R visualization class project

Neha Gupta

5/19/2019

I am a postdoctoral fellow at NICHD, NIH. In our labs, we study regulation of gene expression.

My focus: Translation initiation and mRNA degradation

Interested in understanding the roles of Pat1 and Dhh1 in translational control in yeast.

Performed RNA-seq and ribosome footprint profiling of WT, pat1 deletion, and pat1dhh1 double deletion strains. (dhh1 deletion experiemtns were recently published from our lab). Biological triplicates.

Bioinformatics analysis - removing rRNA sequences, aligning sequences to genomes, counting read alignments, etc.

**Step 1:** Checking correlation between biological replicates (Raw read counts). Combined all counts from all samples in 1 file.

library(tidyverse)  
allcount <- read\_csv("raw\_count\_all.csv") #reading combined raw read count file  
head(allcount)

## # A tibble: 6 x 31  
## Genes mRNA\_dhh\_1\_S9\_r… mRNA\_dhh\_2\_S10\_… mRNA\_pat\_1\_rna\_… mRNA\_pat\_2\_rna\_…  
## <chr> <int> <int> <int> <int>  
## 1 YAL0… 131 40 351 266  
## 2 YAL0… 4 0 86 63  
## 3 YAL0… 277 98 1224 981  
## 4 YAL0… 326 103 1305 1145  
## 5 YAL0… 1069 389 4609 4532  
## 6 YAL0… 319 152 1577 1382  
## # ... with 26 more variables: mRNA\_pat\_3\_rna\_counts.txt <int>,  
## # mRNA\_pd\_1\_rna\_counts.txt <int>, mRNA\_pd\_2\_rna\_counts.txt <int>,  
## # mRNA\_pd\_3\_rna\_counts.txt <int>, mRNA\_scd6\_1\_S8\_rna\_counts.txt <int>,  
## # mRNA\_scd6\_2\_S12\_rna\_counts.txt <int>, mRNA\_WT\_1\_rna\_counts.txt <int>,  
## # mRNA\_WT\_2\_rna\_counts.txt <int>, mRNA\_WT\_3\_rna\_counts.txt <int>,  
## # mRNA\_WTdhh\_1\_S7\_rna\_counts.txt <int>,  
## # mRNA\_WTdhh\_2\_S11\_rna\_counts.txt <int>,  
## # ribo\_dhh\_1\_S3\_fp\_counts.txt <int>, ribo\_dhh\_2\_S4\_fp\_counts.txt <int>,  
## # ribo\_pat\_1\_fp\_counts.txt <int>, ribo\_pat\_2\_fp\_counts.txt <int>,  
## # ribo\_pat\_3\_fp\_counts.txt <int>, ribo\_pd\_1\_fp\_counts.txt <int>,  
## # ribo\_pd\_2\_fp\_counts.txt <int>, ribo\_pd\_3\_fp\_counts.txt <int>,  
## # ribo\_scd6\_1\_S2\_fp\_counts.txt <int>,  
## # ribo\_scd6\_2\_S6\_fp\_counts.txt <int>, ribo\_WT\_1\_fp\_counts.txt <int>,  
## # ribo\_WT\_2\_fp\_counts.txt <int>, ribo\_WT\_3\_fp\_counts.txt <int>,  
## # ribo\_WTdhh\_1\_S1\_fp\_counts.txt <int>,  
## # ribo\_WTdhh\_2\_S5\_fp\_counts.txt <int>

**Changing column names** The column names are too long with extra text. mRNA = RNA-seq samples ribo = ribo-seq samples pd = pat1dhh1

raw\_count <- allcount %>%  
 set\_names(str\_replace\_all(names(.), '\_counts.txt', '')) %>%  
 set\_names(str\_replace(names(.), '\_rna', '')) %>%  
 set\_names(str\_replace(names(.), '\_fp','')) %>%  
 set\_names(str\_replace(names(.), "\_1", "\_repA")) %>%  
 set\_names(str\_replace(names(.), "\_2", "\_repB")) %>%  
 set\_names(str\_replace(names(.), "\_3", "\_repC"))  
  
  
colnames(raw\_count)

## [1] "Genes" "mRNA\_dhh\_repA\_S9" "mRNA\_dhh\_repB\_S10"   
## [4] "mRNA\_pat\_repA" "mRNA\_pat\_repB" "mRNA\_pat\_repC"   
## [7] "mRNA\_pd\_repA" "mRNA\_pd\_repB" "mRNA\_pd\_repC"   
## [10] "mRNA\_scd6\_repA\_S8" "mRNA\_scd6\_repB\_S12" "mRNA\_WT\_repA"   
## [13] "mRNA\_WT\_repB" "mRNA\_WT\_repC" "mRNA\_WTdhh\_repA\_S7"   
## [16] "mRNA\_WTdhh\_repB\_S11" "ribo\_dhh\_repA\_S3" "ribo\_dhh\_repB\_S4"   
## [19] "ribo\_pat\_repA" "ribo\_pat\_repB" "ribo\_pat\_repC"   
## [22] "ribo\_pd\_repA" "ribo\_pd\_repB" "ribo\_pd\_repC"   
## [25] "ribo\_scd6\_repA\_S2" "ribo\_scd6\_repB\_S6" "ribo\_WT\_repA"   
## [28] "ribo\_WT\_repB" "ribo\_WT\_repC" "ribo\_WTdhh\_repA\_S1"   
## [31] "ribo\_WTdhh\_repB\_S5"

**Replicate analysis**

First, I did correlation analysis between replicates. Usually, in our lab, we do 2 biological replicates, but I am working with three biological replicates. I am still doing pair-wise correlation, but on a look out for a better statistical test to analyze three replicates together.

library(ggpubr)

This is replicate 1 and 2 analysis of ribo-seq samples of pat1 deletion.

ggplot(raw\_count, aes(x = ribo\_pat\_repA , y = ribo\_pat\_repB)) +  
 geom\_point(color = "darkgrey") + stat\_cor(aes(label = paste(..rr.label..))) + #no p-value label  
 scale\_y\_log10(labels = scales::number\_format(big.mark = '')) + #bigmark get rids of space between numbers  
 scale\_x\_log10(labels = scales::number\_format(big.mark = '')) +  
 theme\_bw() +  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14, face = "bold.italic"),  
 axis.text = element\_text(size = 12)) +  
 coord\_cartesian(clip = 'off') +  
 labs(x = "Replicate 1", y = "Replicate 2",  
 title = expression("pat1"\*Delta),  
 subtitle = "Ribosome density")

## Warning: Transformation introduced infinite values in continuous y-axis

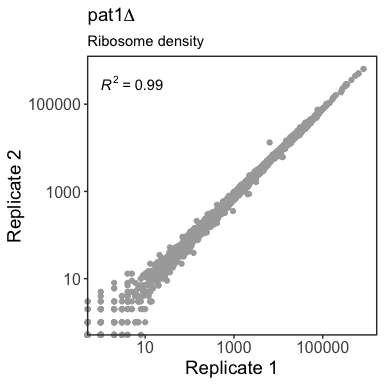
## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 185 rows containing non-finite values (stat\_cor).

