P-Value Visualizations

Abigail Seeger

January 11, 2021

Contents

Load necessary packages

```r
library(dplyr)
library(tidyverse)
library(ggplot2)
### Lemon is used in ggplot2 -
### facet_rep_grid modification
library(lemon)
library(data.table)
library(ggthemes)
library(extrafont)
### Routliers is used for outliersmad to
### find outliers
library(Routliers)
library(stringi)
library(wesanderson)
library(viridis)
library(reshape2)
library(sm)
library(lme4)
library(lmerTest)
library(lsmeans)
library(car)
library(ggcorrplot)
library(preprocessCore)
library(grid)
```

Read in the data

```
depi_data <- read.table("C:/Users/Owner/Documents/Research/Shiu_Lab/Shiu_Lab_R/Data/Clean_DEPI_Data_V2.
    sep = ",", header = TRUE)
```

Functions

Note that I modified this function to use the normalized_value when calcuating p-values, not the measured value.

```
p_value <- function(data_frame) {
    ### Initialize an empty data frame
    out = data.frame()
    ### For each genotype:
    for (i in unique(filter(data_frame, genotype !=
        "Col0")$genotype)) {
        ### We don't want to make comparisons of WT
        ### to itself - this could impact FDR
        ### correction
        indiv_data <- data_frame %>% ### Focus on each time point and
        ### measurement
        group_by(time_point, measurement) %>%
            ### Create a column with the number of WT
        ### individual plants and the number of
        ### plants for each genotype Use this later
        ### to calculate effect size
        mutate(n_genotype = length(normalized_value[genotype ==
            i]), n_wt = length(normalized_value[genotype ==
            "Col0"])) %>% ### Create a column of p-values using a
        ### nonparametric Wilcox test

        ### Default set to exact = TRUE, because
        ### our sample sizes are too small to use a
        ### normal approximation But, when there
        ### are ties in the values (i.e. one value
        ### appears twice in the ranking process),
        ### wilcox.test returns to the normal
        ### approximation and spits out a warning
        ### message This may be a problem - include
        ### correct = FALSE to stop this from
        ### happening
        ### Paired = FALSE, because the Col0 plants
        ### are independet from each genotype
        ### Correct = FALSE turns off the
        ### continuity correction
        mutate(p = (wilcox.test(normalized_value[genotype ==
            i], normalized_value[genotype ==
            "Col0"], correct = FALSE, paired = FALSE))$p.value) %>%

        ### Add a column with each genotype
        mutate(genotype = i) %>% # mutate(number = unique(filter(feb_data,
        # genotype == i)$number))%>%
        # mutate(number_2 =
        # unique(filter(feb_data, genotype ==
        # i)$number_2))%>%
        select(time_point, genotype, measurement,
            day, p, n_wt, n_genotype)


        ### Add individual information to the main
        ### data frame
        out <- rbind(as.data.frame(indiv_data),
            out)
    }
    return(out)
}
```

Note that I removed the effect size calculations from this function in order to save time when running the code.

```
corrected_p_value <- function(data_frame) {
    out <- data_frame %>% ### For some reason, I have multiple copies
    ### of each row
    distinct() %>% ### Group by time point and measurement -
    ### we are correcting by the number of
    ### genotypes
    group_by(time_point, measurement) %>%
        mutate(p_adj = p.adjust(p, method = "fdr"))
    ### Gather the data to make it easier to
    ### plot according to whether p was
    ### adjusted
    out <- gather(out, type, p, p, p_adj) %>%
        arrange(genotype, time_point)
    return(out)
}
```

```
add_number <- function(data_frame) {
    ### First, if the genotype is Col0 (only
    ### genotype with length 4), assign 0 as
    ### number Else, assign number as genotype
    ### with 'mpk' removed Example: mpk1 will
    ### be 1, mpk1-17 will be 1-17
    data_frame <- data_frame %>% mutate(number = ifelse(genotype !=
        "Col0", (stri_sub(genotype, 4, length(genotype))),
        0))
    ### Next, for all double mutants, replace
    ### ’-’ with ’0’ Example: 1-17 becomes 1017
    data_frame$number <- as.numeric(gsub("-",
        "0", data_frame$number))
    ### Almost there! There’s a problem with
    ### two single digit double mutants We need
    ### a four digit number to sort correctly
    ### Example: mpk1-3 -> 1-3 -> 103, but we
    ### need it to be 1003 to sort correctly
    data_frame$number[data_frame$number ==
        "103"] <- "1003"
    data_frame$number[data_frame$number ==
        "506"] <- "5006"
    data_frame$number[data_frame$number ==
        "608"] <- "6008"
    data_frame$number[data_frame$number ==
        "609"] <- "6009"
    ### Convert number to a numberic in order
    ### to sort
    data_frame$number <- as.numeric(data_frame$number)
    data_frame <- data_frame %>% arrange(number)
    data_frame <- data_frame %>% mutate(number_2 = number)
    data_frame$number_2[nchar(data_frame$number_2) ==
        4] <- 0
    data_frame$number_2[nchar(data_frame$number_2) ==
        5] <- 0
    return(data_frame)
}
```

