Insights on Epstein-Barr Virus (EBV) Oncogenesis using RNA-Seq

Code

Isabel Dengos

August 2022

Motivations

Importance: Epstein-Barr Virus Leads to many types of lymphoma in immunocompromised patients

Study oncogenesis pathways by observing differentially expressed genes during EBV infection

Long term goal: be able to make targeted pathway therapeutics and intervention methods

Recreate the analysis done in the paper: RNA Sequencing Analysis of Gene Expression during Epstein-Barr Infection of Primary B Lymphocytes by Wang et al.[1]

Pre-Data Analysis

Followed the process of primary resting B lymphocytes (RBLs) to lymphoblastoid cell line (LCL) after infection of Epstein-Barr Virus

Analyzed differential gene expression of 3 different donors for 0, 2, 4, 7, 14, 21 and 28 days

Libraries were already generated using Illumina

Cluster files from supplemental material Wang et al [1]

Hide

```
getwd()

#clear workspace
rm(list=ls())
```

Hide

```
#Load_libraries
library("DESeq2")
library(AnnotationDbi)
library(org.Hs.eg.db)
library(EnhancedVolcano)
library("pheatmap")
library(maser)
library(dplyr)
library(cluster)
library(factoextra)
library(gplots)
library(NbClust)
#library(tidyverse)
library(readxl)
```

Data Analysis

Hide

```
#Count_File_Generation
file.list <- list.files( path = "./STAR/", pattern = "*ReadsPerGene.out.tab$")
counts.files <- lapply(paste('./STAR/', file.list,sep=''), read.table, skip = 4)
counts <- as.data.frame( sapply( counts.files, function(x) x[ ,2] ) )
colnames(counts) <- file.list
row.names(counts) <- counts.files[[1]]$V1

#Should change the column names here so human readable, will make analysis/sorting ea
sier later
colnames(counts) <- c('d0_Donor1','d0__Donor2','d0_Donor3','d2_Donor1','d2_Donor2','d
2_Donor3','d4_Donor1','d4_Donor2','d4_Donor3','d7_Donor1','d7_Donor2','d7_Donor3','d1
4_Donor1','d14_Donor2','d14_Donor3','d21_Donor1','d21_Donor2','d21_Donor3','d28_Donor
1','d28_Donor2','d28_Donor3')
```

Hide

```
#Get the cluster files undone
cluster.file <- read_excel("~/achm511/idengos/proj2/jvi.00226-19-sd001.xlsx", skip =
1)

#Clean up ensemble IDs
cluster.file$gene_id <- gsub("\\..*","",cluster.file$gene_id)
```

Hide

```
#Conditions_Metadata
condition <- c(rep("Day0",3), rep("Day2",3), rep("Day4",3), rep("Day7",3), rep("Day1
4",3), rep("Day21",3), rep("Day28",3))
sampleTable <- data.frame(sampleName = file.list, condition = condition)
dds <- DESeqDataSetFromMatrix(countData = counts, colData = sampleTable, design = ~ c
ondition)

#The gosh darn order of the dds is off
dds$condition <- factor(dds$condition, levels=c("Day0","Day2","Day4","Day7","Day14","
Day21","Day28"))
```

Hide

```
#Running_and_writing_data
output <- DESeq(dds)
output_df <- as.data.frame(results(output))
```

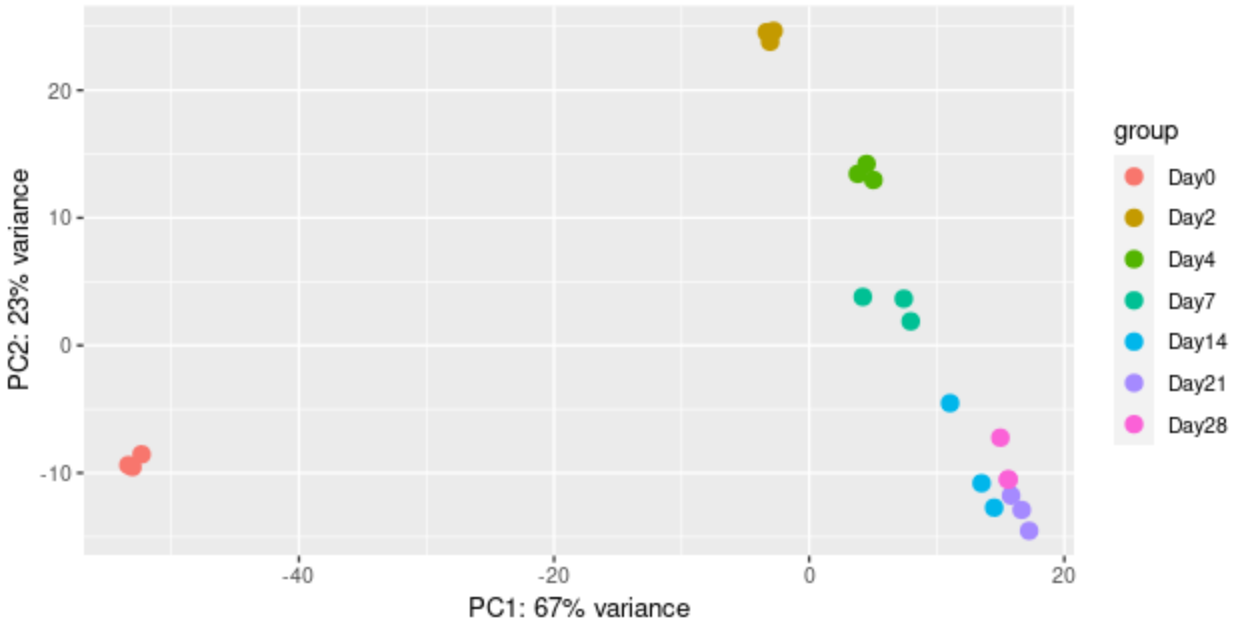
Hide

```
#Ensemble Ids for whole to gene names
IDS <- row.names(output_df)
output_df$Symbol <-   mapIds(org.Hs.eg.db, IDS, 'SYMBOL', 'ENSEMBL') #maps ensemble I
Ds to gene names
```

PCA Plot

Hide

```
#PCA
vsd <- vst(dds, blind=FALSE)
#png("PCAplot.png")
plotPCA(vsd, intgroup=c("condition"))
```



Hide

```
#dev.off()
```

PCA plot showing drastic changes between Day 0 and the rest of the data

Hide

```
results_Control_Day2 <- results(output, contrast=c("condition","Day2","Day0"))
results_Control_Day2_PValue <- results_Control_Day2[order(results_Control_Day2$pad
j),]
write.csv(as.data.frame(results_Control_Day2_PValue), file="Control_Day2_DE.csv")
```

Hide

```
results_Control_Day28 <- results(output, contrast=c("condition","Day28","Day0"))
results_Control_Day28_PValue <- results_Control_Day28[order(results_Control_Day28$pad
j),]
write.csv(as.data.frame(results_Control_Day28_PValue), file="Control_Day28_DE.csv")
```

Hide

```
#Converting_ENSEMBLE_IDs_to_Gene_IDs
Day2_Control <- read.table("Control_Day2_DE.csv", header = TRUE, sep =',')
IDs <- c(Day2_Control$X)
Day2_Control$Symbol <- mapIds(org.Hs.eg.db, IDs, 'SYMBOL', 'ENSEMBL') #maps ensemble
IDs to gene names
rownames(Day2_Control) <- Day2_Control$X
```

Hide

```
#Converting_ENSEMBLE_IDs_to_Gene_IDs
Day28_Control <- read.table("Control_Day28_DE.csv", header = TRUE, sep =',')
IDs <- c(Day28_Control$X)
Day28_Control$Symbol <- mapIds(org.Hs.eg.db, IDs, 'SYMBOL', 'ENSEMBL') #maps ensemble
IDs to gene names
rownames(Day28_Control) <- Day28_Control$X
```

