QBS 103: Final Project

Riya Mehta

22 August 2023

Contents

1 Chosen Gene: *BPI*

The datasets analyzed in this project are collected from RNA-seq and mass spectrometry of diverse COVID-19 patients [1]. Different genes are involved in the data collection process. *BPI* or Bactericidal Permeability Increasing protein, my chosen gene, plays a role in the immune response against bacterial infections. It can kill the bacteria by targeting their cell membranes and can also increase permeability of the cell membranes. Additionally, *BPI* encodes a lipopolysaccharide binding protein which helps with regulating inflammatory responses [2].

2 Summary Statistics of *BPI*

Summary Statistics for *BPI* (n = 99)		
Variable	**Statistic**	**Value**
Age	Mean (SD)	61.64 (15.41)
Ventilator-Free Days	Mean (SD)	19.17 (11.6)
Procalcitonin (ng/mL)	Mean (SD)	3.15 (9.93)
Female	n (%)	41 (41.41)
Male	n (%)	58 (58.59)
With Mechanical Ventilation	n (%)	46 (46.46)
Without Mechanical Ventilation	n (%)	53 (53.54)
In ICU	n (%)	58 (58.59)
Not in ICU	n (%)	41 (41.41)

Table 1: Subject demographics and clinical characteristics

This table (Table 1) displays the summary statistics for the chosen gene, *BPI*. It shows different variables including age, ventilator-free days, procalcitonin, sex, mechanical ventilation, and ICU status along with the values associated with each variable.

