Peyton Hall
QBIO490: Directed Research - Multi-Omic Analysis
13 October 2023

R Review Project

Part I. Review Questions

General Concepts:

1. TCGA is The Cancer Genome Atlas, it is important because it provides publicly available longitudinally collected genomic, epigenomic, proteomic, and transcriptomic data. Due to the large amount and accessibility of the data, TCGA has transformed understanding of the genetic changes associated with various cancers.
2. Some strengths of TCGA are that patient information is publicly available, increasing accessibility. Some weaknesses are missing data (which can be compensated for by creating is.na Boolean mask).

Coding Skills:

1. Ls to show current contents of directory, Cd to change directory, Git status to check git status, git add "file name" to add file, git status to find where file is, git commit -m "alternate name for file", git status to see current status of file, git push to send to GitHub repository, and git status one last time to make sure the file is in the repository.
2. To use a package in R, one must load by library(package). If the package is not in R already, one must use the following code:
   ```
   if (!require("package", quietly = TRUE))
     BiocManager::install("package")
   library(package)
   ```
3. To use a Bioconductor package in R, one might use
   ```
   if (!require("Bioconductor", quietly = TRUE))
     BiocManager::install("Bioconductor")
   library(Bioconductor)
   ```
4. Boolean indexing is adding a Boolean value (true or false) to a row or column within a dataframe. An application of this is creating a Boolean mask.

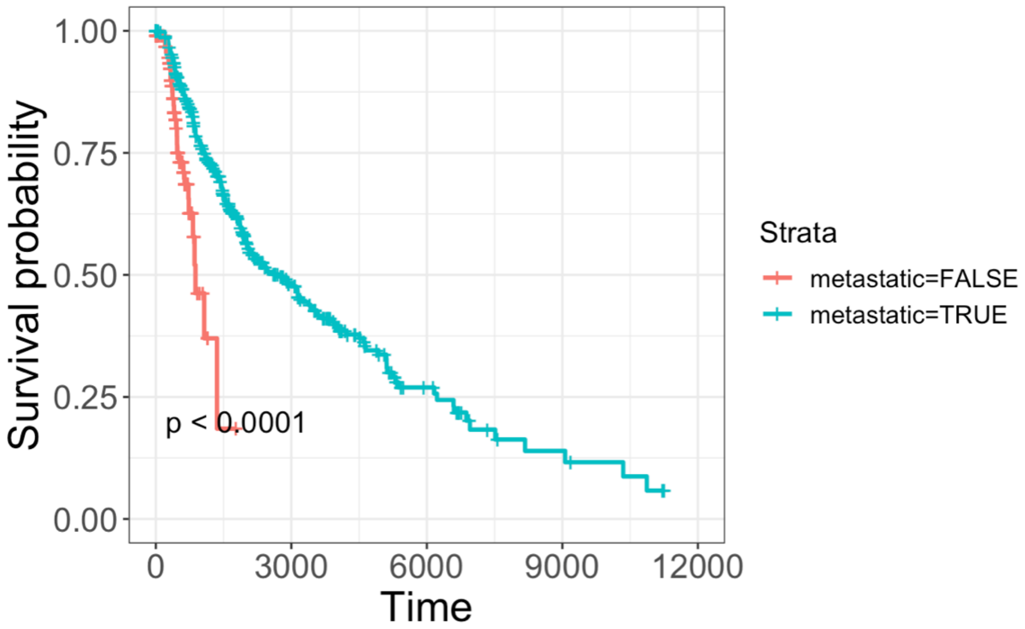
Patient_Barcode	Race	Gender	Tumor_status	Days_at_index
TCGA-02-200-2999-299873	BLACK OR AFRICAN AMERICAN	FEMALE	WITH TUMOR	676
TCGA-02-200-2999-299873	WHITE	FEMALE	TUMOR FREE	754

a. with_tumor_mask <- ifelse(clinical$tumor_status == "WITH TUMOR", T, F)
   clinical <- clinical[With_tumor_mask,]
   This creates a mask where anything not "WITH TUMOR" in tumor status is false. The second line applies it to the data frame.

b. Gender_na_mask <- !is.na(clinical$gender)
   Clinical <- clinical[gender_na_mask,]
   This creates a mask that removes na values from the Gender column and the second line applies the mask to the data frame.

