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)  
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v tibble 2.1.3 v purrr 0.3.2  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

load data.

life\_stage\_gene\_expression <- readRDS("../../data/processed\_data/life\_stage\_gene\_expression.rds")  
Base\_J\_H3V\_gene\_expression <- readRDS("../../data/processed\_data/Base\_J\_H3V\_gene\_expression.rds")

life\_stage\_gene\_expression

## # A tibble: 9,694 x 5  
## `Gene ID` `Mean Slender` `Mean Stumpy` `Mean Early` `Mean Late`  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Tb927.4.4520 3.77 4.17 3.99 2.06  
## 2 Tb927.7.3810 127. 121. 175. 153.   
## 3 Tb927.8.7820 2.3 9.75 1.7 1.02  
## 4 Tb927.8.2310 61.8 40.7 63.6 54.2   
## 5 Tb927.11.15150 70.7 37.0 106. 74.5   
## 6 Tb927.11.1560 64.0 62.9 188. 127.   
## 7 Tb927.4.360 127. 93.9 75.0 42.2   
## 8 Tb927.7.1320 253. 319. 391. 280.   
## 9 Tb927.7.1340 252. 315. 393. 288.   
## 10 Tb927.11.9530 331. 275. 404. 259.   
## # ... with 9,684 more rows

glimpse(life\_stage\_gene\_expression)

## Observations: 9,694  
## Variables: 5  
## $ `Gene ID` <chr> "Tb927.4.4520", "Tb927.7.3810", "Tb927.8.7820",...  
## $ `Mean Slender` <dbl> 3.77, 127.05, 2.30, 61.79, 70.74, 64.05, 127.35...  
## $ `Mean Stumpy` <dbl> 4.17, 121.06, 9.75, 40.70, 37.02, 62.92, 93.88,...  
## $ `Mean Early` <dbl> 3.99, 175.37, 1.70, 63.58, 105.90, 188.21, 74.9...  
## $ `Mean Late` <dbl> 2.06, 153.45, 1.02, 54.15, 74.52, 126.94, 42.21...

First, I want to know which genes have up-regulated expression in each of four T. brucei life stages. To answer this question, if a certain gene has the highest transcirpts read in one of the life stage, then I regard the gene to be over-represented in that life stage. For exmaple, for gene Tb927.4.4520, the transcripts reads for slender form, stumpy form, Early and late procyclic form are 3.77, 4.17, 3.99, and 2.06, respectivelt. This gene is upregulated in the stumpy stage according to the definition. I need to do some data transformation to find out in which life stage a gene has the highest expression.

life\_stage\_gene\_expression <- life\_stage\_gene\_expression %>%  
 gather("life stage", "value", 2, 3, 4, 5, convert = TRUE)  
life\_stage\_gene\_expression

## # A tibble: 38,776 x 3  
## `Gene ID` `life stage` value  
## <chr> <chr> <dbl>  
## 1 Tb927.4.4520 Mean Slender 3.77  
## 2 Tb927.7.3810 Mean Slender 127.   
## 3 Tb927.8.7820 Mean Slender 2.3   
## 4 Tb927.8.2310 Mean Slender 61.8   
## 5 Tb927.11.15150 Mean Slender 70.7   
## 6 Tb927.11.1560 Mean Slender 64.0   
## 7 Tb927.4.360 Mean Slender 127.   
## 8 Tb927.7.1320 Mean Slender 253.   
## 9 Tb927.7.1340 Mean Slender 252.   
## 10 Tb927.11.9530 Mean Slender 331.   
## # ... with 38,766 more rows

life\_stage\_gene\_expression <- life\_stage\_gene\_expression %>%  
 group\_by(`Gene ID`) %>%  
 top\_n(1, value)  
life\_stage\_gene\_expression

## # A tibble: 9,704 x 3  
## # Groups: Gene ID [9,694]  
## `Gene ID` `life stage` value  
## <chr> <chr> <dbl>  
## 1 Tb927.4.360 Mean Slender 127.   
## 2 Tb927.11.16430 Mean Slender 133.   
## 3 Tb927.10.14530 Mean Slender 210.   
## 4 Tb927.11.11490 Mean Slender 117.   
## 5 Tb927.7.5540 Mean Slender 104.   
## 6 Tb927.10.6450 Mean Slender 110.   
## 7 Tb927.11.11390 Mean Slender 52.2  
## 8 Tb927.7.1530 Mean Slender 111.   
## 9 Tb927.9.5370 Mean Slender 45.6  
## 10 Tb927.3.5630 Mean Slender 94.6  
## # ... with 9,694 more rows

Now, I will split the table into 4 small stables, each including genes up-regulated in a T. brucei life stage.