3 Corrected Figures

3.1 Histogram

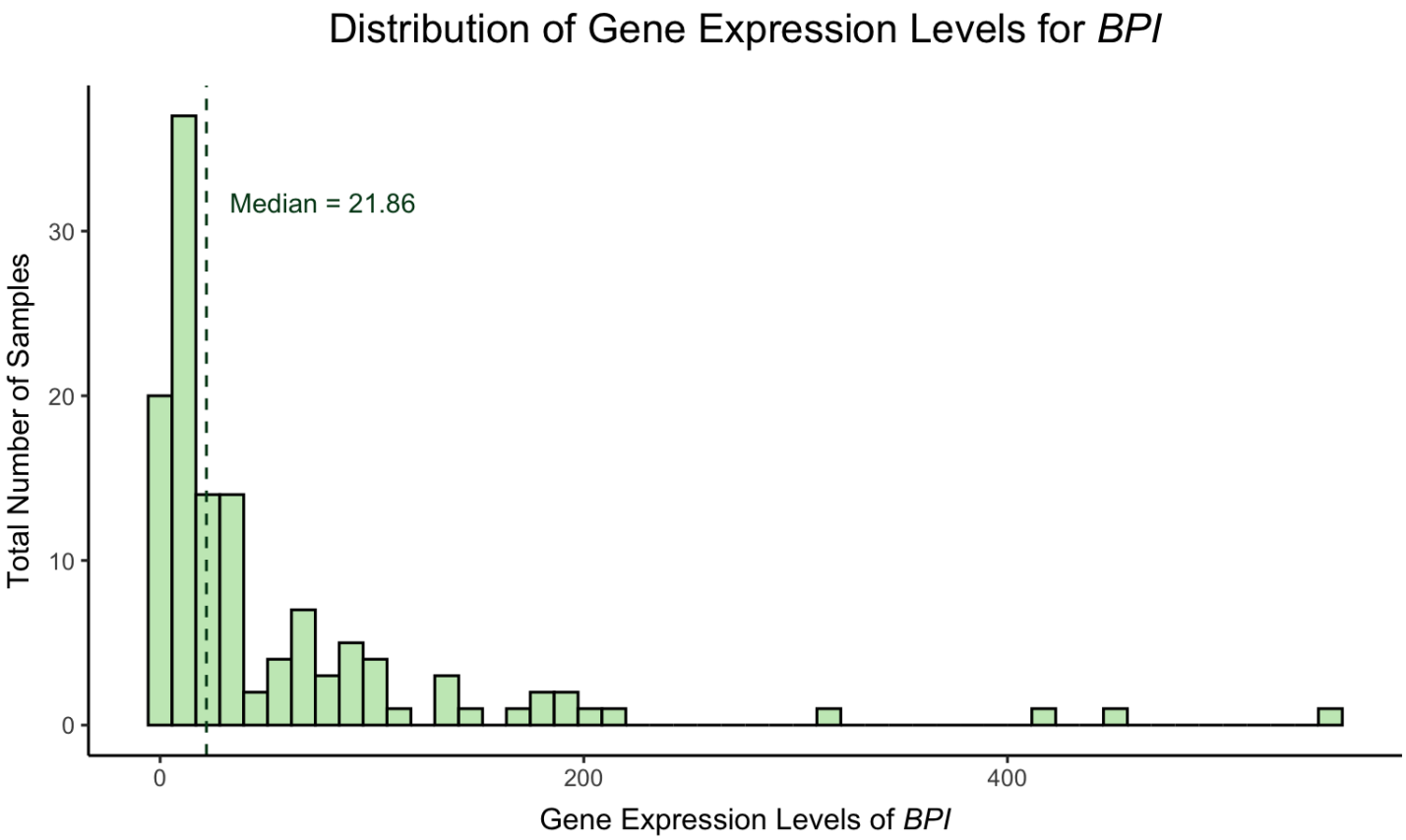

Figure 1: Histogram including all feedback from previous submissions

The histogram above (Figure 1) displays the distribution of *BPI* gene expression levels and the number of samples present in the dataset. There is a right skew present, meaning that there are more samples with lower gene expression levels with a smaller amount of samples showing higher gene expression levels. The tail displays the outliers on the right side of the graph, and shows us that the gene is not expressed at the same level for all individuals. The majority of the samples are also seen below the threshold of 200, and the median gene expression of the selected is calculated to be 21.86.

