Uncovering Expression Patterns in Lupus: PCA and Heatmap Visualization of SLE Gene Expression Data

Isaiah Dominguez

Objective:

To identify transcriptomic differences between SLE patients and healthy individuals using gene expression microarray data from the NCBI GEO database.

Background / Problem Statement:

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease with complex immune dysregulation. Identifying differentially expressed genes may uncover biomarkers or molecular pathways involved in disease progression. Microarray analysis allows high-throughput profiling of gene expression in patient-derived samples.

Goal: Analyze gene expression differences in peripheral blood samples using statistical modeling and visualization in R.

Dataset Overview:

Dataset: GSE10325 (Gene Expression Omnibus)

- **Samples:** 67 total (39 SLE, 28 Healthy Controls)
- **Tissue Source:** Peripheral blood (lymphocytes
- **Platform:** Affymetrix Human Genome U133A Array (GPL96)
- **Data Type:** Raw expression values

Analysis Pipeline

Tools Used:

- R (GEOquery, limma, prcomp, pheatmap)

Load Required Libraries

```
library(GEOquery)
```

```
Loading required package: Biobase
```

```
Loading required package: BiocGenerics
```

```
Attaching package: 'BiocGenerics'

The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
    table, tapply, union, unique, unsplit, which.max, which.min

Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.

Setting options('download.file.method.GEOquery'='auto')

Setting options('GEOquery.inmemory.gpl'=FALSE)

library(limma)


Attaching package: 'limma'

The following object is masked from 'package:BiocGenerics':

    plotMA

library(pheatmap)
```

Download and Prepare Gene Expression Data

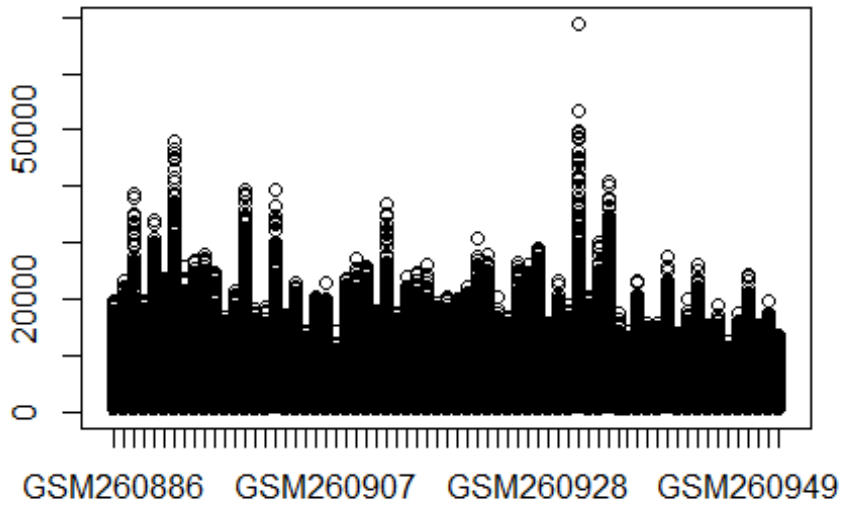
- This section retrieves the SLE gene expression dataset GSE10325 from the GEO database.
- exprs_data contains the raw gene expression matrix (genes × samples).
- pheno_data holds the sample metadata, including group labels.

Basic exploratory plots and summaries are used to inspect the raw expression data

- `boxplot()` shows the distribution of expression values across samples.
- `summary()` provides descriptive stats for the data set.

- `hist()` displays a histogram of overall expression levels to assess scale and normalization needs.

```
gse <- getGEO("GSE10325", GSEMatrix = TRUE)

Found 1 file(s)

GSE10325_series_matrix.txt.gz

exprs_data <- exprs(gse[[1]])
pheno_data <- pData(gse[[1]])
boxplot(exprs_data)
```