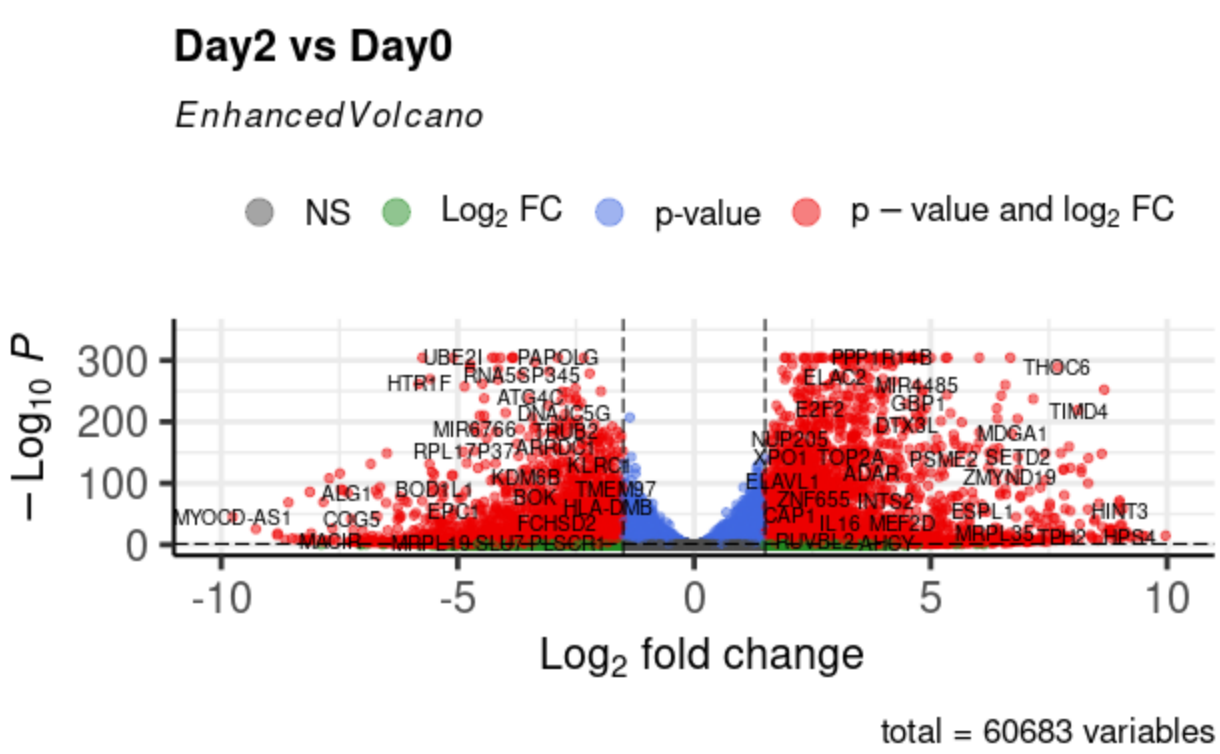
Volcano Plot

Volcano Plot Day 2 vs Day 0

Hide

```
#Volcano_Plot Day 2 vs 0

#png("VolDay2_0.png")
EnhancedVolcano(results_Control_Day2,
                lab = Day2_Control$Symbol,
                x = "log2FoldChange",
                y = "padj",
                pCutoff = 0.05,
                FCcutoff = 1.5,
                pointSize = 1.5,
                labSize = 3.0,
               title = "Day2 vs Day0",
               xlim = c(-10,10),
               ylim = c(0,350))
```

```
Warning: One or more p-values is 0. Converting to 10^-1 * current lowest non-zero p-v
alue...
```



Hide

```
#dev.off()
```

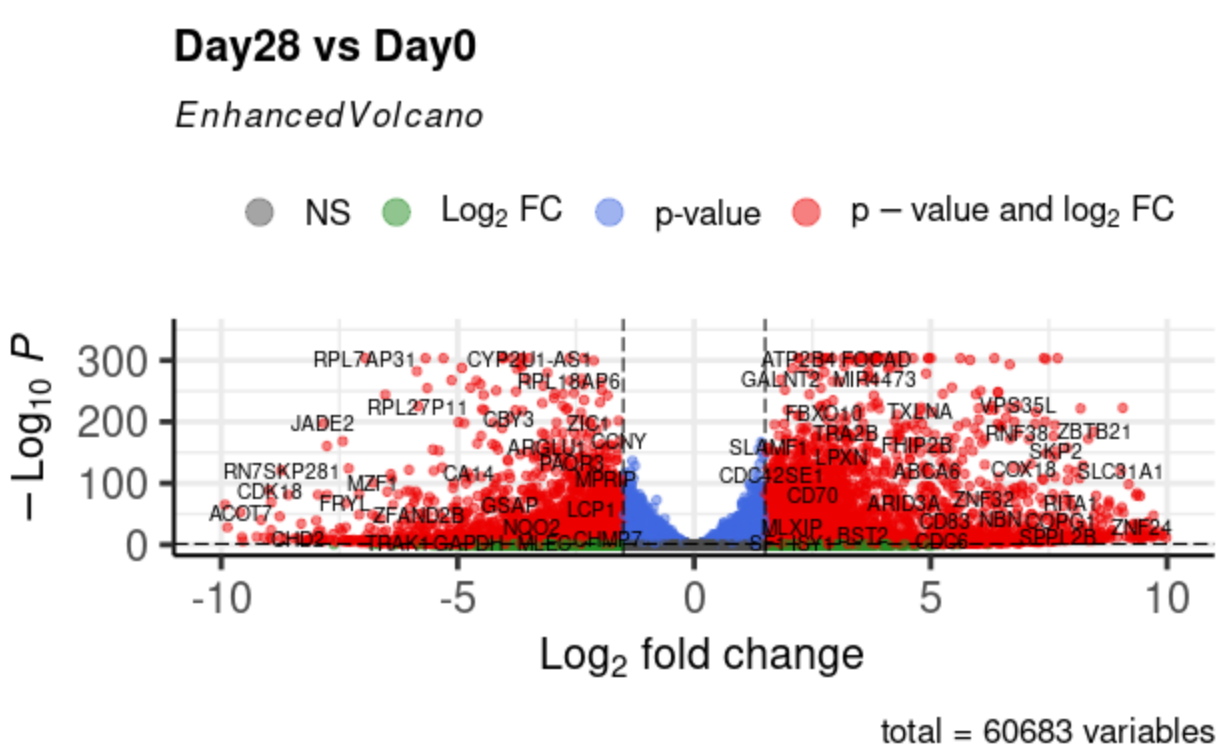
Volcano plot from Day 0 to Day 2. Extreme number of statistically important expressed genes.

Volcano Plot Day 28 vs Day 0

Hide

```
#Volcano_Plot Day 28 vs 0

#png("VolDay28_0.png")
EnhancedVolcano(results_Control_Day28,
                lab = Day28_Control$Symbol,
                x = "log2FoldChange",
                y = "padj",
                pCutoff = 0.05,
                FCcutoff = 1.5,
                pointSize = 1.5,
                labSize = 3.0,
               title = "Day28 vs Day0",
               xlim = c(-10,10),
               ylim = c(0,350))
```

```
Warning: One or more p-values is 0. Converting to 10^-1 * current lowest non-zero p-v
alue...
```