Gene\_upregualted\_Slender <- life\_stage\_gene\_expression %>%  
 filter(`life stage` == "Mean Slender")  
Gene\_upregualted\_Stumpy <- life\_stage\_gene\_expression %>%  
 filter(`life stage` == "Mean Stumpy")  
Gene\_upregualted\_Early <- life\_stage\_gene\_expression %>%  
 filter(`life stage` == "Mean Early")  
Gene\_upregualted\_Late <- life\_stage\_gene\_expression %>%  
 filter(`life stage` == "Mean Late")

take a look at them

Gene\_upregualted\_Slender

## # A tibble: 2,612 x 3  
## # Groups: Gene ID [2,612]  
## `Gene ID` `life stage` value  
## <chr> <chr> <dbl>  
## 1 Tb927.4.360 Mean Slender 127.   
## 2 Tb927.11.16430 Mean Slender 133.   
## 3 Tb927.10.14530 Mean Slender 210.   
## 4 Tb927.11.11490 Mean Slender 117.   
## 5 Tb927.7.5540 Mean Slender 104.   
## 6 Tb927.10.6450 Mean Slender 110.   
## 7 Tb927.11.11390 Mean Slender 52.2  
## 8 Tb927.7.1530 Mean Slender 111.   
## 9 Tb927.9.5370 Mean Slender 45.6  
## 10 Tb927.3.5630 Mean Slender 94.6  
## # ... with 2,602 more rows

Gene\_upregualted\_Stumpy

## # A tibble: 2,853 x 3  
## # Groups: Gene ID [2,853]  
## `Gene ID` `life stage` value  
## <chr> <chr> <dbl>  
## 1 Tb927.4.4520 Mean Stumpy 4.17  
## 2 Tb927.8.7820 Mean Stumpy 9.75  
## 3 Tb927.11.11680 Mean Stumpy 577.   
## 4 Tb927.3.950 Mean Stumpy 56.9   
## 5 Tb927.5.3950 Mean Stumpy 115.   
## 6 Tb927.5.980 Mean Stumpy 303.   
## 7 Tb927.8.6970 Mean Stumpy 134.   
## 8 Tb927.4.5350 Mean Stumpy 135.   
## 9 Tb927.3.3070 Mean Stumpy 58.3   
## 10 Tb927.7.4710 Mean Stumpy 69.3   
## # ... with 2,843 more rows

Gene\_upregualted\_Early

## # A tibble: 2,819 x 3  
## # Groups: Gene ID [2,819]  
## `Gene ID` `life stage` value  
## <chr> <chr> <dbl>  
## 1 Tb927.7.3810 Mean Early 175.   
## 2 Tb927.8.2310 Mean Early 63.6  
## 3 Tb927.11.15150 Mean Early 106.   
## 4 Tb927.11.1560 Mean Early 188.   
## 5 Tb927.7.1320 Mean Early 391.   
## 6 Tb927.7.1340 Mean Early 393.   
## 7 Tb927.11.9530 Mean Early 404.   
## 8 Tb927.11.6870 Mean Early 145.   
## 9 Tb927.6.1570 Mean Early 175.   
## 10 Tb927.5.300b Mean Early 16.9  
## # ... with 2,809 more rows

Gene\_upregualted\_Late

## # A tibble: 1,420 x 3  
## # Groups: Gene ID [1,420]  
## `Gene ID` `life stage` value  
## <chr> <chr> <dbl>  
## 1 Tb927.8.6060 Mean Late 868.   
## 2 Tb11.v5.1012 Mean Late 41.4  
## 3 Tb927.9.10580 Mean Late 74.6  
## 4 Tb927.6.4540 Mean Late 132.   
## 5 Tb927.8.2480 Mean Late 43.4  
## 6 Tb927.3.1840 Mean Late 213.   
## 7 Tb927.3.3340 Mean Late 106.   
## 8 Tb927.10.13300 Mean Late 59.1  
## 9 Tb927.11.10870 Mean Late 41.3  
## 10 Tb927.10.5360 Mean Late 869.   
## # ... with 1,410 more rows

Save them to the result folder.

saveRDS(Gene\_upregualted\_Early, file = "../../results/Gene\_upregulated\_Early.rds")  
saveRDS(Gene\_upregualted\_Late, file = "../../results/Gene\_upregulated\_Late.rds")  
saveRDS(Gene\_upregualted\_Slender, file = "../../results/Gene\_upregulated\_Slender.rds")  
saveRDS(Gene\_upregualted\_Stumpy, file = "../../results/Gene\_upregulated\_Stumpy.rds")

