Comparison between DESeq2 Normalization and RUVg Normalization

Maritza Morales

2024-12-10

Contents

Background Information

- Upon starting in the Duval Lab, I was given RNA seq data that had already need normalized with using two methods:
  - DESeq2
  - RUVg
- I was tasked with comparing the 2 normalization methods via the following methods:
  - Basic Statistics
    - Mean Gene Expression
    - Median Gene Expression
    - Max Gene Expression
  - Correlation Plots
    - Correlation plots were generated using GSVA score data for both normalization methods
- This task served two purposes:
  1. The lab had wanted to compare both methods, but due to the current project load when I started this project was put on the back burner. My arrival meant I could take some of those projects that had been deprioritized and work on them, helping to decrease the pending project work load.

  2. This project allowed me to become familiar and comfortable with utilizing R and RStudio. Prior to the start of the program I had extremely limited coding experience. I was very up front with

Dr. Duval about this when she interviewed me. This project was her way of allowing me the time I needed to became more familar with the skills I needed to be successful in the lab.

Cell Lines Per Cancer

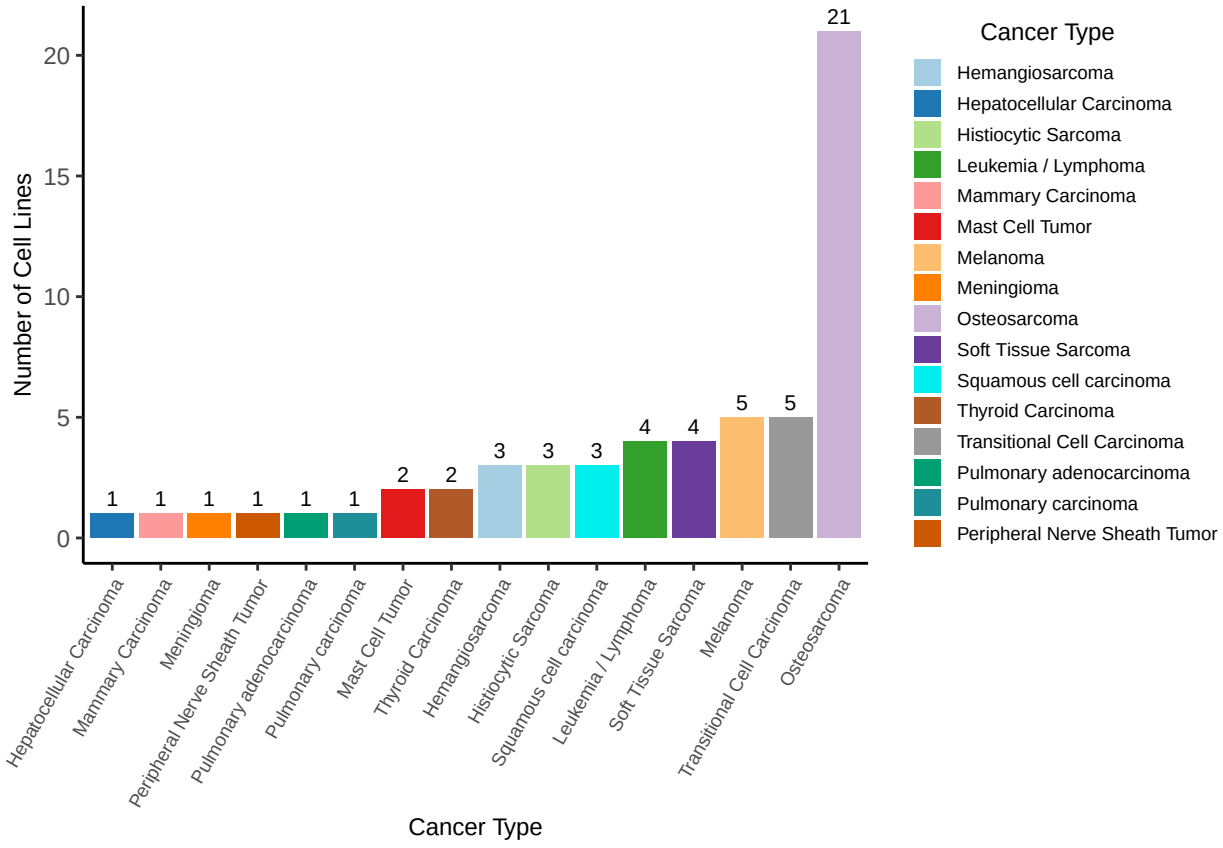

Figure 1: Bar plot illustrating the distribution of 58 cancer cell lines across 16 different cancer types. Osteosarcoma has the highest number of cell lines (21), highlighting its prevalence in cancer research. This plot can be used to identify cancer types with readily available cell lines for further study.