## Warning: package 'bindrcpp' was built under R version 3.4.4



There are about 185 missing values from stat\_Cor. log of 0 will give infinite value, thats why? But stat\_cor is before log transformation of axes, shouldn’t it take precedence?

anyNA(raw\_count)

## [1] FALSE

dim(raw\_count)

## [1] 5647 31

raw\_count1 <- raw\_count %>%   
 select(ribo\_pat\_repA, ribo\_pat\_repB) %>%   
 filter(ribo\_pat\_repA > 0, ribo\_pat\_repB > 0)  
dim(raw\_count1)

## [1] 5462 2

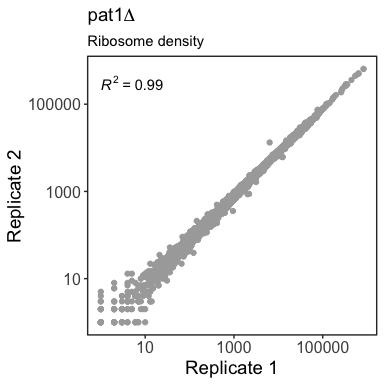
OK, so those missing values are due to counts = 0. Testing correlation

cor.test(raw\_count$ribo\_pat\_repA, raw\_count$ribo\_pat\_repB, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: raw\_count$ribo\_pat\_repA and raw\_count$ribo\_pat\_repB  
## t = 1806.6, df = 5645, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9990901 0.9991802  
## sample estimates:  
## cor   
## 0.9991364

Now, make a graph with 0 counts removed. The warning is gone.

ggplot(raw\_count1, aes(x = ribo\_pat\_repA , y = ribo\_pat\_repB)) +  
 geom\_point(color = "darkgrey") + stat\_cor(aes(label = paste(..rr.label..))) +  
 scale\_y\_log10(labels = scales::number\_format(big.mark = '')) +  
 scale\_x\_log10(labels = scales::number\_format(big.mark = '')) +  
 theme\_bw() +  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14, face = "italic"),  
 axis.text = element\_text(size = 12)) +  
 coord\_cartesian(clip = 'off') +  
 labs(x = "Replicate 1", y = "Replicate 2",  
 title = expression("pat1"\*Delta),  
 subtitle = "Ribosome density")



#ggsave("pat1\_ribo\_repAB.pdf", height = 4, width = 4)

Similarly, do for the RNA-seq samples

ggplot(raw\_count, aes(x = mRNA\_pat\_repA , y = mRNA\_pat\_repB)) +  
 geom\_point(color = "darkgrey") + stat\_cor(aes(label = paste(..rr.label..))) +  
 scale\_y\_log10(labels = scales::number\_format(big.mark = '')) +  
 scale\_x\_log10(labels = scales::number\_format(big.mark = '')) +  
 theme\_bw() +  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14),  
 axis.text = element\_text(size = 12)) +  
 coord\_cartesian(clip = 'off') +  
 labs(x = "Replicate 1", y = "Replicate 2",  
 title = expression("pat1"\*Delta),  
 subtitle = "mRNA density")

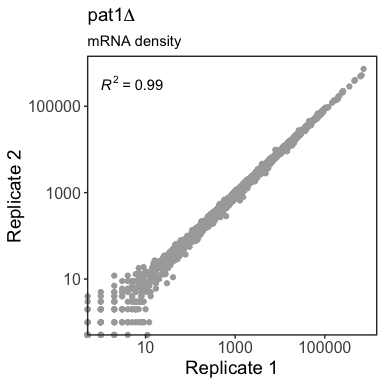
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 40 rows containing non-finite values (stat\_cor).



Did the same analysis for all replicates of WT and mutants, found a correlation of 0.98-0.99.

Proceed with DeSeq2 analysis (Bioconductors), and started looking for differential expression. We have a script in the lab for Deseq2 analysis. Changed minimum read counts for RNA-seq to 128 and ribo-seq to 32. The script generates 6 files and the information needs to be combined for subsequent analysis.

First, I merged 6 or so files together for each pair.

#pat\_norm\_count <- read.delim("./pat1\_WT/pat1\_vs\_WT\_Normalized\_counts.txt", header = TRUE, sep = "\t")  
#pat\_mRNA <- read.delim("./pat1\_WT/pat1\_mRNA\_change\_vs\_WT.txt", header = TRUE, sep = "\t")  
#pat\_ribo <- read.delim("./pat1\_WT/pat1\_Ribo\_change\_vs\_WT.txt", header = TRUE, sep = "\t")  
  
#pat\_patTE <- read.delim("./pat1\_WT/pat1\_TE\_with\_WT.txt", header = TRUE, sep = "\t")  
  
#pat\_WTTE <- read.delim("./pat1\_WT/WT\_TE\_with\_pat1.txt", header = TRUE, sep = "\t")  
  
#pat\_TEchange <- read.delim("./pat1\_WT/TE\_change\_pat1\_vs\_WT.txt", header = TRUE, sep = "\t")  
  
#First, looking at the normalized read count file.  
  
  
#head(pat\_norm\_count)  
#colnames(pat\_norm\_count)  
#colSums(pat\_norm\_count)  
  
  
#Changing column names:  
  
  
#pat\_norm\_count\_1 <- pat\_norm\_count %>%  
 #set\_names(str\_replace\_all(names(.), '\_counts.txt', '')) %>%  
 #set\_names(str\_replace(names(.), '\_rna', '')) %>%  
 #set\_names(str\_replace(names(.), '\_fp','')) %>%  
 #set\_names(str\_replace(names(.), "\_1", "\_repA")) %>%  
 #set\_names(str\_replace(names(.), "\_2", "\_repB")) %>%  
 #set\_names(str\_replace(names(.), "\_3", "\_repC"))  
#colnames(pat\_norm\_count\_1)  
  
#The number of rows are equal, and same gene are represented in the same order so can using cbind.  
  
