Rscript03: Analysis microbial performance at biome level

Angel Rain

2025-06-24

Contents

1 Load data

1.1 Load packages

```
knitr::opts_chunk$set(dev = "cairo_pdf")
rm(list=ls())
library(readxl)
library(ggplot2)
library(stringr)
library(dplyr)
library(ggrepel)
library(rstatix)
library(ggfortify)
library(cowplot)
library(tidyr)
library(scales)
library(ggbreak)
library(moments)
library(dendextend)
library(ggExtra)
library(purrr)
library(ggtext)
library(kableExtra)
library(FSA)      # install.packages("FSA") if needed

cbp1 <- c("red","#E69F00","#0072B2","#009E73","#56B4E9","#F0E442","#100000")
```

1.2 Load Rating-data from Rscript02

```
load(file = "../Scripts/Elo_MAPdata_from_Script02.RData")
str(df_MAP_Elo)
```

```
## 'data.frame':    1816294 obs. of  5 variables:
##  $ player_id  : chr  "90_1;96_1;97_1;98_1;99_1" "90_1;96_1;97_1;98_1;99_1" "90_1;96_1;97_1;98_1;99_1
##  $ biome      : chr  "animal.3" "plant.2" "freshwater.3" "freshwater.2" ...
##  $ Category   : chr  "Corrected" "Classic" "Corrected" "Classic" ...
##  $ mean_rating: num  802 1000 794 1000 1000 ...
##  $ sd_rating  : num  8.089 0.028 6.7699 0.0505 0.0473 ...
```

```
tibble(df_MAP_Elo)
```

```
## # A tibble: 1,816,294 x 5
##    player_id                biome        Category  mean_rating sd_rating
##    <chr>                    <chr>        <chr>           <dbl>     <dbl>
##  1 90_1;96_1;97_1;98_1;99_1 animal.3     Corrected        802.    8.09
##  2 90_1;96_1;97_1;98_1;99_1 plant.2      Classic         1000.    0.0280
##  3 90_1;96_1;97_1;98_1;99_1 freshwater.3 Corrected        794.    6.77
##  4 90_1;96_1;97_1;98_1;99_1 freshwater.2 Classic         1000.    0.0505
##  5 90_1;96_1;97_1;98_1;99_1 freshwater.1 Classic         1000.    0.0473
```

```
##  6 90_1;96_1;97_1;98_1;99_1 animal.2     Classic         1000.    0.0179
##  7 90_1;96_1;97_1;98_1;99_1 freshwater.1 Corrected        869.    8.93
##  8 90_1;96_1;97_1;98_1;99_1 plant.3      Classic         1000.    0.0536
##  9 90_1;96_1;97_1;98_1;99_1 plant.2      Corrected        702.   11.7
## 10 90_1;96_1;97_1;98_1;99_1 freshwater.2 Corrected        812.    9.64
## # i 1,816,284 more rows
```

1.2.1 Cheching results

```
print (paste("Number of unique OTUs =",length(unique(df_MAP_Elo$player_id))))
```

```
## [1] "Number of unique OTUs = 124772"
```

1.3 Load metadata

```
df.elo.diversity<-read_xlsx("../data/Supplementary_tables_ms.xlsx",
          sheet="Table S2",skip=1)

tibble(df.elo.diversity)
```

```
## # A tibble: 24 x 10
##      `#` MAP_biomes   Life.Style      MAP_biomesFullName   n.samples.initial
##    <dbl> <chr>        <chr>           <chr>                            <dbl>
##  1     1 airborne.1   Free-living     airborne                          1595
##  2     2 animal.1     Host-associated animal-urogenital                29317
##  3     3 animal.2     Host-associated animal-proximalgut              180797
##  4     4 animal.3     Host-associated animal-distalgut                125143
##  5     5 animal.4     Host-associated animal-oral                      37433
##  6     6 animal.5     Host-associated animal-skin                      40488
##  7     7 animal.6     Host-associated animal-respiratory               10360
##  8     8 freshwater.1 Free-living     freshwater-sediments             19946
##  9     9 freshwater.2 Free-living     freshwater-water                 46327
## 10    10 freshwater.3 Free-living     freshwater-biofilm                5067
## # i 14 more rows
## # i 5 more variables: n.samples.1000reads.3OTUS <dbl>, gamma.diversity <dbl>,
## #   `Elo.coeffcient (a)` <dbl>, `Elo.coeffcient (a).error` <dbl>,
## #   `Elo.coeffcient (a).pvalue` <dbl>
```

