Ian Sam
V00990790

STAT 469 Assignment #4

Introduction

This report evaluates the clustering patterns of a bulk RNA-seq gene expression matrix using hierarchical clustering and k-means clustering. The objective is to explore how patients and genes can be grouped based on their expression profiles and to visualize these groupings using heatmaps.

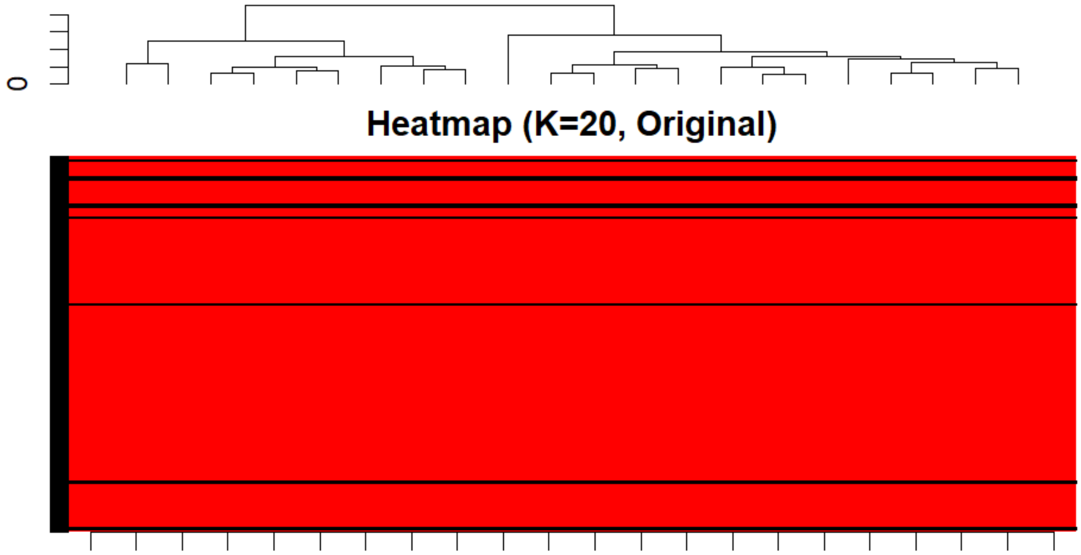
Hierarchical clustering with complete linkage is applied to the patients to identify similarities in gene expression across samples. K-means clustering is applied to the genes using three different cluster counts (K = 20, 30, 40), with 100 random initializations for each K. The analysis is performed using both the original gene expression values and log-transformed values to assess the effect of data scaling on clustering outcomes.

Six heatmaps are generated in total—three for the original data and three for the log-transformed data. Each heatmap includes a patient dendrogram and boundaries indicating gene clusters, allowing for a detailed comparison of clustering results across methods and transformations.

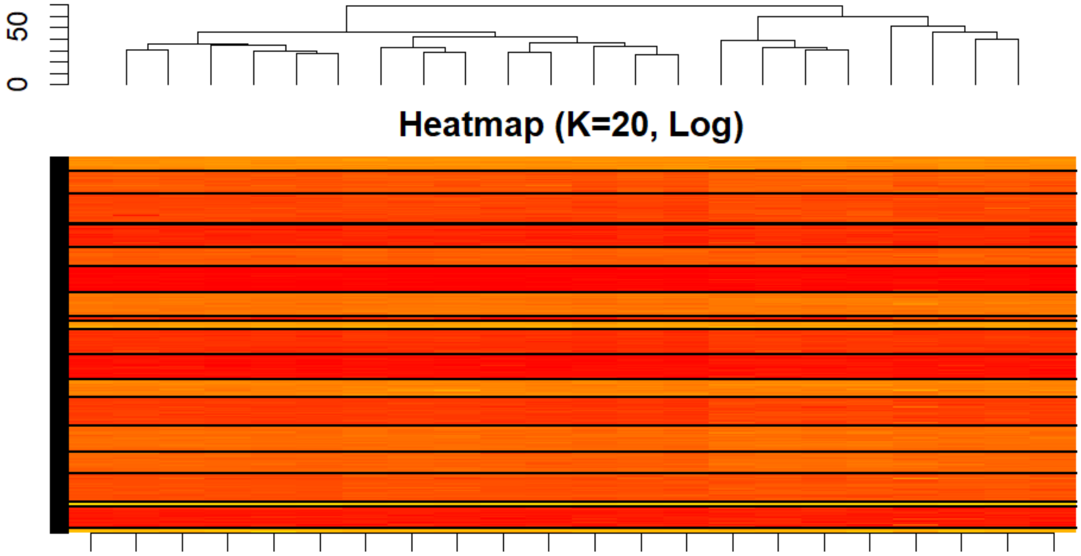
Methods Used

This analysis uses hierarchical clustering and k-means clustering to explore structure in a bulk RNA-seq gene expression matrix. Hierarchical clustering is applied to the patients (columns of the matrix) using the hclust function in R with complete linkage. K-means clustering is applied to the genes (rows of the matrix) using the kmeans function, with three different values of K (20, 30, and 40) and 100 random starts for each K to ensure stability in cluster assignments.