Calculate p-values

```
### Create subsets of the data for each
### month
dec_data <- depi_data %>% filter(month ==
    "Dec")
jan_data <- depi_data %>% filter(month ==
    "Jan")
feb_data <- depi_data %>% filter(month ==
    "Feb")

### Calculate the p-values for each month
dec_p_values <- p_value(dec_data)
jan_p_values <- p_value(jan_data)
feb_p_values <- p_value(feb_data)

### Correct the p-values for each month
corrected_dec_p_values <- corrected_p_value(dec_p_values)
corrected_jan_p_values <- corrected_p_value(jan_p_values)
corrected_feb_p_values <- corrected_p_value(feb_p_values)
```

P-value visualizations

December

First, create subsets of the data for each measurement.

```
december_npq <- corrected_dec_p_values %>%
    filter(measurement == "npq", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))

december_phi2 <- corrected_dec_p_values %>%
    filter(measurement == "phi2", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))

december_leafarea <- corrected_dec_p_values %>%
    filter(measurement == "leafarea", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))
```

Modify the bins and properly sort the bins and genotypes for the p-value heat maps.

Note that because leaf area has no significant p-values, moving forward I will only create visualizations for npq and phi2:

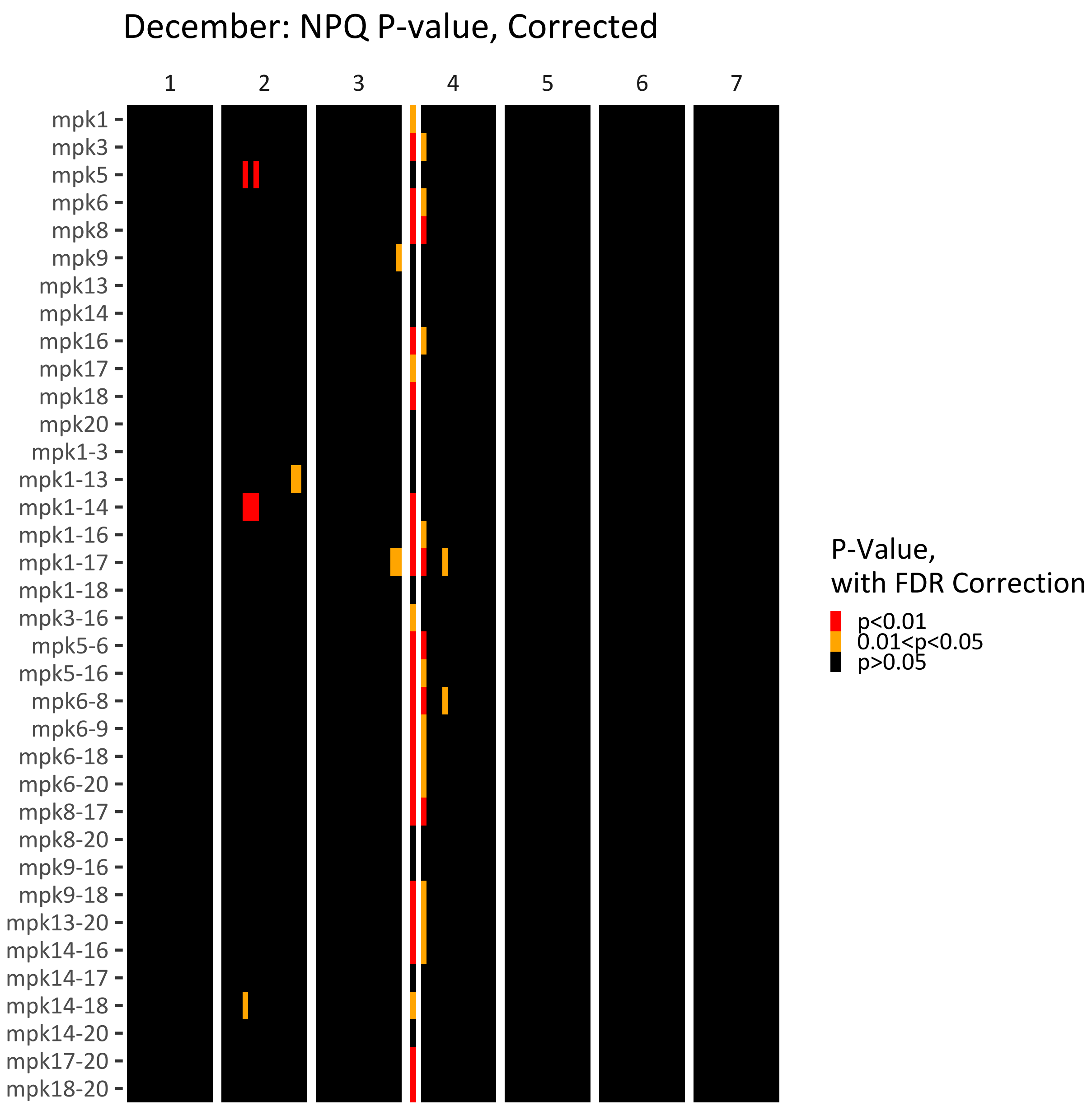
```
summary(december_leafarea$p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1210  0.9112  0.9772  0.9416  1.0000  1.0000
```

