GLM: mixed effects logisitc regression.

Jacob Westaway

Last updated on 2021-04-17

Load packages.

```
sapply(c("lme4", "aods3", "tidyverse", "ggplot2", "MuMIn", "gridExtra", "effects",
         "car", "emmeans", "ggpubr", "DESeq2", "phyloseq", "sjPlot", "grid"),
         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(.) == "TTTCGAATCATTCACAATGGGGGAAACCCTGATGGTGCAACGCCGCGTGGGGGATGAAGGCCTTCGGGTTGTAAACC
             rownames(.) ==  "TGAGGAATATTGGACAATGGACGAAAGTCTGATCCAGCCATGCCGCGTGCAGGATGACGGCCCTATGGGTTGTA
             rownames(.) == "TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCGCGTGGAGGATGACACATTTCGGTGCGTAAA
             rownames(.) == "TGGGGAATCTTAGACAATGGGGGAAACCCTGATCTAGCCATGCCGCGTGAGTGACGAAGGCCTTAGGGTCGTAAA
             rownames(.) == "TGAGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGCAGGAAGAATGCCCTATGGGTTGTAA
  base::t() %>%
  as.data.frame() %>%
  dplyr::rename("Coraliomargarita" = "TTTCGAATCATTCACAATGGGGGAAACCCTGATGGTGCAACGCCGCGTGGGGGATGAAGGCCTTC
         "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.

```
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()
```

Explore relationship between parasite and variables

```
metadata_with_taxa %>%
  ggplot(aes(x = ddPCR, y = Rainfall)) +
  geom_point() +
  scale_x_log10() +
  geom_smooth(method = "lm", se = T) +
  scale_y_log10()

metadata_with_taxa %>%
  ggplot(aes(x = ddPCR, y = RDO_Conc_mgL)) +
  geom_point() +
  geom_smooth(method = "lm", se = T)
```

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

```
# 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, Rainfall, Salinity_PSU,
         NS4_marine_group, Salinirepens, Marivivens,
         Coraliomargarita, Arcobacter) %>%
  corvif()
```

```
##
##
## Variance inflation factors
##
##                      GVIF
## Temperature_C     3.865617
## pH                8.620872
## RDO_Conc_mgL      8.375105
```

```
## Rainfall          2.595200
## Salinity_PSU      1.205985
## NS4_marine_group  1.220025
## Salinirepens     57.220675
## Marivivens       82.939413
## Coraliomargarita  2.334099
## Arcobacter       12.115178
```

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

```
##
##
## Variance inflation factors
##
##                       GVIF
## RDO_Conc_mgL     1.498519
## Rainfall         2.081334
## Salinity_PSU     1.106708
## NS4_marine_group 1.174154
## Coraliomargarita 1.249510
## Arcobacter       1.910704
```

Fit Model.

```
global <- lme4::glmer(parasite_burden ~ RDO_Conc_mgL +
                      Rainfall + 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 + Rainfall + Salinity_PSU + NS4_marine_group +
##     Coraliomargarita + Arcobacter + (1 | Date)
##    Data: glm_data
##
##      AIC      BIC   logLik deviance df.resid
##     96.3    118.9    -40.1     80.3      117
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1700  0.0024  0.1371  0.3398  1.3108
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Date   (Intercept) 1.053    1.026
```

```
## Number of obs: 125, groups:  Date, 25
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.2683     0.8104   4.033 5.51e-05 ***
## RDO_Conc_mgL       -3.0959     1.4676  -2.109   0.0349 *
## Rainfall           -2.2670     1.1148  -2.034   0.0420 *
## Salinity_PSU       -1.1376     0.9097  -1.251   0.2111
## NS4_marine_group   -2.8276     1.3099  -2.159   0.0309 *
## Coraliomargarita    8.6003     3.4593   2.486   0.0129 *
## Arcobacter         -0.9684     0.8625  -1.123   0.2615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) RDO_C_ Ranfll Sl_PSU NS4_m_ Crlmrg
## RDO_Cnc_mgL -0.467
## Rainfall    -0.359  0.556
## Salinty_PSU -0.249  0.116  0.331
## NS4_mrn_grp -0.327  0.137  0.029  0.181
## Coralimrgrt  0.744 -0.242 -0.103 -0.200 -0.478
## Arcobacter  -0.148  0.376 -0.266 -0.044  0.223 -0.034
```

Goodness of fit and R2.