1.4 Load alpha diversity data

```
#Load files per biome
diversity.files<-list.files(path="../data/Diversity_alpha/",pattern="AlphaRaw.*")

list.tmp<-list()

for (i in 1:length(diversity.files)){
  df.tmp<-read.csv(paste0("../data/Diversity_alpha/",diversity.files[i]))
```

```
  df.tmp$biome<-str_split_fixed(str_remove_all(diversity.files[i],".csv"),"[.]",2)[,2]
  df.tmp<-df.tmp[,c(1:6)]
  names(df.tmp)<-c("sample","richness","shannon","pielou","simpson","biome")
  list.tmp[[i]]<-df.tmp
  rm(df.tmp) # save some memory
}

#Combine datasets
df.diversity<-do.call(rbind,list.tmp)
rm(list.tmp,diversity.files) #Remove unused files
```

1.5 Setting rating data.frame

```
# Add full names here:
df_MAP_Elo<-merge(df.elo.diversity,df_MAP_Elo,by.x="MAP_biomes",by="biome")
df_MAP_Elo$MAP_biomesFullName<-as.factor(df_MAP_Elo$MAP_biomesFullName)
str(df_MAP_Elo)
```

```
## 'data.frame':    1816294 obs. of  14 variables:
##  $ MAP_biomes               : chr  "airborne.1" "airborne.1" "airborne.1" "airborne.1" ...
##  $ #                        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Life.Style               : chr  "Free-living" "Free-living" "Free-living" "Free-living" ...
##  $ MAP_biomesFullName       : Factor w/ 24 levels "airborne","animal-distalgut",..: 1 1 1 1 1 1 1 1
##  $ n.samples.initial        : num  1595 1595 1595 1595 1595 ...
##  $ n.samples.1000reads.3OTUS: num  1449 1449 1449 1449 1449 ...
##  $ gamma.diversity          : num  24141 24141 24141 24141 24141 ...
##  $ Elo.coeffcient (a)       : num  1.02 1.02 1.02 1.02 1.02 ...
##  $ Elo.coeffcient (a).error : num  0.00012 0.00012 0.00012 0.00012 0.00012 0.00012 0.00012 0.00012 0
##  $ Elo.coeffcient (a).pvalue: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ player_id                : chr  "90_602;96_5854;97_7045;98_8712;99_11631" "90_1728;96_20467;97_254
##  $ Category                 : chr  "Classic" "Corrected" "Classic" "Classic" ...
##  $ mean_rating              : num  1000 866 1000 1000 866 ...
##  $ sd_rating                : num  0.0837 3.1459 0.0219 0.2918 3.1449 ...
```

```
#Re-order biome names
levels_sorted<-c("animal-urogenital","animal-proximalgut","animal-distalgut",
                 "animal-oral","animal-skin","animal-respiratory",
                 "plant-rhizosphere","plant-phyllosphere","plant-endosphere",
                 "plant-spermosphere","airborne","freshwater-sediments",
                 "freshwater-water","freshwater-biofilm","freshwater-peatlands(peat/bog)",
                 "freshwater-peatlands(water)", "saline-sediments","saline-water",
                 "saline-biofilm", "soil-agricultural","soil-desert",
                 "soil-tundra","soil-forest","soil-grassland")

#Set biome full names as a factor
df_MAP_Elo$MAP_biomesFullName<-factor(df_MAP_Elo$MAP_biomesFullName,levels_sorted)
```