Correlation

Correlation Plot

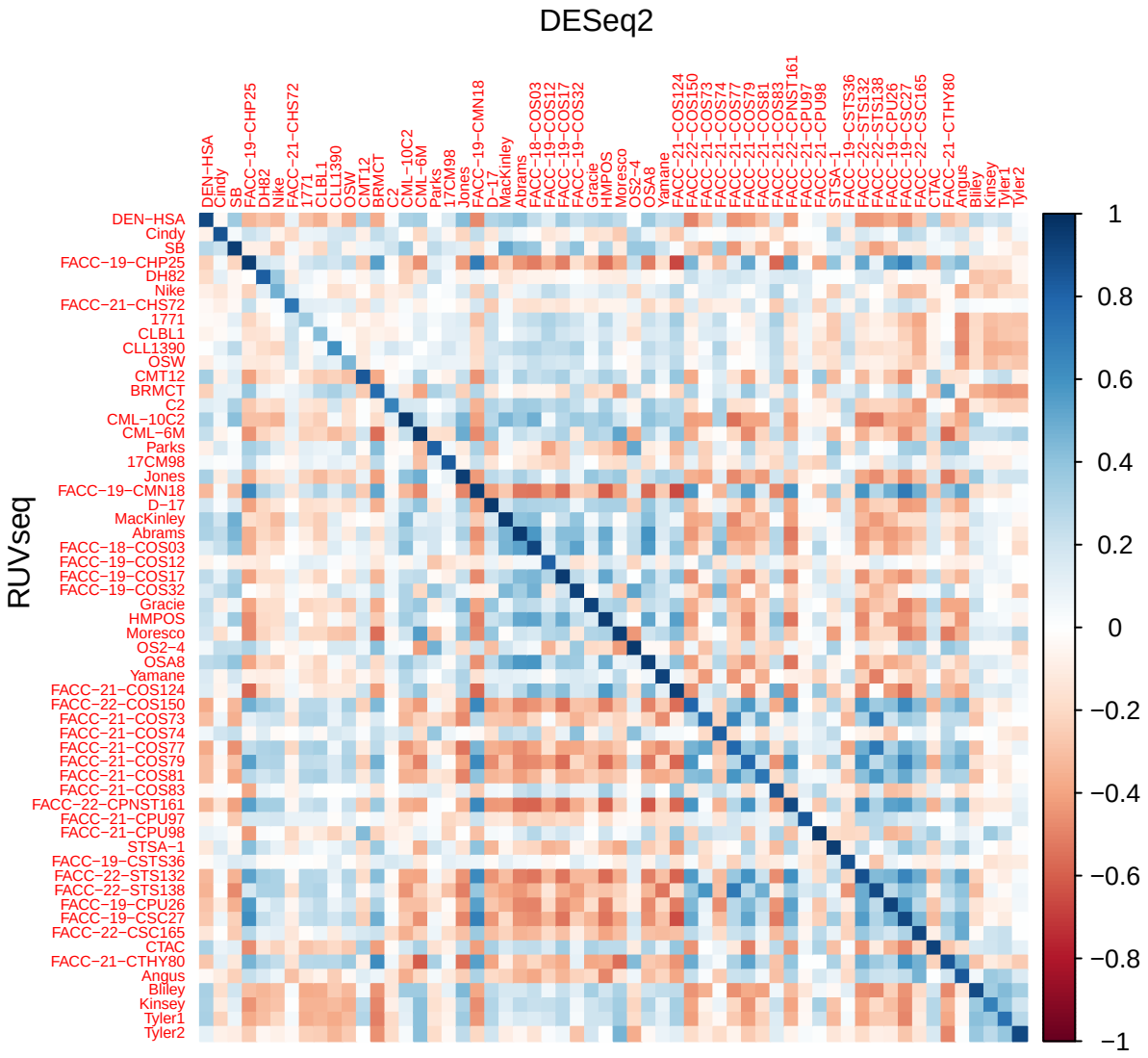

Figure 2: Heatmap visualization of the correlation coefficients between RUVg and DESeq2 normalization methods for gene set variation analysis (GSVA) data. Each cell in the matrix represents the correlation coefficient for a pair of cancer cell lines, with RUVg normalization on the y-axis and DESeq2 normalization on the x-axis. The color intensity reflects the strength and direction of the correlation (typically red for positive, blue for negative). This heatmap allows for a detailed examination of the relationships between the two normalization methods across all 58 cancer cell lines, potentially revealing systematic biases or trends depending on the cancer type.

