Script01: Microbial occurrence

Angel Rain

2025-06-13

Contents

1 Load data **1**

1.1 Load packages 1

1.2 Load and plot results from server OTU99 1

1.3 OTUs biome occurrence 2

1.3.1 Describe OTUs occurrence per life style. Figure 2A 3

1.4 Describe unique OTU biome origin 5

1.4.1 Figure 2B 6

1 Load data

1.1 Load packages

```
rm(list=ls())
library(dplyr)
library(stringr)
library(purrr)
library(ggplot2)
library(tidyr)
library(data.table)
library(ggtext)
library(readxl)
library(cowplot)
library(kableExtra)
```

1.2 Load and plot results from server OTU99

```
# Count all OTUs
dt <- fread("../data/Occurrence/otu_biome_presence_summary.csv")    # reads very quickly
setnames(dt, "biomes_list", "raw")                # rename for clarity

# split & unlist in one go:
```

```
dt_long <- dt[, .(biome = unlist(strsplit(raw, ";"))),
              by = .(otu, n_biomes)]
dt_long$biome<-as.factor(dt_long$biome)
levels(dt_long$biome)
```

```
##  [1] "airborne.1"   "animal.1"     "animal.2"     "animal.3"     "animal.4"
##  [6] "animal.5"     "animal.6"     "freshwater.1" "freshwater.2" "freshwater.3"
## [11] "peatland.1"   "peatland.2"   "plant.1"      "plant.2"      "plant.3"
## [16] "plant.4"      "saline.1"     "saline.2"     "saline.3"     "soil.1"
## [21] "soil.2"       "soil.3"       "soil.4"       "soil.5"
```

```
head(dt_long)
```

```
##                                    otu n_biomes      biome
##                                 <char>    <int>     <fctr>
## 1: 90_45;96_1693;97_2027;98_2492;99_3254       22 airborne.1
## 2: 90_45;96_1693;97_2027;98_2492;99_3254       22   animal.1
## 3: 90_45;96_1693;97_2027;98_2492;99_3254       22   animal.2
## 4: 90_45;96_1693;97_2027;98_2492;99_3254       22   animal.3
## 5: 90_45;96_1693;97_2027;98_2492;99_3254       22   animal.4
## 6: 90_45;96_1693;97_2027;98_2492;99_3254       22   animal.5
```

```
#Load biome metadata
metadata<-read_xlsx("../data/Supplementary_tables_ms.xlsx",
          sheet="Table S2")
head(metadata)
```

```
## # A tibble: 6 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
## # i 5 more variables: n.samples.1000reads.3OTUS <dbl>, gamma.diversity <dbl>,
## #   Elo.coef <dbl>, Elo.coef.error <dbl>, Elo.pvalue <dbl>
```

1.3 OTUs biome occurrence

```
#Summarize counts
df.summary<-aggregate(otu~n_biomes,data=unique(dt_long[,c(1:2)]),length)
#Get percentages
df.summary$otu.perc<-df.summary$otu/sum(df.summary$otu)*100
#Get the cumulative sun of the percentages
df.summary$cum.sum<-cumsum(df.summary$otu.perc)
kbl(df.summary,,digits=1)%>%
  kable_styling(full_width = FALSE, position = "center")
```

n_biomes	otu	otu.perc	cum.sum
1	13997	11.2	11.2
2	15052	12.1	23.3
3	12161	9.7	33.0
4	10494	8.4	41.4
5	8926	7.2	48.6
6	8004	6.4	55.0
7	7073	5.7	60.7
8	6277	5.0	65.7
9	5632	4.5	70.2
10	4976	4.0	74.2
11	4527	3.6	77.8
12	4140	3.3	81.2
13	3656	2.9	84.1
14	3279	2.6	86.7
15	2909	2.3	89.0
16	2643	2.1	91.2
17	2391	1.9	93.1
18	2046	1.6	94.7
19	1831	1.5	96.2
20	1535	1.2	97.4
21	1293	1.0	98.5
22	934	0.7	99.2
23	646	0.5	99.7
24	350	0.3	100.0

1.3.1 Describe OTUs occurrence per life style. Figure 2A

Distribution of OTUs according to number of biomes in which they were found. Bars (left axis): number of OTUs occurring in one or more biomes (see Methods). Line and symbols (right axis): cumulative percentage of OTUs in increasing numbers of biomes. Inset pie chart: proportions of OTUs exclusively present in host-associated or free-living biomes, or found in both types.

