Statistical analysis of data from an anthocyanin measurement experiment in *A. thaliana* seedlings

CrisprCat

Contents

Introduction

This script describes the statistical analysis of an experiment in which anthocyanin levels in 5 day old *A. thaliana* seedlings grown in red light were determined. The anthocyanin levels of 3 different mutants and wildtype seedlings were analysed. To identify statistical significant differences among the means of anthocyanin levels of the mutants an One-Way ANOVA is used.

The data used for this example is available under DOI: 10.1038/s41477-020-0725-0 and is the source data of Figure 1d.

Setup

Installing and loading required packages

```
# Store package names, required for the analysis in a vector
packages <- c("tidyverse", "car", "multcomp")

# Install packages that are not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
```

```
}

# Load packages
invisible(lapply(packages, library, character.only = TRUE))
```

Load the data

The data is usually stored in a .csv file with one column describing the "Gentoype" and a second column describing the measured value "anthocyanin_level".

```
# Read the .csv file
antho <- read.csv(file = "anthocyanin_levels.CSV", sep = ";", header = T)

# Change the class of Genotype from character to factor
antho$Genotype <- as.factor(antho$Genotype)

# Transform the data
#   The log transformation is required so that the data meets the assumptions for
#   an ANOVA
antho$T_log <- log(antho$anthocyanin_level)
```

Statistical analysis

Key assumptions for the One-Way ANOVA are that the data
* follows a normal distribution
* shows homogeneity of variances

As this biological data is independently and randomly sampled from a population I assume that these requirements are met after log transformation. However, it is also possible to test for these assumptions.

Test the assumption of equal variances

With the Brown-Forsythe test the null hypothesis, that the variances of the analysed groups (anthocyanin levels in each genotype) are equal, is tested.

Brown-Forsythe test

The Brown-Forsythe test is more robust to not gaussian distributed data than the Levene's test.

```
# Test for homoscedasticity
leveneTest(antho$T_log ~ antho$Genotype, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3   2.001 0.1545
##       16
```

As the p-value (Pr) is > 0.05, the null hypothesis cannot be rejected.

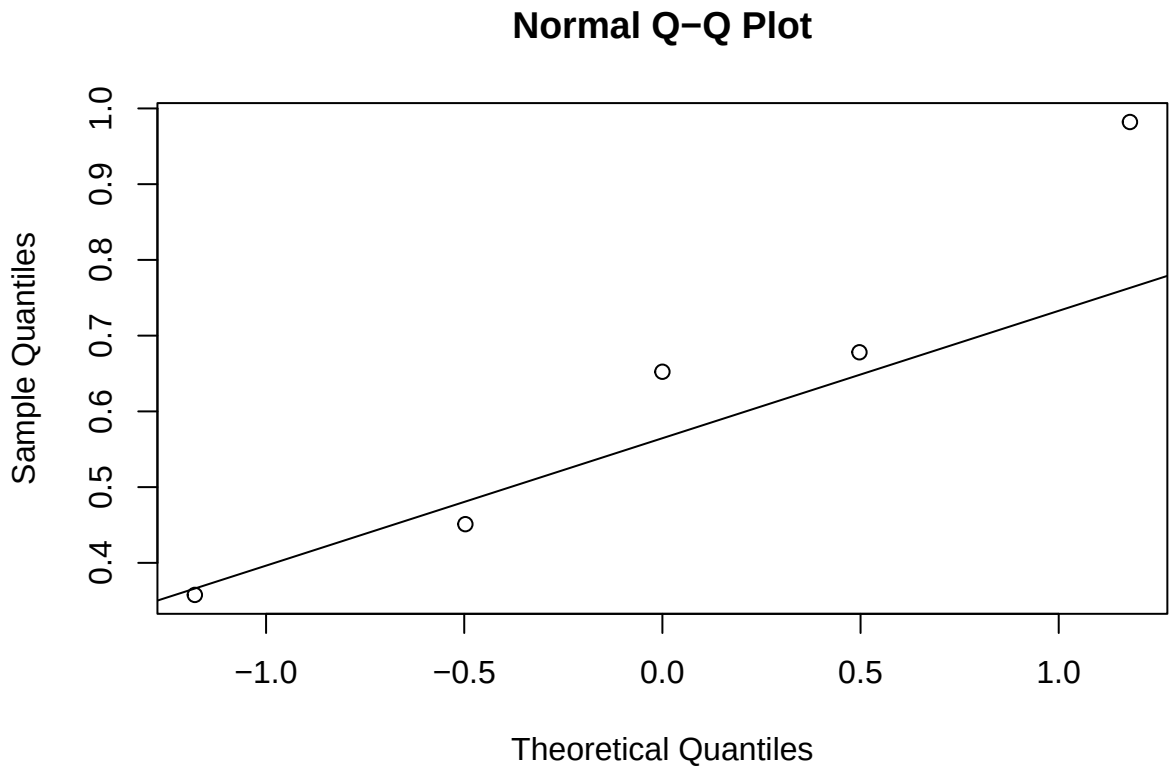
Test the assumption of normal distribution

