GLM: mixed effects logisitc regression.

Jacob Westaway

Last updated on 2021-03-25

Load packages.

```
sapply(c("lme4", "aods3", "tidyverse", "ggplot2", "MuMIn", "gridExtra",
         "car", "emmeans", "ggpubr", "DESeq2", "phyloseq"),
         require, character.only = TRUE)
```

Use DESeq2 to determine what taxa to include in the model.

Define function for calculating geometric means.

```
calc_geo_means <- function(deseq_object){
# geometric mean
  gm_mean = function(x, na.rm = TRUE){
    exp(sum(log(x[x > 0]), na.rm = na.rm) / length(x))
  }
  geoMeans <- apply(counts(deseq_object), 1, gm_mean)
# size factors
  estimateSizeFactors(deseq_object, geoMeans = geoMeans)
}
```

Define function to subset out taxa with small counts and low occurance (count of at least 10 in 60 or more samples).

```
deseq_filter <- function(deseq_object){
  nc <- counts(deseq_object, normalized = TRUE)
  filtered <- rowSums(nc >= 10) >= 60 # filter = abundance of 10 in 60 samples.
  deseq_object[filtered,]
}
```

Define function to extract significant results from a DESeq2 LRT test

```
get_deseq_res_lrt <- function(deseq_object){
  res = results(deseq_object)
  res = res[order(res$padj, na.last = NA), ]
  sigtab = res[(res$padj < 0.01), ]
  sigtab = cbind(as(sigtab, "data.frame"),
         as(tax_table(ps3.Microbiome)[rownames(sigtab), ], "matrix"))
  sigtab %>%
  arrange(padj) %>%
  select("log2FoldChange", "lfcSE", "padj", "Genus")
}
```

Run the DESeq2 analysis

```
phyloseq_to_deseq2(ps3.Microbiome, ~ parasite_burden) %>%
    calc_geo_means() %>%
    deseq_filter() %>%
    DESeq(fitType = "local", test = "LRT", reduced = ~ 1) %>%
    get_deseq_res_lrt() %>%
    remove_rownames()
```

```
##   log2FoldChange     lfcSE         padj             Genus
## 1      2.4128359 0.4344837 0.0001295705 Coraliomargarita
## 2     -1.4205522 0.4161264 0.0044270235       Arcobacter
## 3     -0.8046972 0.2452338 0.0047818441 NS4_marine_group
## 4     -2.2217426 0.7173547 0.0047818441     Salinirepens
## 5     -2.3186334 0.7394662 0.0047818441       Marivivens
```

Create a new dataframe that includes the transformed abundances of the significant genera above.

```
metadata_with_taxa <- subset_samples(ps3, Type == "Microbiome") %>%
  sample_data() %>%
  unclass() %>%
  as.data.frame() %>%
  mutate(ID = paste0("AM", ID)) %>%
  left_join(
    (phyloseq_to_deseq2(ps3.Microbiome, ~ parasite_burden) %>%
    calc_geo_means() %>%
    counts(normalized = TRUE) %>%
    as.data.frame() %>%
    filter(rownames(.) == "TTTCGAATCATTCACAATGGGGAAACCCTGATGGTGCAACGCCGCGTGGGGGATGAAGGCCTTCGGGTTGTAAAC
             rownames(.) ==  "TGAGGAATATTGGACAATGGACGAAAGTCTGATCCAGCCATGCCGCGTGCAGGATGACGGCCCTATGGGTTGT
             rownames(.) == "TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCGCGTGGAGGATGACACATTTCGGTGCGTAA
             rownames(.) == "TGGGGAATCTTAGACAATGGGGGAAACCCTGATCTAGCCATGCCGCGTGAGTGACGAAGGCCTTAGGGTCGTAA
             rownames(.) == "TGAGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGCAGGAAGAATGCCCTATGGGTTGTA
  base::t() %>%
  as.data.frame() %>%
  dplyr::rename("Coraliomargarita" = "TTTCGAATCATTCACAATGGGGAAACCCTGATGGTGCAACGCCGCGTGGGGGATGAAGGCCTTC
         "NS4_marine_group" = "TGAGGAATATTGGACAATGGACGAAAGTCTGATCCAGCCATGCCGCGTGCAGGATGACGGCCCTATGGGTTG
         "Arcobacter" = "TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCGCGTGGAGGATGACACATTTCGGTGCGTAAACTC
         "Marivivens" = "TGGGGAATCTTAGACAATGGGGGAAACCCTGATCTAGCCATGCCGCGTGAGTGACGAAGGCCTTAGGGTCGTAAAGCT
         "Salinirepens" = "TGAGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGCAGGAAGAATGCCCTATGGGTTGTAAA
  rownames_to_column(var = "ID")), by = "ID")
```