  
#str(pat\_mRNA)  
#str(pat\_norm\_count\_1)  
#str looks good  
  
#anyNA(pat\_mRNA)  
#anyNA(pat\_norm\_count\_1)  
# No missing values  
  
#new\_pat1 <- cbind(pat\_mRNA, pat\_norm\_count\_1)  
#head(new\_pat1)

Similarly, I combined all 6 files, changed the column names, removed duplicate or unnecessary column to create a table for all files together. Saved it as a txt and csv file.

pat\_combined <- read\_csv("pat1\_combined\_data.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## TEchange\_gene\_ID = col\_character(),  
## gene\_name = col\_character(),  
## WTTE\_gene\_ID = col\_character(),  
## patTE\_gene\_ID = col\_character(),  
## ribo\_gene\_ID = col\_character(),  
## mRNA\_gene\_ID = col\_character()  
## )

## See spec(...) for full column specifications.

head(pat\_combined)

## # A tibble: 6 x 33  
## TEchange\_gene\_ID gene\_name TEchange\_log2FC `TEchange\_p-val` TEchange\_FDR  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 YAL067C SEO1 0.197 0.366 0.611   
## 2 YAL063C FLO9 0.422 0.0158 0.0840   
## 3 YAL054C ACS1 1.06 0.0000000181 0.000000820  
## 4 YAL049C AIM2 0.196 0.0226 0.107   
## 5 YAL047C SPC72 -0.0134 0.897 0.956   
## 6 YAL046C AIM1 -0.0473 0.573 0.768   
## # ... with 28 more variables: WTTE\_log2FC <dbl>, `WTTE\_p-val` <dbl>,  
## # WTTE\_FDR <dbl>, WTTE\_gene\_ID <chr>, patTE\_log2FC <dbl>,  
## # `patTE\_p-val` <dbl>, patTE\_FDR <dbl>, patTE\_gene\_ID <chr>,  
## # ribo\_log2FC <dbl>, `ribo\_p-val` <dbl>, ribo\_FDR <dbl>,  
## # ribo\_gene\_ID <chr>, mRNA\_log2FC <dbl>, `mRNA\_p-val` <dbl>,  
## # mRNA\_FDR <dbl>, mRNA\_gene\_ID <chr>, mRNA\_WT\_repA <dbl>,  
## # mRNA\_WT\_repB <dbl>, mRNA\_WT\_repC <dbl>, mRNA\_pat\_repA <dbl>,  
## # mRNA\_pat\_repB <dbl>, mRNA\_pat\_repC <dbl>, ribo\_WT\_repA <dbl>,  
## # ribo\_WT\_repB <dbl>, ribo\_WT\_repC <dbl>, ribo\_pat\_repA <dbl>,  
## # ribo\_pat\_repB <dbl>, ribo\_pat\_repC <dbl>

Now, I need to calculate the means for the mRNA and ribo-reads.

pat1\_combined <- pat\_combined %>%   
 select(-'TEchange\_p-val', -'WTTE\_p-val', -WTTE\_gene\_ID, -'patTE\_p-val', -patTE\_gene\_ID, -'ribo\_p-val', -ribo\_gene\_ID, -'mRNA\_p-val', -mRNA\_gene\_ID) %>%   
 mutate(mRNA\_WT\_mean = rowMeans(select(.,c("mRNA\_WT\_repA", "mRNA\_WT\_repB", "mRNA\_WT\_repC")))) %>%   
 mutate(mRNA\_pat1\_mean = rowMeans(select(.,c("mRNA\_pat\_repA", "mRNA\_pat\_repB", "mRNA\_pat\_repC")))) %>%   
 mutate(ribo\_WT\_mean = rowMeans(select(.,c("ribo\_WT\_repA", "ribo\_WT\_repB", "ribo\_WT\_repC")))) %>%  
 mutate(ribo\_pat1\_mean = rowMeans(select(.,c("ribo\_pat\_repA", "ribo\_pat\_repB", "ribo\_pat\_repC"))))  
  
head(pat1\_combined)

## # A tibble: 6 x 28  
## TEchange\_gene\_ID gene\_name TEchange\_log2FC TEchange\_FDR WTTE\_log2FC  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 YAL067C SEO1 0.197 0.611 -1.42   
## 2 YAL063C FLO9 0.422 0.0840 -2.64   
## 3 YAL054C ACS1 1.06 0.000000820 -2.14   
## 4 YAL049C AIM2 0.196 0.107 0.844  
## 5 YAL047C SPC72 -0.0134 0.956 -0.675  
## 6 YAL046C AIM1 -0.0473 0.768 0.865  
## # ... with 23 more variables: WTTE\_FDR <dbl>, patTE\_log2FC <dbl>,  
## # patTE\_FDR <dbl>, ribo\_log2FC <dbl>, ribo\_FDR <dbl>, mRNA\_log2FC <dbl>,  
## # mRNA\_FDR <dbl>, mRNA\_WT\_repA <dbl>, mRNA\_WT\_repB <dbl>,  
## # mRNA\_WT\_repC <dbl>, mRNA\_pat\_repA <dbl>, mRNA\_pat\_repB <dbl>,  
## # mRNA\_pat\_repC <dbl>, ribo\_WT\_repA <dbl>, ribo\_WT\_repB <dbl>,  
## # ribo\_WT\_repC <dbl>, ribo\_pat\_repA <dbl>, ribo\_pat\_repB <dbl>,  
## # ribo\_pat\_repC <dbl>, mRNA\_WT\_mean <dbl>, mRNA\_pat1\_mean <dbl>,  
## # ribo\_WT\_mean <dbl>, ribo\_pat1\_mean <dbl>

**mRNA changes in pat1 deletion vs WT**

Now, lets look at mRNA changes. First, collect the required columns otherwise the plot gets too big as seen with TE changes.

mRNA\_pat <- pat1\_combined %>%   
 select(mRNA\_log2FC, mRNA\_FDR, mRNA\_WT\_mean, mRNA\_pat1\_mean)  
head(mRNA\_pat)

## # A tibble: 6 x 4  
## mRNA\_log2FC mRNA\_FDR mRNA\_WT\_mean mRNA\_pat1\_mean  
## <dbl> <dbl> <dbl> <dbl>  
## 1 0.720 8.93e- 8 154. 252.  
## 2 1.45 1.02e- 59 327. 891.  
## 3 2.32 1.81e-103 213. 1060.  
## 4 1.28 3.38e- 93 1524. 3695.  
## 5 1.15 5.53e- 63 556. 1236.  
## 6 1.64 5.03e-150 454. 1414.

