Part 1

General Concepts

1. The Cancer Genome Atlas or TCGA is a cancer genomics program started by the National Cancer Institute and National Human Genome Research Institute. The program collects cancer patient data and standardized and deidentifies it such that it can be placed in a publicly available data set that contains 20,000 samples spanning 33 different types of cancers. TCGA contains various types of information such as molecular mutation information to clinical data on each sample. It also covers various areas from genomics, to gene expression in transcriptomics, and more. Resultantly, TCGA is important for several reasons. Firstly, TCGA provides an incredibly rich data set for researchers to use to better understand various cancers. Next, this has led to more effective treatments for cancers and resulted in different approaches with how patients are treated in the clinic. The TCGA also allows for constant improvement within health and science technologies as well.

2. TCGA has several strengths. TCGA is a public dataset meaning that everyone has access to it and can use it without needing any special type of permissions. This makes it convenient and easy to utilize. The TCGA is also a large data set containing a large set of patients and samples spanning many different types of cancers and patient types. TCGA also has several weaknesses compared to a private dataset or other datasets. A private dataset usually has much better funding which also results in more cleaning done to the data which leads to better quality available. Many times, TCGA also lacks long term follow up data or lacks an overall response in the data.

Coding Skills

1. To save to my github, I begin by navigating to the folder that I have connected to Git in my terminal using the “cd” command with the path (e.g. cd /Users/kevinliu/desktop/qbio_490/QBIO_490_kevinliu). Next, I will use “git add” followed by the filename that I want to add. Then, I will use “git commit -m [message about upload]” before lastly using “git push”.

2. First, I must have the package downloaded. To do so, I can use “install.packages('[package name]’)”. Next, I will use “library([package name])” so that I will be able to access and use the package within R.

3. To use Bioconductor, I will first use “if (!require("BiocManager", quietly = TRUE))”, then “install.packages("BiocManager")”, before lastly using “BiocManager::install(version = "3.17")”.

4. Boolean indexing is where we create an index of true and false based on some vector or statement that we specify. This will assign each term that we pass through a true or false index value based on the order it was passed through. This becomes useful in boolean masking where this mask or list of trues and falses can be used to generate a new list or data frame that only exclude the falses or vice versa.

5. #We can begin by building a simple data frame.

   example <- data.frame(color = c("blue", "green", "yellow", "red",NA), number = c("1","2","54","65","65"))

   #The next step we will use an ifelse() statement to create a mask, the mask will assign rows that have 65 as true and the rest false.

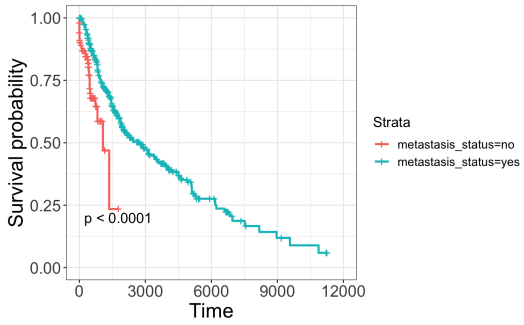
   number_mask <- ifelse(example$number == "65", T, F)

   #Lastly, we can use our mask and apply its values to our original dataframe to create a new one containing only rows with the number 65 and the same data in the other column. This process of matching based on our mask is boolean indexing.

   example <- example[number_mask,]

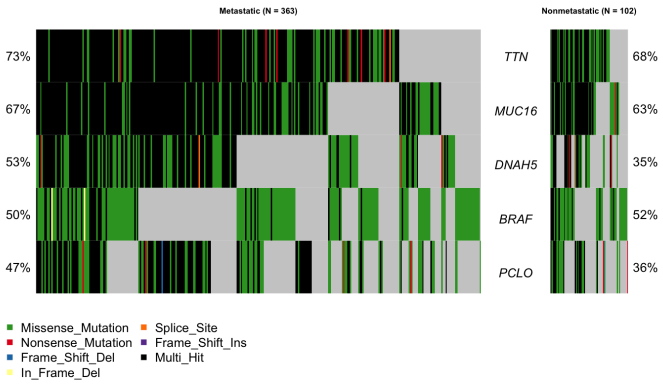
Part 3

1.