2 OTUs ratings description at biome level

```
# 1) define your color map for Life.Style
style_cols <- c(`Free-living`     = "blue2",`Host-associated` = "red3")

# 2) build a named vector of HTML-wrapped labels
label_vec <- with(df_MAP_Elo,setNames(
    sprintf("<span style='color:%s;'>%s</span>",
            style_cols[Life.Style],
            MAP_biomesFullName),MAP_biomesFullName))

## Setting plot for absence corrected rating or BB-score
plot.descriptive.corrected.Elo<-df_MAP_Elo[df_MAP_Elo$Category=="Corrected",]%>%
  ggplot(aes(x = MAP_biomesFullName, y = (mean_rating))) +
  geom_point(stat = "identity",shape=20,size=0.5,alpha=0.75,
             position = position_dodge2(0.5),aes(colour=Life.Style)) +
  stat_summary(fun="mean",size=0.8,geom="point")+
  coord_flip() +
    theme_bw()+ theme(panel.grid.major= element_blank(), panel.grid.minor= element_blank())+
    theme(text = element_text(size=11))+
  labs(title="A)", x = NULL,y = "BB-score")+
  scale_x_discrete(labels = label_vec)+
  theme(panel.grid      = element_blank(),
        text            = element_text(size = 11),
        axis.text.y     = element_markdown(),           # render the <span> tags
        legend.position = "none")+
  scale_colour_manual(values=c("blue2","red3"))+
  theme(plot.title = element_text(face="bold",vjust=-4,hjust=0.01,size=12))+
  theme(plot.margin = margin(t=-4, r=6, l=1, 0),
        plot.title= element_text(margin = margin(t = 0,b=0)))

## Setting plot for Elo-rating
plot.descriptive.classic.Elo<-df_MAP_Elo[df_MAP_Elo$Category=="Classic",]%>%
  ggplot(aes(x = MAP_biomesFullName, y = (mean_rating))) +
  geom_point(stat = "identity",shape=20,size=0.5,alpha=0.75,
             position = position_dodge2(0.5),aes(colour=Life.Style)) +
  stat_summary(fun="mean",size=0.8,geom="point")+
  coord_flip() +
    theme_bw()+ theme(panel.grid.major= element_blank(), panel.grid.minor= element_blank())+
    theme(text = element_text(size=11))+
  labs(title="", x = NULL,y = "Classic Elo-Rating")+
  scale_x_discrete(labels = label_vec)+
  theme(panel.grid      = element_blank(),
        text            = element_text(size = 11),
        axis.text.y     = element_markdown(),           # render the <span> tags
        legend.position = "none")+
  scale_colour_manual(values=c("blue2","red3"))+
  scale_y_continuous(limits=c(900,1250))   # keep numeric labels
```

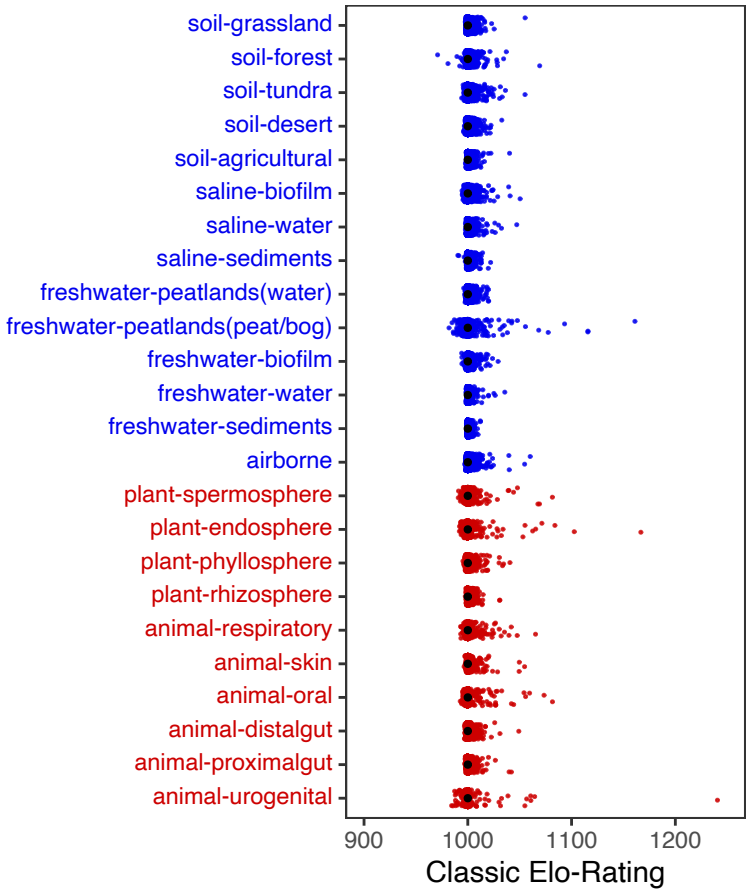
MAP_biomesFullName	mean_rating	sd_rating
animal-urogenital	1000	2.5
animal-proximalgut	1000	0.5
animal-distalgut	1000	0.5
animal-oral	1000	1.2
animal-skin	1000	0.6
animal-respiratory	1000	1.1
plant-rhizosphere	1000	0.4
plant-phyllosphere	1000	0.6
plant-endosphere	1000	2.1
plant-spermosphere	1000	1.4
airborne	1000	1.0
freshwater-sediments	1000	0.2
freshwater-water	1000	0.3
freshwater-biofilm	1000	0.5
freshwater-peatlands(peat/bog)	1000	3.5
freshwater-peatlands(water)	1000	0.6
saline-sediments	1000	0.4
saline-water	1000	0.5
saline-biofilm	1000	1.0
soil-agricultural	1000	0.4
soil-desert	1000	0.7
soil-tundra	1000	1.1
soil-forest	1000	0.7
soil-grassland	1000	0.6