The assumption of normal distribution of the data in each group = (genotype) can be graphically tested with a QQ-plot or with the Shapiro-Wilk's test.

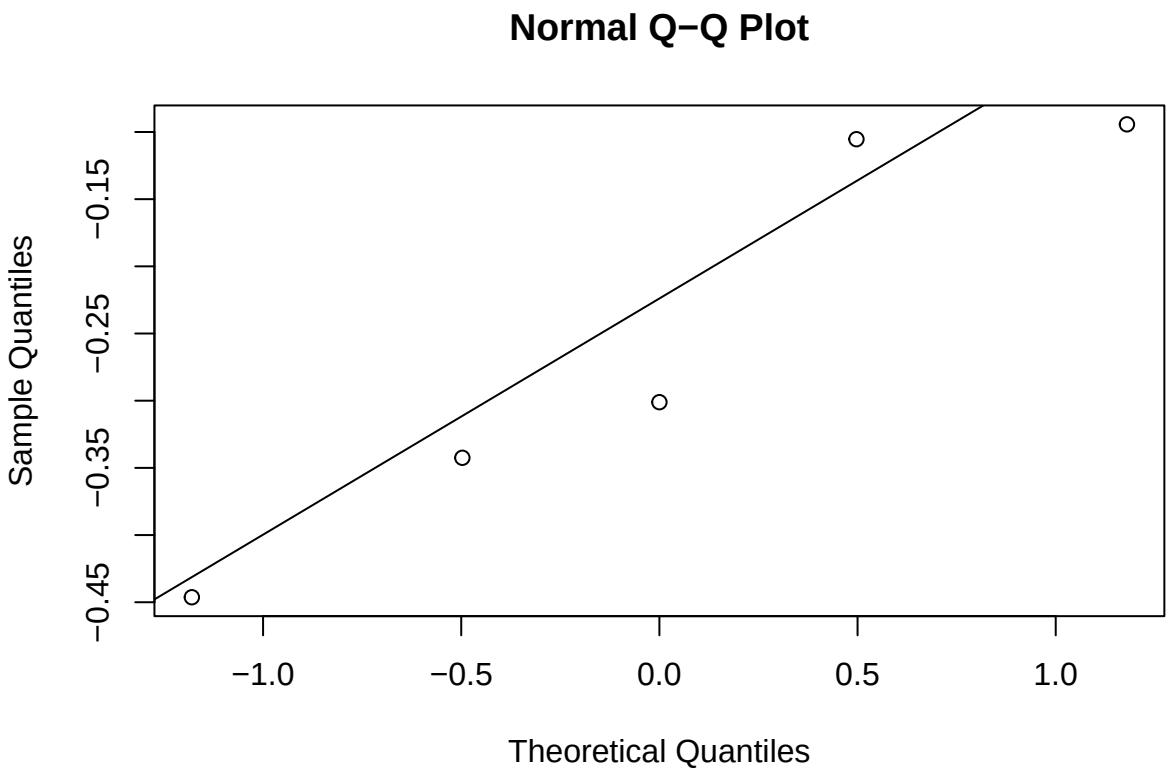
QQ-plot

When the points in the QQ-plot show linearity it suggests the data is normally distributed.