Hide

```
#dev.off()
```

Volcano Plot Day 28 vs Day 21

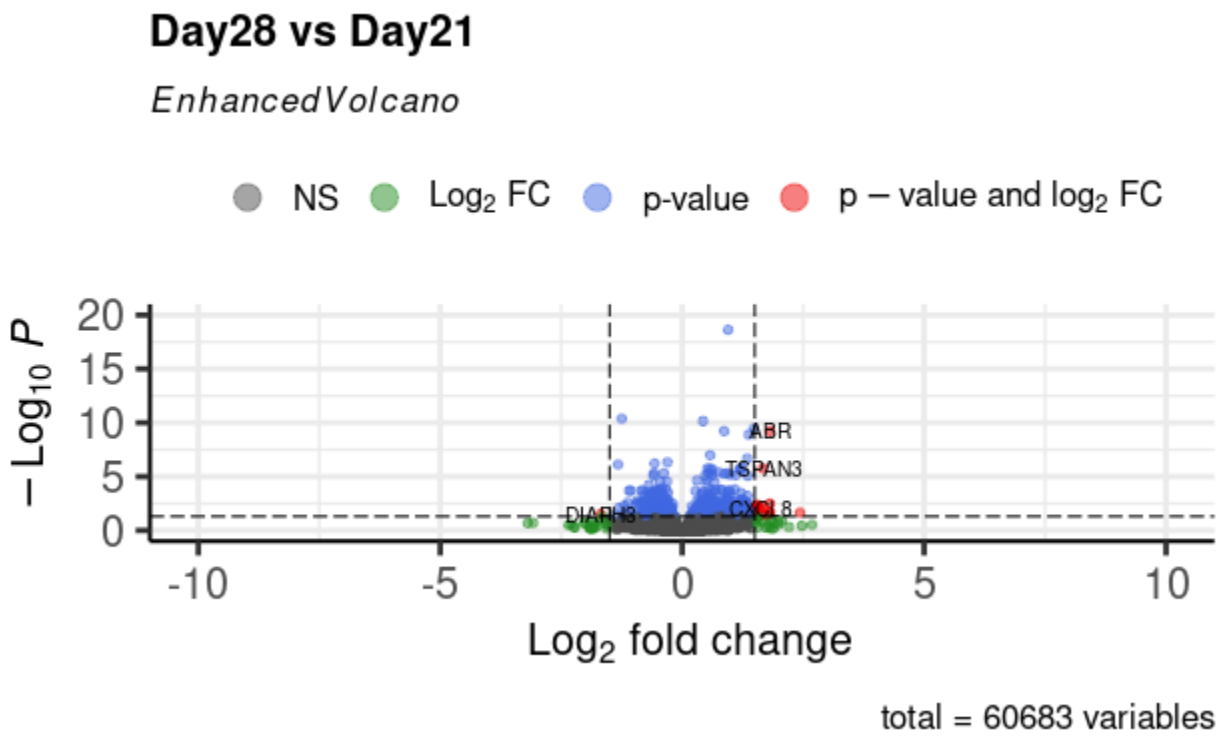
Hide

```
results_Control_DayClose <- results(output, contrast=c("condition","Day28","Day21"))
results_Control_DayClose_PValue <- results_Control_DayClose[order(results_Control_Day
Close$padj),]

write.csv(as.data.frame(results_Control_DayClose_PValue), file="Control_DayClose_DE.c
sv")


#Converting_ENSEMBLE_IDs_to_Gene_IDs
DayClose_Control <- read.table("Control_DayClose_DE.csv", header = TRUE, sep =',')
IDs <- c(DayClose_Control$X)
DayClose_Control$Symbol <- mapIds(org.Hs.eg.db, IDs, 'SYMBOL', 'ENSEMBL') #maps ensem
ble IDs to gene names
```

```
'select()' returned 1:many mapping between keys and columns
```

Hide

```
rownames(DayClose_Control) <- DayClose_Control$X

#Volcano_Plot Day 28 vs 0

#png("VolDay28_21.png")
EnhancedVolcano(results_Control_DayClose,
                lab = DayClose_Control$Symbol,
                x = "log2FoldChange",
                y = "padj",
                pCutoff = 0.05,
                FCcutoff = 1.5,
                pointSize = 1.5,
                labSize = 3.0,
               title = "Day28 vs Day21",
               xlim = c(-10,10),
               ylim = c(0,20))
```



Hide

```
#dev.off()
```

Volcano plot from Day 21 to Day 28. Less differential expression compared to day 0 to day 2. Can be explained by closer clustering within PCA; biologically, the virus has taken root fully

Hide

```
rld <- rlog(dds, blind=FALSE)
```

Heatmaps

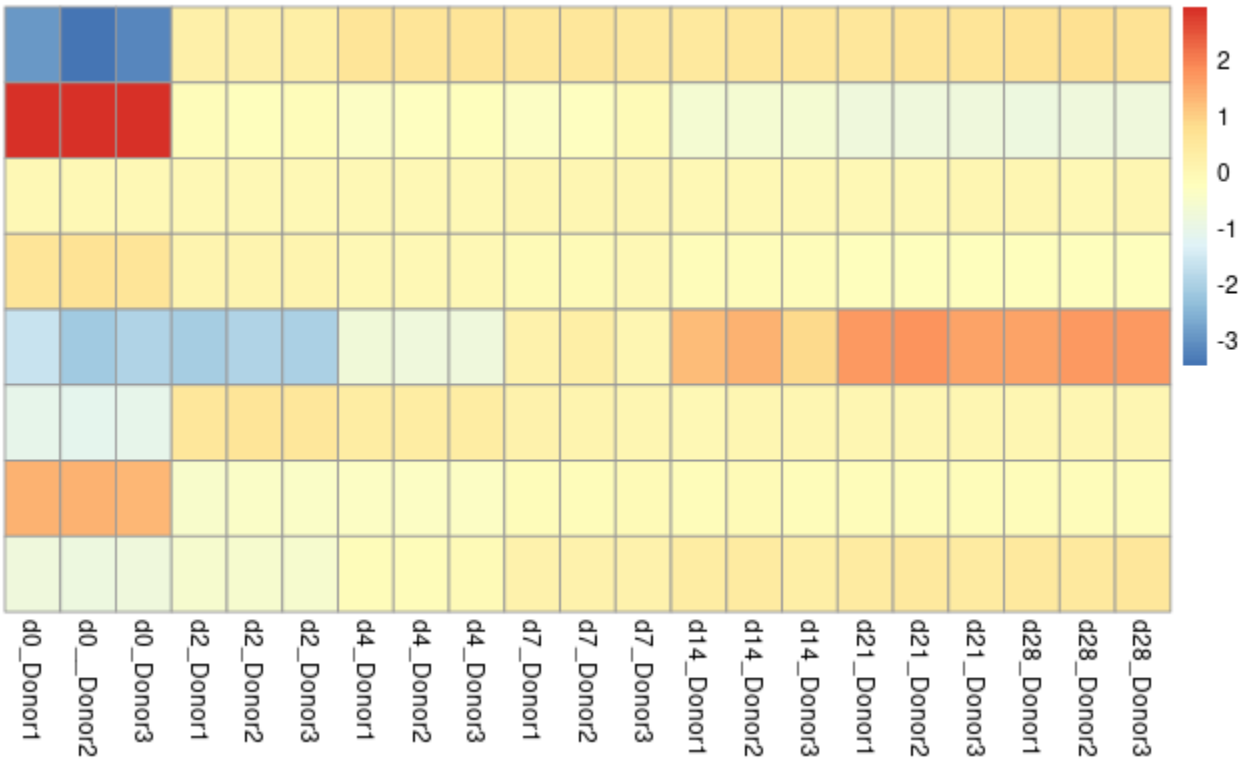
Overall Heatmap

Hide

```
Omat <- assay(rld)
```

Hide

```
df0 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df0) <- "Condition"
row.names(df0) <- sampleTable$sampleName
Omat <- Omat - rowMeans(Omat)
#png("heatmap_clust1.png")
pheatmap(Omat, annottation_col=df0, cluster_rows=F, show_rownames=F, cluster_cols=F,
kmeans_k=8)
```