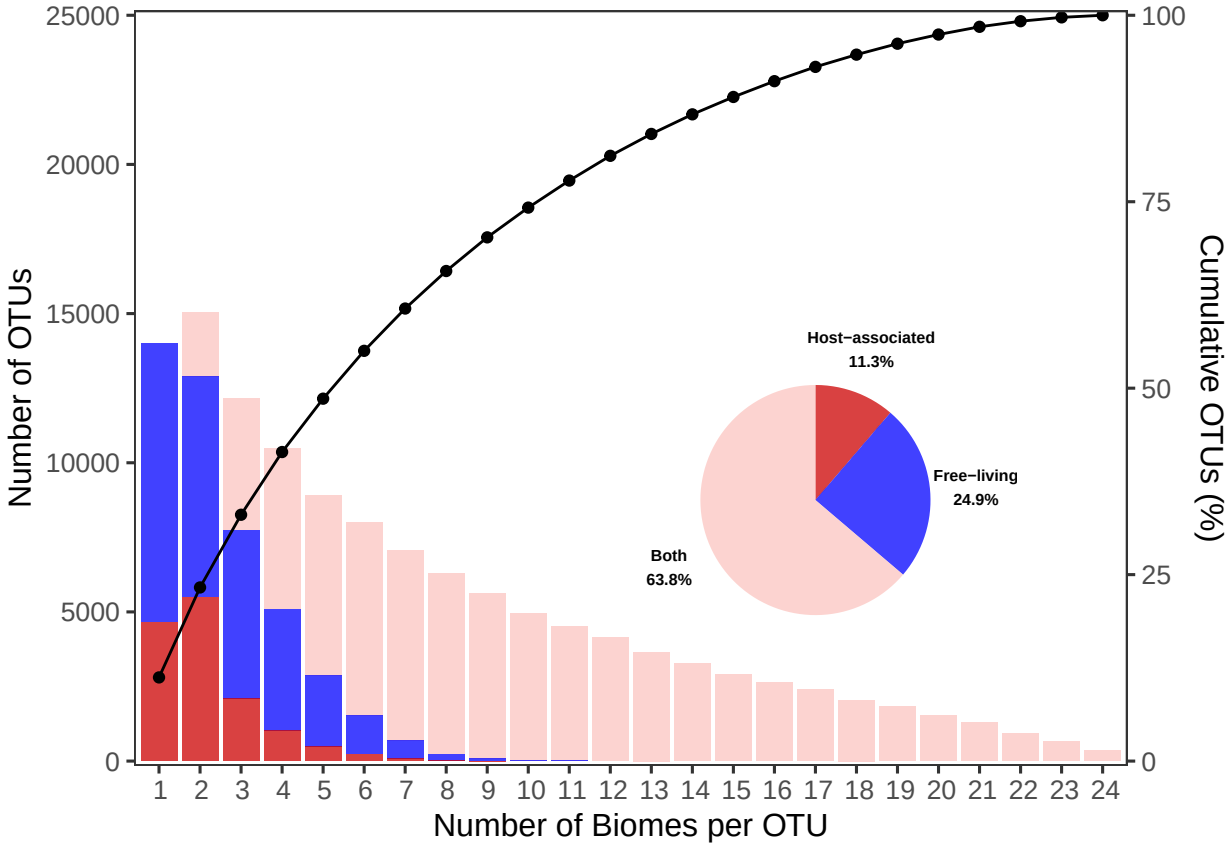
```
dt_long.common<-dt_long

dt_long.common<-merge(dt_long.common,metadata[,c(1:3)],by.x="biome",by.y="MAP_biomes")


dt_otu_life_style <- dt_long.common %>%
  group_by(otu) %>%
  summarise(
    n_biomes  = n_distinct(biome),                                    # how many distinct biomes
    n_styles  = n_distinct(Life.Style),                               # how many distinct lifestyl
    styles    = paste(sort(unique(Life.Style)), collapse = " & "),    # e.g. "Free-living" or "Fre
    category  = case_when(                                            # categorize based on those
      n_styles == 1 & styles == "Free-living"     ~ "Free-living only",
      n_styles == 1 & styles == "Host-associated" ~ "Host-associated only",
      TRUE~ "Both"),
    .groups = "drop")

dt_otu_life_style$category<-as.factor(dt_otu_life_style$category)
```

```r
dt_otu_life_style$category<-factor(dt_otu_life_style$category,
                                   c("Both","Free-living only","Host-associated only"))
# how Many OTUs per category?
aggregate(n_biomes~ category, data=dt_otu_life_style,length)
```

```
##               category n_biomes
## 1                 Both    79577
## 2     Free-living only    31111
## 3 Host-associated only    14084
```

```r
df <- tibble::tibble(
  category = c("Both", "Free-living", "Host-associated"),
  otu_count = c(79577, 31111, 14084)
) %>%
  arrange(desc(category)) %>%
  mutate(
    fraction = otu_count / sum(otu_count),
    ymax = cumsum(fraction),
    ymin = c(0, head(ymax, n = -1)),
    label_pos = (ymax + ymin) / 2,#*2+pi,
    label = paste0(category, "\n", round(fraction * 100, 1), "%")
  )