Histogram

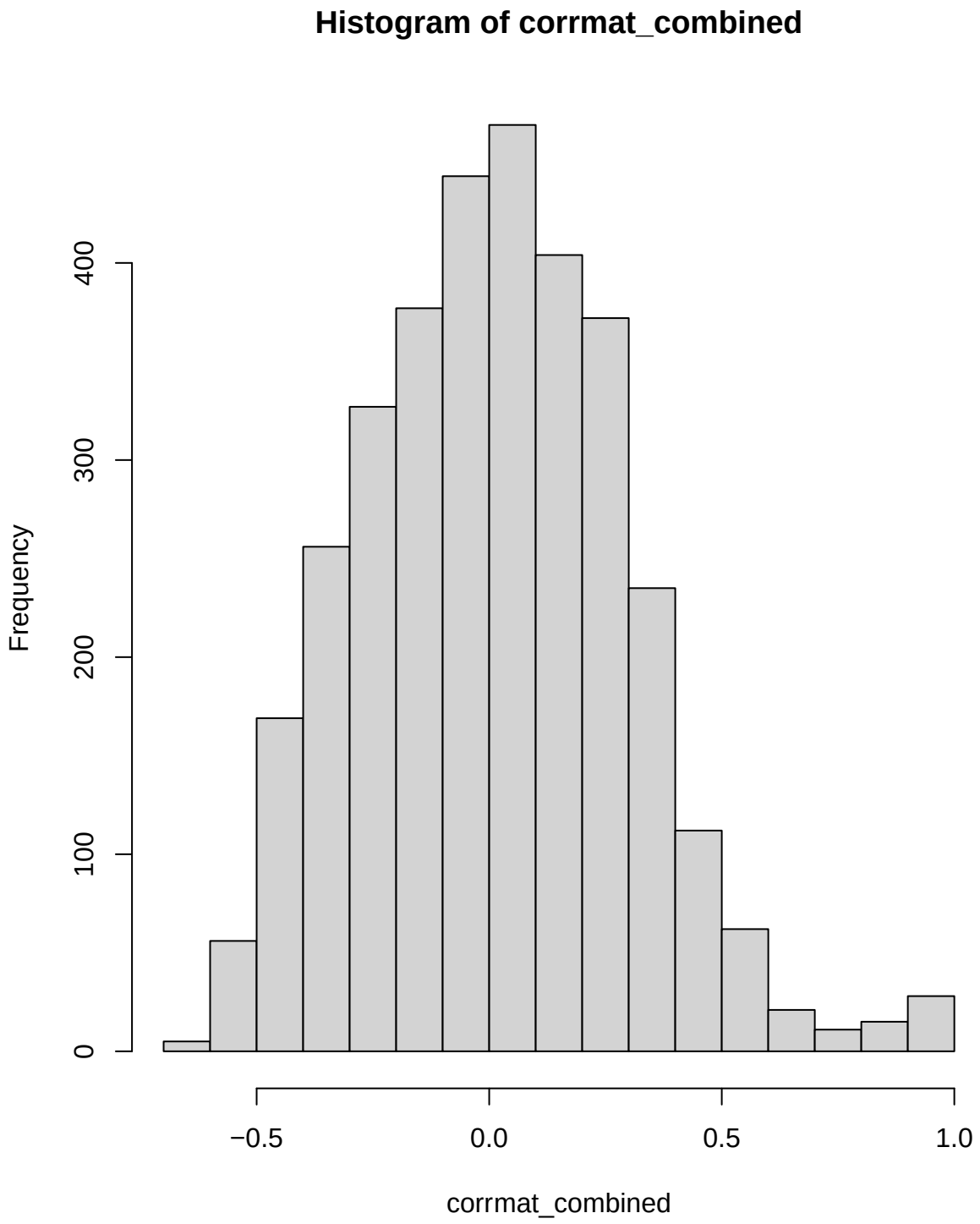

Figure 3: Histogram visualization comparing RUVg and DESeq2 normalization methods for gene set variation analysis (GSVA) data. The x-axis represents the correlation coefficient values, indicating the strength and direction of the association between the two normalization methods. The y-axis shows the frequency of each correlation value across all 58 cancer cell lines. This histogram allows for a comprehensive analysis of the distribution of correlation coefficients, highlighting potential biases or trends favoring one normalization method over the other for GSVA analysis in this cancer cell line dataset.

Correlation Bar Plot

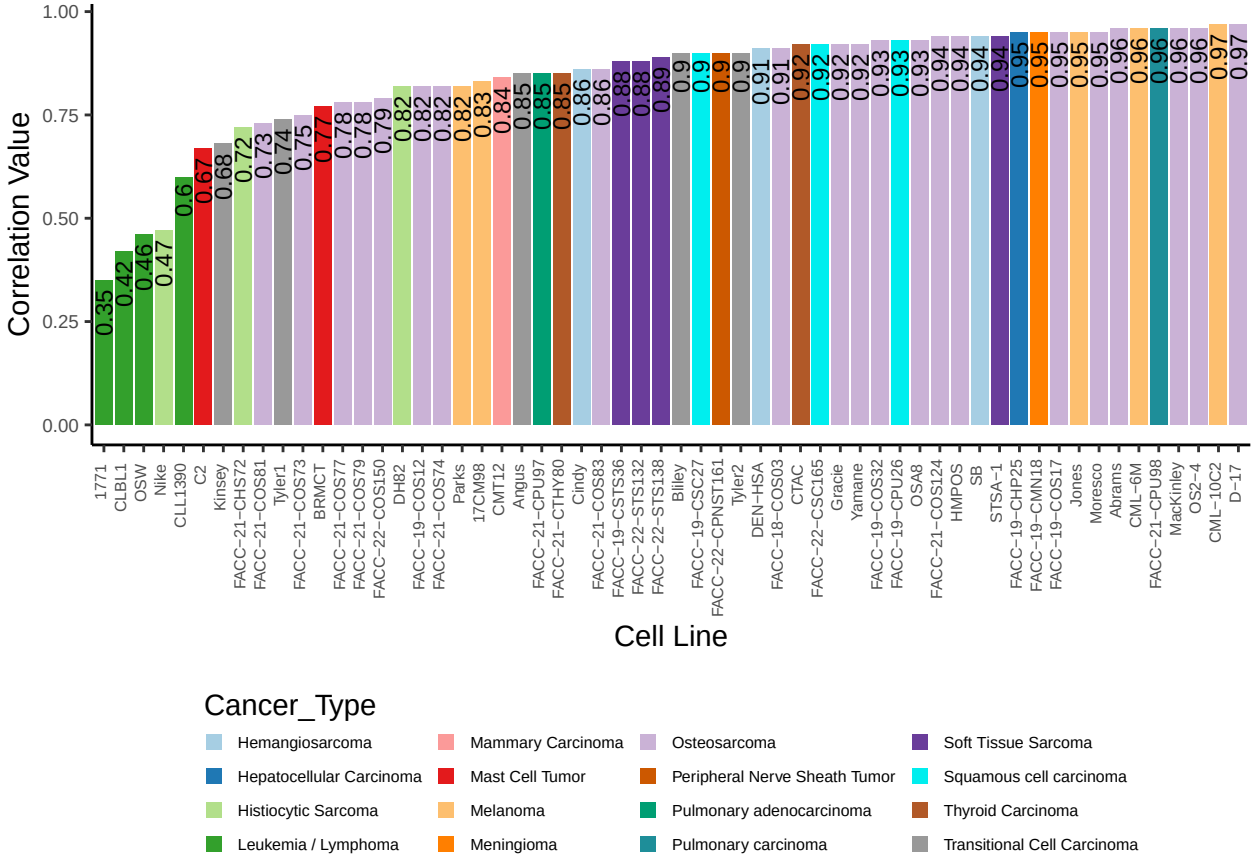

Figure 4: Bar plot visualizing the correlation coefficients between RUVseq and DESeq normalization methods for gene set variation analysis (GSVA) on 58 cancer cell lines. Each bar is colored according to the cancer type it represents, allowing for quick identification of trends between normalization methods and specific cancer types. This visualization facilitates the comparison of normalization techniques, highlighting which method yields more consistent results (higher correlation) for GSVA analysis within different cancer types.

