Step2_RNA_Analysis

Shashank Gupta

2023-01-18

Contents

Load all the packages used in the analysis

```
library("ranacapa")
library("phyloseq")
library("ggplot2")
library("stringr")
library("plyr")
library("reshape2")
library("reshape")
library("dplyr")
library("tidyr")
library("doBy")
library("plyr")
library("microbiome")
library("ggpubr")
library("vegan")
library("tidyverse")
library("magrittr")
library("cowplot")
library("dendextend")
library("WGCNA")
library("metagenomeSeq")
library("decontam")
library("RColorBrewer")
library("ampvis2")
library("ggpubr")
library("formatR")
```

Load data

```
setwd("/Users/shashankgupta/Desktop/ImprovAFish/github/ImprovaFish/Metagenomics")
raw <- import_biom("/Users/shashankgupta/Desktop/ImprovAFish/exported-feature-table/feature-table_taxono
tree <- read_tree("/Users/shashankgupta/Desktop/ImprovAFish/exported-feature-table/tree.nwk")
refseq <- Biostrings::readDNAStringSet("/Users/shashankgupta/Desktop/ImprovAFish/exported-feature-table/
dat <- read.table("/Users/shashankgupta/Desktop/ImprovAFish/metadata.txt", header = TRUE,row.names = 1,
# Merge into one complete phyloseq object
all <- merge_phyloseq(raw, sample_data(dat), tree, refseq)
```

Cleaning the data

```
tax <- data.frame(tax_table(all), stringsAsFactors = FALSE)
tax <- tax[,1:7] # No info in col 8-15
# Set informative colnames
colnames(tax) <- c("Kingdom", "Phylum","Class","Order","Family","Genus", "Species")
tax.clean <- data.frame(row.names = row.names(tax),
                        Kingdom = str_replace(tax[,1], "d__",""),
                        Phylum = str_replace(tax[,2], "p__",""),
                        Class = str_replace(tax[,3], "c__",""),
                        Order = str_replace(tax[,4], "o__",""),
                        Family = str_replace(tax[,5], "f__",""),
                        Genus = str_replace(tax[,6], "g__",""),
                        Species = str_replace(tax[,7], "s__",""),
                        stringsAsFactors = FALSE)
tax.clean[is.na(tax.clean)] <- ""
# - Clean rank by rank
# Kingdom - Remove the unassigned completely
# Phylum
# Class
# Remove extra info about origin from some bacteria
# Remove all fields that contain "uncultured", "Unknown" or "Ambigious"
bad <- c("Ambiguous_taxa","uncultured", "Subgroup_21")
tax.clean[tax.clean$Class %in% bad,3:7] <- ""

# Order
bad <- c("0319-6G20","1-20","11-24", "ADurb.Bin180","D8A-2", "Group_1.1c", "JGI_0000069-P22","Marine_Gr
         "Pla3_lineage","Run-SP154",
         "Ambiguous_taxa", "uncultured", "UBA10353_marine_group",
         "Subgroup_17", "SAR86_clade", "SAR11_clade", "SAR202_clade", "Chloroplast")
tax.clean[tax.clean$Order %in% bad,4:7] <- ""

# Family
bad <- c("Ambiguous_taxa","11-24","67-14", "uncultured", "SAR116_clade", "Run-SP154",
         "Marine_Group_II", "env.OPS_17", "SAR116_clade", "S085", "S-70", "NS9_marine_group", "Mitochon
tax.clean[tax.clean$Family %in% bad,5:7] <- ""

# Genus
bad <- c("Ambiguous_taxa","Unknown_Family","uncultured","Subgroup_10", "1174-901-12", "67-14")
tax.clean[tax.clean$Genus %in% bad,6:7] <- ""

# Species
bad <- c("Ambiguous_taxa","marine_metagenome","low_GC","wastewater_metagenome","unidentified",
         "uncultured_synthetic", "uncultured_organism")
tax.clean[tax.clean$Species %in% bad,6:7] <- ""
```

```
#tax.clean[grepl("uncultured", tax.clean$Species),"Species"] <- ""
#tax.clean[grepl("unidentified", tax.clean$Species),"Species"] <- ""

# Remove remove ".", change "-" and " " to "_"
for (i in 1:ncol(tax.clean)){
  tax.clean[,i] <- str_replace_all(tax.clean[,i], "[.]","")
  tax.clean[,i] <- str_replace_all(tax.clean[,i], "[(]","")
  tax.clean[,i] <- str_replace_all(tax.clean[,i], "[)]","")
  tax.clean[,i] <- str_replace_all(tax.clean[,i], "-","_")
  tax.clean[,i] <- str_replace_all(tax.clean[,i], " ","_")
}

for (i in 1:7){ tax.clean[,i] <- as.character(tax.clean[,i])}
# File holes in the tax table
for (i in 1:nrow(tax.clean)){
  #  Fill in missing taxonomy
  if (tax.clean[i,2] == ""){
    kingdom <- paste("Kingdom_", tax.clean[i,1], sep = "")
    tax.clean[i, 2:7] <- kingdom
  } else if (tax.clean[i,3] == ""){
    phylum <- paste("Phylum_", tax.clean[i,2], sep = "")
    tax.clean[i, 3:7] <- phylum
  } else if (tax.clean[i,4] == ""){
    class <- paste("Class_", tax.clean[i,3], sep = "")
    tax.clean[i, 4:7] <- class
  } else if (tax.clean[i,5] == ""){
    order <- paste("Order_", tax.clean[i,4], sep = "")
    tax.clean[i, 5:7] <- order
  } else if (tax.clean[i,6] == ""){
    family <- paste("Family_", tax.clean[i,5], sep = "")
    tax.clean[i, 6:7] <- family
  } else if (tax.clean[i,7] == ""){
    tax.clean$Species[i] <- paste("Genus_",tax.clean$Genus[i], sep = "_")
  }
}

rm(bad, class, family, i, kingdom,new,order,phylum,uncul)

tax_table(all) <- as.matrix(tax.clean)
all
```