# Plot with accurate label placement
pie.plot<-ggplot(df, aes(ymax = ymax, ymin = ymin, xmax = 4, xmin = 2, fill = category)) +
  geom_rect(alpha = 0.75) +
  geom_text(aes(x = 4.8, y = label_pos, label = label,fontface="bold"),
            color = "black", size = 2.2) +
  coord_polar(theta = "y") +
  #xlim(c(0, 4)) +
  scale_fill_manual(values = c("#fcc5c0", "blue", "red3"), guide = FALSE) +
  theme_void()

pie_grob <- ggplotGrob(pie.plot)

plot.panel<-dt_otu_life_style %>%
  # 1. tally one row per otu-category-n_biomes combination
  count(n_biomes, category, name = "otu_count") %>%
  # 2. plot
  ggplot(aes(x = factor(n_biomes), y = otu_count)) +
    geom_col(aes(fill=category),alpha=0.75) + scale_fill_manual(values=c("#fcc5c0","blue","red3"),name=
    #facet_grid(~ category,scales="free_x",space = "free_x") +
    labs(
      x     = "Number of Biomes per OTU",
      y     = "Number of OTUs") +
    theme_bw(base_size = 12) +
    theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank(),
      axis.text.x = element_text(angle = 0, hjust = 0.4),
      strip.text  = element_text(face = "bold"),legend.position ="none")+# c(0.75, 0.35))+
  scale_y_continuous(expand=c(0,1000),limits = c(0,20000))

#Add piechart
plot.panel<-plot.panel +
```

```
  annotation_custom(grob = pie_grob,
                    xmin = 12, xmax = 22,
                    ymin = 2000, ymax = 15500)

#Add secondary axis cumulative % OTUs
plot.panel<-plot.panel+geom_line(data=df.summary,aes(n_biomes,cum.sum*250))+
  geom_point(data=df.summary,aes(n_biomes,cum.sum*250))+
  scale_y_continuous(sec.axis = sec_axis(trans = ~./250,name="Cumulative OTUs (%)"),expand = c(0,150))

plot.panel
```



