Part 1: Review Questions

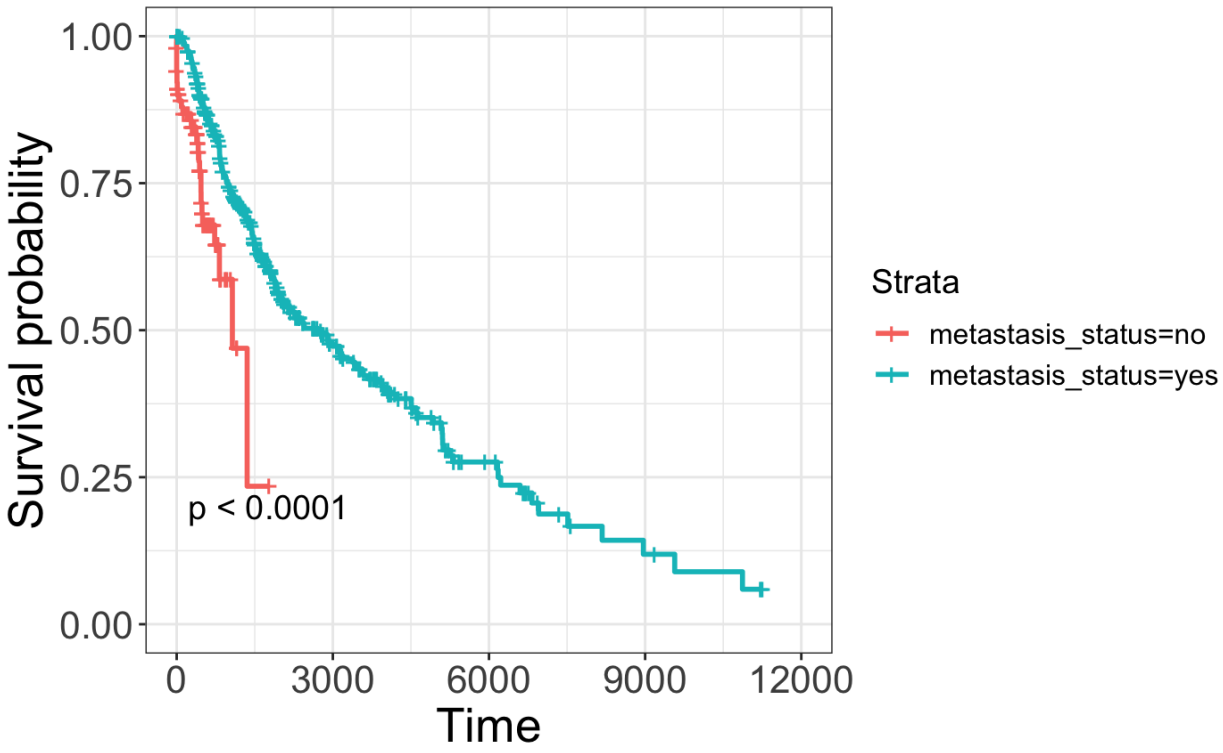
1) TCGA is a multi-omic dataset organized by the National Cancer Institute and the National Human Genome Research Institute. TCGA allows us to explore a wide range of genes across a large patient sample and understand how genomics and transcriptomics impact the progression of various cancers. Data from TCGA is de-identified to protect patient privacy, and so different samples are identified via unique patient barcodes.
2) Strengths and weaknesses of TCGA:
    a) Strengths: numerous clinical variables included for 33 cancer types; over 20000 patient samples
    b) Weaknesses: often many NA values → require processing of data prior to actual analysis

Coding skills:

1) Saving to github repository:
    a) cd into local repository
    b) Check git status to see which local files need to be pushed
    c) git add *file*
    d) git commit -m “message”
    e) git push
2) First, we need to install the package if it hasn’t been installed already. Then we must load it using library(library_name).
3) Bioconductor package:
    a) if (!require("BiocManager", quietly = TRUE))
    b) install.packages("BiocManager")
    c) library(BiocManager)
4) Boolean indexing = applying a vector of boolean values to either a column or row in a particular dataframe
    a) True = select data, false = ignore data
5) df [age] == c (‘young, ‘old’, ‘old’, ‘young’, ‘old’, ‘young’)
6) age_mask <- c (F, T, T, F, F, T)
7) Select for all old patients:
    a) rows = age_mask, columns = all
    b) df[age_mask,]
    c) df$old <- ifelse(‘old’, T, F)

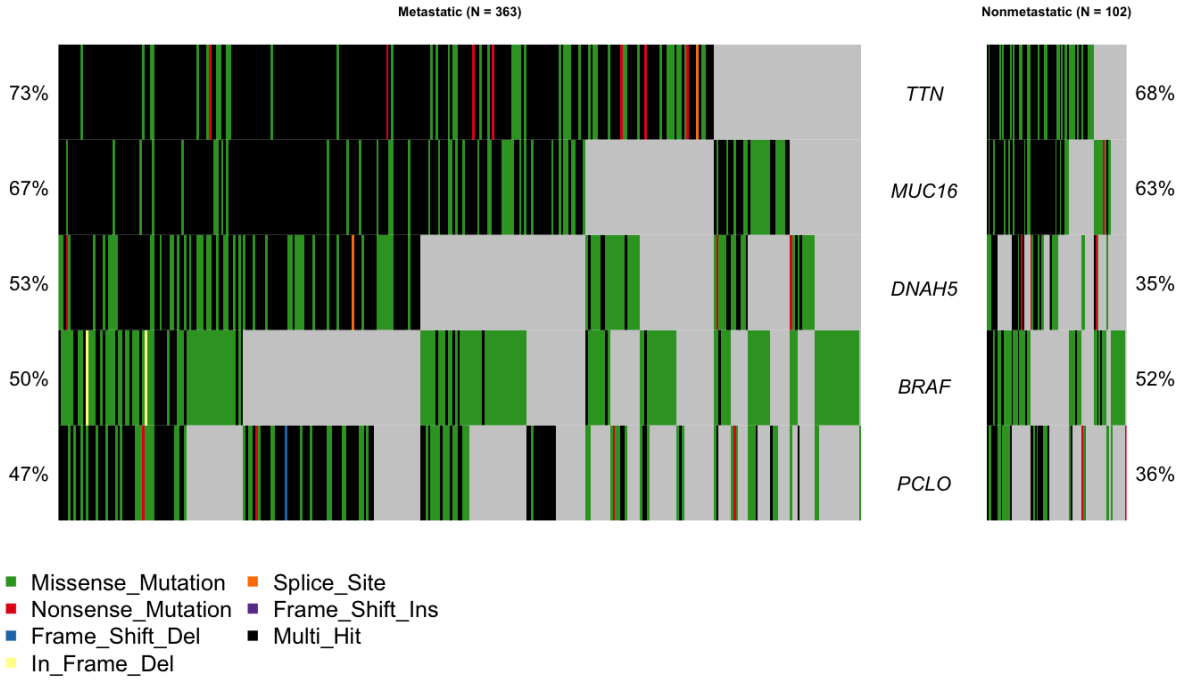
Part 3: SKCM Analysis

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



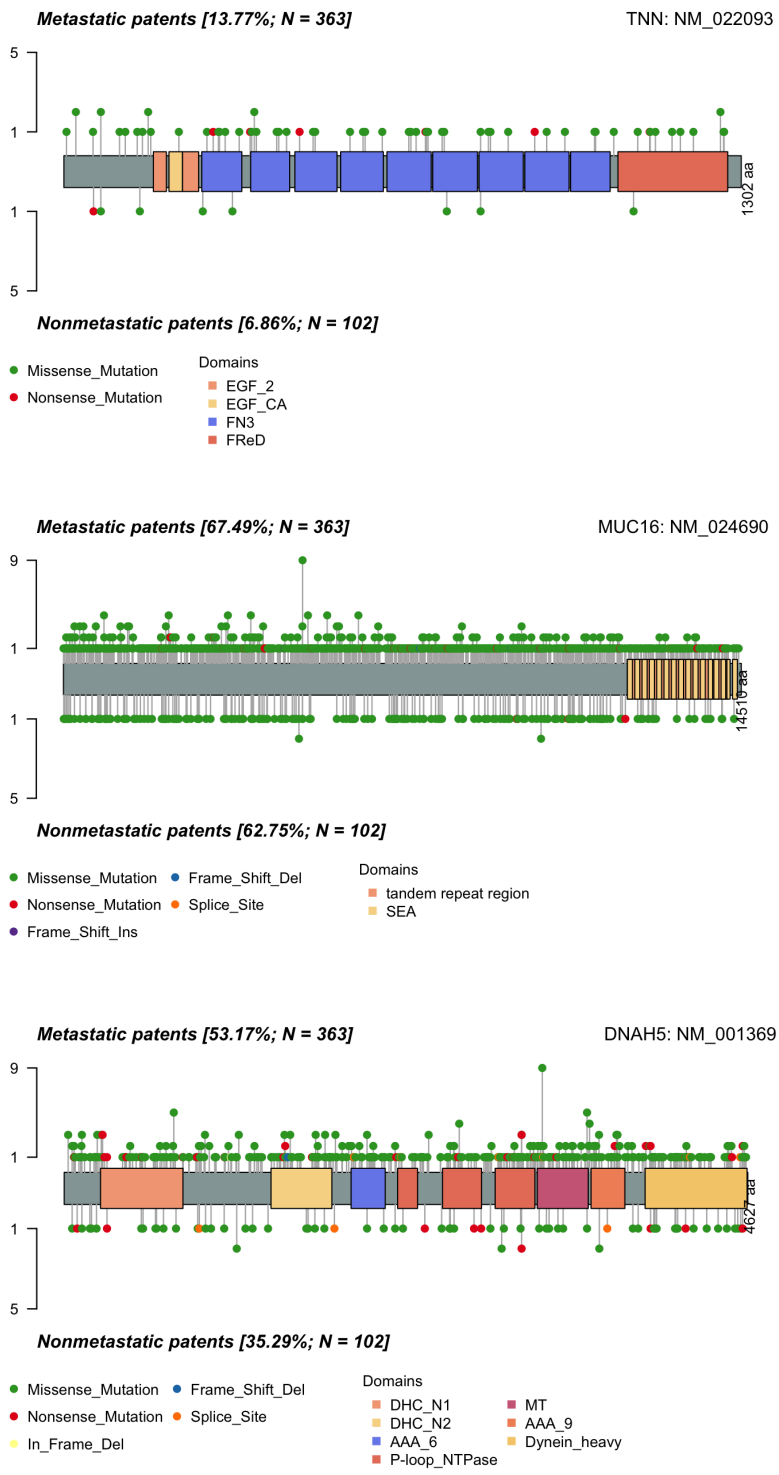
This is a Kaplan Meier plot, which shows two curves plotting survival probability. In this case, we see that patients with metastatic cancer have an overall lower survival probability as opposed to patients without metastasis. Because the p-value is less than 0.0001 and therefore less than 0.05, we have evidence to believe that this plot shows a statistically significant relationship between survival probability and metastasis status. One conclusion we cannot draw is how long they live, only that one is less likely to survive than the other. Literature supports the fact that the mobility of the cancer plays a big role in patient survival.

2) Mutation differences between metastatic and non-metastatic patients for multiple genes.



This is a co-oncoplot, which allows us to look at the most highly mutated genes. We see that though TTN and MUC16 are most mutated in both metastatic and nonmetastatic cancers, DNAH5 and PCLO are most differentially mutated between the two, with higher mutation rates in metastatic SKCM. Most of the mutations appear to be green and black, indicating both missense and multi-hit mutations. One conclusion we cannot draw is how differentially expressed they are as this co-oncoplot looks at mutation data, not transcriptomic data. Literature shows that TTN is the most commonly mutated gene skin cutaneous melanoma.

