# Lab 5: Mini Project

Beibei Xian

Fall 2025

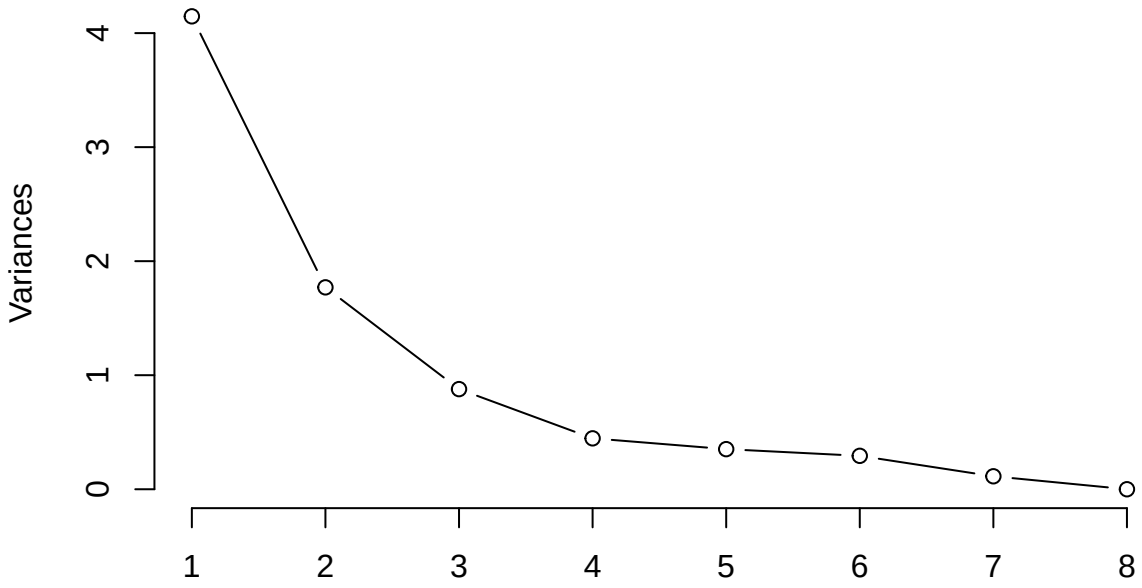
## 0. Introduction

Our group BARJJ, consisting of Dizon, Jericho V., Chen, Ronglin, Idrovo, Jose A., Shah, Ahna P., and Xian, Beibei, conducted an in-depth analysis of the growing threat of the Zika virus, a mosquito-borne illness known to cause severe birth defects such as microcephaly in unborn babies. Our primary goal was to analyze RNA-Seq data from mouse cortical development to identify windows of susceptibility, pinpointing critical stages when the brain is most vulnerable to viral infection.

Beyond microcephaly, our study extended to broader brain diseases, including cognitive disorder, mental depression, anxiety disorders, communication disorders, and migraines. By examining gene expression patterns across key developmental time points, we aimed to uncover genetic markers and periods associated with these conditions. Identifying these susceptibility windows provides insights into early-stage interventions, helping us understand not only how Zika disrupts brain development but also how genetic factors contribute to a range of brain diseases.

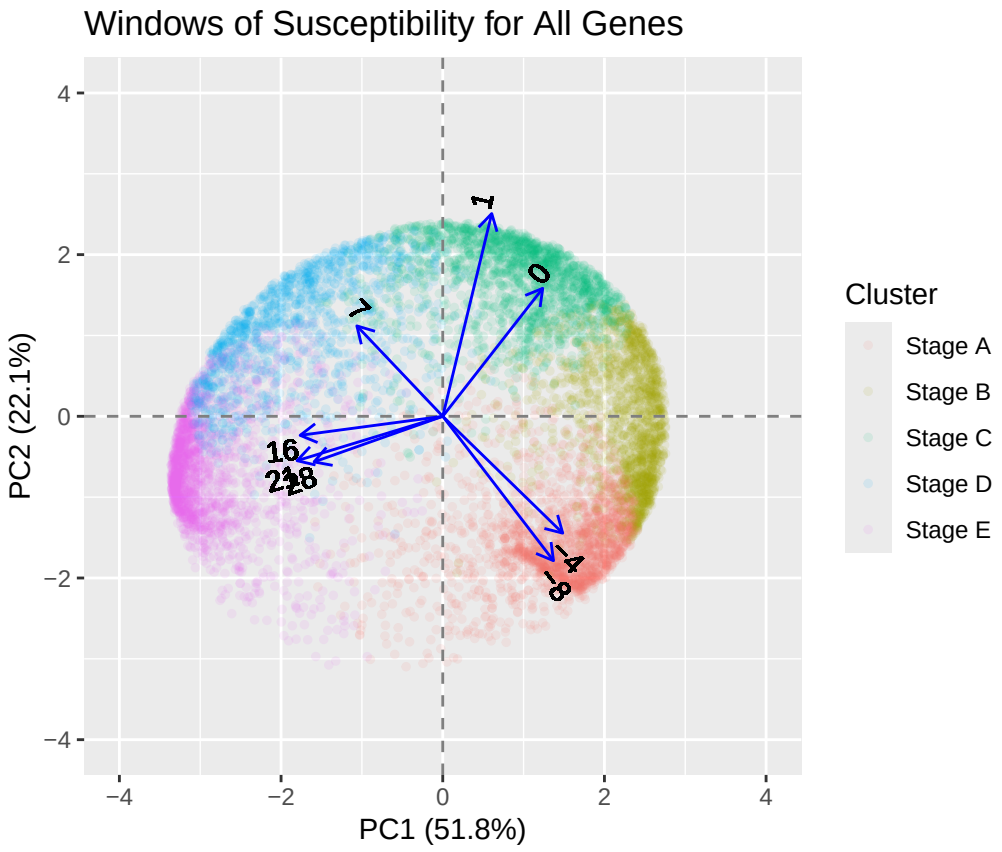
## 1. Stages of Development Analysis

```
# Load the data
mouse.df <- read.csv("~/MATP-4400/data/MouseHomologData.csv", row.names = 1)
colnames(mouse.df) <- c("-8", "-4", "0", "1", "7", "16", "21", "28")
mouse.matrix <- as.matrix(mouse.df)

# Perform PCA
mouse.pca <- prcomp(mouse.matrix, retx = TRUE, center = TRUE, scale = TRUE)

# Scree plot
screeplot(mouse.pca, type = "lines", main = 'Explained Variance of Mouse Genes')
```

