Lab 5: Mini Project

Manal Hajjam

2/17/2020

0. Introduction

Name: Manal Hajjam Group Members: Sarah DeCelie, Michael Landolfi, Brandyn Sigouin, Tracy Chen

The goal of this project to do a PCA analysis of different diseases such as Zikamicrocephaly (Zika disease) and a disease I chose to study about, Craniosynostosis. In addition, I created my own visual analysis to compare the transcriptome levels of each gene that is affected by the disease and how ti fluxiates throughtought the days after birth.

1. Stages of Development Analysis

```
# Read in the data and create a dataframe
Mouse.df <-read.csv("~/MATP-4400/data/MouseHomologData.csv", row.names = 1)
# Create a matrix for our analysis

#take out first column

Mouse.matrix <- as.matrix(Mouse.df)
#Mouse.matrix <- Mouse.matrix[,-1]

# Summarize; note the scaling
summary(Mouse.df)
```

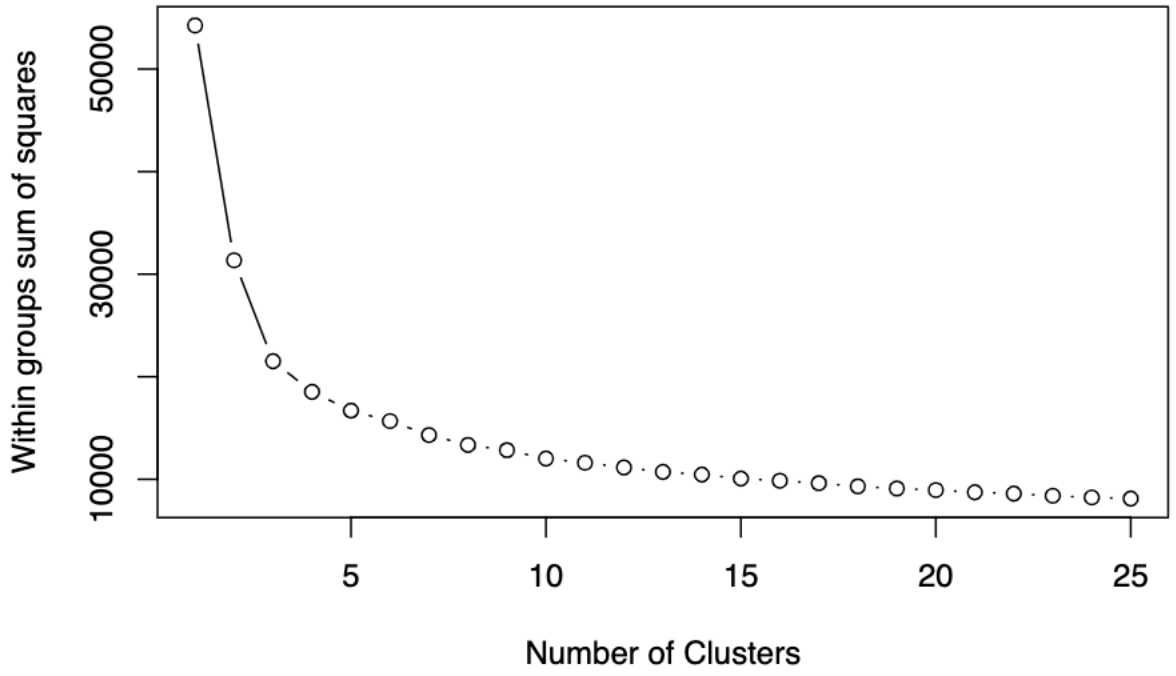
```
##     DayNeg8           DayNeg4             Day0            DayPos1
##  Min.   :-2.2436   Min.   :-2.3882   Min.   :-2.3330   Min.   :-2.02584
##  1st Qu.:-0.9138   1st Qu.:-1.0365   1st Qu.:-0.5185   1st Qu.:-0.52867
##  Median :-0.1530   Median :-0.5435   Median : 0.1894   Median :-0.07383
##  Mean   : 0.2260   Mean   :-0.3335   Mean   : 0.3118   Mean   : 0.13705
##  3rd Qu.: 1.4918   3rd Qu.: 0.3735   3rd Qu.: 1.0578   3rd Qu.: 0.76547
##  Max.   : 2.4749   Max.   : 2.4710   Max.   : 2.4739   Max.   : 2.47368
##     DayPos7           DayPos16          DayPos21          DayPos28
##  Min.   :-1.58200   Min.   :-1.7886   Min.   :-1.7495   Min.   :-2.04906
##  1st Qu.:-0.44628   1st Qu.:-0.7918   1st Qu.:-0.7839   1st Qu.:-0.75671
##  Median :-0.04786   Median :-0.4784   Median :-0.4840   Median :-0.44015
##  Mean   : 0.18491   Mean   :-0.2320   Mean   :-0.1980   Mean   :-0.09629
##  3rd Qu.: 0.71298   3rd Qu.: 0.3390   3rd Qu.: 0.4180   3rd Qu.: 0.51383
##  Max.   : 2.45886   Max.   : 2.4749   Max.   : 2.4749   Max.   : 2.47487
```

```
# Demonstrate the scaling by viewing the norm
norm(rowMeans(Mouse.matrix))
```

```
## [1] 6.127969e-08
```

```r
wssplot <- function(data, nc=25, seed=20){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")}

wssplot(Mouse.matrix, nc=25)
```

```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
```