```
gof(global)
```

```
## D  = 65.0151, df = 117, P(>D) = 0.9999742
## X2 = 66.4648, df = 117, P(>X2) = 0.9999534
```

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

```
##                    R2m       R2c
## theoretical 0.8153717 0.8601263
## delta       0.7856542 0.8287777
```

Backwards selection.

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

```
## Single term deletions
##
## Model:
## parasite_burden ~ RDO_Conc_mgL + Rainfall + Salinity_PSU + NS4_marine_group +
##     Coraliomargarita + Arcobacter + (1 | Date)
##                   npar   dAIC
## <none>                 0.7328
## RDO_Conc_mgL         1 3.9858
```

```
## Rainfall            1 2.9845
## Salinity_PSU        1 0.1624
## NS4_marine_group    1 4.4872
## Coraliomargarita    1 9.1156
## Arcobacter          1 0.0000
```

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

dfun(drop1(global2))
```

```
## Single term deletions
##
## Model:
## parasite_burden ~ RDO_Conc_mgL + Rainfall + NS4_marine_group +
##     Coraliomargarita + (1 | Date)
##                   npar   dAIC
## <none>                  0.0000
## RDO_Conc_mgL         1 1.5281
## Rainfall             1 3.8868
## NS4_marine_group     1 1.6445
## Coraliomargarita     1 6.2654
```

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

```
## Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
##  Family: binomial  ( logit )
## Formula: parasite_burden ~ RDO_Conc_mgL + Rainfall + NS4_marine_group +      Coraliomargarita + (1 |
##    Data: glm_data
##
##      AIC      BIC   logLik deviance df.resid
##     95.0    112.0    -41.5     83.0      119
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.54592  0.00405  0.14635  0.26281  1.15672
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  Date   (Intercept) 2.104    1.451
## Number of obs: 125, groups:  Date, 25
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.2107     0.8724   3.680 0.000233 ***
## RDO_Conc_mgL       -2.6399     1.5552  -1.697 0.089603 .
## Rainfall           -2.5841     1.2157  -2.126 0.033543 *
## NS4_marine_group   -2.2684     1.2993  -1.746 0.080835 .
## Coraliomargarita    7.6272     3.5139   2.171 0.029963 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation of Fixed Effects:
##             (Intr) RDO_C_ Ranfll NS4_m_
## RDO_Cnc_mgL -0.390
## Rainfall    -0.355  0.720
## NS4_mrn_grp -0.257 -0.003 -0.018
## Coralimrgrt  0.682 -0.176 -0.011 -0.467
```

Goodness of fit and R2.

```
gof(global2)
```

```
## D  = 60.5056, df = 119, P(>D) = 0.9999983
## X2 = 53.5577, df = 119, P(>X2) = 1
```

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

```
##                    R2m       R2c
## theoretical 0.7365926 0.8393405
## delta       0.7059230 0.8043928
```

Export summary as a table

```
tab_model(global2, show.se = TRUE, string.se = "Standard Error", show.ci = FALSE,
          show.re.var = FALSE, show.ngroups = FALSE, show.icc = FALSE,
          title = "Generalised logistic regression model",
          file = "C:/Users/Jacob/Desktop/Other_Projects/Aquaculture_Microbiome/Outputs/GLM.doc")
```

