QBS 103: Final Project

Yihan (Ethan) Zhang

August 27, 2024

Contents

1 Introduction

1.1 Abstract

The dataset comprises 126 samples, each representing a different individual, and contains expression data for 100 genes. Among these, the gene "ABCA1" was selected for detailed analysis. ABCA1 is a key gene involved in cholesterol metabolism, crucial for generating HDL-C particles that remove cholesterol from cells. Defects in ABCA1, like in Tangier disease, lead to low HDL-C levels and cholesterol buildup, contributing to atherosclerosis. Increasing ABCA1 expression is a potential strategy to reduce the risk of atherosclerotic cardiovascular disease (ASCVD).[1] The expression values in the dataset are numerical representations of the gene expression levels across different samples, which include various patient characteristics such as age, sex, and ICU status. The primary goal is to visualize the expression patterns of "A1BG" using different types of plots, including bar plots, scatter plots, and box plots, to explore any potential correlations between gene expression and patient characteristics.

2 Methods

2.1 Data Source

The dataset used for this analysis was obtained from publicly available sources. [2]

2.2 R version

It was processed using R version 3.6.2 (2019-12-12).

2.3 R Packages

Various R packages were employed, including tidyverse for data manipulation and visualization,[3] knitr and kableExtra for table formatting,[4] pheatmap for generating heatmaps,[5] and ggplot2 for creating visualizations.[6]

2.4 Clustering Algorithm

The heatmap clustering was performed using the Euclidean distance method, which is well-suited for continuous data such as gene expression levels. Clustering algorithms and color palettes were applied to visualize the patterns and relationships within the data effectively.[7]

3 Results

3.1 Summary Table

This table provides a summary of demographic and clinical characteristics for patients who were in the ICU compared to those who were not. The variables include age, ferritin levels,

CRP levels, sex distribution, and Charlson Comorbidity Scores. The mean age and biomarker levels (ferritin and CRP) are consistent between both groups. However, there are noticeable differences in the distribution of Charlson Scores and sex between the groups, with a higher proportion of males and higher Charlson Scores in the ICU group. (Table 1).

Variable-mean(sd)	In ICU	Not In ICU
Age	63.45 (14)	58.67 (17.82)
Ferritin (ng/mL)	935.32 (1019.02)	715.75 (1067.55)
CRP (mg/L)	149.57 (105.54)	109.4 (94.38)
Sex-n(%)		
male	41 (62.12%)	33 (55%)
female	24 (36.36%)	27 (45%)
unknown	1 (1.52%)	0 (0%)
Charlson Score-n(%)		
0	1 (1.52%)	10 (16.67%)
1	10 (15.15%)	9 (15%)
2	14 (21.21%)	11 (18.33%)
3	13 (19.7%)	3 (5%)
4	6 (9.09%)	9 (15%)
5	6 (9.09%)	5 (8.33%)
6	2 (3.03%)	6 (10%)
7	8 (12.12%)	4 (6.67%)
8	3 (4.55%)	3 (5%)
9	2 (3.03%)	0 (0%)
11	1 (1.52%)	0 (0%)

Table 1: Demographic and clinical characteristics of patients stratified by ICU status. This table shows the mean age, ferritin, and CRP levels, as well as the distribution of sex and Charlson Comorbidity Scores among ICU and non-ICU patients.

3.2 Histogram

This histogram displays the distribution of expression levels for the gene 'ABCA1' across the dataset. The gene expression levels are varied, with a higher frequency observed at lower expression levels and a tapering frequency at higher expression levels, indicating a right-skewed distribution. (Figure 1).