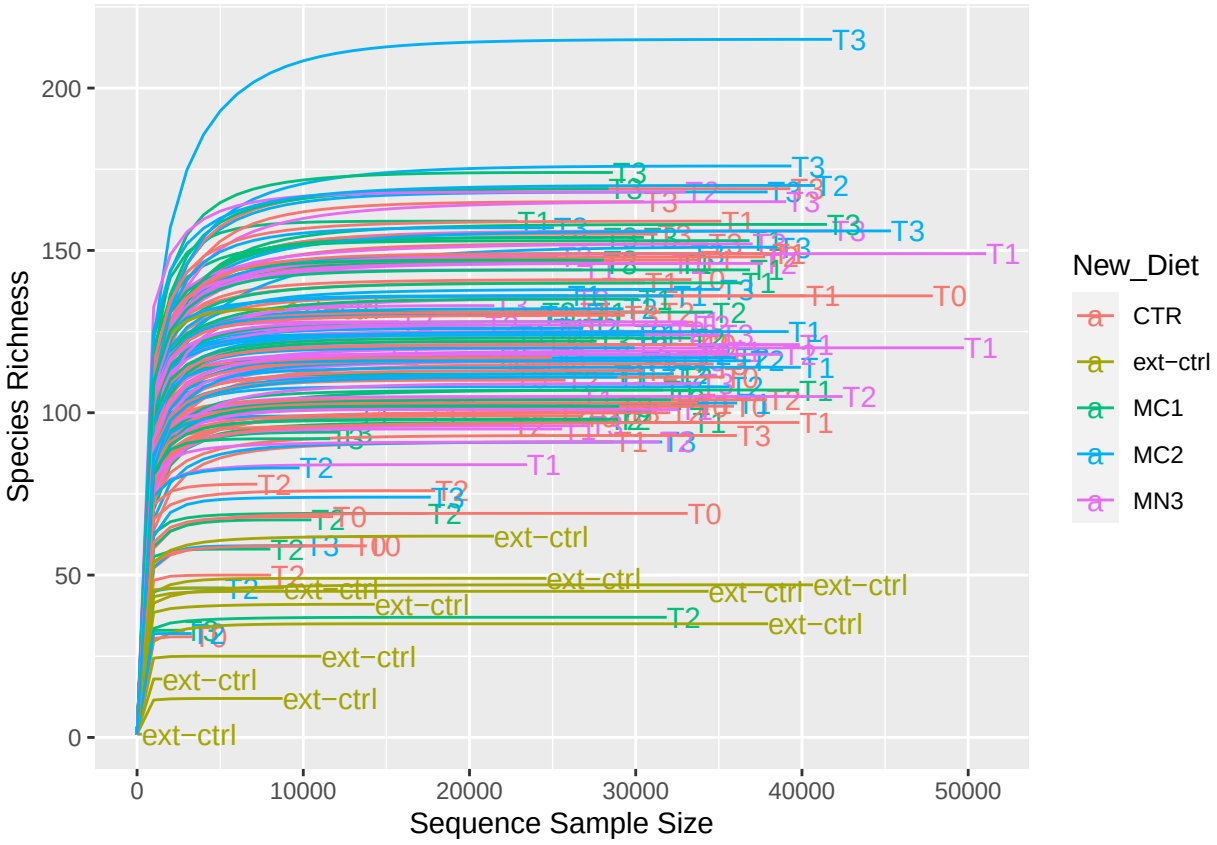
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 7521 taxa and 168 samples ]
## sample_data() Sample Data:       [ 168 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 7521 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 7521 tips and 7465 internal nodes ]
## refseq()      DNAStringSet:      [ 7521 reference sequences ]
```

```
#Remove Unnassigned
#all.clean <- subset_taxa(all, Kingdom != "Archaea")
#all.clean <- subset_taxa(all.clean, Kingdom != "Eukaryota")
all.clean <- subset_taxa(all, Kingdom != "Unassigned")
all.clean <- prune_taxa(taxa_sums(all.clean) > 0, all.clean)
all.clean
```

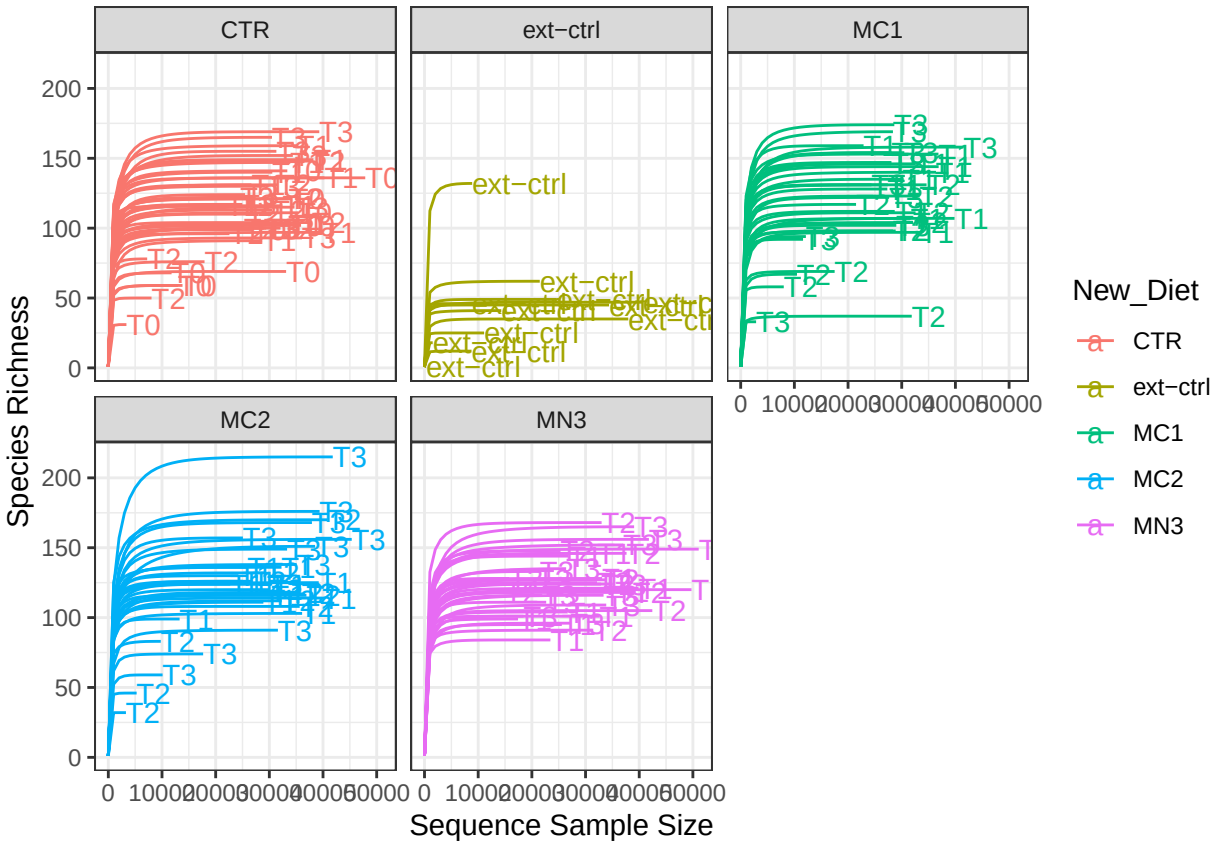
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 7481 taxa and 168 samples ]
## sample_data() Sample Data:       [ 168 samples by 12 sample variables ]
## tax_table()   Taxonomy Table:    [ 7481 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 7481 tips and 7425 internal nodes ]
## refseq()      DNAStringSet:      [ 7481 reference sequences ]
```

