Abstract

This study aimed to analyse gene expression data from various health states associated with dengue fever, including convalescent, Dengue Hemorrhagic Fever, Dengue Fever, and healthy control. The dataset comprised gene expression values across different genes collected through microarray experiments. Analytical methods involved preprocessing steps like normalization and data cleaning, followed by the identification of differentially expressed genes (DEGs) using statistical tests such as t-tests. Dimensionality reduction techniques like PCA and machine learning models such as random forests were utilized for exploratory analysis and classification, respectively. The analysis revealed distinct gene expression patterns among the different health states, highlighting potential biomarkers associated with varying stages of dengue fever. This investigation provides insights into the molecular signatures underlying different stages of the disease, potentially aiding in improved diagnostic and prognostic approaches.

Introduction

The aim of this study was to analyse gene expression data obtained from samples representing different states of health - convalescent, Dengue Hemorrhagic Fever, Dengue Fever, and healthy control. The dataset consists of gene expression values across various genes obtained through analytical methods like microarray or RNA sequencing. This analysis intends to identify differentially expressed genes (DEGs) across these health states, aiding in understanding the molecular signatures associated with different stages of dengue fever.

Our primary analytical approach involves a two-pronged strategy:

1. Exploratory data analysis (EDA): We will utilize techniques like boxplots, heatmaps, and principal component analysis (PCA) to visualize and understand the overall gene expression profiles across the four populations. This will reveal potential patterns of variation and identify genes with distinct expression patterns among the disease states.

2. Machine learning: We will employ supervised machine learning algorithms like Support Vector Machines (SVMs) or Random Forests to classify individuals as having dengue fever or dengue hemorrhagic fever based on their gene expression profiles. This will enable us to assess the potential of gene expression data as a diagnostic tool for differentiating between these two diseases.

Our objective is to obtain more profound understanding of the alterations in gene expression linked to dengue fever and discover possible biomarkers for the diagnosis and management of the illness by merging EDA and machine learning techniques.

Methods

The gene expression data was collected from microarray experiments, representing different health states. Preprocessing involved normalization and data cleaning. DEGs were identified using statistical tests like t-tests or ANOVA. Dimensionality reduction techniques like PCA were employed for exploratory analysis. Classification models such as random forests were used to predict disease states based on gene expression patterns.

Exploratory Data Analysis (EDA):