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

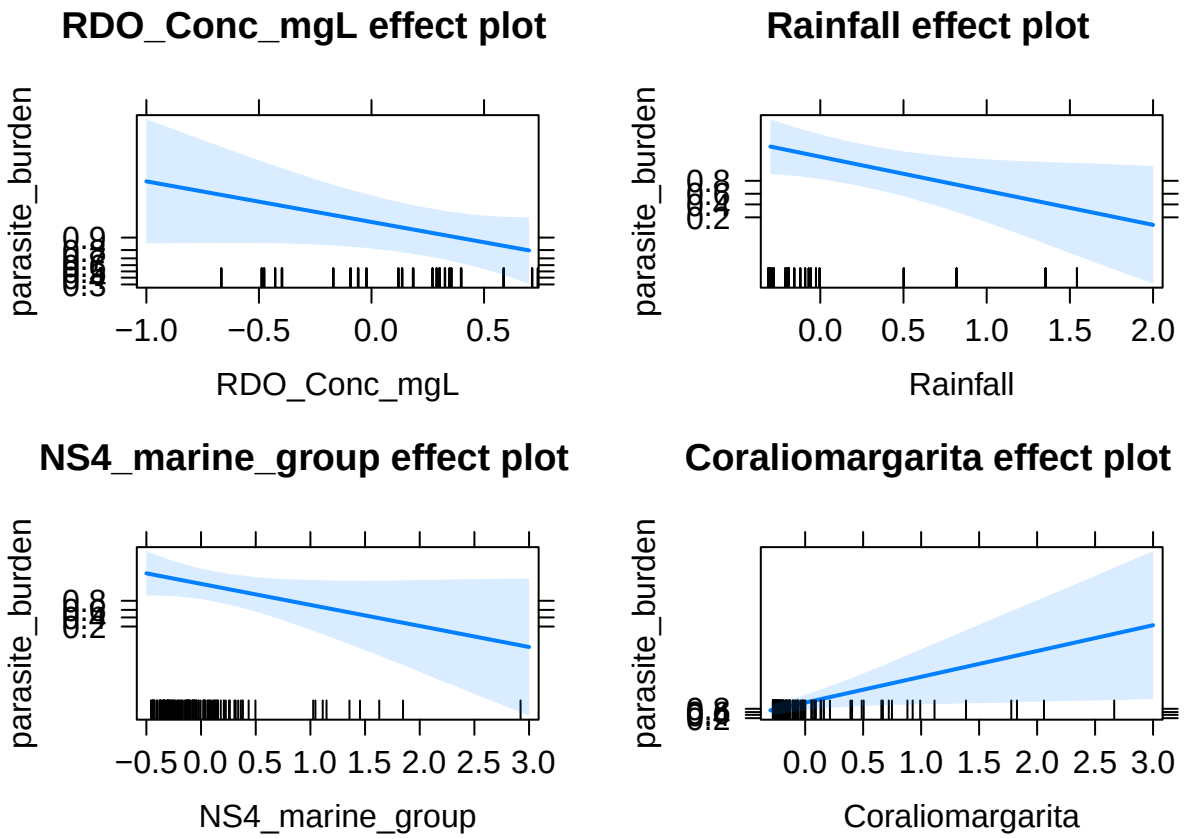
```
lme4::glmer(parasite_burden ~ RDO_Conc_mgL + Salinity_PSU +
                      Rainfall + NS4_marine_group +
                      Coraliomargarita + (1|Date),
                      data = glm_data, family = "binomial") %>%
  tab_model(show.se = TRUE, string.se = "Standard Error", show.ci = FALSE,
            show.re.var = FALSE, show.ngroups = FALSE, show.icc = FALSE,
            title = "Generalised logistic regression model",
            file = "GLM2.doc")
```

```
lme4::glmer(parasite_burden ~ RDO_Conc_mgL + Arcobacter +
                      Rainfall + NS4_marine_group +
                      Coraliomargarita + (1|Date),
                      data = glm_data, family = "binomial") %>%
  tab_model(show.se = TRUE, string.se = "Standard Error", show.ci = FALSE,
            show.re.var = FALSE, show.ngroups = FALSE, show.icc = FALSE,
            title = "Generalised logistic regression model",
            file = "GLM3.doc")
```

Visualisation.