Rarefaction plot

```
p <- ranacapa::ggrare(all.clean, step = 1000, color = "New_Diet", label = "samplingTime", se = FALSE, pa
```




```
p + facet_wrap(~New_Diet)+ theme_bw()
```



Raw data Observed richness

```
shannon.div <- estimate_richness(all.clean, measures = c("Shannon", "Simpson", "Observed","Chao1"))
sampledata1<- data.frame(sample_data(all.clean))
row.names(shannon.div) <- gsub("[.]","-", row.names(shannon.div))
sampleData <- merge(sampledata1, shannon.div, by = 0 , all = TRUE)
cols  <- c(brewer.pal(8,"Set1"), brewer.pal(7,"Dark2"),brewer.pal(7,"Set2"),brewer.pal(12,"Set3"),brewe

sampleData$New_Diet <- factor(sampleData$New_Diet, levels=c('ext-ctrl', 'CTR', 'MC1', 'MC2', 'MN3'))
levels(sampleData$New_Diet)
```

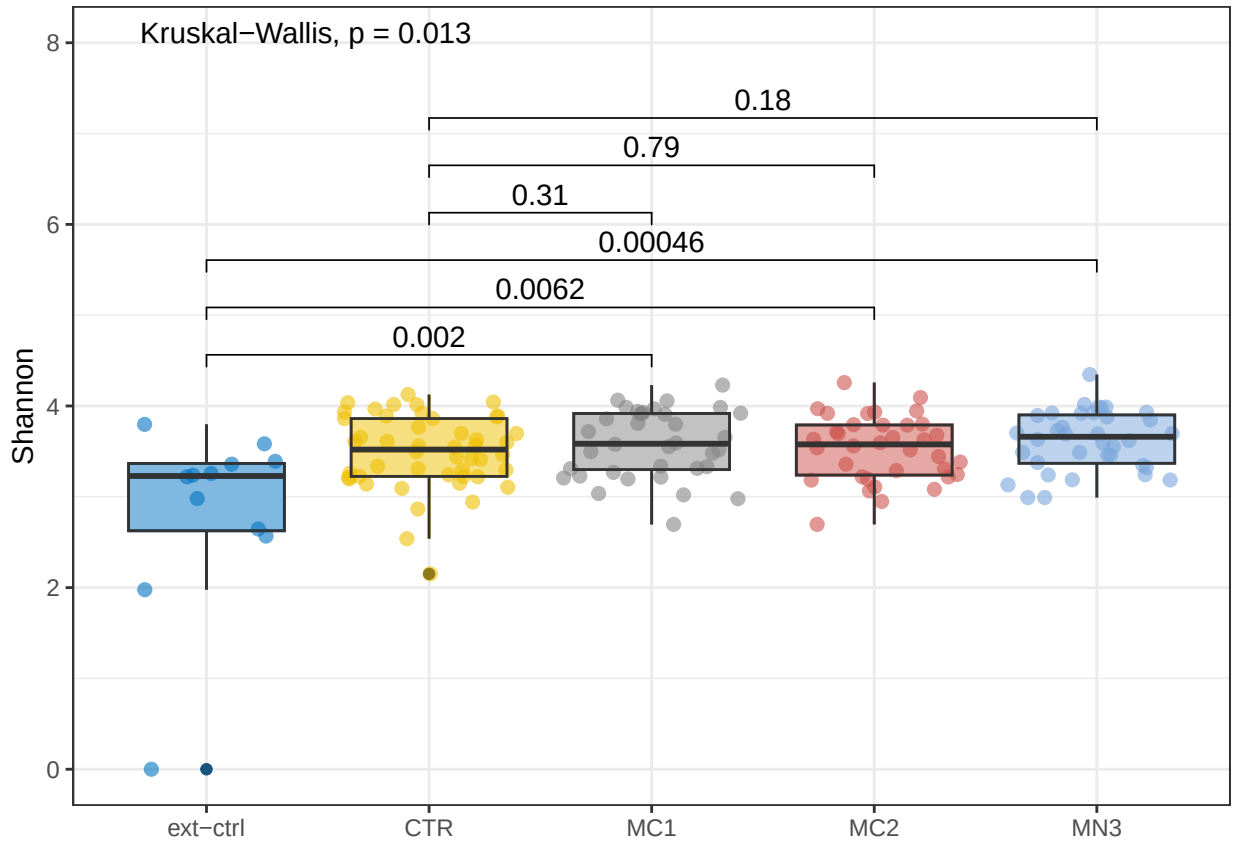
```
## [1] "ext-ctrl" "CTR"      "MC1"      "MC2"      "MN3"
```

```
my_comparisons <- list( c("ext-ctrl", "MC1"), c("ext-ctrl", "MC2"), c("ext-ctrl", "MN3"),
                        c("CTR", "MC1"), c("CTR", "MC2"),
                        c("CTR", "MN3"))
p1 <- ggboxplot(sampleData, x = "New_Diet", y = "Observed",
                color = "New_Diet", palette = "jco", legend = "none")+
  stat_compare_means(comparisons = my_comparisons)+ # Add pairwise comparisons p-value
  stat_compare_means(label.y = 400) +
  geom_jitter(aes(colour = New_Diet), size = 2, alpha = 0.6) +
  geom_boxplot(aes(fill = New_Diet), width=0.7, alpha = 0.5) +
  theme_bw() +  theme(legend.position="none",axis.title.x=element_blank())    +   scale_fill_manual(valu
```

```
## [1] FALSE
```

