Bee microbiome gene expression

Sasha Mikheyev

11/7/2019

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(sleuth)
library(readxl)
library(ggsignif)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
sample_id <- dir(file.path("data/kallisto"))
metadata <- read_xlsx("data/RNA_seq_sample_metafile.xlsx")
t2g <- read_tsv("data/gene2isoform.txt.gz", col_names = c("gene_id", "target_id"))
```

```
## Parsed with column specification:
## cols(
##   gene_id = col_double(),
##   target_id = col_character()
## )
```

```
genesAnnotation <- read_tsv("data/genes.txt", col_names = c("beebase", "target_id", "description"), col_

stressGenes <- read_xlsx("data/doublet.xlsx") %>% mutate(gene_id = as.character(NCBI_ID)) %>% dplyr::sel
treatments <- c("Tetracycline", "Glyphosate", "Chlorothalonil")
```

Pre-treatment samples only

We're going to look at post-stress samples to see if we have a similar pattern to the whole model. There were not enough samples for pre-treatment chlorothalonil, so we'll use the other two only as a check.

```r
waldPre <- function(trt = "Tetracycline") {
  # b > 0 genes were higher in co-evolved microbes
  dat <-  metadata %>% dplyr::filter((grepl(trt, treatment) | history == "control") & time_stress == "be
  so <- sleuth_prep(dat, extra_bootstrap_summary = T, target_mapping = t2g, aggregation_column = 'gene_
  so <- sleuth_fit(so, ~ history , 'full')
  so <- sleuth_wt(so, 'historystress_co_evolved')
  results <- sleuth_results(so, test = "historystress_co_evolved", pval_aggregate = F) %>% mutate(gene_
  return(left_join(results, genesAnnotation, by = c("gene_id" = "target_id"))  %>% dplyr::select(gene_i
}
```

```r
dgePre <- list()
for (trt in c("Tetracycline", "Glyphosate", "Chlorothalonil"))
    dgePre[[trt]] <- waldPre(trt)
```

```
## Warning in check_num_cores(num_cores): It appears that you are running Sleuth from within Rstudio.
## Because of concerns with forking processes from a GUI, 'num_cores' is being set to 1.
## If you wish to take advantage of multiple cores, please consider running sleuth from the command lin
```

```
## reading in kallisto results
```

```
## dropping unused factor levels
```

```
## .........
## normalizing est_counts
## 11524 targets passed the filter
## normalizing tpm
## merging in metadata
## summarizing bootstraps
## .........
## fitting measurement error models
## shrinkage estimation
## 2 NA values were found during variance shrinkage estimation due to mean observation values outside o
## The LOESS fit will be repeated using exact computation of the fitted surface to extrapolate the miss
## These are the target ids with NA values: NM_001011607.2, XM_016917951.1
## computing variance of betas
```

```
## Warning in check_num_cores(num_cores): It appears that you are running Sleuth from within Rstudio.
## Because of concerns with forking processes from a GUI, 'num_cores' is being set to 1.
## If you wish to take advantage of multiple cores, please consider running sleuth from the command lin
```

```
## reading in kallisto results
## dropping unused factor levels
## ........
## normalizing est_counts
## 12583 targets passed the filter
## normalizing tpm
## merging in metadata
## summarizing bootstraps
## ........
## fitting measurement error models
## shrinkage estimation
## 1 NA values were found during variance shrinkage estimation due to mean observation values outside o
## The LOESS fit will be repeated using exact computation of the fitted surface to extrapolate the miss
## These are the target ids with NA values: XM_006563653.3
## computing variance of betas
```

```
## Warning in check_num_cores(num_cores): It appears that you are running Sleuth from within Rstudio.
## Because of concerns with forking processes from a GUI, 'num_cores' is being set to 1.
## If you wish to take advantage of multiple cores, please consider running sleuth from the command line
```

