

Chapter 3: ITS data analysis using the R package PhyloSeq

Emily Giroux

4/17/2019

Load the relevant images, then save this chapter's image as a separate image to retain environment data specific to the ITS processing and analysis workflow.

```
# Set the name for this chapter's image:
chptImage    <- "ecobiomics_ITS_analysis.RData"

# Save this chapter's image:
save.image(paste(imageDirPath, chptImage, sep = ""))
```

When re-starting a session, you can quickly load up the image by running the chunk below:

```
sharedPath <- "/isilon/cfia-ottawa-fallowfield/users/girouxeml/PIRL_working_directory/"
analysis   <- "ecobiomics/"
sharedPathAn <- paste(sharedPath, analysis, sep = "")
imageDirPath <- "/home/CFIA-ACIA/girouxeml/GitHub_Repos/r_environments/ecobiomics/"
chptImageA <- "ecobiomics_ITS.RData"
load(paste(imageDirPath, chptImageA, sep = ""))
chptImage    <- "ecobiomics_ITS_analysis.RData"
save.image(paste(imageDirPath, chptImage, sep = ""))
```

Let's get familiar with our phyloseq object created at the end of our sequencing sample processing chapter:

Below I am using the `ps`, rather than `phySeq` objects. Recall from the last chunks of Chapter 2, the `phySeq` object is a `phyloseq` object without the `fitGTRtreeinfo`, while the `ps` object was created leveraging the `fitGTR` information. The `phySeq` object can be used instead of the `ps` object if the `optim.pml` command wasn't run.

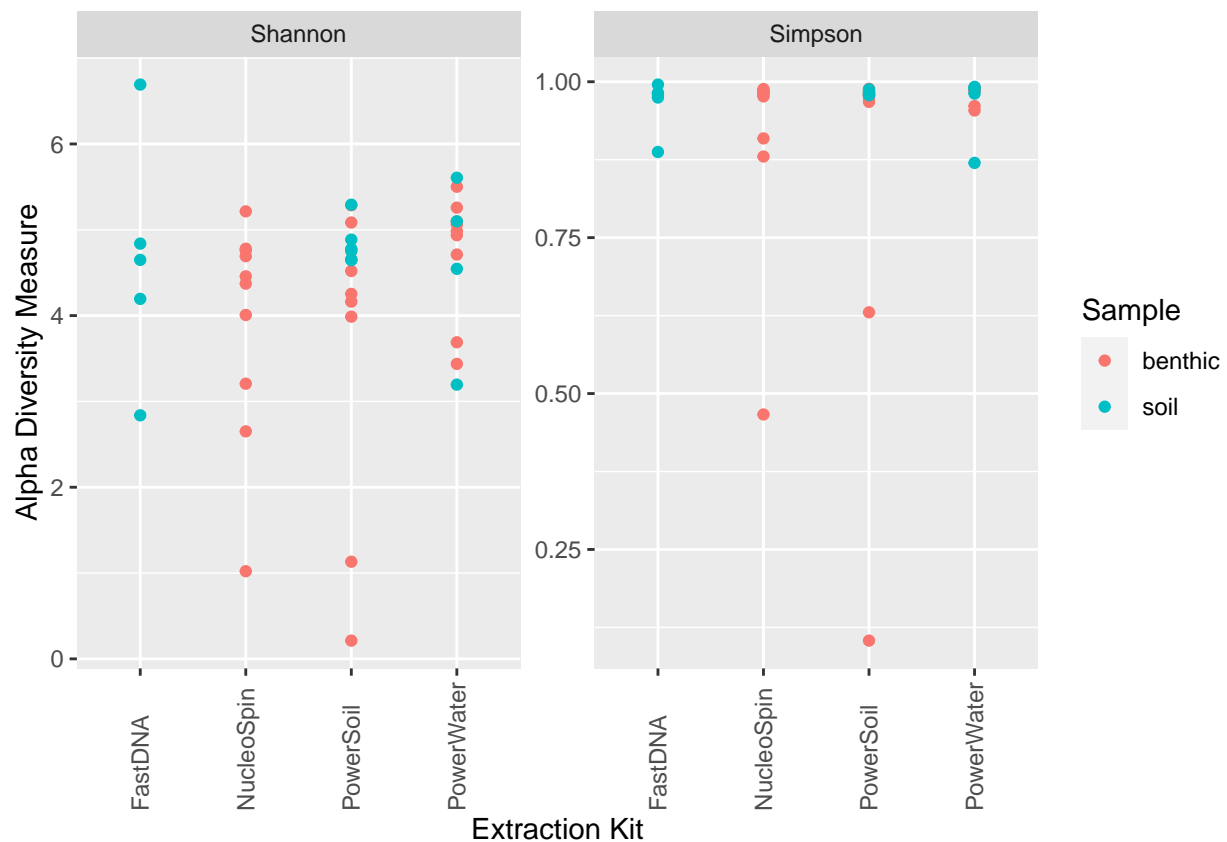
Here is if we filter based on having to know the species:

```
library("phyloseq")
# table(tax_table(ps)[, "Genus"], exclude = NULL)
t <- table(tax_table(ps)[, "Genus"], exclude = NULL)
# head(t[order(-t)])
t[order(-t)][2:10]
```

g__Acidea	g__Bannoa	g__Mortierella	g__Trichoderma	g__Venturia
1429	512	197	99	60
g__Malassezia	g__Tetracladium	g__Myrmecridium	g__Alatospora	
57	50	46	41	

Visualize alpha-diversity, phylum:

```
library("phyloseq")
library("ggplot2")
library("cowplot")
plot_richness(ps,
  x = "ExtractionKit",
  measures = c("Shannon", "Simpson"),
  color = "Sample") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Extraction Kit")
```



Prevalence evaluation for species:

```
library("phyloseq")
prevDf <- apply(X = otu_table(ps),
  MARGIN = ifelse(taxa_are_rows(ps),
```

```

                                yes = 1, no = 2),
    FUN = function(x){sum(x>0)})

prevDf <- data.frame(Prevalence = prevDf,
                    TotalAbundance = taxa_sums(ps),
                    tax_table(ps))

prevalenceTblSpp <- plyr::ddply(prevDf, "Species",
                                function(df1){
                                    cbind(mean(df1$Prevalence),
                                            sum(df1$Prevalence))})
colnames(prevalenceTblSpp) <- c("Species", "Mean", "Sum")
head(prevalenceTblSpp)
head(prevalenceTblSpp[order(-prevalenceTblSpp[,3]),])

```

	Species	Mean	Sum
1	s__abeliceae	5.000000	5
2	s__acerophilum	1.000000	1
3	s__acicola	1.000000	2
4	s__aculeatus	2.666667	8
5	s__acuminata	3.400000	102
6	s__adusta	1.000000	4

	Species	Mean	Sum
427	<NA>	1.422086	8834
124	s__extrema	1.000000	1429
271	s__ogasawarensis	1.000000	512
349	s__schulzeri	3.465116	149
172	s__harzianum	1.359551	121
295	s__piceae-abietis	6.562500	105

Prevalence evaluation for phyla:

```

library("phyloseq")
prevDf <- apply(X = otu_table(ps),
               MARGIN = ifelse(taxa_are_rows(ps),
                               yes = 1, no = 2),
               FUN = function(x){sum(x>0)})

prevDf <- data.frame(Prevalence = prevDf,
                    TotalAbundance = taxa_sums(ps),
                    tax_table(ps))

prevalenceTblPhyla <- plyr::ddply(prevDf, "Phylum",
                                   function(df1){
                                       cbind(mean(df1$Prevalence),
                                               sum(df1$Prevalence))})

