Yang\_data\_analysis

Load required packages

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(broom)

load data.

baseJ\_KO\_gene\_change <- readRDS("../../data/processed\_data/baseJ\_KO\_gene\_change.rds")  
H3V\_KO\_gene\_change <- readRDS("../../data/processed\_data/H3V\_KO\_gene\_change.rds")  
double\_KO\_gene\_change <- readRDS("../../data/processed\_data/double\_KO\_gene\_change.rds")   
life\_stage\_gene\_expression <- readRDS("../../data/processed\_data/life\_stage\_gene\_expression.rds")

First, I ask if genes whose expressions were affected by base J knockout are related to certain stages of T. brucei life cycle.

baseJ\_KO\_gene\_change

## # A tibble: 36 x 10  
## mutant name orientation `wt value` `mutant value` annotation `Q value`  
## <chr> <chr> <chr> <dbl> <dbl> <chr> <dbl>  
## 1 J<U+2206> Tb92~ sense 3.34 16.8 expressio~ 0.00431  
## 2 J<U+2206> Tb92~ sense 3.13 16.8 expressio~ 0.00140  
## 3 J<U+2206> Tb92~ sense 4.50 25.7 expressio~ 0.00379  
## 4 J<U+2206> Tb92~ sense 0.394 5.86 hypotheti~ 0.00431  
## 5 J<U+2206> Tb92~ sense 0.0789 0.885 hypotheti~ 0.0304   
## 6 J<U+2206> Tb92~ sense 3.56 17.6 receptor-~ 0.00431  
## 7 J<U+2206> Tb92~ sense 1.10 12.0 hypotheti~ 0.0191   
## 8 J<U+2206> Tb92~ sense 0.645 9.26 hypotheti~ 0.0402   
## 9 J<U+2206> Tb92~ sense 1.31 6.48 variant s~ 0.0101   
## 10 J<U+2206> Tb92~ sense 2.96 14.5 hypotheti~ 0.0187   
## # ... with 26 more rows, and 3 more variables: `distance to 5' HT  
## # SSR` <chr>, `distance to 3' HT SSR` <chr>, `distance to cSSR  
## # edge` <chr>

life\_stage\_gene\_expression

## # A tibble: 9,694 x 10  
## `Gene ID` `Mean Slender` `Mean Stumpy` `Mean Early` `Mean Late`  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Tb927.4.~ 3.77 4.17 3.99 2.06  
## 2 Tb927.7.~ 127. 121. 175. 153.   
## 3 Tb927.8.~ 2.3 9.75 1.7 1.02  
## 4 Tb927.8.~ 61.8 40.7 63.6 54.2   
## 5 Tb927.11~ 70.7 37.0 106. 74.5   
## 6 Tb927.11~ 64.0 62.9 188. 127.   
## 7 Tb927.4.~ 127. 93.9 75.0 42.2   
## 8 Tb927.7.~ 253. 319. 391. 280.   
## 9 Tb927.7.~ 252. 315. 393. 288.   
## 10 Tb927.11~ 331. 275. 404. 259.   
## # ... with 9,684 more rows, and 5 more variables: `Product  
## # Description` <chr>, `FC St/Sl` <dbl>, `FC St/Early` <dbl>, `FC  
## # St/Late` <dbl>, `FC Early/Late` <dbl>

BaseJ\_related\_genes <- baseJ\_KO\_gene\_change$name  
cluster1 <- life\_stage\_gene\_expression %>%  
 filter(`Gene ID` %in% BaseJ\_related\_genes)  
cluster1

## # A tibble: 27 x 10  
## `Gene ID` `Mean Slender` `Mean Stumpy` `Mean Early` `Mean Late`  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Tb927.3.~ 6.84 25 147. 42.5   
## 2 Tb927.1.~ 7.42 21.4 12.8 3.07  
## 3 Tb927.2.~ 19.8 44.1 17.4 11.8   
## 4 Tb927.2.~ 12.1 33.9 14.0 4.61  
## 5 Tb927.10~ 0.44 0.46 0.55 0.5   
## 6 Tb927.10~ 0.44 0.46 0.55 0.5   
## 7 Tb927.10~ 6.33 3.24 0.55 1.01  
## 8 Tb927.11~ 2.74 0.92 1.68 1.02  
## 9 Tb927.11~ 0.44 0.46 0.55 0.5   
## 10 Tb927.2.~ 1.79 5.57 1.13 1.02  
## # ... with 17 more rows, and 5 more variables: `Product  
## # Description` <chr>, `FC St/Sl` <dbl>, `FC St/Early` <dbl>, `FC  
## # St/Late` <dbl>, `FC Early/Late` <dbl>

To see if any of these genes are up-regulationed by a certain cell cycle stage, I need to compare the ratios of expression values between different life stages, as presented in the last columns in cluster1 table. If the ratio for a specific gene is greater than 2, then I consider that gene to be up-regulated.

cluster1 %>%  
 filter(`FC St/Sl` > 2) %>%  
 select(`Gene ID`) %>%  
 count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 10

cluster1 %>%  
 filter(`FC St/Early` > 2) %>%  
 select(`Gene ID`) %>%  
 count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 10

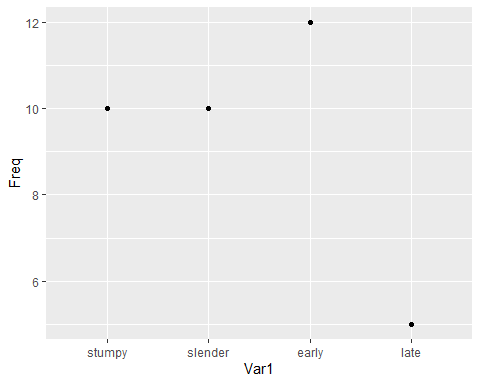
cluster1 %>%  
 filter(`FC St/Late` > 2) %>%  
 select(`Gene ID`) %>%  
 count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 12

cluster1 %>%  
 filter(`FC Early/Late` > 2) %>%  
 select(`Gene ID`) %>%  
 count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 5

number\_genes <- c(10, 10, 12, 5)  
genes\_affected\_baseJ\_KO <- matrix(number\_genes, ncol = 1, byrow = TRUE)  
rownames(genes\_affected\_baseJ\_KO) <- c("stumpy", "slender", "early", "late")  
genes\_affected\_baseJ\_KO <- as.table(genes\_affected\_baseJ\_KO)  
genes\_affected\_baseJ\_KO <- as.data.frame(genes\_affected\_baseJ\_KO)  
genes\_affected\_baseJ\_KO %>%  
 ggplot(aes(Var1, Freq)) + geom\_point()



From this plot, it seems that base J KO had the least impact on gene expression of late procyclic form of T. brucei.