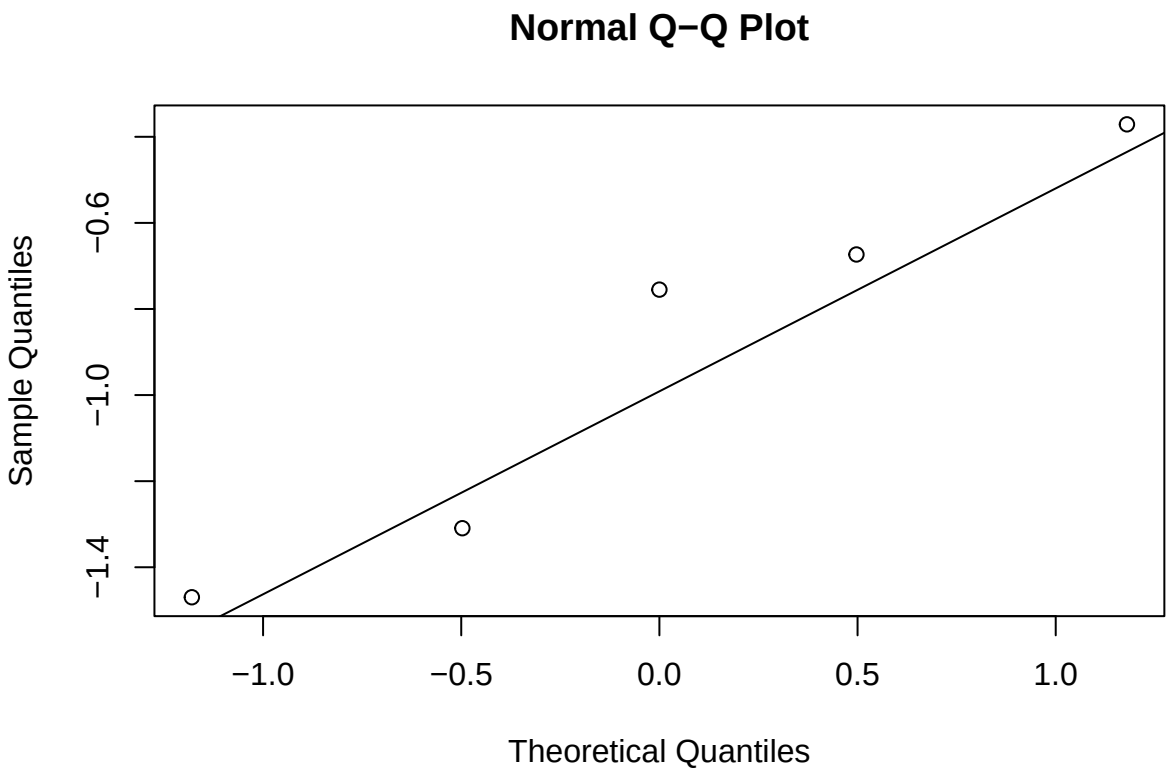
```
# Create a QQ-plot for WT anthocyanin levels
qqnorm(antho$T_log[antho$Genotype == 'WT'])
qqline(antho$T_log[antho$Genotype == 'WT'])
```