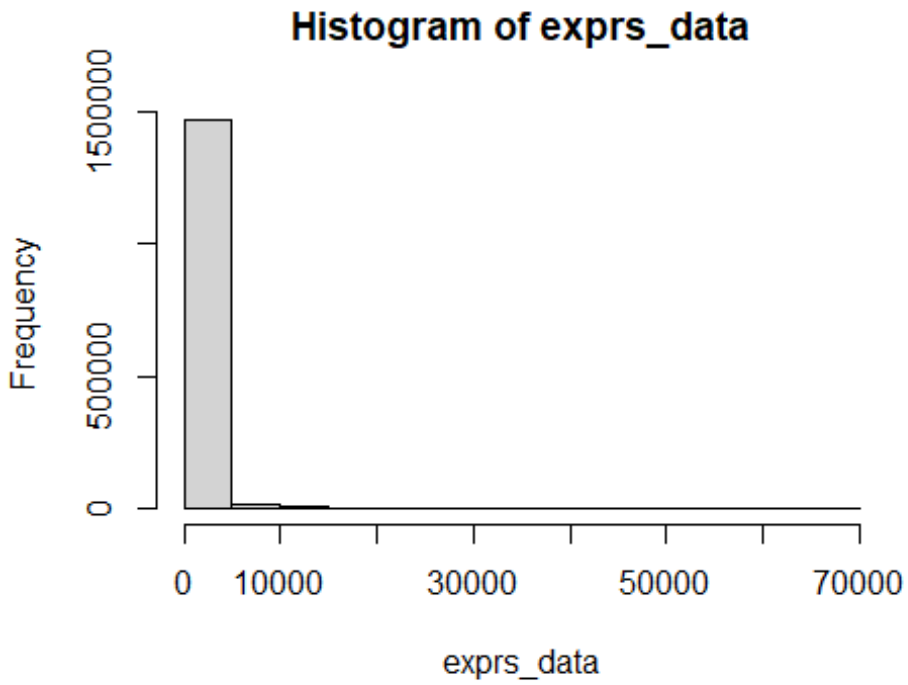

```
summary(exprs_data)

   GSM260886         GSM260887          GSM260888         GSM260889
 Min.   :    0.3   Min.   :    0.20   Min.   :    0.5   Min.   :    0.2
 1st Qu.:   23.3   1st Qu.:   24.55   1st Qu.:   29.4   1st Qu.:   30.2
 Median :   80.7   Median :   83.20   Median :  101.6   Median :   93.7
 Mean   :  401.0   Mean   :  421.11   Mean   :  458.7   Mean   :  405.0
 3rd Qu.:  290.9   3rd Qu.:  291.50   3rd Qu.:  271.4   3rd Qu.:  290.8
 Max.   :19606.0   Max.   :23358.00   Max.   :38671.0   Max.   :19822.1
   GSM260890         GSM260891          GSM260892         GSM260893
 Min.   :    0.7   Min.   :    0.2   Min.   :    0.4   Min.   :    0.1
 1st Qu.:   29.4   1st Qu.:   23.5   1st Qu.:   32.8   1st Qu.:   22.6
 Median :   96.2   Median :   81.2   Median :  108.3   Median :   78.8
 Mean   :  465.8   Mean   :  436.0   Mean   :  525.5   Mean   :  430.9
 3rd Qu.:  289.1   3rd Qu.:  287.9   3rd Qu.:  284.9   3rd Qu.:  258.6
```

```
Max.   :34014.9   Max.   :23613.7   Max.   :47850.7   Max.   :25749.3
  GSM260894         GSM260895         GSM260896         GSM260897
Min.   :    0.2   Min.   :    0.4   Min.   :    0.2   Min.   :    0.1
1st Qu.:   22.8   1st Qu.:   24.7   1st Qu.:   25.5   1st Qu.:   17.6
Median :   78.9   Median :   85.3   Median :   89.3   Median :   67.5
Mean   :  442.3   Mean   :  449.2   Mean   :  424.6   Mean   :  387.8
3rd Qu.:  290.8   3rd Qu.:  290.4   3rd Qu.:  266.5   3rd Qu.:  290.5
Max.   :26711.9   Max.   :28100.9   Max.   :24748.6   Max.   :16946.3
  GSM260898         GSM260899         GSM260900         GSM260901
Min.   :    0.3   Min.   :    0.1   Min.   :    0.2   Min.   :    0.1
1st Qu.:   20.5   1st Qu.:   38.3   1st Qu.:   19.9   1st Qu.:   18.1
Median :   72.8   Median :  139.4   Median :   72.9   Median :   69.2
Mean   :  416.3   Mean   :  773.2   Mean   :  396.5   Mean   :  397.5
3rd Qu.:  281.9   3rd Qu.:  528.4   3rd Qu.:  281.8   3rd Qu.:  278.8
Max.   :21551.2   Max.   :39391.9   Max.   :18348.2   Max.   :18454.3
  GSM260902         GSM260903         GSM260904         GSM260905
Min.   :    0.2   Min.   :    0.2   Min.   :    0.1   Min.   :    0.5
1st Qu.:   41.7   1st Qu.:   20.8   1st Qu.:   25.0   1st Qu.:   23.4
Median :  149.4   Median :   73.4   Median :   84.2   Median :   84.2
Mean   :  752.2   Mean   :  389.2   Mean   :  416.3   Mean   :  372.3
3rd Qu.:  547.9   3rd Qu.:  290.1   3rd Qu.:  295.4   3rd Qu.:  283.8
Max.   :39557.8   Max.   :17308.5   Max.   :22834.1   Max.   :14268.2
  GSM260906         GSM260907         GSM260908         GSM260909
Min.   :    0.2   Min.   :    0.2   Min.   :    0.2   Min.   :    0.1
1st Qu.:   24.9   1st Qu.:   24.2   1st Qu.:   27.3   1st Qu.:   26.0
Median :   84.2   Median :   87.0   Median :   90.6   Median :   88.6