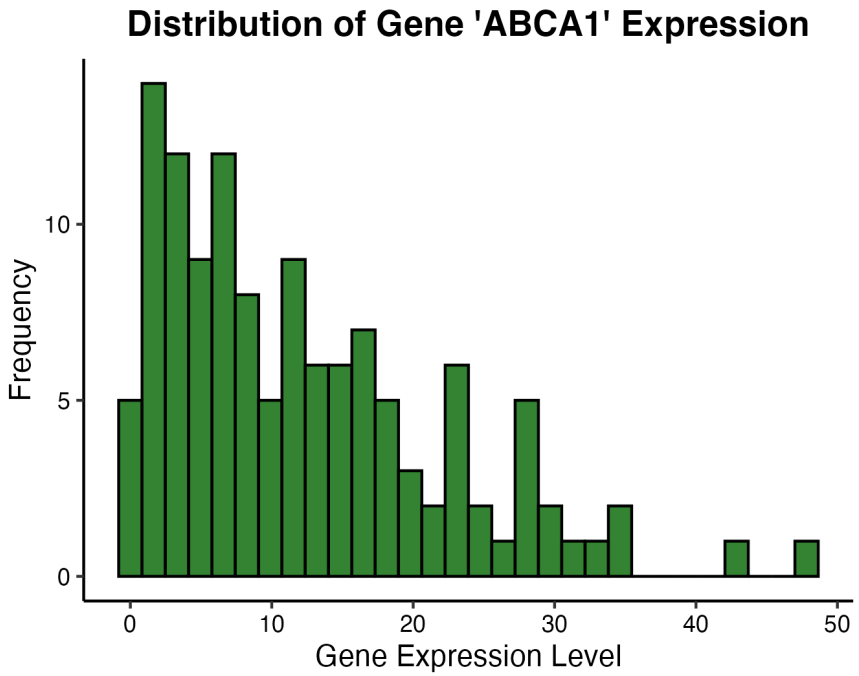

Figure 1: Distribution of Gene 'ABCA1' Expression Levels Across the Sample Set.

3.3 Scatter plot

This scatterplot visualizes the relationship between ABCA1 gene expression levels and participant age. The plot reveals a wide distribution of gene expression levels across different ages, with no clear linear relationship. The color gradient indicates higher gene expression levels in red and lower levels in green, with most data points concentrated in the lower expression range. (Figure 2).

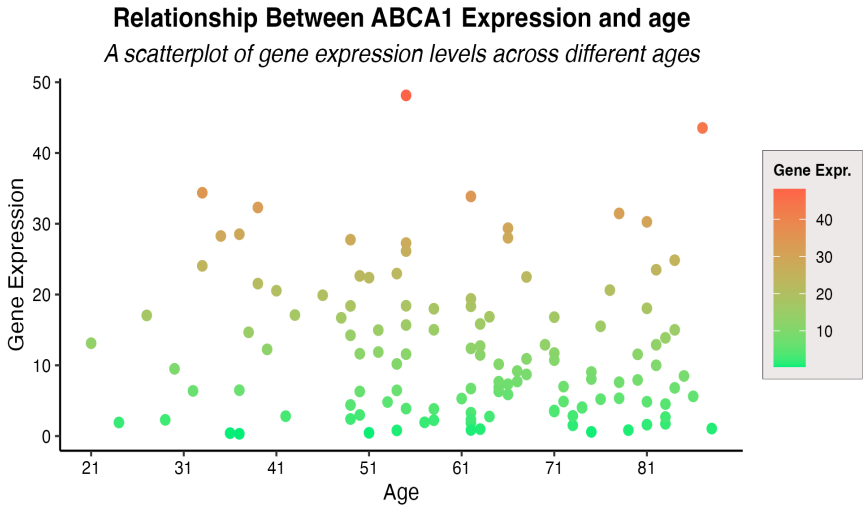

Figure 2: Scatterplot of ABCA1 Gene Expression Levels Across Different Ages.

3.4 Boxplot

This boxplot compares the distribution of gene expression levels by ICU status and sex. The plot shows that males generally have higher gene expression levels than females, with more variability in the "no ICU" group. The plot also highlights a few outliers, particularly among males in the "yes ICU" group. (Figure 3).