3.2 Scatterplot

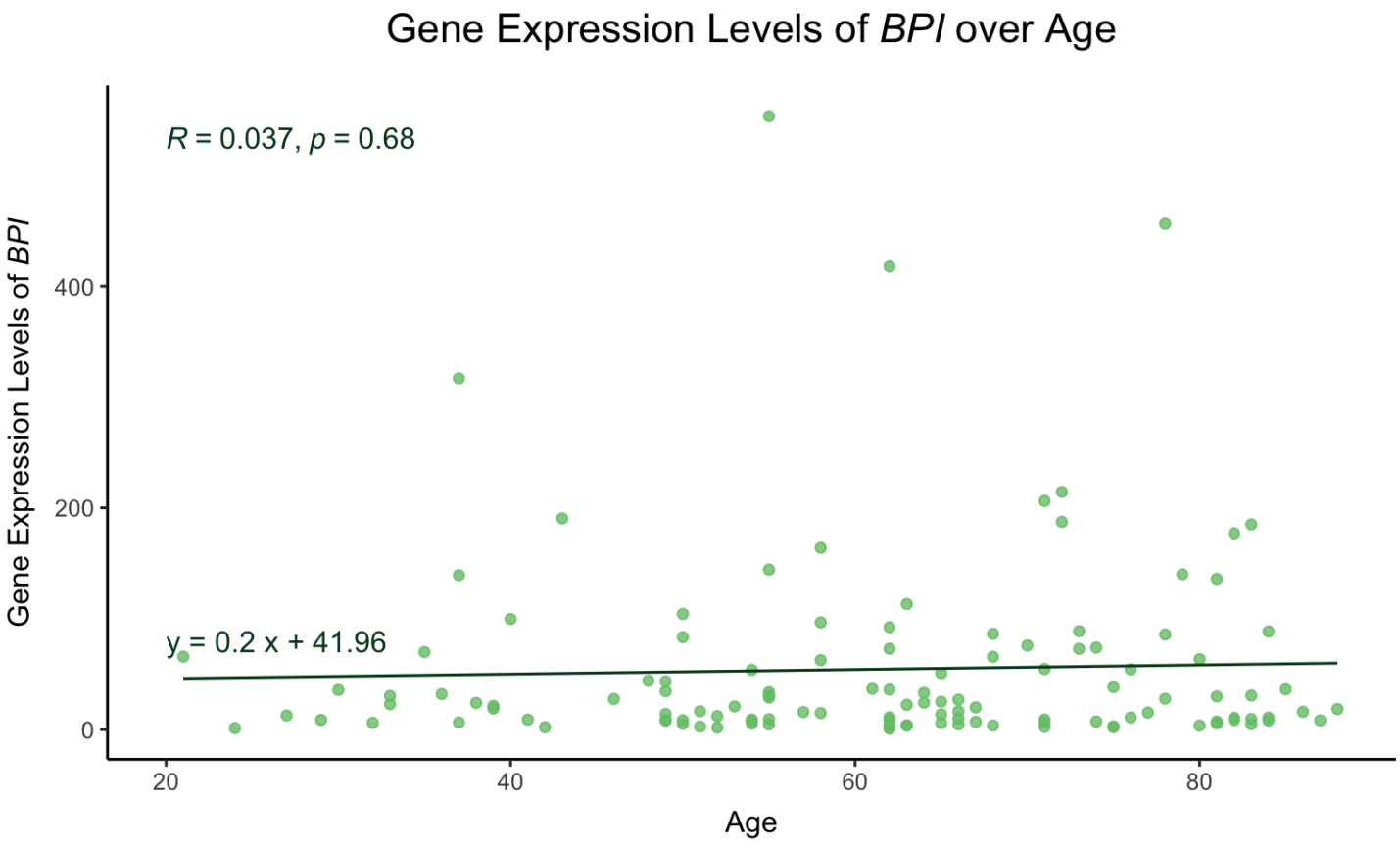

Figure 2: Scatterplot including all feedback from previous submissions

The scatter above (Figure 2) displays the relationship between *BPI* gene expression and ages of individuals sampled. There is a slight positive correlation between the two continuous variables (the slope of the line of best fit is around 0.2). The points seem to cluster towards the lower gene expression levels, specifically around ages 50 to 85. The y-intercept of the line of best fit for this graph is at 41.96, and there are a few outliers seen scattered towards higher gene expressions. The R value is 0.037, which means that there is a weak correlation between the two variables. The p value is 0.68, which means that we fail to reject the null hypothesis that there is no correlation between the two variables.

