phyloseq_demo

Anna DeVeaux

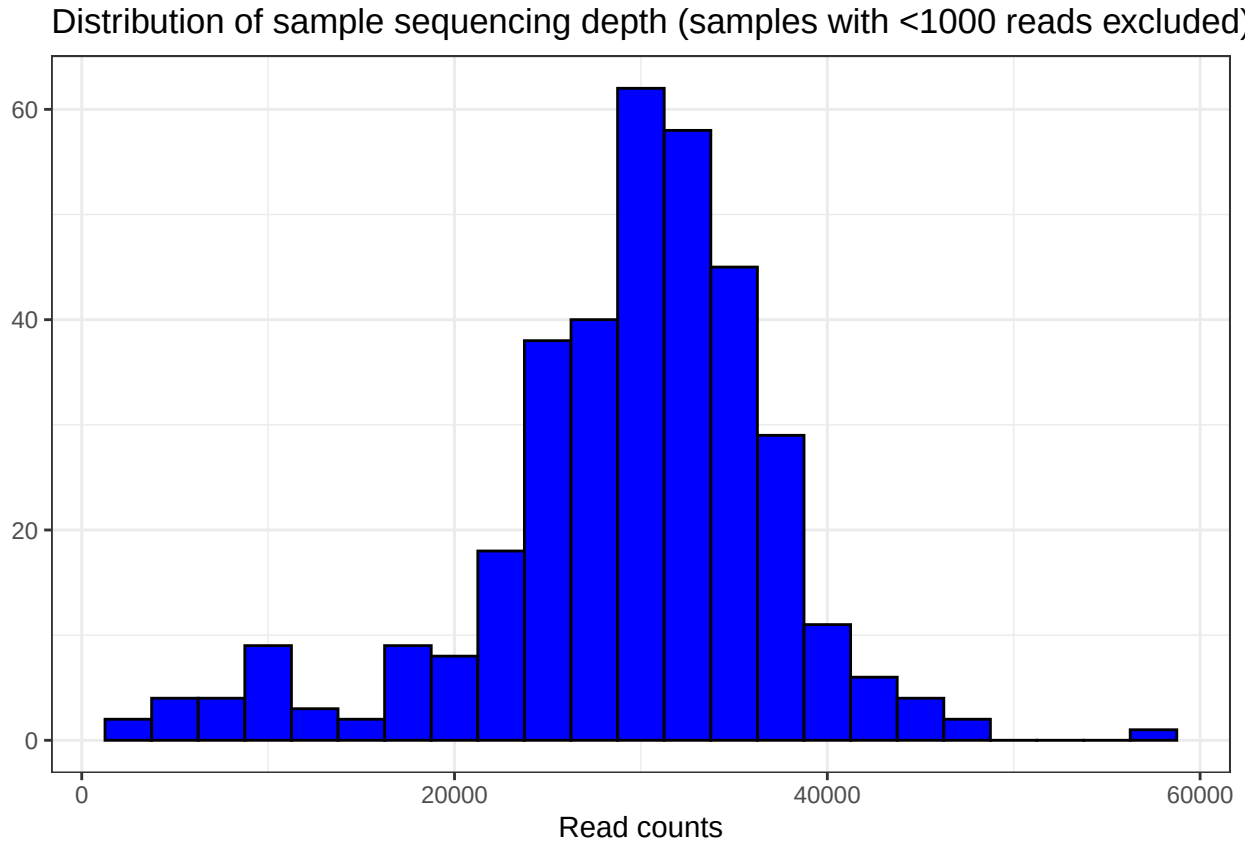
10/22/2019

```
taxfile <- "~/Box Sync/EEB447_Microbes_in_the_Wild/2019/Plastics_Group/final.FWDB.Silva.taxonomy.tsv_ed
metafile <- "~/Box Sync/EEB447_Microbes_in_the_Wild/2019/Plastics_Group/Erie-Plastics_New_Metadata.csv"
otufile <- "~/Box Sync/EEB447_Microbes_in_the_Wild/2019/Plastics_Group/FinalOTUs.otu.table"


plast_tax <- read.table(taxfile, header=F, sep="\t", fill=T)
rownames(plast_tax) <- plast_tax$V1
plast_tax <- subset(plast_tax, select = -c(V1))
taxmat <- as.matrix(plast_tax)   # converts plast_tax from a list into a matrix
colnames(taxmat) <- c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "Species", "Subspecies")
# View(taxmat)

plast_meta <- read.table(metafile, header=T, sep = ",")
rownames(plast_meta) <- plast_meta$Sample_ID   # why don't we subset this out after making it row names
plast_meta_df <- as.data.frame(plast_meta)  # convert to data frame, not matrix, b/c data =/= numerical
# View(plast_meta)

plast_otu <- read.table(otufile, header=T, sep="\t")
rownames(plast_otu) <- plast_otu$OTUid
plast_otu <- subset(plast_otu, select = -c(OTUid))
otumat <- as.matrix(plast_otu)     # again, matrix
# View(plast_otu)
```

```
plast_phylo <- phyloseq(otu_table(otumat, taxa_are_rows = T), tax_table(taxmat), sample_data(plast_meta

colnames(tax_table(plast_phylo)) <- c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "Species
tax_table(plast_phylo) <- cbind(tax_table(plast_phylo), "OTU" = row.names(tax_table(plast_phylo)))
```