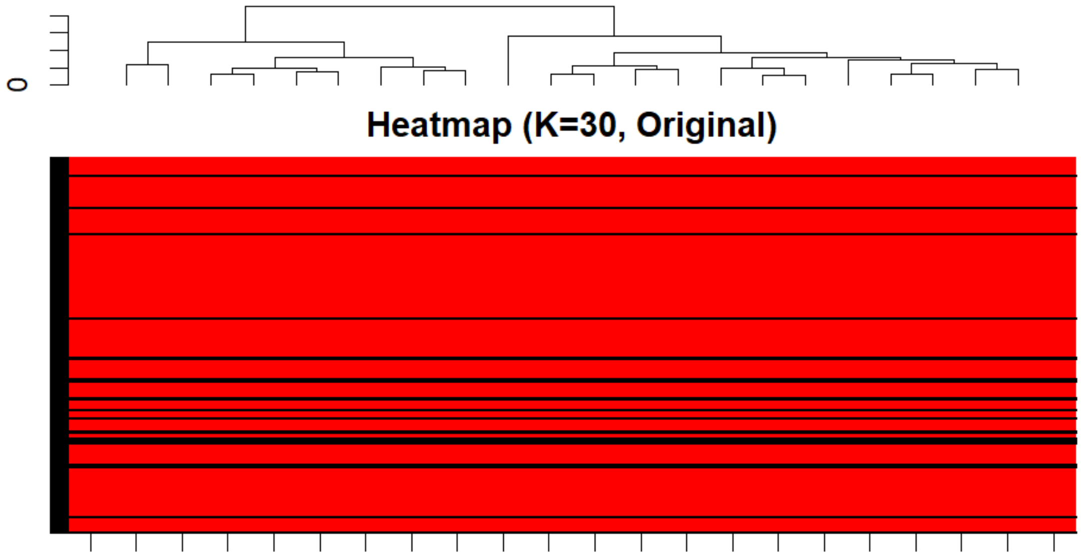
The gene expression matrix is analyzed on both the original scale and after log transformation. For each combination of K and transformation type, the data matrix is reordered based on clustering results and visualized as a heatmap using the image function in R. Each heatmap includes a dendrogram representing patient clusters and horizontal lines to indicate the boundaries of gene clusters identified by k-means.



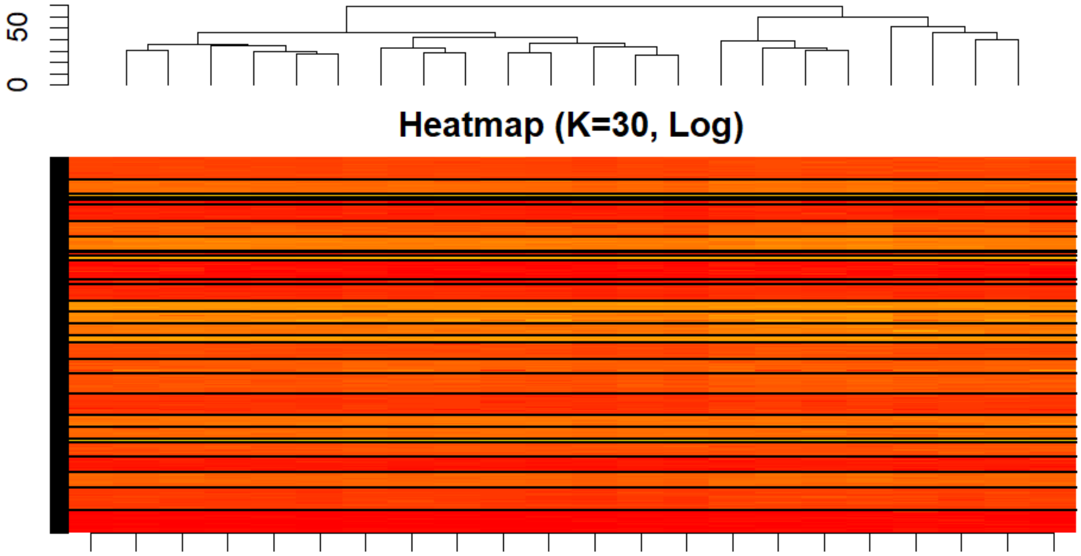
Heatmap (K=20, Original)



