Part 1: Review Questions

General Concepts

1. What is TCGA and why is it important?

TCGA is The Cancer Genome Atlas, a cancer genomics program that characterizes over 20,000 primary and matched normal samples. TCGA is important because it provides rich genomics data to researchers and scientists. Research using TCGA has deepened our understanding of cancer, advanced health and science technologies, and transformed cancer treatments.

2. What are some strengths and weaknesses of TCGA?

Strengths of TCGA are that it provides large, comprehensive cancer data and is an open access resource. Weaknesses of TCGA are that it lacks clinical diversity and that there is limited diversity and lifestyle data.

Coding Skills

1. What commands are used to save a file to your GitHub repository?

- Open Ubuntu/terminal and navigate to wherever qbio_490_name is located on your computer.
- Type "git status" to verify that you have untracked changes (this means you have modified the file i.e. you finished your homework)
- Use "git add filename" to stage files to push to GitHub (you can use "git add ." to stage all files)
- Type "git commit -m message" and replace "message" with something more descriptive (e.g. "week3_homework")
- Finally, enter "git push" and type in your password (if you made one) to push all of the changes to GitHub

2. What command(s) must be run in order to use a package in R?

```
install.packages("name")
library(name)
```

3. What command(s) must be run in order to use a Bioconductor package in R?

```
install.packages("BiocManager")
BiocManager::install("name")
library(name)
```

4. What is boolean indexing? What are some applications of it?

Boolean indexing uses boolean values to filter data. Elements where the condition is True are selected. For example, boolean indexing can be used to filter by age.

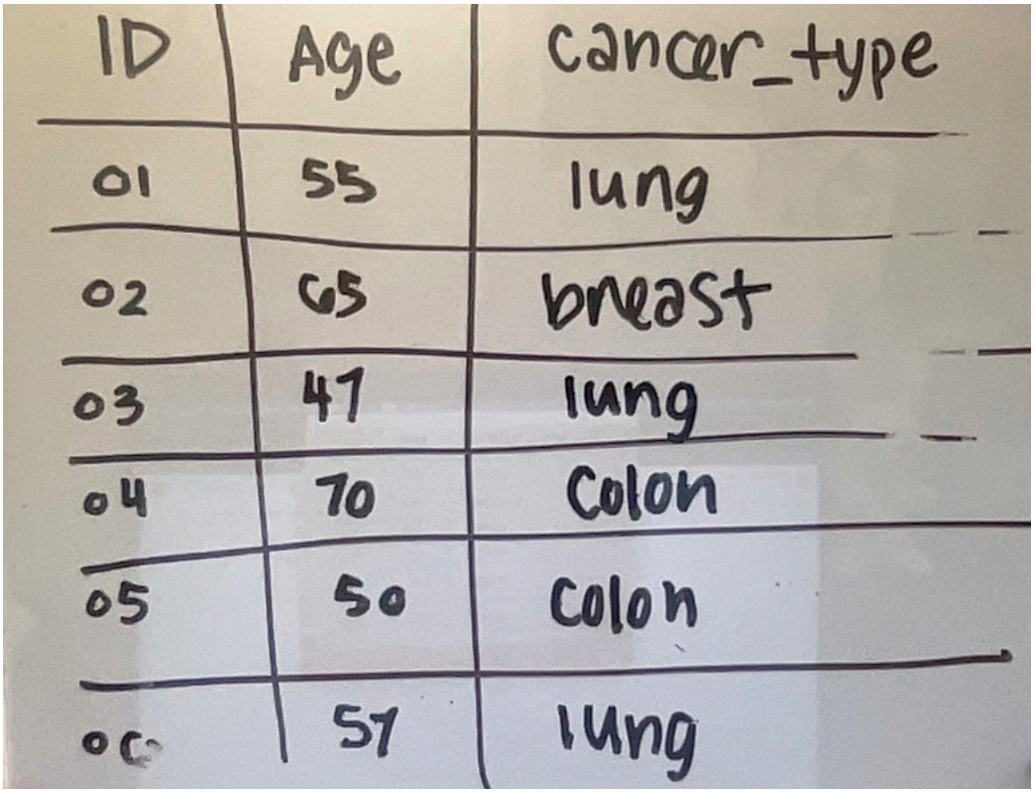
5. Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.
   a. an ifelse() statement

```
clinical$age <- ifelse(clinical$age_at_diagnosis <= 35, "Young",
                       ifelse(clinical$age_at_diagnosis >= 50, "Old",
                              "Middle"))
```

#creates a new column that categorizes patients as young if age at diagnosis is less than or equal to 35, old if age at diagnosis is greater than or equal to 50, and middle otherwise.

   b. boolean indexing

```
old_lung_cancer <- clinical[clinical$age == "Old" & clinical$cancer_type == "Lung", ]
# Select rows where age is "Old" and cancer type is "Lung"
```