3.3 Boxplot

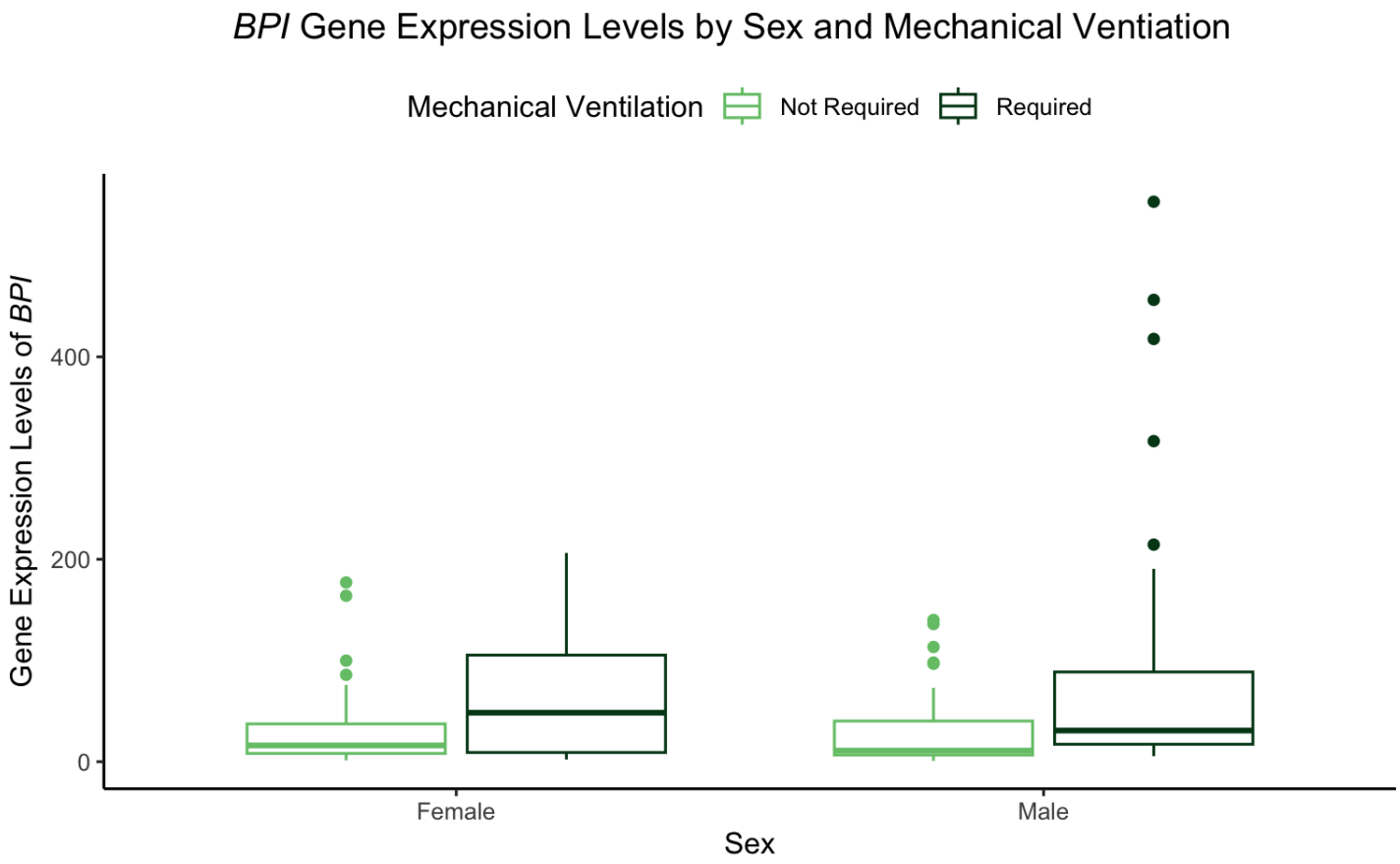

Figure 3: Boxplot including all feedback from previous submissions

The boxplot above (Figure 3) displays the distribution of gene expression of *BPI* across sexes and mechanical ventilation categories (required ventilation and did not require ventilation). It is observed that the samples with mechanical ventilation have a higher mean gene expression level for both males and females. The groups that required mechanical ventilation have a higher IQR when compared to the groups that did not require mechanical ventilation. Females with and without mechanical ventilation have a higher mean gene expression level than males with and without mechanical ventilation, respectively.