```

```
colnames(prevalenceTblPhyla) <- c("Phylum", "Mean", "Sum")
prevalenceTblPhyla[order(-prevalenceTblPhyla[,3]),]
```

	Phylum	Mean	Sum
2	p__Ascomycota	1.700448	9480
4	p__Basidiomycota	1.189139	2584
14	<NA>	1.164004	2186
10	p__Mortierellomycota	1.777228	359
6	p__Chytridiomycota	1.292857	181
7	p__Glomeromycota	1.111111	30
11	p__Mucoromycota	1.230769	16
13	p__Rozellomycota	1.428571	10
1	p__Aphelidiomycota	1.500000	6
12	p__Olpidiomycota	1.000000	6
5	p__Blastocladiomycota	1.666667	5
9	p__Monoblepharomycota	1.000000	4
8	p__Kickxellomycota	1.000000	3
3	p__Basidiobolomycota	1.000000	1

From the above calculations, there are a few low-abundance Phylas that appear in less than 10 samples:

Aphelidiomycota
 Basidiobolomycota
 Blastocladiomycota
 Kickxellomycota
 Monoblepharomycota
 Olpidiomycota
 Rozellomycota

Filter entries with unidentified Phylum, or those phyla that appear in less than 10 samples:

```
library("phyloseq")
phylas <- subset(prevalenceTblPhyla, prevalenceTblPhyla$Sum < 10)
ps1 <- subset_taxa(ps, !Phylum %in% phylas$Phylum)
rank_names(ps1)
head(table(tax_table(ps1)[, "Phylum"], exclude = NULL))
head(table(tax_table(ps1)[, "Species"], exclude = NULL))

t2 <- table(tax_table(ps1)[, "Species"], exclude = NULL)
t2[order(-t2)][2:10]
```

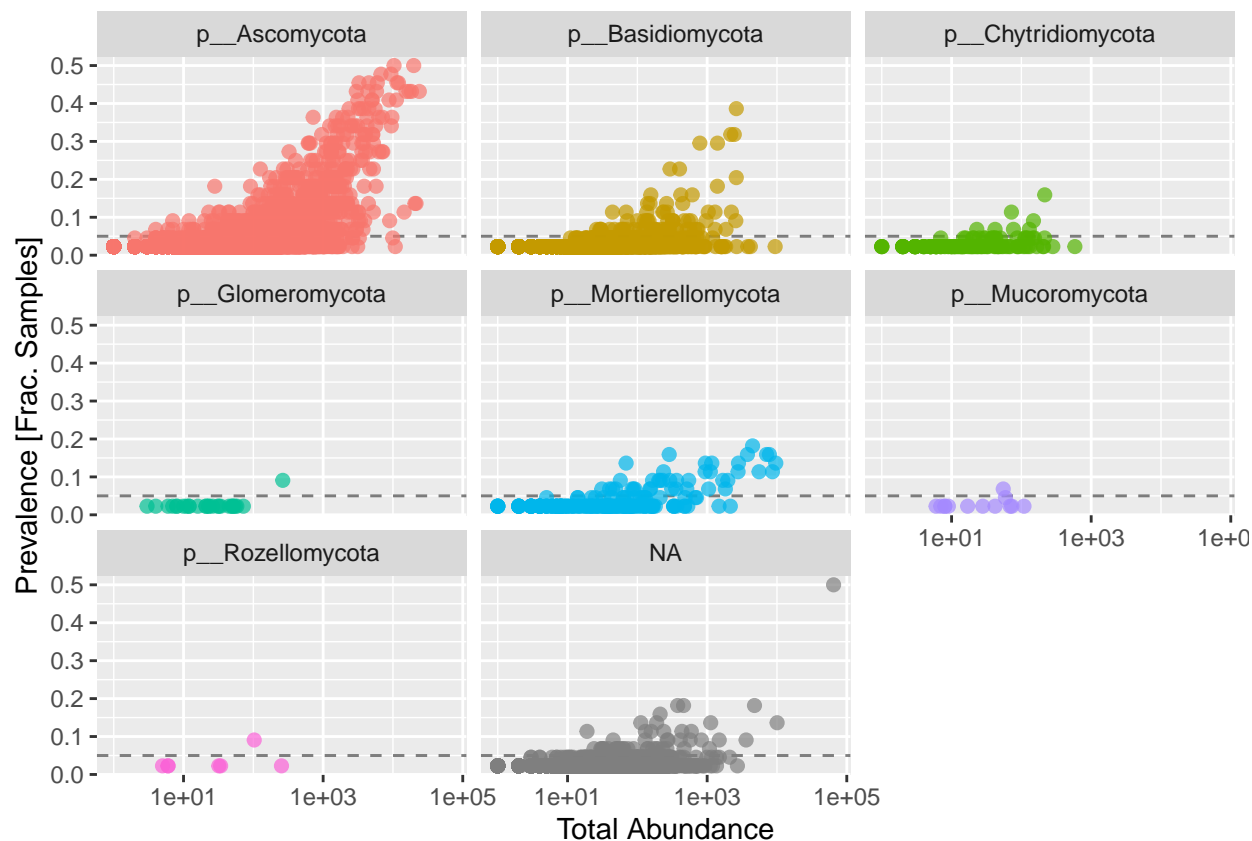
```
[1] "Kingdom" "Phylum" "Class" "Order" "Family" "Genus" "Species"
```

p__Ascomycota	p__Basidiomycota	p__Chytridiomycota
5575	2173	140
p__Glomeromycota	p__Mortierellomycota	p__Mucoromycota
27	202	13

s__abeliceae	s__acerophilum	s__acicola	s__aculeatus	s__acuminata
1	1	2	3	30
s__adusta				
4				
s__extrema	s__ogasawarensis	s__harzianum	s__minutissima	
1429	512	89	49	
s__schulzeri	s__elongata	s__hyalina	s__acuminata	
43	33	33	30	
s__aquaticus				
30				

Plot Phylum:

```
library("phyloseq")
library("ggplot2")
prevDf1 <- subset(prevDf, Phylum %in% get_taxa_unique(ps1, "Phylum"))
ggplot(prevDf1, aes(TotalAbundance, Prevalence / nsamples(ps1),
                    color = Phylum)) +
  geom_hline(yintercept = 0.05, alpha = 0.5, linetype = 2) +
  geom_point(size = 2, alpha = 0.7) +
  scale_x_log10() + xlab("Total Abundance") + ylab("Prevalence [Frac. Samples]") +
  facet_wrap(~Phylum) + theme(legend.position = "none")
```