Next, I filtered the data with FDR < 0.05 and fold-change of 1.5 or 2 or greater. I used this information on the graph.

mpat\_filtered\_up2 <- mRNA\_pat %>%   
 filter(mRNA\_log2FC > 1) %>%   
 filter(mRNA\_FDR < 0.05)  
dim(mpat\_filtered\_up2)

## [1] 364 4

mpat\_filtered\_up1 <- mRNA\_pat %>%   
 filter(mRNA\_log2FC > 0.58) %>%   
 filter(mRNA\_FDR < 0.05)  
dim(mpat\_filtered\_up1)

## [1] 756 4

mpat\_filtered\_down2 <- mRNA\_pat %>%   
 filter(mRNA\_log2FC < -1) %>%   
 filter(mRNA\_FDR < 0.05)  
dim(mpat\_filtered\_down2)

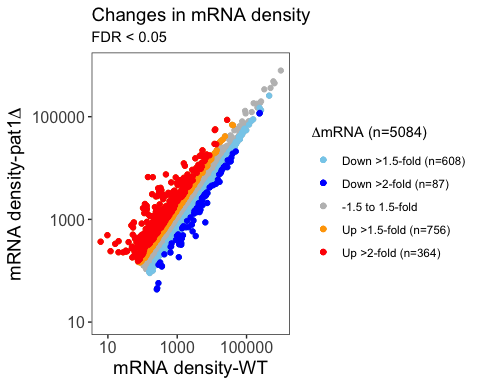
## [1] 87 4

mpat\_filtered\_down1 <- mRNA\_pat %>%   
 filter(mRNA\_log2FC < -0.58) %>%   
 filter(mRNA\_FDR < 0.05)  
dim(mpat\_filtered\_down1)

## [1] 608 4

Visualization of the data.

cols1 <- c("Nochange" = "grey", "Up2" = "red", "Down2" = "blue", "Up1" = "orange", "Down1" = "skyblue")  
mRNA\_p <- ggplot(data = mRNA\_pat, aes(x = mRNA\_WT\_mean, y = mRNA\_pat1\_mean)) +  
 geom\_point(aes(color = "Nochange")) +  
 scale\_y\_log10(limits = c(10,1000000), labels = scales::number\_format(big.mark = ''), breaks = c(10,1000,100000)) +  
 scale\_x\_log10(labels = scales::number\_format(big.mark = ''), breaks = c(10,1000,100000)) +   
 theme\_bw() +  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14),  
 axis.text = element\_text(size = 12)) +  
 geom\_point(data = mRNA\_pat %>% filter(mRNA\_log2FC > 0.58 & mRNA\_FDR <= 0.05), aes(color = "Up1")) +  
 geom\_point(data = mRNA\_pat %>% filter(mRNA\_log2FC < -0.58 & mRNA\_FDR <= 0.05), aes(color = "Down1")) +  
 geom\_point(data = mRNA\_pat %>% filter(mRNA\_log2FC > 1 & mRNA\_FDR <= 0.05), aes(color = "Up2")) +  
 geom\_point(data = mRNA\_pat %>% filter(mRNA\_log2FC < -1 & mRNA\_FDR <= 0.05), aes(color = "Down2")) +  
scale\_color\_manual(name = expression(Delta\*"mRNA (n=5084)"), values = cols1,  
 labels = c("Down >1.5-fold (n=608)", "Down >2-fold (n=87)", "-1.5 to 1.5-fold", "Up >1.5-fold (n=756)", "Up >2-fold (n=364) "))  
  
   
mRNA\_p + labs(x = "mRNA density-WT", y = expression("mRNA density-pat1"\*Delta),  
 title = "Changes in mRNA density",  
 subtitle = "FDR < 0.05")



Here, I am looking at TE changes. TE (translational efficiency) is ribo counts/mRNA count, and change is TE = TE mutant/TE WT.

Again calculating the number of genes upregulated or downregulated in pat1 mutant strain.

pat\_filtered\_up2 <- pat1\_combined %>%   
 filter(TEchange\_log2FC > 1) %>%   
 filter(TEchange\_FDR < 0.05)  
dim(pat\_filtered\_up2)

## [1] 60 28

pat\_filtered\_up1 <- pat1\_combined %>%   
 filter(TEchange\_log2FC > 0.58) %>%   
 filter(TEchange\_FDR < 0.05)  
dim(pat\_filtered\_up1)

## [1] 123 28

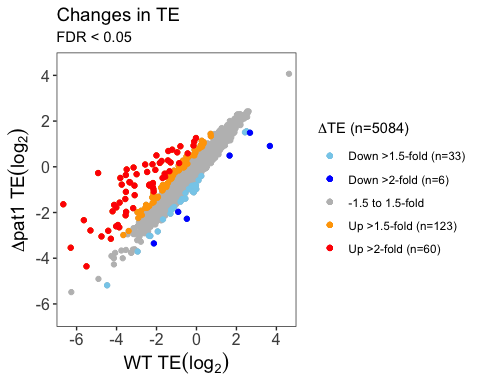
pat\_filtered\_down2 <- pat1\_combined %>%   
 filter(TEchange\_log2FC < -1) %>%   
 filter(TEchange\_FDR < 0.05)  
dim(pat\_filtered\_down2)

## [1] 6 28

pat\_filtered\_down1 <- pat1\_combined %>%   
 filter(TEchange\_log2FC < -0.58) %>%   
 filter(TEchange\_FDR < 0.05)  
dim(pat\_filtered\_down1)

## [1] 33 28

cols1 <- c("Nochange" = "grey", "Up2" = "red", "Down2" = "blue", "Up1" = "orange", "Down1" = "skyblue")  
p <- ggplot(data = pat1\_combined, aes(x = WTTE\_log2FC, y = patTE\_log2FC)) +  
 geom\_point (aes(color = "Nochange")) +  
 scale\_x\_continuous(limits = c(-7,5), breaks = c(-6,-4,-2,0,2,4,6), expand = c(0,0)) +  
 scale\_y\_continuous(limits = c(-7,5), breaks = c(-6,-4,-2,0,2,4,6), expand = c(0,0)) +  
 theme\_bw() +  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14),  
 axis.text = element\_text(size = 12)) +  
 geom\_point(data = pat1\_combined %>% filter(TEchange\_log2FC > 0.58 & TEchange\_FDR <= 0.05), aes(color = "Up1")) +  
 geom\_point(data = pat1\_combined %>% filter(TEchange\_log2FC < -0.58 & TEchange\_FDR <= 0.05), aes(color = "Down1")) +  
 geom\_point(data = pat1\_combined %>% filter(TEchange\_log2FC > 1 & TEchange\_FDR <= 0.05), aes(color = "Up2")) +  
 geom\_point(data = pat1\_combined %>% filter(TEchange\_log2FC < -1 & TEchange\_FDR <= 0.05), aes(color = "Down2")) +  
scale\_color\_manual(name = expression(Delta\*"TE (n=5084)"), values = cols1,  
 labels = c("Down >1.5-fold (n=33)", "Down >2-fold (n=6)", "-1.5 to 1.5-fold", "Up >1.5-fold (n=123)", "Up >2-fold (n=60) "))  
  
   
p + labs(x = expression("WT TE"(log[2])), y = expression(Delta\*"pat1 TE"(log[2])),  
 title = "Changes in TE",  
 subtitle = "FDR < 0.05")