```
## reading in kallisto results
## dropping unused factor levels
## .........
## normalizing est_counts
## 11309 targets passed the filter
## normalizing tpm
## merging in metadata
## summarizing bootstraps
## .........
## fitting measurement error models
## shrinkage estimation
## 4 NA values were found during variance shrinkage estimation due to mean observation values outside of
## The LOESS fit will be repeated using exact computation of the fitted surface to extrapolate the miss
## These are the target ids with NA values: XM_006566164.3, XM_016911284.2, NM_001011607.2, XM_01691795
## computing variance of betas
```

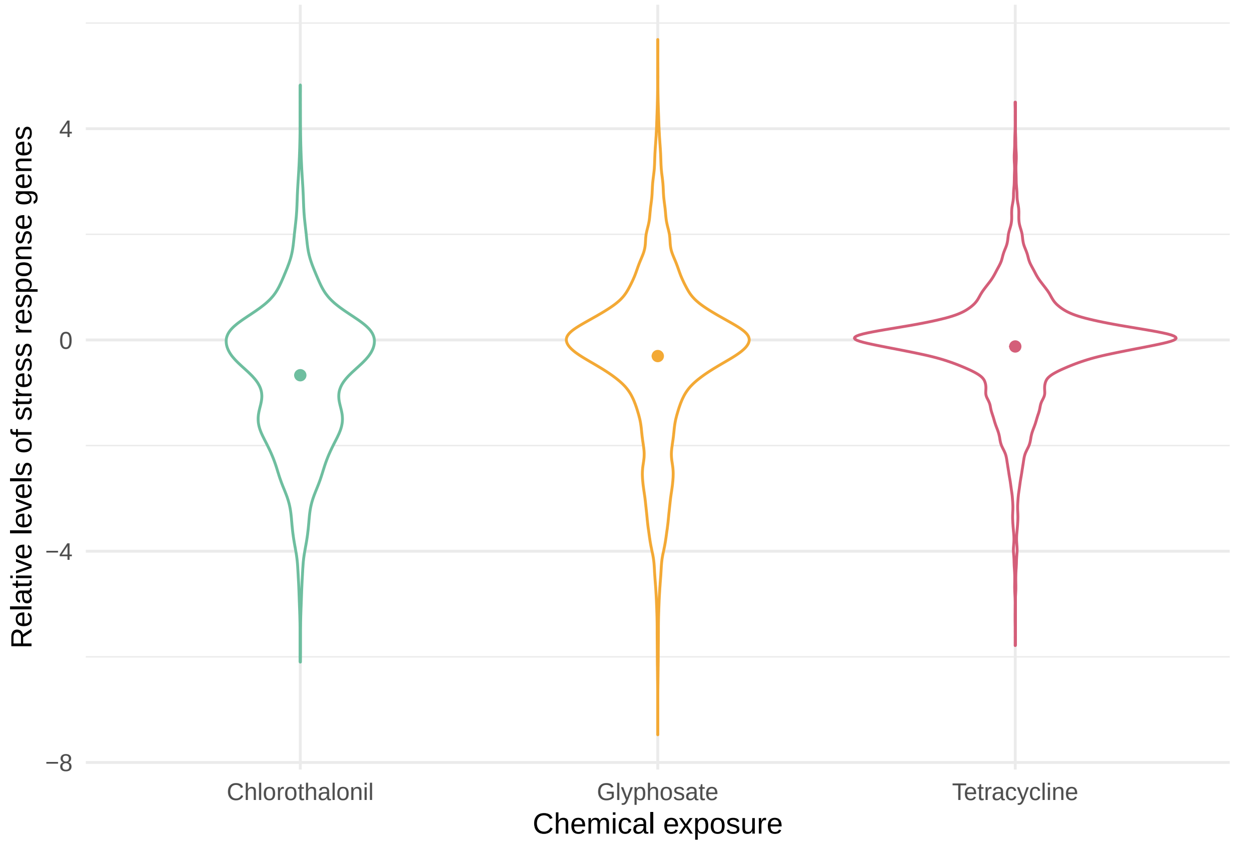
```r
(immuneGenesPre <- rbind(
  stressGenes %>% left_join(dgePre[["Chlorothalonil"]], by =  "gene_id" ) %>% mutate(treatment = "Chloro
  stressGenes %>% left_join(dgePre[["Tetracycline"]], by =  "gene_id" ) %>% mutate(treatment = "Tetracyc
  stressGenes %>% left_join(dgePre[["Glyphosate"]], by =  "gene_id" ) %>% mutate(treatment = "Glyphosate
) )%>%
  ggplot(aes(treatment, b, color = treatment))  + geom_violin() + theme_minimal() + stat_summary(fun = m
```

