Phyloseq

Laura Grace Schaerer

4/2/2019

# packages used

library(phyloseq)  
library(csv)  
library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.3.0   
## ✔ tibble 2.0.1 ✔ dplyr 0.8.0.1  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1   
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(vegan)

## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.5-2

# set seed so workflow can be repeated

set.seed(81)

# Phyloseq

### load in meta data

meta <- as.csv("/home/lgschaer/DARPA/phyloseq/input\_phyloseq/meta.F.csv", row.names = 1, header = TRUE, sep = ",", check.names = TRUE, stringsAsFactors = TRUE)  
head(meta)

## sampletype location region filtertype datatype  
## BUS1 openwater Busan Asia pre openwater  
## BUS10 openwater Busan Asia pre openwater  
## BUS10post openwater Busan Asia post openwater  
## BUS11 openwater Busan Asia pre openwater  
## BUS11post openwater Busan Asia post openwater  
## BUS12 openwater Busan Asia pre openwater

### add title to SampleID column

meta2 <- rownames\_to\_column(meta, var = "SampleID")  
head(meta2)

## SampleID sampletype location region filtertype datatype  
## 1 BUS1 openwater Busan Asia pre openwater  
## 2 BUS10 openwater Busan Asia pre openwater  
## 3 BUS10post openwater Busan Asia post openwater  
## 4 BUS11 openwater Busan Asia pre openwater  
## 5 BUS11post openwater Busan Asia post openwater  
## 6 BUS12 openwater Busan Asia pre openwater

### load in data sequence table

sequence\_table <- readRDS("/home/datadrop/techtmann\_lab/port\_microbiome/phyloseq\_ready\_files/seqtabmergedNoC.rds")  
colnames(sequence\_table) <- NULL  
sequence\_table <- as.data.frame(sequence\_table)  
sequence\_table[1:8,1:8]

## V1 V2 V3 V4 V5 V6 V7 V8  
## BUS1 0 0 0 122 0 0 0 0  
## BUS10 22 0 0 70 0 0 0 0  
## BUS10post 0 0 0 140 0 0 0 0  
## BUS11 0 0 5 144 0 0 0 6  
## BUS11post 0 0 0 262 0 0 0 0  
## BUS12 20 0 18 140 0 0 0 12  
## BUS12post 0 0 0 179 0 0 0 0  
## BUS13 0 0 7 39 0 0 0 0

sequence\_table <- rownames\_to\_column(sequence\_table, var = "SampleID")  
sequence\_table[1:10,1:10]

## SampleID V1 V2 V3 V4 V5 V6 V7 V8 V9  
## 1 BUS1 0 0 0 122 0 0 0 0 136  
## 2 BUS10 22 0 0 70 0 0 0 0 90  
## 3 BUS10post 0 0 0 140 0 0 0 0 101  
## 4 BUS11 0 0 5 144 0 0 0 6 90  
## 5 BUS11post 0 0 0 262 0 0 0 0 178  
## 6 BUS12 20 0 18 140 0 0 0 12 118  
## 7 BUS12post 0 0 0 179 0 0 0 0 149  
## 8 BUS13 0 0 7 39 0 0 0 0 26  
## 9 BUS13post 0 0 0 96 0 0 0 0 108  
## 10 BUS14 7 0 6 30 0 0 0 5 70

### join metadata to sequence table so we can filter samples

seq\_table <- meta2 %>%   
 left\_join(sequence\_table, by = "SampleID") %>%  
 column\_to\_rownames(var = "SampleID")  
seq\_table[1:10,1:10]

## sampletype location region filtertype datatype V1 V2 V3 V4 V5  
## BUS1 openwater Busan Asia pre openwater 0 0 0 122 0  
## BUS10 openwater Busan Asia pre openwater 22 0 0 70 0  
## BUS10post openwater Busan Asia post openwater 0 0 0 140 0  
## BUS11 openwater Busan Asia pre openwater 0 0 5 144 0  
## BUS11post openwater Busan Asia post openwater 0 0 0 262 0  
## BUS12 openwater Busan Asia pre openwater 20 0 18 140 0  
## BUS12post openwater Busan Asia post openwater 0 0 0 179 0  
## BUS13 openwater Busan Asia pre openwater 0 0 7 39 0  
## BUS13post openwater Busan Asia post openwater 0 0 0 96 0  
## BUS14 openwater Busan Asia pre openwater 7 0 6 30 0