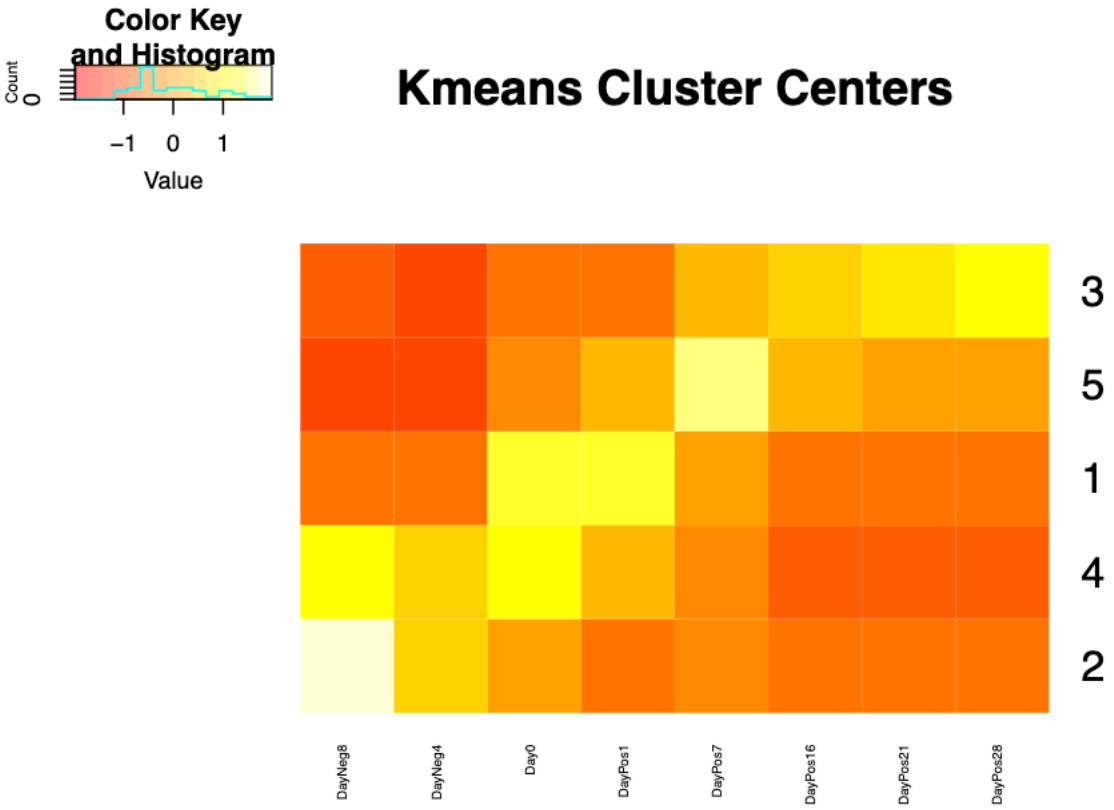



```r
# Calculate the PCA
my.pca <- prcomp(Mouse.matrix, retx=TRUE, center=TRUE, scale=TRUE)
# Summarize, to see the complete PCA result
summary(my.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0364 1.3306 0.9370 0.66801 0.59296 0.54132 0.33712
## Proportion of Variance 0.5184 0.2213 0.1098 0.05578 0.04395 0.03663 0.01421
## Cumulative Proportion  0.5184 0.7397 0.8494 0.90521 0.94917 0.98579 1.00000
##                           PC8
## Standard deviation     3.306e-11
## Proportion of Variance 0.000e+00
## Cumulative Proportion  1.000e+00
```

Looking at the graph above we can see that there is an elbow in the graph at 5. This means that we should have 5 clusters in our data.