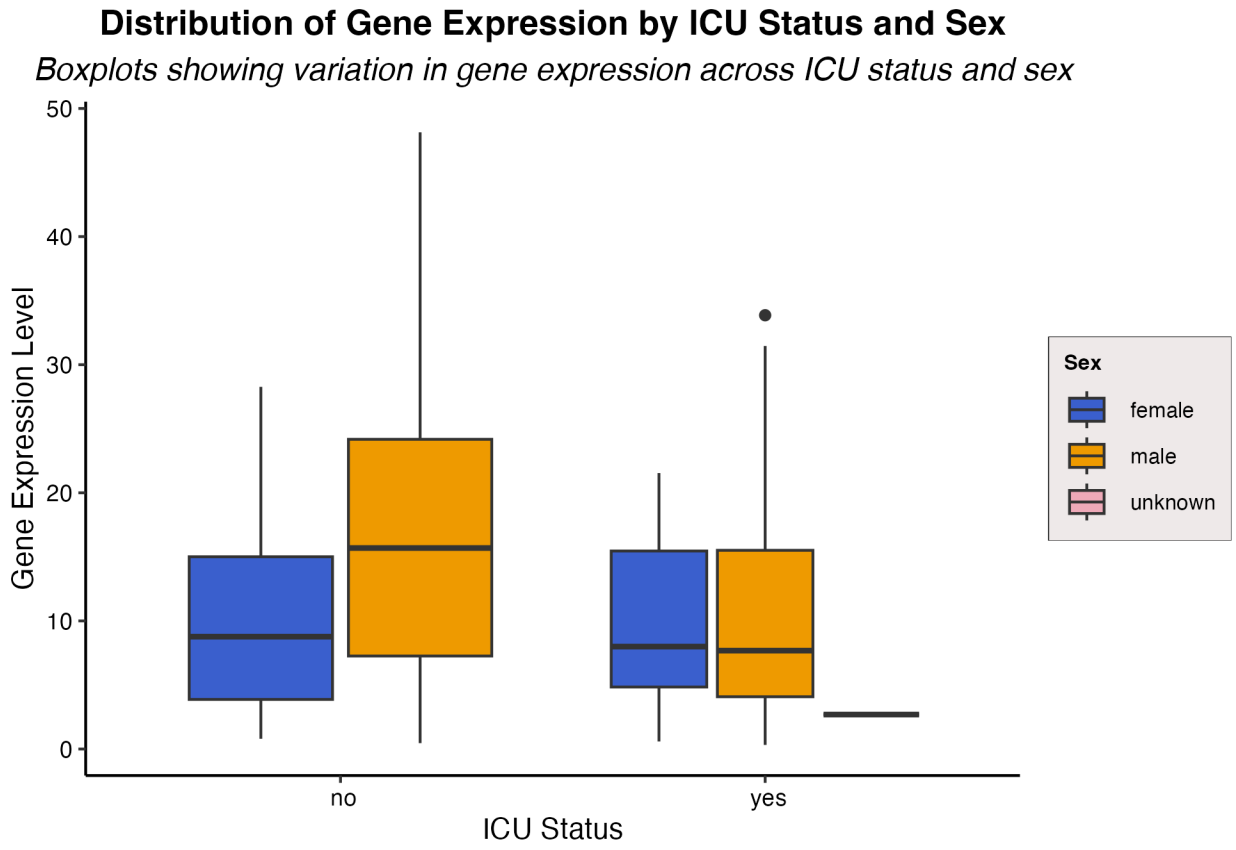

Figure 3: Boxplot of Gene Expression Levels by ICU Status and Sex, Showing Variation Across Groups.

3.5 Heatmap

This heatmap visualizes the expression levels of selected genes across different samples, with hierarchical clustering applied to both genes and samples. The color gradient represents gene expression levels, with blue indicating lower expression and red indicating higher expression. The samples are annotated by sex and ICU status, allowing for the identification of patterns or clusters related to these variables. (Figure 4).