DESeq2

PCA Plot

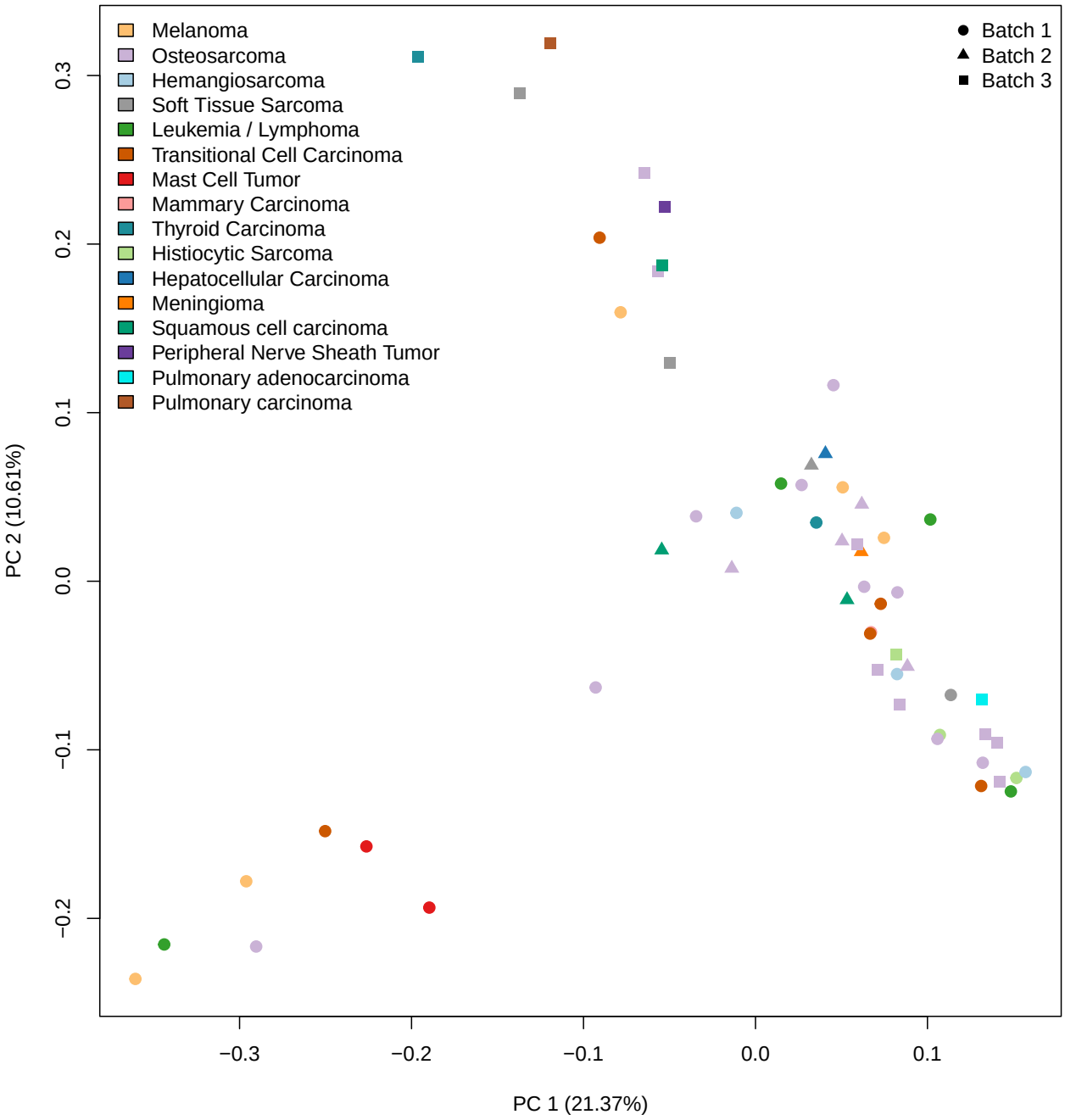

Figure 5: PCA plot showing the expression values of the DESeq normalized data organized by the cancer type.