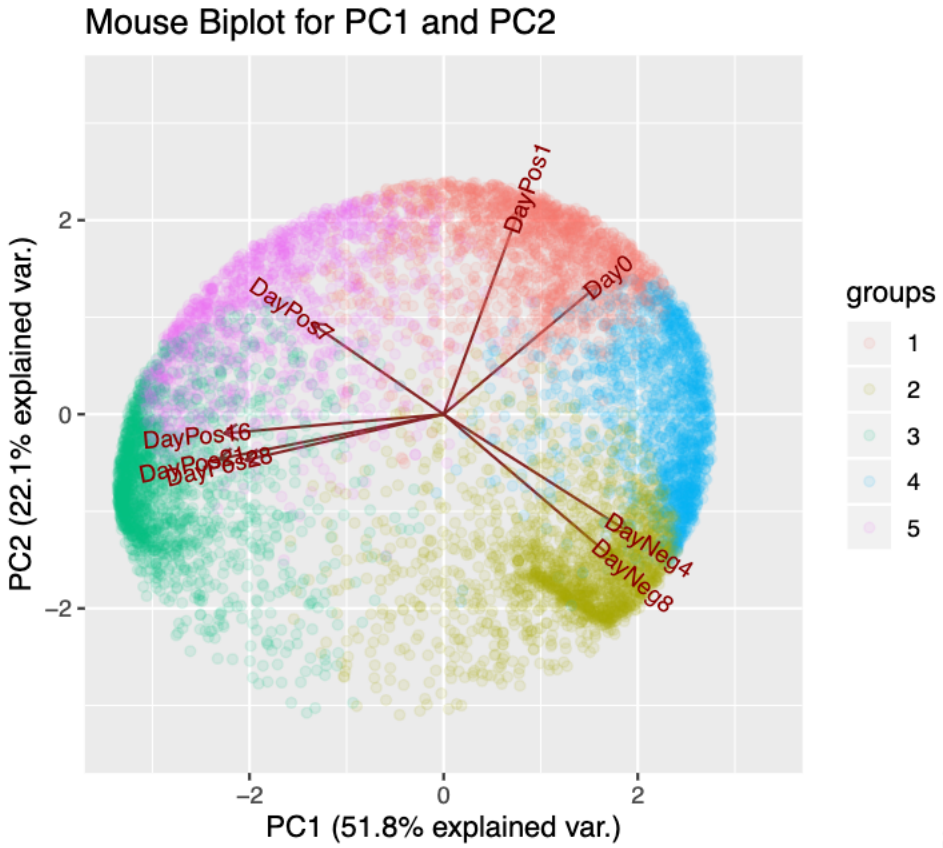
```
set.seed(300)
km <-kmeans(Mouse.matrix, 5)

heatmap.2(km$centers,
          scale = "none",
          dendrogram = "none",
          Colv=FALSE,
          cexCol=0.5,
          main = "Kmeans Cluster Centers",
          trace ="none")
```



Day -8: Cluster 2 Day 0: Cluster 1 Day 0: Cluster 4 Day 7: Cluster 5 Day 28: Cluster 3

```
# Calculate x and y scale limits for the biplot
t<-max(abs(my.pca$x[,1:2]))
# Generate the biplot using ggbiplot
p <- ggbiplot(my.pca,
              choices=c(1,2),
              alpha=.1,
              varname.adjust=0.5,
              obs.scale = 1,
              groups=as.factor(km$cluster))
p + ggtitle('Mouse Biplot for PC1 and PC2') + xlim(-t,t) + ylim(-t,t)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Stage A: 2 Stage B: 4 Stage C: 1 Stage D: 5 Stage E: 3

2. Windows of Susceptibility Analysis of Zika

```
# Read in the dataset; create the matrix

zika.df <-read.csv("~/MATP-4400/data/Zikamicrocephaly_data.csv",row.names = 1)
zika_symbols <- intersect(as.character(zika.df$symbol),
                          as.character(rownames(Mouse.df)))
#zika.matrix <- as.matrix(zika.df)
```