This is a Kaplan-Meier survival plot. It shows the likelihood of a patient surviving overtime depending on if the patient exhibits metastasis or not. We can conclude that TCGA SKCM patients with metastasis are less likely to survive at 1500 days or less compared to non metastatic patients since the metastatic curve lies below the non metastatic curve. What we cannot draw conclusions for is any survival probability for metastatic patients compared to non metastatic patients after around 1600 days since we no longer have any data for metastatic patients past that time. Our results are supported by other research studies which show a similar plot. In another study on TCGA SKCM patients with a focus on metastasis, high risk and low risk groups were used instead to plot, which also resulted in similar features on the plot (Huang et al., 2022). Their low risk featured a similar long survival time with decreasing survival probability while the high risk exhibited a steep drop in survival in a short time with no further data after. This makes sense as having metastasis is definitely higher risk of death for patients and having it also likely decreases survival time meaning we can infer that not many patients make it past where the plot data ends.

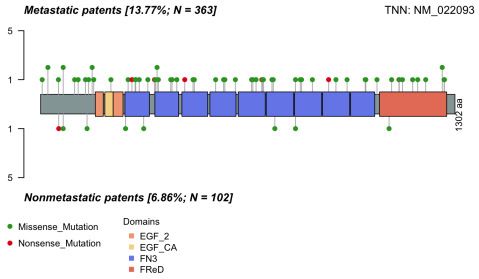
2.



This is a co-oncoplot. It shows the percentage of mutations and types for the same genes but in metastatic patients versus non metastatic patients. The plot here shows the highest mutated genes for the two. From our plot, we can conclude that there is a high number of Multi Hit mutations in the gene MUC16 in both sets of patients. This is shown

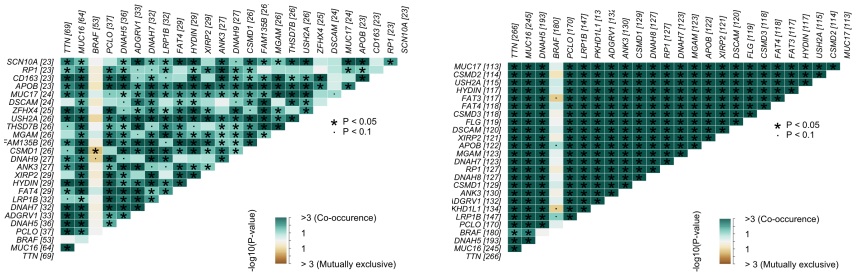
by the black bars that are prevalent for both groups. Looking at the differences in percentages of mutations, although it is higher for metastatic patients in MUC16 we cannot conclude that the higher mutation percentage is caused by the metastasis of the cancer. We don't have that type of evidence and can only infer that. Looking at other literature regarding SKCM, it was shown that the genes show a high number of overall mutations compared to other cancers (Guan et al., 2015). This further brings into question whether the high mutation rates we see for metastasis patients can be correlated to the spread of cancer or not.

3.



This is a co-lollipop plot. For a chosen gene, the plot shows the mutation counts at various locations on the gene. The boxes represent the domains of the protein and the axis gives the location of the amino acids. Top and bottom directions represent the different groups being compared. From our plot, we can conclude that for the gene TNN, metastasis patients show a greater frequency of mutations, especially in the domains like FReD. This is shown by the green dots connected by lines pointed up compared to the number pointed down. What we cannot conclude is that metastatic patients are more mutated than non metastatic patients. This is because there are less non metastatic patients with mutations here than metastatic so frequency is not enough to provide evidence for that conclusion. Our results can be supported by other research as well. Our plot shows an abundance of missense mutations along the whole gene which is supported by plots in other papers that also show an overwhelming amount of missense mutations for the gene TTN (Oh et al., 2020).

