Foundations of Data Science Final Project

Samira Rahman

October 2025

1 Introduction

This analysis explores RNA-seq gene expression data from The Cancer Genome Atlas (TCGA), focusing on understanding expression variability across different sample types, including metastatic, primary tumor, and solid tissue normal samples. The dataset contains normalized count data for hundreds of genes across multiple patient samples.

For this project, two genes were selected for deeper analysis:

- ***ENSG00000000419.13*** **(Gene1)**: a gene associated with cellular metabolic and structural functions.
- ***ENSG00000000460.17*** **(Gene2)**: another gene showing moderate to high expression variability across tumor samples.

Studying these genes helps illustrate how expression differs across cancer sample types, providing biological insights into transcriptional changes associated with tumor progression and metastasis.

2 Methods

All analyses were performed using **R version 4.5.1**. The following R packages were used:

- **ggplot2 (v4.0.0)**: for generating histograms, scatter plots, violin plots, and bar charts.
- **ComplexHeatmap (v2.24.1)**: to create annotated heatmaps for multi-gene visualization.
- **circlize (v0.4.16)**: as a dependency for ComplexHeatmap color mapping.
- **xtable (v1.8.4)**: LaTeX compatible output formatting.

The RNA-seq **count matrix** and associated **metadata file** were read into R. Counts were extracted for two specific genes, converted to numeric vectors, and combined with the metadata covariate *sample_type*. Visualizations were saved as PNG files. Summary statistics (mean, median, standard deviation, minimum, and maximum) were calculated for both selected genes and formatted as a LaTeX table for this document.

3 Results

A series of plots were generated to explore and visualize expression trends:

3.1 Histogram

This histogram shows the **distribution of expression counts** for the gene *ENSG0000000419.13* across all samples. Most samples have low to moderate expression levels, with a sharp peak around lower count values and a long right tail. This right-skewed distribution indicates that while most samples express this gene at low levels, a few exhibit very high expression, suggesting potential overexpression in certain tumor samples.

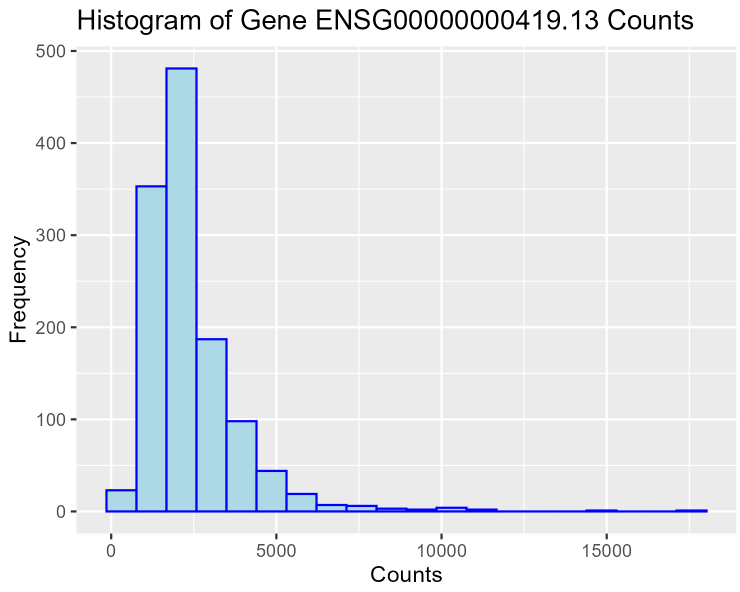

Figure 1: Histogram of Gene ENSG00000000419.13 Counts.

3.2 Scatter Plot

The scatter plot comparing *ENSG00000000419.13* and *ENSG00000000460.17* showed a positive correlation between their expression levels across samples, suggesting potential co-regulation or shared biological pathways. The regression line with confidence interval (gray shading) supports this trend.

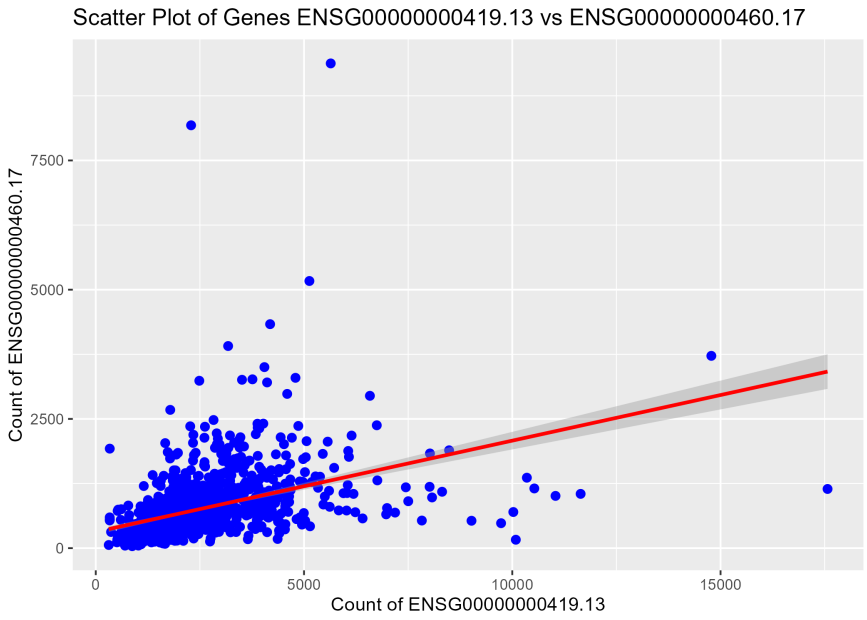

Figure 2: Scatter plot comparing expression of ENSG00000000419.13 and ENSG00000000460.17.