Centre and scale numeric variables using defined function.

```r
centre_and_scale <- function(data1){
# get numeric variables
data2 <- data1 %>%
  select_if(is.numeric)
# centering and scaling over variables
data3 <- sapply(data2, function(x) scale(x, center=T, scale = 2*sd(x))) %>%
  as.data.frame() %>%
  rownames_to_column("RowID")
# join scaled/centred data to non-numeric data
data1 %>%
  select_if(negate(is.numeric)) %>%
  rownames_to_column("RowID") %>%
  left_join(data3, by = "RowID") %>%
  select(-RowID)
}

glm_data <- metadata_with_taxa %>% centre_and_scale()
```

Test for multicollinearity: define the `corvif()`function that takes metadata and creates a linear model to see if any collinearity exists between variables.

```r
# myvif
myvif <- function(mod) {
  v <- vcov(mod)
  assign <- attributes(model.matrix(mod))$assign
  if (names(coefficients(mod)[1]) == "(Intercept)") {
    v <- v[-1, -1]
    assign <- assign[-1]
  } else warning("No intercept: vifs may not be sensible.")
  terms <- labels(terms(mod))
  n.terms <- length(terms)
  if (n.terms < 2) stop("The model contains fewer than 2 terms")
  if (length(assign) > dim(v)[1] ) {
    diag(tmp_cor)<-0
    if (any(tmp_cor==1.0)){
      return("Sample size is too small, 100% collinearity is present")
    } else {
      return("Sample size is too small")
    }
  }
  R <- cov2cor(v)
  detR <- det(R)
  result <- matrix(0, n.terms, 3)
  rownames(result) <- terms
  colnames(result) <- c("GVIF", "Df", "GVIF^(1/2Df)")
  for (term in 1:n.terms) {
    subs <- which(assign == term)
    result[term, 1] <- det(as.matrix(R[subs, subs])) * det(as.matrix(R[-subs, -subs])) / detR
    result[term, 2] <- length(subs)
  }
  if (all(result[, 2] == 1)) {
    result <- data.frame(GVIF=result[, 1])
  } else {
    result[, 3] <- result[, 1]^(1/(2 * result[, 2]))
  }
  invisible(result)
}

# corvif
corvif <- function(data) {
  data <- as.data.frame(data)

  form    <- formula(paste("fooy ~ ",paste(strsplit(names(data)," "),collapse = " + ")))
  data  <- data.frame(fooy = 1 + rnorm(nrow(data)) ,data)
  lm_mod  <- lm(form,data) # runs linear model with above formula and metadata

  cat("\n\nVariance inflation factors\n\n")
  print(myvif(lm_mod))
}
```

```
# full
glm_data %>%
  select(Temperature_C, pH,
         RDO_Conc_mgL, RainGauge_mm, Salinity_PSU,
         NS4_marine_group, Salinirepens, Marivivens,
         Coraliomargarita, Arcobacter) %>%
  corvif()
```

```
##
##
## Variance inflation factors
##
##                       GVIF
## Temperature_C     5.367213
## pH               11.337891
## RDO_Conc_mgL      3.764943
## RainGauge_mm     14.346282
## Salinity_PSU      1.552170
## NS4_marine_group  1.423795
## Salinirepens     60.342632
## Marivivens       86.567244
## Coraliomargarita  2.014515
## Arcobacter       11.919902
```

```
# final model
glm_data %>%
  select(RDO_Conc_mgL, RainGauge_mm,
         Salinity_PSU, NS4_marine_group,
         Coraliomargarita, Arcobacter) %>%
  corvif()
```

```
##
```

```
##
## Variance inflation factors
##
##                      GVIF
## RDO_Conc_mgL     2.129486
## RainGauge_mm     2.014933
## Salinity_PSU     1.280985
## NS4_marine_group 1.115909
## Coraliomargarita 1.211361
## Arcobacter       1.289042
```