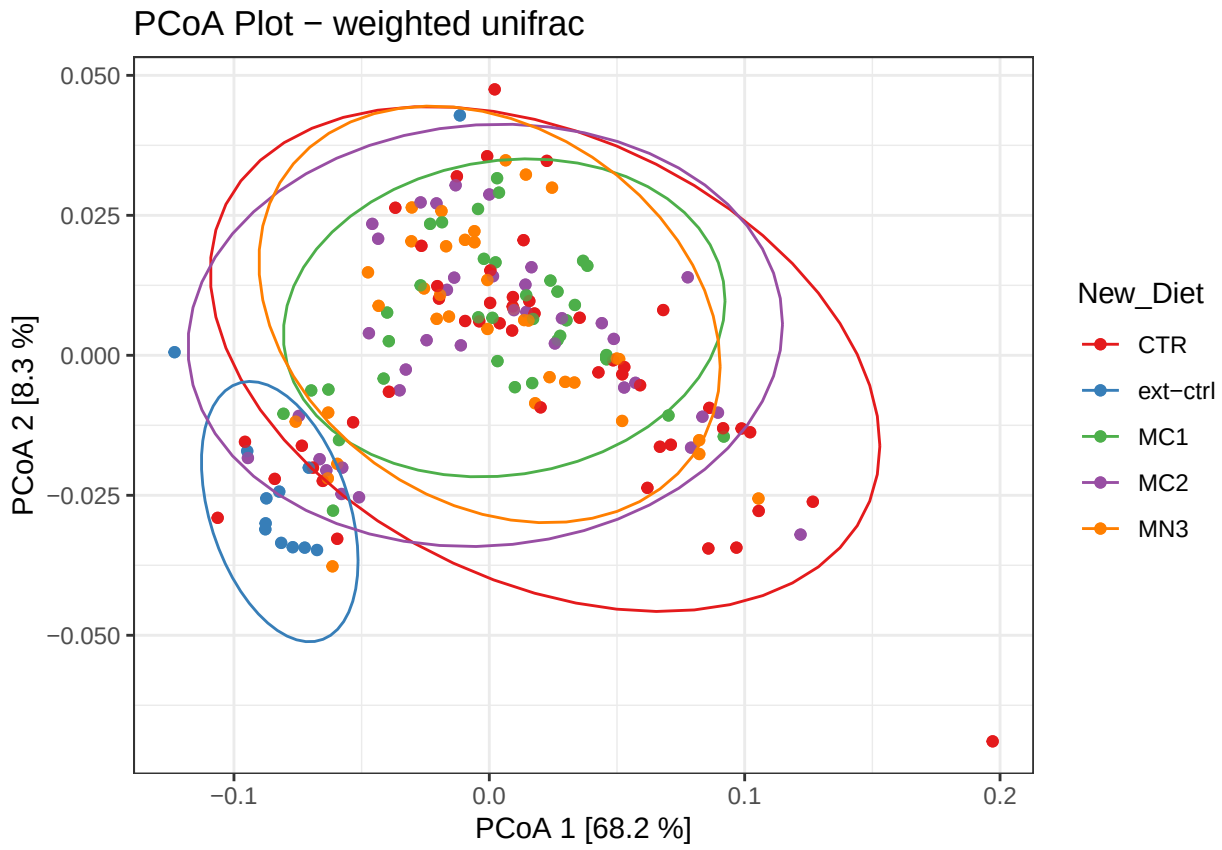
```
ggboxplot(sampleData, x = "New_Diet", y = "Shannon",
          color = "New_Diet", palette = "jco", legend = "none")+
  stat_compare_means(comparisons = my_comparisons)+ # Add pairwise comparisons p-value
  stat_compare_means(label.y = 8) +
  geom_jitter(aes(colour = New_Diet), size = 2, alpha = 0.6) +
  geom_boxplot(aes(fill = New_Diet), width=0.7, alpha = 0.5) +
  theme_bw() +  theme(legend.position="none",axis.title.x=element_blank())     # Add global p-value
```

```
## [1] FALSE
```

