data processing

packages loading

library(readxl)  
library(tidyverse)

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

## v ggplot2 3.2.1 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## 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 the data.

life\_stage\_gene\_expression <- readxl::read\_excel("../../data/raw\_data/12864\_2018\_4600\_MOESM2\_ESM.xlsx")

The “life\_stage\_gene\_expression” file contains mean expression values (reads per million) for all transcripts from four life-cycle stages of Trypanosoma brucei brucei (slender form, stumpy form, ealy and late procyclic forms). cDNA libraries were prepared from poly(A)-selected RNA, and were sequenced using Illumina Hiseq sequencing system. Reads obtained from the RNA-seq were aligned to the T. b. brucei 927 reference genome version 5. Data was obtained from ***Naguleswaran, A., Doiron, N., & Roditi, I. (2018). RNA-Seq analysis validates the use of culture-derived Trypanosoma brucei and provides new markers for mammalian and insect life-cycle stages. BMC genomics, 19(1), 227.***

glimpse(life\_stage\_gene\_expression)

## Observations: 9,694  
## Variables: 10  
## $ `Gene ID` <chr> "Tb927.4.4520", "Tb927.7.3810", "Tb927.8...  
## $ `Mean Slender` <dbl> 3.77, 127.05, 2.30, 61.79, 70.74, 64.05,...  
## $ `Mean Stumpy` <dbl> 4.17, 121.06, 9.75, 40.70, 37.02, 62.92,...  
## $ `Mean Early` <dbl> 3.99, 175.37, 1.70, 63.58, 105.90, 188.2...  
## $ `Mean Late` <dbl> 2.06, 153.45, 1.02, 54.15, 74.52, 126.94...  
## $ `Product Description` <chr> "'Cold-shock' DNA-binding domain contain...  
## $ `FC St/Sl` <dbl> 1.11, 0.95, 4.24, 0.66, 0.52, 0.98, 0.74...  
## $ `FC St/Early` <dbl> 1.05, 0.69, 5.74, 0.64, 0.35, 0.33, 1.25...  
## $ `FC St/Late` <dbl> 2.02, 0.79, 9.56, 0.75, 0.50, 0.50, 2.22...  
## $ `FC Early/Late` <dbl> 1.94, 1.14, 1.67, 1.17, 1.42, 1.48, 1.78...

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>

The file “life\_stage\_gene\_expression” contains 10 variables with 9694 observations. The variables include gene accession number (gene ID), description of each genes and mean transcript values from four life cycle stages.The last four columns are the calculated ratios of transcripts expression values between pairs of the 4 life stages. The table is tidy, and contains all the information I need.

Read another file into R.

Base\_J\_H3V\_gene\_expression <- readxl::read\_excel("../../data/raw\_data/journal.pgen.1005762.s012.XLSX")

glimpse(Base\_J\_H3V\_gene\_expression)

## Observations: 237  
## Variables: 10  
## $ mutant <chr> "J<U+2206>", "J<U+2206>", "J<U+2206>", "J<U+2206>", "J<U+2206>", "J<U+2206>", "J...  
## $ name <chr> "Tb927.1.275", "Tb927.2.660", "Tb927.2...  
## $ orientation <chr> "sense", "sense", "sense", "sense", "s...  
## $ `wt value` <dbl> 3.33511593, 3.12746650, 4.50293087, 0....  
## $ `mutant value` <dbl> 16.7892277, 16.7800227, 25.6751093, 5....  
## $ annotation <chr> "expression site-associated gene (ESAG...  
## $ `Q value` <dbl> 0.004311201, 0.001401953, 0.003788336,...  
## $ `distance to 5' HT SSR` <chr> "NA", "99493", "20113", "68938", "NA",...  
## $ `distance to 3' HT SSR` <chr> "9262", "8482", "9306", "793559", "229...  
## $ `distance to cSSR edge` <chr> "22961", "13830", "14609", "22068", "5...

Base\_J\_H3V\_gene\_expression

## # A tibble: 237 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 227 more rows, and 3 more variables: `distance to 5' HT  
## # SSR` <chr>, `distance to 3' HT SSR` <chr>, `distance to cSSR  
## # edge` <chr>

The file “Base\_J\_H3V\_gene\_expression” display the genes whose transcript levels were significant different (as analyzed by a T test using Benjamimi and Hochberg correction, p value < 0.05) between different mutant cells and wild type cells. The mutant cells used in the study include wild type cells, cells with base J knockout (KO), H3V KO, and cells with both base J and H3V KO. Similar to the study presented above, cDNA libraries were prepared form poly(A)-selected RNA and sequenced on an Illumina HiSeq 2000 sequencer. Reads obtained were aligned to the T. brucei 927 version 5 reference genome. Data was obtained from ***Schulz, D., Zaringhalam, M., Papavasiliou, F. N., & Kim, H. S. (2016). Base J and H3. V regulate transcriptional termination in Trypanosoma brucei. PLoS genetics, 12(1), e1005762.*** Again luckily, This table is tidy, except that there is special symbol in “mutant” column representing KO that cannot be displayed correctly here.

I want to do some cross comparison between table 1 and table 2. My primary question is whether base J or H3V ablation leads to upregulation of gene characteristic to certain T. brucei life stages. To make it easier for comparison, I will split table 2 to 3 tables, with each of them displaying genes with significantly different expression values than wild type cells in three different mutant cells.

baseJ\_KO\_gene\_change <- Base\_J\_H3V\_gene\_expression %>%  
 filter(mutant == "J∆")  
H3V\_KO\_gene\_change <- Base\_J\_H3V\_gene\_expression %>%  
 filter(mutant == "H3.V∆")  
double\_KO\_gene\_change <- Base\_J\_H3V\_gene\_expression %>%  
 filter(mutant == "J∆H3.V∆")

head(baseJ\_KO\_gene\_change)

## # A tibble: 0 x 10  
## # ... with 10 variables: mutant <chr>, name <chr>, orientation <chr>, `wt  
## # value` <dbl>, `mutant value` <dbl>, annotation <chr>, `Q value` <dbl>,  
## # `distance to 5' HT SSR` <chr>, `distance to 3' HT SSR` <chr>,  
## # `distance to cSSR edge` <chr>

head(H3V\_KO\_gene\_change)

## # A tibble: 0 x 10  
## # ... with 10 variables: mutant <chr>, name <chr>, orientation <chr>, `wt  
## # value` <dbl>, `mutant value` <dbl>, annotation <chr>, `Q value` <dbl>,  
## # `distance to 5' HT SSR` <chr>, `distance to 3' HT SSR` <chr>,  
## # `distance to cSSR edge` <chr>

head(double\_KO\_gene\_change)

## # A tibble: 0 x 10  
## # ... with 10 variables: mutant <chr>, name <chr>, orientation <chr>, `wt  
## # value` <dbl>, `mutant value` <dbl>, annotation <chr>, `Q value` <dbl>,  
## # `distance to 5' HT SSR` <chr>, `distance to 3' HT SSR` <chr>,  
## # `distance to cSSR edge` <chr>