```
## Warning: Ignoring unknown parameters: fun
```

```
## Warning: Removed 26715 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 26715 rows containing non-finite values (stat_summary).
```

```
## No summary function supplied, defaulting to `mean_se()
```



```r
ggsave("plots/immine.pdf", width = 5, height = 3)
```

```
## Warning: Removed 26715 rows containing non-finite values (stat_ydensity).

## Warning: Removed 26715 rows containing non-finite values (stat_summary).

## No summary function supplied, defaulting to `mean_se()
```

```r
wilcox.test(immuneGenesPre %>% filter(treatment == "Chlorothalonil") %>% pull(b))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  immuneGenesPre %>% filter(treatment == "Chlorothalonil") %>%     pull(b)
## V = 8730900, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(immuneGenesPre %>% filter(treatment == "Tetracycline") %>% pull(b))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  immuneGenesPre %>% filter(treatment == "Tetracycline") %>% pull(b)
## V = 16869000, p-value = 8.731e-11
## alternative hypothesis: true location is not equal to 0
```

```r
wilcox.test(immuneGenesPre %>% filter(treatment == "Glyphosate") %>% pull(b))
```

```
##
##  Wilcoxon signed rank test with continuity correction
```

```
##
## data:  immuneGenesPre %>% filter(treatment == "Glyphosate") %>% pull(b)
## V = 17609000, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

```
immuneGenesPre %>% filter(treatment == "Chlorothalonil") %>% pull(b) %>% na.omit() %>% mean()
```

```
## [1] -0.6678534
```

```
immuneGenesPre %>% filter(treatment == "Glyphosate") %>% pull(b) %>% na.omit() %>% mean()
```

```
## [1] -0.3049754
```

```
immuneGenesPre %>% filter(treatment == "Tetracycline") %>% pull(b) %>% na.omit() %>% mean()
```

```
## [1] -0.1234829
```

So, the bees in chemical-exposed treatments have lower stress gene levels than control bees.

```
waldStress <- function(trt = "Tetracycline") {
  # b > 0 genes were higher before stress
  dat <-  metadata %>% dplyr::filter(grepl(trt, treatment)) %>%  dplyr::select(sample, time_stress) %>%
  so <- sleuth_prep(dat, extra_bootstrap_summary = T, target_mapping = t2g, aggregation_column = 'gene_
  so <- sleuth_fit(so, ~ time_stress , 'full')
  so <- sleuth_wt(so, 'time_stressbefore_stress')
  results <- sleuth_results(so, test = "time_stressbefore_stress", pval_aggregate = F) %>% mutate(gene_
  return(left_join(results, genesAnnotation, by = c("gene_id" = "target_id"))  %>% dplyr::select(gene_i
}
```

```
dgeStress <- list()
for (trt in c("Tetracycline", "Glyphosate", "Chlorothalonil"))
    dgeStress[[trt]] <- waldStress(trt)
```

```
## Warning in check_num_cores(num_cores): It appears that you are running Sleuth from within Rstudio.
## Because of concerns with forking processes from a GUI, 'num_cores' is being set to 1.
## If you wish to take advantage of multiple cores, please consider running sleuth from the command line

## reading in kallisto results

## dropping unused factor levels

## .........
## normalizing est_counts
## 11612 targets passed the filter
## normalizing tpm
## merging in metadata
## summarizing bootstraps
## .........
## fitting measurement error models
## shrinkage estimation
## 4 NA values were found during variance shrinkage estimation due to mean observation values outside of
## The LOESS fit will be repeated using exact computation of the fitted surface to extrapolate the missi
## These are the target ids with NA values: XM_006566935.3, XR_003304694.1, NM_001011607.2, XM_01691795
## computing variance of betas

## Warning in check_num_cores(num_cores): It appears that you are running Sleuth from within Rstudio.
## Because of concerns with forking processes from a GUI, 'num_cores' is being set to 1.
## If you wish to take advantage of multiple cores, please consider running sleuth from the command line

## reading in kallisto results
## dropping unused factor levels
## .......
## normalizing est_counts
## 11460 targets passed the filter
## normalizing tpm
## merging in metadata
## summarizing bootstraps
## .......
## fitting measurement error models
## shrinkage estimation
## 2 NA values were found during variance shrinkage estimation due to mean observation values outside of
## The LOESS fit will be repeated using exact computation of the fitted surface to extrapolate the missi
## These are the target ids with NA values: XM_016912198.2, XR_120124.4
## computing variance of betas
```

```
## Warning in check_num_cores(num_cores): It appears that you are running Sleuth from within Rstudio.
## Because of concerns with forking processes from a GUI, 'num_cores' is being set to 1.
## If you wish to take advantage of multiple cores, please consider running sleuth from the command line
```

```
## reading in kallisto results
## dropping unused factor levels
## ........
## normalizing est_counts
## 10645 targets passed the filter
## normalizing tpm
## merging in metadata
## summarizing bootstraps
## ........
## fitting measurement error models
## shrinkage estimation
## 5 NA values were found during variance shrinkage estimation due to mean observation values outside of
## The LOESS fit will be repeated using exact computation of the fitted surface to extrapolate the missi
## These are the target ids with NA values: XM_026439362.1, XR_001702501.2, XR_412580.3, NM_001011607.2
## computing variance of betas
```