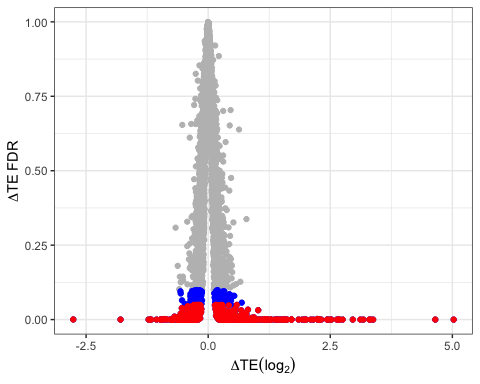


For summary statistics, sometimes only the FDR cutoff is used, since fold-change can be arbitarary. Since, there aren’t “many” genes that undergo changes according to fold-change cutoff, I wanted to take a look at relationship between FC and FDR. This is like an upside version of volcano plot.

FDR\_TE <- pat1\_combined %>%  
 select(TEchange\_log2FC, TEchange\_FDR)   
FDR\_TE2 <- FDR\_TE %>%  
 filter(TEchange\_FDR <0.05) %>%   
 filter(TEchange\_log2FC > 0)  
 #filter(TEchange\_log2FC < 0)   
  
dim(FDR\_TE2)

## [1] 377 2

#0.1= 473 up 0.05 = 377  
#0.1 = 559 down 0.05 = 401  
 ggplot(FDR\_TE, aes(TEchange\_log2FC, TEchange\_FDR)) +  
 geom\_point(color="grey") +  
 theme\_bw() +  
 geom\_point(data = FDR\_TE %>% filter (TEchange\_FDR < 0.1), color = "blue") +  
 geom\_point(data = FDR\_TE %>% filter (TEchange\_FDR < 0.05), color = "red") +  
 labs(y = expression(Delta\*"TE FDR"), x = expression(Delta\*"TE" (log[2])))



I am looking at the relationship of genes that are up- or down-regulated in Pat1 deletion strain with dhh1 and pat1dhh1 deletion strains. Are they also affected in these strains?

pat <- read\_csv("pat1\_combined\_data.csv")  
dhh <- read\_csv("dhh1\_combined\_data.csv")  
pd <- read\_csv("pat1dhh1\_combined\_data.csv")

Since all three files have the same/similar column names, I am changing to distinguish them. Here, only analysis with mRNA samples is shown. Similar analysis is done with TE change data.

#head(pat)  
pat1 <- pat %>%   
 select(TEchange\_gene\_ID, mRNA\_log2FC, mRNA\_FDR)   
  
names(pat1)[1] <- "pat\_gene\_ID"  
names(pat1)[2] <- "pat\_mRNA\_FC"  
names(pat1)[3] <- "pat\_mRNA\_FDR"  
  
head(pat1)

## # A tibble: 6 x 3  
## pat\_gene\_ID pat\_mRNA\_FC pat\_mRNA\_FDR  
## <chr> <dbl> <dbl>  
## 1 YAL067C 0.720 8.93e- 8  
## 2 YAL063C 1.45 1.02e- 59  
## 3 YAL054C 2.32 1.81e-103  
## 4 YAL049C 1.28 3.38e- 93  
## 5 YAL047C 1.15 5.53e- 63  
## 6 YAL046C 1.64 5.03e-150

#head(dhh)  
dhh1 <- dhh %>%   
 select(TEchange\_gene\_ID, mRNA\_log2FC, mRNA\_FDR)   
  
names(dhh1)[1] <- "dhh\_gene\_ID"  
names(dhh1)[2] <- "dhh\_mRNA\_FC"  
names(dhh1)[3] <- "dhh\_mRNA\_FDR"  
  
head(dhh1)

## # A tibble: 6 x 3  
## dhh\_gene\_ID dhh\_mRNA\_FC dhh\_mRNA\_FDR  
## <chr> <dbl> <dbl>  
## 1 YAL067C -0.630 0.0250   
## 2 YAL063C 0.118 0.729   
## 3 YAL054C 1.35 0.0000959  
## 4 YAL049C 0.00480 0.980   
## 5 YAL047C -0.117 0.624   
## 6 YAL046C -0.186 0.381

#head(pd)  
pd1 <- pd %>%   
 select(TEchange\_gene\_ID, mRNA\_log2FC, mRNA\_FDR)   
  
names(pd1)[1] <- "pd\_gene\_ID"  
names(pd1)[2] <- "pd\_mRNA\_FC"  
names(pd1)[3] <- "pd\_mRNA\_FDR"  
  
head(pd1)

## # A tibble: 6 x 3  
## pd\_gene\_ID pd\_mRNA\_FC pd\_mRNA\_FDR  
## <chr> <dbl> <dbl>  
## 1 YAL063C 0.808 2.25e- 8  
## 2 YAL054C 2.94 5.44e-18  
## 3 YAL049C -0.0651 4.07e- 1  
## 4 YAL047C -0.00845 9.44e- 1  
## 5 YAL046C 0.302 2.99e- 3  
## 6 YAL044C 0.597 5.61e-17

#str(pd1)

Ideally, there shouldn’t be any NA or 0 in the dataset.

anyNA(pat1)

## [1] FALSE

anyNA(dhh1)

## [1] FALSE

anyNA(pd1)

## [1] FALSE

Since, in this dataset, I am looking at the pat1 regulated genes, I am combining genes that are present in pat1. The other dataset have overlapping but distinct set of genes.

comb1 <- left\_join(pat1, pd1, by=c("pat\_gene\_ID"="pd\_gene\_ID"))  
  
head(comb1)

## # A tibble: 6 x 5  
## pat\_gene\_ID pat\_mRNA\_FC pat\_mRNA\_FDR pd\_mRNA\_FC pd\_mRNA\_FDR  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 YAL067C 0.720 8.93e- 8 NA NA   
## 2 YAL063C 1.45 1.02e- 59 0.808 2.25e- 8  
## 3 YAL054C 2.32 1.81e-103 2.94 5.44e-18  
## 4 YAL049C 1.28 3.38e- 93 -0.0651 4.07e- 1  
## 5 YAL047C 1.15 5.53e- 63 -0.00845 9.44e- 1  
## 6 YAL046C 1.64 5.03e-150 0.302 2.99e- 3

combined <- left\_join(comb1, dhh1, by=c("pat\_gene\_ID"="dhh\_gene\_ID"))  
  
head(combined)