Plot effects

```
plot(allEffects(global2))
```



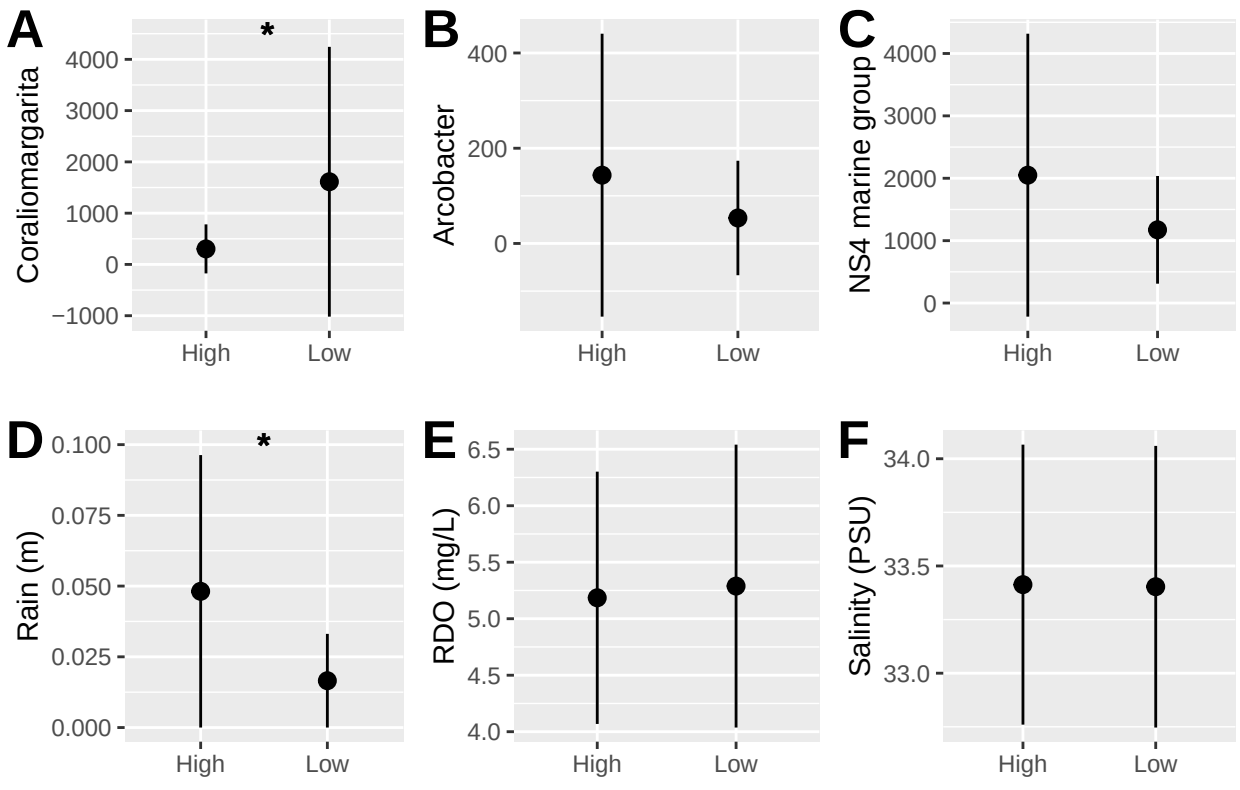
Calculate the mean and standard deviation for each variable.

```
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))
```

