Statistics fact sheet

Chun-Hui Gao

2019-05-21

Statistics were performed maily with the stats and rstatix package in R version 3.5.3. First of all, we loaded the packages and the tidyverse package.

library(tidyverse)  
library(stats)  
library(rstatix)  
  
# order of NPs  
np\_levels <- c("CK","Fe2O3", "ZnO", "CeO2", "Fe3O4", "Al2O3", "CuO", "TiO2")

# Figure 3: Comparison of Ct values

This part is for the comparision of Ct values. Ct values for each RT-PCR reaction was stored in “RT-PCR\_results.csv”. The experiment has repeat four times, so each group has 4 Ct values.

The following code block read in the result.

results <- read\_csv("data/RT-PCR\_results.csv")  
results$nps <- factor(results$nps, levels = np\_levels)

Analysis then can be performed for Ex-taq or Phusion separately.

## Fig. 3A Ex-taq.

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with “\*\*” in the model summary.

data <- filter(results,enzyme=="extaq")  
aov <- aov(CT~nps,data=data)  
summary(aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## nps 4 9.860 2.4650 6.607 0.00284 \*\*  
## Residuals 15 5.596 0.3731   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As the ANOVA test is significant, we can compute Tukey HSD (**Tukey Honest Significant Differences**, R function: TukeyHSD()) for performing multiple pairwise-comparison between the means of groups.

TukeyHSD(aov)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = CT ~ nps, data = data)  
##   
## $nps  
## diff lwr upr p adj  
## Fe2O3-CK 0.12075 -1.2129075 1.45440754 0.9985015  
## Fe3O4-CK -0.00550 -1.3391575 1.32815754 1.0000000  
## Al2O3-CK -1.35275 -2.6864075 -0.01909246 0.0460398  
## CuO-CK 0.81050 -0.5231575 2.14415754 0.3699970  
## Fe3O4-Fe2O3 -0.12625 -1.4599075 1.20740754 0.9982169  
## Al2O3-Fe2O3 -1.47350 -2.8071575 -0.13984246 0.0271207  
## CuO-Fe2O3 0.68975 -0.6439075 2.02340754 0.5210514  
## Al2O3-Fe3O4 -1.34725 -2.6809075 -0.01359246 0.0471491  
## CuO-Fe3O4 0.81600 -0.5176575 2.14965754 0.3637444  
## CuO-Al2O3 2.16325 0.8295925 3.49690754 0.0012519

It can be seen from the output, that the difference between CK and Al2O3 is significant with an adjusted p-value of 0.046.

The ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. It’s possible to use **Levene’s test** to check the homogeneity of variances. The function leveneTest() [in car package] will be used:

car::leveneTest(CT~nps,data=data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 1.0645 0.408  
## 15

From the output above we can see that the p-value is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different NP groups.

Normality assumption can be checked by the **Shapiro-Wilk test** on the ANOVA residuals.

shapiro.test(x=residuals(aov))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(aov)  
## W = 0.96983, p-value = 0.7514

The output above finds no indication that normality is violated (W = 0.97, p = 0.75). We can assume normality.

## Fig. 3B Phusion

Normality assumption can be checked by the **Shapiro-Wilk test** on the ANOVA residuals as well.

data <- filter(results,enzyme=="phusion")  
aov <- aov(CT~nps,data=data)  
shapiro.test(x=residuals(aov))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(aov)  
## W = 0.81303, p-value = 0.00136

The output above finds that normality is violated (W = 0.81, p = 0.001). We can not assume normality.

Therefore, we have to use **Kruskal-Wallis rank sum test** for this data, as this test can be used when ANOVA assumptions are not met.

kruskal.test(CT~nps,data=data)

##   
## Kruskal-Wallis rank sum test  
##   
## data: CT by nps  
## Kruskal-Wallis chi-squared = 16.614, df = 4, p-value = 0.002297

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between differnt NP treatments.

From the output of the **Kruskal-Wallis test**, we know that there is a significant difference between groups, but we don’t know which pairs of groups are different.