1.4 Describe unique OTU biome origin

```
#Extract biome-unique OTUs
uniq.otu.biomes<-dt[dt$n_biomes==1,]

uniq.otu.biomes$biome<-as.factor(uniq.otu.biomes$raw)
levels(uniq.otu.biomes$biome)
```

```
##  [1] "airborne.1"   "animal.1"     "animal.2"     "animal.3"     "animal.4"
##  [6] "animal.5"     "animal.6"     "freshwater.1" "freshwater.2" "freshwater.3"
## [11] "peatland.1"   "peatland.2"   "plant.1"      "plant.2"      "plant.3"
## [16] "plant.4"      "saline.1"     "saline.2"     "saline.3"     "soil.1"
```

Life.Style	MAP_biomesFullName	otu	gamma.diversity	perc.unique	perc.unique.biome
Free-living	airborne	64	24141	0.5	0.3
Host-associated	animal-urogenital	63	14039	0.5	0.4
Host-associated	animal-proximalgut	1686	52407	12.0	3.2
Host-associated	animal-distalgut	882	48089	6.3	1.8
Host-associated	animal-oral	538	26565	3.8	2.0
Host-associated	animal-skin	651	52082	4.7	1.2
Host-associated	animal-respiratory	56	19591	0.4	0.3
Free-living	freshwater-sediments	2643	81865	18.9	3.2
Free-living	freshwater-water	1011	77726	7.2	1.3
Free-living	freshwater-biofilm	410	38546	2.9	1.1
Free-living	freshwater-peatlands(peat/bog)	13	8546	0.1	0.2
Free-living	freshwater-peatlands(water)	183	33662	1.3	0.5
Host-associated	plant-rhizosphere	425	53293	3.0	0.8
Host-associated	plant-phyllosphere	186	39370	1.3	0.5
Host-associated	plant-endosphere	135	17164	1.0	0.8
Host-associated	plant-spermosphere	35	16914	0.3	0.2
Free-living	saline-sediments	1325	46514	9.5	2.8
Free-living	saline-water	1917	46833	13.7	4.1
Free-living	saline-biofilm	211	18596	1.5	1.1
Free-living	soil-agricultural	690	60857	4.9	1.1
Free-living	soil-desert	288	27806	2.1	1.0
Free-living	soil-tundra	56	19728	0.4	0.3
Free-living	soil-forest	325	43586	2.3	0.7
Free-living	soil-grassland	204	40227	1.5	0.5

```
## [21] "soil.2"       "soil.3"       "soil.4"       "soil.5"
```

```
#Summarize the n of OTUS per biome
agg.unique.otu.biome<-aggregate(otu~ raw,data=uniq.otu.biomes,length)

#merge with biome metadata
agg.unique.otu.biome<-uniq.otu.biomes<-merge(agg.unique.otu.biome,metadata,by.x="raw",by.y="MAP_biomes"

#Calcualte how many of the unique OTU come from each biome
agg.unique.otu.biome$perc.unique<-agg.unique.otu.biome$otu/sum(agg.unique.otu.biome$otu)*100

#Calcualte how many of OTUs are unique from the total species pool
agg.unique.otu.biome$perc.unique.biome<-agg.unique.otu.biome$otu/agg.unique.otu.biome$gamma.diversity*1

kbl(agg.unique.otu.biome[, c(4,5,2,8,12,13)],digits=1)|>
  kable_styling(full_width = FALSE, position = "center")
```