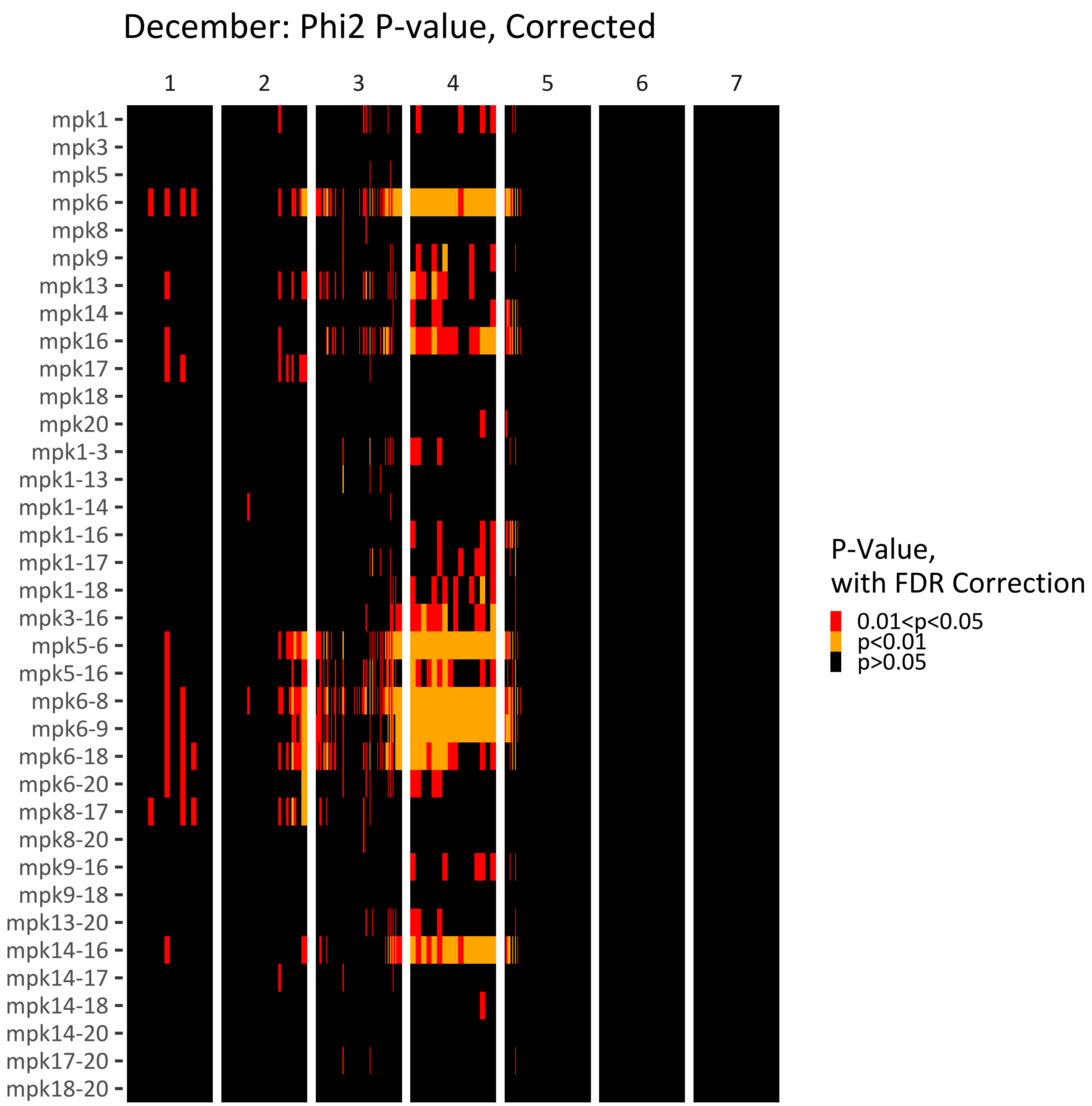
```
### In order to use these bins as the fill
### in a heat map, convert to a factor
december_npq$bin <- as.factor(december_npq$bin)
december_phi2$bin <- as.factor(december_phi2$bin)

december_npq <- add_number(december_npq)
december_phi2 <- add_number(december_phi2)

### We want p<0.1 to be first in the
### legend, so refactor with p<0.01 as the
### first term
december_npq$bin <- relevel(december_npq$bin,
    "p<0.01")
december_phi2$bin <- relevel(december_phi2$bin,
    "0.01<p<0.05")

### Reorder by number so heat map has WT
### first, then single, then double mutants
december_npq$genotype <- reorder(december_npq$genotype,
    december_npq$number)
december_phi2$genotype <- reorder(december_phi2$genotype,
    december_phi2$number)
```

Create heat maps for the adjusted p-values.

```
##### ----- Dec p-value heat map - NPQ ------
ggplot(data = december_npq, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "December: NPQ P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + # scale_x_continuous(breaks =
# round(c(0,15,24,39.5,48,63.7,72,87,96,112,
# 120,135,144,159,168,183,192,207,216,231,240,255,264,279),0))+
theme_tufte(base_family = "Calibri", base_size = 50) +
    theme(strip.background.y = element_blank(),
        strip.text.y = element_blank(), axis.title.x = element_blank(),
        axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        panel.spacing = unit(0, "lines")) +
    scale_fill_manual(values = c("red", "orange",
        "black"))
```



```
##### ----- Dec p-value heat map - phi2
##### ------
ggplot(data = december_phi2, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "December: Phi2 P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + theme_tufte(base_family = "Calibri",
    base_size = 50) + theme(strip.background.y = element_blank(),
    strip.text.y = element_blank(), axis.title.x = element_blank(),
    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
    panel.spacing = unit(0, "lines")) + scale_fill_manual(values = c("red",
    "orange", "black"))
```