Fit Model.

```
global <- lme4::glmer(parasite_burden ~ RDO_Conc_mgL +
                      RainGauge_mm + Salinity_PSU + NS4_marine_group +
                      Coraliomargarita + Arcobacter + (1|Date),
                      data = glm_data, family = "binomial")
summary(global)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: parasite_burden ~ RDO_Conc_mgL + RainGauge_mm + Salinity_PSU +      NS4_marine_group + Coral
##    Data: glm_data
##
##      AIC      BIC   logLik deviance df.resid
##     96.1    118.7    -40.0     80.1      117
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7228  0.0014  0.1103  0.3675  1.2989
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Date   (Intercept) 1.11     1.053
## Number of obs: 125, groups:  Date, 25
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.3156     0.8294   3.998  6.4e-05 ***
## RDO_Conc_mgL       0.6852     1.2019   0.570  0.56864
## RainGauge_mm       3.4972     1.7195   2.034  0.04196 *
## Salinity_PSU       1.0268     1.3898   0.739  0.46002
## NS4_marine_group  -2.7317     1.2540  -2.178  0.02938 *
## Coraliomargarita   9.0863     3.5003   2.596  0.00944 **
## Arcobacter        -1.4981     0.8915  -1.680  0.09290 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) RDO_C_ RnGg_m Sl_PSU NS4_m_ Crlmrg
## RDO_Cnc_mgL   0.071
## RainGaug_mm   0.451  0.368
## Salinty_PSU  0.235  0.275  0.709
## NS4_mrn_grp -0.305 -0.114 -0.033  0.104
## Coralimrgrt  0.757  0.055  0.166  0.008 -0.453
## Arcobacter  -0.243  0.432 -0.361 -0.230  0.073  0.022
```

Goodness of fit and R2.

```
gof(global)
```

```
## D  = 64.4143, df = 117, P(>D) = 0.99998
## X2 = 66.1975, df = 117, P(>X2) = 0.9999582
```

```
r.squaredGLMM(global)
```

```
##                   R2m       R2c
## theoretical 0.8219462 0.8668561
## delta       0.7933804 0.8367295
```