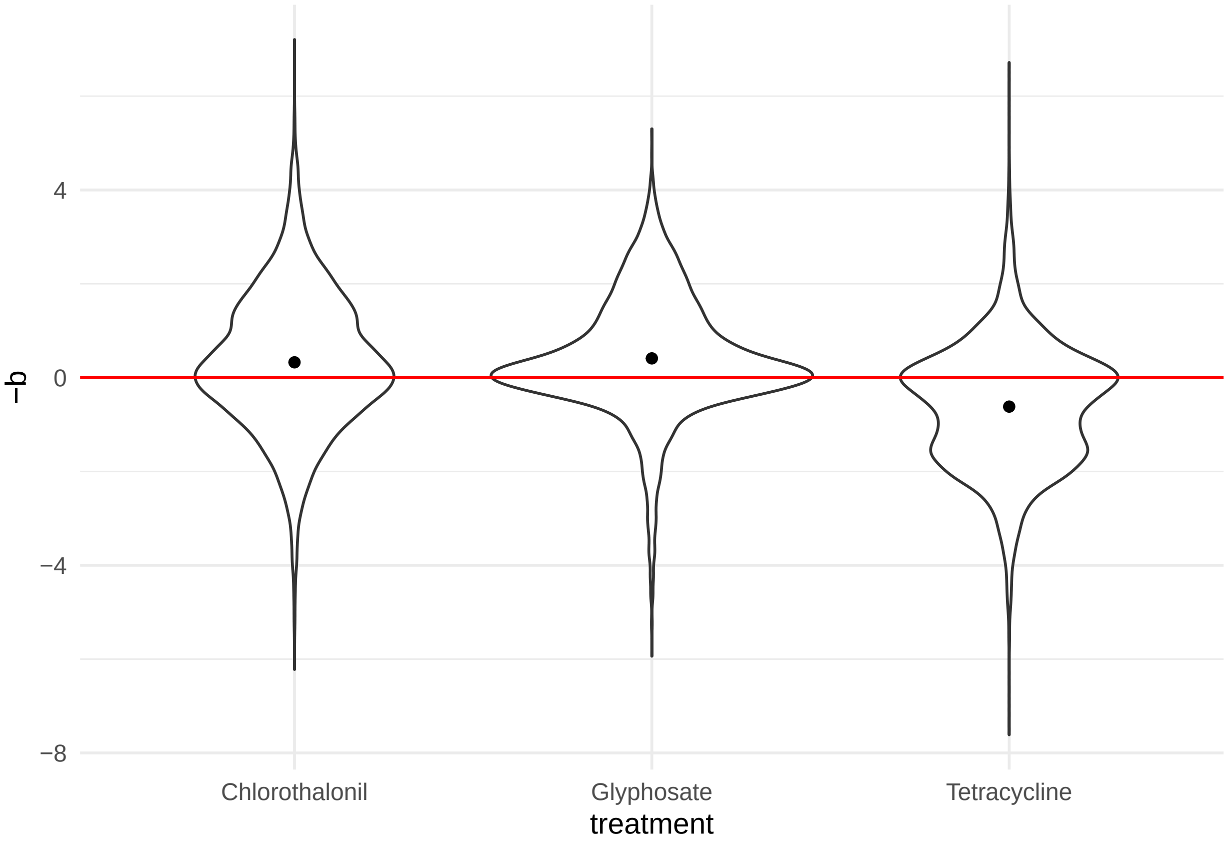
```r
(immuneGenesStress <- rbind(
  stressGenes %>% left_join(dgeStress[["Chlorothalonil"]], by =  "gene_id" ) %>% mutate(treatment = "Chl
  stressGenes %>% left_join(dgeStress[["Tetracycline"]], by =  "gene_id" ) %>% mutate(treatment = "Tetra
  stressGenes %>% left_join(dgeStress[["Glyphosate"]], by =  "gene_id" ) %>% mutate(treatment = "Glypho
) )%>% ggplot(aes(treatment, -b))  + geom_violin() + theme_minimal() + geom_hline(yintercept= 0, color =
```

```
## Warning: Ignoring unknown parameters: fun
```

```
## Warning: Removed 27799 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 27799 rows containing non-finite values (stat_summary).
```

```
## No summary function supplied, defaulting to `mean_se()
```



```
wilcox.test(immuneGenesStress %>% filter(treatment == "Chlorothalonil") %>% pull(b))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  immuneGenesStress %>% filter(treatment == "Chlorothalonil") %>%     pull(b)
## V = 12044000, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(immuneGenesStress %>% filter(treatment == "Tetracycline") %>% pull(b))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  immuneGenesStress %>% filter(treatment == "Tetracycline") %>%     pull(b)
## V = 28114000, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(immuneGenesStress %>% filter(treatment == "Glyphosate") %>% pull(b))
```