### subset to only include sampletypes we want

subset\_all <- seq\_table %>%  
 subset(sampletype=="openwater"|sampletype=="bilge"|sampletype=="swab")  
  
subset\_all$sampletype <- NULL  
subset\_all$location <- NULL  
subset\_all$region <- NULL  
subset\_all$filtertype <- NULL  
subset\_all$datatype <- NULL  
  
subset\_all[1:10,1:10]

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10  
## BUS1 0 0 0 122 0 0 0 0 136 0  
## BUS10 22 0 0 70 0 0 0 0 90 0  
## BUS10post 0 0 0 140 0 0 0 0 101 0  
## BUS11 0 0 5 144 0 0 0 6 90 0  
## BUS11post 0 0 0 262 0 0 0 0 178 0  
## BUS12 20 0 18 140 0 0 0 12 118 0  
## BUS12post 0 0 0 179 0 0 0 0 149 0  
## BUS13 0 0 7 39 0 0 0 0 26 0  
## BUS13post 0 0 0 96 0 0 0 0 108 0  
## BUS14 7 0 6 30 0 0 0 5 70 0

### trim sequence table

m <- (colSums(subset\_all, na.rm=TRUE) != 0) #T if colSum is not 0, F otherwise  
seqtaballtrim\_nz <- subset\_all[, m] #all the non-zero columns  
seqtaballtrim\_z <- subset\_all[, !m] #all the zero columns  
seqtaballtrim\_nz\_2<-seqtaballtrim\_nz[,colSums(seqtaballtrim\_nz>0)>1,drop=FALSE]  
  
sequence\_table\_all <- seqtaballtrim\_nz\_2 #change to more intuitive name

### import taxonomy table:

taxa\_table <- readRDS("/home/datadrop/techtmann\_lab/port\_microbiome/phyloseq\_ready\_files/taxasp.rds")  
taxa\_table\_t<-t(taxa\_table)  
colnames(taxa\_table\_t) <- NULL  
taxa\_table\_re\_t<-t(taxa\_table\_t)  
taxa\_table\_re\_t[1:5,1:5]

## Kingdom Phylum Class Order   
## [1,] "Bacteria" "Cyanobacteria" "Oxyphotobacteria" "Synechococcales"   
## [2,] "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rhodobacterales"   
## [3,] "Bacteria" "Cyanobacteria" "Oxyphotobacteria" "Synechococcales"   
## [4,] "Bacteria" "Bacteroidetes" "Bacteroidia" "Flavobacteriales"  
## [5,] "Bacteria" "Proteobacteria" "Alphaproteobacteria" "SAR11\_clade"   
## Family   
## [1,] "Cyanobiaceae"   
## [2,] "Rhodobacteraceae"  
## [3,] "Cyanobiaceae"   
## [4,] "Cryomorphaceae"   
## [5,] "Clade\_III"

## creating phyloseq object

colnames(sequence\_table\_all) <- NULL

### phyloseq all samples

sdata\_all = sample\_data(meta)  
sdata\_all[1:5,1:5]

## sampletype location region filtertype datatype  
## BUS1 openwater Busan Asia pre openwater  
## BUS10 openwater Busan Asia pre openwater  
## BUS10post openwater Busan Asia post openwater  
## BUS11 openwater Busan Asia pre openwater  
## BUS11post openwater Busan Asia post openwater

seqtab\_all = otu\_table(sequence\_table\_all, taxa\_are\_rows = FALSE)  
taxa\_all = tax\_table(taxa\_table\_re\_t)  
dim(taxa\_all)