Mean   :  404.2   Mean   :  408.7   Mean   :  356.0   Mean   :  430.4
3rd Qu.:  291.0   3rd Qu.:  294.6   3rd Qu.:  278.5   3rd Qu.:  298.9
Max.   :20335.9   Max.   :22783.3   Max.   :14275.8   Max.   :23587.0
  GSM260910         GSM260911         GSM260912         GSM260913
Min.   :    0.3   Min.   :    0.1   Min.   :    0.2   Min.   :    0.20
1st Qu.:   28.9   1st Qu.:   29.6   1st Qu.:   20.4   1st Qu.:   34.25
Median :   96.8   Median :   95.0   Median :   74.1   Median :  112.20
Mean   :  441.1   Mean   :  435.5   Mean   :  397.4   Mean   :  467.47
3rd Qu.:  298.4   3rd Qu.:  265.5   3rd Qu.:  264.5   3rd Qu.:  296.45
Max.   :27105.8   Max.   :25871.5   Max.   :18005.6   Max.   :36764.50
  GSM260914         GSM260915         GSM260916         GSM260917
Min.   :    0.2   Min.   :    0.1   Min.   :    0.5   Min.   :    0.1
1st Qu.:   25.8   1st Qu.:   27.0   1st Qu.:   35.9   1st Qu.:   23.9
Median :   85.7   Median :   89.1   Median :  115.4   Median :   83.8
Mean   :  394.9   Mean   :  425.2   Mean   :  409.9   Mean   :  432.8
3rd Qu.:  288.1   3rd Qu.:  295.2   3rd Qu.:  298.6   3rd Qu.:  288.6
Max.   :17460.2   Max.   :24087.6   Max.   :24866.1   Max.   :26124.5
  GSM260918         GSM260919         GSM260920         GSM260921
Min.   :    0.2   Min.   :    0.2   Min.   :    0.4   Min.   :    0.1
1st Qu.:   19.4   1st Qu.:   24.3   1st Qu.:   27.1   1st Qu.:   22.7
Median :   71.9   Median :   79.4   Median :   84.3   Median :   80.0
Mean   :  402.7   Mean   :  406.8   Mean   :  404.1   Mean   :  416.3
3rd Qu.:  261.9   3rd Qu.:  280.1   3rd Qu.:  276.4   3rd Qu.:  250.3
Max.   :19154.9   Max.   :20392.3   Max.   :20005.1   Max.   :22251.6
```

```
  GSM260922         GSM260923         GSM260924         GSM260925
Min.   :    0.5   Min.   :    0.1   Min.   :    0.10   Min.   :    0.1
1st Qu.:   31.1   1st Qu.:   26.4   1st Qu.:   28.45   1st Qu.:   24.8
Median :   99.5   Median :   88.5   Median :   87.60   Median :   85.7
Mean   :  452.6   Mean   :  459.0   Mean   :  397.14   Mean   :  385.1
3rd Qu.:  285.1   3rd Qu.:  285.6   3rd Qu.:  244.45   3rd Qu.:  291.4
Max.   :30878.8   Max.   :28048.2   Max.   :20378.10   Max.   :16862.9
  GSM260926         GSM260927         GSM260928         GSM260929
Min.   :    0.4   Min.   :    0.1   Min.   :    0.3   Min.   :    0.30
1st Qu.:   29.1   1st Qu.:   22.5   1st Qu.:   26.2   1st Qu.:   23.05
Median :   93.6   Median :   80.5   Median :   91.1   Median :   80.30
Mean   :  422.4   Mean   :  434.0   Mean   :  452.8   Mean   :  383.78
3rd Qu.:  293.1   3rd Qu.:  294.3   3rd Qu.:  298.4   3rd Qu.:  273.15
Max.   :26605.2   Max.   :26196.0   Max.   :28850.5   Max.   :15917.90
  GSM260930         GSM260931         GSM260932         GSM260933
Min.   :    0.2   Min.   :    0.1   Min.   :    0.5   Min.   :    0.1
1st Qu.:   21.6   1st Qu.:   23.4   1st Qu.:   43.0   1st Qu.:   19.9