```
# Define cluster_pvals; DO NOT CHANGE!
k <- 5
cluster_pvals <- function(k, km, myplot.df) {
  # Inputs: k, km, myplot.df
  # Returns: results (dataframe with clusters, pvalues, logodds)
  # 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))
  # use dplyr to calculate counts for each cluster
  temp <- myplot.df %>%
        dplyr::group_by(cluster) %>%
        dplyr::count(name="freq")
```

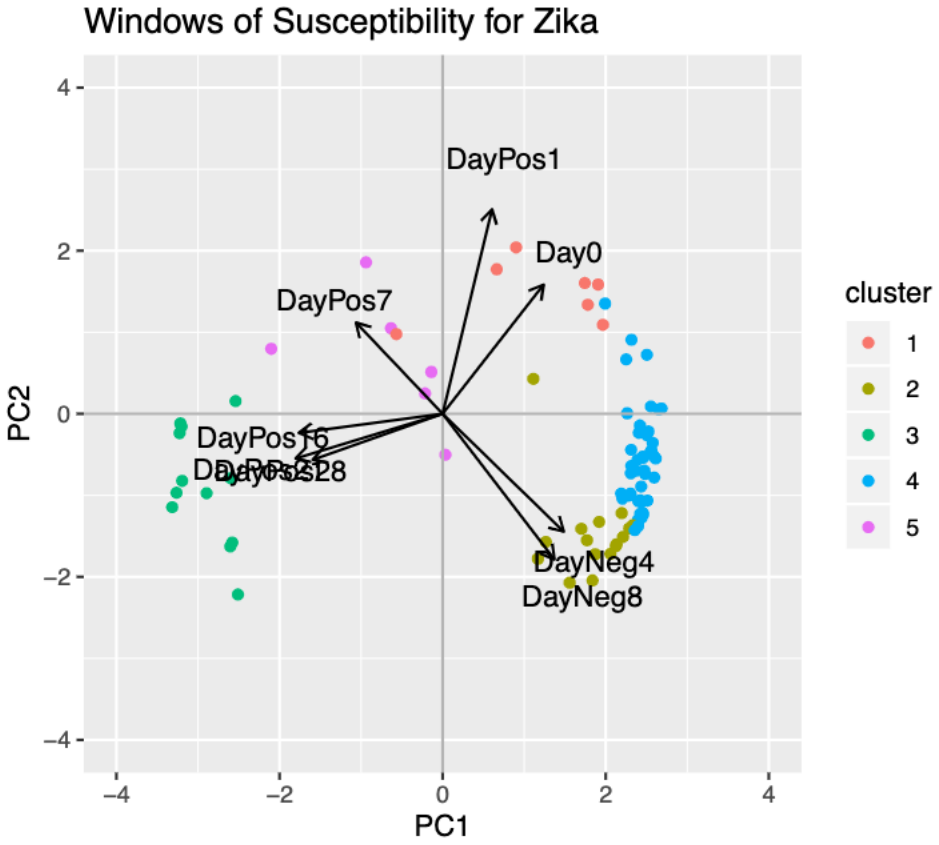
```
  classdisease[temp$cluster] <- temp$freq
  classlogodds <- zeros(k,2)
  totaldisease <- sum(classdisease)
  totalall <- sum(classall)
  # Calculate the log odds ratio for the disease
  for (i in 1:k) {
     n11 <- classdisease[i] +1 # genes in disease in cluster i
     n21 <- totaldisease- classdisease[i] +1 # genes in disease not in cluster i
     n12 <- classall[i]-n11+1 # genes not in disease and in cluster i
     n22 <- totalall- n11-n21 -n12+1; # genes not in disease and not in cluster
     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)
}
```

```
plot.df <- cbind.data.frame(my.pca$x, cluster=as.factor(km$cluster))
myplot.df<-plot.df[zika_symbols,]
# Apply cluster_pvals using the parameters just generated
clusters <- cluster_pvals(k, km, myplot.df)
threshold <- 0.1  # Normally set to 0.1
# Helper function to determine enrichment
enriched <- function(p.value,logodds,p.threshold=0.1){
  if ((p.value <= p.threshold) && (logodds > 0)) {
    return(TRUE)
  } else {return(FALSE)}
}
# Evaluate across our results; create new column
clusters$enriched <- mapply(enriched, clusters$pvalue, clusters$logodds,threshold)

# View results
clusters
```

```
##   cluster       pvalue    logodds enriched
## 1       1 2.636400e-03 -1.0289834    FALSE
## 2       2 6.139669e-01 -0.1544676    FALSE
## 3       3 9.582396e-02 -0.5265662    FALSE
## 4       4 4.101429e-13  1.5996365     TRUE
## 5       5 4.003807e-02 -0.7896979    FALSE
```

```
plot.df <- cbind.data.frame(my.pca$x, cluster=as.factor(km$cluster))
myplot.df<-plot.df[zika_symbols,]

p <- ggplot() +
   geom_point(data=myplot.df, aes(x=PC1, y=PC2,colour=cluster)) +
   coord_fixed(ratio=1) +
   geom_hline(yintercept = 0, color = "gray70") +
   geom_vline(xintercept = 0, color = "gray70") +
   xlim(-4,4) +ylim(-4,4) +
   ggtitle('Windows of Susceptibility for Zika') +