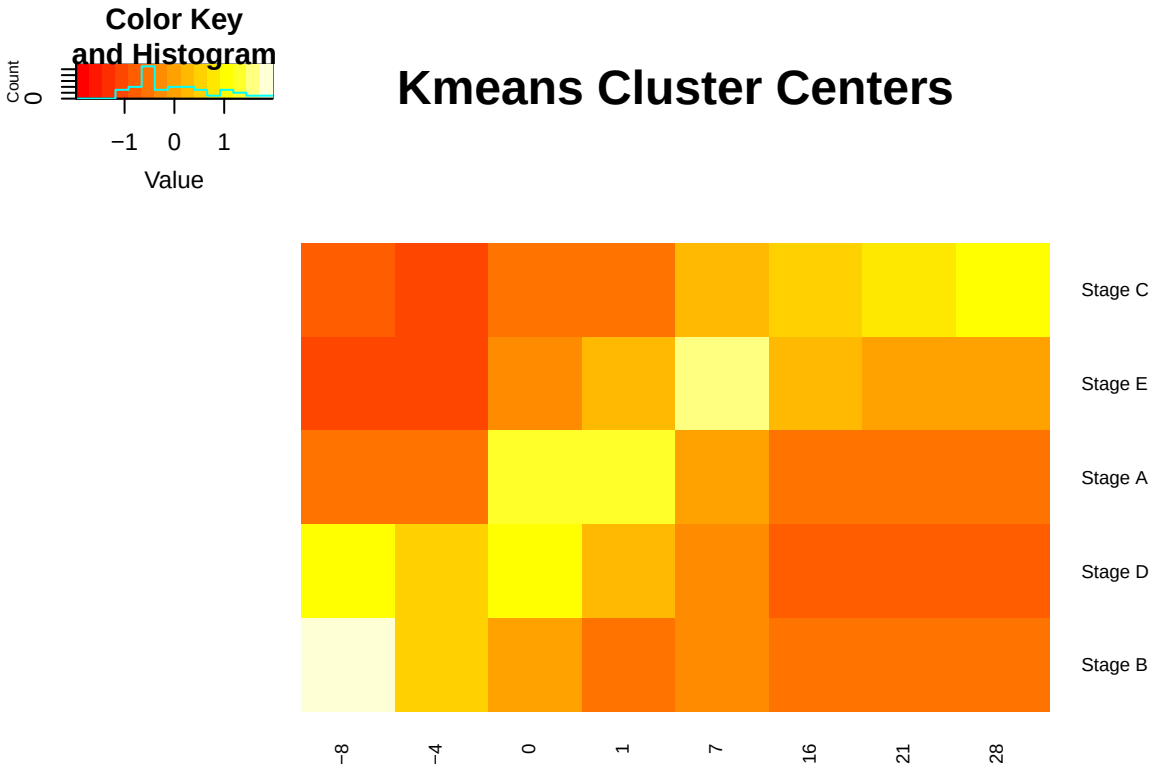



```r
# K-means clustering
set.seed(300)
mouse.km <- kmeans(mouse.df, 5)

# Rename clusters and centers
mouse.km$cluster <- factor(mouse.km$cluster,
                           levels = c("2", "4", "1", "5", "3"),
                           labels = c("Stage A", "Stage B", "Stage C", "Stage D", "Stage E"))
rownames(mouse.km$centers) <- as.factor(c("Stage A", "Stage B", "Stage C", "Stage D", "Stage E"))

# Generate biplot
generateBiplot <- function(pca, km, colnames, vector.length = 4, alpha.setting = 0.1, disease.name) {
  # Extract the scores and loadings
  scores <- as.data.frame(pca$x)
  loadings <- as.data.frame(pca$rotation)
  scores$cluster <- as.factor(km$cluster)

  t <- 1.2 * max(abs(pca$x[, 1:2]))
  eigs <- mouse.pca$sdev^2

  # Plot the biplot
  biplot.result <- ggplot(scores, aes(x = PC1, y = PC2, color = cluster)) +
    geom_point(size = 1, alpha = alpha.setting) +
    scale_color_discrete(name = "Cluster") +
    geom_hline(yintercept = 0, linetype = "dashed", color = "gray50") +
    geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
```

```
    coord_equal() +
    geom_segment(
      data = loadings,
      aes(x = 0, y = 0, xend = PC1 * vector.length, yend = PC2 * vector.length),
      arrow = arrow(length = unit(0.25, "cm")), color = "blue"
    ) +
    xlab(paste0("PC1 (", round(eigs[1] / sum(eigs) * 100, 1), "%)")) +
    ylab(paste0("PC2 (", round(eigs[2] / sum(eigs) * 100, 1), "%)")) +
    ggtitle(paste('Windows of Susceptibility for', disease.name)) +
    xlim(-t, t) +
    ylim(-t, t)  # title plot and make square
  for (i in 1:nrow(loadings)) {
    angle <- atan(loadings$PC2[i] / loadings$PC1[i])
    biplot.result <- biplot.result +
      geom_text(
        x = loadings$PC1[i] * vector.length,
        y = loadings$PC2[i] * vector.length,
        label = rownames(loadings)[i], angle = angle * 180 / pi,
        hjust = ifelse(loadings$PC1[i] < 0, 1.1, -0.1),
        vjust = ifelse(loadings$PC2[i] < 0, 1.1, -0.1),
        color = "black"
      )
  }

  biplot.result
}

# Generate the biplot
generateBiplot(pca = mouse.pca, km = mouse.km, colnames = colnames(mouse.df), disease.name = "All Genes
```