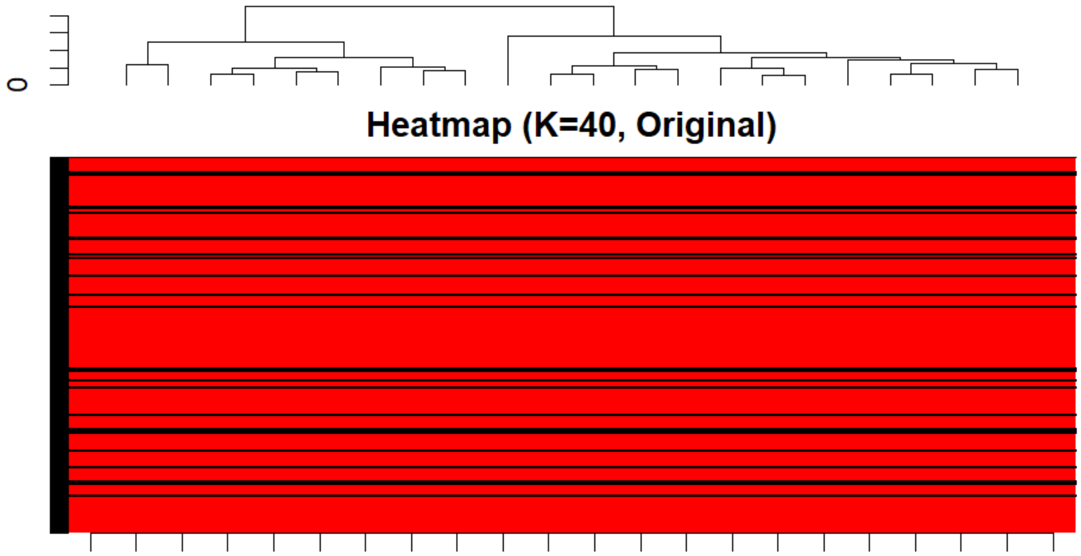
Heatmap (K=20, Log)

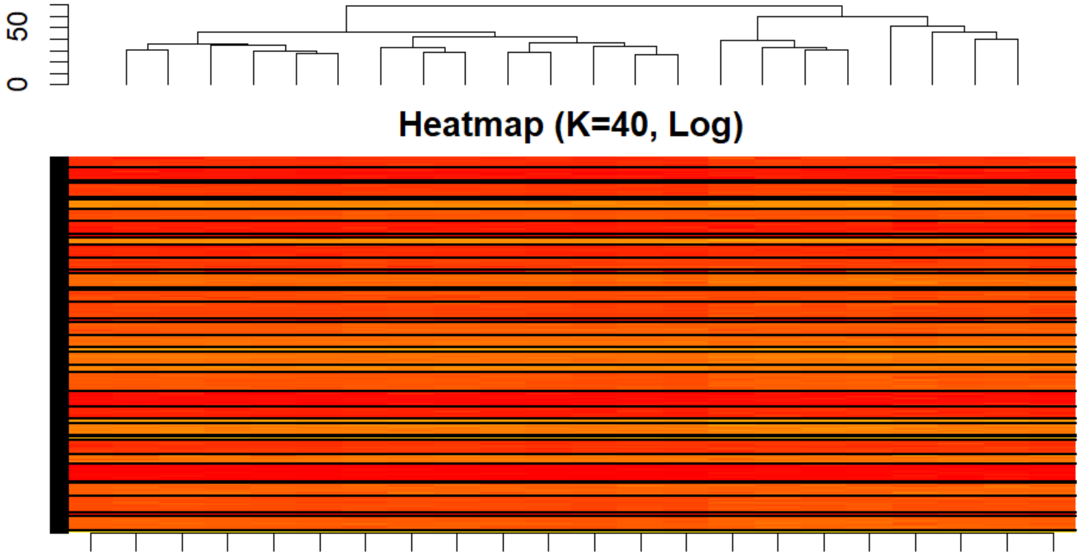


Heatmap (K=30, Original)



Heatmap (K=30, Log)


Heatmap (K=40, Original)


Heatmap (K=40, Log)

Results

Six heatmaps were generated to visualize clustering patterns under different settings: three using the original gene expression values and three using log-transformed values. For each transformation, k-means clustering was applied with K = 20, 30, and 40 to identify gene clusters, and hierarchical clustering was used to group patients.

In the original scale heatmaps, strong saturation limited the ability to distinguish expression differences, especially for higher expression values. In contrast, the log-transformed heatmaps revealed clearer structure. As K increased, gene clusters became more refined, particularly in the log-transformed condition, where smaller gene modules were more visible.

Several log-transformed heatmaps, especially at K = 30 and K = 40, displayed distinct horizontal bands, indicating that k-means successfully grouped together genes with similar expression profiles. The patient dendrograms also showed consistent subgroupings across different K values, suggesting that certain patients share similar underlying gene expression patterns. Overall, the log transformation improved interpretability and revealed meaningful structure that was difficult to observe in the original scale.

Conclusion

This analysis demonstrates that applying hierarchical clustering to patients and k-means clustering to genes can reveal meaningful structure in a bulk RNA-seq gene expression matrix. While the original-scale data provided limited interpretability due to saturation effects, the log-transformed data produced clearer visual patterns and more distinct gene clusters. As the number of gene clusters increased from K = 20 to K = 40, the heatmaps displayed more refined groupings, particularly in the log-transformed condition. Based on these results, using a log transformation and a higher number of gene clusters (K = 30 or 40) appears to be most effective for capturing and visualizing expression patterns across patients and genes.