It’s possible to use the function pairwise.wilcox.test() to calculate pairwise comparisons between group levels with corrections for multiple testing.

pairwise.wilcox.test(data$CT,data$nps,p.adjust.method = "BH")

##   
## Pairwise comparisons using Wilcoxon rank sum test   
##   
## data: data$CT and data$nps   
##   
## CK Fe2O3 ZnO Fe3O4  
## Fe2O3 0.036 - - -   
## ZnO 0.036 0.036 - -   
## Fe3O4 0.036 0.036 0.222 -   
## CuO 0.036 0.036 0.036 0.686  
##   
## P value adjustment method: BH

The pairwise comparison shows that, CK and four NPs conditions are all significant different (p < 0.05).

# Figure 4: Comparision of overall error rates

# error rate and snp frequency were estimated using amplicon sequencing and mothur software  
error\_rate <- read\_csv("data/error\_rate.csv")

## Warning: Missing column names filled in: 'X1' [1]

# read meta data of sequencing library  
meta <- read\_tsv("data/meta.txt")  
meta$nps <- factor(meta$nps, levels = np\_levels)  
  
# join data  
error\_rate <- left\_join(error\_rate,meta)

First of all, we compared the error rate of normal PCR (CK+) between Ex taq and Phusion.

data <- filter(error\_rate, nps=="CK")  
aov <- aov(error\_rate~enzyme,data)  
summary(aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## enzyme 1 5.388e-07 5.388e-07 93.81 0.000636 \*\*\*  
## Residuals 4 2.300e-08 5.700e-09   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = error\_rate ~ enzyme, data = data)  
##   
## $enzyme  
## diff lwr upr p adj  
## phusion-extaq -0.0005993083 -0.0007711094 -0.0004275072 0.0006384

# check normality  
car::leveneTest(error\_rate~enzyme,data)

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to  
## factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.6171 0.4761  
## 4

shapiro.test(x=residuals(aov))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(aov)  
## W = 0.88689, p-value = 0.3022

From the above results we can find the error rate are significant between enzymes in normal PCR.

The analysis is similar to that has been mentioned on the above.

## Fig. 4A Ex-taq

data <- error\_rate %>% filter(enzyme=="extaq")  
aov <- aov(error\_rate~nps,data)  
summary(aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## nps 4 5.923e-08 1.481e-08 2.521 0.107  
## Residuals 10 5.873e-08 5.873e-09

As the p-value is greater than the significance level 0.05, we can conclude that there is no significant differences of error rate between differnt NP treatments.

car::leveneTest(error\_rate~nps,data=data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.4847 0.747  
## 10

shapiro.test(x=residuals(aov))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(aov)  
## W = 0.91966, p-value = 0.1904

The output above finds ANOVA assumptions is valid, therefore, the test result is confidential.

## Fig. 4B Phusion

data <- error\_rate %>% filter(enzyme=="phusion")  
aov <- aov(error\_rate~nps,data)  
summary(aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## nps 4 4.424e-09 1.106e-09 0.793 0.556  
## Residuals 10 1.395e-08 1.395e-09

As the p-value is greater than the significance level 0.05, we can conclude that there is no significant differences of error rate between differnt NP treatments.

car::leveneTest(error\_rate~nps,data=data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.2578 0.8984  
## 10

shapiro.test(x=residuals(aov))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(aov)  
## W = 0.95771, p-value = 0.6526

The output above finds ANOVA assumptions is validty, therefore, the test result is confidential.

# Figure 5: Comparision of SNV frequencies

snp\_freq <- read\_csv("data/snp\_freq.csv")  
snp\_freq <- left\_join(snp\_freq,meta)

## Fig. 5A Ex-taq

data <- snp\_freq %>% filter(enzyme=="extaq")  
data %>% group\_by(snv) %>%   
 anova\_test(freq~nps) %>%  
 mutate(p.adj=p.adjust(p,method = "BH"))

## # A tibble: 12 x 7  
## snv .y. term statistic p method p.adj  
## <chr> <chr> <chr> <dbl> <dbl> <chr> <dbl>  
## 1 AC freq nps 0.639 0.65 Anova 0.709  
## 2 AG freq nps 6.26 0.0086 Anova 0.103  
## 3 AT freq nps 3.01 0.072 Anova 0.264  
## 4 CA freq nps 1.92 0.18 Anova 0.33   
## 5 CG freq nps 2.50 0.11 Anova 0.264  
## 6 CT freq nps 1.72 0.22 Anova 0.33   
## 7 GA freq nps 1.81 0.2 Anova 0.33   
## 8 GC freq nps 1.06 0.43 Anova 0.516  
## 9 GT freq nps 1.04 0.43 Anova 0.516  
## 10 TA freq nps 2.56 0.1 Anova 0.264  
## 11 TC freq nps 3.80 0.04 Anova 0.24   
## 12 TG freq nps 0.309 0.87 Anova 0.87

Since multiple comparisions were conducted to different groups, we need to adjust the p-values. Here, the **Benjamini & Hochberg (“BH”) adjustment method** was applied.