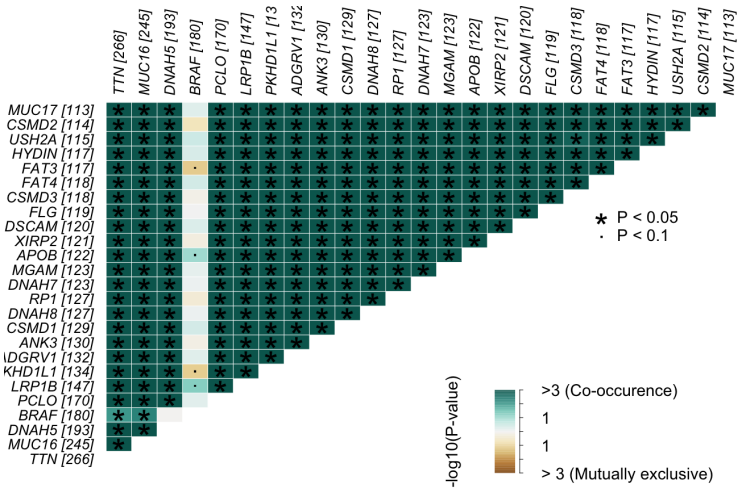
3) Mutation differences for specific genes of interest



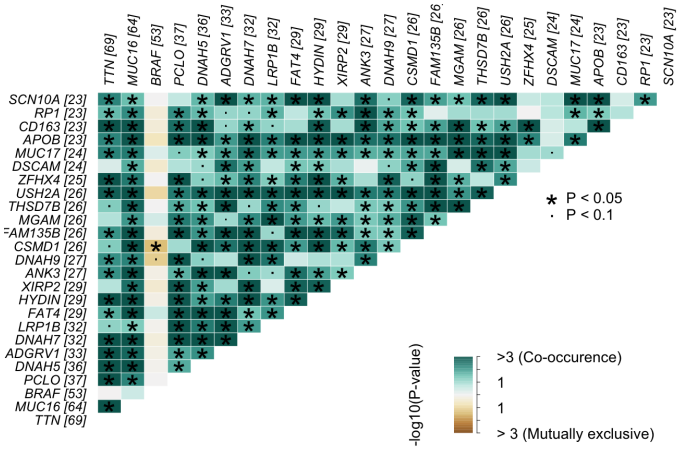
In order to understand mutational differences between metastatic and nonmetastatic patients for different genes of interest, I used co-lollipop plots. I looked at TTN, MUC16, and DNAH5 as they were either the most highly mutated or most differentially mutated between the two groups. TTN and DNAH5

all appear more highly mutated in metastatic patients here, while it is difficult to tell with MUC16. The majority green color for all three plots implies that missense mutations are the most likely type of mutations of this gene. We cannot make any major conclusions about mutations between metastatic and non-metastatic patients for MUC16 and DNAH5 because there doesn't appear to be a big difference in the two groups.

4) Co-occurrence or mutual exclusion common gene mutations: one for metastatic patients, one for non-metastatic patients