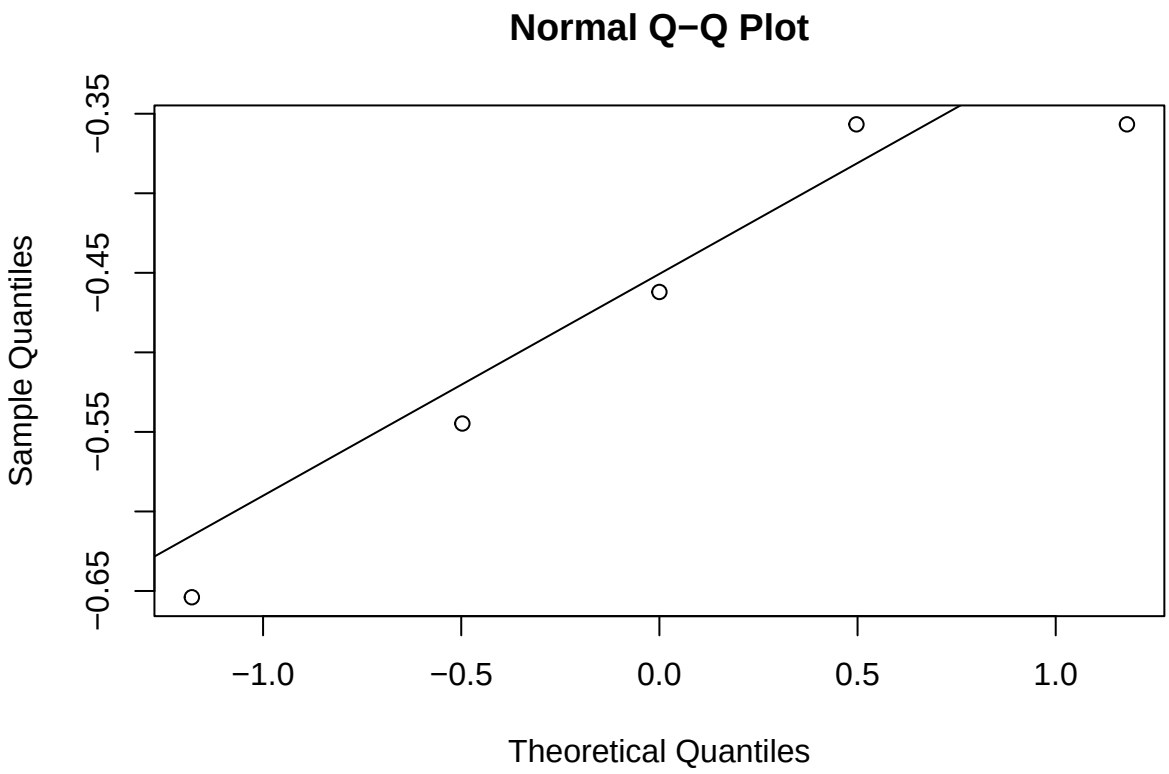

```
# Create a QQ-plot for hy5 anthocyanin levels
qqnorm(antho$T_log[antho$Genotype == 'hy5'])
qqline(antho$T_log[antho$Genotype == 'hy5'])
```




```
# Create a QQ-plot for bbx202122 anthocyanin levels
qqnorm(antho$T_log[antho$Genotype == 'bbx202122'])
qqline(antho$T_log[antho$Genotype == 'bbx202122'])
```




```
# Create a QQ-plot for bbx202122hy5 anthocyanin levels
qqnorm(antho$T_log[antho$Genotype == 'bbx202122hy5'])
qqline(antho$T_log[antho$Genotype == 'bbx202122hy5'])
```



Shapiro-Wilk's test

With the Shapiro-Wilk's test the null hypothesis, that the the samples in each analysed group (anthocyanin levels in each genotype) follow a normal distribution, is tested.

```
# Test for normal distribution
shapiro.test(antho$T_log[antho$Genotype == 'WT'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  antho$T_log[antho$Genotype == "WT"]
## W = 0.94811, p-value = 0.7237
```

```
# Test for normal distribution
shapiro.test(antho$T_log[antho$Genotype == 'hy5'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  antho$T_log[antho$Genotype == "hy5"]
## W = 0.89675, p-value = 0.3922
```

```
# Test for normal distribution
shapiro.test(antho$T_log[antho$Genotype == 'bbx202122'])
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  antho$T_log[antho$Genotype == "bbx202122"]
## W = 0.92631, p-value = 0.5714
```

```
# Test for normal distribution
shapiro.test(antho$T_log[antho$Genotype == 'bbx202122hy5'])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  antho$T_log[antho$Genotype == "bbx202122hy5"]
## W = 0.90745, p-value = 0.4524
```

As the p-value (Pr) in each group is > 0.05, the null hypothesis cannot be rejected.

One-Way ANOVA

With the One-Way ANOVA the null hypothesis, that there are no statistical significant differences among the means of the anthocyanin levels of the mutants, is tested.

```
# Fit an Analysis of Variance model
res.aov <- aov(T_log ~ Genotype, data = antho)

# Show the results of the fitted model
summary(res.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Genotype     3  6.289  2.0965   27.15 1.63e-06 ***
## Residuals   16  1.235  0.0772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value (Pr) is < 0.05, the null hypothesis can be rejected.

Multiple comparisons

To determine if the means of specific groups are statistically significant different from each other the Tukey HSD post-hoc test is computed.