Backwards selection.

```
dfun(drop1(global))
```

```
## Single term deletions
##
## Model:
## parasite_burden ~ RDO_Conc_mgL + RainGauge_mm + Salinity_PSU +
##     NS4_marine_group + Coraliomargarita + Arcobacter + (1 | Date)
##                  npar    dAIC
## <none>                 1.6820
## RDO_Conc_mgL        1  0.0000
## RainGauge_mm        1  5.9821
## Salinity_PSU        1  0.2820
## NS4_marine_group    1  6.0044
## Coraliomargarita    1 12.1818
## Arcobacter          1  3.1062
```

```
global2 <- lme4::glmer(parasite_burden ~
                      RainGauge_mm + NS4_marine_group +
                      Coraliomargarita + Arcobacter + (1|Date),
                      data = glm_data, family = "binomial")

dfun(drop1(global2))
```

```
## Single term deletions
##
## Model:
## parasite_burden ~ RainGauge_mm + NS4_marine_group + Coraliomargarita +
##     Arcobacter + (1 | Date)
##                    npar    dAIC
## <none>                   0.0000
## RainGauge_mm        1  4.3061
## NS4_marine_group    1  4.9300
## Coraliomargarita    1 10.4848
## Arcobacter          1  3.1587
```

```r
summary(global2) # final model
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [’glmerMod’]
##  Family: binomial  ( logit )
## Formula: parasite_burden ~ RainGauge_mm + NS4_marine_group + Coraliomargarita +      Arcobacter + (1
##    Data: glm_data
##
##      AIC      BIC   logLik deviance df.resid
##     92.8    109.8    -40.4     80.8      119
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6948  0.0013  0.1258  0.3659  1.3029
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Date   (Intercept) 1.203    1.097
## Number of obs: 125, groups:  Date, 25
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.2829     0.8172   4.017 5.89e-05 ***
## RainGauge_mm       2.6840     1.1690   2.296  0.02167 *
## NS4_marine_group  -2.8162     1.2250  -2.299  0.02151 *
## Coraliomargarita   9.1247     3.4737   2.627  0.00862 **
## Arcobacter        -1.6202     0.7425  -2.182  0.02910 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##             (Intr) RnGg_m NS4_m_ Crlmrg
## RainGaug_mm  0.440
## NS4_mrn_grp -0.344 -0.116
## Coralimrgrt  0.778  0.253 -0.453
## Arcobacter  -0.247 -0.501  0.211 -0.026
```

Goodness of fit and R2.

```r
gof(global2)
```

```
## D  = 64.2268, df = 119, P(>D) = 0.9999902
## X2 = 65.7842, df = 119, P(>X2) = 0.9999808
```

```
r.squaredGLMM(global2)
```

```
##                   R2m       R2c
## theoretical 0.8184623 0.8670722
## delta       0.7900621 0.8369853
```

Reintgeration to calculate estimates, standard error and p values for each of the variables removed during backwards selection.

```
lme4::glmer(parasite_burden ~ RDO_Conc_mgL +
              RainGauge_mm + NS4_marine_group +
              Coraliomargarita + Arcobacter + (1|Date),
            data = glm_data, family = "binomial") %>%
  summary()
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: parasite_burden ~ RDO_Conc_mgL + RainGauge_mm + NS4_marine_group +
##     Coraliomargarita + Arcobacter + (1 | Date)
##    Data: glm_data
##
##      AIC      BIC   logLik deviance df.resid
##     94.7    114.5    -40.3     80.7      118
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0485  0.0012  0.1237  0.3640  1.3081
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Date   (Intercept) 1.128    1.062
## Number of obs: 125, groups:  Date, 25
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.2739     0.8104   4.040 5.34e-05 ***
## RDO_Conc_mgL       0.4161     1.1494   0.362   0.7173
## RainGauge_mm       2.7956     1.2051   2.320   0.0204 *
## NS4_marine_group  -2.8762     1.2462  -2.308   0.0210 *
## Coraliomargarita   9.1835     3.4949   2.628   0.0086 **
## Arcobacter        -1.4497     0.8544  -1.697   0.0898 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) RDO_C_ RnGg_m NS4_m_ Crlmrg
## RDO_Cnc_mgL  0.014
## RainGaug_mm  0.420  0.294
## NS4_mrn_grp -0.337 -0.149 -0.140
## Coralimrgrt  0.780  0.054  0.247 -0.459
## Arcobacter  -0.201  0.519 -0.257  0.095  0.013
```

```r
lme4::glmer(parasite_burden ~ Salinity_PSU +
              RainGauge_mm + NS4_marine_group +
              Coraliomargarita + Arcobacter + (1|Date),
            data = glm_data, family = "binomial") %>%
  summary()
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: parasite_burden ~ Salinity_PSU + RainGauge_mm + NS4_marine_group +
##     Coraliomargarita + Arcobacter + (1 | Date)
##    Data: glm_data
##
##      AIC      BIC   logLik deviance df.resid
##     94.4    114.2    -40.2     80.4      118
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3196  0.0016  0.1065  0.3622  1.2929
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Date   (Intercept) 1.221    1.105
## Number of obs: 125, groups:  Date, 25
##
## Fixed effects:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.3167     0.8339   3.977 6.97e-05 ***
## Salinity_PSU         0.8309     1.3520   0.615  0.53885
## RainGauge_mm         3.1913     1.5967   1.999  0.04564 *
## NS4_marine_group    -2.6655     1.2345  -2.159  0.03084 *
## Coraliomargarita     8.9981     3.4588   2.601  0.00928 **
## Arcobacter          -1.7468     0.8149  -2.144  0.03206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) Sl_PSU RnGg_m NS4_m_ Crlmrg
## Salinty_PSU   0.205
## RainGaug_mm   0.455  0.662
## NS4_mrn_grp  -0.309  0.147 -0.003
## Coralimrgrt   0.754 -0.017  0.169 -0.446
## Arcobacter   -0.303 -0.381 -0.608  0.147 -0.012
```