Median :   78.1   Median :   82.9   Median :  153.3   Median :   72.9
Mean   :  409.7   Mean   :  393.7   Mean   :  812.1   Mean   :  409.2
3rd Qu.:  287.1   3rd Qu.:  286.8   3rd Qu.:  530.9   3rd Qu.:  283.9
Max.   :23127.7   Max.   :18953.0   Max.   :68860.6   Max.   :20392.9
  GSM260934         GSM260935         GSM260936         GSM260937
Min.   :    0.3   Min.   :    0.9   Min.   :    0.1   Min.   :    0.1
1st Qu.:   23.3   1st Qu.:   42.9   1st Qu.:   17.9   1st Qu.:   21.7
Median :   81.5   Median :  151.8   Median :   65.4   Median :   72.4
Mean   :  445.7   Mean   :  783.9   Mean   :  379.6   Mean   :  371.7
3rd Qu.:  287.6   3rd Qu.:  542.6   3rd Qu.:  282.1   3rd Qu.:  268.3
Max.   :30121.4   Max.   :40956.6   Max.   :17643.7   Max.   :14326.7
  GSM260938         GSM260939         GSM260940         GSM260941
Min.   :    0.1   Min.   :    0.2   Min.   :    0.1   Min.   :    0.2
1st Qu.:   29.8   1st Qu.:   16.9   1st Qu.:   21.0   1st Qu.:   40.0
Median :  114.3   Median :   62.4   Median :   77.6   Median :  143.8
Mean   :  702.0   Mean   :  378.2   Mean   :  375.4   Mean   :  720.1
3rd Qu.:  499.1   3rd Qu.:  281.9   3rd Qu.:  284.8   3rd Qu.:  527.8
Max.   :23384.2   Max.   :15607.4   Max.   :15590.0   Max.   :27432.5
  GSM260942         GSM260943         GSM260944         GSM260945
Min.   :    0.1   Min.   :    0.2   Min.   :    0.3   Min.   :    0.2
1st Qu.:   21.5   1st Qu.:   26.4   1st Qu.:   40.1   1st Qu.:   16.5
Median :   74.9   Median :   89.7   Median :  135.7   Median :   60.0
Mean   :  370.3   Mean   :  393.8   Mean   :  726.0   Mean   :  377.4
3rd Qu.:  284.3   3rd Qu.:  276.4   3rd Qu.:  503.2   3rd Qu.:  283.4
Max.   :13979.3   Max.   :20058.5   Max.   :26283.8   Max.   :15382.3
  GSM260946          GSM260947          GSM260948          GSM260949
Min.   :    0.20   Min.   :    0.1   Min.   :    0.3   Min.   :    0.2
1st Qu.:   18.85   1st Qu.:   15.1   1st Qu.:   25.7   1st Qu.:   27.2
Median :   68.70   Median :   59.7   Median :   82.8   Median :   90.0
Mean   :  388.43   Mean   :  359.4   Mean   :  385.9   Mean   :  431.8
3rd Qu.:  284.80   3rd Qu.:  266.4   3rd Qu.:  286.3   3rd Qu.:  282.8
Max.   :19044.60   Max.   :12557.4   Max.   :17451.5   Max.   :24511.2
  GSM260950         GSM260951         GSM260952
```

```
 Min.   :    0.1   Min.   :    0.2   Min.   :    0.3
 1st Qu.:   23.3   1st Qu.:   21.8   1st Qu.:   21.5
 Median :   81.7   Median :   75.5   Median :   69.8
 Mean   :  376.2   Mean   :  397.3   Mean   :  365.6
 3rd Qu.:  294.9   3rd Qu.:  275.9   3rd Qu.:  250.2
 Max.   :15302.5   Max.   :19738.2   Max.   :13638.5
```

```
hist(exprs_data)
```




```
pheno_data$Group <- ifelse(pheno_data$source_name_ch1 == "Peripheral blood
from individuals with systemic lupus erythematosus", "SLE", "Healthy")
table(pheno_data$Group)
```