Prune out samples with less than 1000 reads and then look at sequencing depth

```
plast_phylo_pruned <- prune_samples(sample_sums(plast_phylo) > 1000, plast_phylo)

sample_sum_pruned_df <- data.frame(sum = sample_sums(plast_phylo_pruned))

ggplot(sample_sum_pruned_df, aes(x = sum)) +
  geom_histogram(color = "black", fill = "blue", binwidth= 2500) +
  ggtitle("Distribution of sample sequencing depth (samples with <1000 reads excluded)") +
  xlab("Read counts") +
  theme(axis.title.y = element_blank())
```




```
smin <- min(sample_sums(plast_phylo_pruned)) %>% print
```

```
## [1] 2409
```

```
smean <- mean(sample_sums(plast_phylo_pruned)) %>% print
```

```
## [1] 29386.33
```

```
smax <- max(sample_sums(plast_phylo_pruned)) %>% print
```

```
## [1] 56793
```

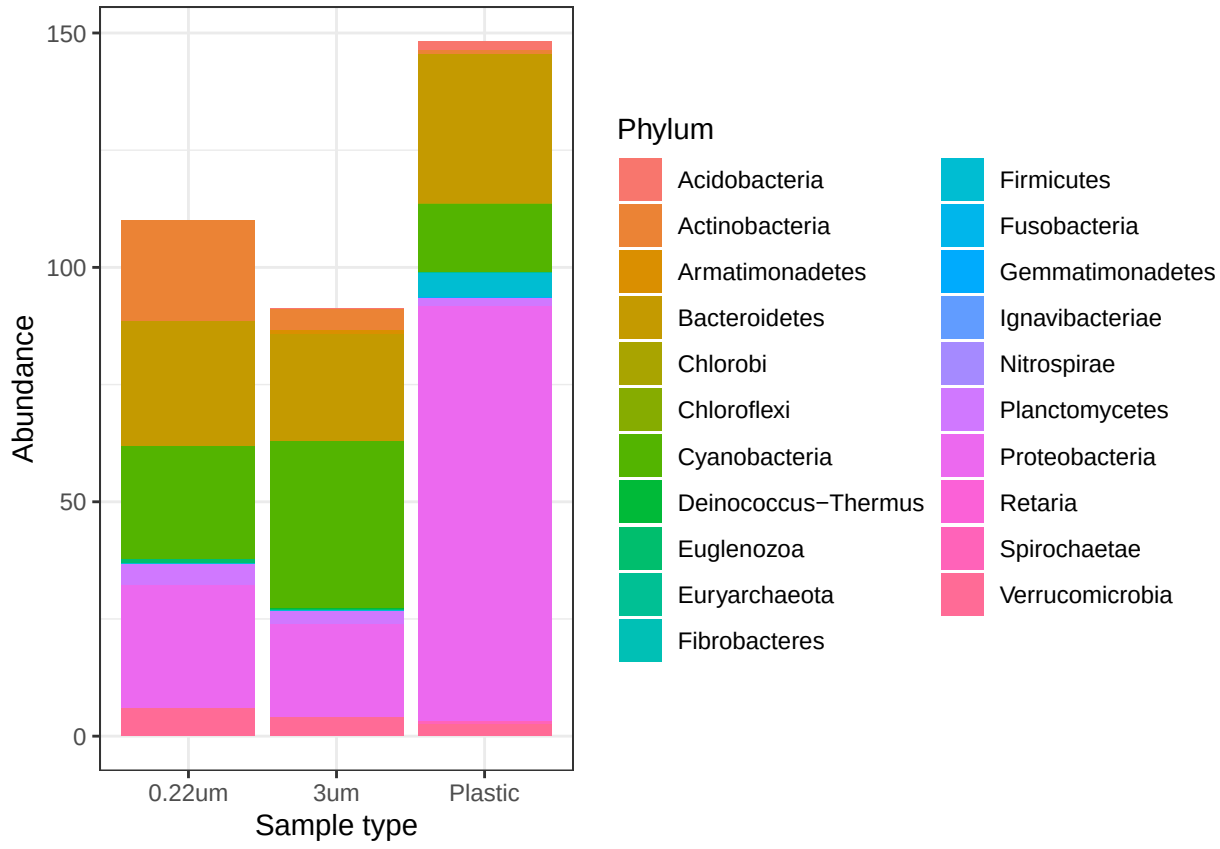
These data look like they represent samples from plastic or filters (0.22um, 3um) collected from the Great Lakes or rivers feeding in to it

Make a stacked barplot of the three sites next to each other (pool locations...)

```
# First "melt" to long format for ggplotting
# prune out phyla below 2% in each sample

plast_phylum <- plast_phylo %>%
  tax_glom(taxrank="Phylum") %>%
  transform_sample_counts(function(x) {x/sum(x)} ) %>%
  psmelt() %>%   # don't know what this does
  filter(Abundance > 0.02) %>%
  arrange(Phylum)
```