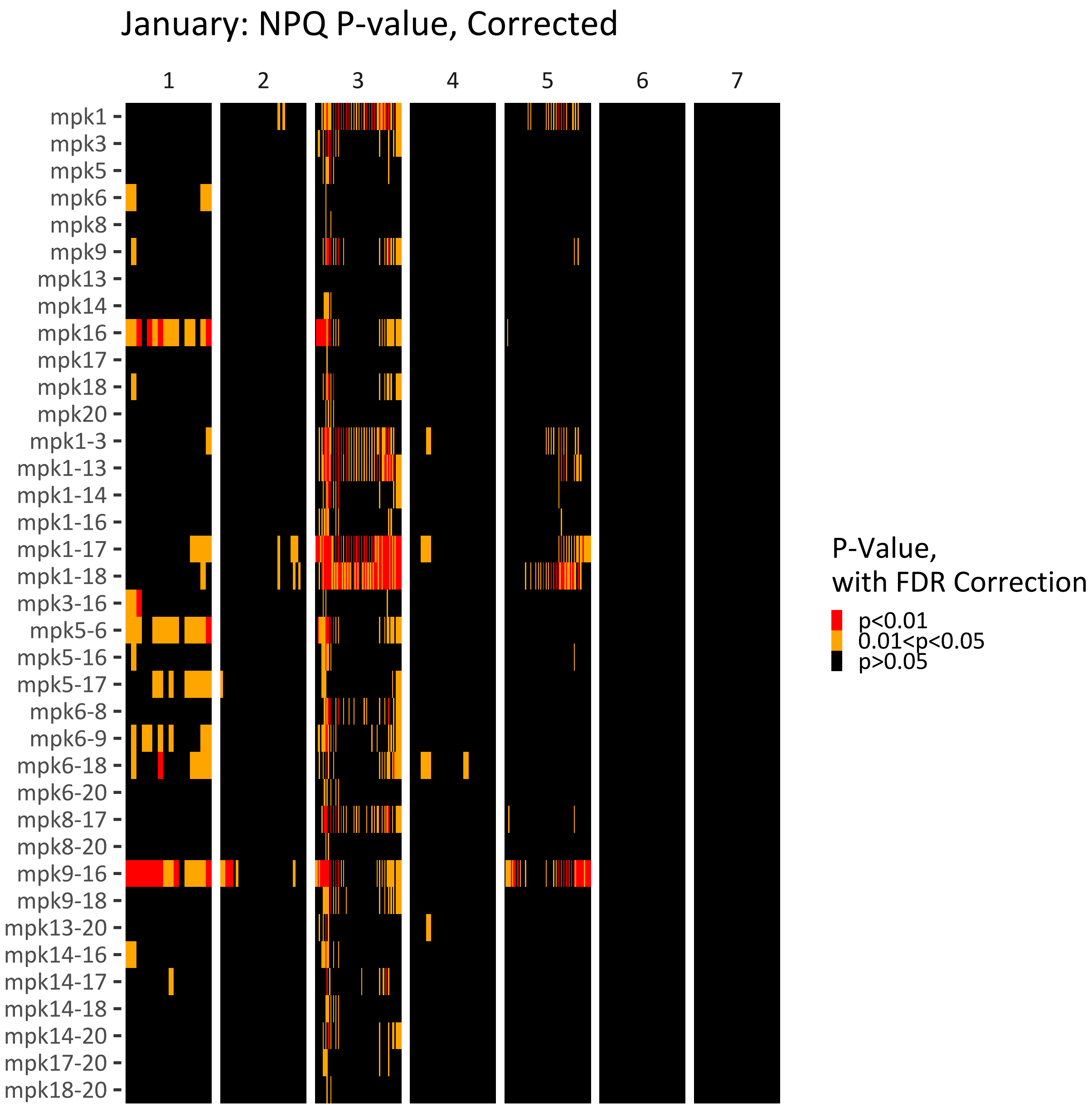
January

```
january_npq <- corrected_jan_p_values %>%
    filter(measurement == "npq", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))

january_phi2 <- corrected_jan_p_values %>%
    filter(measurement == "phi2", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))

january_leafarea <- corrected_jan_p_values %>%
    filter(measurement == "leafarea", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))
```

```
### In order to use these bins as the fill
### in a heat map, convert to a factor
january_npq$bin <- as.factor(january_npq$bin)
january_phi2$bin <- as.factor(january_phi2$bin)
january_leafarea$bin <- as.factor(january_leafarea$bin)

january_npq <- add_number(january_npq)
january_phi2 <- add_number(january_phi2)
january_leafarea <- add_number(january_leafarea)


### We want p<0.1 to be first in the
### legend, so refactor with p<0.01 as the
### first term
january_npq$bin <- relevel(january_npq$bin,
    "p<0.01")
january_phi2$bin <- relevel(january_phi2$bin,
    "0.01<p<0.05")
january_leafarea$bin <- relevel(january_leafarea$bin,
    "p<0.01")

### Reorder by number so heat map has WT
### first, then single, then double mutants
january_npq$genotype <- reorder(january_npq$genotype,
    january_npq$number)
january_phi2$genotype <- reorder(january_phi2$genotype,
    january_phi2$number)
january_leafarea$genotype <- reorder(january_leafarea$genotype,
    january_leafarea$number)
```

