Load Required Libraries

library(GEOquery) library(DESeq2) library(ggplot2) library(pheatmap) library(clusterProfiler) library(org.Hs.eg.db) library(enrichplot) library(umap)

1. Data Download and Import

Load GEO dataset

geo_id <- "GSE21942" gse <- getGEO(geo_id, GSEMatrix = TRUE) data <- exprs(gse[[1]]) metadata <- pData(gse[[1]])

2. Data Preprocessing

Log transformation and filtering

data <- log2(data + 1) keep <- rowSums(data > 10) > (0.5 * ncol(data)) data_filtered <- data[keep,]

Normalization using Variance Stabilizing Transformation (VST)

dds <- DESeqDataSetFromMatrix(countData = data_filtered, colData = metadata, design = ~ condition) dds <- estimateSizeFactors(dds) data_vst <- vst(dds)

Boxplot for Normalization Check

boxplot(assay(data_vst), col = rainbow(ncol(data_vst)), main = "VST Normalized Data")

3. Exploratory Data Analysis

PCA Plot

pca <- prcomp(t(assay(data_vst))) autoplot(pca, data = metadata, colour = 'condition')

Heatmap of Top 50 Variable Genes

top_var_genes <- head(order(rowVars(assay(data_vst)), decreasing = TRUE), 50)
pheatmap(assay(data_vst)[top_var_genes,], scale = "row", annotation_col = metadata)

UMAP Plot

umap_res <- umap(t(assay(data_vst))) ggplot(data.frame(UMAP1 = umap_res$layout[,1], UMAP2 = umap_res$layout[,2], Condition = metadata$condition), aes(x=UMAP1, y=UMAP2, color=Condition)) + geom_point()

4. Differential Expression Analysis

dds <- DESeq(dds) res <- results(dds, alpha = 0.05, lfcThreshold = 1) degs <- res[which(res$padj < 0.05 & abs(res$log2FoldChange) > 1),]
write.csv(as.data.frame(degs), "DEGs.csv")

5. Functional Enrichment Analysis

Convert gene symbols to Entrez IDs

gene_list <- rownames(degs) entrez_ids <- mapIds(org.Hs.eg.db, keys = gene_list, column = "ENTREZID", keytype = "SYMBOL", multiVals = "first")

GO and KEGG Enrichment Analysis

ego <- enrichGO(gene = entrez_ids, OrgDb = org.Hs.eg.db, keyType = "ENTREZID", readable = TRUE) kegg <- enrichKEGG(gene = entrez_ids, organism = 'hsa') write.csv(as.data.frame(ego), "GO_Enrichment.csv") write.csv(as.data.frame(kegg), "KEGG_Enrichment.csv")

Visualization

dotplot(ego) + ggtitle("GO Enrichment Analysis") dotplot(kegg) + ggtitle("KEGG Pathway Analysis")

6. Gene Expression Visualization

Volcano Plot

volcano <- ggplot(as.data.frame(res), aes(x = log2FoldChange, y = -log10(padj))) + geom_point(aes(color = padj < 0.05 & abs(log2FoldChange) > 1)) + theme_minimal() print(volcano)

Boxplot for Top Genes

top_genes <- rownames(head(degs[order(degs$log2FoldChange, decreasing = TRUE),], 10)) gene_expr <- assay(data_vst)[top_genes,] melted_data <- reshape2::melt(gene_expr) ggplot(melted_data, aes(x=Var2, y=value, fill=Var2)) + geom_boxplot() + theme_minimal()