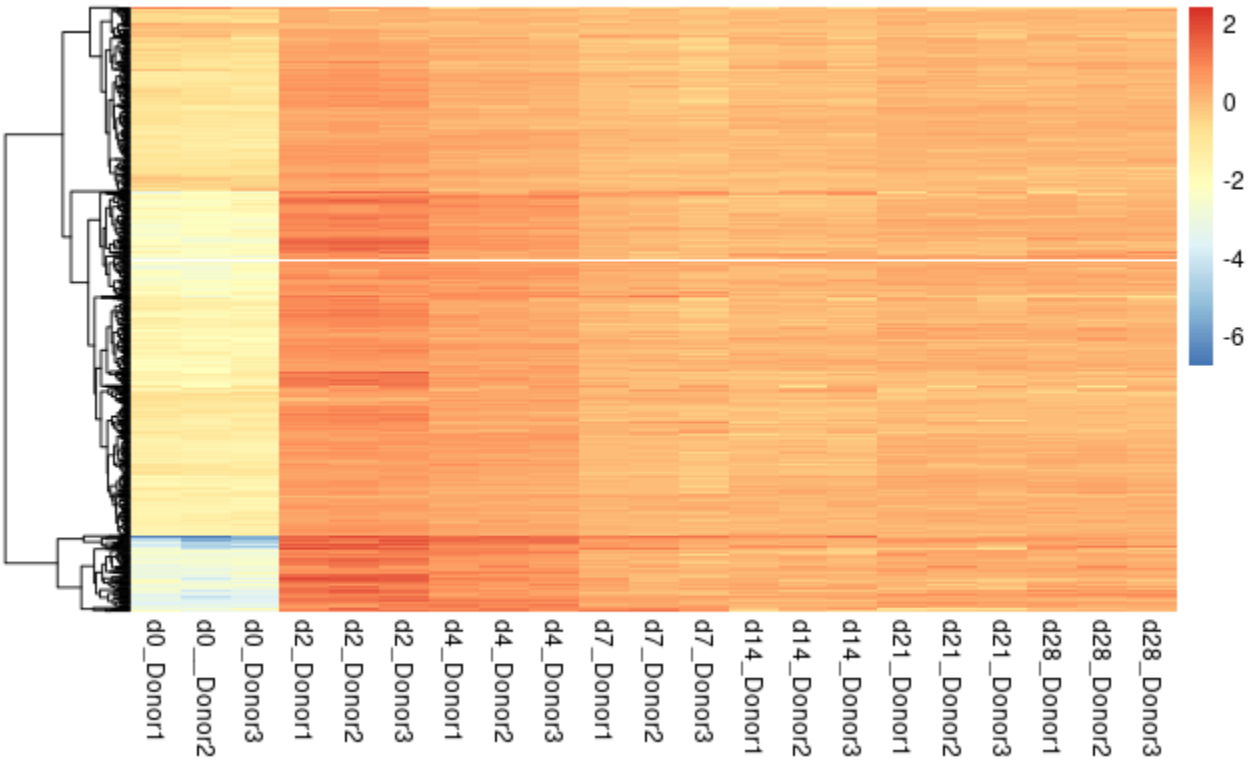
Hide

```
#dev.off()
```

Cluster 1

Hide

```
clust1_genes <- filter(cluster.file, group == "1" )

clust1_gene_list <- select(clust1_genes, gene_id)

clust1_gene_list <- unlist(clust1_gene_list)

#some of the genes are missing from the dds but in the clustering file from the xls
x?!?! had to remove those 2
con <- which(clust1_gene_list %in% row.names(output_df))
clust1_gene_list <- clust1_gene_list[con]


mathett <- assay(rld)[clust1_gene_list,]

df1 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df1) <- "Condition"
row.names(df1) <- sampleTable$sampleName
mathett <- mathett - rowMeans(mathett)
#png("heatmap_clust1.png")
ht1 = pheatmap(mathett, annottation_col=df1, cluster_rows=T, show_rownames=F, cluster
_cols=F)
```



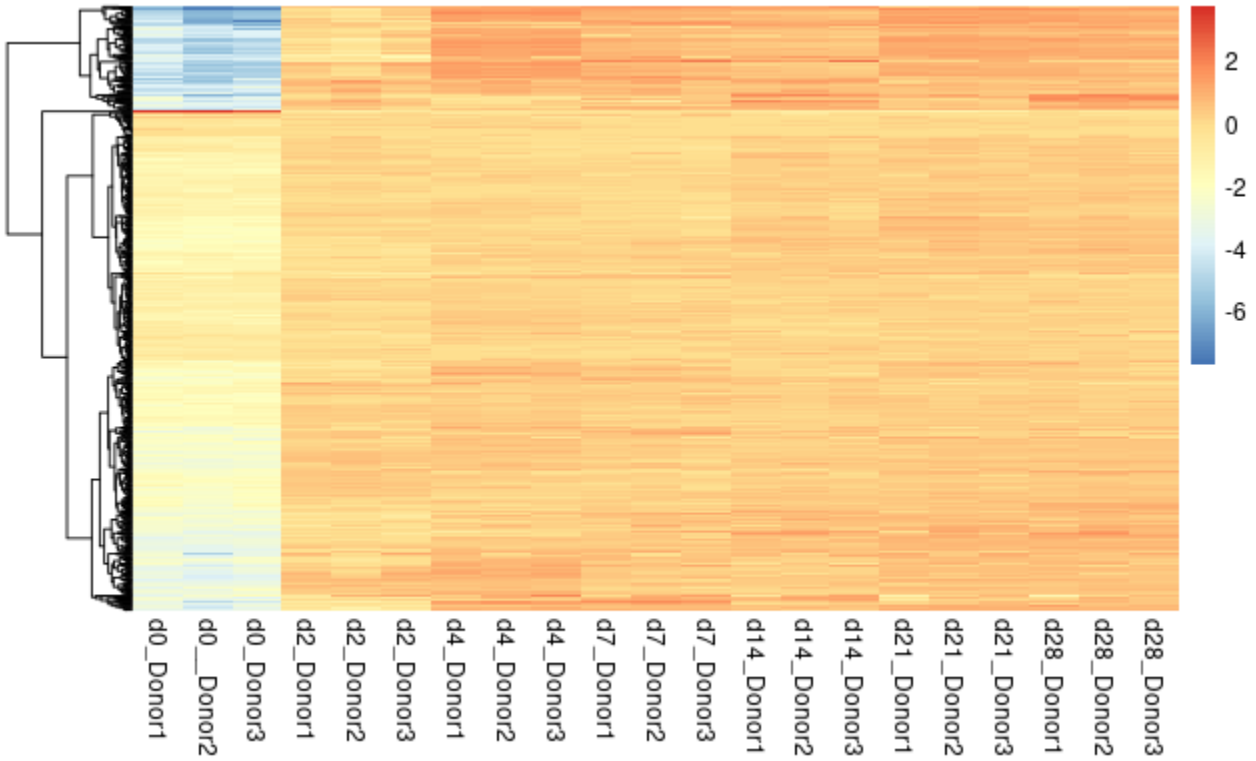
Hide

```
#dev.off()
```

Cluster 2

Hide

```
clust2_genes <- filter(cluster.file, group == "2" )

clust2_gene_list <- select(clust2_genes, gene_id)

clust2_gene_list <- unlist(clust2_gene_list)

con <- which(clust2_gene_list %in% row.names(output_df))
clust2_gene_list <- clust2_gene_list[con]

mat2 <- assay(rld)[clust2_gene_list,]

df2 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df2) <- "Condition"
row.names(df2) <- sampleTable$sampleName
mat2 <- mat2 - rowMeans(mat2)

#png("heatmap_clust2.png")
ht2 = pheatmap(mat2, annottation_col=df2, cluster_rows=T, show_rownames=F, cluster_co
ls=F)
```



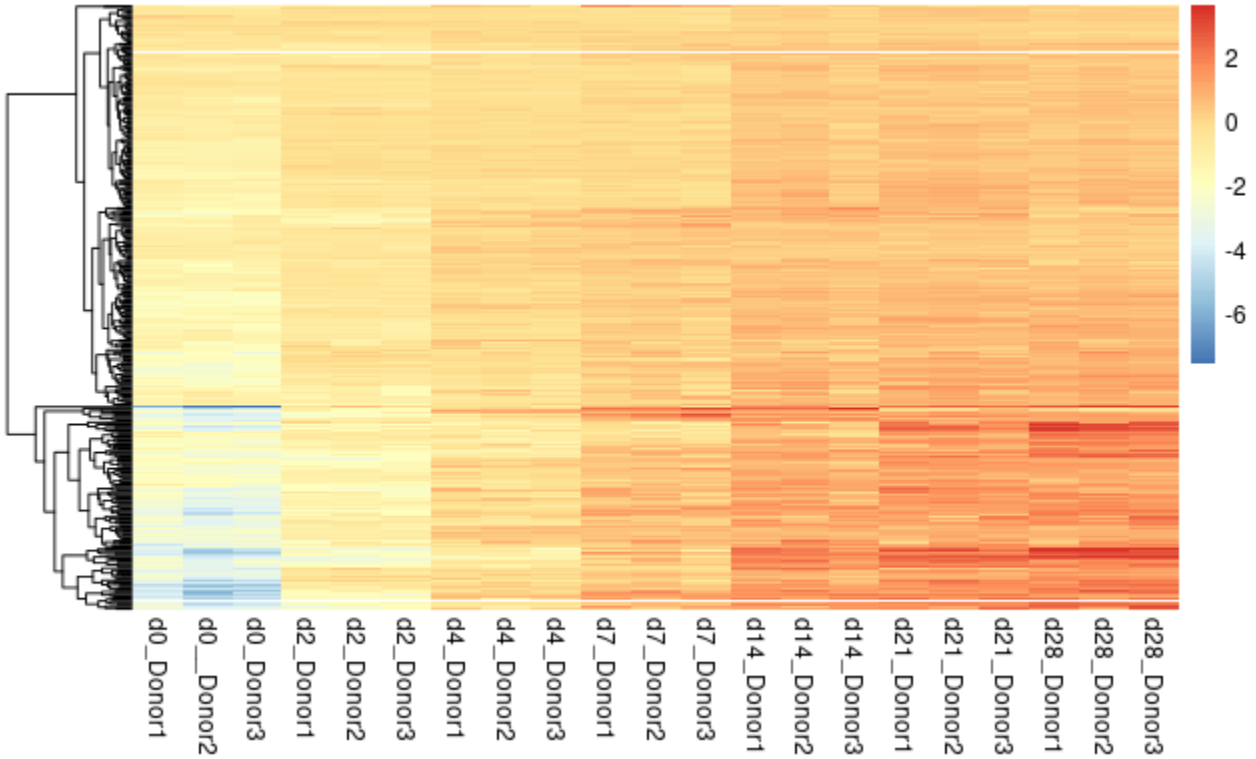
Hide

```
#dev.off()
```

Cluster 3

Hide

```
clust3_genes <- filter(cluster.file, group == "3" )

clust3_gene_list <- select(clust3_genes, gene_id)

clust3_gene_list <- unlist(clust3_gene_list)

con <- which(clust3_gene_list %in% row.names(output_df))
clust3_gene_list <- clust3_gene_list[con]

mat3 <- assay(rld)[clust3_gene_list,]

df3 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df3) <- "Condition"
row.names(df3) <- sampleTable$sampleName
mat3 <- mat3 - rowMeans(mat3)

#png("heatmap_clust3.png")
pheatmap(mat3, annottation_col=df3, cluster_rows=T, show_rownames=F, cluster_cols=F)
```