## # A tibble: 6 x 7  
## pat\_gene\_ID pat\_mRNA\_FC pat\_mRNA\_FDR pd\_mRNA\_FC pd\_mRNA\_FDR dhh\_mRNA\_FC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 YAL067C 0.720 8.93e- 8 NA NA -0.630   
## 2 YAL063C 1.45 1.02e- 59 0.808 2.25e- 8 0.118   
## 3 YAL054C 2.32 1.81e-103 2.94 5.44e-18 1.35   
## 4 YAL049C 1.28 3.38e- 93 -0.0651 4.07e- 1 0.00480  
## 5 YAL047C 1.15 5.53e- 63 -0.00845 9.44e- 1 -0.117   
## 6 YAL046C 1.64 5.03e-150 0.302 2.99e- 3 -0.186   
## # ... with 1 more variable: dhh\_mRNA\_FDR <dbl>

dim(combined)

## [1] 5084 7

mRNA <- combined %>%   
 filter(pat\_mRNA\_FDR < 0.05 & pat\_mRNA\_FC > 0)  
dim(mRNA)

## [1] 1771 7

mRNA\_2 <- mRNA %>%  
 select(pat\_gene\_ID, pat\_mRNA\_FC, dhh\_mRNA\_FC, pd\_mRNA\_FC) %>%  
 set\_names(str\_replace(names(.), "\_mRNA\_FC", "1")) %>%   
 set\_names(str\_replace(names(.), "pd1", "pat1dhh1")) %>%   
   
  
 gather(sample, change, -pat\_gene\_ID)  
   
head(mRNA\_2)

## # A tibble: 6 x 3  
## pat\_gene\_ID sample change  
## <chr> <chr> <dbl>  
## 1 YAL067C pat1 0.720  
## 2 YAL063C pat1 1.45   
## 3 YAL054C pat1 2.32   
## 4 YAL049C pat1 1.28   
## 5 YAL047C pat1 1.15   
## 6 YAL046C pat1 1.64

anyNA(mRNA\_2)

## [1] TRUE

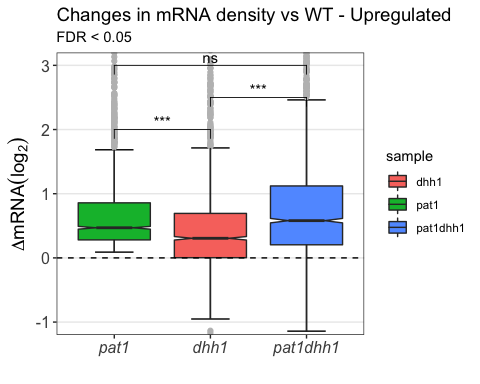
mRNA\_3 <- mRNA\_2 %>% drop\_na()  
  
anyNA(mRNA\_3)

## [1] FALSE

dim(mRNA\_3)

## [1] 5094 3

my\_comparisons <- list(c('pat1','dhh1'),  
c('dhh1','pat1dhh1'),  
c('pat1','pat1dhh1'))  
  
symnum.arg <- list(cutpoints = c(0, 0.001, 0.01, 0.05, 1), symbols = c("\*\*\*", "\*\*", "\*", "ns"))  
  
ggplot(mRNA\_3, aes(x= sample, y=change, fill = sample)) +  
 stat\_boxplot(geom ='errorbar', width = 0.4) +  
 geom\_boxplot(notch = TRUE, outlier.color = "grey") +  
 stat\_compare\_means(data = mRNA\_3, method = "wilcox.test" , comparisons = my\_comparisons, label.y = c(2,2.5,3), tip.length = 0.01, symnum.args = symnum.arg, inherit.aes = FALSE) +  
 coord\_cartesian(ylim = c(-1,3)) + # it acts as zooming into the plot, doesn't change the underlying data for stat\_compare\_mean  
 #scale\_y\_continuous(limits = c(-1,3)) + #this acually scales the data, throwing statistics off  
 scale\_x\_discrete(limits = c("pat1", "dhh1", "pat1dhh1")) +  
 theme\_bw() +  
 theme(panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14), panel.grid.major.x = element\_blank(),  
 axis.text = element\_text(size = 12), axis.text.x = element\_text(face = "italic")) +  
 geom\_hline(yintercept = 0, color = "black", linetype = "dashed" ) +  
 labs(x = "", y = expression(Delta\*"mRNA"(log[2])),  
 title = "Changes in mRNA density vs WT - Upregulated",  
 subtitle = "FDR < 0.05")



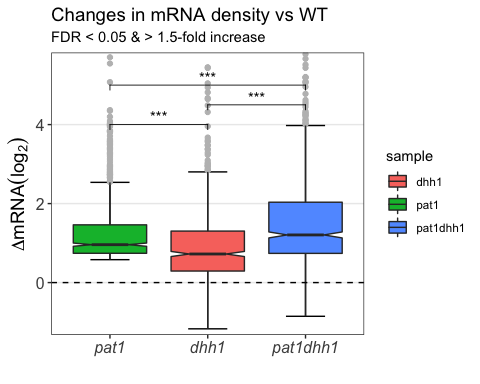
mRNA\_4 <- combined %>%   
 filter(pat\_mRNA\_FDR < 0.05 & pat\_mRNA\_FC > 0.58) %>%   
 select(pat\_gene\_ID, pat\_mRNA\_FC, dhh\_mRNA\_FC, pd\_mRNA\_FC) %>%  
 set\_names(str\_replace(names(.), "\_mRNA\_FC", "1")) %>%   
 set\_names(str\_replace(names(.), "pd1", "pat1dhh1")) %>%   
 gather(sample, change, -pat\_gene\_ID) %>%   
 drop\_na()  
  
head(mRNA\_4)

## # A tibble: 6 x 3  
## pat\_gene\_ID sample change  
## <chr> <chr> <dbl>  
## 1 YAL067C pat1 0.720  
## 2 YAL063C pat1 1.45   
## 3 YAL054C pat1 2.32   
## 4 YAL049C pat1 1.28   
## 5 YAL047C pat1 1.15   
## 6 YAL046C pat1 1.64

dim(mRNA\_4)