2.1 Display biome-level data from Elo-ratings estimations

```
#Calculate the average and standard deviation
tmp.av<-aggregate(mean_rating~MAP_biomesFullName,
                  data=df_MAP_Elo[df_MAP_Elo$Category=="Classic",],mean)
tmp.av$sd_rating<-aggregate(mean_rating~MAP_biomesFullName,
                            data=df_MAP_Elo[df_MAP_Elo$Category=="Classic",],sd)[,2]

tmp.av%>%
  kbl(digits=1)%>%
  kable_styling(full_width = FALSE, position = "center")
```

2.2 Classic Elo-rating Figure S1

Elo-ratings across biomes. Each dot represents the Elo-rating of an individual OTU across soil, freshwater, saline, plant-associated, and animal-associated biomes (blue: free-living; red: host-associated). The vertical line indicates the average Elo-ratins per biome, which converges at 1000 due to the zero-sum game property of Elo-ratings estimations, where per community gains by some OTUs are balanced by losses of others.



3 Diversity indices across biomes

Alpha- and gamma-diversity across biomes. (A) Sample-level richness (OTU count), (B) Shannon diversity, and (C) Pielou's evenness by biome. Boxplots show the median (center line) and the first and third quartiles (bottom and upper limits of boxes); whiskers span the 10th and 90th percentiles. Overlaid points represent the individual samples. (D) gamma-diversity (total OTUs observed across all samples within each biome)