   # This is the extra credit part
   geom_segment(aes(x=0, y=0,
```

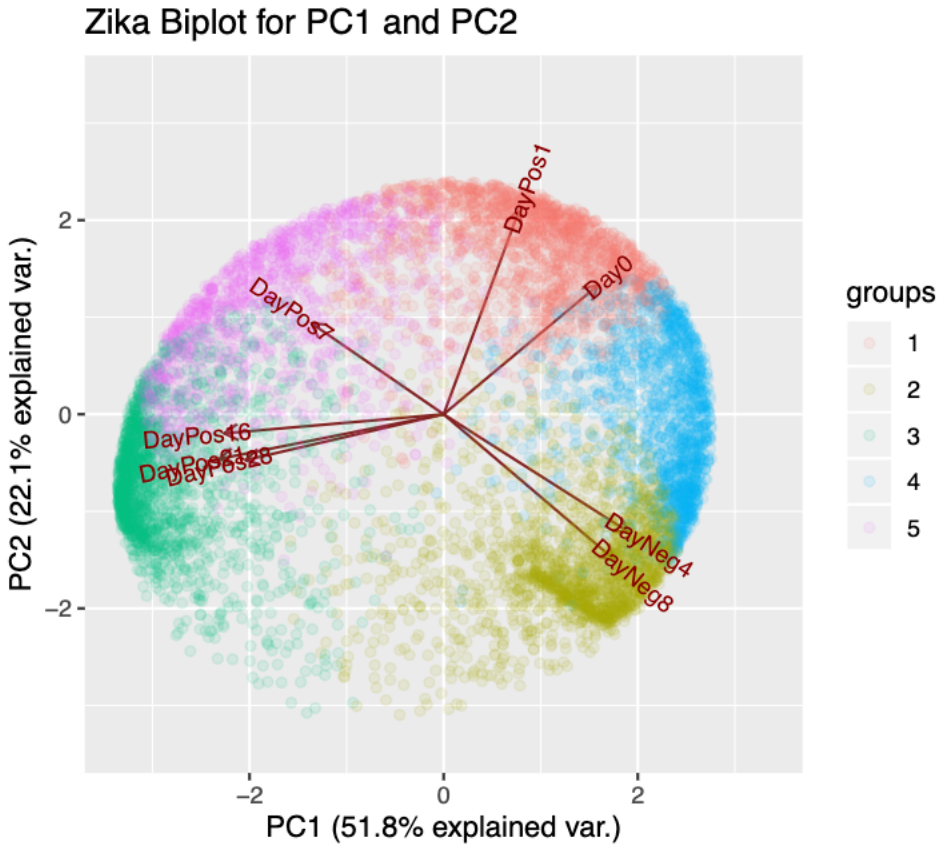
```
                    xend = 4*my.pca$rotation[,1], yend = 4*my.pca$rotation[,2]),
                    arrow = arrow(length = unit(1/2, 'picas'))) +
   geom_text(aes(x = 5*my.pca$rotation[,1],
                 y = 5*my.pca$rotation[,2],
                 label=rownames(my.pca$rotation)),

               size=4)
#legend(x = 3.5, y=14, legend=c("Stage A", "Stage B", "Stage C", "Stage D", "Stage E"),
      # col=c("Cluster 2", "Cluster 4", "Cluster 1", "Cluster 5", "Cluster 3"))
#print the plot
p
```



Stage A: Cluster 2 Stage B: Cluster 4 Stage C: Cluster 1 Stage D: Cluster 5 Stage E: Cluster 3

```
# Calculate x and y scale limits for the biplot
t<-max(abs(my.pca$x[,1:2]))
# Generate the biplot using ggbiplot
p <- ggbiplot(my.pca,
              choices=c(1,2),
              alpha=.1,
              varname.adjust=0.5,
              obs.scale = 1,
              groups=as.factor(km$cluster))
p + ggtitle('Zika Biplot for PC1 and PC2') + xlim(-t,t) + ylim(-t,t)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



3. Windows of Susceptibility Analysis of Craniosynostosis

```
# Read in the dataset; create the matrix

cran.df <-read.csv("~/IDM_work/Craniosynostosis_heat_map_data.csv",row.names = 1)
cran.matrix <- as.matrix(cran.df)
cran_symbols <- intersect(as.character(cran.df$symbol),
                          as.character(rownames(Mouse.df)))
```

```
# Define cluster_pvals; DO NOT CHANGE!
cluster_pvals <- function(k, km, myplot.df) {
  # Inputs: k, km, myplot.df
  # Returns: results (dataframe with clusters, pvalues, logodds)
  # 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))
  # use dplyr to calculate counts for each 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 # genes in disease not in cluster i
     n12 <- classall[i]-n11+1 # genes not in disease and in cluster i
     n22 <- totalall- n11-n21 -n12+1; # genes not in disease and not in cluster
     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)
}
```