Mean Box Plot

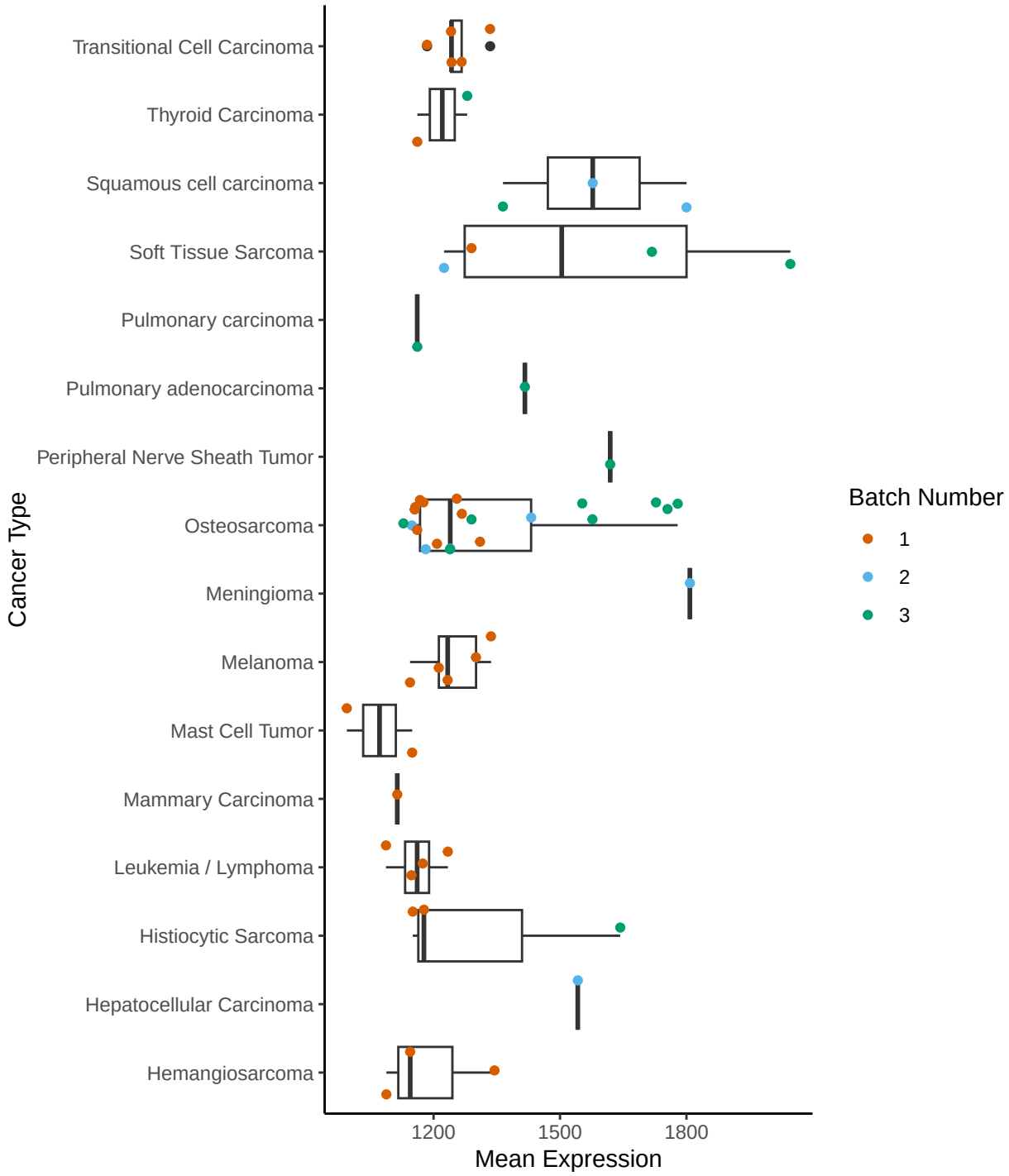

Figure 6: Box plot of the mean expression values calculated using the DESeq normalized data with a jitter. The jitter also illustrates the batch numbers that mean values were calculated from. Entire plot is organized by the cancer type.