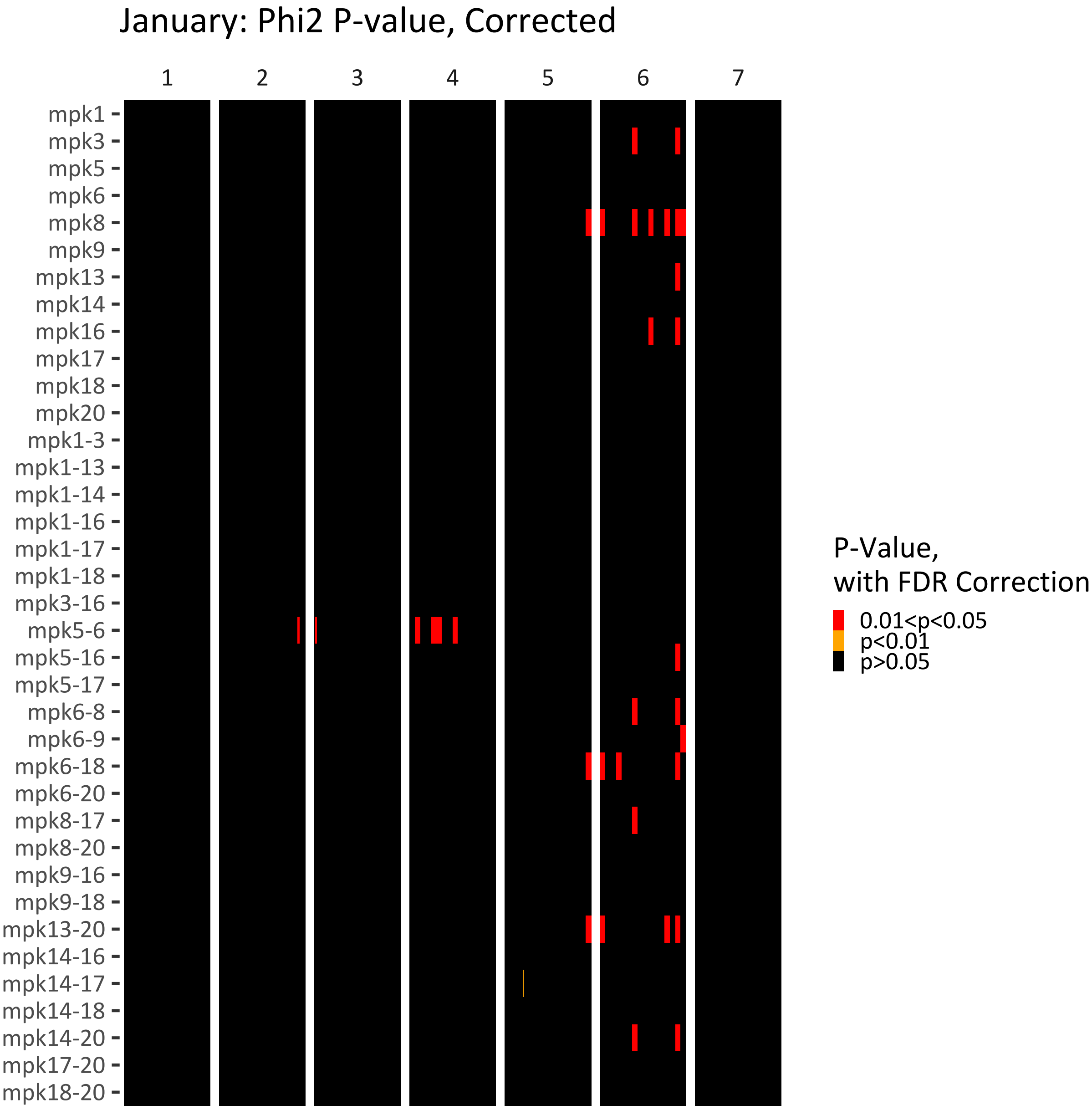
Create heat maps for the adjusted p-values.

```
##### ----- January p-value heat map - NPQ
##### ------
ggplot(data = january_npq, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "January: NPQ P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + # scale_x_continuous(breaks =
# round(c(0,15,24,39.5,48,63.7,72,87,96,112,
# 120,135,144,159,168,183,192,207,216,231,240,255,264,279),0))+
theme_tufte(base_family = "Calibri", base_size = 50) +
    theme(strip.background.y = element_blank(),
        strip.text.y = element_blank(), axis.title.x = element_blank(),
        axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        panel.spacing = unit(0, "lines")) +
    scale_fill_manual(values = c("red", "orange",
        "black"))
```