## [1] 2096 3

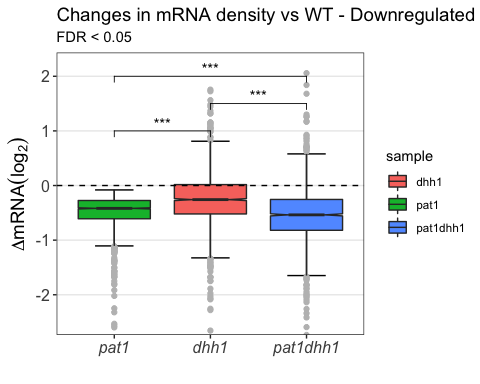
my\_comparisons <- list(c('pat1','dhh1'),  
c('dhh1','pat1dhh1'),  
c('pat1','pat1dhh1'))  
  
symnum.arg <- list(cutpoints = c(0, 0.001, 0.01, 0.05, 1), symbols = c("\*\*\*", "\*\*", "\*", "ns"))  
  
ggplot(mRNA\_4, aes(x= sample, y=change, fill = sample)) +  
 stat\_boxplot(geom ='errorbar', width = 0.4) +  
 geom\_boxplot(notch = TRUE, outlier.color = "grey") +  
 stat\_compare\_means(method = "wilcox.test" , comparisons = my\_comparisons, label.y = c(4,4.5,5), tip.length = 0.01, symnum.args = symnum.arg) +  
 coord\_cartesian(ylim = c(-1,5.5)) +  
 #scale\_y\_continuous(limits = c(-1,5.5)) +  
 scale\_x\_discrete(limits = c("pat1", "dhh1", "pat1dhh1")) +  
 theme\_bw() +  
 theme(panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14), panel.grid.major.x = element\_blank(),  
 axis.text = element\_text(size = 12), axis.text.x = element\_text(face = "italic")) +  
 geom\_hline(yintercept = 0, color = "black", linetype = "dashed" ) +  
 labs(x = "", y = expression(Delta\*"mRNA" (log[2])),  
 title = "Changes in mRNA density vs WT",  
 subtitle = "FDR < 0.05 & > 1.5-fold increase")



mRNA\_5 <- combined %>%   
 filter(pat\_mRNA\_FDR < 0.05 & pat\_mRNA\_FC < 0) %>%   
 select(pat\_gene\_ID, pat\_mRNA\_FC, dhh\_mRNA\_FC, pd\_mRNA\_FC) %>%  
 set\_names(str\_replace(names(.), "\_mRNA\_FC", "1")) %>%   
 set\_names(str\_replace(names(.), "pd1", "pat1dhh1")) %>%   
 gather(sample, change, -pat\_gene\_ID) %>%   
 drop\_na()  
  
dim(mRNA\_5)

## [1] 6531 3

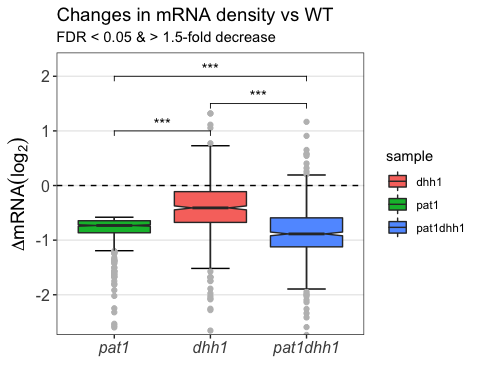
my\_comparisons <- list(c('pat1','dhh1'),  
c('dhh1','pat1dhh1'),  
c('pat1','pat1dhh1'))  
  
symnum.arg <- list(cutpoints = c(0, 0.001, 0.01, 0.05, 1), symbols = c("\*\*\*", "\*\*", "\*", "ns"))  
  
ggplot(mRNA\_5, aes(x= sample, y=change, fill = sample)) +  
 stat\_boxplot(geom ='errorbar', width = 0.4) +  
 geom\_boxplot(notch = TRUE, outlier.color = "grey") +  
 stat\_compare\_means(method = "wilcox.test" , comparisons = my\_comparisons, label.y = c(1,1.5,2), tip.length = 0.01, symnum.args = symnum.arg) +  
 coord\_cartesian(ylim = c(-2.5,2.2)) +  
 #scale\_y\_continuous(limits = c(-2.5,2.2)) +  
 scale\_x\_discrete(limits = c("pat1", "dhh1", "pat1dhh1")) +  
 theme\_bw() +  
 theme(panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14), panel.grid.major.x = element\_blank(),  
 axis.text = element\_text(size = 12), axis.text.x = element\_text(face = "italic")) +  
 geom\_hline(yintercept = 0, color = "black", linetype = "dashed" ) +  
 labs(x = "", y = expression(Delta\*"mRNA" (log[2])),  
 title = "Changes in mRNA density vs WT - Downregulated",  
 subtitle = "FDR < 0.05")



mRNA\_6 <- combined %>%   
 filter(pat\_mRNA\_FDR < 0.05 & pat\_mRNA\_FC < -0.58) %>%   
 select(pat\_gene\_ID, pat\_mRNA\_FC, dhh\_mRNA\_FC, pd\_mRNA\_FC) %>%  
 set\_names(str\_replace(names(.), "\_mRNA\_FC", "1")) %>%   
 set\_names(str\_replace(names(.), "pd1", "pat1dhh1")) %>%   
 gather(sample, change, -pat\_gene\_ID) %>%   
 drop\_na()  
 #mutate(sample = fct\_relevel(sample, "pat\_mRNA\_FC", "dhh\_mRNA\_FC", "pd\_mRNA\_FC"))  
  
dim(mRNA\_6)

## [1] 1809 3

my\_comparisons <- list(c('pat1','dhh1'),  
c('dhh1','pat1dhh1'),  
c('pat1','pat1dhh1'))  
  
symnum.arg <- list(cutpoints = c(0, 0.001, 0.01, 0.05, 1), symbols = c("\*\*\*", "\*\*", "\*", "ns"))  
  
ggplot(mRNA\_6, aes(x= sample, y=change, fill = sample)) +  
 stat\_boxplot(geom ='errorbar', width = 0.4) +  
 geom\_boxplot(notch = TRUE, outlier.color = "grey") +  
 stat\_compare\_means(method = "wilcox.test" , comparisons = my\_comparisons, label.y = c(1,1.5,2), tip.length = 0.01, symnum.args = symnum.arg) +  
 coord\_cartesian(ylim = c(-2.5,2.2)) +  
 # scale\_y\_continuous(limits = c(-2.5,2.2)) +  
 scale\_x\_discrete(limits = c("pat1", "dhh1", "pat1dhh1")) +  
 theme\_bw() +  
 theme(panel.grid.minor = element\_blank(), axis.title = element\_text(size = 14), plot.title = element\_text(size = 14), panel.grid.major.x = element\_blank(),  
 axis.text = element\_text(size = 12), axis.text.x = element\_text(face = "italic")) +  
 geom\_hline(yintercept = 0, color = "black", linetype = "dashed" ) +  
 labs(x = "", y = expression(Delta\*"mRNA" (log[2])),  
 title = "Changes in mRNA density vs WT",  
 subtitle = "FDR < 0.05 & > 1.5-fold decrease")



There seems to be an overlap between mRNAs that are regulated by pat1 are also regulated by dhh1.