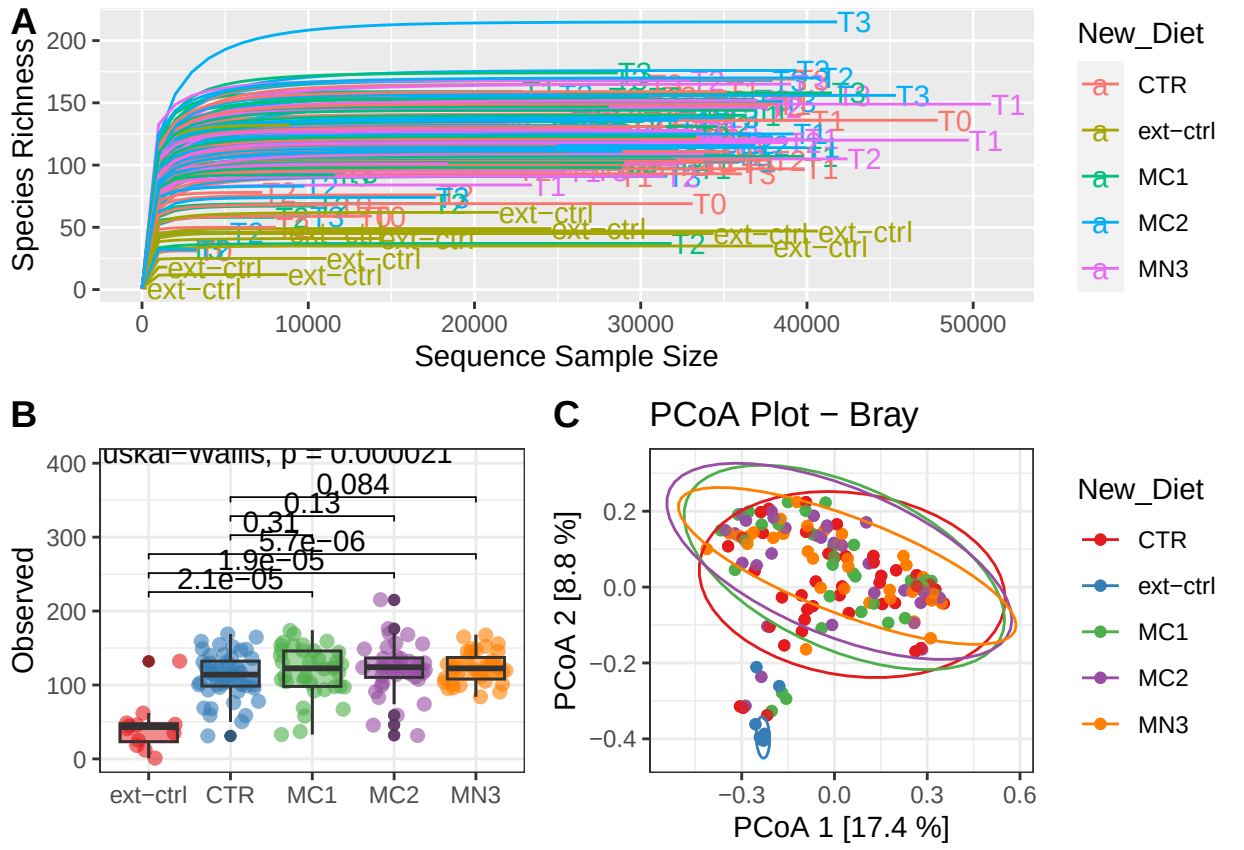



```
set.seed(1)
PCoA_bray <- ordinate(physeq = all.clean, method = "PCoA", distance = "bray")
PCoA_bray_plot<- plot_ordination(
  physeq = all.clean,
  ordination = PCoA_bray,
  color = "New_Diet"
) +
  geom_point(shape = 19, alpha=0.7) + theme_bw() + ggtitle("PCoA Plot - Bray") +
  xlab("PCoA 1 [17.4 %]") + ylab("PCoA 2 [8.8 %]") + stat_ellipse()+   scale_fill_manual(values =cols) +


PCoA_wunifrac <- ordinate(physeq = all.clean, method = "PCoA", distance = "wunifrac")
PCoA_wunifrac_plot<- plot_ordination(
  physeq = all.clean,
  ordination = PCoA_wunifrac,
  color = "New_Diet"
```

```
) +
  geom_point(shape = 19, alpha=0.7) + theme_bw() + ggtitle("PCoA Plot - weighted unifrac") +
  xlab("PCoA 1 [68.2 %]") + ylab("PCoA 2 [8.3 %]") + stat_ellipse()+   scale_fill_manual(values =cols) +

PCoA_wunifrac_plot
```




```
bottom_row <- plot_grid(p1, PCoA_bray_plot, labels = c('B', 'C'), align = 'h', rel_widths = c(1, 1.3))
plot_grid(p, bottom_row, labels = c('A', ''), ncol = 1, rel_heights = c(1, 1.2))
```



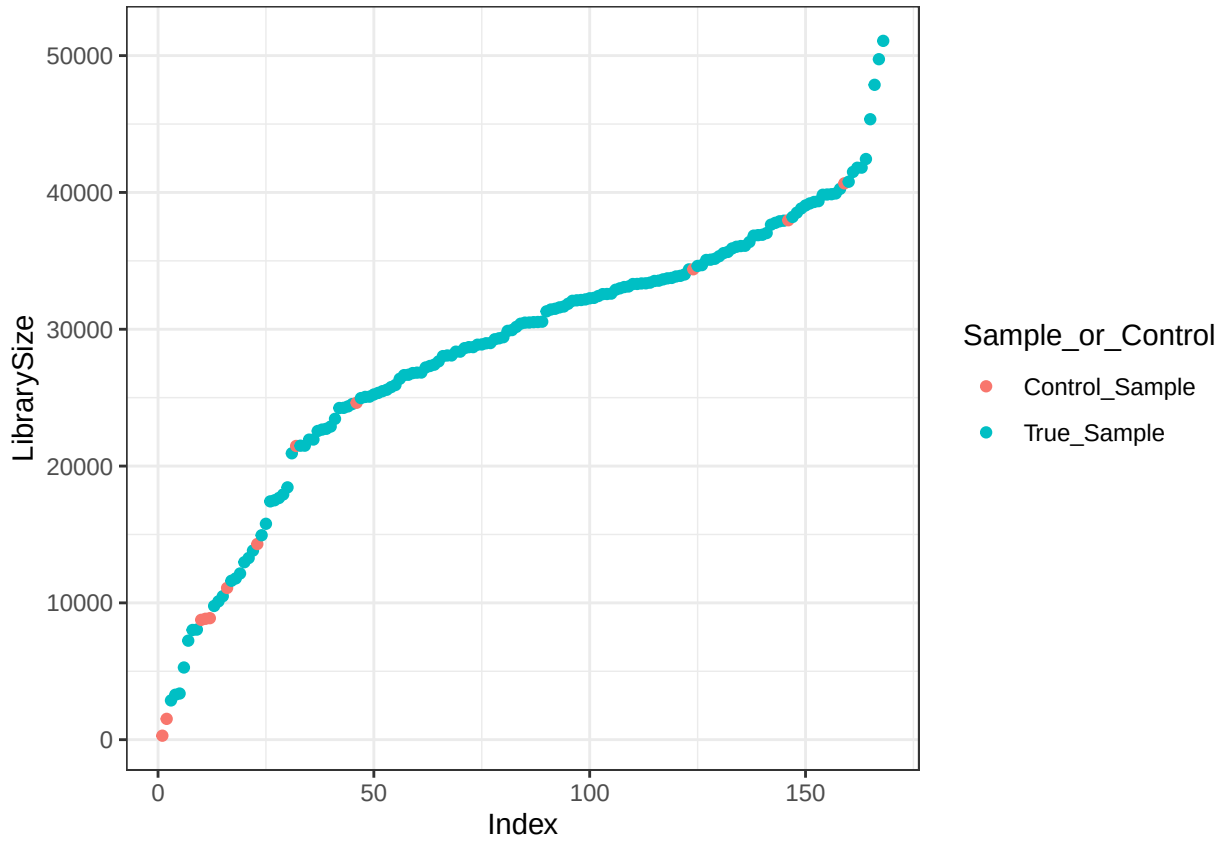

```
sampledf <- data.frame(sample_data(all.clean))
bcdist <- phyloseq::distance(all.clean, method="bray",normalized=TRUE)
adonis2(bcdist ~ New_Diet,
        data = sampledf, permutations = 9999)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 9999
##
## adonis2(formula = bcdist ~ New_Diet, data = sampledf, permutations = 9999)
##           Df SumOfSqs      R2      F Pr(>F)
## New_Diet   4    3.572 0.07596 3.3498 0.0001 ***
## Residual 163   43.450 0.92404
## Total    167   47.021 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

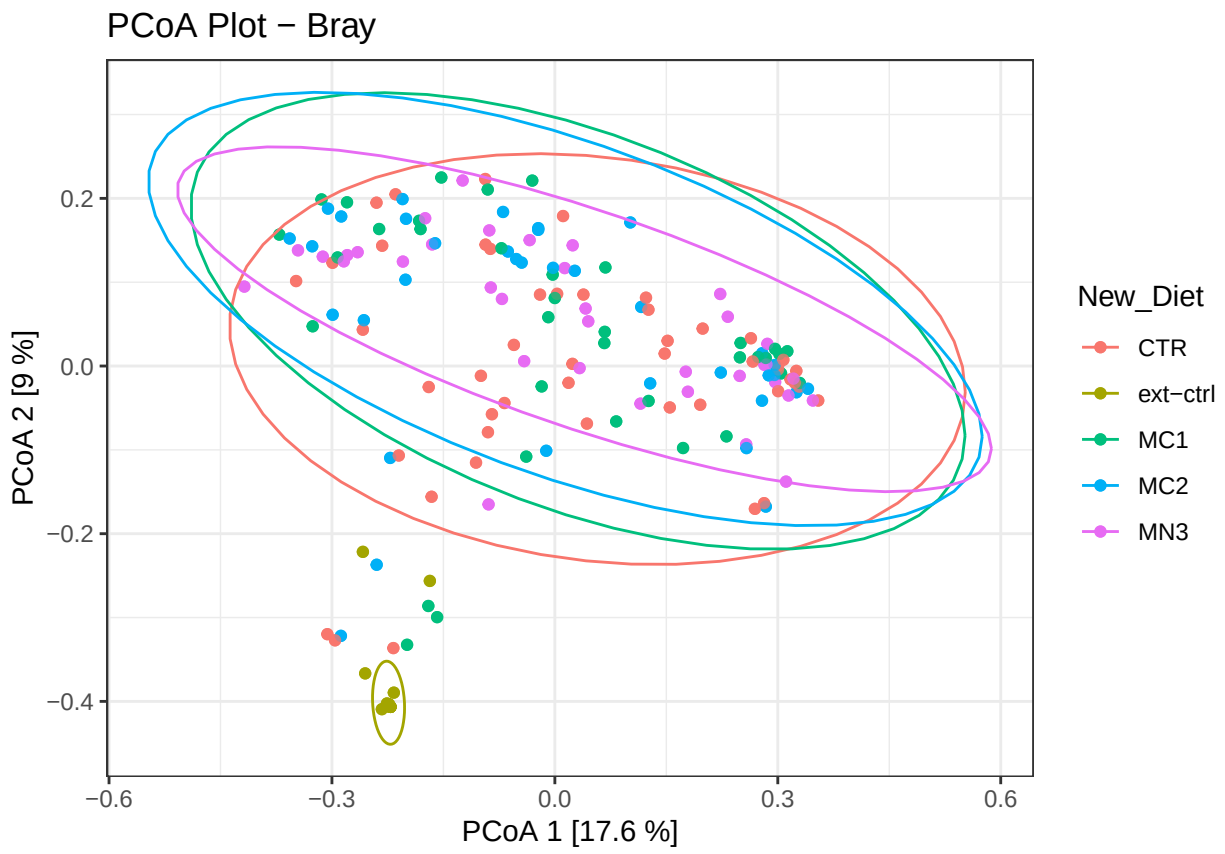
Contamination removal