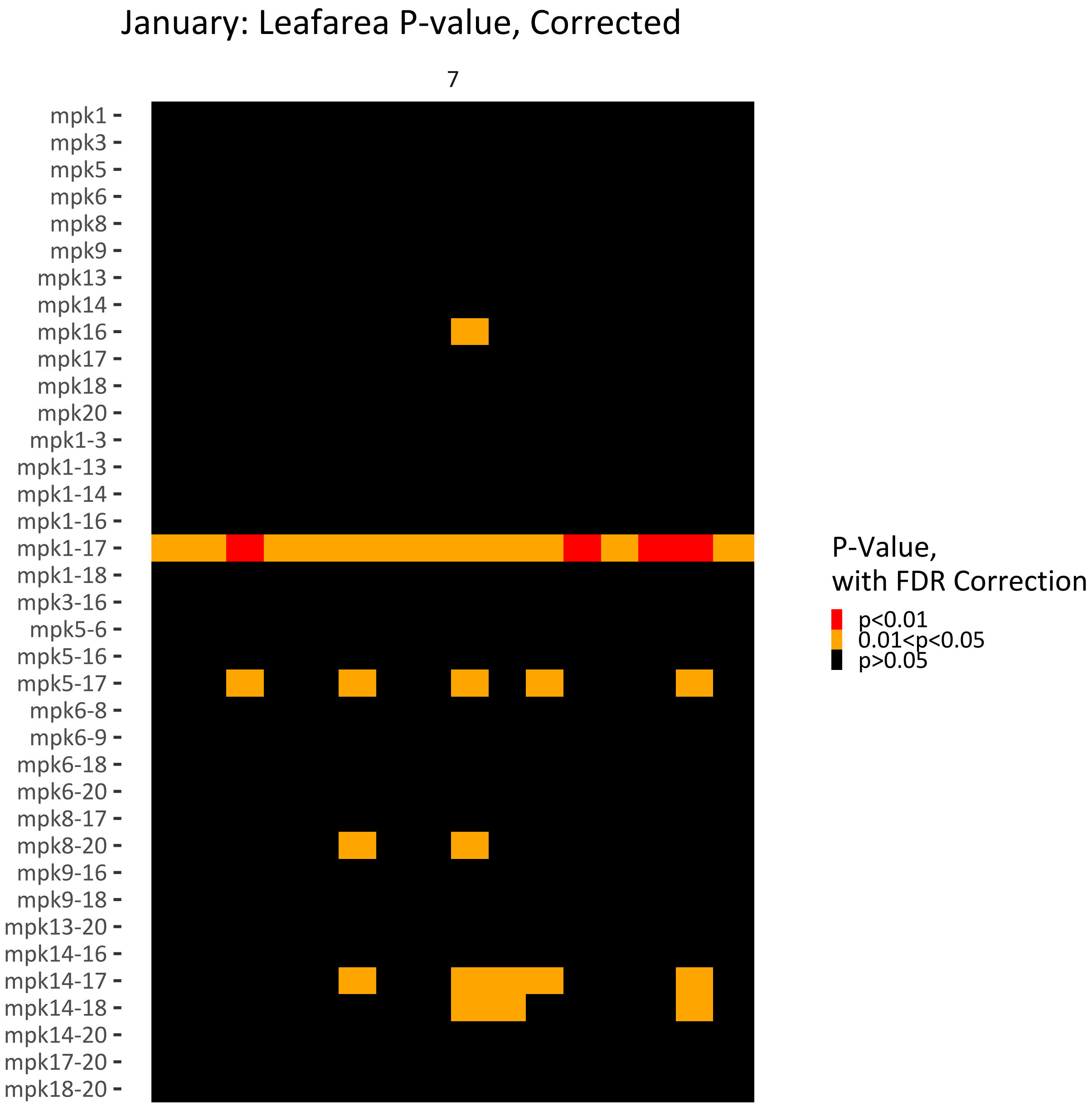


```
##### ----- January p-value heat map - phi2
##### ------
ggplot(data = january_phi2, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "January: Phi2 P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + theme_tufte(base_family = "Calibri",
```

```r
    base_size = 50) + theme(strip.background.y = element_blank(),
    strip.text.y = element_blank(), axis.title.x = element_blank(),
    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
    panel.spacing = unit(0, "lines")) + scale_fill_manual(values = c("red",
    "orange", "black"))
```



```r
##### ----- January p-value heat map - phi2
##### ------
ggplot(data = january_leafarea, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "January: Leafarea P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + theme_tufte(base_family = "Calibri",
    base_size = 50) + theme(strip.background.y = element_blank(),
    strip.text.y = element_blank(), axis.title.x = element_blank(),
    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
    panel.spacing = unit(0, "lines")) + scale_fill_manual(values = c("red",
    "orange", "black"))
```



February

```
february_npq <- corrected_feb_p_values %>%
    filter(measurement == "npq", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))

february_phi2 <- corrected_feb_p_values %>%
    filter(measurement == "phi2", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))

february_leafarea <- corrected_feb_p_values %>%
    filter(measurement == "leafarea", type ==
        "p_adj") %>% mutate(bin = case_when((p <
    0.01) ~ "p<0.01", (p > 0.01 & p < 0.05) ~
    "0.01<p<0.05", (p > 0.05) ~ "p>0.05"))
```