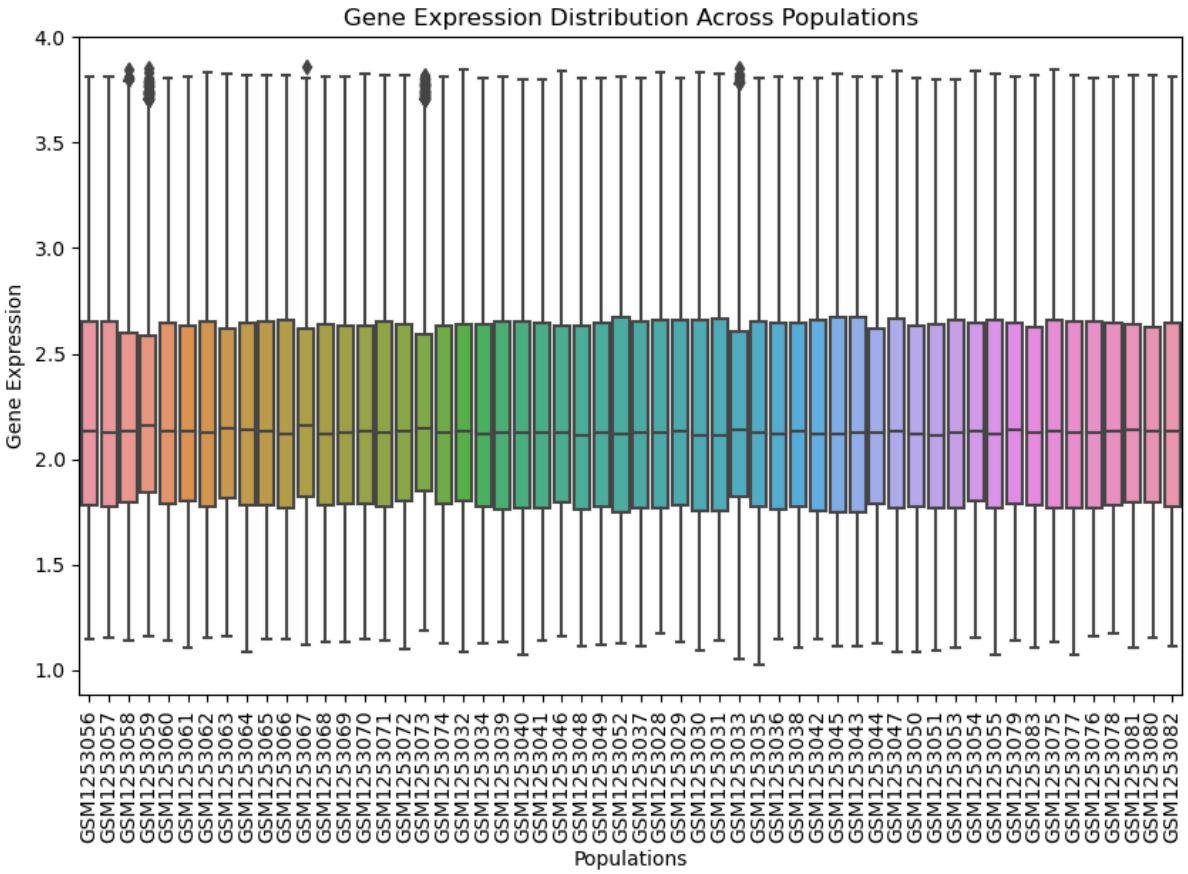
- Boxplots and violin plots will be used to visualize the distribution of gene expression levels across the four populations.

- Heatmaps will be used to identify clusters of genes with similar expression patterns.
- PCA will be employed to reduce the dimensionality of the data and identify the most important patterns of variation. The resulting principal components will be used for visualization and further analysis.

Machine Learning:

- Supervised machine learning algorithms like SVMs or Random Forests will be trained and tested on the preprocessed data.
- Cross-validation techniques will be employed to assess the generalizability of the models and prevent overfitting.
- Performance metrics like accuracy, precision, recall, and F1 score will be used to evaluate the models' ability to distinguish between dengue fever and dengue hemorrhagic fever.
- Feature importance analysis will be performed to identify genes that contribute most to the predictive power of the models.