Then, I wanted to test if it is true for all mRNAs.

Heatmap

Collecting appropriate data.

head(combined)

## # A tibble: 6 x 7  
## pat\_gene\_ID pat\_mRNA\_FC pat\_mRNA\_FDR pd\_mRNA\_FC pd\_mRNA\_FDR dhh\_mRNA\_FC  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 YAL067C 0.720 8.93e- 8 NA NA -0.630   
## 2 YAL063C 1.45 1.02e- 59 0.808 2.25e- 8 0.118   
## 3 YAL054C 2.32 1.81e-103 2.94 5.44e-18 1.35   
## 4 YAL049C 1.28 3.38e- 93 -0.0651 4.07e- 1 0.00480  
## 5 YAL047C 1.15 5.53e- 63 -0.00845 9.44e- 1 -0.117   
## 6 YAL046C 1.64 5.03e-150 0.302 2.99e- 3 -0.186   
## # ... with 1 more variable: dhh\_mRNA\_FDR <dbl>

mRNA\_h <- combined %>%   
 select(-pat\_mRNA\_FDR,-pd\_mRNA\_FDR, -dhh\_mRNA\_FDR, -pat\_gene\_ID) %>%   
 set\_names(str\_replace(names(.), "\_mRNA\_FC", "")) %>%   
 drop\_na()   
  
mRNA\_h1 <- mRNA\_h[,c(1,3,2)]  
head(mRNA\_h1)

## # A tibble: 6 x 3  
## pat dhh pd  
## <dbl> <dbl> <dbl>  
## 1 1.45 0.118 0.808   
## 2 2.32 1.35 2.94   
## 3 1.28 0.00480 -0.0651   
## 4 1.15 -0.117 -0.00845  
## 5 1.64 -0.186 0.302   
## 6 1.49 0.492 0.597

head(mRNA\_h)

## # A tibble: 6 x 3  
## pat pd dhh  
## <dbl> <dbl> <dbl>  
## 1 1.45 0.808 0.118   
## 2 2.32 2.94 1.35   
## 3 1.28 -0.0651 0.00480  
## 4 1.15 -0.00845 -0.117   
## 5 1.64 0.302 -0.186   
## 6 1.49 0.597 0.492

dim(mRNA\_h)

## [1] 4777 3

str(mRNA\_h)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 4777 obs. of 3 variables:  
## $ pat: num 1.45 2.32 1.28 1.15 1.64 ...  
## $ pd : num 0.80831 2.9375 -0.06509 -0.00845 0.3018 ...  
## $ dhh: num 0.1182 1.353 0.0048 -0.1171 -0.1859 ...

mRNA\_h2 <- data.matrix(mRNA\_h1)  
head(mRNA\_h2)

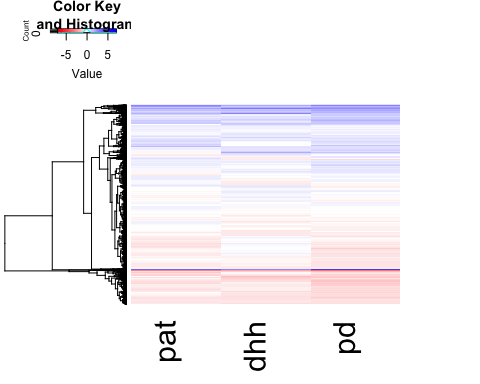
## pat dhh pd  
## [1,] 1.447031 0.118170173 0.80830607  
## [2,] 2.317785 1.352995000 2.93750030  
## [3,] 1.276213 0.004801524 -0.06509376  
## [4,] 1.153527 -0.117116751 -0.00844742  
## [5,] 1.638963 -0.185914323 0.30179834  
## [6,] 1.488186 0.492428158 0.59659985

str(mRNA\_h2)

## num [1:4777, 1:3] 1.45 2.32 1.28 1.15 1.64 ...  
## - attr(\*, "dimnames")=List of 2  
## ..$ : NULL  
## ..$ : chr [1:3] "pat" "dhh" "pd"

library(gplots)  
library(RColorBrewer)

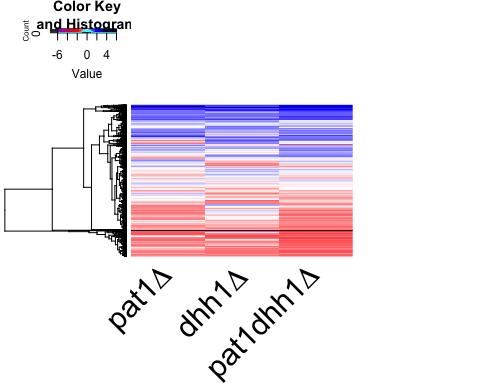
#pdf(file = "mRNA\_changes.pdf")  
heatmap.2(mRNA\_h2, dend="row",   
 col=colorRampPalette(c("red","white","blue"), bias=1.01)(n=100),  
 labRow=c(""), trace="none", Colv = FALSE)



#dev.off()

The heatmap doesn’t reflect the changes of 1.5 fold well (too white). Changing color scheme

my\_palette <- colorRampPalette(c("purple","red","white","blue","black"))(n=499)  
col\_breaks <- c(seq(-6,-3,length=100),  
 seq(-2.9,-0.51,length=100),  
 seq(-0.5,0.5,length=100),  
 seq(0.51,2.9,length=100),   
 seq(3,6,length=100))   
anno <- c(expression("pat1"\*Delta), expression("dhh1"\*Delta), expression("pat1dhh1"\*Delta))  
#my\_palette <- colorRampPalette(rev(brewer.pal(8, "Spectral")))(n=60)  
#pdf(file ="hm2.pdf")  
heatmap.2(mRNA\_h2, dend="row",   
 col= my\_palette,  
 labRow=c(""), trace="none", Colv = FALSE, breaks= col\_breaks, labCol = anno, srtCol = 45,  
 margins = c(8,8))



#dev.off()