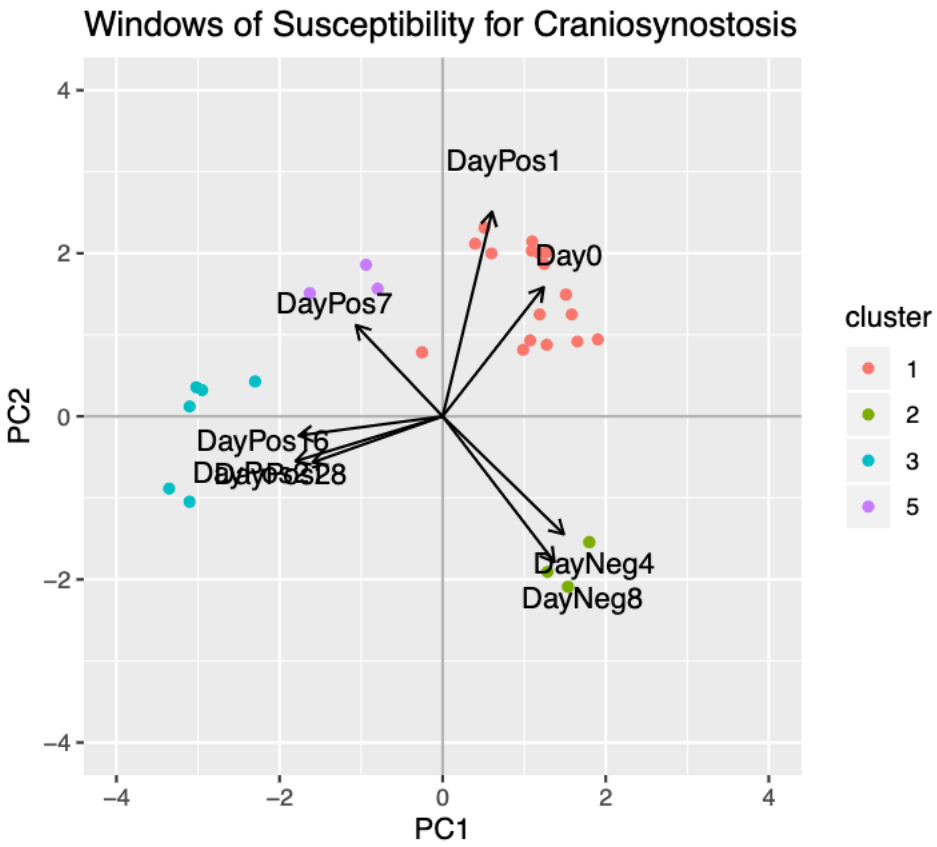
Applying the Helper Function and display the results

```
plot.df <- cbind.data.frame(my.pca$x, cluster=as.factor(km$cluster))
myplot.df<-plot.df[cran_symbols,]

# Apply cluster_pvals using the parameters just generated
clusters <- cluster_pvals(k, km, myplot.df)

threshold <- 0.1  # Normally set to 0.1

# Helper function to determine enrichment
enriched <- function(p.value,logodds,p.threshold=0.1){
  if ((p.value <= p.threshold) && (logodds > 0)) {
    return(TRUE)
  } else {return(FALSE)}
}
# Evaluate across our results; create new column
clusters$enriched <- mapply(enriched, clusters$pvalue, clusters$logodds,threshold)

# View results
clusters
```

```
##   cluster       pvalue     logodds enriched
## 1       1 8.755564e-06  1.64017255     TRUE
## 2       2 2.804323e-01 -0.67818530    FALSE
## 3       3 1.000000e+00  0.02516653    FALSE
## 4       4 3.280933e-02 -1.91929384    FALSE
## 5       5 1.000000e+00 -0.22054277    FALSE
```

```
plot.df <- cbind.data.frame(my.pca$x, cluster=as.factor(km$cluster))
myplot.df<-plot.df[cran_symbols,]

p <- ggplot() +
   geom_point(data=myplot.df, aes(x=PC1, y=PC2,colour=cluster)) +
   coord_fixed(ratio=1) +
   geom_hline(yintercept = 0, color = "gray70") +
```

```
  geom_vline(xintercept = 0, color = "gray70") +
  xlim(-4,4) +ylim(-4,4) +
  ggtitle('Windows of Susceptibility for Craniosynostosis') +
  # This is the extra credit part
  geom_segment(aes(x=0, y=0,
                   xend = 4*my.pca$rotation[,1], yend = 4*my.pca$rotation[,2]),
                   arrow = arrow(length = unit(1/2, 'picas'))) +
  geom_text(aes(x = 5*my.pca$rotation[,1],
                y = 5*my.pca$rotation[,2],
                label=rownames(my.pca$rotation)),

              size=4)

#print the plot
p
```