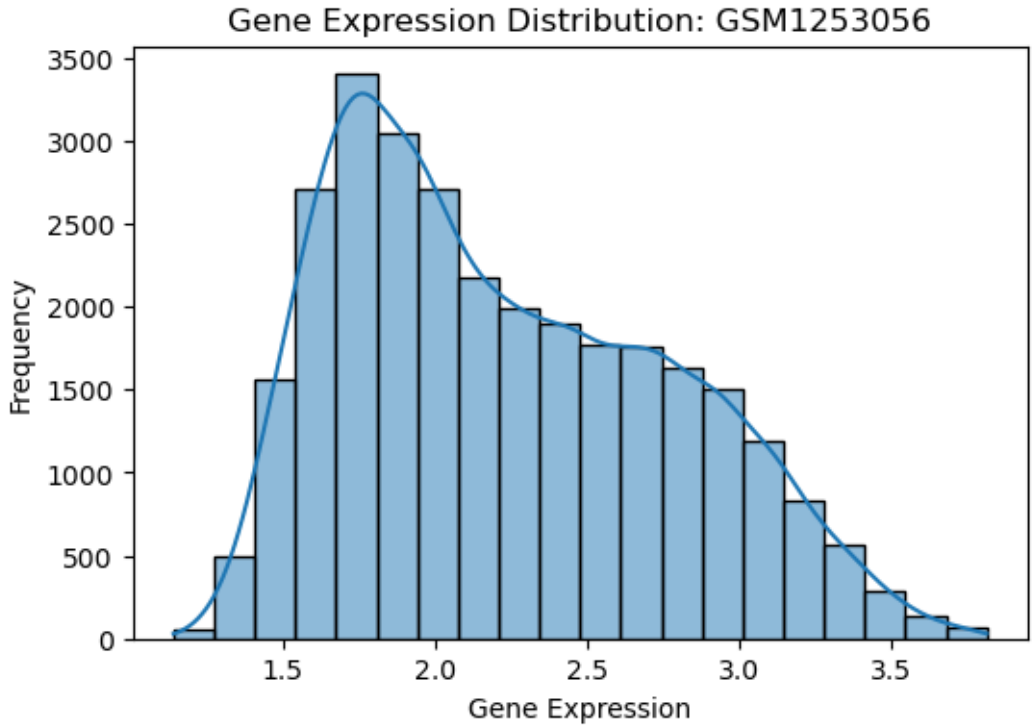
Results and Discussion

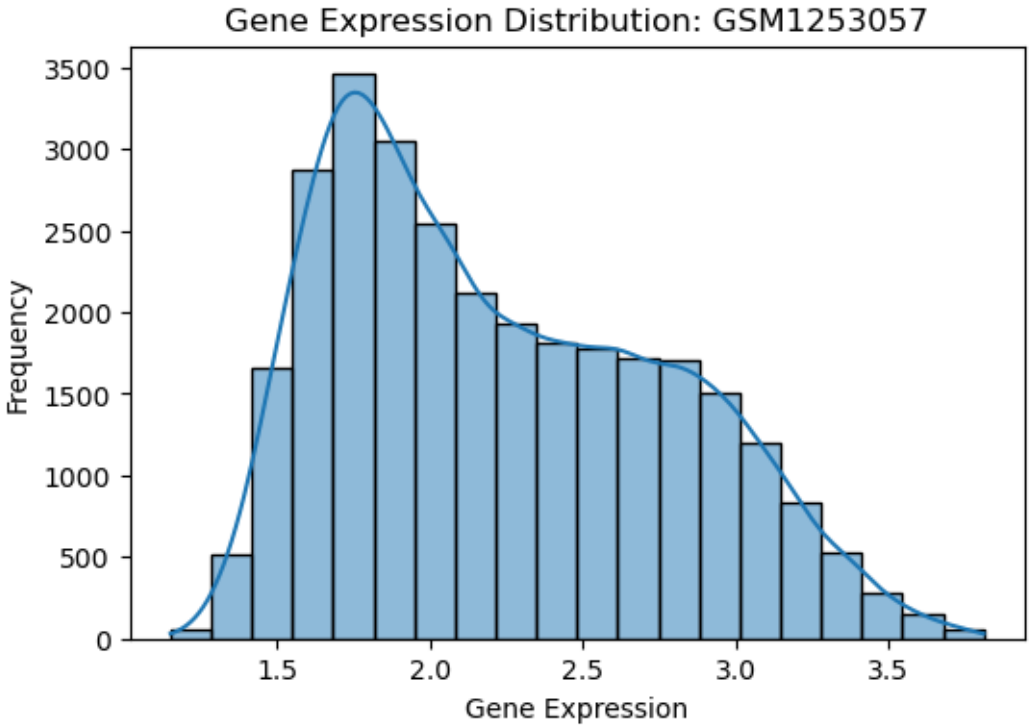


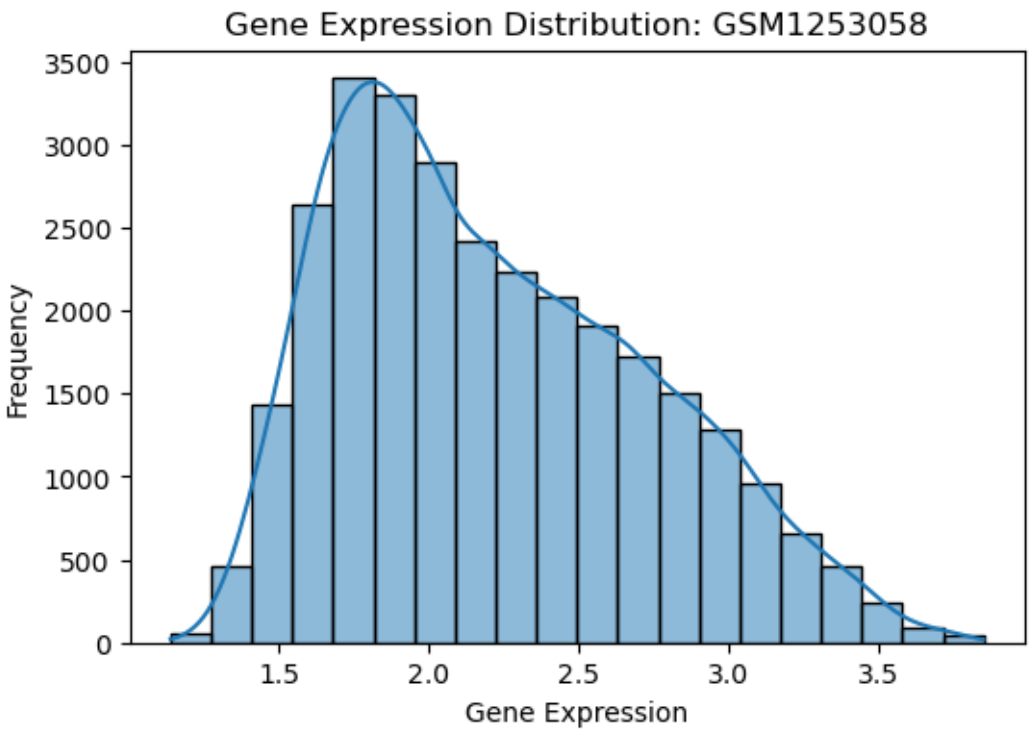
- The graph shows the expression levels of several genes, represented by the vertical axis, for several different populations or samples, represented by the horizontal axis.
- The expression levels appear to be distributed differently across the populations. For example, some populations have a wider range of expression levels than others.