4 Heatmap

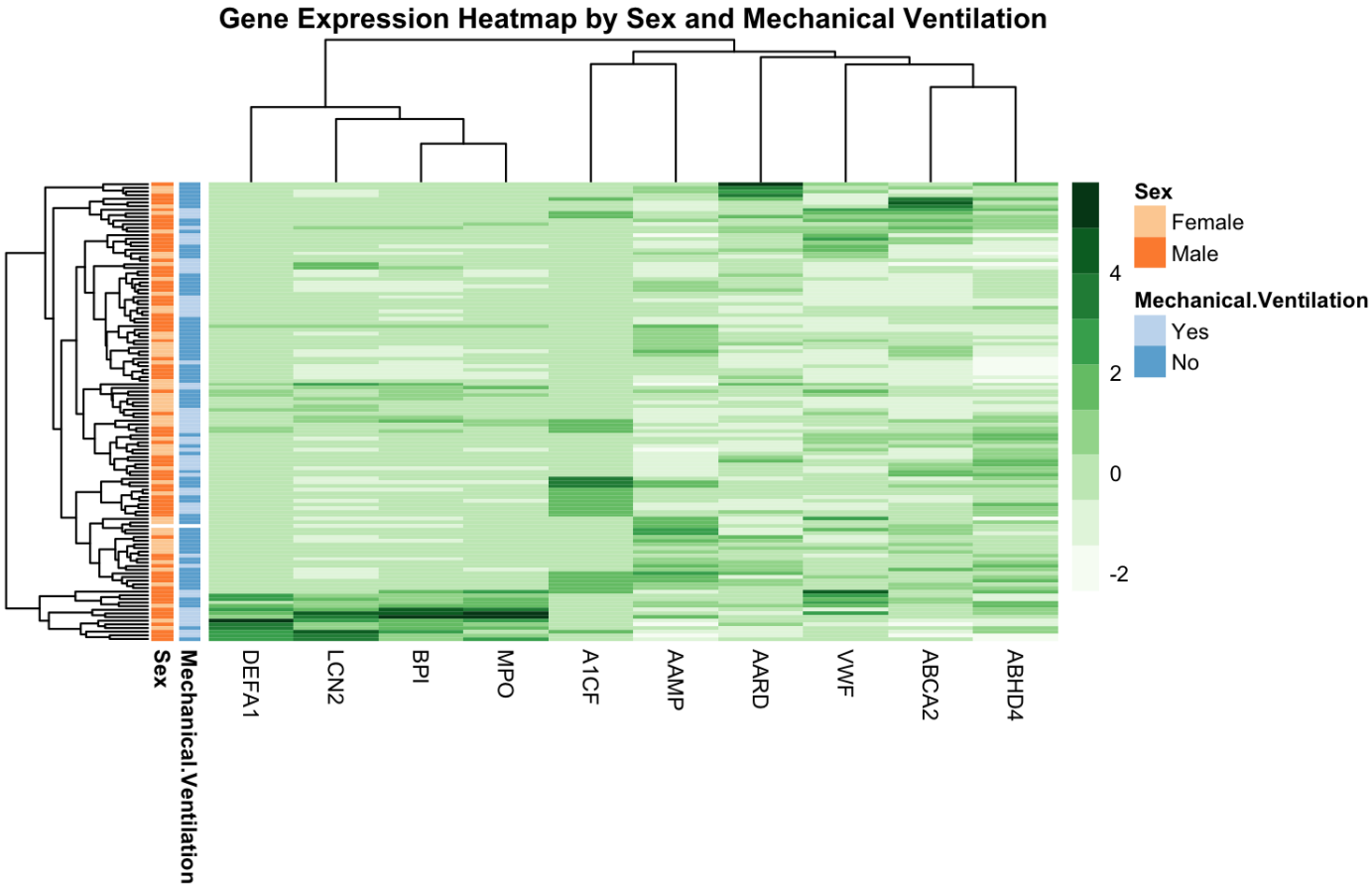

Figure 4: Heatmap created from ten genes

The heatmap above (Figure 4) displays patterns in gene expression, sex, and mechanical ventilation for each individual sampled. The intensity of the green color in the center represents how high the scaled gene expression is per sample. Euclidean clustering was done on the data. The tracking bars on the side show a differentiation in mechanical ventilation and sex, which are two columns down on the left end of the map. We can see the correlations between the ventilation and sex as well as gene expression from this.