```
df <- as.data.frame(sample_data(all.clean)) # Put sample_data into a ggplot-friendly data.frame
df$Sample_or_Control <- ifelse( df$New_Diet  %in% c("ext-ctrl"), "Control_Sample", "True_Sample")
sample_data(all.clean) <- df
df$LibrarySize <- sample_sums(all.clean)
df <- df[order(df$LibrarySize),]
df$Index <- seq(nrow(df))
```

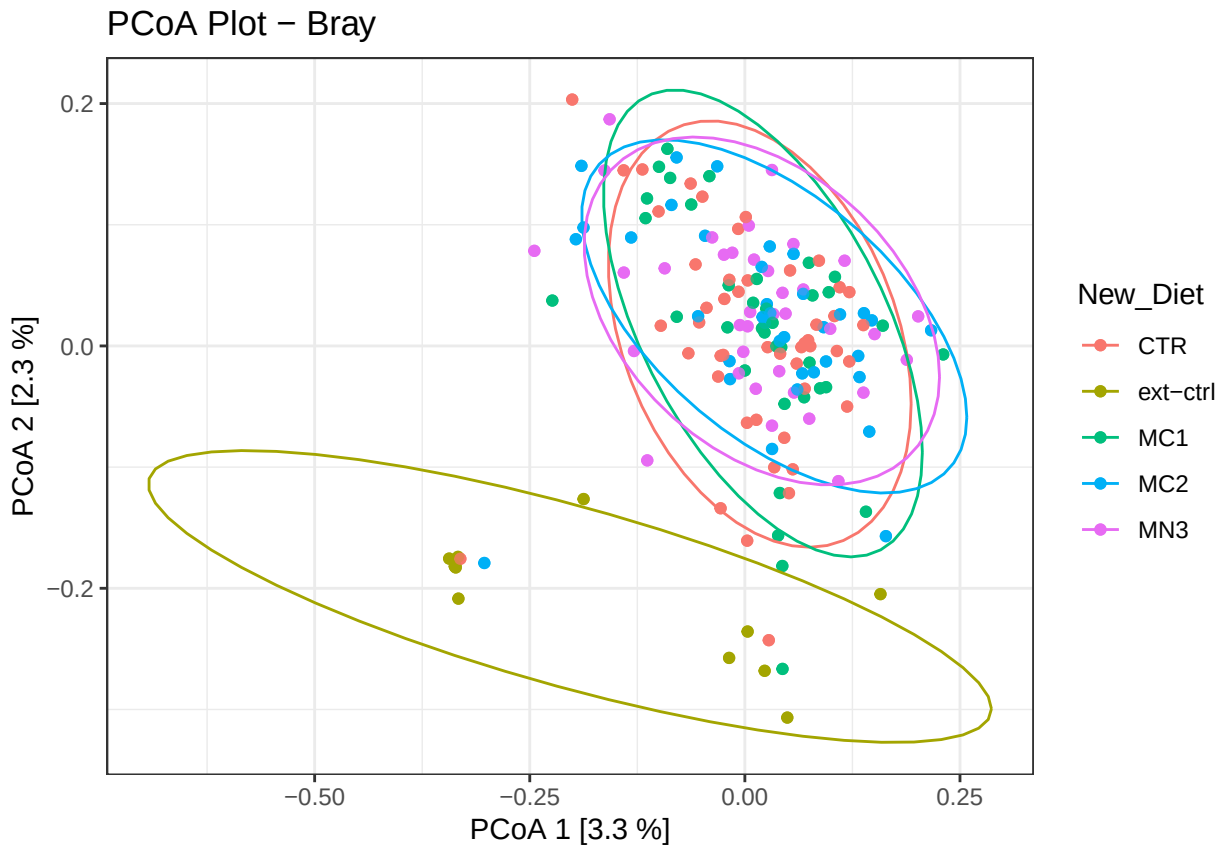
```
ggplot(data=df, aes(x=Index, y=LibrarySize, color=Sample_or_Control)) + geom_point() + theme_bw()
```