Base\_J\_H3V\_gene\_expression

## # A tibble: 237 x 4  
## mutant name `wt value` `mutant value`  
## <chr> <chr> <dbl> <dbl>  
## 1 J<U+2206> Tb927.1.275 3.34 16.8   
## 2 J<U+2206> Tb927.2.660 3.13 16.8   
## 3 J<U+2206> Tb927.2.910 4.50 25.7   
## 4 J<U+2206> Tb927.2.1290 0.394 5.86   
## 5 J<U+2206> Tb927.3.4010 0.0789 0.885  
## 6 J<U+2206> Tb927.4.130 3.56 17.6   
## 7 J<U+2206> Tb927.4.140 1.10 12.0   
## 8 J<U+2206> Tb927.4.5090 0.645 9.26   
## 9 J<U+2206> Tb927.5.3990 1.31 6.48   
## 10 J<U+2206> Tb927.6.110 2.96 14.5   
## # ... with 227 more rows

First, I want to ask the question that if there is any gene that gets upregulated in both J KO and H3V KO.

Table\_J <- Base\_J\_H3V\_gene\_expression %>%  
 filter(mutant == "J∆")  
Table\_H <- Base\_J\_H3V\_gene\_expression %>%  
 filter(mutant == "H3.V∆")  
Table\_J %>%  
 filter(name %in% Table\_H$name)

## # A tibble: 0 x 4  
## # ... with 4 variables: mutant <chr>, name <chr>, `wt value` <dbl>,  
## # `mutant value` <dbl>

Table\_H

## # A tibble: 0 x 4  
## # ... with 4 variables: mutant <chr>, name <chr>, `wt value` <dbl>,  
## # `mutant value` <dbl>

The result indicates that 2 genes were upregulated in both base J and H3V KO.

affected\_genes <- matrix(c(34, 6, 2), ncol = 3, byrow = TRUE)  
colnames(affected\_genes) <- c("base J alone", "H3V alone", "either")  
rownames(affected\_genes) <- c("number of genes")  
affected\_genes <- as.table(affected\_genes)  
affected\_genes

## base J alone H3V alone either  
## number of genes 34 6 2

saveRDS(affected\_genes, file = "../../results/affected\_genes.rds")

The list of genes whose expression got upregulated in base J and H3V double knockout.

Table\_JH <- Base\_J\_H3V\_gene\_expression %>%  
 filter(mutant == "J∆H3.V∆")

Next, I ask the question that whether the genes whose expression get upregulated in Base J and/or H3.V KO are enriched in specific T. brucei life stage.

Table\_J %>%  
 filter(name %in% Gene\_upregualted\_Early$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_J %>%  
 filter(name %in% Gene\_upregualted\_Late$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_J %>%  
 filter(name %in% Gene\_upregualted\_Slender$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_J %>%  
 filter(name %in% Gene\_upregualted\_Stumpy$`Gene ID`) %>%  
 nrow()

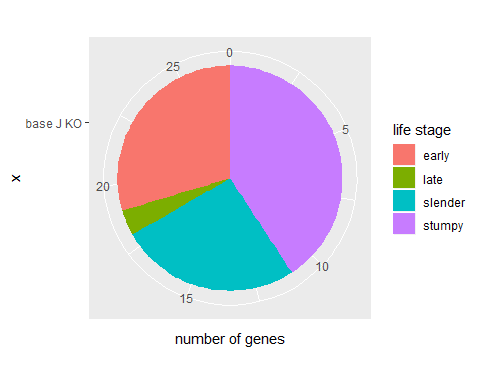
## [1] 0

It shows here that Base J KO affected 8, 1, 7, 11 genes whose expressions are the highest in early, late procyclic form, slender form and stumpy form respectively. Make a pie chart to visualize the results.

base\_J <- matrix(c(8, 1, 7, 11), ncol = 4, byrow = TRUE)  
colnames(base\_J) <- c("early", "late", "slender", "stumpy")  
rownames(base\_J) <- c("number of genes")  
base\_J <- as.data.frame(base\_J)  
base\_J <- base\_J %>%  
 gather("life stage", "number of genes", 1, 2, 3, 4)  
base\_J

## life stage number of genes  
## 1 early 8  
## 2 late 1  
## 3 slender 7  
## 4 stumpy 11

pie\_base\_J <- base\_J %>%  
 ggplot(aes(x = "base J KO", y = `number of genes`, fill = `life stage`)) + geom\_bar(width = 1, stat = "identity") +coord\_polar("y", start=0)  
pie\_base\_J



ggsave(filename = "../../results/Pie\_base\_J.png", plot = pie\_base\_J)

## Saving 5 x 4 in image

Repeat the same analysis for H3V single KO.