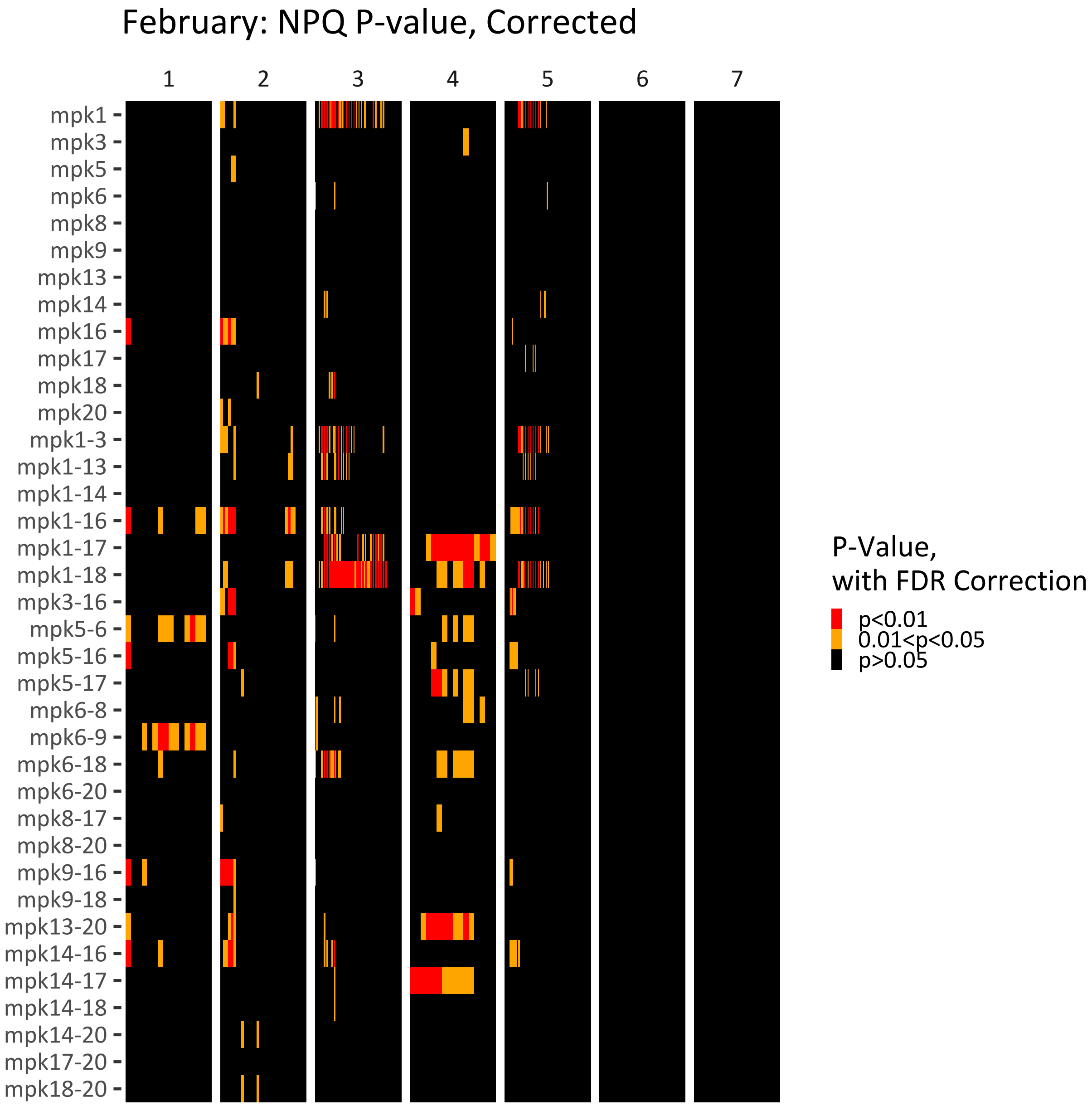
Once again, it looks like the p-values for leafarea are never less than 0.05:

```
summary(february_leafarea$p)
```