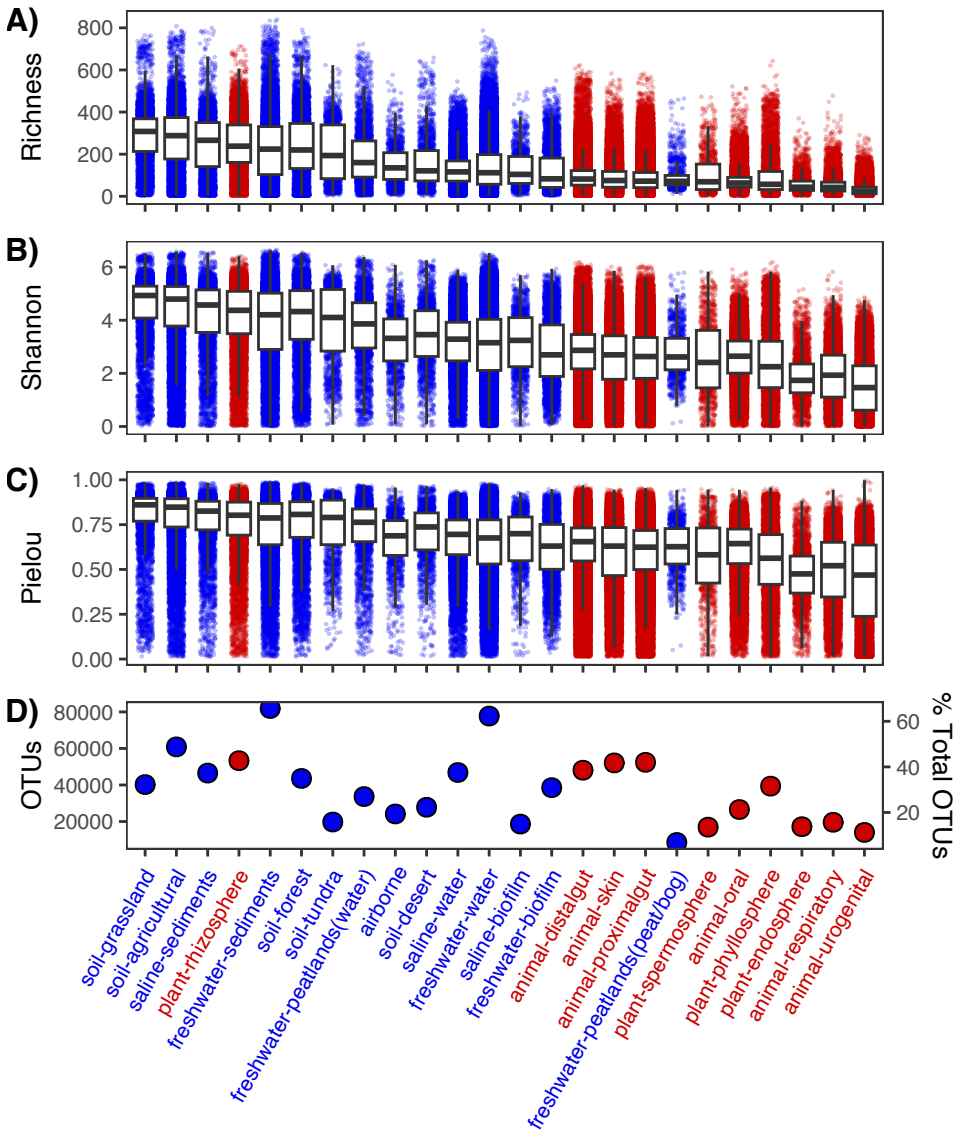
```
# Add diversity indeces
df.diversity<-merge(df.diversity,df.elo.diversity[,c(2,3,4,7)],by.x="biome",by.y="MAP_biomes")

# Plot richness
plot.R<-df.diversity%>%
ggplot(aes(x=reorder(MAP_biomesFullName,-richness,median),richness))+
  geom_point(stat = "identity",shape=1,size=0.07,alpha=0.25,
             position = position_dodge2(0.5),aes(colour=Life.Style)) +
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())+
  theme(text= element_text(size = 10),
        axis.text.x = element_blank(),
        legend.position = "none")+
  scale_colour_manual(values=c("blue2","red3"))+
```

```
  labs(y="Richness",x=NULL)

#Plot Shannon index
plot.H<-df.diversity%>%
ggplot(aes(x=reorder(MAP_biomesFullName,-richness,median),shannon))+
  geom_point(stat = "identity",shape=1,size=0.07,alpha=0.25,
             position = position_dodge2(0.5),aes(colour=Life.Style)) +
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())+
  theme(text= element_text(size = 10),
        axis.text.x = element_blank(),           # render the <span> tags
        legend.position = "none")+
  scale_colour_manual(values=c("blue2","red3"))+
  labs(y="Shannon",x=NULL)

#Plot Pielou
plot.Pielou<-df.diversity%>%
ggplot(aes(x=reorder(MAP_biomesFullName,-richness,median),pielou))+
  geom_point(stat = "identity",shape=1,size=0.07,alpha=0.25,
             position = position_dodge2(0.5),aes(colour=Life.Style)) +
  geom_boxplot(outlier.shape = NA)+
  theme_bw()+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())+
 # scale_x_discrete(labels = label_vec)+
  theme(text= element_text(size = 10),
        axis.text.x =element_blank(),           # render the <span> tags
    legend.position = "none")+
    scale_colour_manual(values=c("blue2","red3"))+
  labs(y="Pielou",x=NULL)

#Plot gamma diversity (biome pool)
plot.gamma<-df.diversity%>%
ggplot(aes(x=reorder(MAP_biomesFullName,-richness,median),gamma.diversity))+
  stat_summary(fun = "mean",geom="point",shape=21,size=3,aes(fill=Life.Style))+
  theme_bw()+theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank())+
  scale_x_discrete(labels = label_vec)+
  theme(text= element_text(size = 10),
        axis.text.x = element_markdown(angle=60,hjust=1),           # render the <span> tag
        legend.position = "none")+
  scale_fill_manual(values=c("blue2","red3"))+
  scale_y_continuous(name="OTUs",
                     sec.axis = sec_axis(~./1247.72, name="% Total OTUs")) +
  labs(y="OTUs",x=NULL)
```