## [1] 157986 7

taxa\_all[1:7,1:7]

## Taxonomy Table: [7 taxa by 7 taxonomic ranks]:  
## Kingdom Phylum Class Order   
## sp1 "Bacteria" "Cyanobacteria" "Oxyphotobacteria" "Synechococcales"   
## sp2 "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rhodobacterales"   
## sp3 "Bacteria" "Cyanobacteria" "Oxyphotobacteria" "Synechococcales"   
## sp4 "Bacteria" "Bacteroidetes" "Bacteroidia" "Flavobacteriales"  
## sp5 "Bacteria" "Proteobacteria" "Alphaproteobacteria" "SAR11\_clade"   
## sp6 "Bacteria" "Proteobacteria" "Alphaproteobacteria" "Rhodobacterales"   
## sp7 "Bacteria" "Actinobacteria" "Actinobacteria" "Frankiales"   
## Family Genus Species   
## sp1 "Cyanobiaceae" "Cyanobium\_PCC-6307" NA   
## sp2 "Rhodobacteraceae" "Planktomarina" NA   
## sp3 "Cyanobiaceae" "Cyanobium\_PCC-6307" NA   
## sp4 "Cryomorphaceae" NA NA   
## sp5 "Clade\_III" NA NA   
## sp6 "Rhodobacteraceae" "Planktomarina" "temperata"  
## sp7 "Sporichthyaceae" "hgcI\_clade" NA

seqtab\_all[1:10,1:10]

## OTU Table: [10 taxa and 10 samples]  
## taxa are columns  
## sp1 sp2 sp3 sp4 sp5 sp6 sp7 sp8 sp9 sp10  
## BUS1 0 0 0 122 0 0 0 0 136 0  
## BUS10 22 0 0 70 0 0 0 0 90 0  
## BUS10post 0 0 0 140 0 0 0 0 101 0  
## BUS11 0 0 5 144 0 0 0 6 90 0  
## BUS11post 0 0 0 262 0 0 0 0 178 0  
## BUS12 20 0 18 140 0 0 0 12 118 0  
## BUS12post 0 0 0 179 0 0 0 0 149 0  
## BUS13 0 0 7 39 0 0 0 0 26 0  
## BUS13post 0 0 0 96 0 0 0 0 108 0  
## BUS14 7 0 6 30 0 0 0 5 70 0

phyloseq\_object\_all = phyloseq(seqtab\_all, taxa\_all, sdata\_all)  
phyloseq\_object\_all

## phyloseq-class experiment-level object  
## otu\_table() OTU Table: [ 29372 taxa and 1498 samples ]  
## sample\_data() Sample Data: [ 1498 samples by 5 sample variables ]  
## tax\_table() Taxonomy Table: [ 29372 taxa by 7 taxonomic ranks ]

## normalize data

### Delete samples with a mean of less than 1000

samplesover1000\_all <- subset\_samples(phyloseq\_object\_all, sample\_sums(phyloseq\_object\_all) > 1000)

### Check if there are OTUs with no counts, if so how many?

any(taxa\_sums(samplesover1000\_all) == 0)

## [1] FALSE

sum(taxa\_sums(samplesover1000\_all) == 0)

## [1] 0

### Prune OTUs with no counts

prune\_samplesover1000\_all <- prune\_taxa(taxa\_sums(samplesover1000\_all) > 0, samplesover1000\_all)  
any(taxa\_sums(prune\_samplesover1000\_all) == 0)

## [1] FALSE

### make sure seed is set the same each time, set to 81 here

rarefy\_samplesover1000\_all <- rarefy\_even\_depth(prune\_samplesover1000\_all, rngseed= 81, sample.size = min(sample\_sums(prune\_samplesover1000\_all)))