```
# Perform the post-hoc test
TukeyHSD(res.aov, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = T_log ~ Genotype, data = antho)
##
## $Genotype
##                              diff        lwr       upr     p adj
## bbx202122hy5-bbx202122 0.4408802 -0.0619151 0.9436755 0.0965163
## hy5-bbx202122          0.6577773  0.1549820 1.1605726 0.0086453
## WT-bbx202122           1.5399255  1.0371302 2.0427208 0.0000009
## hy5-bbx202122hy5       0.2168971 -0.2858983 0.7196924 0.6152044
## WT-bbx202122hy5        1.0990453  0.5962499 1.6018406 0.0000619
## WT-hy5                 0.8821482  0.3793529 1.3849435 0.0006557
```

When the p adj value is < 0.05 the null hypothesis, that the means of the two compared groups are not statistically significant different, is rejected.

To assign letters to each group, that indicate groups that are statistically significant different from each other the cld() function is used.

```
# Create a glht (general linear hypotheses) object
Tukey_results = glht(res.aov, linfct = mcp(Genotype = "Tukey"))

# Create a cld (compact letter display) object of the pairwise comparisons
tuk_cld <- cld(Tukey_results, decreasing = TRUE)

# Display the letters assigned to each group
print(tuk_cld)
```

```
##    bbx202122 bbx202122hy5          hy5           WT
##          "c"          "bc"          "b"          "a"
```

Data visualization

Calculate summary statistics

These summary statistics are used to create a meaningful data visualization of the experiments result.

```
# Calculate summary statistics
antho_summary = antho %>%
  group_by(Genotype) %>%
  summarise(mean_antho = mean(anthocyanin_level), # Calculate the mean of anthocyanin
                                                  # levels for each genotype
            sd_antho = sd(anthocyanin_level), # Calculate the standard deviation of
                                              # anthocyanin levels for each genotype
            n_antho = n(), # Calculate the number of measurements of anthocyanin levels
                           # for each genotype
            SE_antho = sd_antho / sqrt(n()), # Calculate the standard error of the mean
                                             # of anthocyanin levels for each genotype
            max_antho = max(anthocyanin_level)) %>% # Define the maximum of anthocyanin
                                                    # levels for each genotype
  # Add a column with the letters indicating statistical significant different groups
  mutate(diff = print(tuk_cld$mcletters$Letters))
```

```
##    bbx202122 bbx202122hy5          hy5           WT
##          "c"          "bc"          "b"          "a"
```

```
# Display the summary statistics
print(antho_summary)
```

```
## # A tibble: 4 x 7
##   Genotype     mean_antho sd_antho n_antho SE_antho max_antho diff
##   <fct>             <dbl>    <dbl>   <int>    <dbl>     <dbl> <chr>
## 1 bbx202122         0.434   0.188        5   0.084       0.69 c
## 2 bbx202122hy5      0.626   0.0780       5   0.0349      0.7  bc
## 3 hy5               0.78    0.120        5   0.0536      0.91 b
## 4 WT                1.91    0.481        5   0.215       2.67 a
```

Create a plot and save it as a pdf