3.1 Diversity indeces Figure S2

Alpha- and gamma-diversity across biomes. (A) Sample-level richness (OTU count), (B) Shannon diversity, and (C) Pielou's evenness by biome. Boxplots show the median (center line) and the first and third quartiles (bottom and upper limits of boxes); whiskers span the 10th and 90th percentiles. Overlaid points represent the individual samples. (D) gamma-diversity (total OTUs observed across all samples within each biome); each point represents one biome.

```
## cairo_pdf
##         2
```



4 Summarize BB-score adata and diversity indices

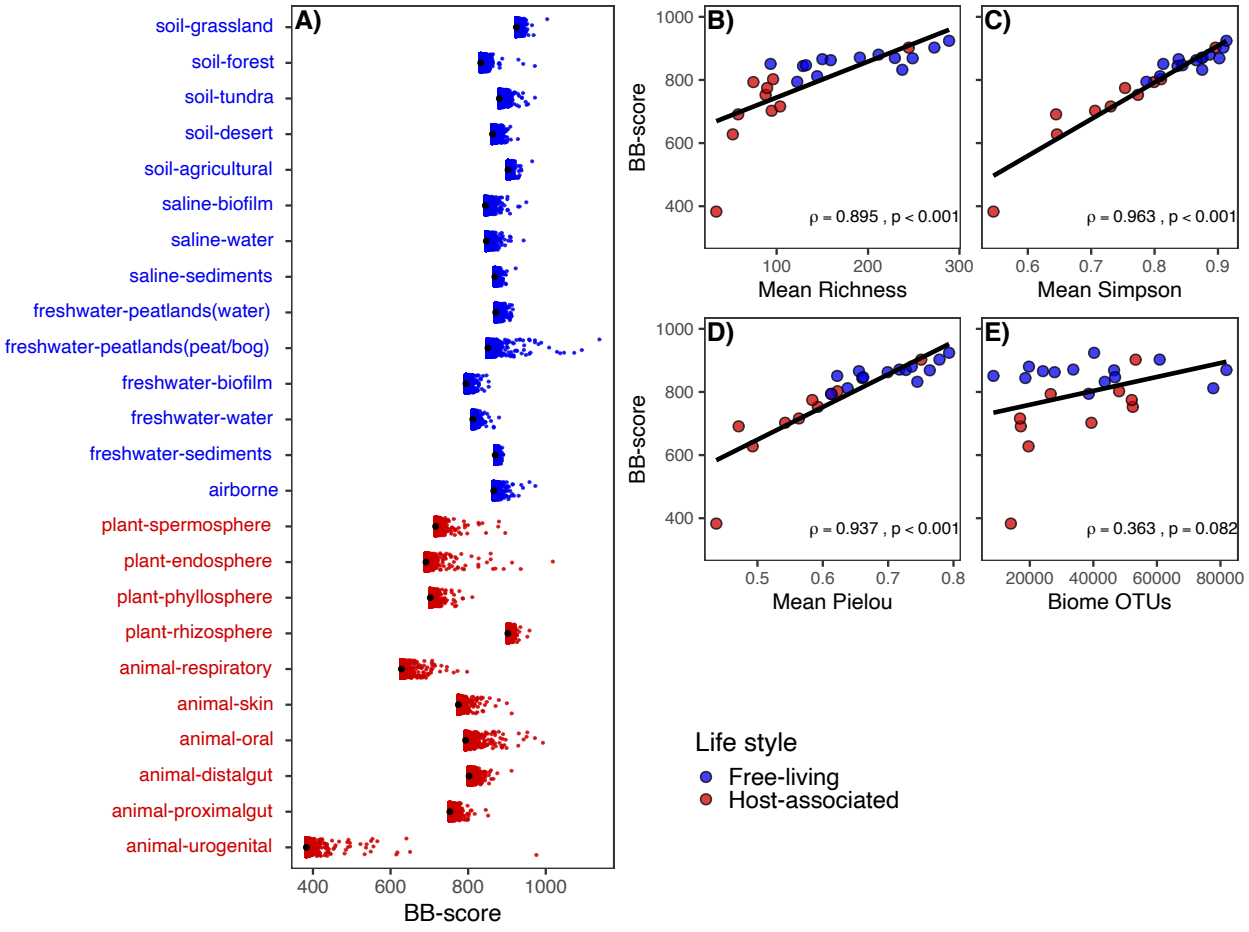
Mean diversity indices (richness, Shannon, Pielou, Simpson,and gamma diversity) and BB-scores aggregated at the biome level are correlated (Spearman rank correlation). .

```
df.diversity.summary<-aggregate(.~MAP_biomesFullName,data=df.diversity[,c(3,4,5,6,8)],mean)
df.elo.summary<-aggregate(mean_rating~MAP_biomesFullName,data=df_MAP_Elo[df_MAP_Elo$Category=="Correcte

df.elo.diversity<-merge(df.elo.diversity,df.diversity.summary,by="MAP_biomesFullName")
df.elo.diversity<-merge(df.elo.diversity,df.elo.summary,by="MAP_biomesFullName")

#Select variables for correlations test
var.to.test<-c("richness","shannon","pielou","simpson",
               "n.samples.1000reads.3OTUS","gamma.diversity")
```

```r
# Results Spearman correlations
spearman_results <- map_dfr(var.to.test, function(var) {
  # Formula for mean_elo ~ var
  formula <- as.formula(paste("mean_rating ~", var))
  # run test
  test <- cor.test(
    df.elo.diversity[[var]],
    df.elo.diversity$mean_rating,
    method = "spearman")
  # extract estimate and p.value
  tibble(
    variable = var,
    rho      = unname(test$estimate),
    p_value  = test$p.value)
})

spearman_results$label <- with(spearman_results,
  paste0("rho==", round(rho, 3),
         ifelse(p_value < 0.001, "~','~p<0.001",
                paste0("~','~p==", round(p_value, 3)))))

# 4. Build one ggplot per variable in a list
plots.diversity <- lapply(seq_len(nrow(spearman_results)), function(i) {
  var  <- spearman_results$variable[i]
  lab  <- spearman_results$label[i]
  #plots
  ggplot(df.elo.diversity,
         aes_string(x = var, y = "mean_rating")) +
    geom_point(aes(fill = Life.Style), shape = 21, size = 2,alpha=0.75) +
    geom_smooth(method = "lm", color = "black", se = FALSE) +
    labs(x = var, y = "BB-score") +ylim(300,1000)+
    theme_bw() +
    theme(text = element_text(size=10),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          legend.position   = "none") +
    scale_fill_manual(values = c("blue2", "red3")) +
    annotate("text",
             x     = Inf,    # place at top-right corner...
             y     = 400,    # ...and then nudge in with hjust/vjust
             label = lab,
             parse = TRUE,
             hjust = 1,
             vjust = 1.25,size=2.75)
})
```