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

## Warning in psmelt(.): The rank names:
## OTU
##  have been renamed to:
## taxa_OTU
## to avoid conflicts with special phyloseq plot attribute names.
```

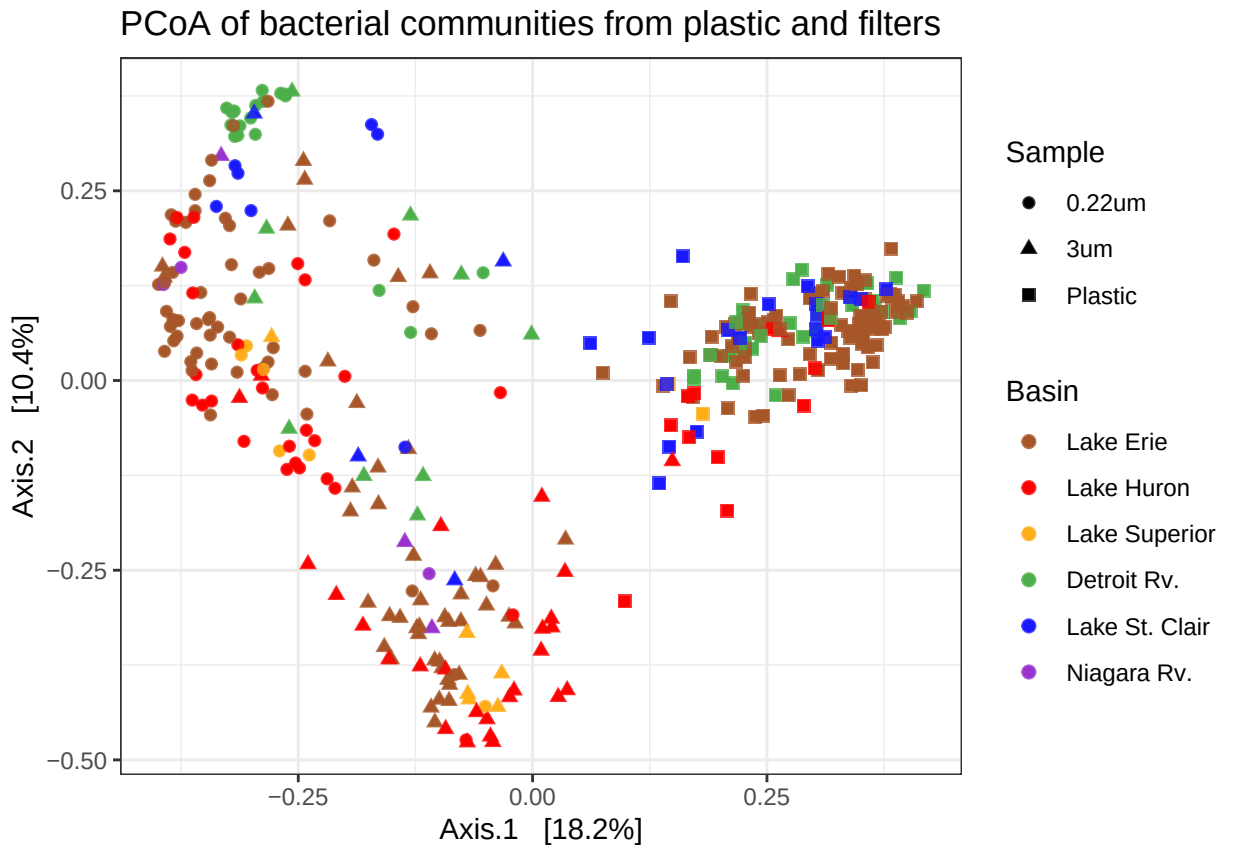
```
ggplot(plast_phylum, aes(x=sample_Sample, y=Abundance, fill=Phylum)) +
  geom_bar(stat="identity") +
  xlab("Sample type")
```



```
  # guides(fill = guide_legend(reverse = T, keywidth=1, keyheight = 1))

# ASK WHY THIS LOOKS THIS WAY (without position = "fill" added to the geom_bar command)
```

Make PCoA plot to look at differences in bacterial communities between different locations... use those as the colors and then use sample types as the shapes

```
# let's see if this works



# Scale reads to even depth
plast_scale <- scale_reads(plast_phylo, n=1000) # WHAT IS THIS ARGUMENT n????? I set a random one

sample_data(plast_scale)$Basin <- factor(
  sample_data(plast_scale)$Basin,
  levels = c("Lake Erie", "Lake Huron", "Lake Superior", "Detroit Rv.", "Lake St. Clair", "Niagara Rv.")
)

# Ordinate
plast_pcoa <- ordinate(
  physeq = plast_scale,
  method = "PCoA",
  distance = "bray"
)

# Plot
plot_ordination(
  physeq = plast_scale,
  ordination = plast_pcoa,
  color = "Basin",
  shape = "Sample",
  title = "PCoA of bacterial communities from plastic and filters",
) +
  scale_color_manual(values = c("#a65628", "red", "#ffae19",
    "#4daf4a", "#1919ff", "darkorchid3")
  ) +
  geom_point(aes(color = Basin), alpha = 0.8, size = 2)
```




```
#geom_point(color = "grey90", size = 1.5)
```