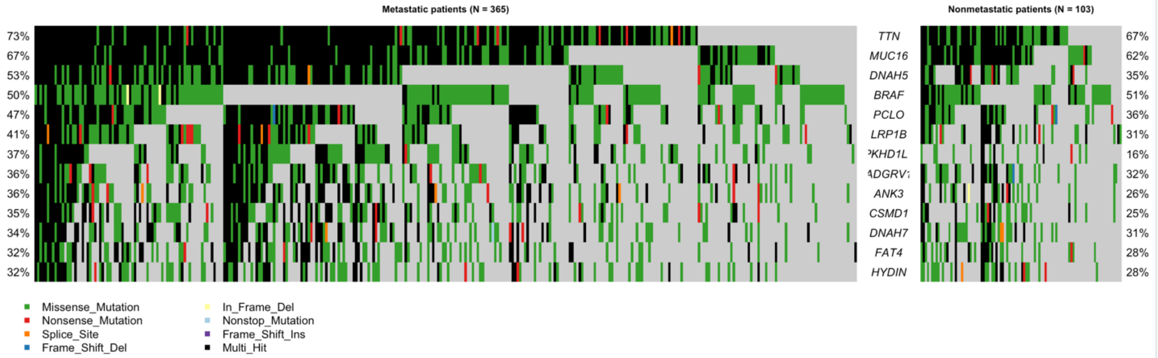
Part II. SKCM Analysis and Part III. Results and Interpretations

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



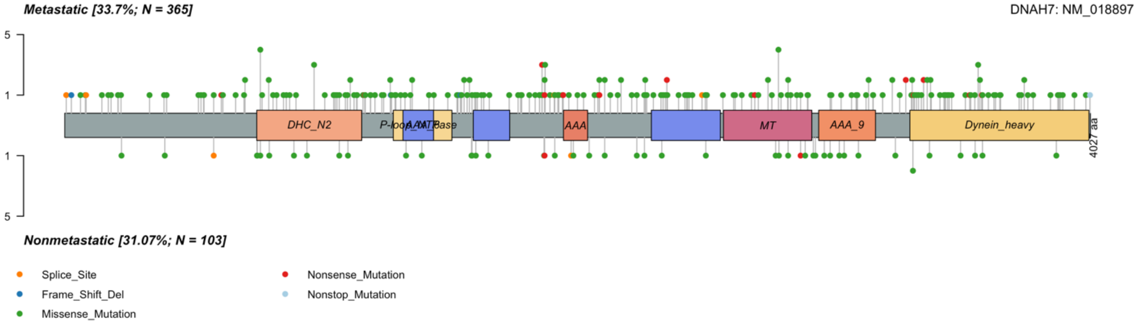
This plot is a KM plot, which shows the probability that the patient may survive up to a certain time. On the x axis, is time in days. On the y axis, is survival probability between 0.00 and 1.00. In this plot, it appears non-metastatic survival probability is higher (stops closer to 0.25) and metastatic drops lower to 0. However, the survival time for non-metastatic appears to be more prolonged (up to 12,000 days). A conclusion I cannot draw is that the non-metastatic patients do not appear past 1,500 days because their cancer was effectively treated, and it was the last they were in the study. This may be due to missing data or other data related issues. Metastasis describes the spread of cancer and metastatic cancer is cancer that has spread throughout the body, beyond the point of cancer origin. Given, the survival for metastatic cancer (stage IV) is lower than for patients with non-metastatic cancer. This supports the KM plot since the survival probability ends higher for non-metastatic patients than non-metastatic patients.