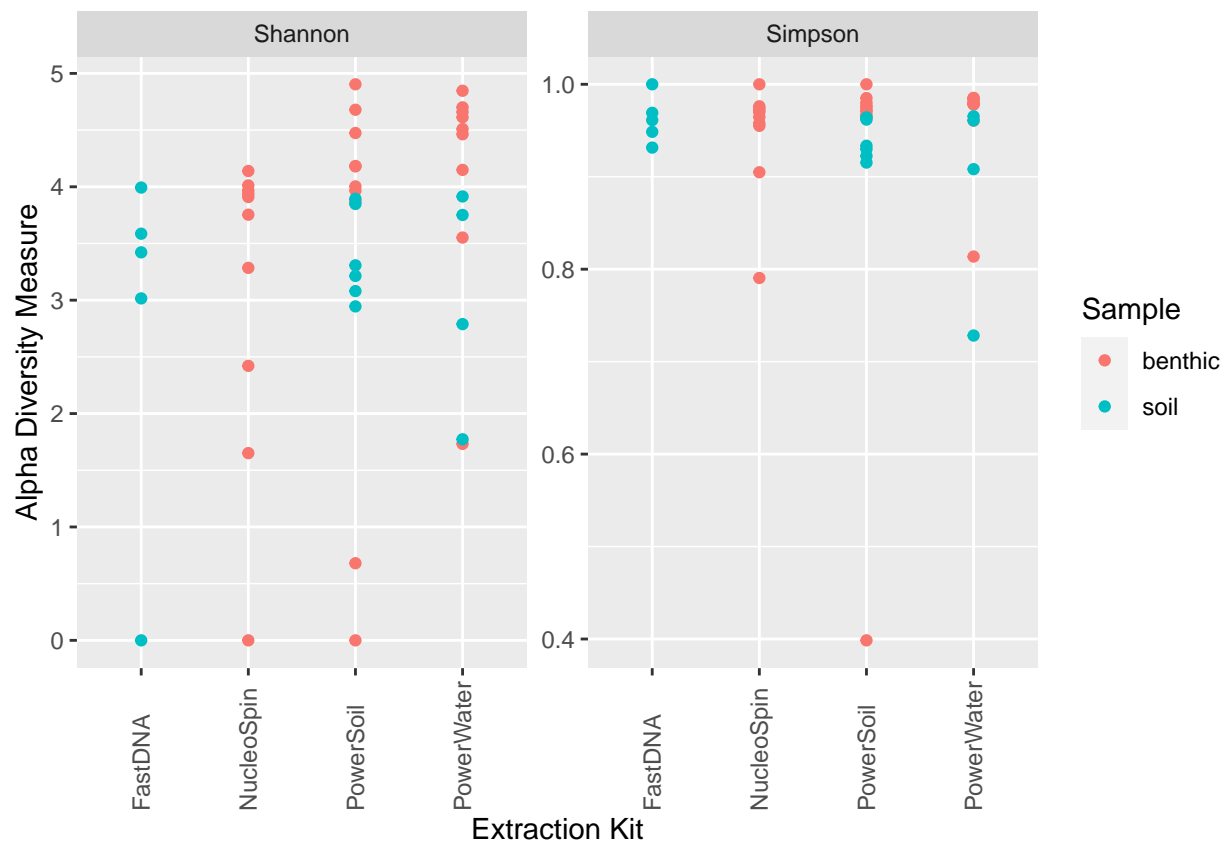
Each point in the above plots is a different taxa, Phylum.

```
prevalenceThreshold = 0.05*nsamples(ps1)
prevalenceThreshold
```

```
[1] 2.2
```

The taxa with a prevalence threshold less than the one set in the above chunk are removed using `prune_taxa` and put into a new phyloseq object, `ps2`, and we look at the resulting richness plot:

```
library("phyloseq")
keepTaxa <- rownames(prevDf1)[(prevDf1$Prevalence >= prevalenceThreshold)]
ps2 <- prune_taxa(keepTaxa, ps1)
rank_names(ps2)
# table(tax_table(ps2)[, "Phylum"], exclude = NULL)
# table(tax_table(ps2)[, "Species"], exclude = NULL)
plot_richness(ps2,
  x = "ExtractionKit",
  measures = c("Shannon", "Simpson"),
  color = "Sample") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(name = "Extraction Kit")
```



```
[1] "Kingdom" "Phylum" "Class" "Order" "Family" "Genus" "Species"
```

Phylum - curiosity:

Curious about the mean and sum prevalence after keeping only taxa passing prevalenceThreshold:

```
library("phyloseq")
prevDf2 <- apply(X = otu_table(ps2),
  MARGIN = ifelse(taxa_are_rows(ps2), yes = 1, no = 2),
  FUN = function(x){sum(x > 0)})

prevDf2 <- data.frame(Prevalence = prevDf2,
  TotalAbundance = taxa_sums(ps2),
  tax_table(ps2))

prevPhylatblThreshold <- plyr::ddply(prevDf2, "Phylum",
  function(df1){
    cbind(mean(df1$Prevalence),
      sum(df1$Prevalence))})
colnames(prevPhylatblThreshold) <- c("Phylum", "Mean", "Sum")
prevPhylatblThreshold
```

```
Phylum    Mean    Sum
```

```

1      p__Ascomycota 6.277439 4118
2      p__Basidiomycota 4.890411 357
3      p__Chytridiomycota 4.000000 28
4      p__Glomeromycota 4.000000 4
5      p__Mortierellomycota 4.450000 178
6      p__Mucoromycota 3.000000 3
7      p__Rozellomycota 4.000000 4
8      <NA> 4.510638 212

```

Note: I am assuming that the mean is the mean number of times the phylum was seen in a sample for all samples in which it was seen, while the sum is the total times it was seen across all samples. Ascomycota was seen a total of 1,857 times, with about 6 occurrences per sample, while Mucoromycota was seen 3 times total and the mean is simply 3 because when it was seen, it was all three in one sample.

Number of unique phyla, genera and species, across all samples:

```

library("phyloseq")
uniqueClasses <- c("Phylum", "Genus", "Species")
for(i in unique(uniqueClasses))
  cat(cat(i), length(phyloseq::get_taxa_unique(ps2, taxonomic.rank = i)), "\n")

```

```

Phylum 8
Genus 139
Species 132

```

The `tax_glom` function of `phyloseq` merges species that have the same taxonomy at certain taxonomic rank, using categorical data. The `tip_glom` function agglomerates tree tips into a single taxa if they are separated by less than a height specified by `h`.

```

library("phyloseq")
ps3 <- phyloseq::tax_glom(ps2, "Genus", NArm = TRUE)
h1 = 0.4
ps4 <- phyloseq::tip_glom(ps2, h = h1)

```

Below we will look at plots of our trees before agglomeration, with agglomeration using `tax_glom`, and with agglomeration by tip separation using `tip_glom`:

```

library("phyloseq")
library("ggplot2")
library("gridExtra")
multiPlotTitleTextSize = 15
p2Tree <- phyloseq::plot_tree(ps2, method = "treeonly",
                             ladderize = "left",
                             title = "Before Agglomeration") +
  ggplot2::theme(plot.title = element_text(size = multiPlotTitleTextSize))