3.3 Violin Plot

The violin plot illustrated the distribution of *ENSG00000000419.13* expression across the three sample types. Expression of this gene appears to increase from normal to primary tumor tissues, suggesting possible upregulation in tumor development, followed by a decrease or stabilization in metastatic tissues. This pattern could imply that the gene is associated with tumor initiation or progression but not necessarily with metastasis itself.

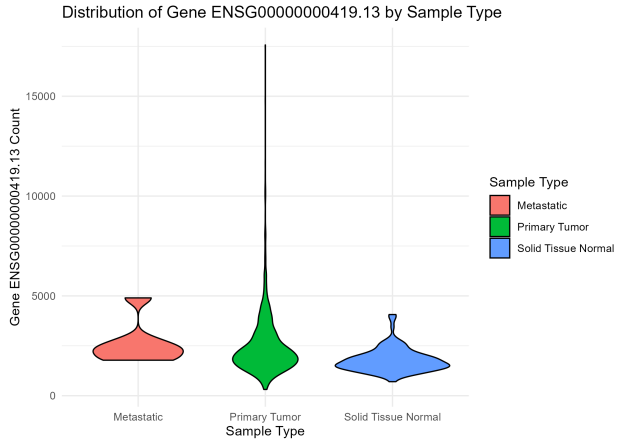

Figure 3: Distribution of Gene ENSG00000000419.13 by Sample Type.

3.4 Heatmap

The heatmap visualized expression levels of **ten selected genes** across all samples, revealing distinct clusters corresponding to **sample type** categories. The top annotation bar provided clear differentiation among Metastatic, Primary Tumor, and Normal samples. Each row represents a gene, and each column represents a sample. The color gradient (blue to red) indicates expression levels, where blue corresponds to low expression and red to high expression. Overall, the heatmap shows that Primary Tumor and Metastatic samples (orange and red bars) exhibit higher gene expression levels for several genes compared to Solid Tissue Normal samples (green bars), which remain mostly blue. Notably, genes like *ENSG00000000419.13* show distinct upregulation in tumor samples, suggesting potential involvement in tumor development or progression.

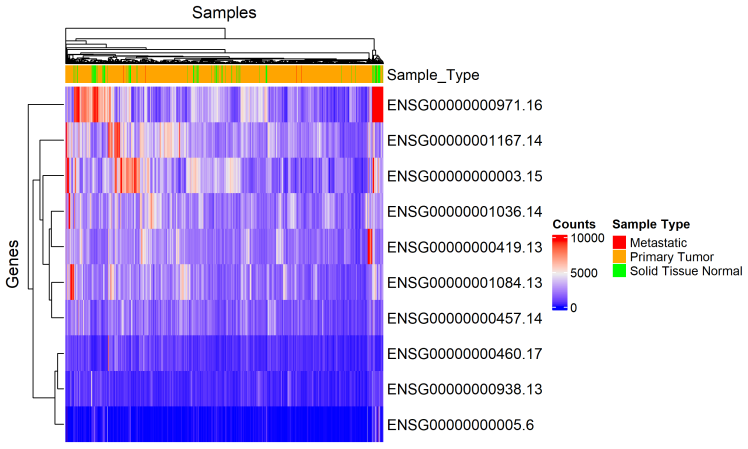

Figure 4: Heatmap of multiple gene expression profiles across sample types.

3.5 Bar Plot

The bar chart displays the average expression of *ENSG00000000419.13* by sample type. We see that **Metastatic** and **Primary Tumor** samples have **higher mean expression** compared to **Solid Tissue Normal**, indicating that this gene may be **upregulated in cancerous tissues** and could play a role in tumor development or progression.

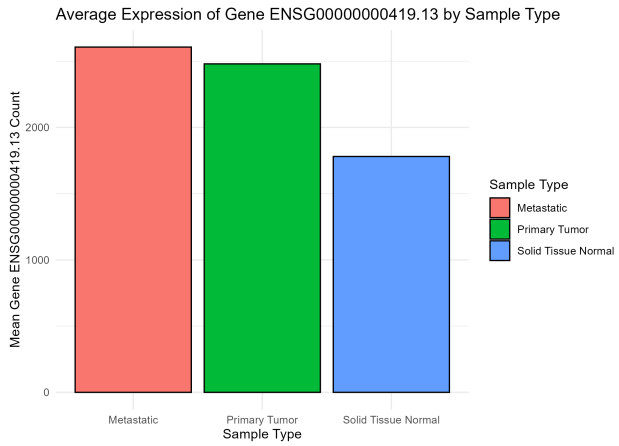

Figure 5: Average expression of Gene ENSG00000000419.13 by sample type.

3.6 Summary Statistics

Summary statistics for both genes showed higher variability and mean values for *ENSG00000000419.13* compared to *ENSG00000000460.17*, consistent with its broader expression range across samples.

	Statistics	ENSG00000000419.13	ENSG00000000460.17
1	Mean	2416.31	741.10
2	Median	2052.00	587.00
3	Standard Deviation	1459.49	627.80
4	Minimum	312.00	38.00
5	Maximun	17569.00	9377.00

Table 1: Summary Statistics of Gene ENSG00000000419.13 and ENSG00000000460.17

4 References

1. Gu, Z. circlize implements and enhances circular visualization in R. Bioinformatics 2014.
2. R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
3. The Cancer Genome Atlas (TCGA). Retrieved from `https://portal.gdc.cancer.gov/`
4. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.