Table\_H %>%  
 filter(name %in% Gene\_upregualted\_Early$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_H %>%  
 filter(name %in% Gene\_upregualted\_Late$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_H %>%  
 filter(name %in% Gene\_upregualted\_Slender$`Gene ID`) %>%  
 nrow()

## [1] 0

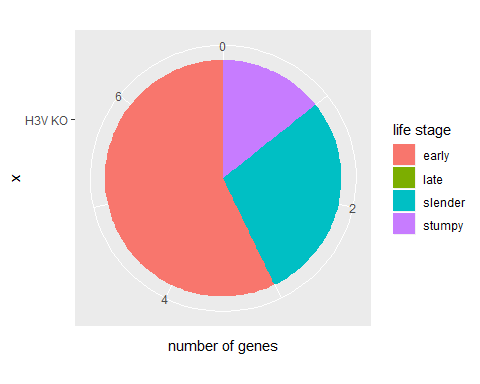
Table\_H %>%  
 filter(name %in% Gene\_upregualted\_Stumpy$`Gene ID`) %>%  
 nrow()

## [1] 0

H3V <- matrix(c(4, 0, 2, 1), ncol = 4, byrow = TRUE)  
colnames(H3V) <- c("early", "late", "slender", "stumpy")  
rownames(H3V) <- c("number of genes")  
H3V <- as.data.frame(H3V)  
H3V <- H3V %>%  
 gather("life stage", "number of genes", 1, 2, 3, 4)  
H3V

## life stage number of genes  
## 1 early 4  
## 2 late 0  
## 3 slender 2  
## 4 stumpy 1

pie\_H3V <- H3V %>%  
 ggplot(aes(x = "H3V KO", y = `number of genes`, fill = `life stage`)) + geom\_bar(width = 1, stat = "identity") +coord\_polar("y", start=0)  
pie\_H3V



ggsave(filename = "../../results/Pie\_H3V.png", plot = pie\_H3V)

## Saving 5 x 4 in image

Repeat the analysis for base J and H3V double knockout.

Table\_JH %>%  
 filter(name %in% Gene\_upregualted\_Early$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_JH %>%  
 filter(name %in% Gene\_upregualted\_Late$`Gene ID`) %>%  
 nrow()

## [1] 0

Table\_JH %>%  
 filter(name %in% Gene\_upregualted\_Slender$`Gene ID`) %>%  
 nrow()

## [1] 0

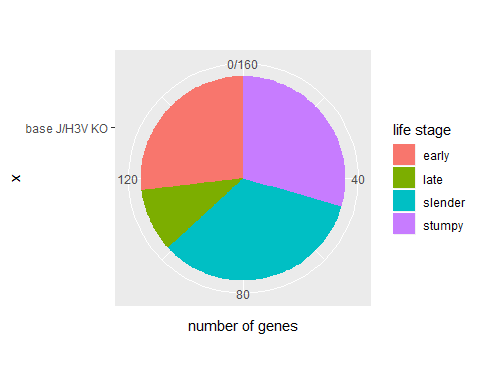
Table\_JH %>%  
 filter(name %in% Gene\_upregualted\_Stumpy$`Gene ID`) %>%  
 nrow()

## [1] 0

JH3V <- matrix(c(43, 16, 54, 47), ncol = 4, byrow = TRUE)  
colnames(JH3V) <- c("early", "late", "slender", "stumpy")  
rownames(JH3V) <- c("number of genes")  
JH3V <- as.data.frame(JH3V)  
JH3V <- JH3V %>%  
 gather("life stage", "number of genes", 1, 2, 3, 4)  
JH3V

## life stage number of genes  
## 1 early 43  
## 2 late 16  
## 3 slender 54  
## 4 stumpy 47

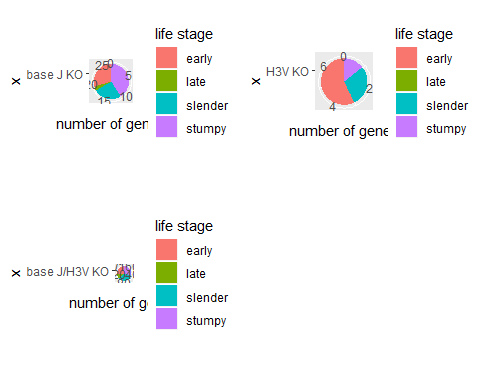
pie\_JH3V <- JH3V %>%  
 ggplot(aes(x = "base J/H3V KO", y = `number of genes`, fill = `life stage`)) + geom\_bar(width = 1, stat = "identity") +coord\_polar("y", start=0)  
pie\_JH3V



ggsave(filename = "../../results/Pie\_JH3V.png", plot = pie\_JH3V)

## Saving 5 x 4 in image

result\_figure <- cowplot::plot\_grid(pie\_base\_J, pie\_H3V, pie\_JH3V)  
result\_figure



ggsave(filename = "../../results/result\_figure.png", plot = result\_figure)

## Saving 5 x 4 in image