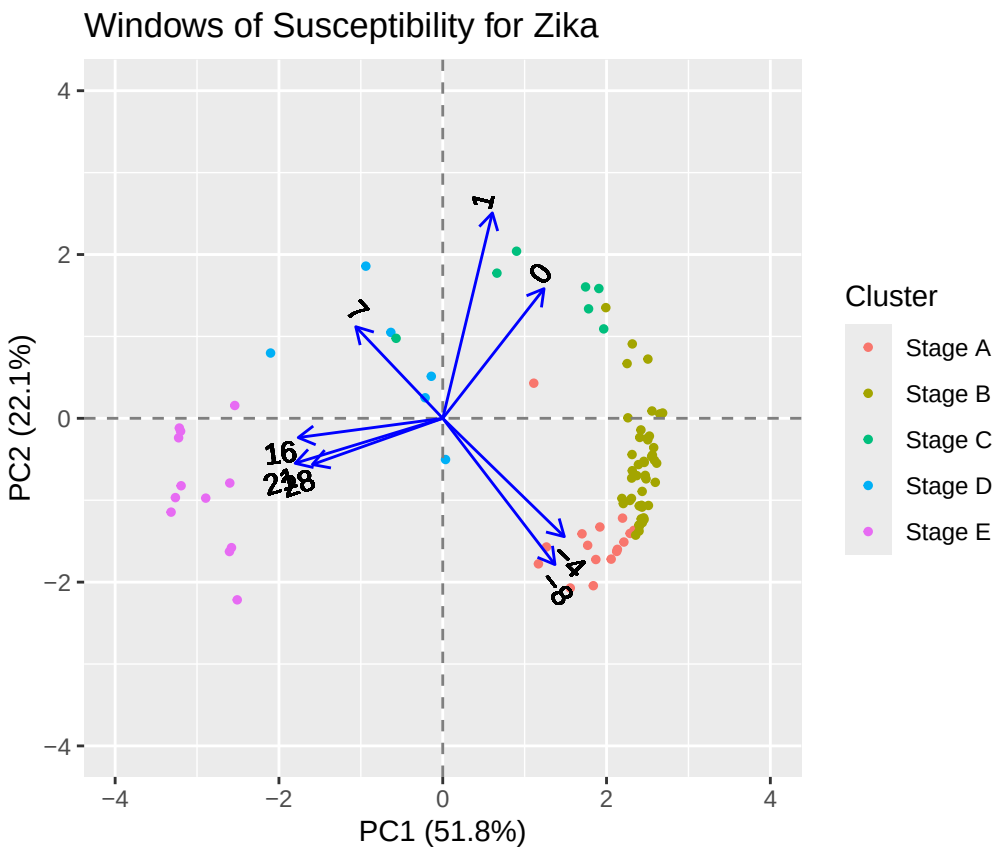



```
heatmap.2(mouse.km$centers,
          scale = "none",
          dendrogram = "none",
          Colv = FALSE,
          cexCol = 0.75,
          cexRow = 0.75,
          main = "Kmeans Cluster Centers",
          trace = "none")
```