5 Hexbin Plot

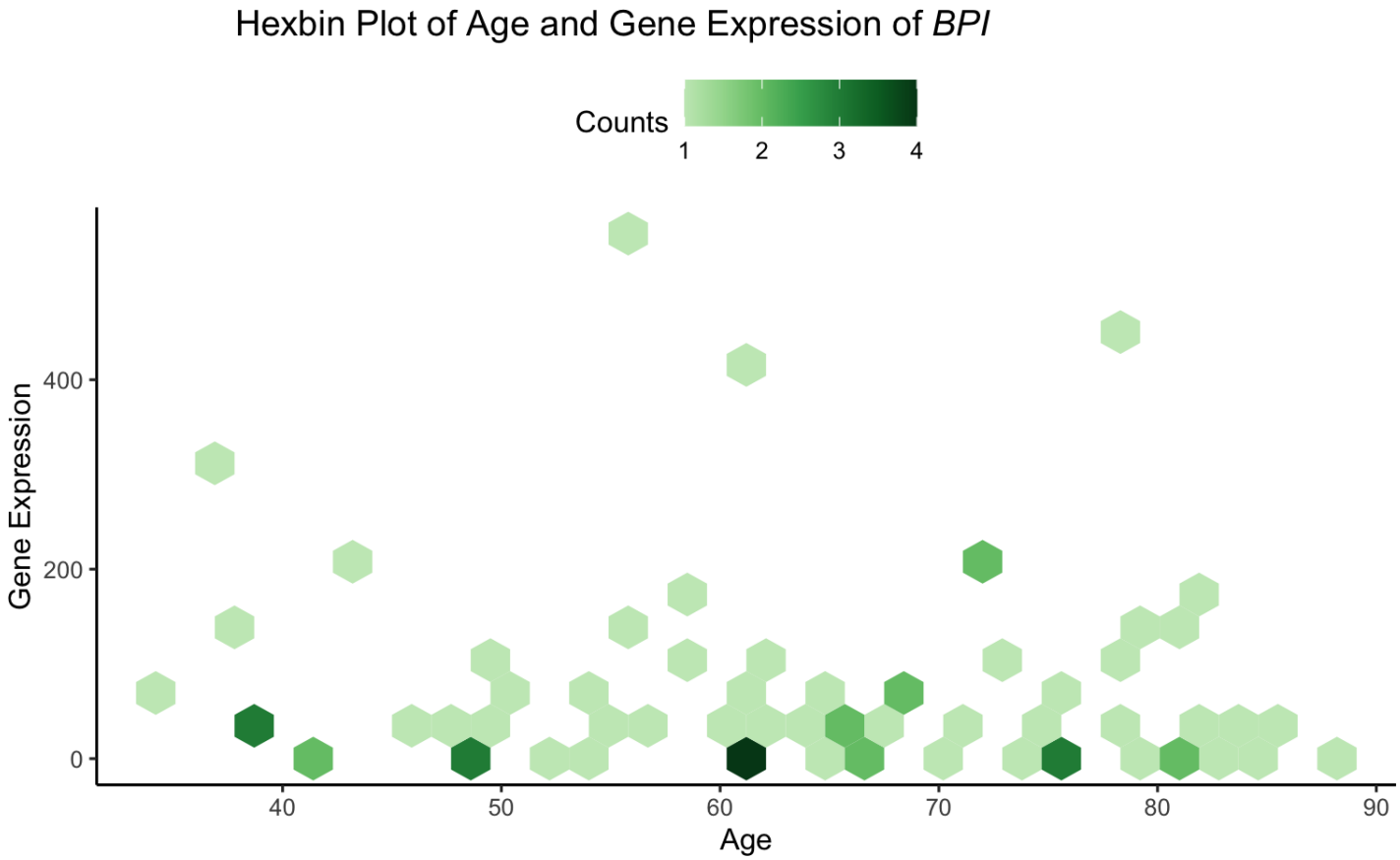

Figure 5: Hexbin plot

The hexbin plot above (Figure 5) helps in visualizing the relationship seen between age and *BPI* gene expression. Each hexagon in the plot is not just one point from the dataset but is clustered by bins. The number of individual samples that are included within a bin is represented by the intensity of the green, as shown by the legend. This plot allows us to understand how gene expression varies based on age.

References

[1] Overmyer, K. A, Shishkova, E, Miller, I. J, Balnis, J, Bernstein, M. N, Peters-Clarke, T. M, Meyer, J. G, Quan, Q, Muehlbauer, L. K, Trujillo, E. A, He, Y, Chopra, A, Chieng, H. C, Tiwari, A, Judson, M. A, Paulson, B, Brademan, D. R, Zhu, Y, Serrano, L. R, Linke, V, Drake, L. A, Adam, A. P, Schwartz, B. S, Singer, H. A, Swanson, S, Mosher, D. F, Stewart, R, Coon, J. J, & Jaitovich, A. (2021) Large-scale multi-omic analysis of covid-19 severity. *Cell Systems* **12**, 23–40.e7.

[2] (2023) Ncbi: Bpi bactericidal permeability increasing protein [homo sapiens (human)].