## `set.seed(81)` was used to initialize repeatable random subsampling.

## Please record this for your records so others can reproduce.

## Try `set.seed(81); .Random.seed` for the full vector

## ...

## 2913OTUs were removed because they are no longer   
## present in any sample after random subsampling

## ...

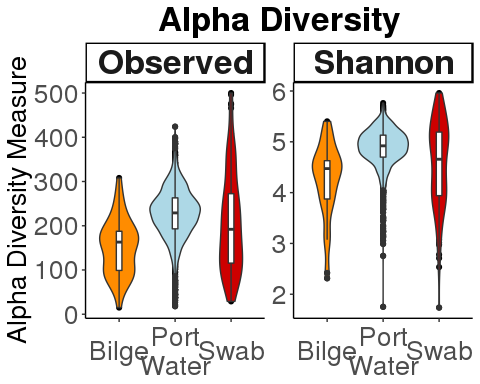
# Alpha diversity

### rename phyloseq object

pso <- rarefy\_samplesover1000\_all

### Violin plot of alpha diversity Observed OTUs and Shannon Diversity

pso %>% #phyloseq object  
 plot\_richness(  
 x = "sampletype", #compare diversity of sampletype  
 measures = c("Observed", "Shannon")) + #choose diversity measures  
 geom\_violin(aes(fill = sampletype), show.legend = FALSE)+ #make violin plot, set fill aes to sampletype  
 geom\_boxplot(width=0.1) + #add boxplot, set width  
 theme\_classic()+ #change theme to classic  
 xlab(NULL)+ #no label on x-axis  
 theme(axis.text.y.left = element\_text(size = 20), #adjust y-axis text  
 axis.text.x = element\_text(size = 20, vjust = .5), #adjust x-axis label position  
 axis.title.y = element\_text(size = 20))+ #adjust y-axis title  
 theme(strip.text = element\_text(face = "bold", size = 25))+ #adjust headings  
 scale\_fill\_manual(values = c("swab" = "red3", "openwater" = "lightblue", "bilge" = "darkorange"))+ #set fill colors  
 scale\_x\_discrete( #change x-axis labels  
 breaks = c("bilge", "openwater", "swab"),   
 labels = c("Bilge", "Port\nWater", "Swab")) +   
 ggtitle("Alpha Diversity") + #add title  
 theme(plot.title=element\_text(size = 25, face = "bold", hjust =.5)) #change title size, face and position



### data manipulation

richness <- pso %>%  
 estimate\_richness(measures = c("Observed", "Shannon")) %>% #specify which measures  
 rownames\_to\_column(var = "SampleID") #add column name to SampleID column  
  
meta2 <- meta %>%  
 rownames\_to\_column(var = "SampleID") #add column name to SampleID column

### finding means and medians

rich\_summary <- richness %>%  
 left\_join(meta2, by = "SampleID") %>% #join to meta data  
 filter(sampletype=="swab"|sampletype=="bilge"|sampletype=="openwater") %>% #filter only sample types we want  
 group\_by(sampletype) %>% #group by sample type  
 summarize( #add columns for stat summaries  
 rangeObs = max(Observed) - min(Observed),  
 meanObserved = mean(Observed),  
 medianObserved = median(Observed),  
 rangeShan = max(Shannon) - min(Shannon),  
 meanShannon = mean(Shannon),  
 medianShannon = median(Shannon)  
 )  
head(rich\_summary) #check that we have the columns we want

## # A tibble: 3 x 7  
## sampletype rangeObs meanObserved medianObserved rangeShan meanShannon  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 bilge 293 152. 163 3.09 4.24  
## 2 openwater 406 226. 230. 4.01 4.87  
## 3 swab 471 207. 192 4.22 4.56  
## # … with 1 more variable: medianShannon <dbl>

# ANOVA

set.seed(81)  
  
pso #phyloseq object

## phyloseq-class experiment-level object  
## otu\_table() OTU Table: [ 26459 taxa and 1469 samples ]  
## sample\_data() Sample Data: [ 1469 samples by 5 sample variables ]  
## tax\_table() Taxonomy Table: [ 26459 taxa by 7 taxonomic ranks ]