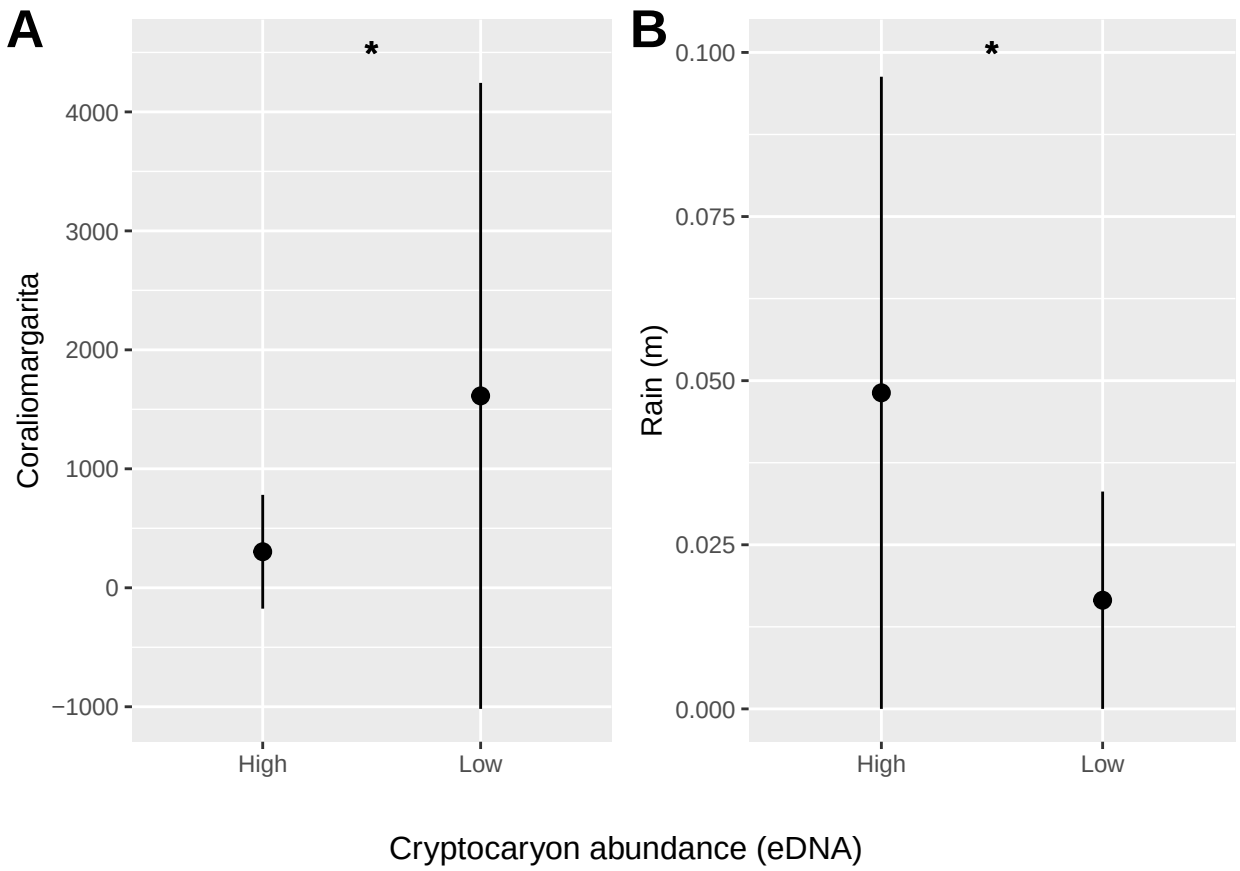
Generate plot.

```
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") +
  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") +
  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 = Rainfall)) +
geom_pointrange(aes(ymin = Rainfall - Rainfall,
                    ymax = Rainfall + Rainfall)) +
  xlab("") +
  ylab("Rain (m)") +
  geom_text(aes(label = "*", y = .1, 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))
)
```




```
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 = Rainfall)) +
geom_pointrange(aes(ymin = Rainfall - Rainfall,
                    ymax = Rainfall + Rainfall)) +
  xlab("") +
  ylab("Rain (m)") +
  geom_text(aes(label = "*", y = .1, x = 1.5, fontface = "bold", size = 20)) +
  theme(legend.position = "none")) %>%
  annotate_figure(fig.lab = "B", fig.lab.face = "bold", fig.lab.size = 20),

ncol = 2, bottom = textGrob("Cryptocaryon abundance (eDNA)", gp = gpar(fontsize = 12))
)
```


A
B
*
*
4000
3000
2000
1000
0
−1000
Coraliomargarita
0.100
0.075
0.050
0.025
0.000
Rain (m)
High
Low
High
Low
Cryptocaryon abundance (eDNA)

Finished.