Median Box Plot

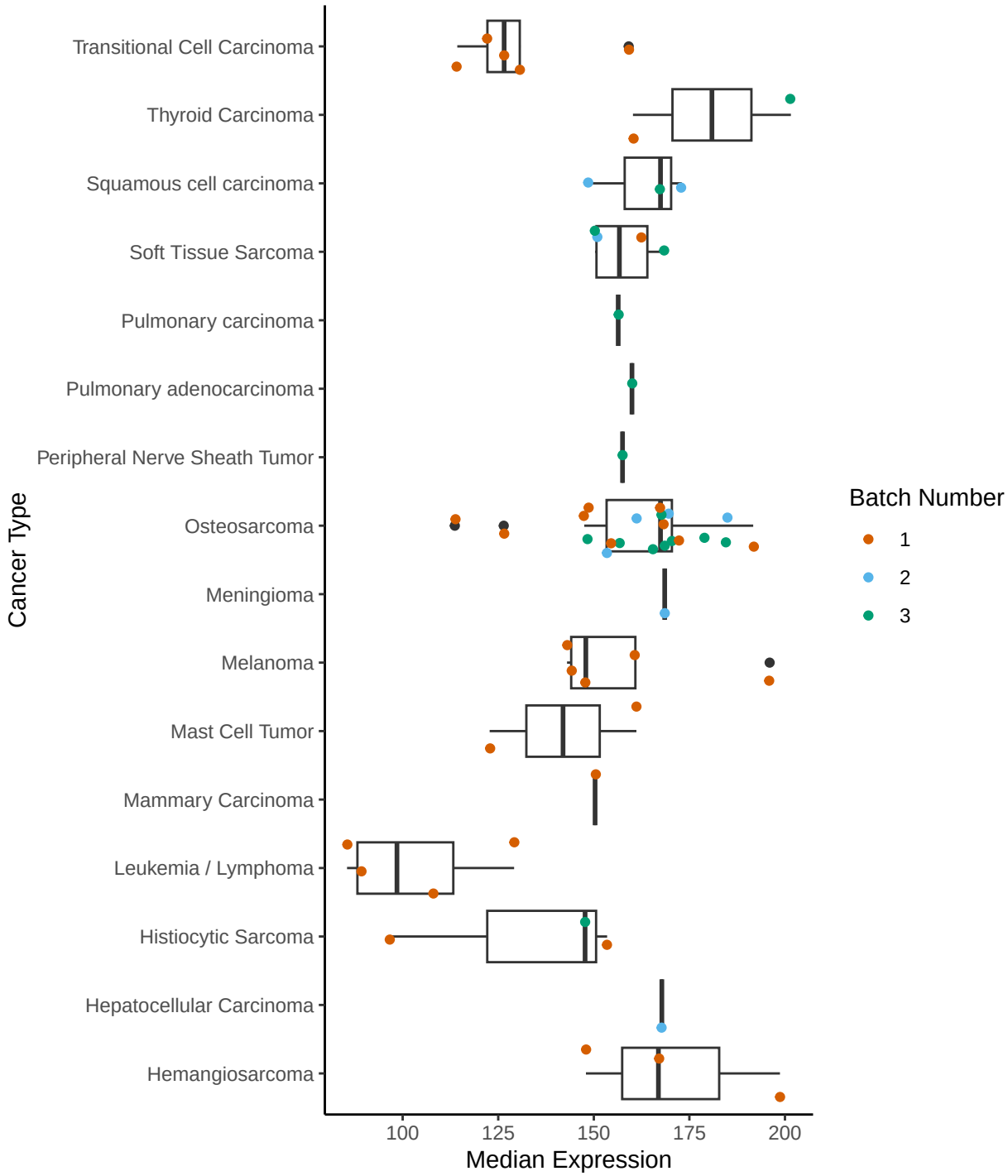

Figure 7: Box plot of the median expression values calculated using the DESeq normalized data with a jitter. The jitter also illustrates the batch numbers that median values were calculated from. Entire plot is organized by the cancer type.

Max Gene Expression

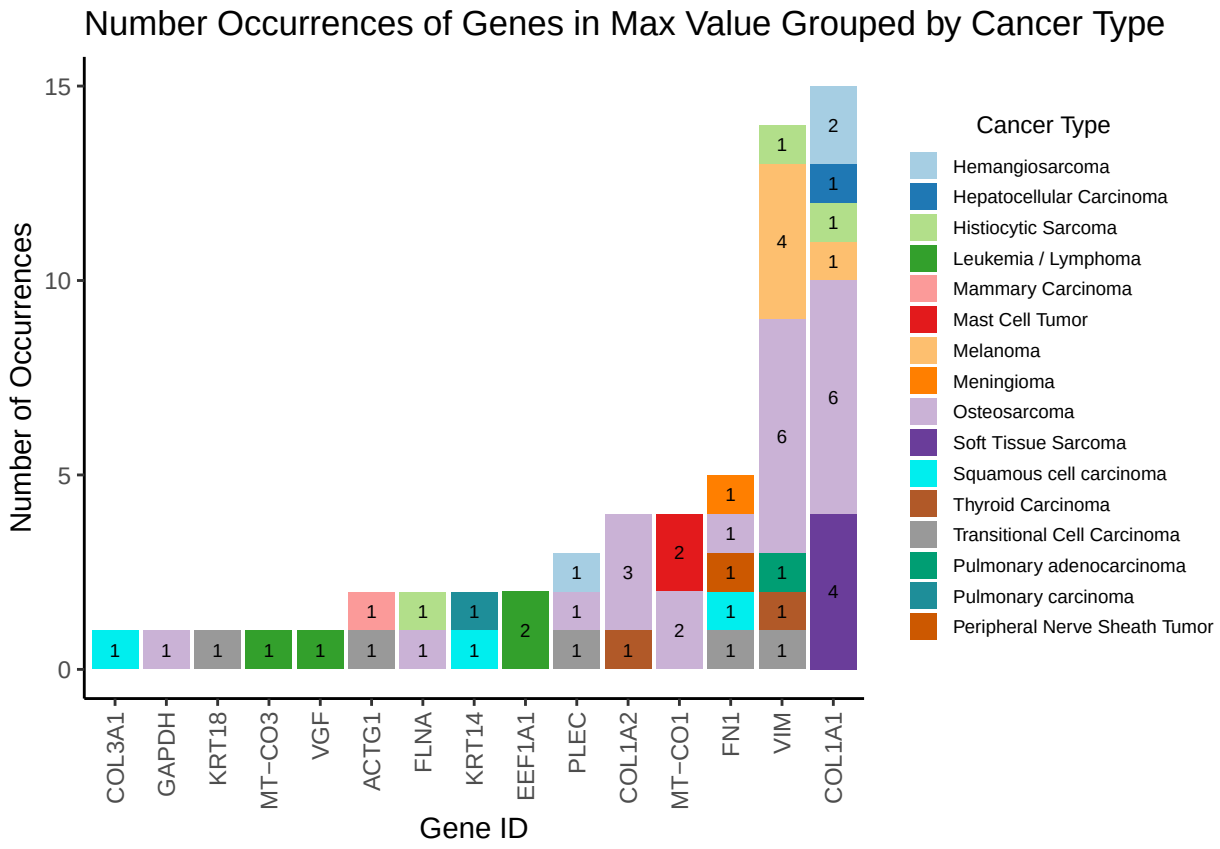

Figure 8: This bar graph visualizes the distribution of genes with the highest expression level across the 58 cell lines. While nearly 20,000 genes were analyzed, only 15 emerged as the most highly expressed gene in at least one cell line. Each bar represents a gene, and its color indicates the specific cancer(s) where that gene showed the highest expression. Refer to Appendix Table 19 for detailed descriptions of these 15 genes.