## 2. Windows of Susceptibility Analysis of Zika

```
disease.df <- read.csv(
  "~/MATP-4400/data/Zikamicrocephaly_data.csv",
  row.names = 1
)
pca.zika <- mouse.pca
pca.zika$x <- mouse.pca$x[rownames(mouse.pca$x) %in% disease.df$symbol, ]
matrix.zika <- as.matrix(disease.df)

# Set up kmeans
km.zika <- mouse.km
km.zika$cluster <- km.zika$cluster[rownames(pca.zika$x)]

# Generate the biplot
generateBiplot(
  pca = pca.zika,
  km = km.zika,
  colnames = colnames(Mouse.df),
  alpha.setting = 1,
  disease.name = "Zika"
)
```




```r
# Define cluster_pvals;
cluster_pvals <- function(k, km, myplot.df) {

  # Set the p-value and logodds to 0
  pvalue   <- zeros(k, 1)
  logodds  <- zeros(k, 1)
  results  <- cbind.data.frame(cluster = 1:k, pvalue, logodds)
  classdisease <- zeros(k, 1)
  classall     <- as.vector(table(km$cluster))

  temp <- myplot.df %>%
    dplyr::group_by(cluster) %>%
    dplyr::count(name = "freq")  # Creates 'freq' column

  classdisease[temp$cluster] <- temp$freq
  classlogodds   <- zeros(k, 2)
  totaldisease   <- sum(classdisease)
  totalall       <- sum(classall)

  # Calculate the log odds ratio for the disease
  for (i in 1:k) {
    n11 <- classdisease[i] + 1        # genes in disease in cluster i
    n21 <- totaldisease - classdisease[i] + 1
    n12 <- classall[i] - n11 + 1
    n22 <- totalall - n11 - n12 + 1
```

```
    res <- fisher.test(matrix(c(n11, n21, n12, n22), 2, 2))
    results[i, ]$pvalue <- res$p.value
    results[i, ]$logodds <- log((n11 * n22) / (n12 * n21))
  }

  return(results)
}
disease_symbols <- intersect(
  as.character(disease.df$symbol),
  as.character(rownames(mouse.df))
)
plot.df  <- cbind.data.frame(
  mouse.pca$x,
  cluster = as.factor(mouse.km$cluster)
)
myplot.df <- plot.df[disease_symbols, ]

# Apply cluster_pvals
clusters <- cluster_pvals(5, mouse.km, myplot.df)
threshold <- 0.1

# Evaluate across our results; create a new column
clusters <- clusters %>%
  mutate(enriched = if_else(pvalue < threshold & logodds > 0, TRUE, FALSE))

# View results
kable(clusters)
```

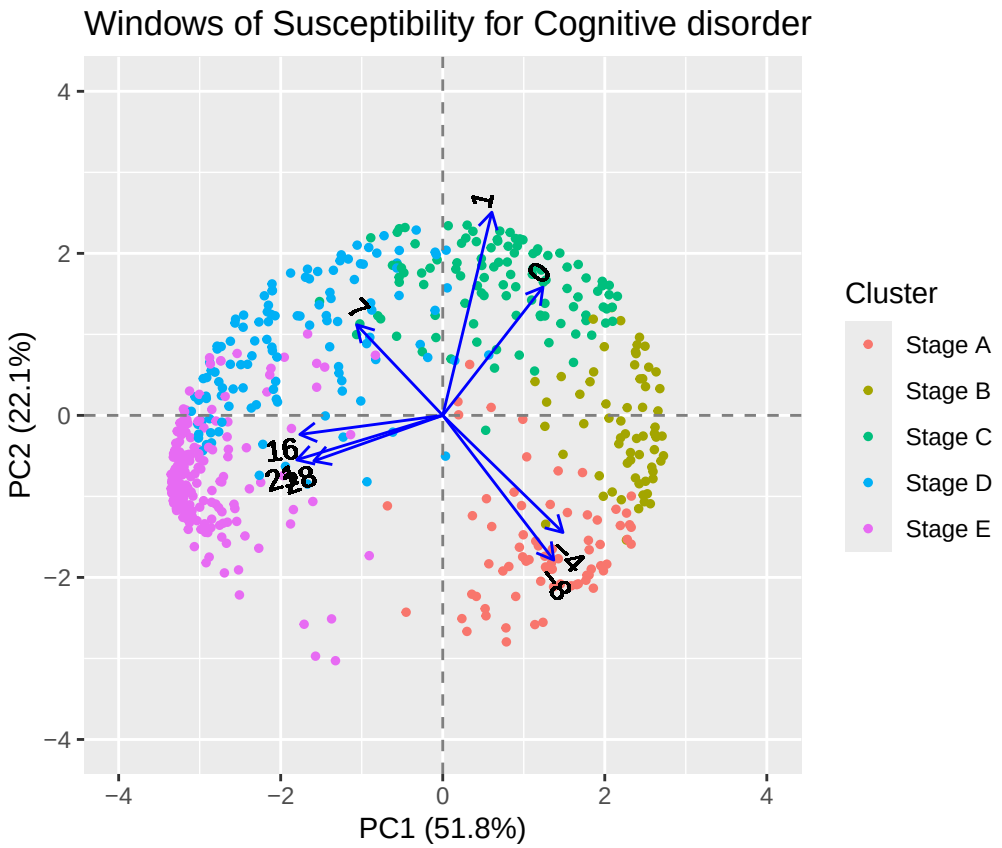
| cluster | pvalue | logodds | enriched |
|---|---|---|---|
| 1 | 0.7027408 | -0.1431217 | FALSE |
| 2 | 0.0000000 | 1.6060558 | TRUE |
| 3 | 0.0026544 | -1.0162582 | FALSE |
| 4 | 0.0402052 | -0.7776957 | FALSE |
| 5 | 0.0954559 | -0.5144924 | FALSE |

The analysis reveals that Cluster 2 is enriched for genes associated with the Zika virus, as indicated by a p-value of 0 and a positive log odds ratio. This suggests that Cluster 2 represents a developmental stage in mice that is particularly susceptible to Zika-induced changes. In contrast, the other clusters (1, 3, 4, and 5) do not show significant enrichment, as their p-values exceed the threshold of 0.1 or their log odds ratios are not positive. These findings implies that Cluster 2 as a critical window of susceptibility, providing valuable insights for further research into Zika-induced microcephaly and potential interventions.