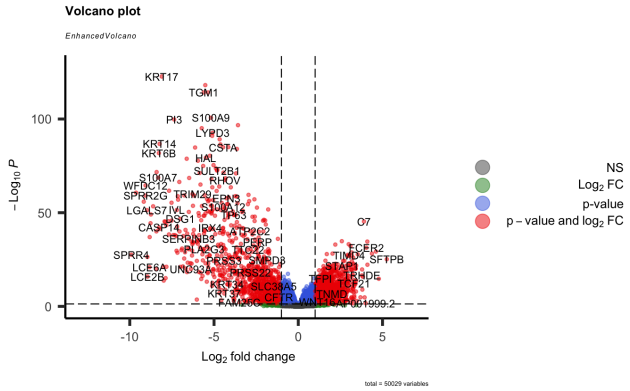
4.



Both of these plots are somatic interactions plot. The one on the left is for non metastatic patients while the one on the right is for metastatic patients. Both show how likely a gene is to mutate with another gene in their respective groups. From our plot, we can conclude that in non metastatic patients, CSMD1 is mutually exclusive from BRAF in that

the two gene mutations are not correlated with each other. This is shown by the significant star in the brown box. Although there are many more concurrent mutations in metastatic patients, we cannot conclude that metastatic patients mutate more than non metastatic patients. This is because we are measuring the likelihood of mutations in genes occurring together and not the frequency of mutations. Looking at other research studies, it is also supported that the BRAF gene stands out in mutations for SKCM cancer. BRAF dominates the mutations in metastatic melanoma (Guan et al. 2015). Potentially, BRAF does not depend on the mutations of other genes to occur which allows it to mutate more freely.

5.



This is a volcano plot. The plot shows the differential expression of genes between metastatic and non metastatic patients when controlling for treatments (chemotherapy, immunotherapy, molecular therapy, and vaccine), race, gender, and vital status. We are using the threshold of 0.05 for significance and log2FoldChange at |1| to determine if a gene is up or down regulated. Looking at our plot, we can conclude that KRT17 is significantly downregulated in metastatic patients compared to non metastatic patients. This is because it lies in the upper left region of our plot. Looking at our plot, we cannot conclude that any gene is more important than another (e.g. KRT17 is more important than SPRR4). KRT17 is more statistically significant while SPRR4 is more downregulated. None of these measures provide evidence that one gene has a greater impact than another. Looking at other literature, our volcano plots compare similarly. In another paper, a volcano plot was generated for metastatic and non metastatic patients, but not controlling for the same covariates. Similarly, the plot showed few nonsignificant genes. However, there were many more upregulated genes that were more significant as well as more significant genes that were neither up or down regulated (Jia et al., 2021). This could potentially be explained by our inclusion of covariates. Our covariates could have had a n association that led to these trends that we did not observe.

References

About the program. ccg - National Cancer Institute. (n.d.). https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history#:~:text=The%20Cancer%20Genome%20Atlas%20(TCGA,biology%2C%20and%20other%20research%20fields.

Guan, J., Gupta, R., & Filipp, F. V. (2015). Cancer Systems Biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep07857

Huang, R., Li, M., Zeng, Z., Zhang, J., Song, D., Hu, P., Yan, P., Xian, S., Zhu, X., Chang, Z., Zhang, J., Guo, J., Yin, H., Meng, T., & Huang, Z. (2022). The identification of prognostic and metastatic alternative splicing in skin cutaneous melanoma. *Cancer Control*, *29*, 107327482110515. https://doi.org/10.1177/10732748211051554

Jia, G., Song, Z., Xu, Z., Tao, Y., Wu, Y., & Wan, X. (2021). Screening of gene markers related to the prognosis of metastatic skin cutaneous melanoma based on logit regression and survival analysis. *BMC Medical Genomics*, *14*(1). https://doi.org/10.1186/s12920-021-00923-0

Oh, J.-H., Jang, S. J., Kim, J., Sohn, I., Lee, J.-Y., Cho, E. J., Chun, S.-M., & Sung, C. O. (2020). Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *Npj Genomic Medicine*, *5*(1). https://doi.org/10.1038/s41525-019-0107-6

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Współczesna Onkologia*, *1A*, 68–77. https://doi.org/10.5114/wo.2014.47136