Metastatic

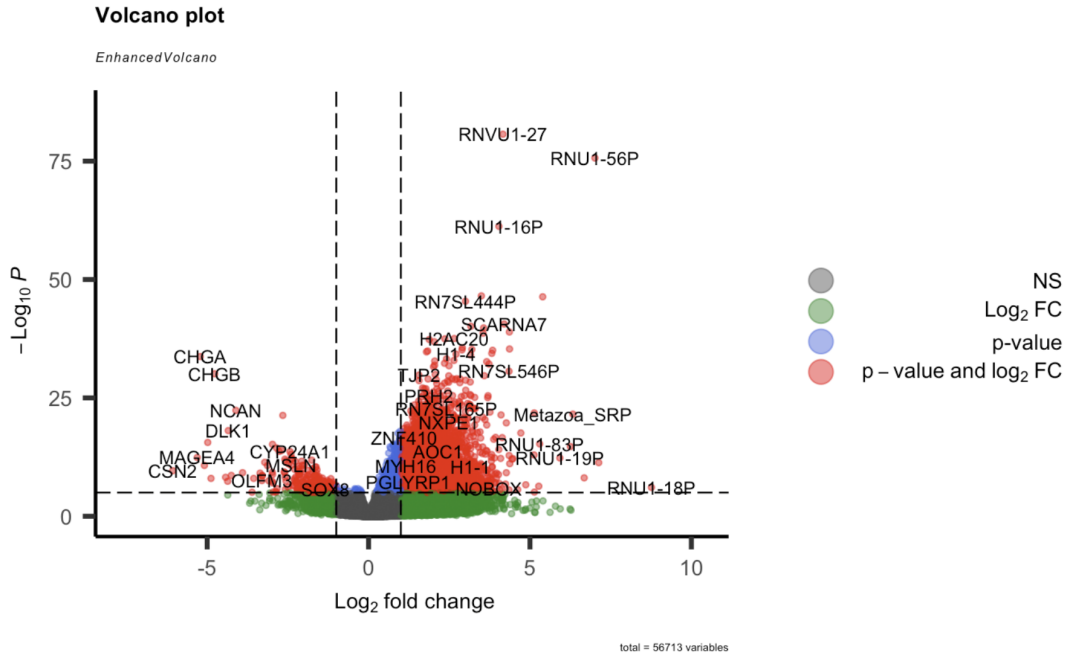


Non-metastatic

I made a somatic interactions plot to understand how different genes associated with cancer might influence each other. In the metastatic somatic interactions plot, almost all the genes are co-occurring, and so we don't really have any interesting conclusions to make. In the non-metastatic somatic interactions plot, CSMD1 and BRAF are apparently mutually exclusive. We cannot draw many conclusions

due to the fact that most of the genes are co-occurring, so nothing stands out apart from the previously mentioned CSMD1 and BRAF. There is literature looking at the interactions between BRAF and other genes (such as CSMD1) to understand how SKCM treatment might affect such interactions.

5) Differential expression between non-metastatic and metastatic patients controlling for treatment effects, race, gender, and vital status



We conducted a differential sequence analysis using DESeq2 to understand how the genetic mutations in metastatic and non-metastatic cancer patients translate to transcriptomic gene expression. None of the genes of interest identified above in neither the co-oncoplots nor the somatic interactions plots appeared in the differential sequence analysis. This means we cannot conclude that the genetic mutations identified above translate to transcriptomic expression, and thus more research is needed to make any conclusions. I couldn't find much literature to support this due to the fact that I wasn't able to make any conclusion from this graph.

References used:

Debugging:
https://stackoverflow.com/questions/46811818/including-function-from-survival-in-r-package
https://rdrr.io/bioc/maftools/src/R/read_maf_dt.R

Part 3:

Number 1: Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A., & Fares, Y. (2020). Molecular principles of metastasis: a hallmark of cancer revisited. *Signal transduction and targeted therapy*, *5*(1), 28. https://doi.org/10.1038/s41392-020-0134-x

Number 2: Wang, Q., Huang, X., Zeng, S., Zhou, R., & Wang, D. (2023). Identification and validation of a TTN-associated immune prognostic model for skin cutaneous melanoma. *Frontiers in genetics*, *13*, 1084937. https://doi.org/10.3389/fgene.2022.1084937

Number 4: Lee, K. H., Goh, J., Kim, Y. J., & Kim, K. (2020). Identification of synthetic chemosensitivity genes paired with BRAF for BRAF/MAPK inhibitors. *Scientific reports*, *10*(1), 20001. https://doi.org/10.1038/s41598-020-76909-2