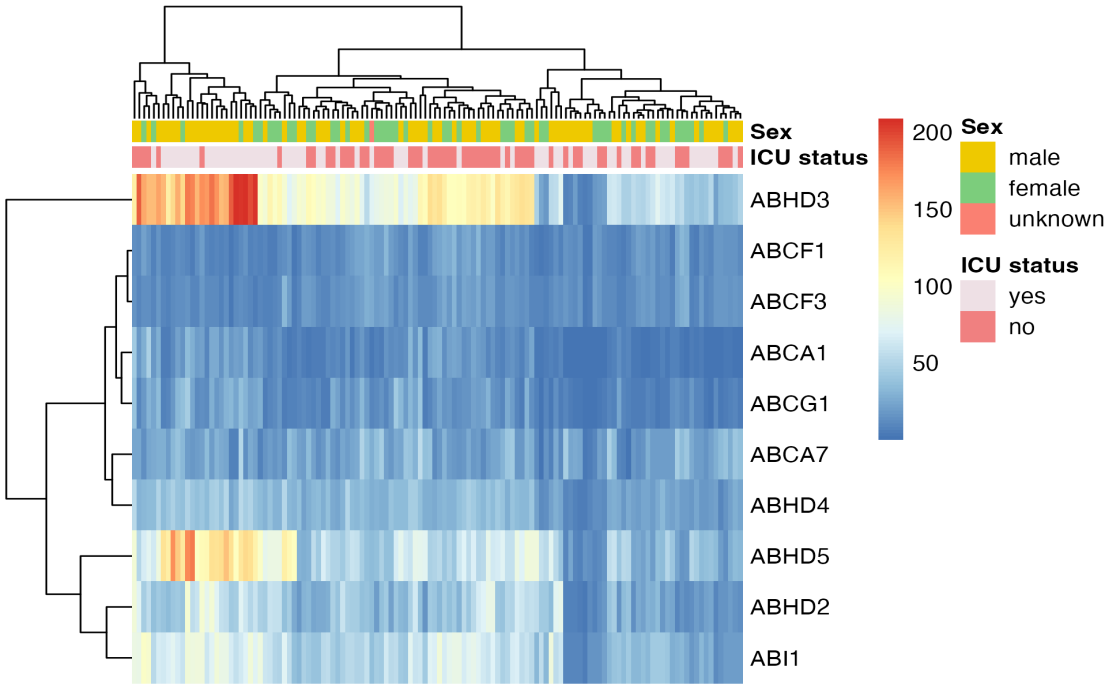

Figure 4: Heatmap of Gene Expression Levels with Hierarchical Clustering by Sample and Gene, Annotated by Sex and ICU Status.

3.6 Violin Plot

The violin plot displays the distribution of gene expression levels for males and females. The plot reveals that males tend to have a wider range and higher levels of gene expression compared to females. (Figure 5).

Figure 5: The plot shows that males generally exhibit a broader range and higher gene expression levels compared to females.

References

[1] Choi, H. Y, Choi, S, Iatan, I, Ruel, I, & Genest, J. (2023) Biomedical advances in abca1 transporter: From bench to bedside.

[2] 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**.

[3] Wickham, H, Averick, M, Bryan, J, Chang, W, McGowan, L, François, R, Grolemund, G, Hayes, A, Henry, L, Hester, J, Kuhn, M, Pedersen, T, Miller, E, Bache, S, Müller, K, Ooms, J, Robinson, D, Seidel, D, Spinu, V, Takahashi, K, Vaughan, D, Wilke, C, Woo, K, & Yutani, H. (2019) Welcome to the tidyverse. *Journal of Open Source Software* **4**.

[4] Zhu, H. (2019) *KableExtra: Construct complex table with 'kable' and pipe syntax*. Vol. 1.

[5] Kolde, R. (2022) Package 'pheatmap': Pretty heatmaps. *R package*.

[6] Tyner, S, Briatte, F, & Hofmann, H. (2017) Network visualization with ggplot2. *R Journal* **9**.

[7] Ultsch, A & Lötsch, J. (2022) Euclidean distance-optimized data transformation for cluster analysis in biomedical data (edotrans). *BMC Bioinformatics* **23**.