## 3. Windows of Susceptibility Analysis of *Cognitive Disorder*

```
# Loadthe dataset and create the matrix
cd_data <- read.csv(
  "Cognitive disorder_heat_map_data.csv",
  row.names = 1
)

# Identify symbols
```

```r
cd_symbols <- intersect(
  as.character(cd_data$symbol),
  as.character(rownames(mouse.df))
)

# Show the cognitive disorder genes on the cluster biplot
cd_pca <- mouse.pca
cd_pca$x <- mouse.pca$x[
  rownames(mouse.pca$x) %in% cd_data$symbol,
]
cd_matrix <- as.matrix(cd_data)

# Set up k-means
cd_km <- mouse.km
cd_km$cluster <- cd_km$cluster[rownames(cd_pca$x)]

# Generate a biplot
generateBiplot(
  pca          = cd_pca,
  km           = cd_km,
  colnames     = colnames(mouse.df),
  alpha.setting = 1,
  disease.name = "Cognitive disorder"
)
```



```r
cd_plot_df <- cbind.data.frame(
  mouse.pca$x,
  cluster = as.factor(mouse.km$cluster)
)

# Subset only the rows with cd_symbols
cd_myplot_df <- cd_plot_df[cd_symbols, ]

# Apply a cluster-p-values function
cd_clusters <- cluster_pvals(5, mouse.km, cd_myplot_df)

cutoff <- 0.1

# Evaluate across results and create a new 'enriched' column
cd_clusters <- cd_clusters %>%
  mutate(enriched = if_else(
    pvalue <= cutoff & logodds > 0,
    TRUE,
    FALSE
  ))

kable(cd_clusters)
```

| cluster | pvalue | logodds | enriched |
|---|---|---|---|
| 1 | 0.0000006 | -0.5966180 | FALSE |
| 2 | 0.0000595 | -0.5207760 | FALSE |
| 3 | 0.4828448 | -0.0835043 | FALSE |
| 4 | 0.0000305 | 0.4661302 | TRUE |
| 5 | 0.0000000 | 0.7912383 | TRUE |

According to the table, we see that clusters 4 and 5 are enriched: both have p-values below 0.1 and positive log-odds, indicating a potential association with cognitive disorder. By contrast, clusters 1, 2, and 3 are not enriched because they either have p-values above 0.1 or negative log-odds (or both).

In the human brain, the enriched cluster is linked to the Upper Layers. In the mouse brain, by contrast, these enriched clusters appear during Neuroectoderm. Although different species reach comparable developmental milestones in distinct ways, this analysis suggests some overlap in how certain gene clusters may influence cognition across mammals.

Cognitive disorders encompass a variety of conditions that affect mental processes such as learning, problem-solving, memory, or decision-making. These challenges can impair adaptive functioning, which includes tasks like communication and independent living.

## 4. Analysis of Genes in *Cognitive Disorder*

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(reshape2)

# Load the dataset
cd_data <- read.csv("Cognitive disorder_heat_map_data.csv", row.names = 1)
```

```r
# Compute mean expression for each gene
cd_data$mean_expression <- rowMeans(cd_data[, 3:ncol(cd_data)])

# Identify one gene far from 0 and one closer to 0
high_variation_gene <- cd_data[which.max(abs(cd_data$mean_expression)), "symbol"]
low_variation_gene <- cd_data[which.min(abs(cd_data$mean_expression)), "symbol"]

# Store selected genes
selected_genes <- c(high_variation_gene, low_variation_gene)
print(selected_genes)
```

```
## [1] "VLDLR" "CAT"
```

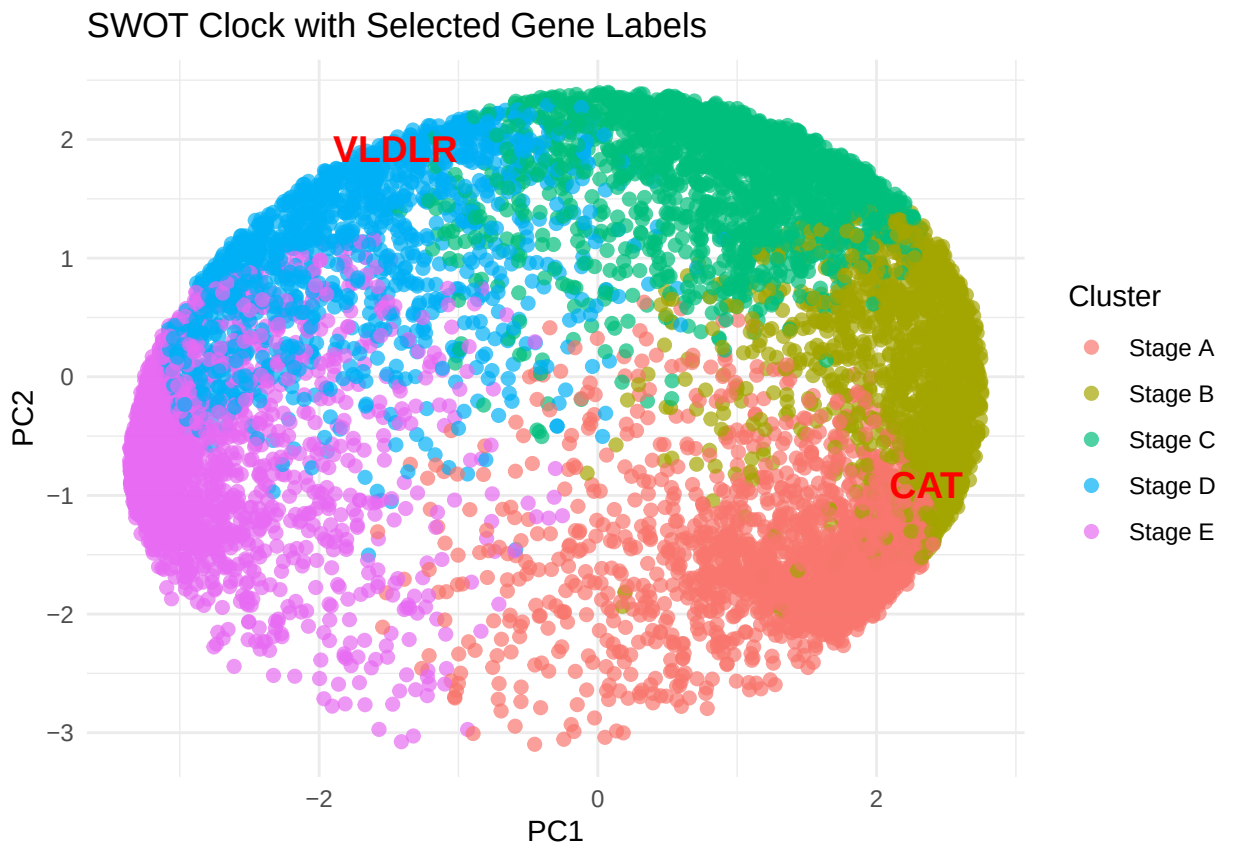
```r
# Extract expression data for selected genes
selected_gene_data <- cd_data[cd_data$symbol %in% selected_genes, ]
time_points <- colnames(cd_data)[3:(ncol(cd_data) - 1)]