Visualisation.

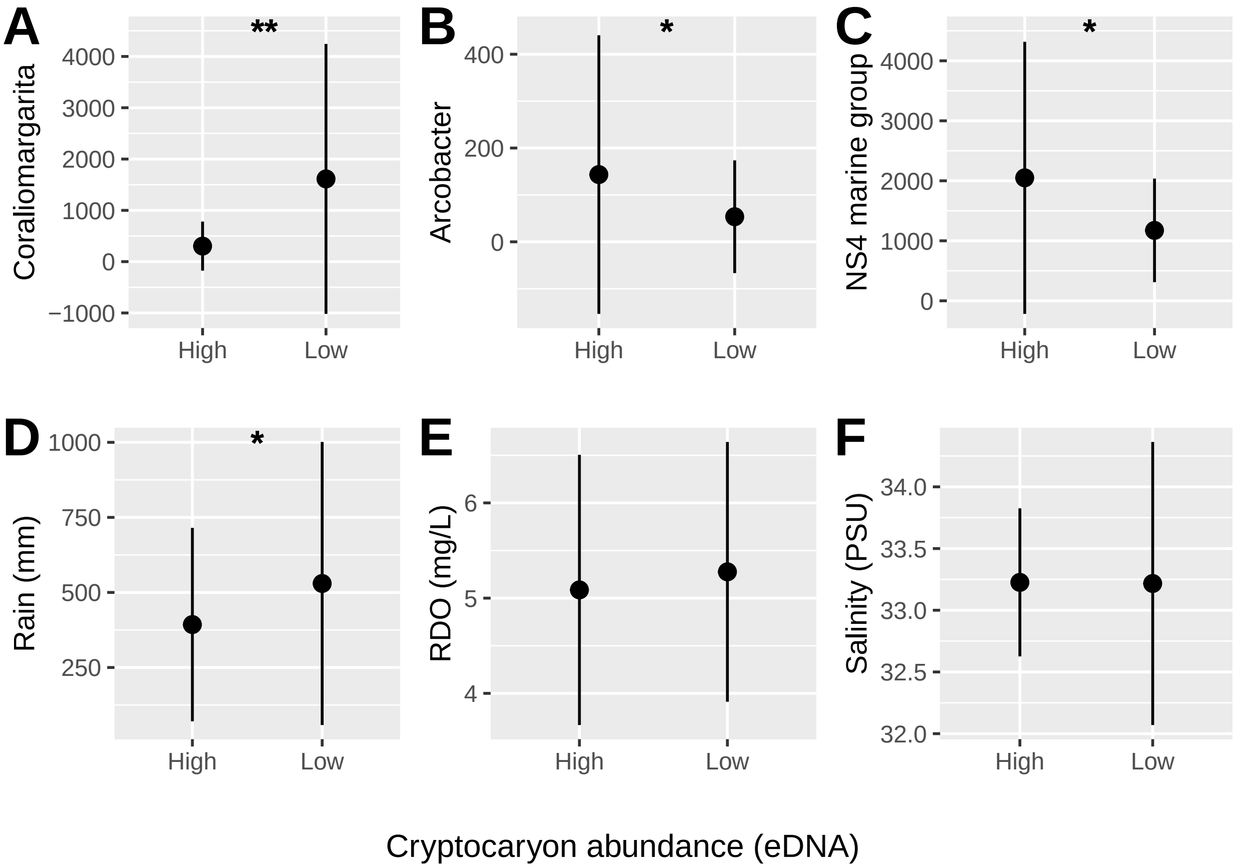
Calculate the mean and standard deviation for each variable.

```r
plot_data <- metadata_with_taxa %>%
  group_by(parasite_burden) %>%
  select_if(is.numeric) %>%
  summarise_all(mean) %>%
  left_join(
    metadata_with_taxa %>%
      group_by(parasite_burden) %>%
      select_if(is.numeric) %>%
      summarise_all(sd) %>%
      rename_at(2:ncol(.), toupper))
```