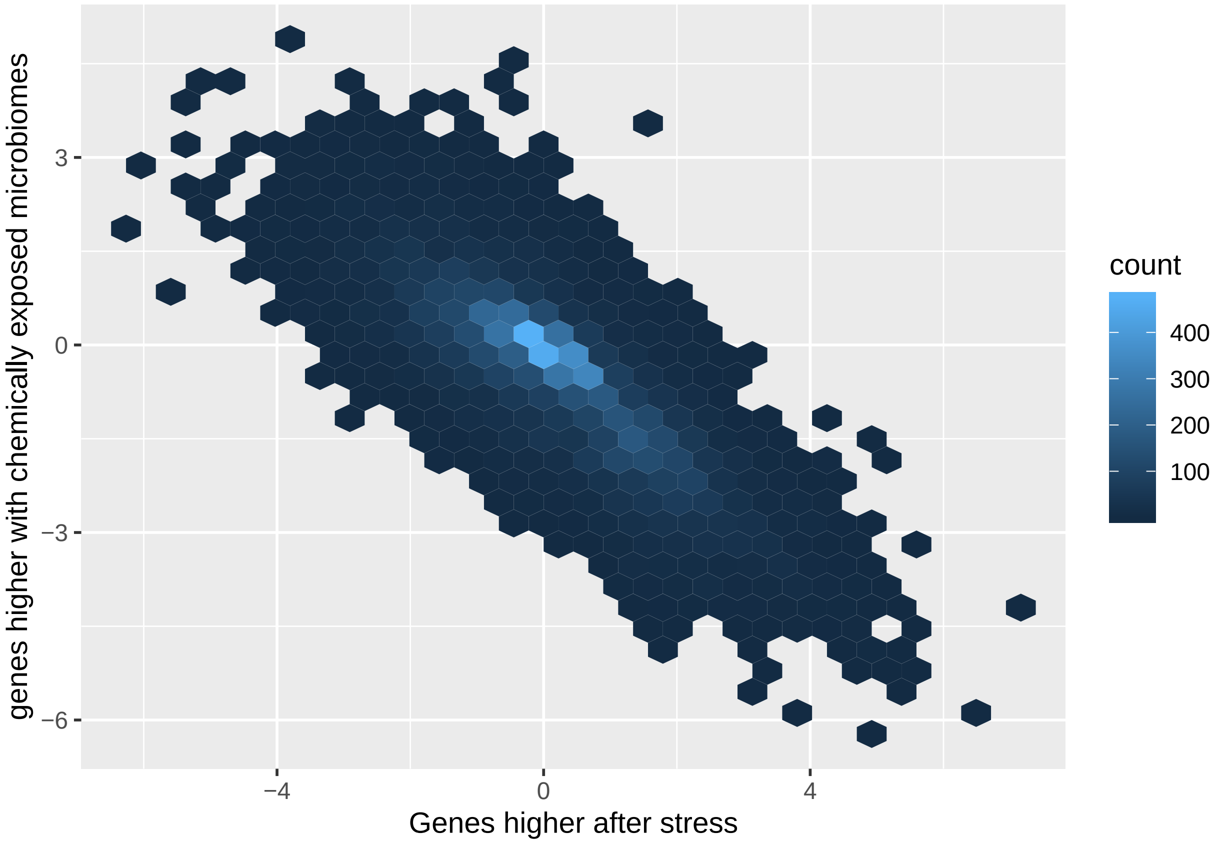
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  immuneGenesStress %>% filter(treatment == "Glyphosate") %>% pull(b)
## V = 10960000, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