# Reshape data for ggplot
melted_gene_data <- melt(selected_gene_data, id.vars = "symbol", measure.vars = time_points)

# Plot expression levels over time
ggplot(melted_gene_data, aes(x = variable, y = value, color = symbol, group = symbol)) +
  geom_line(linewidth = 1) +
  geom_point(linewidth = 2) +
  labs(title = "Expression Levels of Selected Genes Over Time",
       x = "Days",
       y = "Expression Level",
       color = "Gene") +
  theme_minimal()
```

```
## Warning in geom_point(linewidth = 2): Ignoring unknown parameters: ‘linewidth‘
```


```r
cluster_means <- cd_data %>%
  group_by(cluster) %>%
  summarise(across(starts_with("d"), mean, .names = "mean_{.col}")) %>%
  ungroup()

# Convert selected_gene_clusters to a named vector
selected_gene_clusters <- setNames(selected_gene_data$cluster, selected_gene_data$symbol)

# Compute Euclidean distance for each gene from its cluster mean
gene_distances <- sapply(selected_genes, function(gene) {

  # Extract the expression values for the gene
  gene_expression <- as.numeric(selected_gene_data[selected_gene_data$symbol == gene, time_points])

  # Extract the cluster mean for the corresponding cluster
  cluster_number <- as.character(selected_gene_clusters[gene])
  cluster_mean <- as.numeric(cluster_means[cluster_means$cluster == cluster_number, -1])  # Exclude the

  # Compute Euclidean distance
  sqrt(sum((gene_expression - cluster_mean)^2, na.rm = TRUE))  # Handle any potential NA values safely
})

# Convert to data frame for display
gene_distances_df <- data.frame(Gene = selected_genes, Euclidean_Distance = gene_distances)
print(gene_distances_df)
```

```
##        Gene Euclidean_Distance
## VLDLR VLDLR           2.018766
## CAT     CAT           0.907127
```

```r
library(ggplot2)
library(ggrepel)

# Create a dataframe for PCA results with clusters
swot_biplot_df <- data.frame(
  PC1 = mouse.pca$x[,1],
  PC2 = mouse.pca$x[,2],
  Gene = rownames(mouse.pca$x),
  Cluster = as.factor(mouse.km$cluster)
)

# highlight selected genes
swot_biplot_df$Highlight <- ifelse(swot_biplot_df$Gene %in% selected_genes, "Highlighted", "Normal")