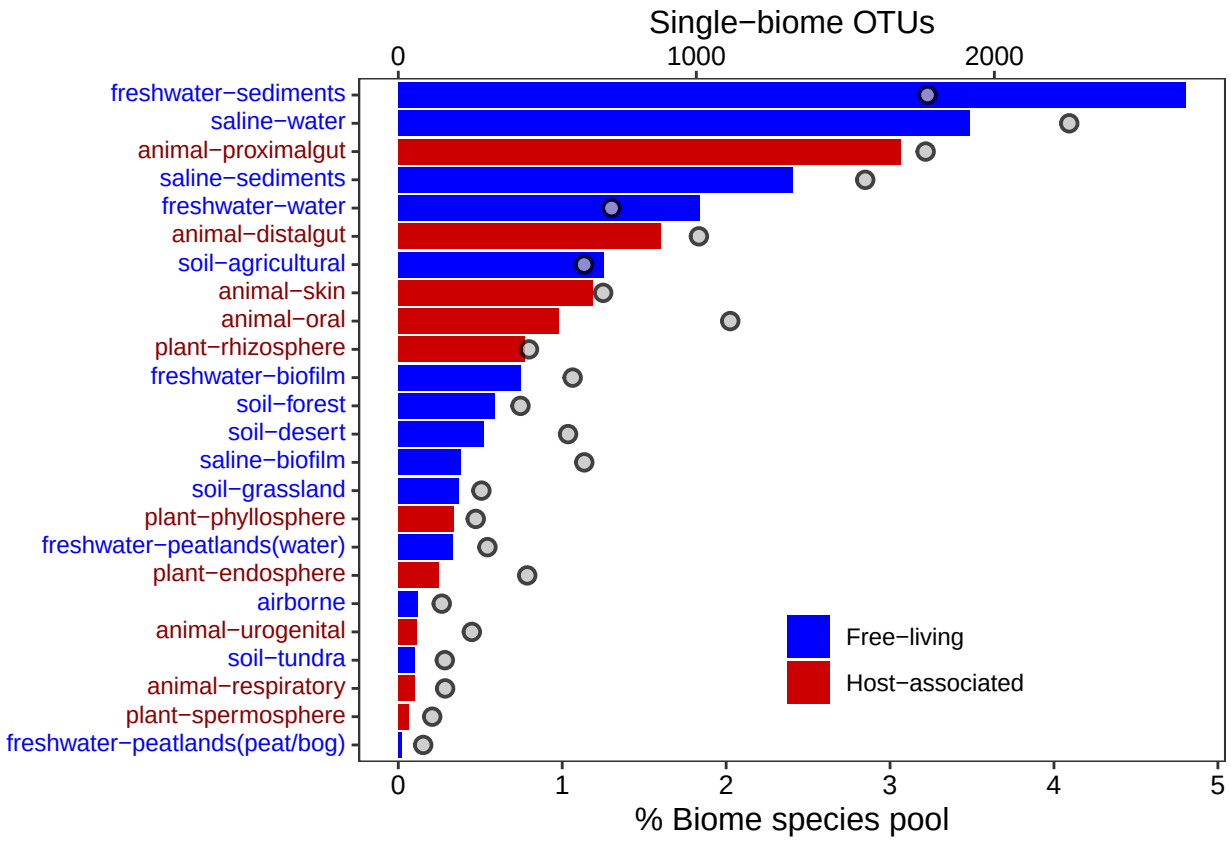
1.4.1 Figure 2B

Contribution of host-associated and free-living OTUs that are limited to a single biome to the total biome species pool. Bars (top axis) indicate the umber of OTUs. The grey symbols (bottom axis) indicate the proportion of total species pool of the respective biome.

```r
# 1) define your color map for Life.Style
style_cols <- c(`Free-living`     = "blue",`Host-associated` = "red4")

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

#Distribution of percentage of biome-unique OTUs
quantile(agg.unique.otu.biome$perc.unique.biome)
```

```
##        0%       25%       50%       75%      100%
## 0.1521179 0.4665182 0.9166129 1.4340671 4.0932676
```

```r
plot.unique.biomes<-agg.unique.otu.biome %>%
  ggplot(aes(x = reorder(MAP_biomesFullName, otu), y = perc.unique.biome), alpha = 0.75, width = 0.75) +
    scale_y_continuous(sec.axis = sec_axis(trans=~.*550, name="Single-biome OTUs"))+
  geom_point(fill="grey",alpha=0.75,shape=21,size=2.25,stroke=1)+
    #geom_boxplot(aes(fill = life_style), outlier.shape = NA,alpha=0.65) +
    theme_bw() +
    scale_colour_manual(name=NULL,values = c("blue","red3")) +
    scale_fill_manual(name=NULL,values= c("blue","red3")) +
    labs(x = NULL, y = "% Biome species pool") +
    theme(
      panel.grid.major      = element_blank(),
      panel.grid.minor      = element_blank(),
      # colour the secondary axis labels/title grey
      axis.title.x.top    = element_text(size=12,color = "black"),
      axis.text.x.top    = element_text(size=10,color = "black"),
      axis.text.x.bottom    = element_text(size=10,color = "black"),
      axis.title.x.bottom    = element_text(size=12,color = "black"),
      legend.position = c(0.63,0.15))+
      coord_flip()+
   scale_x_discrete(labels = label_vec)+theme(axis.text.y     = element_markdown())

plot.unique.biomes
```


Single−biome OTUs
0
1000
2000
freshwater−sediments
saline−water
animal−proximalgut
saline−sediments
freshwater−water
animal−distalgut
soil−agricultural
animal−skin
animal−oral
plant−rhizosphere
freshwater−biofilm
soil−forest
soil−desert
saline−biofilm
soil−grassland
plant−phyllosphere
freshwater−peatlands(water)
plant−endosphere
airborne
animal−urogenital
soil−tundra
animal−respiratory
plant−spermosphere
freshwater−peatlands(peat/bog)
Free−living
Host−associated
0
1
2
3
4
5
% Biome species pool