4.1 Display biome average BB-score and diversity indices (Figure 3)

Baas Becking score (BB-score) across biomes and relationship with diversity metrics. (A) Distribution of BB-scores across soil, freshwater, saline, plant-associated, and animal-associated biomes (blue: free-living; red: host-associated). Correlations between mean BB-scores and community diversity metrics: species richness (B), Shannon index (C), Pielou's evenness (D), Simpson index (E), and gamma-diversity (F). Points represents biome means; black lines show correlation with Spearman's correlation coefficient and significance.


A)
soil-grassland
soil-forest
soil-tundra
soil-desert
soil-agricultural
saline-biofilm
saline-water
saline-sediments
freshwater-peatlands(water)
freshwater-peatlands(peat/bog)
freshwater-biofilm
freshwater-water
freshwater-sediments
airborne
plant-spermosphere
plant-endosphere
plant-phyllosphere
plant-rhizosphere
animal-respiratory
animal-skin
animal-oral
animal-distalgut
animal-proximalgut
animal-urogenital
400
600
800
1000
BB-score
B)
BB-score
1000
800
600
400
ρ = 0.895 , p < 0.001
100
200
300
Mean Richness
C)
ρ = 0.963 , p < 0.001
0.6
0.7
0.8
0.9
Mean Simpson
D)
BB-score
1000
800
600
400
ρ = 0.937 , p < 0.001
0.5
0.6
0.7
0.8
Mean Pielou
E)
ρ = 0.363 , p = 0.082
20000
40000
60000
80000
Biome OTUs
Life style
Free-living
Host-associated