```



```

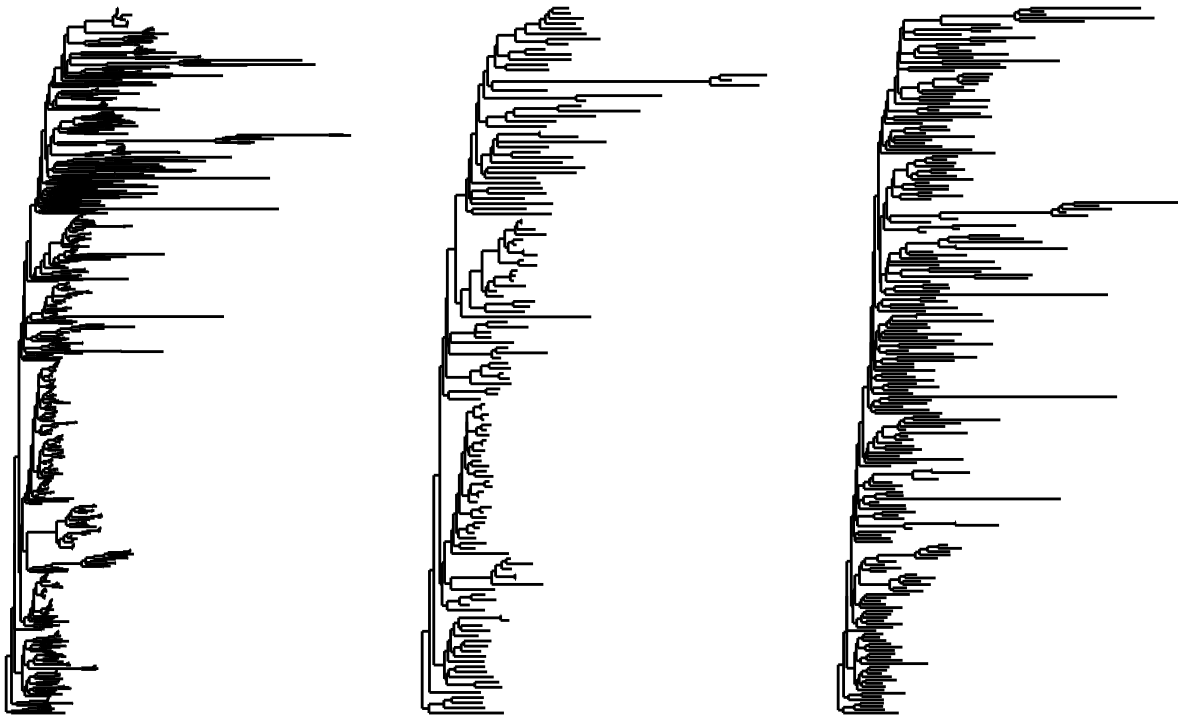
p3Tree <- phyloseq::plot_tree(ps3, method = "treeonly",
                             ladderize = "left", title = "By Genus") +
  ggplot2::theme(plot.title = element_text(size = multiPlotTitleTextSize))

p4Tree <- phyloseq::plot_tree(ps4, method = "treeonly",
                             ladderize = "left", title = "By Height") +
  ggplot2::theme(plot.title = element_text(size = multiPlotTitleTextSize))
gridExtra::grid.arrange(nrow = 1, p2Tree, p3Tree, p4Tree)

```

Before Agglomeration By Genus

By Height



From here on we will continue using the `ps2` phyloseq object, that has had the 'NA', low-abundance, and prevalence threshold filters applied.

```

library("phyloseq")
phyloseq::plot_bar(ps2,
                   x = "sample_Sample",
                   fill = "Phylum",
                   facet_grid = ~ExtractionKit)

```

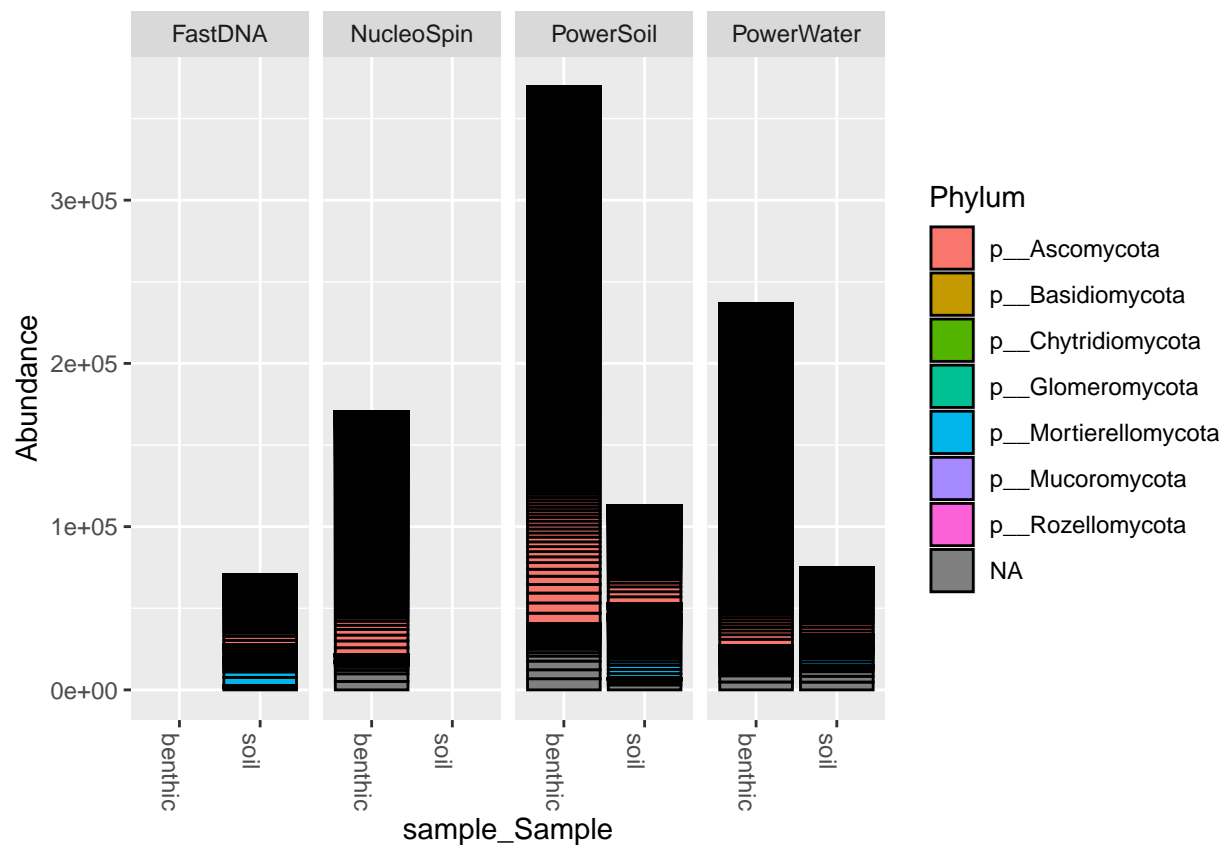
Warning in `psmelt(physeq)`: The sample variables:

Sample

have been renamed to:

sample_Sample

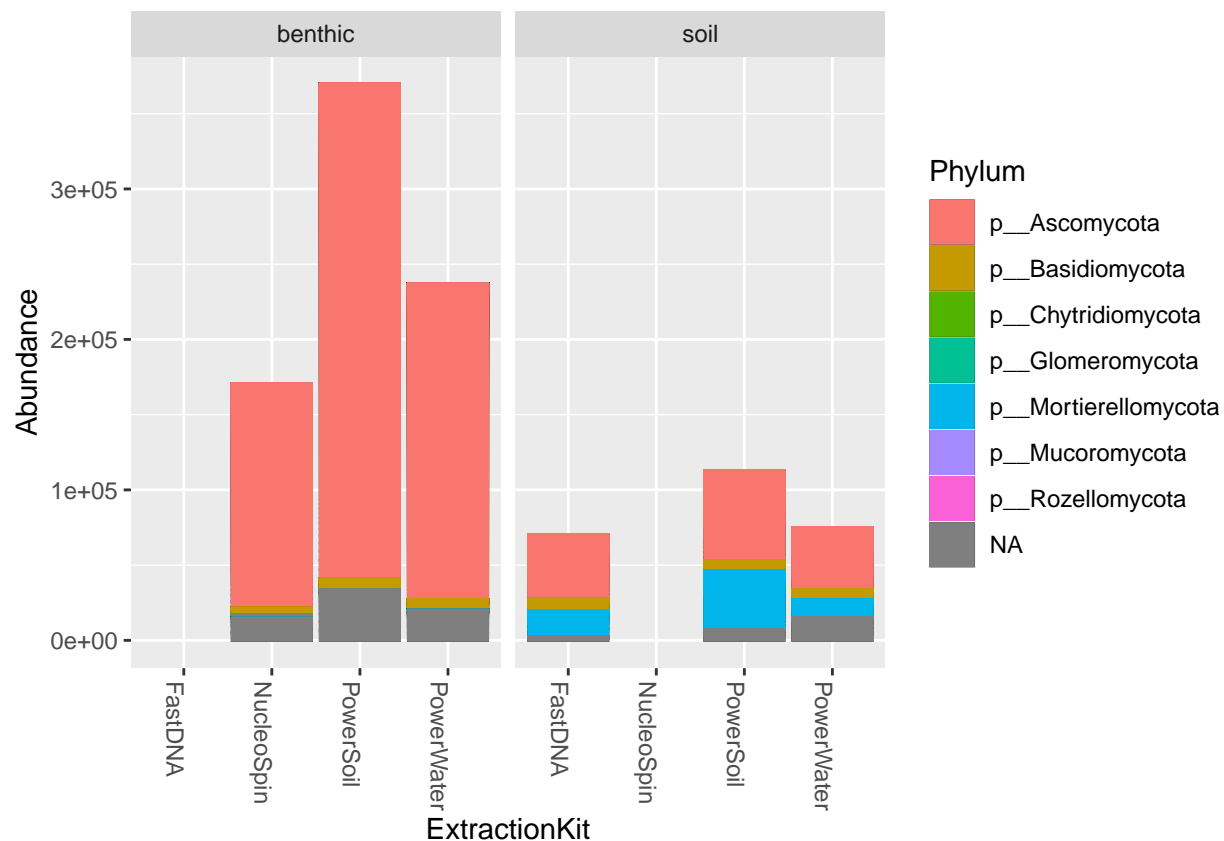
to avoid conflicts with special phyloseq plot attribute names.



```
library("phyloseq")
library("ggplot2")
plotPhylum <- phyloseq::plot_bar(ps2, x = "ExtractionKit", fill = "Phylum",
                                   facet_grid = ~sample_Sample) +
  ylab("Abundance") +
  geom_bar(aes(color = Phylum, fill = Phylum),
           stat = "identity", position = "stack")
```

Warning in psmelt(physeq): The sample variables:
 Sample
 have been renamed to:
 sample_Sample
 to avoid conflicts with special phyloseq plot attribute names.

```
plotPhylum
```



```
library("phyloseq")
library("ggplot2")

topGenus <- names(sort(phyloseq::taxa_sums(ps2), TRUE)[1:41])
taxTabGenus <- cbind(phyloseq::tax_table(ps2), Genus = NA)
taxTabGenus[topGenus, "Genus"] <- as(tax_table(ps2)[topGenus, "Genus"],
                                     "character")

tax_table(ps2) <- phyloseq::tax_table(taxTabGenus)
ps2m <- merge_samples(ps2, "ExtractionKit")
sample_data(ps2m)$ExtractionKit <- levels(sample_data(ps2)$ExtractionKit)
ps2m <- phyloseq::transform_sample_counts(ps2m, function(x) 100 * x/sum(x))

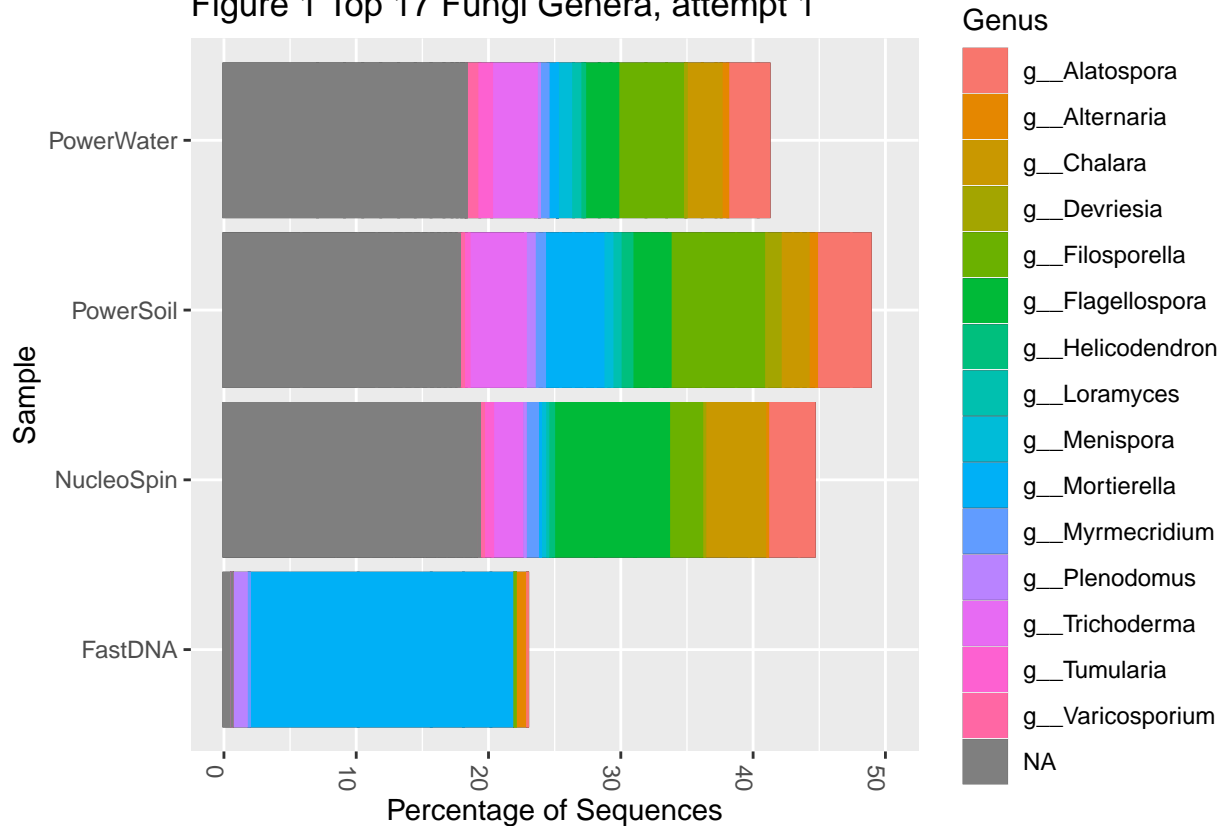
ps2mTop = prune_taxa(topGenus, ps2m)
title = "Figure 1 Top 17 Fungi Genera, attempt 1"
plotGenus <- plot_bar(ps2mTop,
                      #x = "Sample",
                      fill = "Genus",
                      title = title) +

  coord_flip() +
  ylab("Percentage of Sequences") + ylim(0, 50) +
  geom_bar(aes(color = Genus, fill = Genus),
```

```
stat = "identity", position = "stack")
```

```
plotGenus
```