As the p.adj is all greater than the significance level 0.05, we can conclude that there is no significant differences of SNV frequency between different NP treatments.

## Fig. 5B Phusion

data <- snp\_freq %>% filter(enzyme=="phusion")  
data %>% group\_by(snv) %>%   
 anova\_test(freq~nps) %>%  
 mutate(p.adj=p.adjust(p,method = "BH"))

## # A tibble: 12 x 7  
## snv .y. term statistic p method p.adj  
## <chr> <chr> <chr> <dbl> <dbl> <chr> <dbl>  
## 1 AC freq nps 0.0726 0.99 Anova 0.99   
## 2 AG freq nps 1.65 0.24 Anova 0.945  
## 3 AT freq nps 0.758 0.580 Anova 0.945  
## 4 CA freq nps 0.515 0.73 Anova 0.973  
## 5 CG freq nps 1.00 0.45 Anova 0.945  
## 6 CT freq nps 1.29 0.34 Anova 0.945  
## 7 GA freq nps 0.753 0.580 Anova 0.945  
## 8 GC freq nps 0.997 0.45 Anova 0.945  
## 9 GT freq nps 0.667 0.63 Anova 0.945  
## 10 TA freq nps 0.325 0.86 Anova 0.99   
## 11 TC freq nps 0.673 0.63 Anova 0.945  
## 12 TG freq nps 0.161 0.95 Anova 0.99

As the p.adj is greater than the significance level 0.05, we can conclude that there is no significant differences of SNV frequency between differnt NP treatments.

# Session Info

## R version 3.5.3 (2019-03-11)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 17134)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] rstatix\_0.1.0 forcats\_0.4.0 stringr\_1.4.0 dplyr\_0.8.0.1   
## [5] purrr\_0.3.2 readr\_1.3.1 tidyr\_0.8.3 tibble\_2.1.1   
## [9] ggplot2\_3.1.1 tidyverse\_1.2.1  
##   
## loaded via a namespace (and not attached):  
## [1] tidyselect\_0.2.5 xfun\_0.6 haven\_2.1.0   
## [4] lattice\_0.20-38 carData\_3.0-2 colorspace\_1.4-1   
## [7] generics\_0.0.2 htmltools\_0.3.6 yaml\_2.2.0   
## [10] utf8\_1.1.4 rlang\_0.3.4 pillar\_1.3.1   
## [13] foreign\_0.8-71 glue\_1.3.1 withr\_2.1.2   
## [16] modelr\_0.1.4 readxl\_1.3.1 plyr\_1.8.4   
## [19] munsell\_0.5.0 gtable\_0.3.0 cellranger\_1.1.0   
## [22] zip\_2.0.1 rvest\_0.3.3 evaluate\_0.13   
## [25] knitr\_1.22 rio\_0.5.16 curl\_3.3   
## [28] fansi\_0.4.0 broom\_0.5.2 Rcpp\_1.0.1   
## [31] scales\_1.0.0 backports\_1.1.4 jsonlite\_1.6   
## [34] abind\_1.4-5 hms\_0.4.2 digest\_0.6.18   
## [37] openxlsx\_4.1.0 stringi\_1.4.3 grid\_3.5.3   
## [40] cli\_1.1.0 tools\_3.5.3 magrittr\_1.5   
## [43] lazyeval\_0.2.2 crayon\_1.3.4 car\_3.0-2   
## [46] pkgconfig\_2.0.2 data.table\_1.12.2 xml2\_1.2.0   
## [49] lubridate\_1.7.4 assertthat\_0.2.1 rmarkdown\_1.12   
## [52] httr\_1.4.0 rstudioapi\_0.10 R6\_2.4.0   
## [55] nlme\_3.1-137 compiler\_3.5.3