```
# Save the graph as a pdf
pdf("anthocyanin_measurement.pdf", width = 4 , height = 6)
# Create a plot with the mean of the anthocyanin level vs. Genotype
```

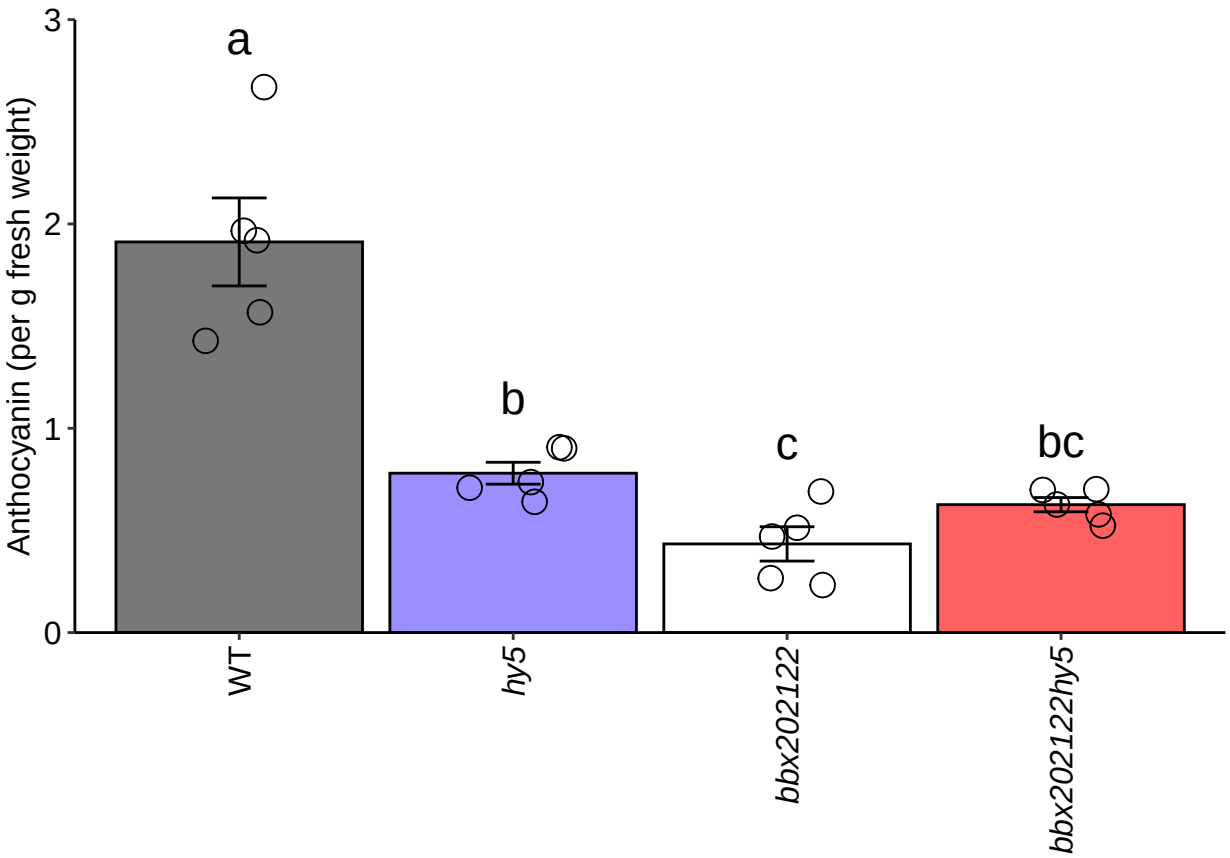
```
antho_plot = ggplot(data = antho_summary, aes(x = Genotype, y = mean_antho)) +
  # Create a bar plot and customize its appearance
  geom_bar(stat = "identity", # Create the bars
           fill = c("#FFFFFF", "#FF6161",
                    "#9B8EFF", "#787878"), # assign individual fill colors to the bars
           color = "black") +
  geom_errorbar(aes(ymin = mean_antho - SE_antho, ymax = mean_antho + SE_antho),
                width = 0.2) + # add error bars
  # Change order of the categories
  scale_x_discrete(limits = c("WT", "hy5", "bbx202122", "bbx202122hy5")) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90,
                                   size = 12,
                                   hjust = 1,
                                   vjust = 0.5,
                                   face = c("plain", "italic", "italic", "italic"),
                                   color = "black"),
        axis.title.y = element_text(size = 12,
                                    color = "black"),
        axis.text.y = element_text(size = 12,
                                   color = "black")) +
  theme(axis.ticks = element_line()) +
  ylab("Anthocyanin (per g fresh weight)") +
  xlab(NULL) +
  scale_y_continuous(breaks = c(0, 1, 2, 3),
                     expand = c(0,0),
                     limits = c(0,3)) +
  geom_text(data = antho_summary,
            aes(y = max_antho, label = diff),
            vjust = -1,
            size = 6,
            color = "black") +
  geom_jitter(data = antho,
              aes(y = anthocyanin_level),
              size = 4,
              color = "black",
              shape = 1,
              width = 0.2)
```

```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```

```
antho_plot
dev.off()
```

```
## pdf
##   2
```

```
# Display the graph
print(antho_plot)
```


Anthocyanin (per g fresh weight)
3
2
1
0
a
b
c
bc
WT
hy5
bbx202122
bbx202122hy5