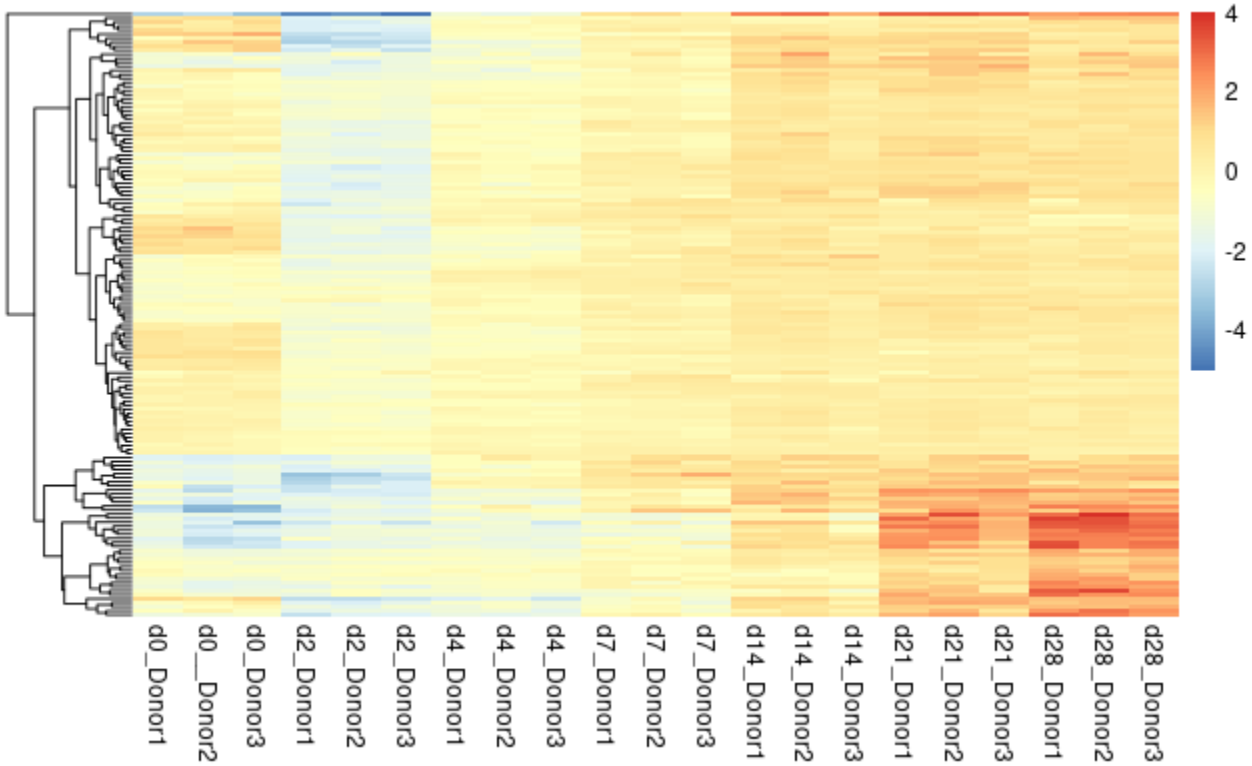
Hide

```
#dev.off()
```

Cluster 4

Hide

```
clust4_genes <- filter(cluster.file, group == "4" )

clust4_gene_list <- select(clust4_genes, gene_id)

clust4_gene_list <- unlist(clust4_gene_list)

con <- which(clust4_gene_list %in% row.names(output_df))
clust4_gene_list <- clust4_gene_list[con]

mat4 <- assay(rld)[clust4_gene_list,]

df4 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df4) <- "Condition"
row.names(df4) <- sampleTable$sampleName
mat4 <- mat4 - rowMeans(mat4)

#png("heatmap_clust4.png")
pheatmap(mat4, annottation_col=df4, cluster_rows=T, show_rownames=F, cluster_cols=F)
```



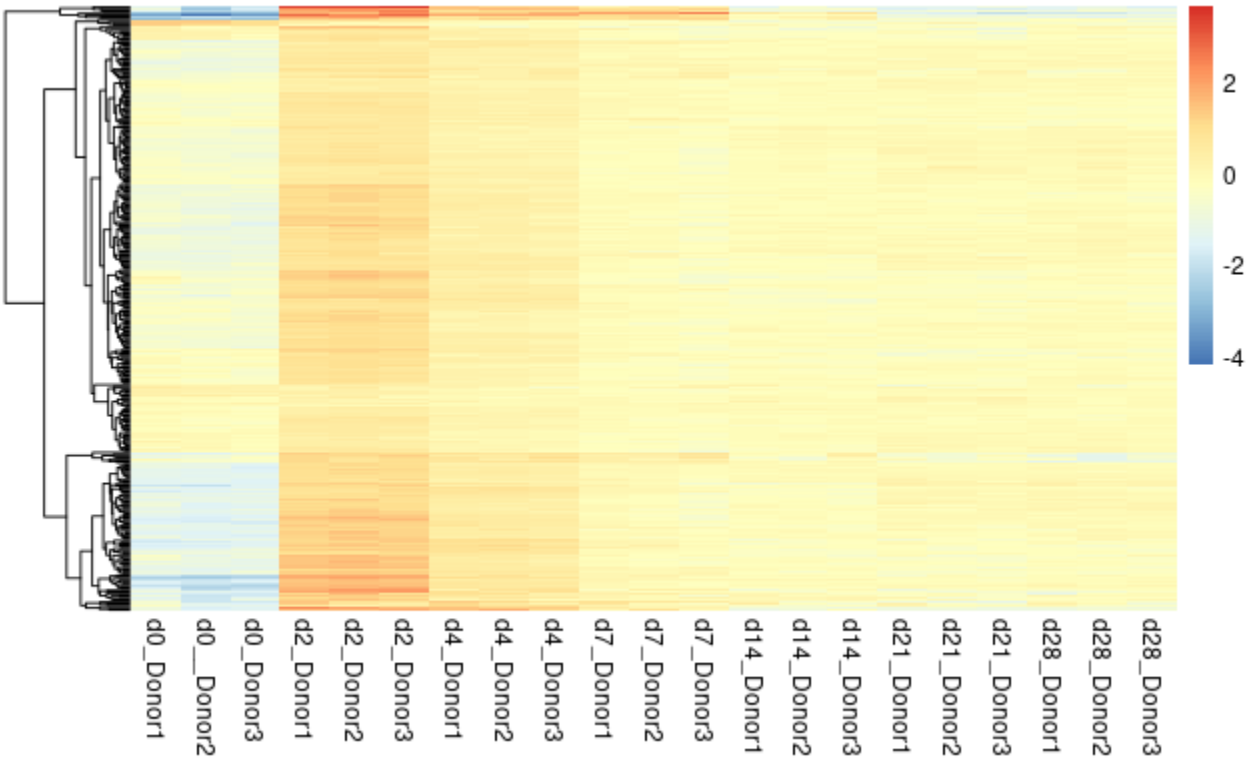
Hide

```
#dev.off()
```

Cluster 5

Hide

```
clust5_genes <- filter(cluster.file, group == "5" )

clust5_gene_list <- select(clust5_genes, gene_id)

clust5_gene_list <- unlist(clust5_gene_list)


con <- which(clust5_gene_list %in% row.names(output_df))
clust5_gene_list <- clust5_gene_list[con]

mat5 <- assay(rld)[clust5_gene_list,]

df5 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df5) <- "Condition"
row.names(df5) <- sampleTable$sampleName
mat5 <- mat5 - rowMeans(mat5)

#png("heatmap_clust5.png")
pheatmap(mat5, annottation_col=df5, cluster_rows=T, show_rownames=F, cluster_cols=F)
```