Figure 1 Top 17 Fungi Genera, attempt 1



```
library("phyloseq")
library("ggplot2")

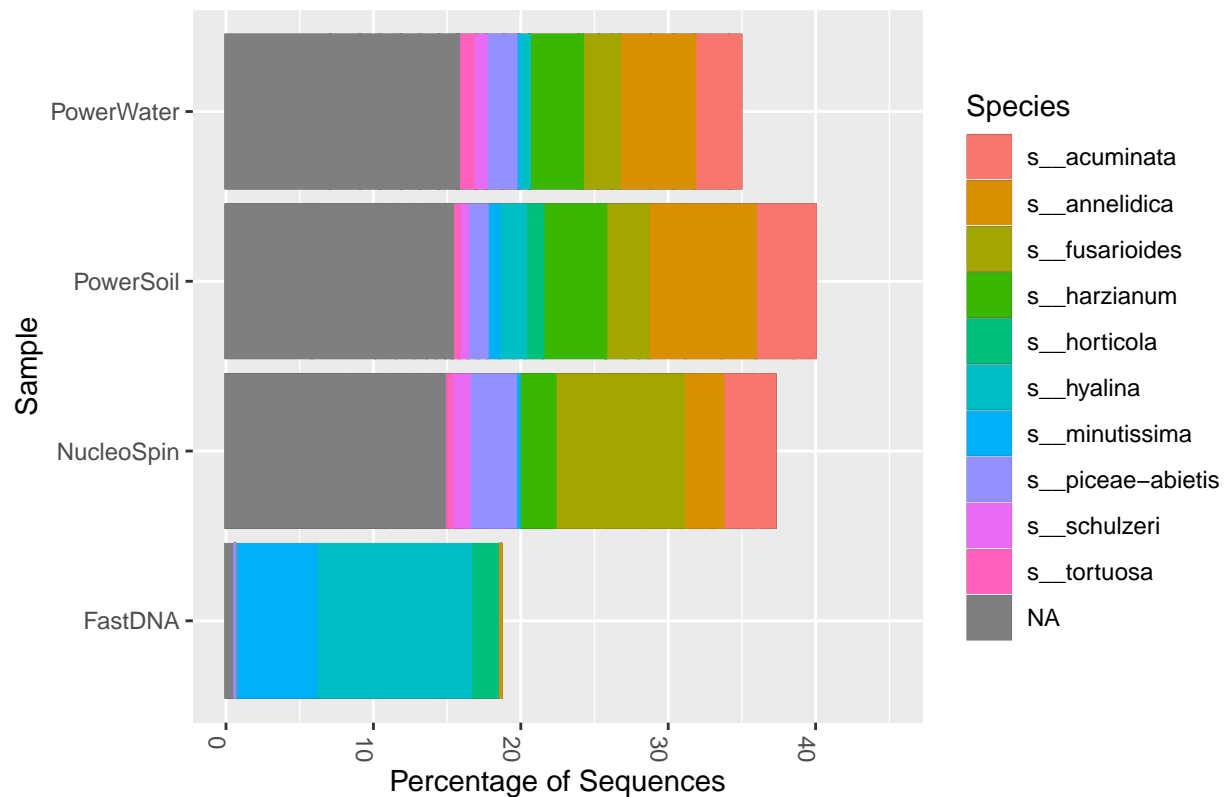
topSpecies <- names(sort(taxa_sums(ps2), TRUE)[1:27])
taxTabSpp <- cbind(phyloseq::tax_table(ps2), Species = NA)
taxTabSpp[topSpecies, "Species"] <- as(phyloseq::tax_table(ps2)[topSpecies, "Species"],
                                         "character")

tax_table(ps2) <- phyloseq::tax_table(taxTabSpp)
ps2mSpp <- phyloseq::merge_samples(ps2, "ExtractionKit")
sample_data(ps2mSpp)$ExtractionKit <- levels(sample_data(ps2)$ExtractionKit)
ps2mSpp <- phyloseq::transform_sample_counts(ps2mSpp, function(x) 100 * x/sum(x))

ps2mSppTop = prune_taxa(topSpecies, ps2mSpp)
title = "Figure 2 Top 17 Fungal Species, attempt 1"
plotSpecies <- plot_bar(ps2mSppTop,
                        x = "Sample",
                        fill = "Species",
                        title = title) +
```

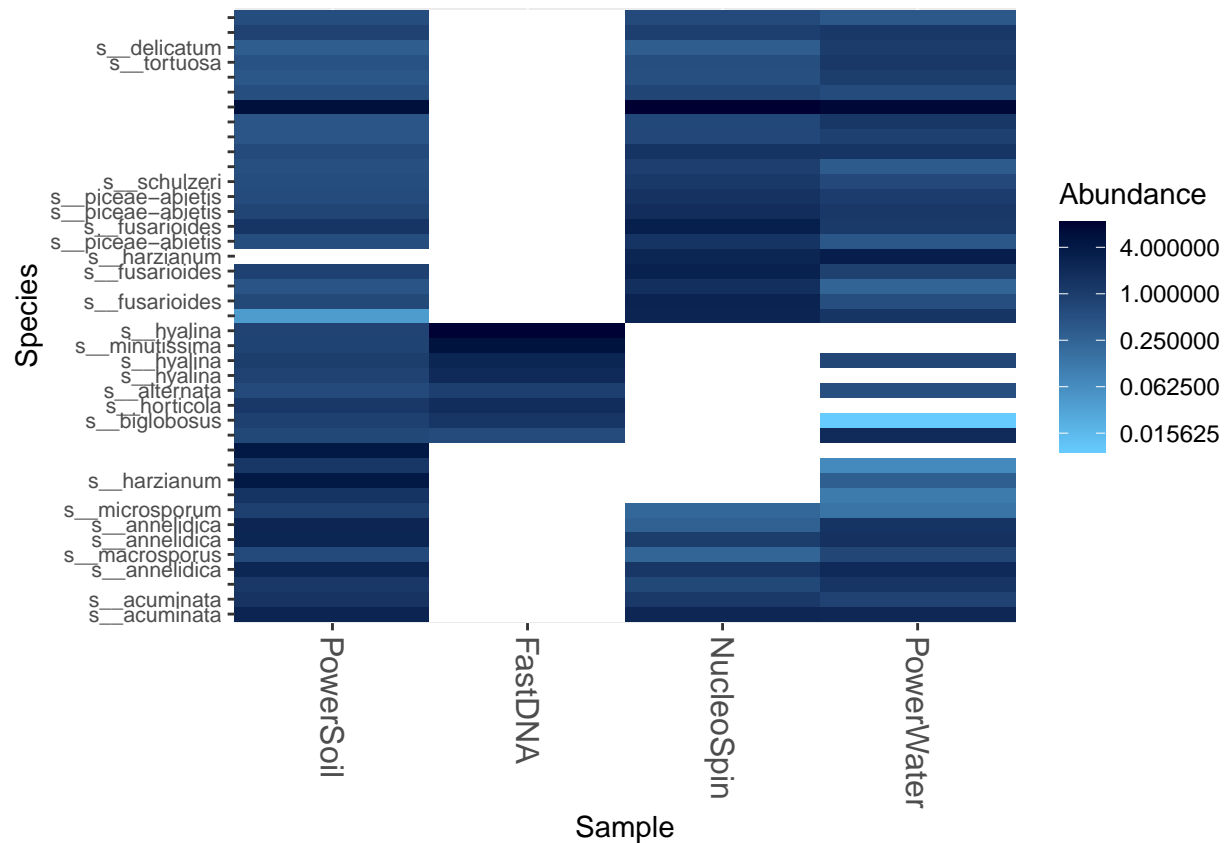
```
coord_flip() +
ylab("Percentage of Sequences") + ylim(0, 45) +
geom_bar(aes(color = Species, fill = Species),
  stat = "identity", position = "stack")
plotSpecies
```