# biplot
ggplot(swot_biplot_df, aes(x = PC1, y = PC2, color = Cluster)) +
  geom_point(size = 2, alpha = 0.7) +  # Plot all genes
  geom_text_repel(data = subset(swot_biplot_df, Gene %in% selected_genes),
                  aes(label = Gene), size = 5, color = "red", fontface = "bold") +  # Label selected ge
  labs(title = "SWOT Clock with Selected Gene Labels",
       x = "PC1",
       y = "PC2") +
  theme_minimal() +
  theme(legend.position = "right")
```




Discussion of Results

The gene expression analysis of VLDLR and CAT shows distinct patterns over time. CAT exhibits a steady increase, peaking at d63, which suggests its involvement in later stages of brain development. In contrast, VLDLR fluctuates, with notable peaks around d49 and d77, indicating it may play a role at multiple developmental stages. These trends provide insight into how each gene contributes to cognitive processes.

Comparing these genes to their cluster means highlights key differences. CAT, with a lower Euclidean distance, follows the overall cluster trend, suggesting it plays a general role in brain development. VLDLR, which is further from the mean, likely has a more specialized function. Its deviation may indicate a stronger link to disease-related processes, making it a potential marker for neurological disorders.

My prompt: "You are a scientific researcher, explain step by step the role of VLDLR in cognitive disorders. Discuss its function in brain development and its connection to cognitive disorder. For CAT, Describe the role of CAT in cognitive disorders, including its function in cognitive disorder."

LLM's Answer:VLDLR (Very Low-Density Lipoprotein Receptor) plays a key role in neuronal migration through the Reelin signaling pathway, which is essential for proper brain development. It helps form the cerebral cortex and cerebellum, supporting learning and memory. Mutations in VLDLR are linked to cerebellar hypoplasia and lissencephaly, causing intellectual disabilities, motor impairments, and coordination issues. Additionally, dysregulated VLDLR signaling has been implicated in Alzheimer's disease, where impaired neuronal connectivity contributes to cognitive decline.

Catalase (CAT) is a key antioxidant enzyme that protects neurons from oxidative stress by breaking down hydrogen peroxide. Low CAT activity is associated with neurodegenerative diseases like Alzheimer's and Parkinson's, where oxidative damage leads to memory loss and impaired cognitive function. Reduced CAT levels are also linked to schizophrenia and bipolar disorder, highlighting its role in brain health. Enhancing CAT activity through antioxidants or gene therapy may help protect against cognitive decline.

My own words:Both VLDLR and CAT are crucial for brain function. VLDLR is essential for neuronal

migration and synaptic function, playing a key role in brain development through the Reelin signaling pathway. Mutations in VLDLR can lead to cognitive disorders like lissencephaly, which cause intellectual and motor impairments. CAT, on the other hand, is a vital antioxidant enzyme that protects neurons from oxidative damage by breaking down hydrogen peroxide. Deficiency in CAT contributes to cognitive impairment in neurodegenerative diseases such as Alzheimer's and Parkinson's, where oxidative stress accelerates neuronal damage and cognitive decline.

## 5. Creative Analysis

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(dplyr)
ppi_data <- read.csv("Cognitive disorder_connection_data.csv", header = FALSE)
# Load heatmap data
heatmap_data <- read.csv("Cognitive disorder_heat_map_data.csv")

gene_mapping <- setNames(heatmap_data$symbol, heatmap_data$index)

# Mapping using mapvalues()
ppi_data <- ppi_data %>%
  mutate(V1 = mapvalues(V1, from = names(gene_mapping), to = gene_mapping, warn_missing = FALSE),
         V2 = mapvalues(V2, from = names(gene_mapping), to = gene_mapping, warn_missing = FALSE)) %>%
  filter(!is.na(V1) & !is.na(V2))

# Name columns
colnames(ppi_data) <- c("Gene1", "Gene2")
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
## The following object is masked from 'package:base':
##
```

```
##     union
```

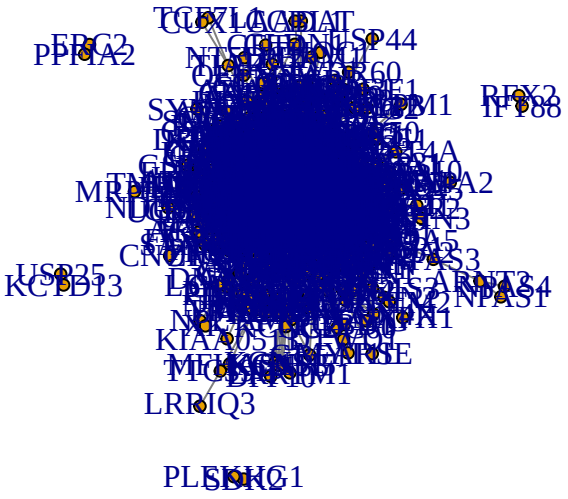
```
str(ppi_data)  # Check structure
```

```
## 'data.frame':    4572 obs. of  2 variables:
##  $ Gene1: chr  "AADAT" "AADAT" "ACTB" "ACTB" ...
##  $ Gene2: chr  "CCBL1" "KMO" "ACTB" "AMY1A" ...
```

```
head(ppi_data) # Preview data
```

```
##   Gene1 Gene2
## 1 AADAT CCBL1
## 2 AADAT   KMO
## 3  ACTB  ACTB
## 4  ACTB AMY1A
## 5  ACTB CASP3
## 6  ACTB  CYCS
```