4. My Creative Analysis

```
cran.df <-read.csv("~/IDM_work/Craniosynostosis_heat_map_data.csv",row.names = 1)
cran.matrix <- as.matrix(cran.df)
cran_symbols <- intersect(as.character(cran.df$symbol),
                          as.character(rownames(Mouse.df)))
```

```
#cluster 1
col0 <- matrix(0, 4, 1)
col7 <- matrix(7, 4, 1)
```

```
col12 <- matrix(12, 4, 1)
col19 <- matrix(19, 4, 1)
col26 <- matrix(26, 4, 1)
col33 <- matrix(33, 4, 1)
col49 <- matrix(49, 4, 1)
col63 <- matrix(63, 4, 1)
col77 <- matrix(77,4,1)

plotc1 <- cbind(c(cran.matrix[1:4,3],cran.matrix[1:4,4],cran.matrix[1:4,5],cran.matrix[1:4,6],cran.matr

#plot(plotc1[,2],plotc1[,1],xlim = c(-1,80), ylim = c(-3,3))
```

```
#cluster 2
col0 <- matrix(0, 2, 1)
col7 <- matrix(7, 2, 1)
col12 <- matrix(12, 2, 1)
col19 <- matrix(19, 2, 1)
col26 <- matrix(26, 2, 1)
col33 <- matrix(33, 2, 1)
col49 <- matrix(49, 2, 1)
col63 <- matrix(63, 2, 1)
col77 <- matrix(77,2,1)

plotc2 <- cbind(c(cran.matrix[5:6,3],cran.matrix[5:6,4],cran.matrix[5:6,5],cran.matrix[5:6,6],cran.matr

#plot(plotc2[,2],plotc2[,1],xlim = c(-1,80), ylim = c(-3,3))
```

```
#cluster 3
col0 <- matrix(0, 9, 1)
col7 <- matrix(7, 9, 1)
col12 <- matrix(12, 9, 1)
col19 <- matrix(19, 9, 1)
col26 <- matrix(26, 9, 1)
col33 <- matrix(33, 9, 1)
col49 <- matrix(49, 9, 1)
col63 <- matrix(63, 9, 1)
col77 <- matrix(77,9,1)

plotc3 <- cbind(c(cran.matrix[7:15,3],cran.matrix[7:15,4],cran.matrix[7:15,5],cran.matrix[7:15,6],cran.

#plot(plotc3[,2],plotc3[,1],xlim = c(-1,80), ylim = c(-3,3))
```

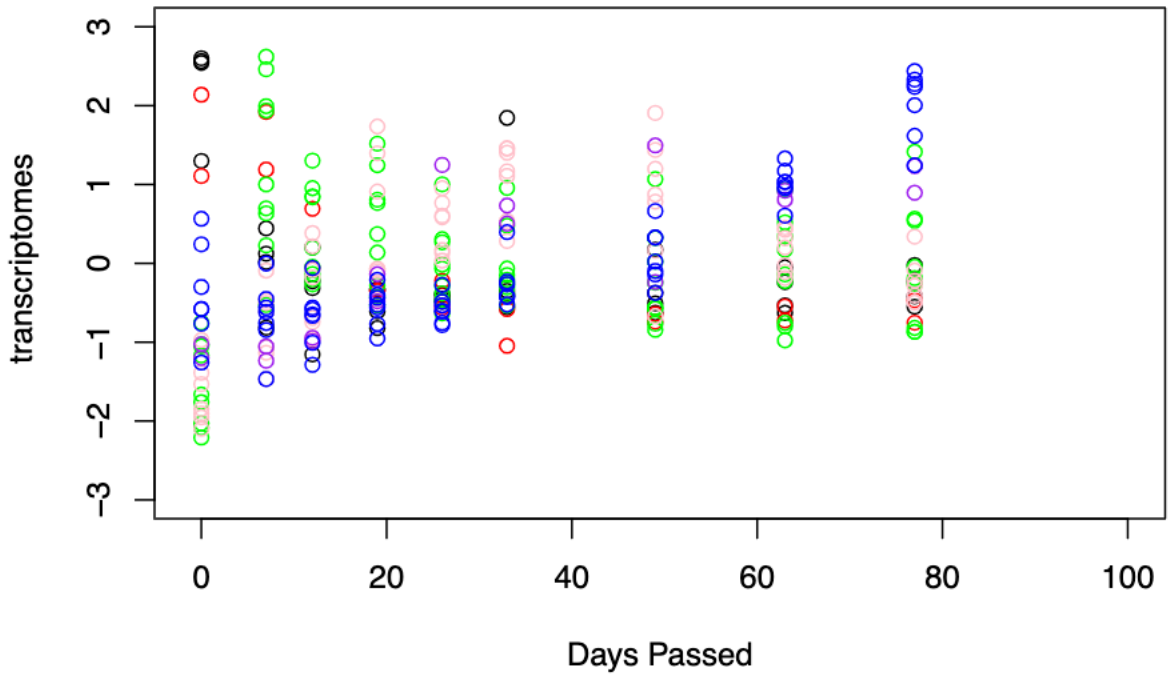
```
#cluster 4
col0 <- matrix(0, 7, 1)
col7 <- matrix(7, 7, 1)
col12 <- matrix(12, 7, 1)
col19 <- matrix(19, 7, 1)
col26 <- matrix(26, 7, 1)
col33 <- matrix(33, 7, 1)
col49 <- matrix(49, 7, 1)
col63 <- matrix(63, 7, 1)
col77 <- matrix(77,7,1)

plotc4 <- cbind(c(cran.matrix[16:22,3],cran.matrix[16:22,4],cran.matrix[16:22,5],cran.matrix[16:22,6],c
```

```
#plot(plotc4[,2],plotc4[,1],xlim = c(-1,80), ylim = c(-3,3))
```

```
#cluster 5
col0 <- matrix(0, 2, 1)
col7 <- matrix(7, 2, 1)
col12 <- matrix(12, 2, 1)
col19 <- matrix(19, 2, 1)
col26 <- matrix(26, 2, 1)
col33 <- matrix(33, 2, 1)
col49 <- matrix(49, 2, 1)
col63 <- matrix(63, 2, 1)
col77 <- matrix(77,2,1)

plotc5 <- cbind(c(cran.matrix[23:24,3],cran.matrix[23:24,4],cran.matrix[23:24,5],cran.matrix[23:24,6],c

#plot(plotc5[,2],plotc5[,1],xlim = c(-1,80), ylim = c(-3,3))
```

```
#cluster 6
col0 <- matrix(0, 7, 1)
col7 <- matrix(7, 7, 1)
col12 <- matrix(12, 7, 1)
col19 <- matrix(19, 7, 1)
col26 <- matrix(26, 7, 1)
col33 <- matrix(33, 7, 1)
col49 <- matrix(49, 7, 1)
col63 <- matrix(63, 7, 1)
col77 <- matrix(77,7,1)

plotc6 <- cbind(c(cran.matrix[25:31,3],cran.matrix[25:31,4],cran.matrix[25:31,5],cran.matrix[25:31,6],c

#plot(plotc6[,2],plotc6[,1],xlim = c(-1,80), ylim = c(-3,3))
```

```
plot(plotc1[,2],plotc1[,1],xlim = c(-1,100), ylim = c(-3,3), xlab = "Days Passed",ylab = " transcriptom
points(plotc2[,2], plotc2[,1], col='red')
points(plotc3[,2], plotc3[,1], col='green')
points(plotc4[,2], plotc4[,1], col='pink')
points(plotc5[,2], plotc5[,1], col='purple')
points(plotc6[,2], plotc6[,1], col='blue')
```