Generate plot.

global <- lme4::glmer(parasite_burden ~ RDO_Conc_mgL + RainGauge_mm + Salinity_PSU + NS4_marine_group + Coraliomargarita + Arcobacter + (1|Date), data = glm_data, family = “binomial”)

```
grid.arrange(
(ggplot(plot_data, aes(x = parasite_burden, y = Coraliomargarita)) +
geom_pointrange(aes(ymin = Coraliomargarita - CORALIOMARGARITA,
                    ymax = Coraliomargarita + CORALIOMARGARITA)) +
  xlab("") +
  ylab("Coraliomargarita") +
  geom_text(aes(label = "**", y = 4500, x = 1.5, fontface = "bold", size = 20)) +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "A", fig.lab.face = "bold", fig.lab.size = 20),

(ggplot(plot_data, aes(x = parasite_burden, y = Arcobacter)) +
geom_pointrange(aes(ymin = Arcobacter - ARCOBACTER,
                    ymax = Arcobacter + ARCOBACTER)) +
  xlab("") +
  ylab("Arcobacter") +
  geom_text(aes(label = "*", y = 450, x = 1.5, fontface = "bold", size = 20)) +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "B", fig.lab.face = "bold", fig.lab.size = 20),

(ggplot(plot_data, aes(x = parasite_burden, y = NS4_marine_group)) +
geom_pointrange(aes(ymin = NS4_marine_group - NS4_MARINE_GROUP,
                    ymax = NS4_marine_group + NS4_MARINE_GROUP)) +
  xlab("") +
  ylab("NS4 marine group") +
  geom_text(aes(label = "*", y = 4500, x = 1.5, fontface = "bold", size = 20)) +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "C", fig.lab.face = "bold", fig.lab.size = 20),

(ggplot(plot_data, aes(x = parasite_burden, y = RainGauge_mm)) +
geom_pointrange(aes(ymin = RainGauge_mm - RAINGAUGE_MM,
                    ymax = RainGauge_mm + RAINGAUGE_MM)) +
  xlab("") +
  ylab("Rain (mm)") +
  geom_text(aes(label = "*", y = 1000, x = 1.5, fontface = "bold", size = 20)) +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "D", fig.lab.face = "bold", fig.lab.size = 20),

(ggplot(plot_data, aes(x = parasite_burden, y = RDO_Conc_mgL)) +
geom_pointrange(aes(ymin = RDO_Conc_mgL - RDO_CONC_MGL,
                    ymax = RDO_Conc_mgL + RDO_CONC_MGL)) +
  xlab("") +
  ylab("RDO (mg/L)") +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "E", fig.lab.face = "bold", fig.lab.size = 20),

(ggplot(plot_data, aes(x = parasite_burden, y = Salinity_PSU)) +
geom_pointrange(aes(ymin = Salinity_PSU - SALINITY_PSU,
                    ymax = Salinity_PSU + SALINITY_PSU)) +
  xlab("") +
  ylab("Salinity (PSU)") +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "F", fig.lab.face = "bold", fig.lab.size = 20),
ncol = 3, bottom = textGrob("Cryptocaryon abundance (eDNA)", gp = gpar(fontsize = 12))
)
```



Finished.