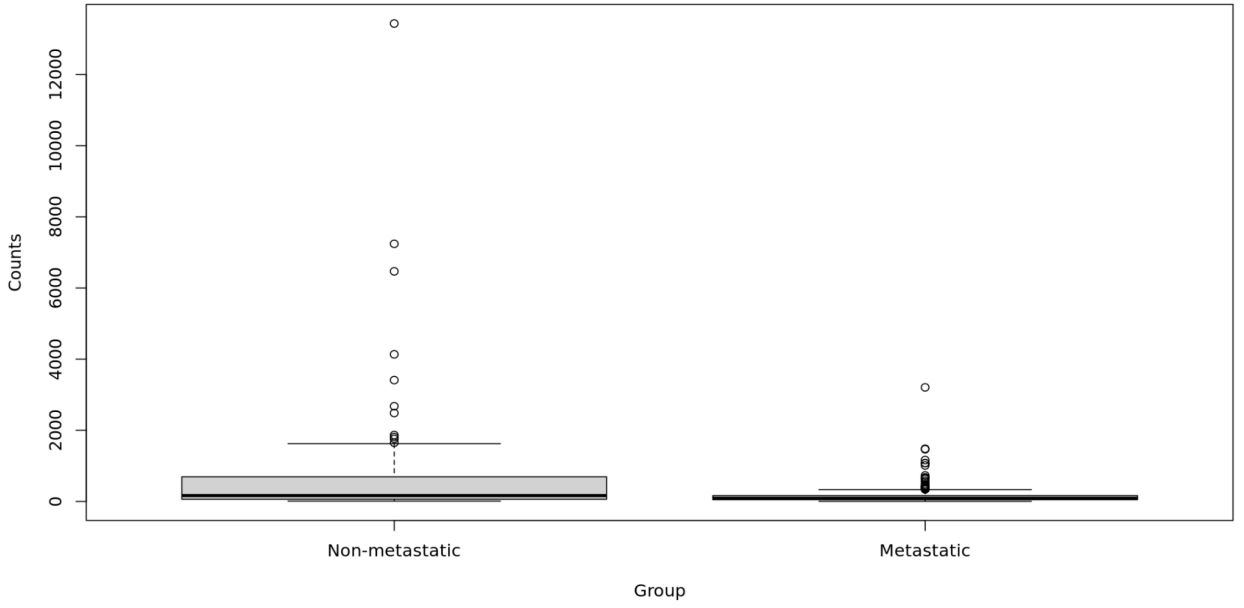
CD164L2:

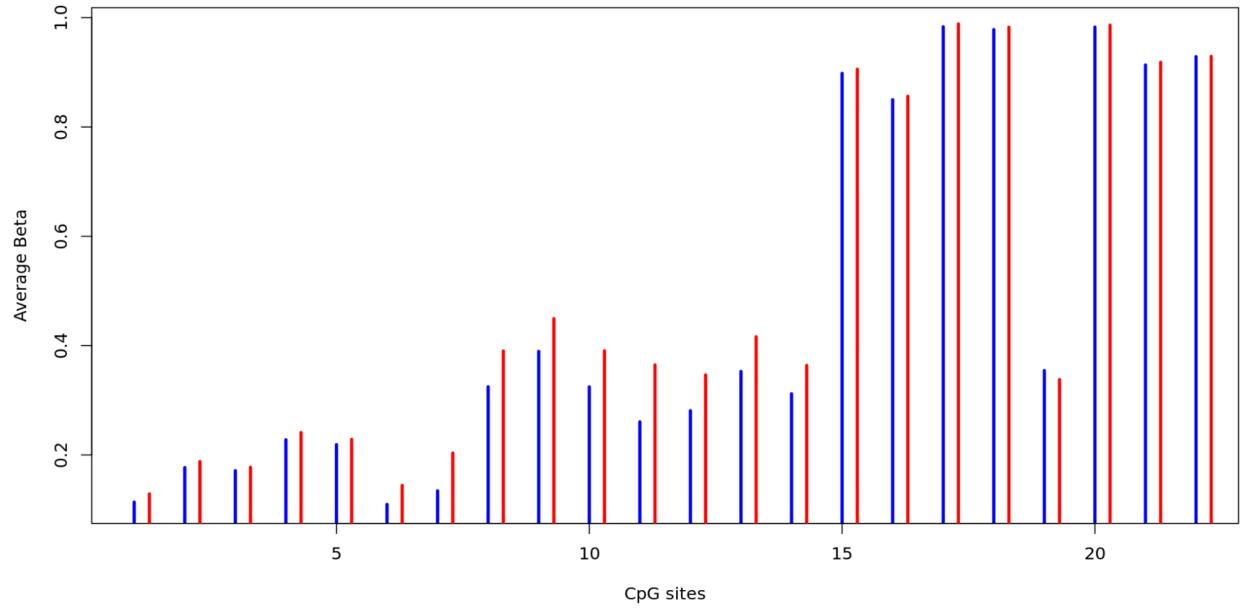




Notes: Shows clear evidence of downregulation in metastatic patients, aside from the outlier near the top; most sites show evidence of hypermethylation in metastatic patients.

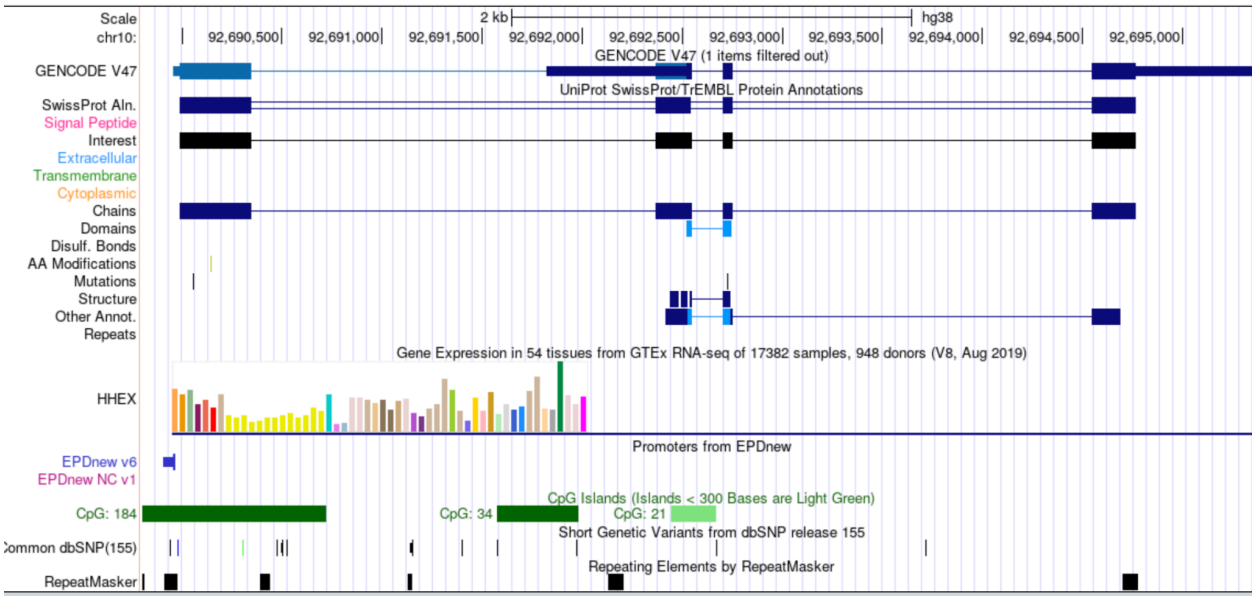
HAS3:



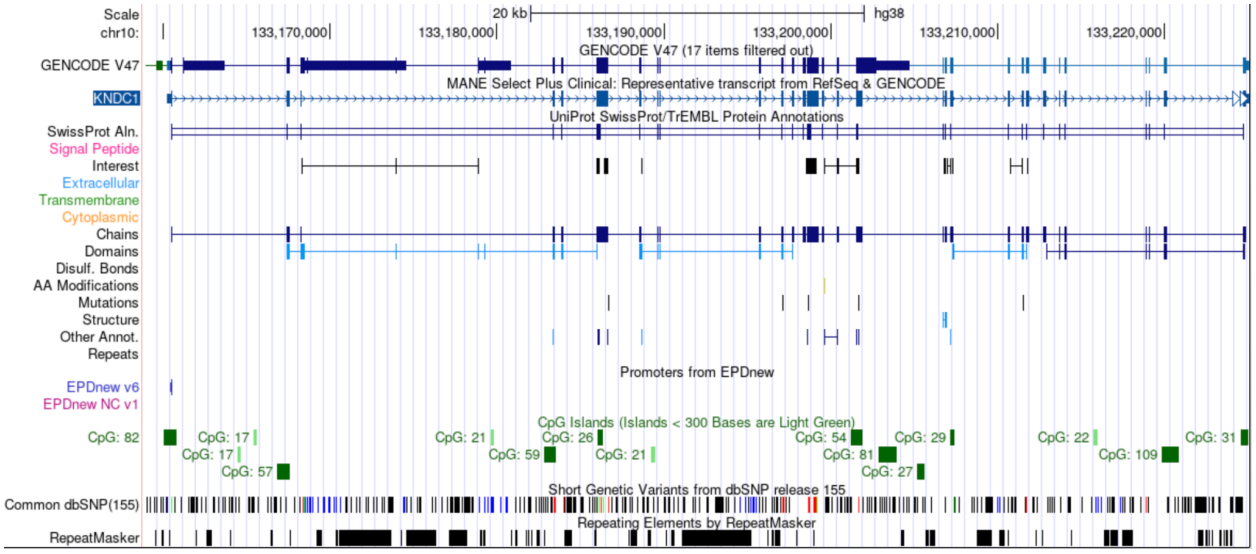


Notes: Shows the clearest evidence of downregulation in metastatic patients; no metastatic outliers are present and the counts range is the largest. The methylation graph also shows clear evidence of hypermethylation in metastatic patients, though the methylation levels are about the same in rightmost CpG sites.

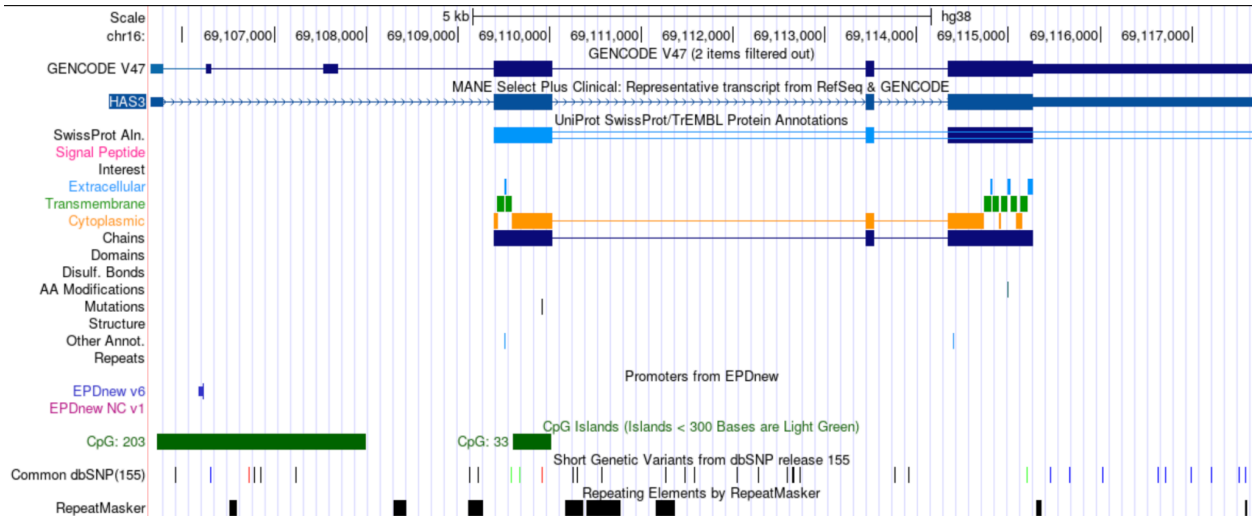
5)
HHEX:


Scale
chr10:
2 kb
hg38
92,690,500
92,691,000
92,691,500
92,692,000
92,692,500
92,693,000
92,693,500
92,694,000
92,694,500
92,695,000
GENCODE V47
GENCODE V47 (1 items filtered out)
UniProt SwissProt/TrEMBL Protein Annotations
SwissProt Aln.
Signal Peptide
Interest
Extracellular
Transmembrane
Cytoplasmic
Chains
Domains
Disulf. Bonds
AA Modifications
Mutations
Structure
Other Annot.
Repeats
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
HHEX
Promoters from EPDnew
EPDnew v6
EPDnew NC v1
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 184
CpG: 34
CpG: 21
Short Genetic Variants from dbSNP release 155
Common dbSNP(155)
Repeating Elements by RepeatMasker
RepeatMasker

KNDC1:


Scale
chr10:
20 kb
hg38
133,170,000
133,180,000
133,190,000
133,200,000
133,210,000
133,220,000
GENCODE V47
GENCODE V47 (17 items filtered out)
KNDC1
MANE Select Plus Clinical: Representative transcript from RefSeq & GENCODE
UniProt SwissProt/TrEMBL Protein Annotations
SwissProt Aln.
Signal Peptide
Interest
Extracellular
Transmembrane
Cytoplasmic
Chains
Domains
Disulf. Bonds
AA Modifications
Mutations
Structure
Other Annot.
Repeats
Promoters from EPDnew
EPDnew v6
EPDnew NC v1
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 82
CpG: 17
CpG: 17
CpG: 57
CpG: 21
CpG: 59
CpG: 26
CpG: 21
CpG: 54
CpG: 81
CpG: 29
CpG: 27
CpG: 22
CpG: 109
CpG: 31
Short Genetic Variants from dbSNP release 155
Common dbSNP(155)
Repeating Elements by RepeatMasker
RepeatMasker

HAS3:



Article:

In the article *Aberrant DNA Methylation Predicts Melanoma-Specific Survival in Patients with Acral Melanoma*, the researchers found that HHEX was differentially methylated in patients with normal primary acral lentiginous melanoma (PALM) compared to parents with metastatic acral lentiginous melanoma. This seems to support my findings for HHEX; changes in methylation can increase the risk of metastasis. The researchers also found that hypermethylation of HHEX was observed in metastatic patients, which slightly differs from what I found (undermethylation), but I did see some CpG sites in my graph where there was hypermethylation in metastatic patients, which agrees with what the researchers found.

References

OpenAI. (2024). *ChatGPT* [Large language model]. https://chatgpt.com

Pradhan, D., Jour, G., Milton, D., Vasudevaraja, V., Tetzlaff, M. T., Nagarajan, P., Curry, J. L., Ivan, D., Long, L., Ding, Y., Ezhilarasan, R., Sulman, E. P., Diab, A., Hwu, W. -J., Prieto, V. G., Torres-Cabala, C. A., & Aung, P. P. (2019). Aberrant DNA Methylation Predicts Melanoma-Specific Survival in Patients with Acral Melanoma. *Cancers*, 11(12), 2031. https://doi.org/10.3390/cancers11122031