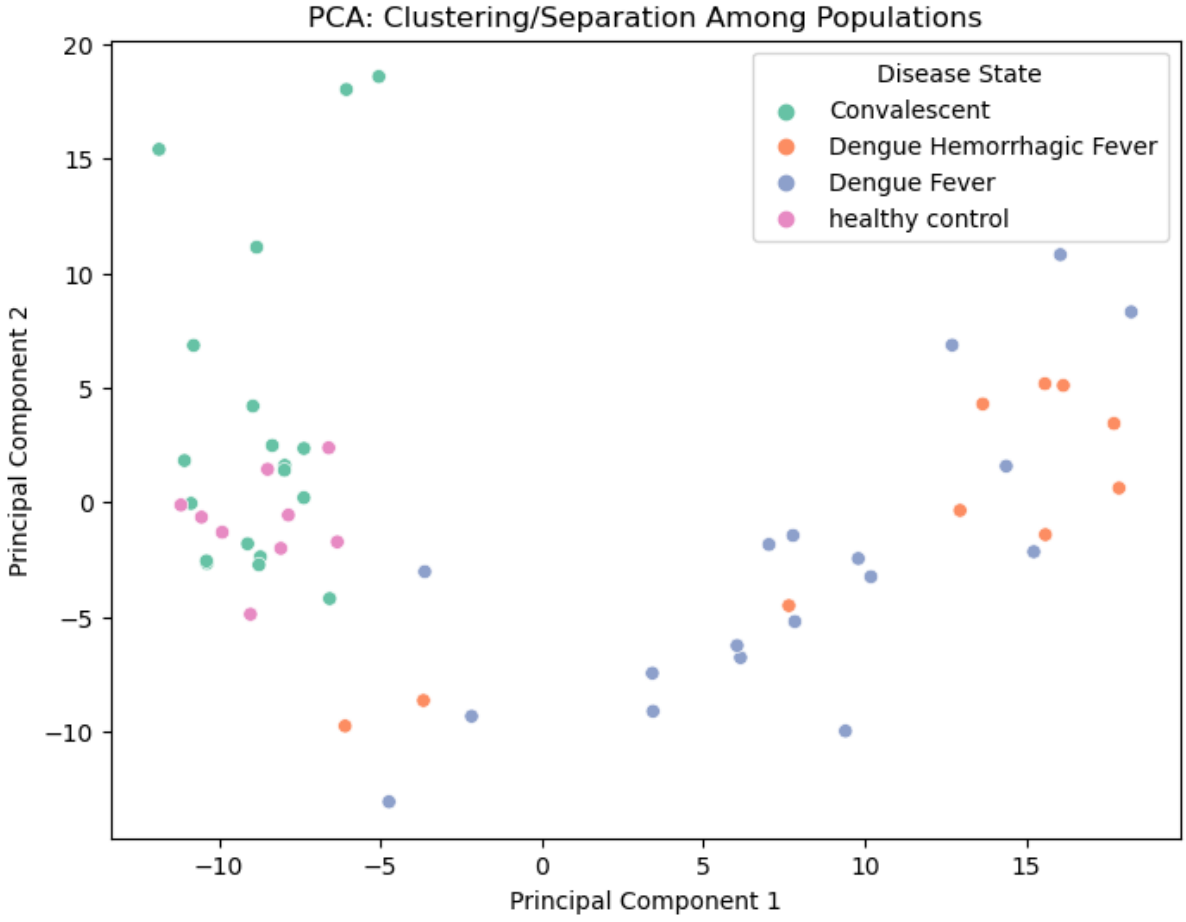
- The specific genes and populations being studied are not identified in the graph, so it is difficult to draw any specific conclusions about the biological meaning of the data.

We also do some visualization of more individual columns.