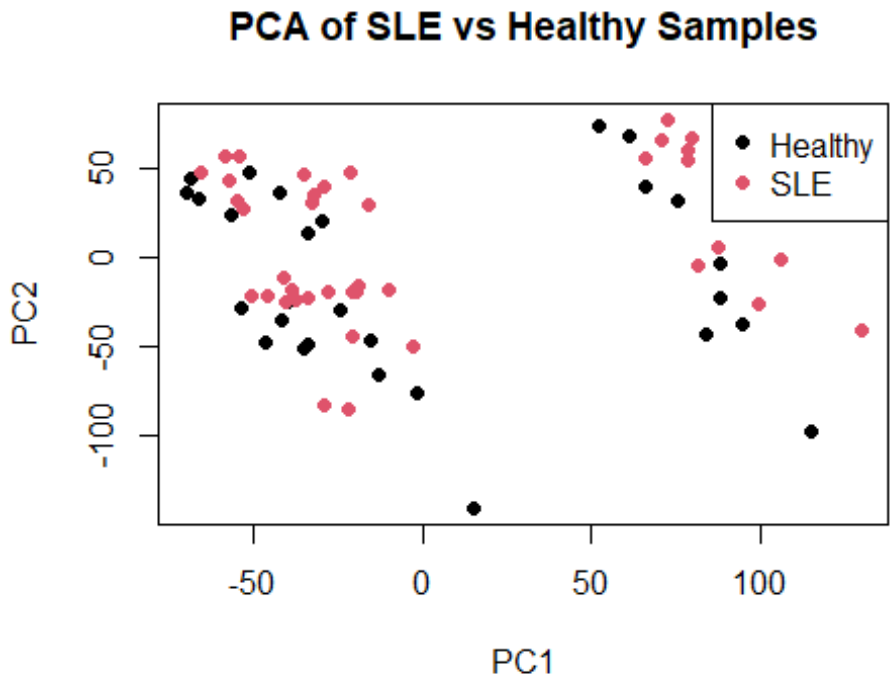
```
Healthy     SLE
     28      39
```

We applied quantile normalization to the raw expression data using the limma package. It ensures that all samples have comparable distributions, which is essential for accurate downstream analysis, especially for micro array data.

PCA is then used to reduce dimensionality and summarize the variance in gene expression across samples.

```
exprs_norm <- normalizeBetweenArrays(exprs_data, method = "quantile")
pca <- prcomp(t(exprs_norm), scale. = TRUE)
plot(pca$x[,1:2],
```

```
     col = as.factor(pheno_data$Group),
     pch = 16,
     xlab = "PC1", ylab = "PC2",
     main = "PCA of SLE vs Healthy Samples")
legend("topright", legend = unique(pheno_data$Group),
       col = 1:2, pch = 16)
```



The PCA analysis of the full gene expression matrix did not reveal a clear separation between SLE and Healthy samples along the first two principal components. Principal Component Analysis did not show clear separation between SLE and healthy samples, indicating that the major sources of variation in this dataset may not be solely attributable to disease status. This highlights the complexity of SLE biology and suggests the need to focus on specific pathways or genes rather than global patterns.

Analysis Pipeline

A variable was created to represent the sample group labels (“SLE” and “Healthy”) for use in the linear modeling process. A design matrix was then constructed without an intercept, allowing each group to have its own coefficient.

Next, a linear model was fitted to the normalized gene expression data using this design matrix. This step estimates the average expression level for each gene within both groups.

To identify differences in expression, a contrast was defined to compare the “SLE” group against the “Healthy” group. This contrast was applied to the fitted model.

Empirical Bayes moderation was then used to stabilize standard errors and enhance statistical power particularly important when working with small sample sizes.