Figure 2 Top 17 Fungal Species, attempt 1



Below I'm testing what a heatmap would look like for taxa abundance across extraction kits:

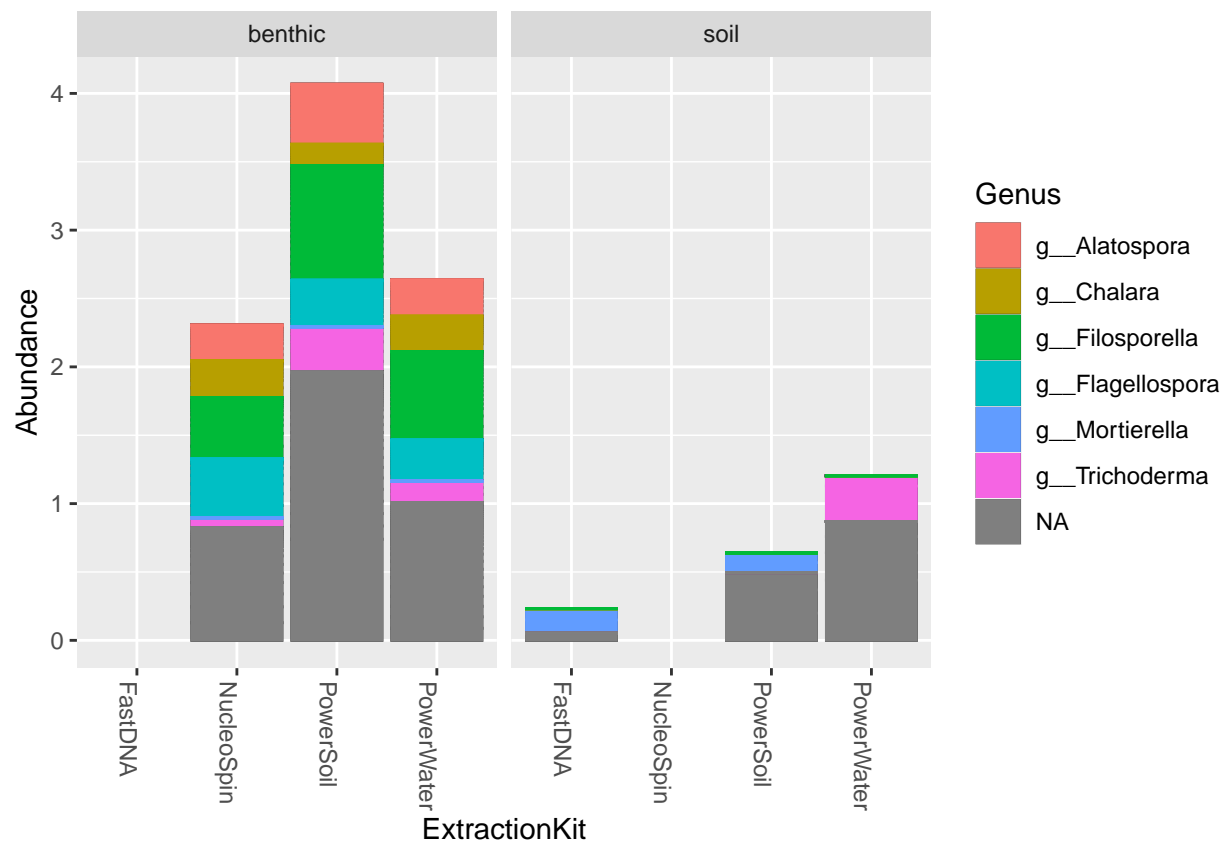
```
library("phyloseq")
library("ggplot2")
heatPlot <- phyloseq::plot_heatmap(ps2mTop, "PCoA", distance="bray",
  sample.label="Sample",
  taxa.label="Species",
  low="#66CCFF", high="#000033", na.value="white") +
  scale_x_discrete(expand=c(0,0))
heatPlot
```



```
library("phyloseq")
library("ggplot2")

top20 <- names(sort(taxa_sums(ps2), decreasing=TRUE))[1:20]
ps.top20 <- phyloseq::transform_sample_counts(ps2, function(OTU) OTU/sum(OTU))
ps.top20 <- phyloseq::prune_taxa(top20, ps.top20)

plotGenus <- phyloseq::plot_bar(ps.top20, x = "ExtractionKit", fill = "Genus",
                                facet_grid = ~sample_Sample)
plotGenus + geom_bar(aes(color = Genus, fill = Genus),
                     stat = "identity",
                     position = "stack")
```



For help see:

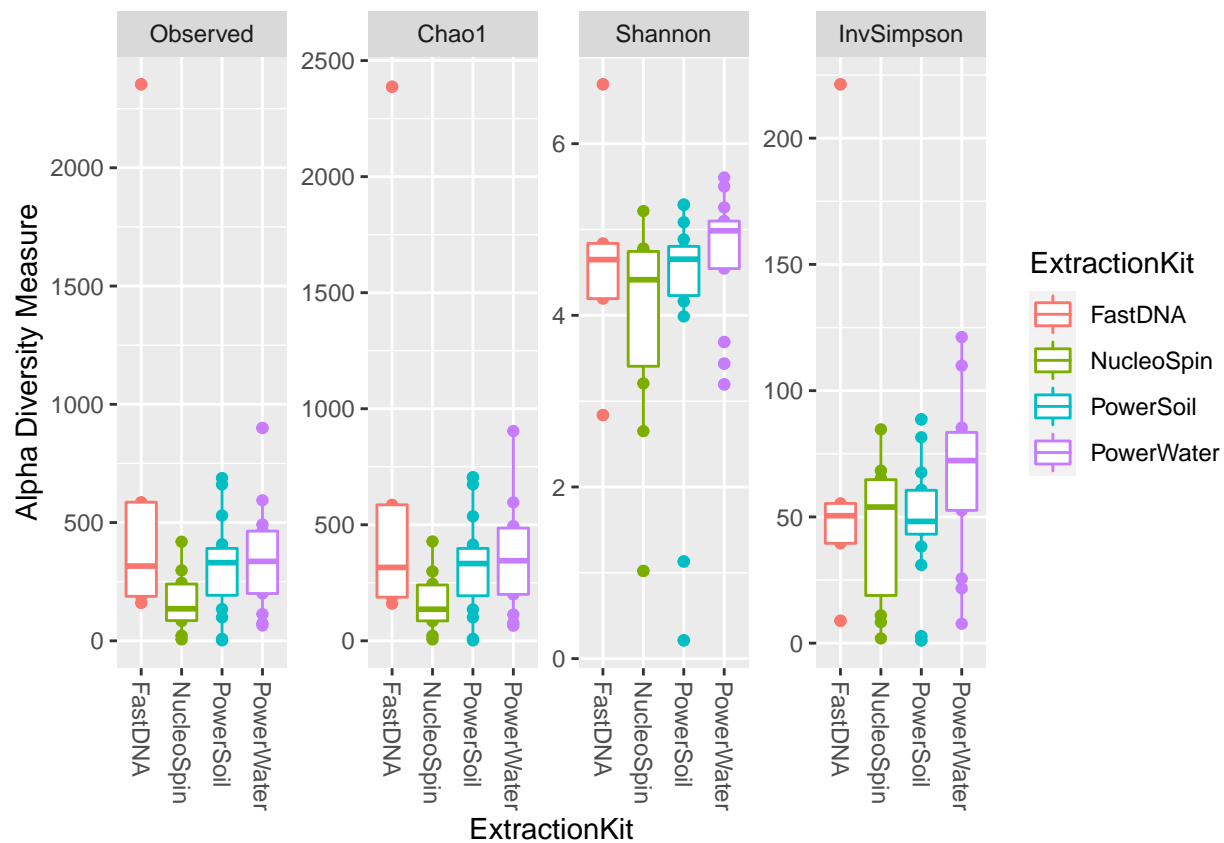
https://www.gdc-docs.ethz.ch/MDA/handouts/MDA20_PhyloseqFormation_Mahendra_Mariadassou.pdf