Examining potential hormetic effects

Hormesis often takes place when previous exposure stimulates genes involved in dealing with the response. If hormesis is responsible for higher survival under chlorothalonil, we would expect that genes expressed post-exposure would be upregulated in the same direction as genes in bees receiving exposed microbiomes pre-exposure.

```
chlorStressGenes <- left_join(dgeStress[["Chlorothalonil"]], dgePre[["Chlorothalonil"]], by =  "target_
chlorStressGenes %>% ggplot(aes(-1*b.x, b.y)) + geom_hex() + xlab("Genes higher after stress") + ylab("
```

```
## Warning: Removed 18091 rows containing non-finite values (stat_binhex).
```



```
with(chlorStressGenes, cor.test(-1*b.x, b.y, method= "s"))
```

```
## Warning in cor.test.default(-1 * b.x, b.y, method = "s"): Cannot compute exact
## p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  -1 * b.x and b.y
## S = 2.8383e+11, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.8116252
```

Sanity checks using tpm

```
read_kallisto <- function(filename) {
  sampleName <- sub("data/kallisto/tsv/(.*).tsv.gz","\\1", filename)
  return(read_tsv(filename) %>%
           select(!!sampleName := tpm))
}
df <- list.files(path = "data/kallisto/tsv", full.names = TRUE) %>%
  lapply(read_kallisto) %>%
  bind_cols()
df$target_id <- list.files(path = "data/kallisto/tsv", full.names = TRUE)[1] %>% read_tsv() %>% select(t

tpm <- gather(df,key="sample", value = "tpm", -25) %>% left_join(metadata)

# tpm of chlor genes before and after stress

chlor <- tpm %>% filter(treatment == "Chlorothalonil") %>% select(target_id, time_stress, tpm) %>% grou

with(left_join(dgeStress[["Chlorothalonil"]], chlor), cor.test(diff, b, method = "s"))

# There is good correlation between tpm and and b estimates. b>0 genes are higher before stress

chlorPre <- tpm %>% dplyr::filter((grepl(trt, "Chlorothalonil") | history == "control") & time_stress ==

with(left_join(dgePre[["Chlorothalonil"]], chlorPre), cor.test(diff, b, method = "s"))

with(left_join(chlor, chlorPre, by = c("target_id")), cor.test(diff.x, diff.y, method = "s"))

#genes that are higher in coevolved were alse higher before stress
```