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



This plot is a Co-Oncoplot. Oncoplots are used to visualize the mutational landscape of patients using a list of specific genes of interest. The genes of interest in this Oncoplot include TTN, MUC16, DNAH5, BRAF, PCLO, LRP1B, PKHD1L, ANK3, CSMD1, DNAH7, FAT4, and HYDIN. The colors represent the type of mutations. A conclusion I can draw about differences between metastatic and nonmetastatic patients is that metastatic patients have more mutations within the TTN gene than nonmetastatic patients (73% vs 67% respectively). One conclusion I cannot draw is that missense mutations are involved in most multi_Hit mutations due to the already high number of missense mutations. It may be possible, but information is not provided on what type of mutations the multi_hit mutations consist of. According to one study, *BRAF* somatic missense mutations are present in in 66% of malignant melanomas and at lower frequency in a wide range of human cancers.

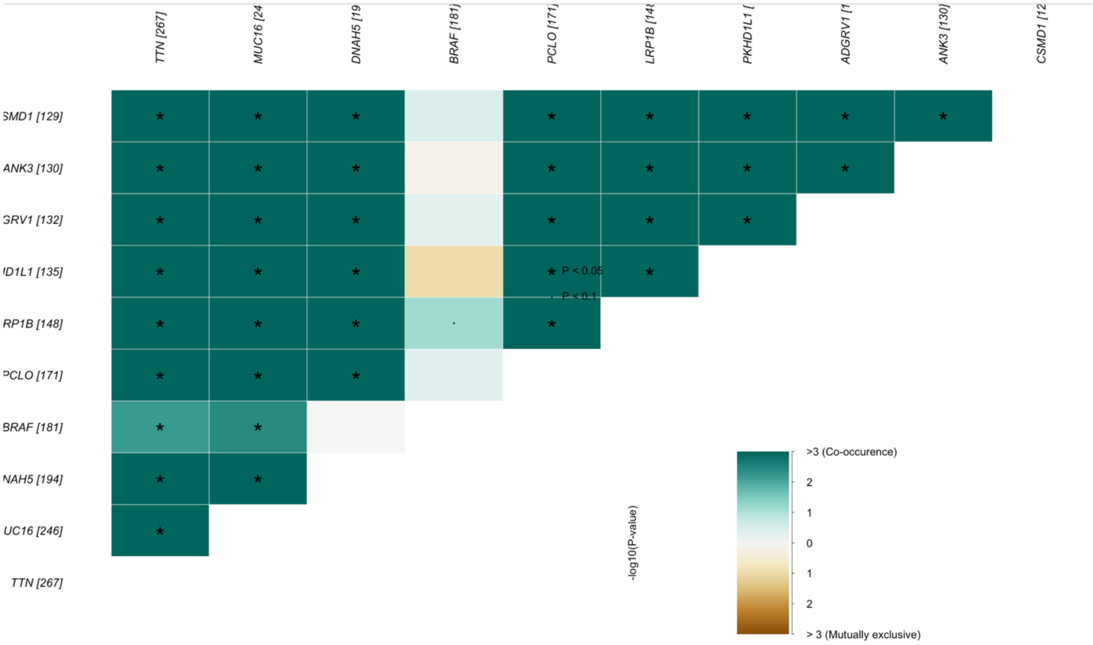
3. *Mutation differences for DNAH7*



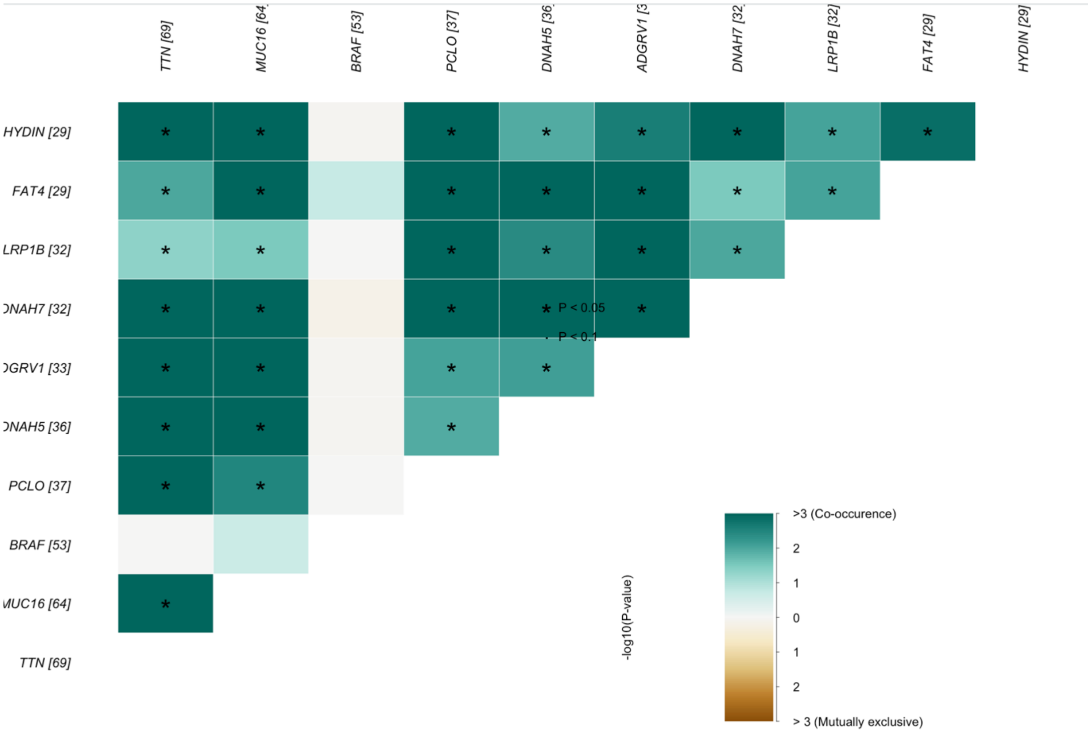
This plot is a lollipop plot, which displays mutation count by gene location. The horizontal axis is the location of the gene (DNAH7), with protein domains indicated. The lollipop colors represent the type of mutation. A conclusion I can draw is that most mutations within the DNAH7 gene is that most across both metastatic and nonmetastatic patients are missense mutations, evident in the large number of green lollipops. A conclusion I cannot draw is the exact location within DNAH7 of most splice site mutations of either metastatic or metastatic since the orange lollipop is not in a specific protein domain. In support of the conclusion, it has been found that missense mutations were the main form of mutation observed in *DNAH7 within a study involving DNAH7 mutations in colorectal cancer patients*.

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

Metastatic:

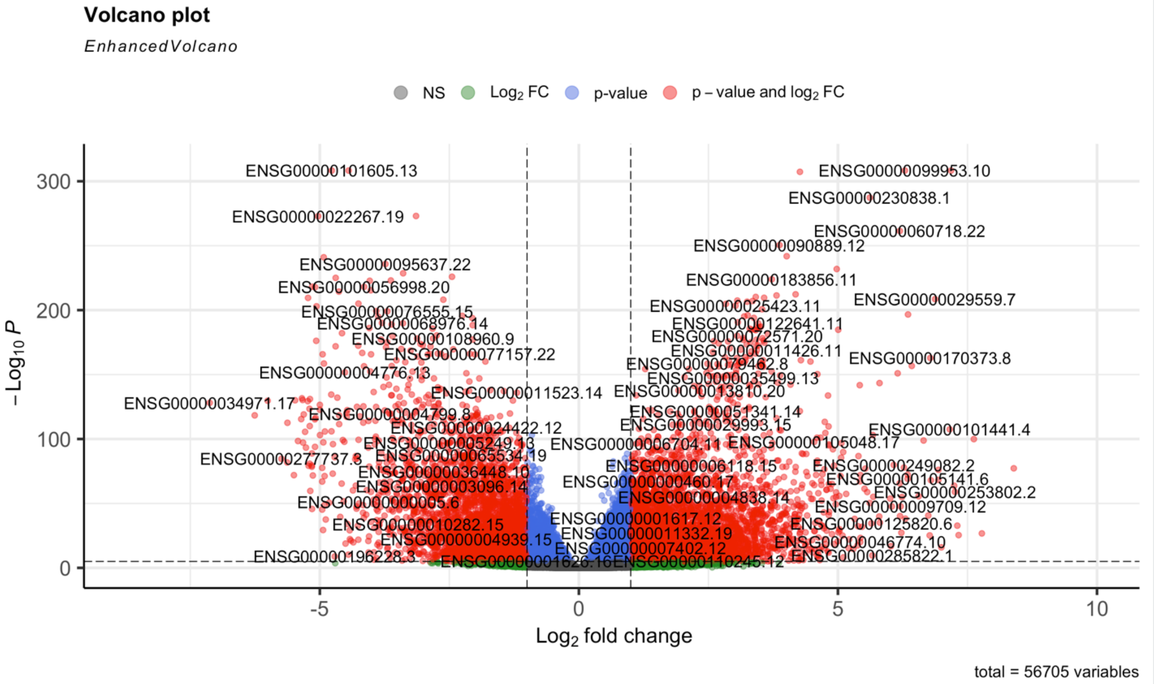


Nonmetastatic:



This plot was a somatic interactions plot, which provides information on gene occurrence in context of other genes. Both axes represent commonly mutated genes, so the plot is essentially n commonly mutated genes vs n commonly mutated genes. The color ranges from genes being mutually exclusive to co-occurring. The opacity is strength of association between gene occurrences. The * represents a p value < 0.05, or a statistically significant relationship between genes. This plot helps find which genes to look at in relation to each other. A conclusion that can be drawn is that there is a strong co-occurrence between MUC16 and TTN in both metastatic and non-metastatic patients due to the >3 co-occurrence color and * between MUC16 and TTN. A conclusion that cannot be drawn is that BRAF and TTN have a significant co-occurrence in both metastatic and non-metastatic cancers since a correlation is only found in metastatic cancers. Within a study correlating MUC4, MUC16, and TTN genes with prognosis and immunotherapy efficacy in gastric cancer, for study's TCGA cohort, the area under ROC curve (AUROC) reached the highest of 0.936 when combined three genes together. However, in the study's FUSCC cohort, AUROC reached the highest of 0.925 for *MUC16* and. This supports the conclusion that MUC16 and TTN have a statistically significant co-occurrence.

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



The plot is a volcano plot, which represents the differential expression of genes. The x-axis represents the fold change, and the y-axis represents p-value. Genes on the top portion of the plot represent significant genes, genes on the bottom portion represent insignificant genes. Genes on the left portion of the graph (negative portion of fold change axis) represent downregulated genes. Genes on the right portion of the graph (positive portion of fold change axis) represent upregulated tumor genes. Within this plot, it can be concluded that ENSG00000099953.10 or gene MMP11 is found to be significantly upregulated in both metastatic and non-metastatic patients. It cannot be concluded that ENSG00000000005.6 or gene TNMD is also significantly upregulated in both metastatic non-metastatic patients due to the negative fold change value and low p-value. Within the discussion section of a study relating MMP11 to proliferation and progression of breast cancer via Smad2 protein interaction, it was concluded there is ample evidence that MMP11 is a regulator of gene expression and complex pathways in cancer, specifically through upregulation in multiple pathways in tumor malignancy (cell proliferation, also overexpressed in protein and mRNA levels).

References:

Davies, H., Bignell, G., Cox, C. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* 417, 949–954 (2002). https://doi.org/10.1038/nature00766

"Metastatic Cancer: When Cancer Spreads" NIH: National Cancer Institute, National Institute of Health, U.S. Department of Health and Human Services. (2020). https://www.cancer.gov/types/metastatic-cancer

Yang, Wenjuan et al. "*DNAH7* mutations benefit colorectal cancer patients receiving immune checkpoint inhibitors." *Annals of translational medicine* vol. 10,24 (2022): 1335. doi:10.21037/atm-22-6166

Yang, Yue et al. "MUC4, MUC16, and TTN genes mutation correlated with prognosis, and predicted tumor mutation burden and immunotherapy efficacy in gastric cancer and pan-cancer." *Clinical and translational medicine* vol. 10,4 (2020): e155. doi:10.1002/ctm2.155

Zhuang, Ying et al. "MMP11 promotes the proliferation and progression of breast cancer through stabilizing Smad2 protein." *Oncology reports* vol. 45,4 (2021): 16. doi:10.3892/or.2021.7967