RUVg

PCA Plot

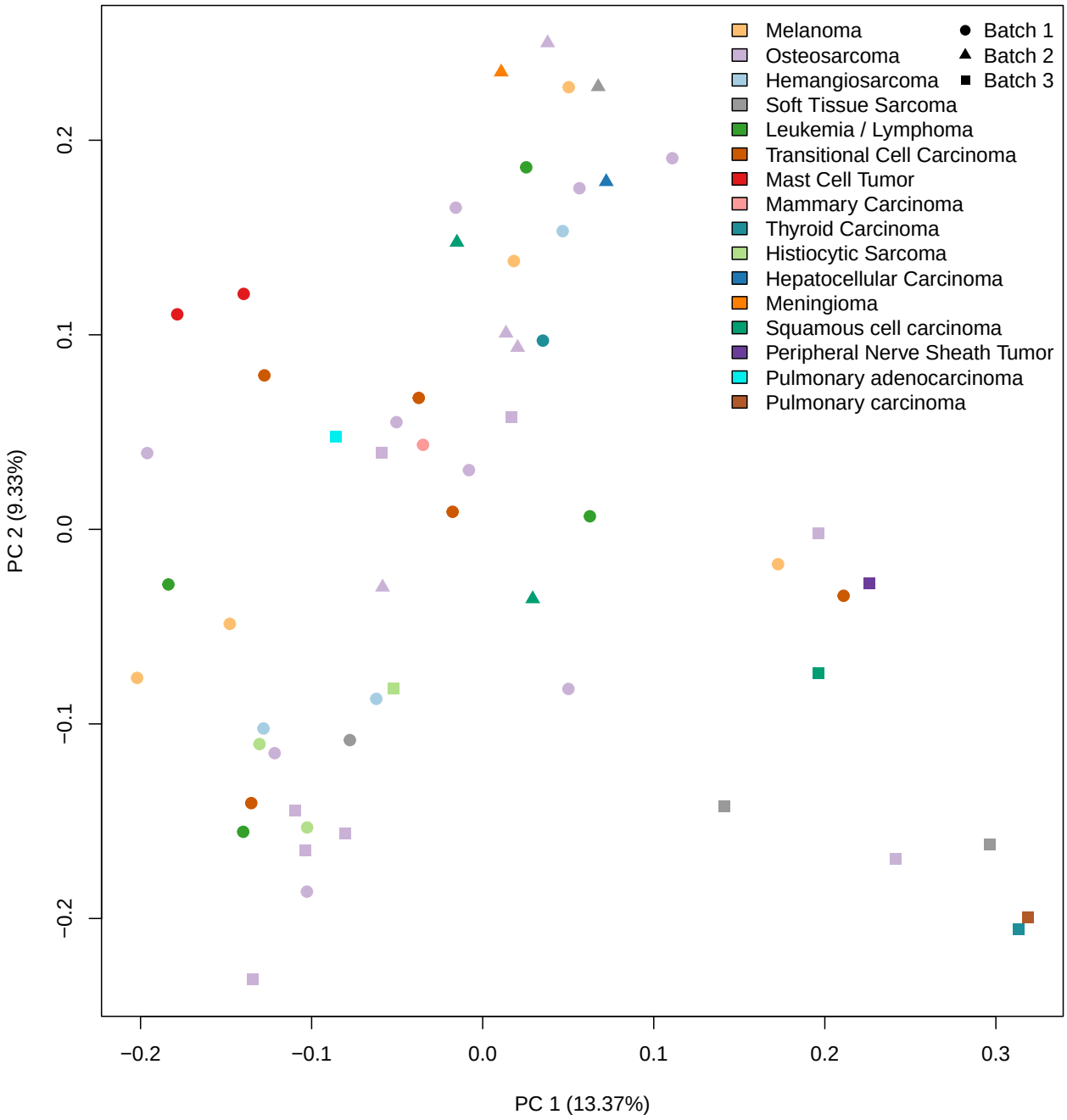

Figure 9: PCA plot showing the expression values of the RUVg normalized data organized by the cancer type.

Mean Box Plot

Figure 10: Box plot of the mean expression values calculated using the RUVg normalized data with a jitter. The jitter also illustrates the batch numbers that mean values were calculated from. Entire plot is organized by the cancer type.