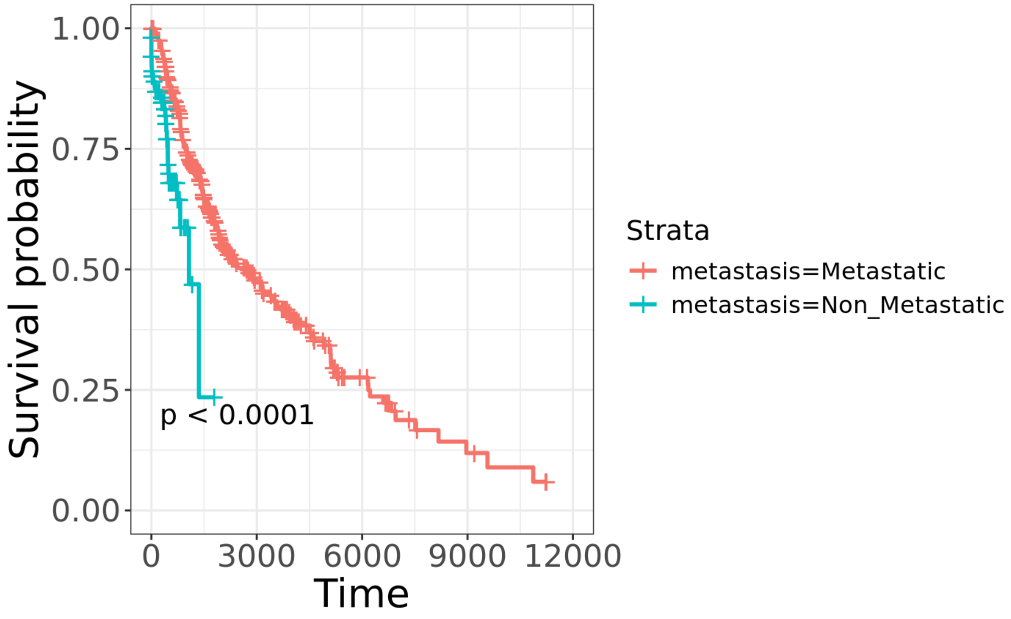


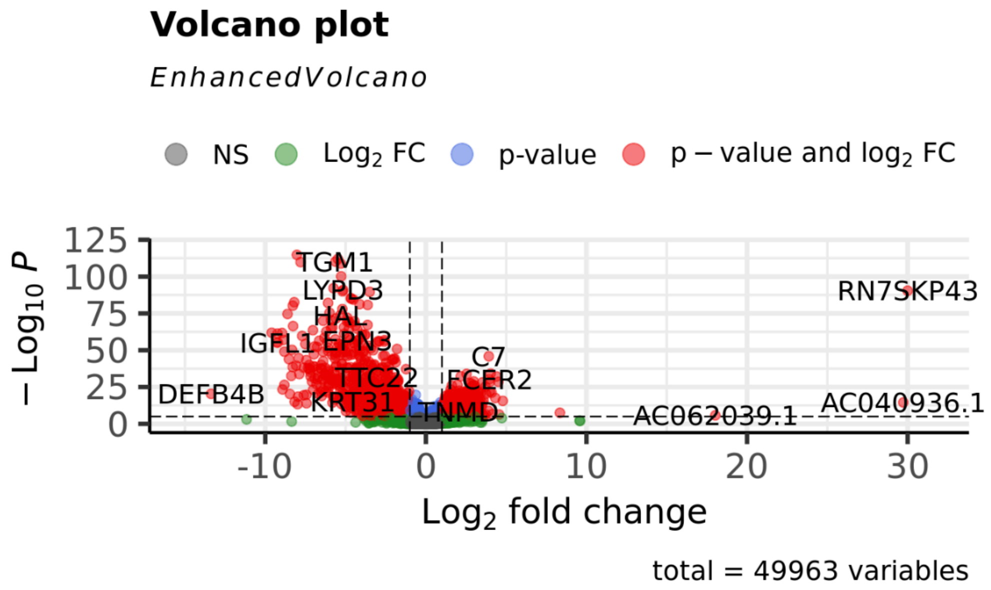
ID	Age	cancer_type
01	55	lung
02	65	breast
03	47	lung
04	70	Colon
05	50	Colon
06	57	lung

Part 3: Results and Interpretations

1. Difference in survival between metastatic and non-metastatic patients
   The KM plot indicates that metastatic patients exhibit higher survival probability than non-metastatic patients ($p < 0.0001$).
2. Expression differences between metastatic and non-metastatic patients
   Many significantly downregulated genes in metastatic patients (e.g., *TGM1*, *LYPD3*, *HAL*, *IGFL1*, *EPN3*, *DEFB4B*, *TTC22*, *KRT31*). These genes might be involved in pathways suppressed during metastasis or as a result of systemic changes in the tumor microenvironment. Fewer significantly upregulated genes in metastatic patients (e.g., *C7*, *FCER2*, *RN7SKP43*, *AC04936*). These genes might play roles in metastatic progression or adaptations to metastatic niches. Few insignificant changes, indicating a generally strong signal of differential expression.
3. Methylation differences between metastatic and non-metastatic patients
   The methylation volcano plot mirrors expression differences, suggesting potential epigenetic regulation of transcription in metastatic versus non-metastatic tumors. Genes with notable methylation differences should be cross-referenced with transcriptional data to identify correlations.
4. Direct comparison of transcriptional activity to methylation status for 10 genes
   Direct comparisons for 10 genes would elucidate whether methylation differences correlate with transcriptional regulation. Hypomethylation in promoters often increases gene expression, while hypermethylation silences it.

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.





References

Gosman, L. M., Țăpoi, D.-A., & Costache, M. (2023). Cutaneous Melanoma: A Review of Multifactorial Pathogenesis, Immunohistochemistry, and Emerging Biomarkers for Early Detection and Management. *International Journal of Molecular Sciences*, *24*(21), 15881. https://doi.org/10.3390/ijms242115881

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), 7857. https://doi.org/10.1038/srep07857