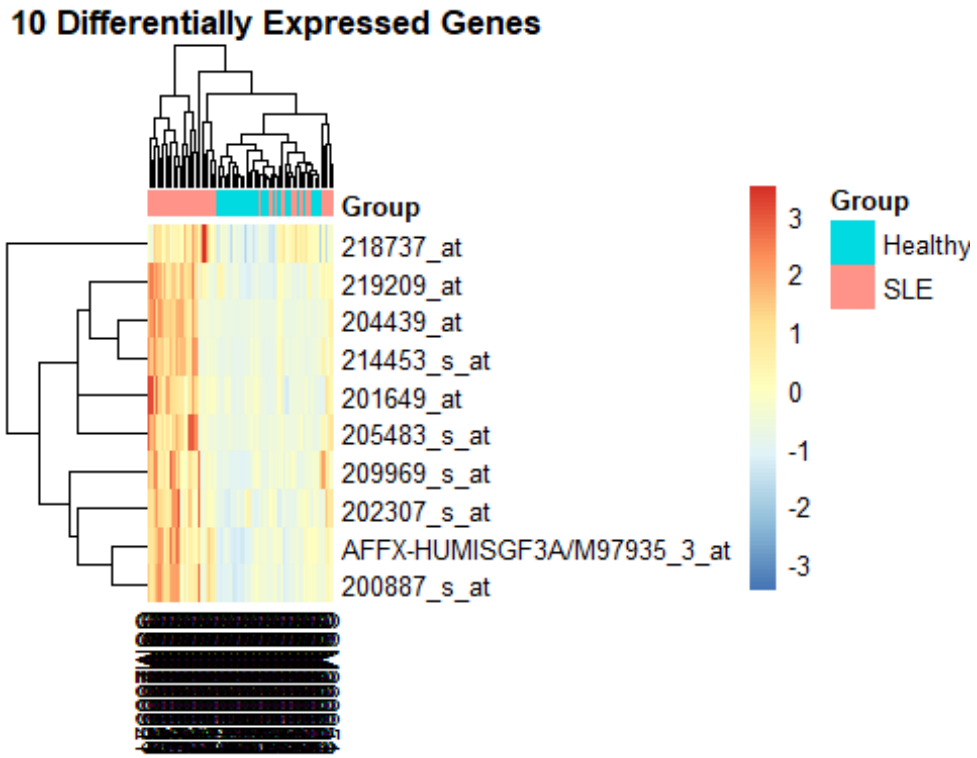
```
group <- factor(pheno_data$Group)
design <- model.matrix(~ 0 + group)
colnames(design) <- levels(group)
fit <- lmFit(exprs_norm, design)
contrast.matrix <- makeContrasts(SLEvsHealthy = SLE - Healthy, levels =
design)
fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)
```

From the results, the top 10 most differentially expressed genes were identified, ranked by statistical significance.

```
top_genes <- topTable(fit2, number = 10)
View(top_genes)
```

Finally, the normalized expression values for these top 10 genes were extracted across all samples. This subset of data was used to generate a heatmap for visualizing expression patterns.

```
top_gene_names <- rownames(top_genes)
heatmap_matrix <- exprs_norm[top_gene_names, ]
annotation_col <- data.frame(Group = pheno_data$Group)
rownames(annotation_col) <- colnames(heatmap_matrix)
pheatmap(heatmap_matrix,
         annotation_col = annotation_col,
         show_rownames = TRUE,
         scale = "row",
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         main = "Top 10 Differentially Expressed Genes")
```



Visualization:

The heat map of the top 10 differentially expressed genes demonstrates consistent expression differences between SLE and Healthy groups. While not sharply contrasted, SLE patients generally exhibit higher expression levels indicated by warmer colors. The clustering pattern supports modest but biologically relevant transcriptional changes associated with lupus.

Key Takeaways & Future Directions:

While individual gene expression differences were modest, the collective behavior of differentially expressed genes points toward immune-related dysregulation a hallmark of systemic lupus erythematosus. These results are consistent with previous studies linking SLE to abnormal immune cell signaling and chronic inflammation. PCA did not show strong global separation between SLE and healthy samples. The heatmap of the top 10 differentially expressed genes revealed subtle yet reproducible group-level differences in expression, with SLE patients tending to exhibit elevated levels across several immune-related genes.