Median Box Plot

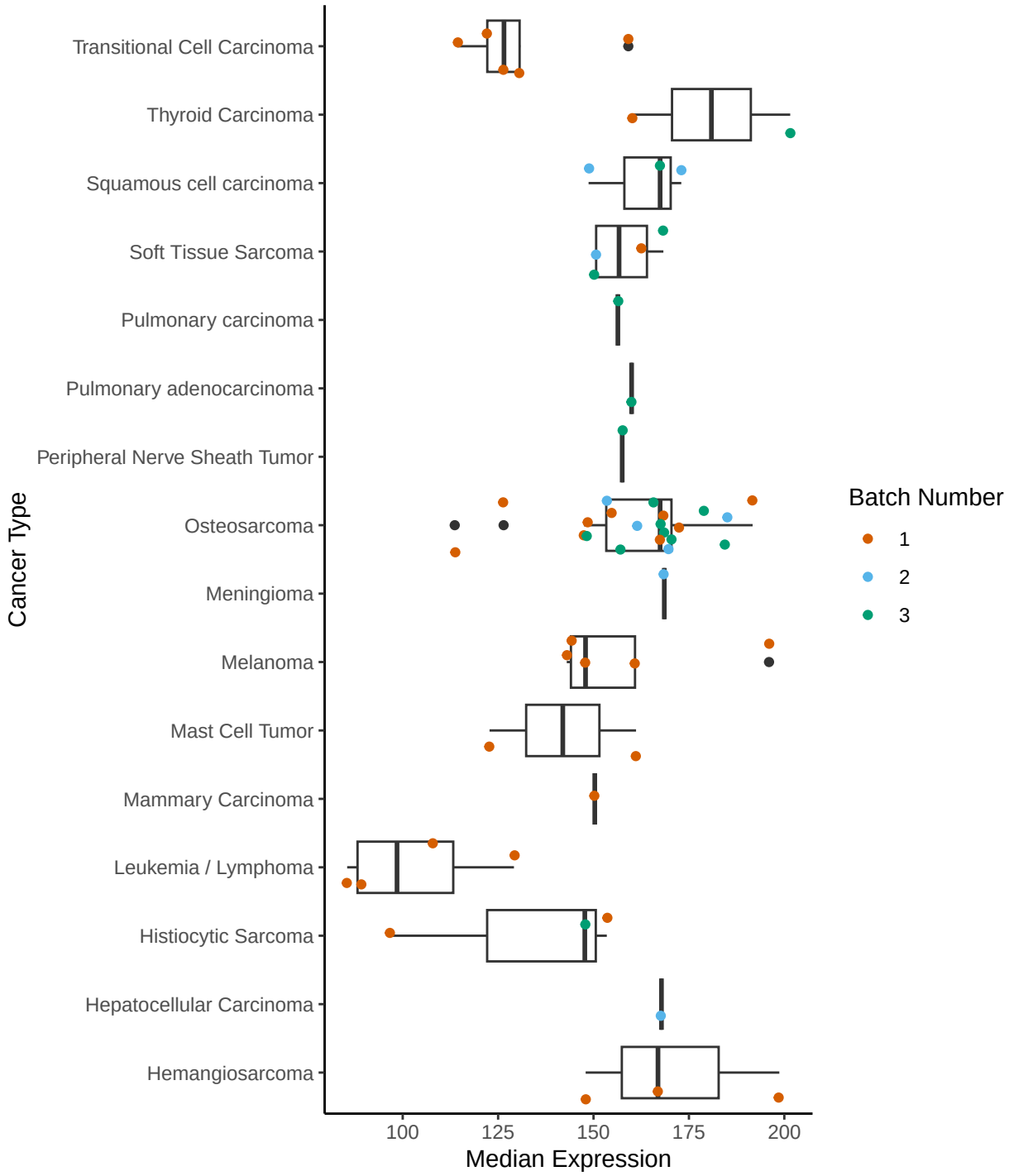

Figure 11: Box plot of the median expression values calculated using the DESeq normalized data with a jitter. The jitter also illustrates the batch numbers that median values were calculated from. Entire plot is organized by the cancer type.

Max Gene Expression

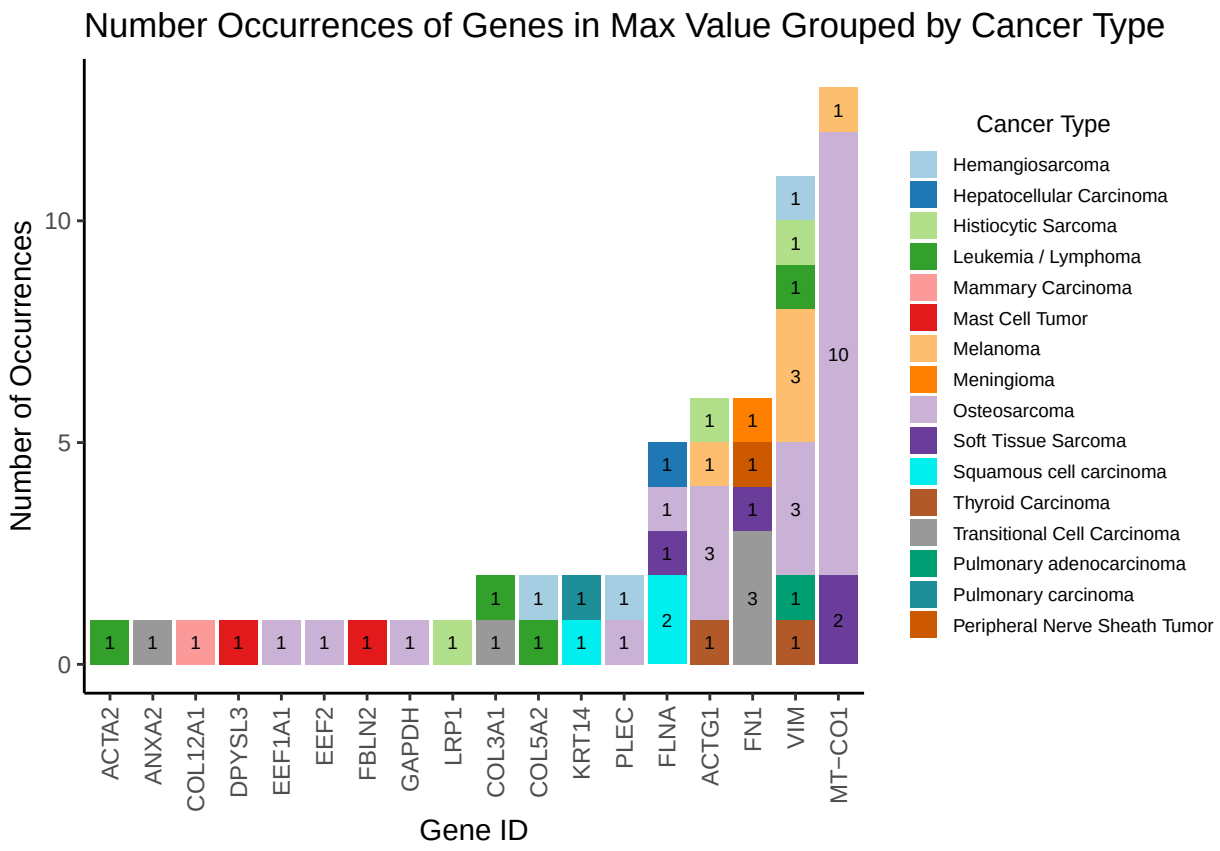

Figure 12: This bar graph visualizes the distribution of genes with the highest expression level across the 58 cell lines. While nearly 20,000 genes were analyzed, only 18 emerged as the most highly expressed gene in at least one cell line. Each bar represents a gene, and its color indicates the specific cancer(s) where that gene showed the highest expression. Refer to Appendix Table 20 for detailed descriptions of these 18 genes.