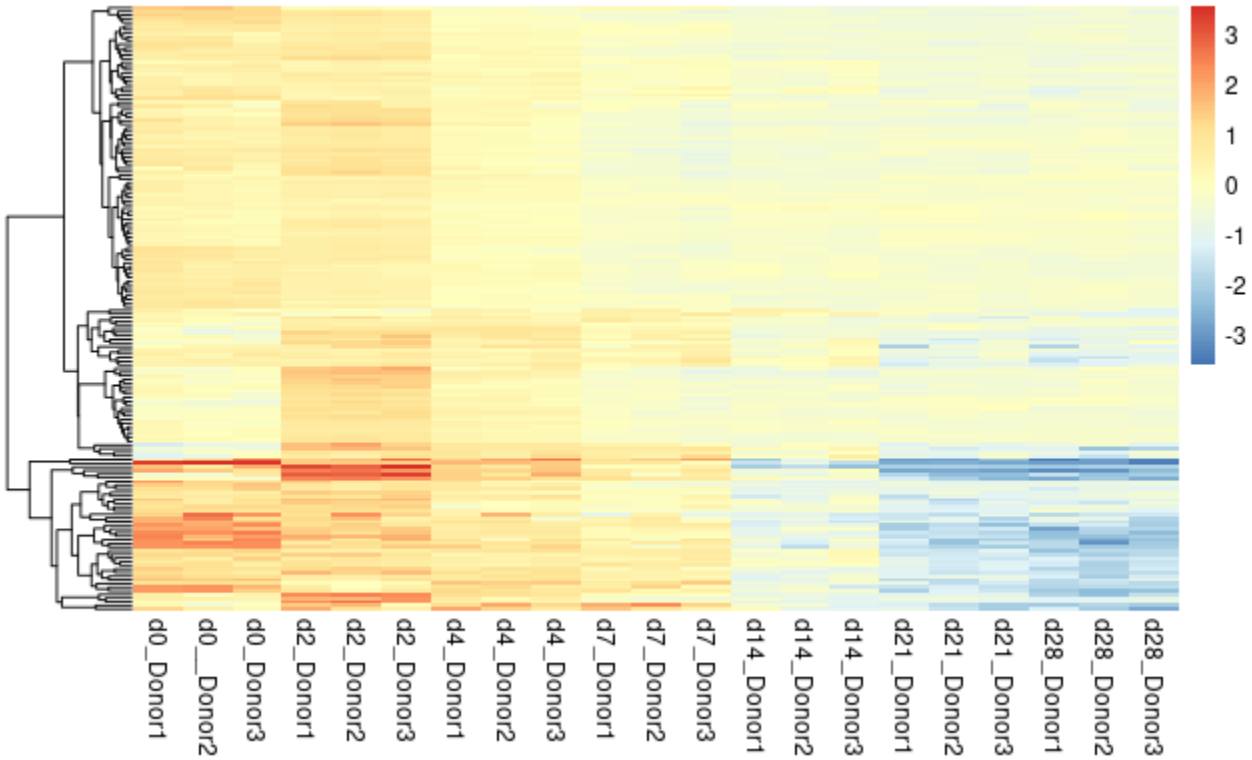
Hide

```
#dev.off()
```

Cluster 6

Hide

```
clust6_genes <- filter(cluster.file, group == "6" )

clust6_gene_list <- select(clust6_genes, gene_id)

clust6_gene_list <- unlist(clust6_gene_list)



con <- which(clust6_gene_list %in% row.names(output_df))
clust6_gene_list <- clust6_gene_list[con]
mat6 <- assay(rld)[clust6_gene_list,]

df6 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df6) <- "Condition"
row.names(df6) <- sampleTable$sampleName
mat6 <- mat6 - rowMeans(mat6)

#png("heatmap_clust6.png")
pheatmap(mat6, annottation_col=df6, cluster_rows=T, show_rownames=F, cluster_cols=F)
```



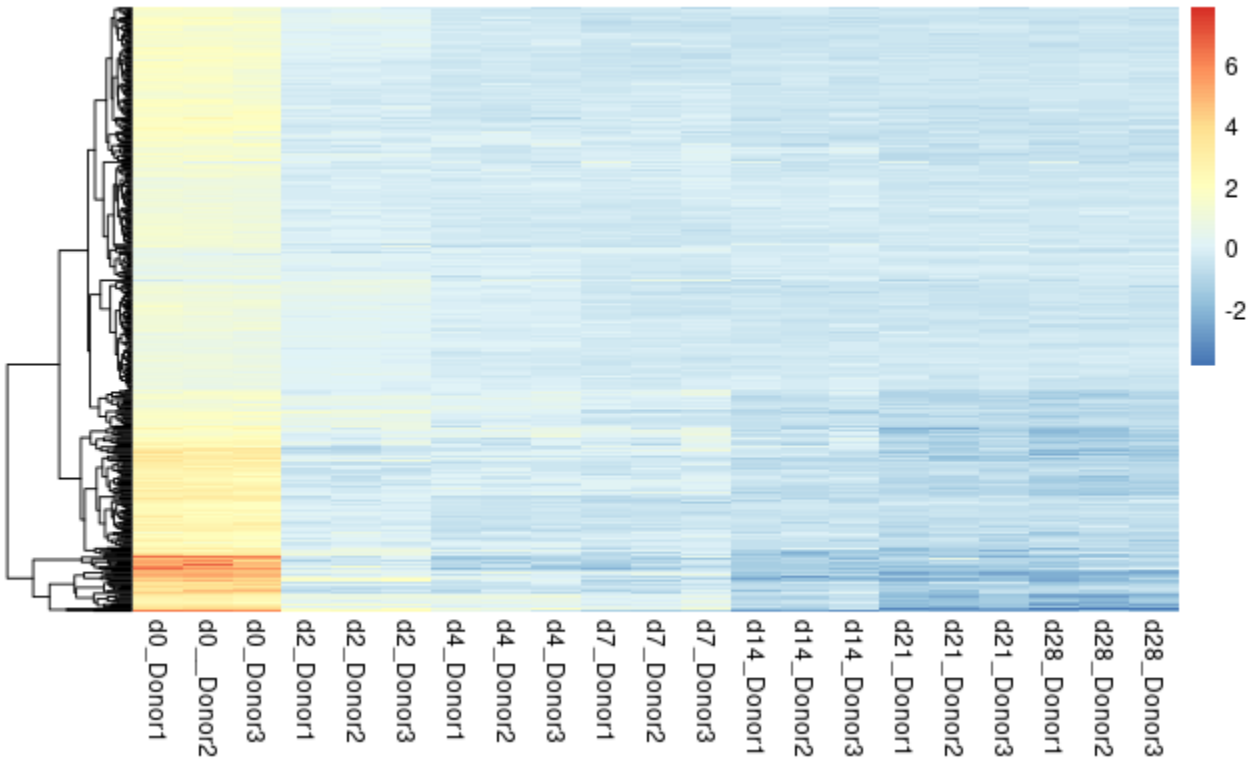
Hide

```
#dev.off()
```

Cluster 7

Hide

```
clust7_genes <- filter(cluster.file, group == "7" )

clust7_gene_list <- select(clust7_genes, gene_id)

clust7_gene_list <- unlist(clust7_gene_list)


#some of the genes are missing from the dds but in the clustering file from the xls
x?!?! had to remove those 2
con <- which(clust7_gene_list %in% row.names(output_df))
clust7_gene_list <- clust7_gene_list[con]

mat7 <- assay(rld)[clust7_gene_list,]

df7 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df7) <- "Condition"
row.names(df7) <- sampleTable$sampleName
mat7 <- mat7 - rowMeans(mat7)

#png("heatmap_clust7.png")
pheatmap(mat7, annottation_col=df7, cluster_rows=T, show_rownames=F, cluster_cols=F)
```