```
# Identify Contaminants - Prevalence
sample_data(all.clean)$is.neg <- sample_data(all.clean)$Sample_or_Control == "Control_Sample"
contamdf.prev <- isContaminant(all.clean, method="prevalence", neg="is.neg")
#table(contamdf.prev$contaminant)
#head(which(contamdf.prev$contaminant))

#more aggressive classification threshold rather than the default. i.e., 0.05
contamdf.prev05 <- isContaminant(all.clean, method="prevalence", neg="is.neg", threshold=0.5)
#table(contamdf.prev05$contaminant)
all.noncontam <- prune_taxa(!contamdf.prev05$contaminant, all.clean)

set.seed(1)
PCoA_bray <- ordinate(physeq = all.noncontam, method = "PCoA", distance = "bray")
PCoA_bray_plot<- plot_ordination(
  physeq = all.noncontam,
  ordination = PCoA_bray,
  color = "New_Diet"
) +
  geom_point(shape = 19, alpha=0.7) + theme_bw() + ggtitle("PCoA Plot - Bray") +
  xlab("PCoA 1 [17.6 %]") + ylab("PCoA 2 [9 %]") + stat_ellipse()

PCoA_bray_plot
```




```
PCoA_wunifrac <- ordinate(physeq = all.noncontam, method = "PCoA", distance = "unifrac")
PCoA_wunifrac_plot<- plot_ordination(
  physeq = all.noncontam,
  ordination = PCoA_wunifrac,
  color = "New_Diet"
) +
  geom_point(shape = 19, alpha=0.7) + theme_bw() + ggtitle("PCoA Plot - Bray") +
  xlab("PCoA 1 [3.3 %]") + ylab("PCoA 2 [2.3 %]") + stat_ellipse()
PCoA_wunifrac_plot
```




```
psdata <- subset_samples(all.noncontam, Sample_or_Control=="True_Sample")
psdata <- prune_taxa(taxa_sums(psdata) > 0, psdata)
psdata
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 7065 taxa and 156 samples ]
## sample_data() Sample Data:       [ 156 samples by 14 sample variables ]
## tax_table()   Taxonomy Table:    [ 7065 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 7065 tips and 7015 internal nodes ]
## refseq()      DNAStringSet:      [ 7065 reference sequences ]
```

Rarefaction curves

Inspect the number of reads per sample and compare to rarefaction curves

```
sample_data(psdata)$reads <- unlist(sample_sums(psdata))
dat1 <- data.frame(sample_data(psdata))
#table(dat1$reads)

# Set a cutoff where Shannon diversity dont increase, and observed increases markedly slower.
# At the same time don't throw too many samples out. Set the cutoff value accordingly
cutoff <- 0

# See which samples have been removed
#dat1[dat1$reads < cutoff,]

# Remove the samples with fewer reads than the cutoff
```

```
psdata.p <- prune_samples(sample_sums(psdata) > cutoff, psdata)
psdata.p <- prune_taxa(taxa_sums(psdata.p) >0, psdata.p)
psdata.p
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 7065 taxa and 156 samples ]
## sample_data() Sample Data:       [ 156 samples by 15 sample variables ]
## tax_table()   Taxonomy Table:    [ 7065 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 7065 tips and 7015 internal nodes ]
## refseq()      DNAStringSet:      [ 7065 reference sequences ]
```

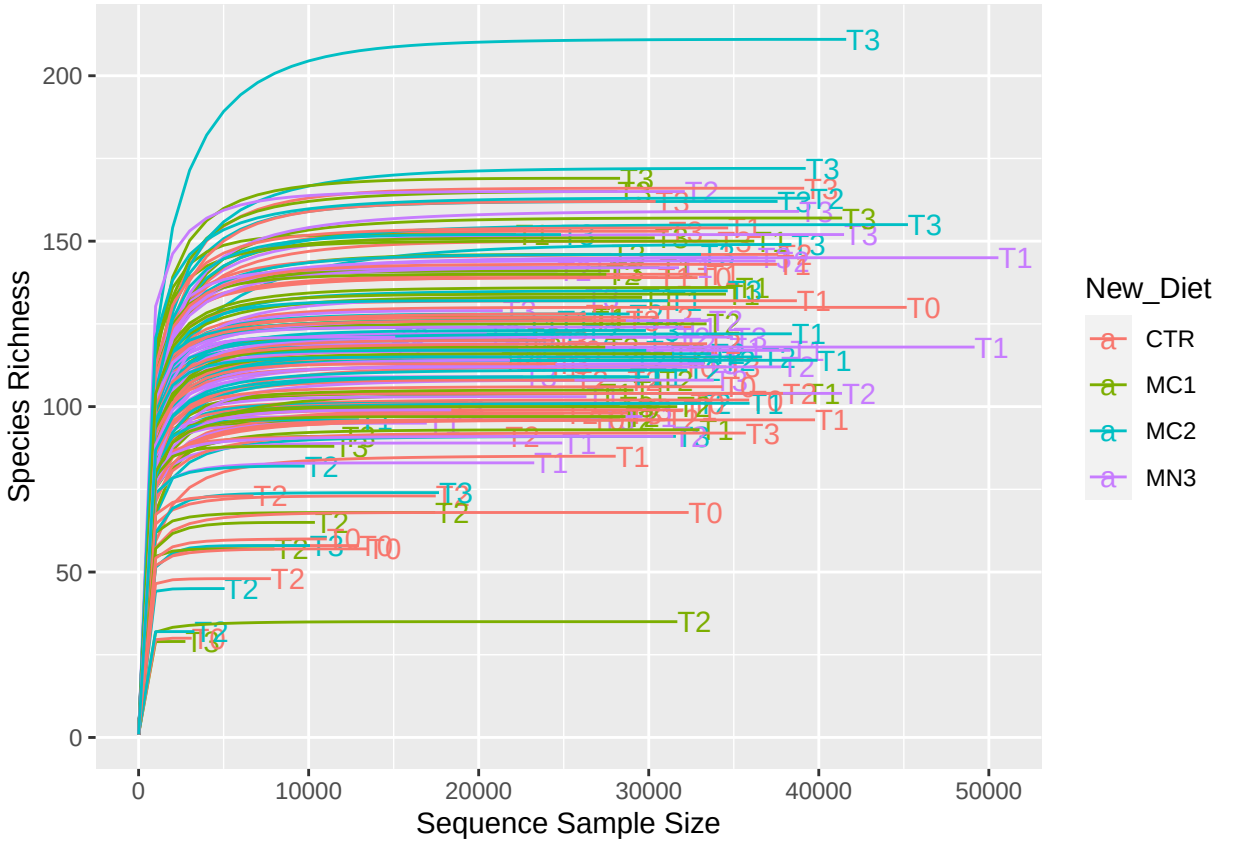
```
# Rarefy the samples using the function multiple_rarefy
psdata.r <- multiple_rarefy(psdata.p)
#psdata.r = rarefy_even_depth(psdata.p, rngseed=1, sample.size=0.9*min(sample_sums(psdata.p)), replace=
#psdata.r = rarefy_even_depth(psdata.p)
```