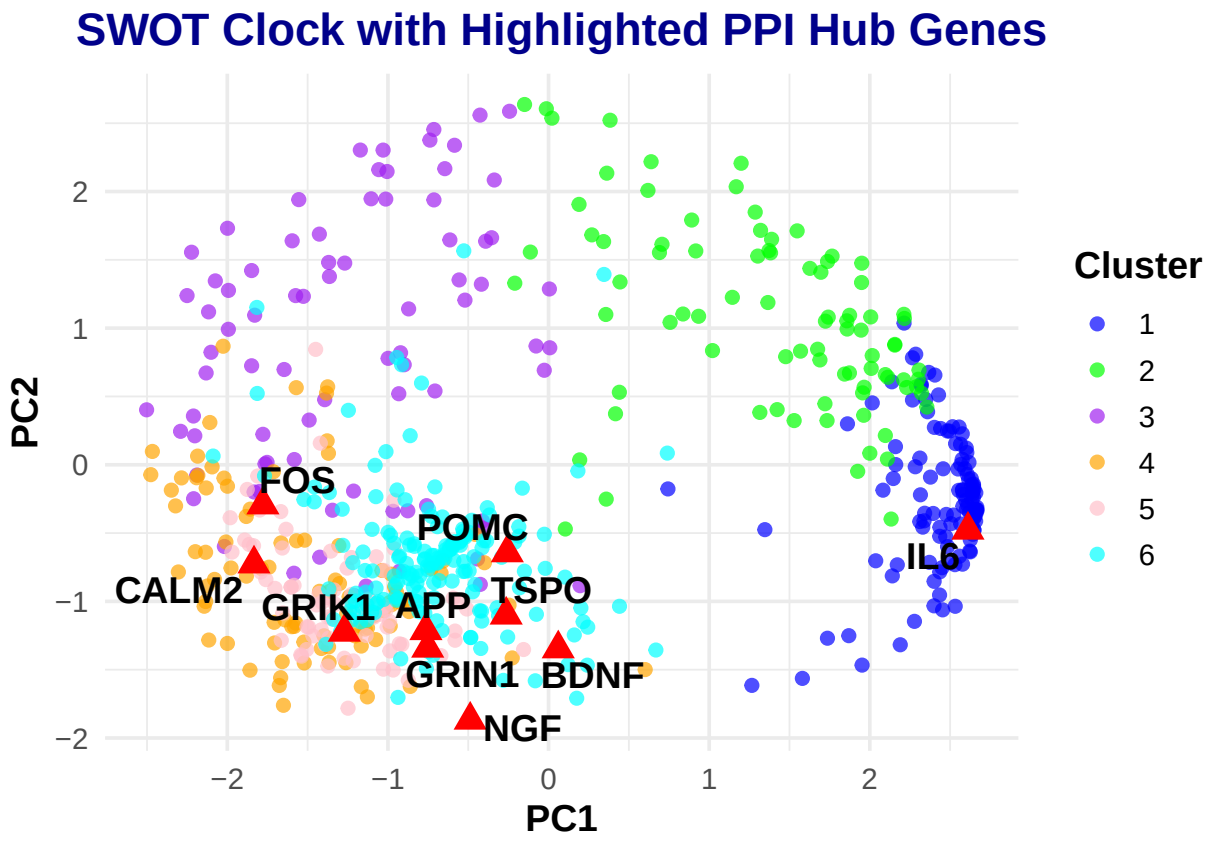
```
colnames(ppi_data) <- c("Gene1", "Gene2")

ppi_data <- ppi_data %>% filter(Gene1 != Gene2)

# Create an interaction graph
ppi_graph <- graph_from_data_frame(ppi_data, directed = FALSE)

# Plot the PPI network
plot(ppi_graph, vertex.size = 5, edge.color = "gray50", main = "Protein-Protein Interaction Network")
```



```r
# Find top 10 hub genes (most connections)
hub_genes <- names(sort(degree(ppi_graph), decreasing = TRUE)[1:10])

# Extract SWOT data for hub genes
swot_biplot_df <- heatmap_data %>% filter(symbol %in% hub_genes)

# Print top hub genes
print(hub_genes)
```

```
##  [1] "TSPO"  "CALM2" "FOS"   "BDNF"  "GRIN1" "APP"   "IL6"   "GRIK1" "NGF"
## [10] "POMC"
```

```r
library(ggplot2)
library(ggrepel)

# Define a color palette
color_palette <- c("blue", "green", "purple", "orange", "pink", "cyan", "yellow")

# Generate SWOT Clock with enhanced colors
ggplot(heatmap_data, aes(x = d0, y = d7, color = factor(cluster))) +
  geom_point(alpha = 0.7, size = 2) +  # Base points with cluster colors
  scale_color_manual(values = color_palette) +  # Assign colors
  geom_point(data = swot_biplot_df, aes(x = d0, y = d7), color = "red", size = 4, shape = 17) +  # Hub
  geom_text_repel(data = swot_biplot_df, aes(label = symbol), color = "black", size = 5, fontface = "bo
  labs(title = "SWOT Clock with Highlighted PPI Hub Genes",
       x = "PC1",
       y = "PC2",
       color = "Cluster") +
  theme_minimal(base_size = 14) +  # Clean theme with larger font
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5, color = "darkblue", face = "bold"),
        axis.title = element_text(face = "bold"),
        legend.title = element_text(face = "bold"))
```




This analysis integrated Protein-Protein Interaction (PPI) data with the SWOT Clock to highlight key hub genes associated with cognitive disorders. The PPI network visualization revealed a dense interaction map, where highly connected genes were difficult to distinguish. By filtering for the top hub genes and mapping them onto the SWOT Clock, we identified key regulatory genes, such as BDNF, GRIN1, and NGF, that play crucial roles in neuronal signaling and synaptic plasticity.

The SWOT Clock provided a clearer spatial distribution of these hub genes across different developmental clusters, emphasizing their potential windows of susceptibility. Notably, genes such as IL6 appeared in distinct clusters, suggesting their involvement in specific developmental stages or neuroinflammatory responses.

## 6. Conclusions

This study identified key windows of susceptibility in mouse cortical development to Zika-induced microcephaly and cognitive disorders. Through PCA and k-means clustering, five developmental stages were classified, with Cluster 2 showing significant enrichment for genes associated with Zika infection. This suggests that this stage is particularly vulnerable to viral disruptions, marking a critical period for potential interventions. In contrast, clusters 4 and 5 were linked to cognitive disorders, indicating distinct gene expression patterns associated with neurodevelopmental impairments.

Further analysis of specific genes, such as VLDLR and CAT, highlighted their distinct roles in brain development and disease susceptibility. VLDLR is crucial for neuronal migration, with mutations linked to lissencephaly and intellectual disabilities, while CAT protects neurons from oxidative stress, with deficiencies associated with Alzheimer's and Parkinson's disease. Additionally, integrating Protein-Protein Interaction (PPI) networks with the SWOT Clock identified key regulatory genes like BDNF, GRIN1, and NGF, emphasizing their roles in synaptic plasticity and neuronal signaling.