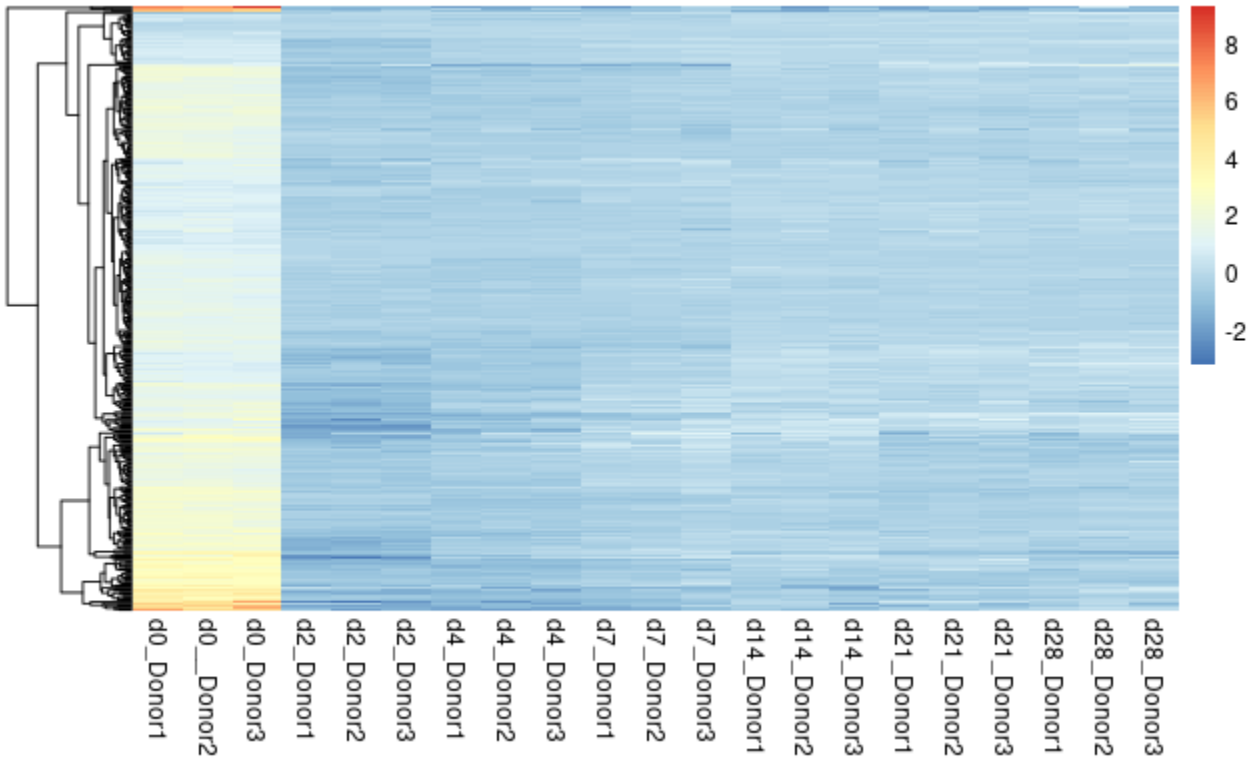
Hide

```
#dev.off()
```

Cluster 8

Hide

```
clust8_genes <- filter(cluster.file, group == "8" )

clust8_gene_list <- select(clust8_genes, gene_id)

clust8_gene_list <- unlist(clust8_gene_list)


con <- which(clust8_gene_list %in% row.names(output_df))
clust8_gene_list <- clust8_gene_list[con]

mat8 <- assay(rld)[clust8_gene_list,]

df8 <- as.data.frame(colData(dds)[,c("condition")])
colnames(df8) <- "Condition"
row.names(df8) <- sampleTable$sampleName
mat8 <- mat8 - rowMeans(mat8)

#png("heatmap_clust8.png")
pheatmap(mat8, annottation_col=df8, cluster_rows=T, show_rownames=F, cluster_cols=F)
```



Hide

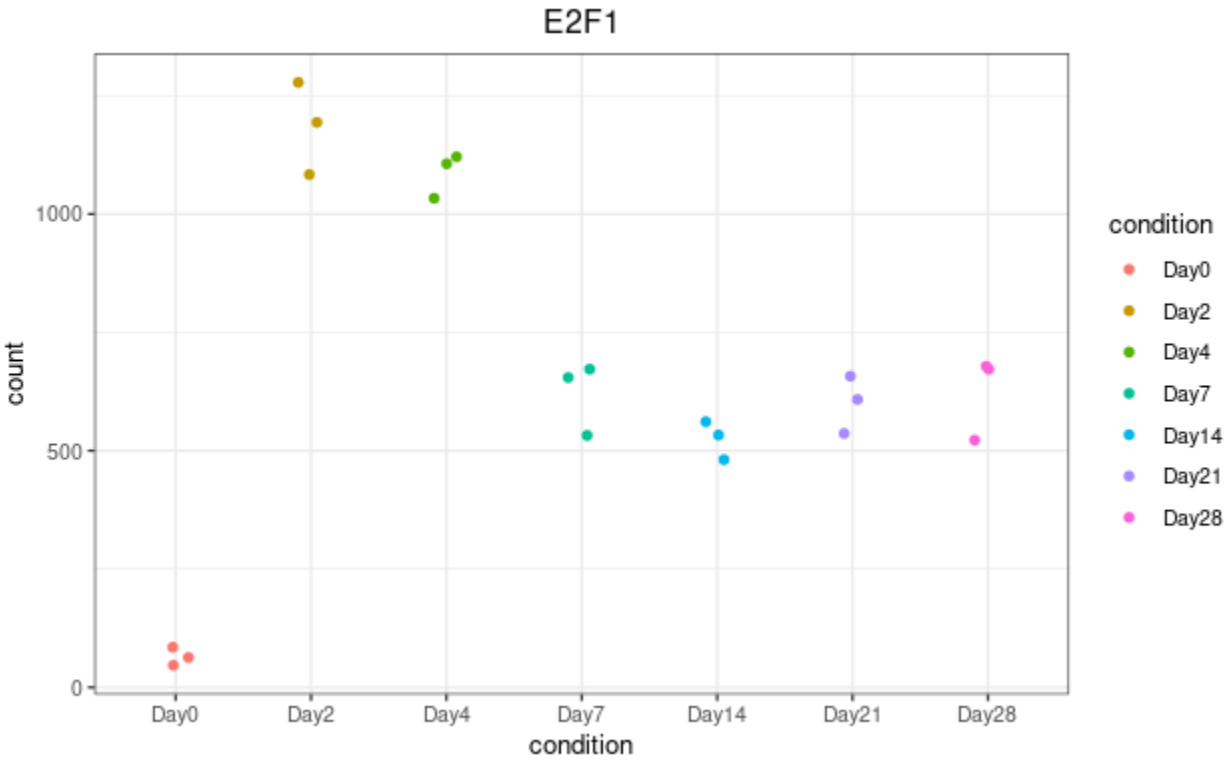
```
#dev.off()
```

Normalized Counts

E2F1

Hide

```
lookie <- cluster.file %>%
  filter(gene_name == "E2F1") %>%
  pull(gene_id)


d <- plotCounts(dds, gene= lookie, intgroup="condition", returnData = TRUE)

#png("E2F1.png")
ggplot(d, aes(x = condition, y = count, color = condition)) +
    geom_point(position=position_jitter(w = 0.1,h = 0)) +

    theme_bw() +
    ggtitle("E2F1") +
    theme(plot.title = element_text(hjust = 0.5))
```