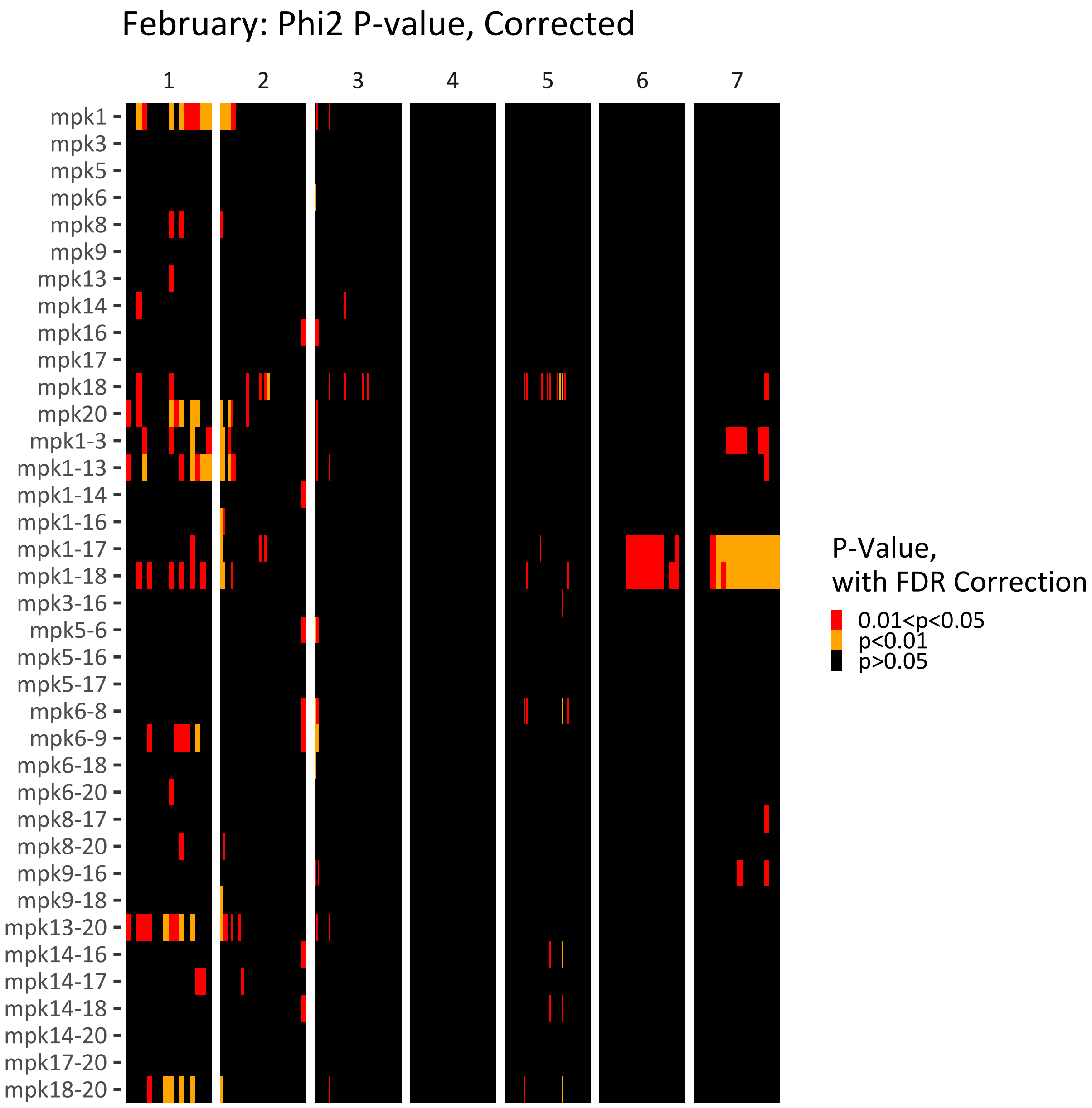
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1364  0.7992  0.9315  0.8752  0.9850  1.0000
```

```
### In order to use these bins as the fill
### in a heat map, convert to a factor
february_npq$bin <- as.factor(february_npq$bin)
february_phi2$bin <- as.factor(february_phi2$bin)

february_npq <- add_number(february_npq)
february_phi2 <- add_number(february_phi2)

### We want p<0.1 to be first in the
### legend, so refactor with p<0.01 as the
### first term
february_npq$bin <- relevel(february_npq$bin,
    "p<0.01")
february_phi2$bin <- relevel(february_phi2$bin,
    "0.01<p<0.05")

### Reorder by number so heat map has WT
### first, then single, then double mutants
february_npq$genotype <- reorder(february_npq$genotype,
    february_npq$number)
february_phi2$genotype <- reorder(february_phi2$genotype,
    february_phi2$number)
```

```
##### ----- February p-value heat map - NPQ
##### ------
ggplot(data = february_npq, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "February: NPQ P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + # scale_x_continuous(breaks =
# round(c(0,15,24,39.5,48,63.7,72,87,96,112,
# 120,135,144,159,168,183,192,207,216,231,240,255,264,279),0))+
theme_tufte(base_family = "Calibri", base_size = 50) +
    theme(strip.background.y = element_blank(),
        strip.text.y = element_blank(), axis.title.x = element_blank(),
        axis.text.x = element_blank(), axis.ticks.x = element_blank(),
        panel.spacing = unit(0, "lines")) +
    scale_fill_manual(values = c("red", "orange",
        "black"))
```



```
##### ----- February p-value heat map - phi2
##### ------
ggplot(data = february_phi2, aes(x = time_point,
    y = genotype, fill = bin)) + labs(fill = "P-Value, \nwith FDR Correction",
    x = "Hours", y = NULL, title = "February: Phi2 P-value, Corrected") +
    geom_tile(height = 1.65, width = 1.1) +
    facet_grid(genotype ~ day, scales = "free",
        switch = "y") + theme_tufte(base_family = "Calibri",
    base_size = 50) + theme(strip.background.y = element_blank(),
    strip.text.y = element_blank(), axis.title.x = element_blank(),
    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
    panel.spacing = unit(0, "lines")) + scale_fill_manual(values = c("red",
    "orange", "black"))
```


February: Phi2 P-value, Corrected
mpk1
mpk3
mpk5
mpk6
mpk8
mpk9
mpk13
mpk14
mpk16
mpk17
mpk18
mpk20
mpk1-3
mpk1-13
mpk1-14
mpk1-16
mpk1-17
mpk1-18
mpk3-16
mpk5-6
mpk5-16
mpk5-17
mpk6-8
mpk6-9
mpk6-18
mpk6-20
mpk8-17
mpk8-20
mpk9-16
mpk9-18
mpk13-20
mpk14-16
mpk14-17
mpk14-18
mpk14-20
mpk17-20
mpk18-20
P-Value,
with FDR Correction
0.01<p<0.05
p<0.01
p>0.05