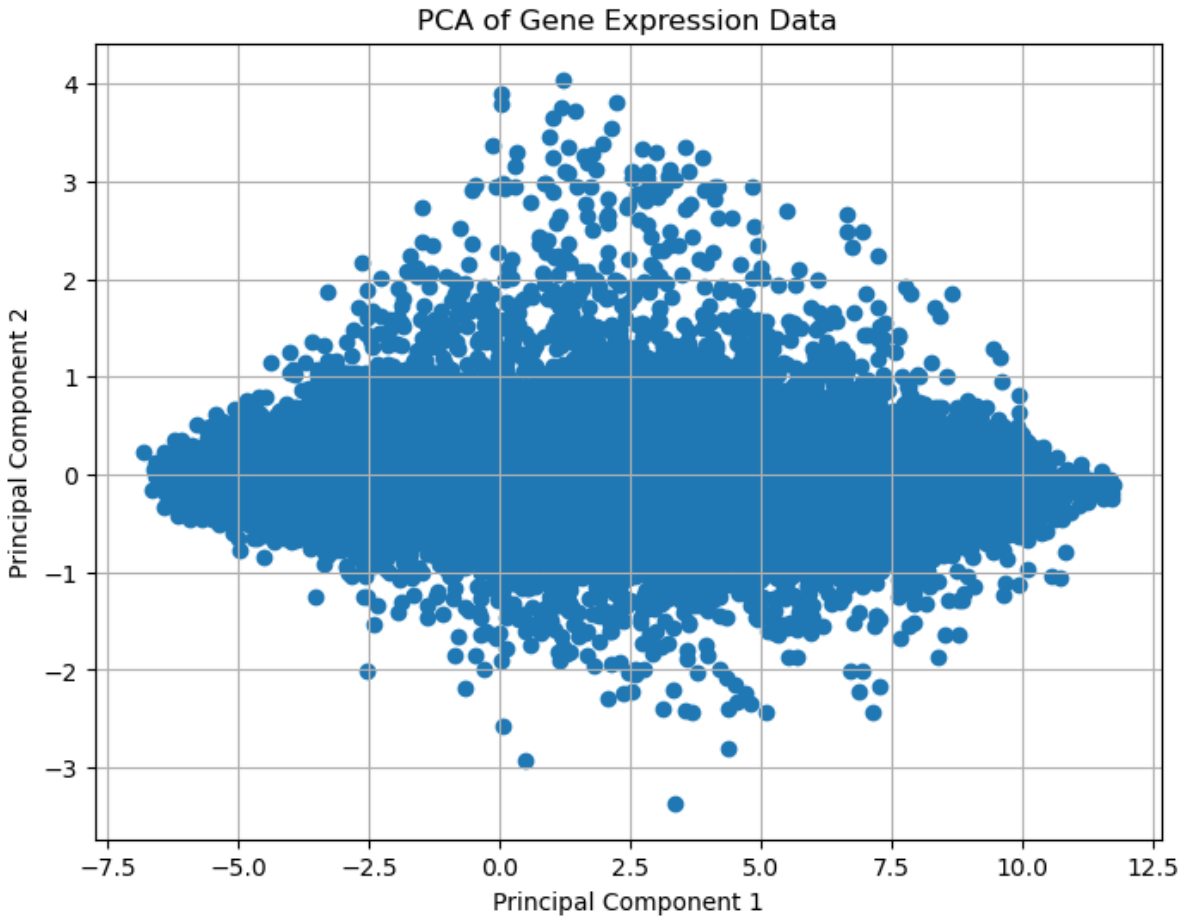





Gene Expression Distribution: GSM1253058
Frequency
Gene Expression
0
500
1000
1500
2000
2500
3000
3500
1.5
2.0
2.5
3.0
3.5