Hide

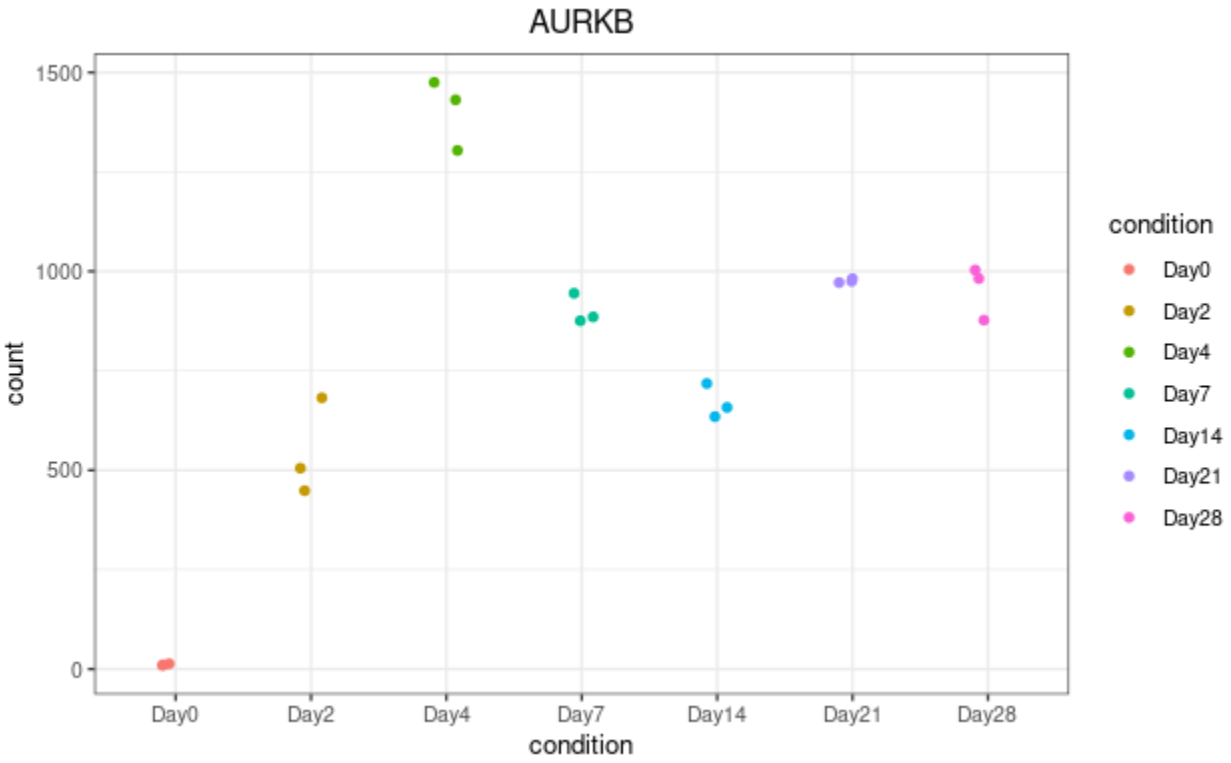
```
#dev.off()
```

Normalized plot count for E2F1, a transcription activator involved in cell cycle regulation [5], showing peak upregulation 2 days post infection and maintained at a lower rate afterwards. This gene is from cluster 1 and exhibits the same pattern of temporal expression as those in the rest of the cluster.

AURKB

Hide

```
lookie <- cluster.file %>%
  filter(gene_name == "AURKB") %>%
  pull(gene_id)


d <- plotCounts(dds, gene= lookie, intgroup="condition", returnData = TRUE)

#png("AURKB.png")
ggplot(d, aes(x = condition, y = count, color = condition)) +
    geom_point(position=position_jitter(w = 0.1,h = 0)) +

    theme_bw() +
    ggtitle("AURKB") +
    theme(plot.title = element_text(hjust = 0.5))
```



Hide

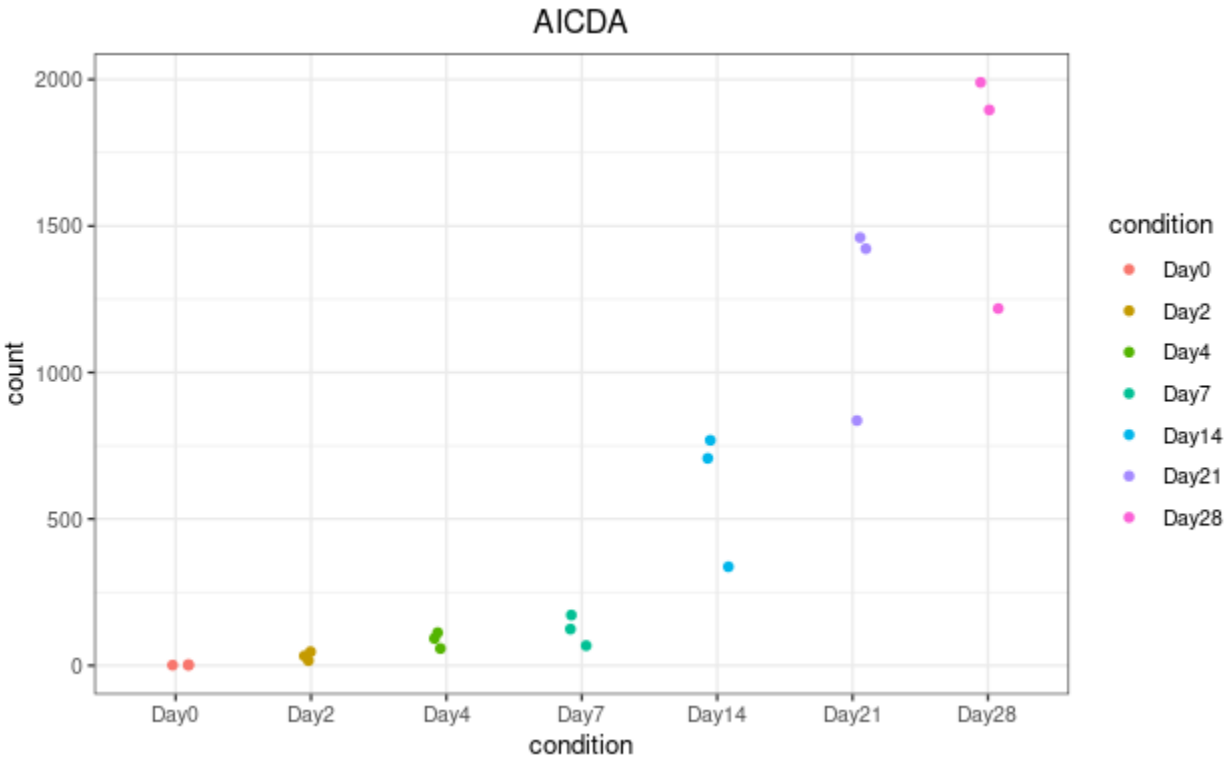
```
#dev.off()
```

Normalized plot count for AURKB. AURKB is a member of serine/threonine kinases necessary for the control of mitosis. [1]. As with other genes found in cluster 2, upregulation can be seen on Day 2 but peaks at Day 4 and then are maintained at a slightly lower level afterwards.

AICDA

Hide

```
lookie <- cluster.file %>%
  filter(gene_name == "AICDA") %>%
  pull(gene_id)


d <- plotCounts(dds, gene= lookie, intgroup="condition", returnData = TRUE)

#png("AICDA.png")
ggplot(d, aes(x = condition, y = count, color = condition)) +
    geom_point(position=position_jitter(w = 0.1,h = 0)) +

    theme_bw() +
    ggtitle("AICDA") +
    theme(plot.title = element_text(hjust = 0.5))
```



Hide

```
#dev.off()
```

Normalized plot count for AICDA, Activation induced cytidine deaminase, which is found in cluster 3. This gene is responsible for somatic hypermutation, class switch and chromosome mutation. Other genes such as EBNAs can induce AICDA expression [3].

Future Directions

Unfortunately, due to summer program time constraints, only the first part of the paper was able to be replicated. With more time, the following goals would be addressed:

- Sequence the EBV genome
    - Observe differentially expressed genes throughout infection
- Understand the clustering algorithm used
    - Not enough computational power originally
- Dive deeper into gene ontology and the connection that up/down regulation mean for the cells and therefore disease progression

The overall goal of the original paper was to understand gene expression during infection to better understand what may be happening during EBV caused oncogenesis. Understanding this process would likely aid in the design of targeted therapeutics with better effectiveness than current therapies. However, the massive size of the data does not lend itself well to current analysis approaches. More classification will be needed before further investigation can begin.

References

1 Wang, C., Li, D., Zhang, L., Jiang, S., Liang, J., Narita, Y., … & Zhao, B. (2019). RNA sequencing analyses of gene expression during Epstein-Barr virus infection of primary B lymphocytes. Journal of virology, 93(13), e00226-19.

2 Lundy, S. K., Klinker, M. W., & Fox, D. A. (2015). Killer B lymphocytes and their fas ligand positive exosomes as inducers of immune tolerance. Frontiers in immunology, 6, 122.

3 Casellas, R., Basu, U., Yewdell, W. T., Chaudhuri, J., Robbiani, D. F., & Di Noia, J. M. (2016). Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. Nature Reviews Immunology, 16(3), 164-176.

4 Zhao, S., Zhang, B., Zhang, Y., Gordon, W., Du, S., Paradis, T., Vincent, M., & Schack, D. v. (2016). Bioinformatics for RNA-Seq Data Analysis. In (Ed.), Bioinformatics - Updated Features and Applications. IntechOpen.

5 Martínez-Balbás, M. A., Bauer, U. M., Nielsen, S. J., Brehm, A., & Kouzarides, T. (2000). Regulation of E2F1 activity by acetylation. The EMBO journal, 19(4), 662-671. https://doi.org/10.5772/63267 (https://doi.org/10.5772/63267)

Acknowledgements

Thank you to the Summer Bioinformatics Fellowship at The RNA Institute for teaching me all the skills and providing the support needed for this project. A special thank you to my group leaders Noah Lefever and Sweta Vangaveti and as always, I couldn't do anything without the guidance of my PI Alan Chen and the support of the rest of the Chen Lab.

Contact information: ijdengos@gmail.com (mailto:ijdengos@gmail.com)