### calculate alpha diversity measures

data\_all <- pso %>%  
 estimate\_richness(measures = c("Observed", "Shannon")) %>% #estimate richness  
 cbind(sample\_data(pso)) #add sample data from phyloseq object  
head(data\_all)

## Observed Shannon sampletype location region filtertype  
## BUS1 274 4.889141 openwater Busan Asia pre  
## BUS10 146 4.386433 openwater Busan Asia pre  
## BUS10post 249 5.162542 openwater Busan Asia post  
## BUS11 240 4.924552 openwater Busan Asia pre  
## BUS11post 272 5.192763 openwater Busan Asia post  
## BUS12 249 4.842337 openwater Busan Asia pre  
## datatype  
## BUS1 openwater  
## BUS10 openwater  
## BUS10post openwater  
## BUS11 openwater  
## BUS11post openwater  
## BUS12 openwater

### Observed ANOVA

anova\_obs <- aov(Observed ~ sampletype, data\_all)  
summary(anova\_obs)

## Df Sum Sq Mean Sq F value Pr(>F)   
## sampletype 2 218923 109461 24.56 3.24e-11 \*\*\*  
## Residuals 1466 6534839 4458   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Shannon ANOVA

anova\_shan <- aov(Shannon ~ sampletype, data\_all)  
summary(anova\_shan)

## Df Sum Sq Mean Sq F value Pr(>F)   
## sampletype 2 22.5 11.27 51.31 <2e-16 \*\*\*  
## Residuals 1466 322.0 0.22   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### tukeyHSD, post-hoc

TukeyHSD(anova\_obs)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Observed ~ sampletype, data = data\_all)  
##   
## $sampletype  
## diff lwr upr p adj  
## openwater-bilge 74.24952 47.42403 101.07501 0.0000000  
## swab-bilge 55.24099 25.00269 85.47930 0.0000575  
## swab-openwater -19.00852 -34.23846 -3.77859 0.0096904

TukeyHSD(anova\_shan)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Shannon ~ sampletype, data = data\_all)  
##   
## $sampletype  
## diff lwr upr p adj  
## openwater-bilge 0.6306499 0.4423468 0.8189530 0.0000000  
## swab-bilge 0.3232648 0.1110052 0.5355244 0.0010607  
## swab-openwater -0.3073851 -0.4142925 -0.2004776 0.0000000

# Beta diversity

## PCoA Plot

### ordination

all\_pcoa <- ordinate(  
 physeq = pso,   
 method = "PCoA",   
 distance = "bray"  
)

### plot

plot\_ordination(  
 physeq = pso, #phyloseq object  
 ordination = all\_pcoa, #ordination  
 shape = "region") + #set shape to region  
 geom\_point(aes(fill = sampletype), size = 3) + #sets fill color to sampletype  
 scale\_fill\_manual( #sets the color to fill your points  
 values = c("swab" = "red3", "openwater" = "lightblue", "bilge" = "darkorange"),   
 labels = c("bilge" ="Bilge", "openwater" = "Port Water", "swab" = "Swab")) +  
 scale\_shape\_manual( #uses a set of fillable shapes  
 values = c("East" = 21, "West" = 22, "Europe" = 23, "Asia" = 24, "Lake\_Summer" = 25, "Lake\_Fall" = 25),  
 labels = c("Asia", "Europe", "Lake Summer", "Lake Fall", "East Coast", "West Coast"))+  
 theme\_classic() + #changes theme, removes grey background  
 theme(   
 legend.text = element\_text(size = 20), #changes legend size  
 legend.title = element\_blank(), #removes legend title  
 legend.position = "right", #positions legend on plot  
 legend.justification = c(1,0), #changes legend justification  
 legend.background = element\_rect(fill = "white", color = "black"))+ #adds black boarder around legend  
 theme(axis.text.y.left = element\_text(size = 20),  
 axis.text.x = element\_text(size = 20),  
 axis.title.x = element\_text(size = 20),  
 axis.title.y = element\_text(size = 20))+  
 guides(fill = guide\_legend(override.aes = list(shape = 21))) #fills legend points based on the fill command