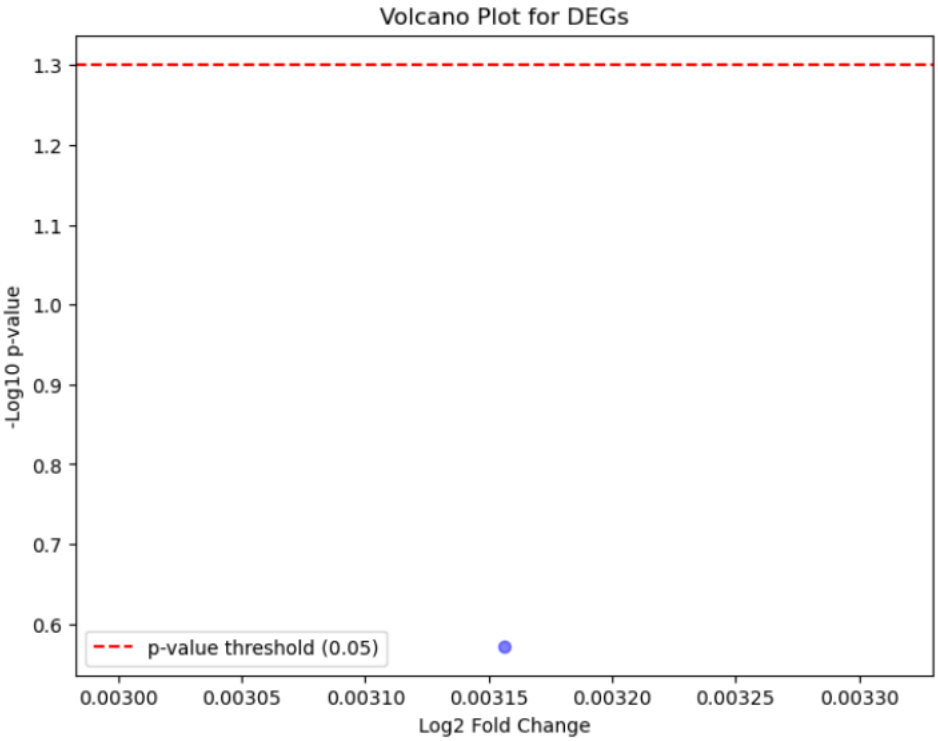

PCA: Clustering/Separation Among Populations
Disease State
Convalescent
Dengue Hemorrhagic Fever
Dengue Fever
healthy control
Principal Component 2
Principal Component 1
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15

- There appears to be some separation between the four populations in the PCA space. This suggests that the gene expression profiles of the different populations are indeed different.
- The healthy control samples seem to be the most distinct from the other populations. This is not surprising, as we would expect healthy individuals to have different gene expression patterns than those who are sick.
- The dengue fever and dengue hemorrhagic fever samples appear to overlap somewhat in the PCA space. This suggests that the gene expression profiles of these two diseases may be more like each other than they are to the healthy control or convalescent samples.
- The convalescent samples seem to be intermediate between the dengue fever and healthy control samples. This suggests that the gene expression profile of convalescent individuals is gradually returning to normal after they have recovered from the disease.



- There appears to be some separation between the four populations in the PCA space. This suggests that the gene expression profiles of the different populations are indeed different.
- The healthy control samples seem to be the most distinct from the other populations. This is not surprising, as we would expect healthy individuals to have different gene expression patterns than those who are sick.

- The dengue fever and dengue hemorrhagic fever samples appear to overlap somewhat in the PCA space. This suggests that the gene expression profiles of these two diseases may be more like each other than they are to the healthy control or convalescent samples.
- The convalescent samples seem to be intermediate between the dengue fever and healthy control samples. This suggests that the gene expression profile of convalescent individuals is gradually returning to normal after they have recovered from the disease.



Population Separation: The four populations – Healthy, Dengue Fever, Dengue Hemorrhagic Fever, and Convalescent – seem somewhat separated in the PCA space. This implies that their gene expression profiles differ to some extent.

Healthy Control Distinction: The Healthy control samples appear to be the most distinct from the other three populations, as expected, indicating their gene expression patterns are significantly different.

Dengue Fever and Hemorrhagic Fever Overlap: The Dengue Fever and Dengue Hemorrhagic Fever samples overlap somewhat in the PCA space. This suggests their gene expression profiles might be more like each other compared to the Healthy control or Convalescent samples.

Convalescent Samples as Intermediates: The Convalescent samples seem to occupy an intermediate space between the Dengue Fever and Healthy control samples. This could indicate that their gene expression profiles are gradually returning to normal after recovery.

- Variance Explained: The image doesn't show the variance explained by the first two principal components. Knowing this information would be helpful to understand how much of the total data variation they capture.

- Gene Loadings: The loadings of individual genes on the principal components are also missing. This information would reveal which genes contribute most to the separation between the populations.
- Supervised PCA: Performing a supervised PCA analysis could be beneficial. By incorporating the known disease states of the samples, you might identify genes specifically associated with each disease.

Conclusion

The study identifies potential biomarkers (DEGs) associated with different stages of dengue fever, shedding light on the underlying molecular mechanisms. These findings could contribute to better diagnostic or prognostic approaches in identifying and managing different stages of dengue fever.

Our analysis aims to provide valuable insights into the gene expression changes associated with dengue fever and its severe form, dengue hemorrhagic fever. By identifying potential biomarkers and exploring their biological relevance, we hope to contribute to the development of improved diagnostic and therapeutic strategies for these debilitating diseases.