Not yet working:

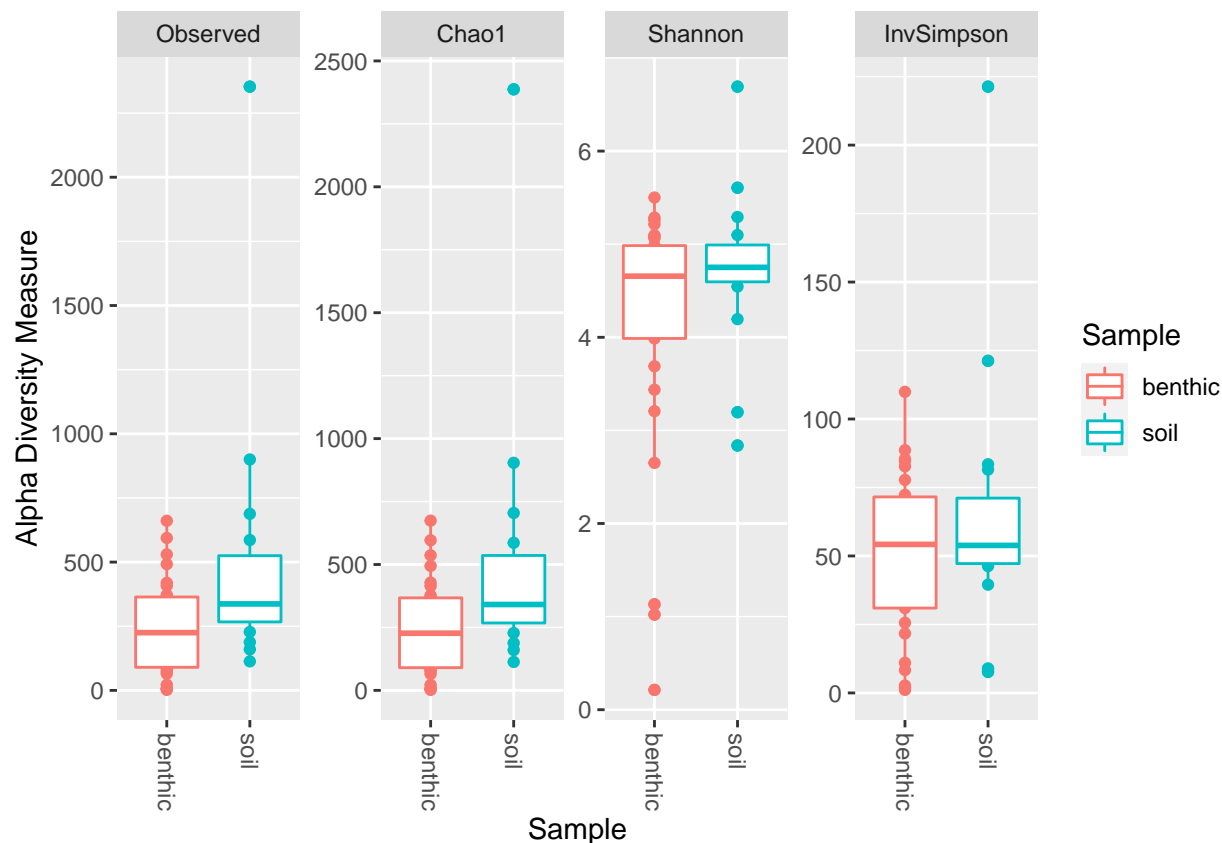
Transform data to proportions as appropriate for Bray-Curtis distances:

```
library("phyloseq")
count_to_prop <- function(x) x/sum(x)
# psProp <- phyloseq::transformSampleCounts(ps2, function(otu) otu/sum(otu))
psProp <- phyloseq::transformSampleCounts(ps2, count_to_prop)
# ordNmDsBray1 <- phyloseq::ordinate(psProp, method="NMDS", distance="bray")
# sample_sums(psProp)[1:5]

p <- plot_richness(ps, color = "ExtractionKit", x = "ExtractionKit",
                  measures = c("Observed", "Chao1", "Shannon", "InvSimpson"))
p <- p + geom_boxplot()
plot(p)
```



```
p <- plot_richness(ps, color = "Sample", x = "Sample",
                   measures = c("Observed", "Chao1", "Shannon", "InvSimpson"))
p <- p + geom_boxplot()
#plot(p)
p
```

Statistical analysis Post hoc comparisons between the four tested methods were made using the Tukey HSD test. OTUs or OTUs pooled at phylum, class, order, family or genera level with different abundances were identified using a generalized linear model where the counts follow an overdispersed Poisson distribution (Kristiansson, Hugenholtz and Dalevi 2009; Jonsson et al.2016). The p-values were corrected for multiple testing using the false discovery rate (FDR) method. The OTU abundance was used for principal component analysis (PCA). Shared OTUs between DNA extraction methods were graphically visualised in Venn diagrams using the corresponding OTU tables exported from QIIME. The hypergeometric distribution was used to test the distribution of gram negatives and gram positives among the taxa identified with the respective four DNA extraction methods. Pearson correlations were used to test for correlations between descriptors of DNA quantity and quality (Table 1), and descriptors of taxonomic diversity (Table 2). The statistical significance for all the analyses was set to $P < 0.05$ or $FDR < 0.05$. All statistical analyses were carried out using the R v.3.2.0 software (R Core Team 2013).

Table 2. Detected 16S rRNA richness and biodiversity from marine periphyton biofilm DNA extracted with the four studied methods.

FastDNA Soil PowerPlant PowerBiofilm PlantDNAzol

n 3 2 3 3 P-values No. of OTUs 666 ± 42 704 ± 58 809 ± 11 791 ± 7 $P < 0.05$ No. of phyla 17 ± 1 17 ± 1 18 ± 0 ns No. of classes 39 ± 1 40 ± 1 40 ± 2 41 ± 1 ns No. of orders 68 ± 3 70 ± 3 71 ± 2 72 ± 1 ns No. of families 91 ± 2 95 ± 6 104 ± 3 106 ± 2 $P < 0.05$ No. of genera 141 ± 4 145 ± 12 159 ± 4 162 ± 3 $P < 0.05$

Each value represents the arithmetic mean \pm standard error of the mean. n: number of replicates. Statistical significance between extraction methods is denoted as $P < 0.05$ (ANOVA).

ns: indicates no statistically significant differences between extraction methods.