```
psdata.r<- transform_sample_counts(psdata.p, function(x) x / sum(x) )
psdata.r <- prune_taxa(taxa_sums(psdata.r) > 0, psdata.r)
psdata.r
```

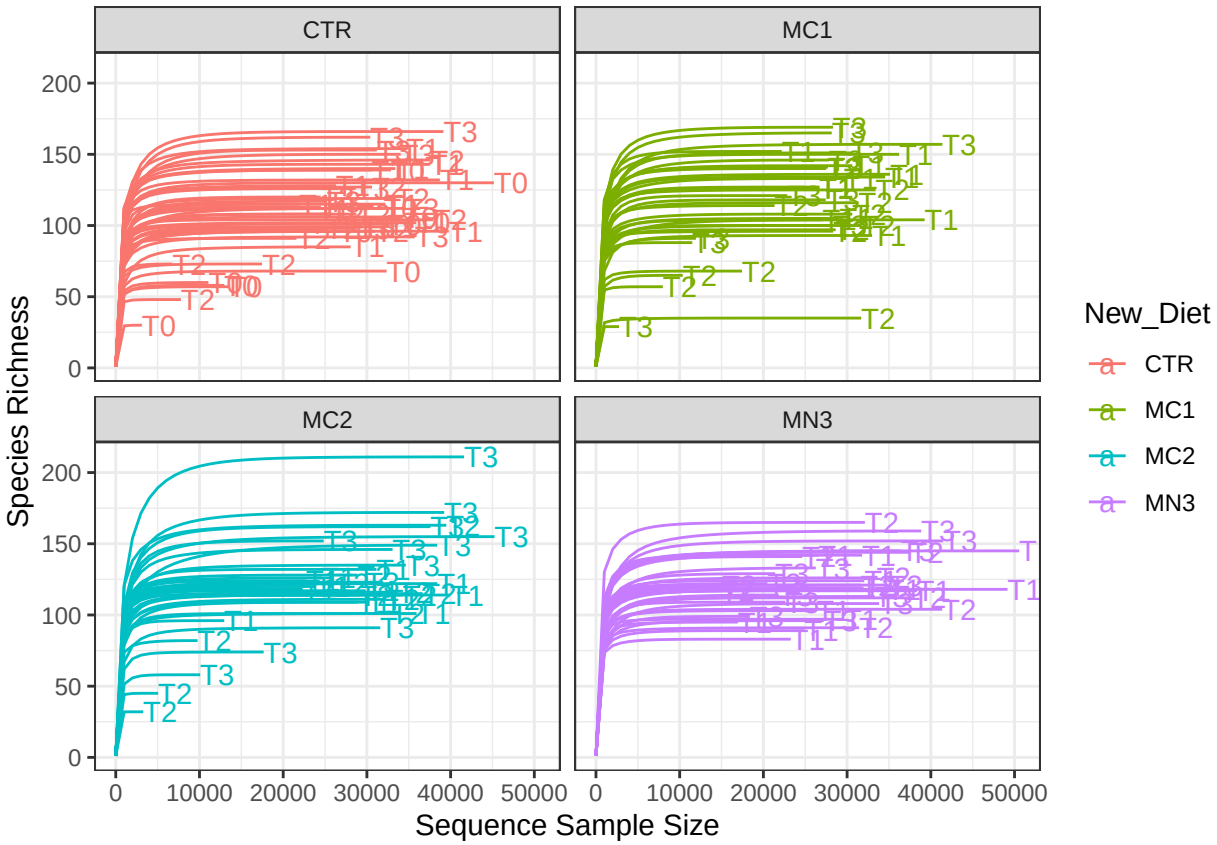
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 7065 taxa and 156 samples ]
## sample_data() Sample Data:       [ 156 samples by 15 sample variables ]
## tax_table()   Taxonomy Table:    [ 7065 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 7065 tips and 7015 internal nodes ]
## refseq()      DNAStringSet:      [ 7065 reference sequences ]
```

```
#table(sample_sums(psdata.r))
rm(all)
rm(all.clean)
rm(all.noncontam)
```

```
p2<- ggrare(psdata.p, step = 1000, color = "New_Diet", label = "samplingTime", se = FALSE)+ facet_wrap(
```


Species Richness
Sequence Sample Size
0
50
100
150
200
0
10000
20000
30000
40000
50000
New_Diet
CTR
MC1
MC2
MN3

```
p2
```


CTR
MC1
MC2
MN3
Species Richness
Sequence Sample Size
0
50
100
150
200
0
10000
20000
30000
40000
50000
New_Diet
CTR
MC1
MC2
MN3