In the scatter plot the different colored circles represent the different clusters in the craniosynostosis data. For cluster 1 (the dots colored black) that contains the genes FGF2, FGF4, FGFR4, SKI, and SPP1 these genes at first have a high transcriptome value, but then they begin to lower at day 7 and stay more or less constant throughout the rest of the days except for day 33 where there is one outlier. The genes are most spread out in day 33 and they are overlapping at day 26. Cluster 2 is represented with the red colored dots. The genes in this cluster include ALPL, FGFR1, NSD1, and SATB2. This is the second smallest cluster in the data with only 4 genes in the data. As for its shape we can see that it starts off with relatively high transcriptome levels and then decreases and by day 33 the level of transcriptome stays abour constant. The genes are most spread apart at day 0 and they are actually exactly the same in day 19. In Cluster 3 the data is marked with green circles. This was the largest cluster in the data set and it contained the genes AXIN2, BMP4, COL2A1, FGFR2, GDF7, GLI3, GPC3, MSX1, PTCH1, SP7, and TGFBR1. This cluster starts off with low transcritpome levels and then it sky rockets at day 7 where day 7 is the peak for this cluster's transcriptome values. Then it decreases a little and it stays at around this level. The genes in this cluster are most different (spread out by transcriptome levels) at day 7 as well, while they are all very close in value at day 33. Cluster 4 represented by the pink circles includes the genes EFNB1, FGF18, FGF8, FGF9, MSX2, NOG, PJA1, POR, and NIC1. This cluster starts off with low transcriptome levels and it slowly increases and it reaches its peak at day 19, then dips a little, then reaches a peak again at day 49 and then it decreases once again. At day 49 the genes are most varied and spread out and at day 63 the genes are clostest together and all amost have the same transcriptome levels. In cluster 5 the data is represented by purple circles. Cluster 5 was the smallest cluster in the data set with only two genes in the cluster: NELL1 and TGFB2. This cluster starts off at about a transcirptome level of -1 and stays around that level until day 19 in which is begins to increase. At day 26 the genes are at approcimately level and stay that way for the rest of the days. At Day 12 the two genes have the same exact transcriptome level and at day 26 and at day 49 the two genes have the most different trasncriptome levels. Lastly, for cluster the data is represented with blue circles. The genes in this cluster are ALX4, FBN1, FGH10, FGF7, TGFB1, and TGFB3. This day starts off with moderate transcriptome levels that dip down very slightly in day 7 and 13 then stays constan for days 19 and 26 then gradully increase throughout the rest of the days. The gene's transcriptome levels are most spread out in day 0 and they are the most similar at day 26. Overall, this scatterplot helped to show the trends in the different genes throughout the 77 days.

5. Conclusions

For the Zika disease Stage B (Cluster 4) is enriched which means at Stage B, between Day 0 and Day -8(before birth), is the window of susceptibility for the Zika disease. For Craniosynostosis the window of suspecntibikty is at Day 0 and Day 1 because Stage B (cluster 1) is enriched.

Extra Credit