Basic_EDA

Anastasia Poluzerova

2023-06-02

```
knitr::opts_chunk$set(fig.width = 10, fig.height = 6)


library('phyloseq')
library('tidyverse')



set.seed(5678)
setwd('/home/nastasista/Metagenomics')
ps <- readRDS("ps.RData")
ps
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 5056 taxa and 24 samples ]
## sample_data() Sample Data:       [ 24 samples by 6 sample variables ]
## tax_table()   Taxonomy Table:    [ 5056 taxa by 6 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 5056 reference sequences ]
```

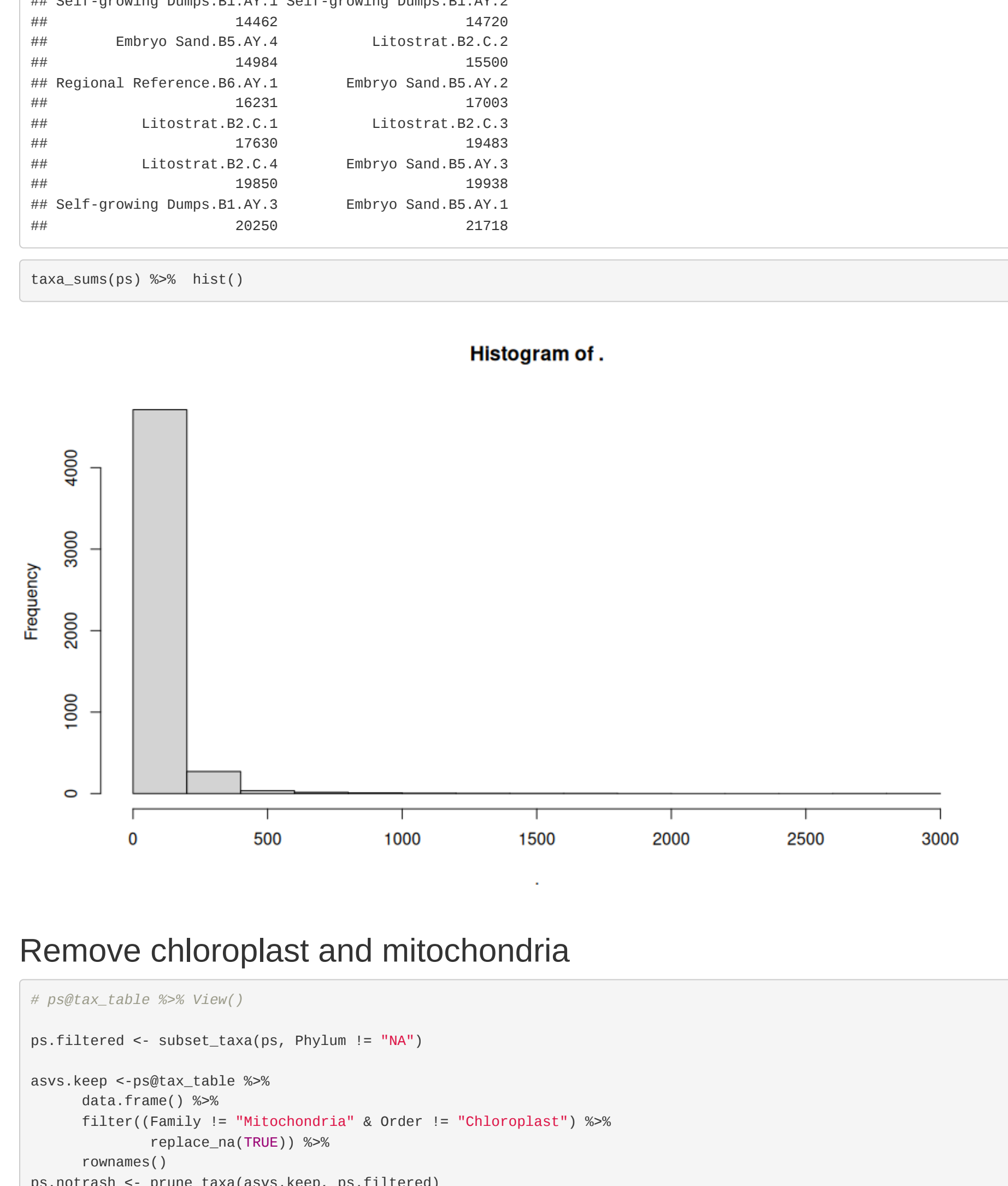
```
ps@tax_table %>% View() #смотрю все ли таксоны аннотированы до Phylum
```

Brief view on samples

```
sample_sums(ps) %>% sort()
```

```
##    Local Reference.B4.AY.3  Coal Mine Terricon.B3.C.2
##                       8157                       8413
##    Local Reference.B4.AY.4  Coal Mine Terricon.B3.C.3
##                       8925                       9839
##  Coal Mine Terricon.B3.C.4    Local Reference.B4.AY.1
##                       9896                      10392
##    Local Reference.B4.AY.2  Coal Mine Terricon.B3.C.1
##                      10827                      11714
## Regional Reference.B6.AY.4 Regional Reference.B6.AY.3
##                      12192                      12724
## Regional Reference.B6.AY.2 Self-growing Dumps.B1.AY.4
##                      12889                      13500
## Self-growing Dumps.B1.AY.1 Self-growing Dumps.B1.AY.2
##                      14462                      14720
##        Embryo Sand.B5.AY.4           Litostrat.B2.C.2
##                      14984                      15500
## Regional Reference.B6.AY.1        Embryo Sand.B5.AY.2
##                      16231                      17003
##           Litostrat.B2.C.1           Litostrat.B2.C.3
##                      17630                      19483
##           Litostrat.B2.C.4        Embryo Sand.B5.AY.3
##                      19850                      19938
## Self-growing Dumps.B1.AY.3        Embryo Sand.B5.AY.1
##                      20250                      21718
```

```
taxa_sums(ps) %>%  hist()
```

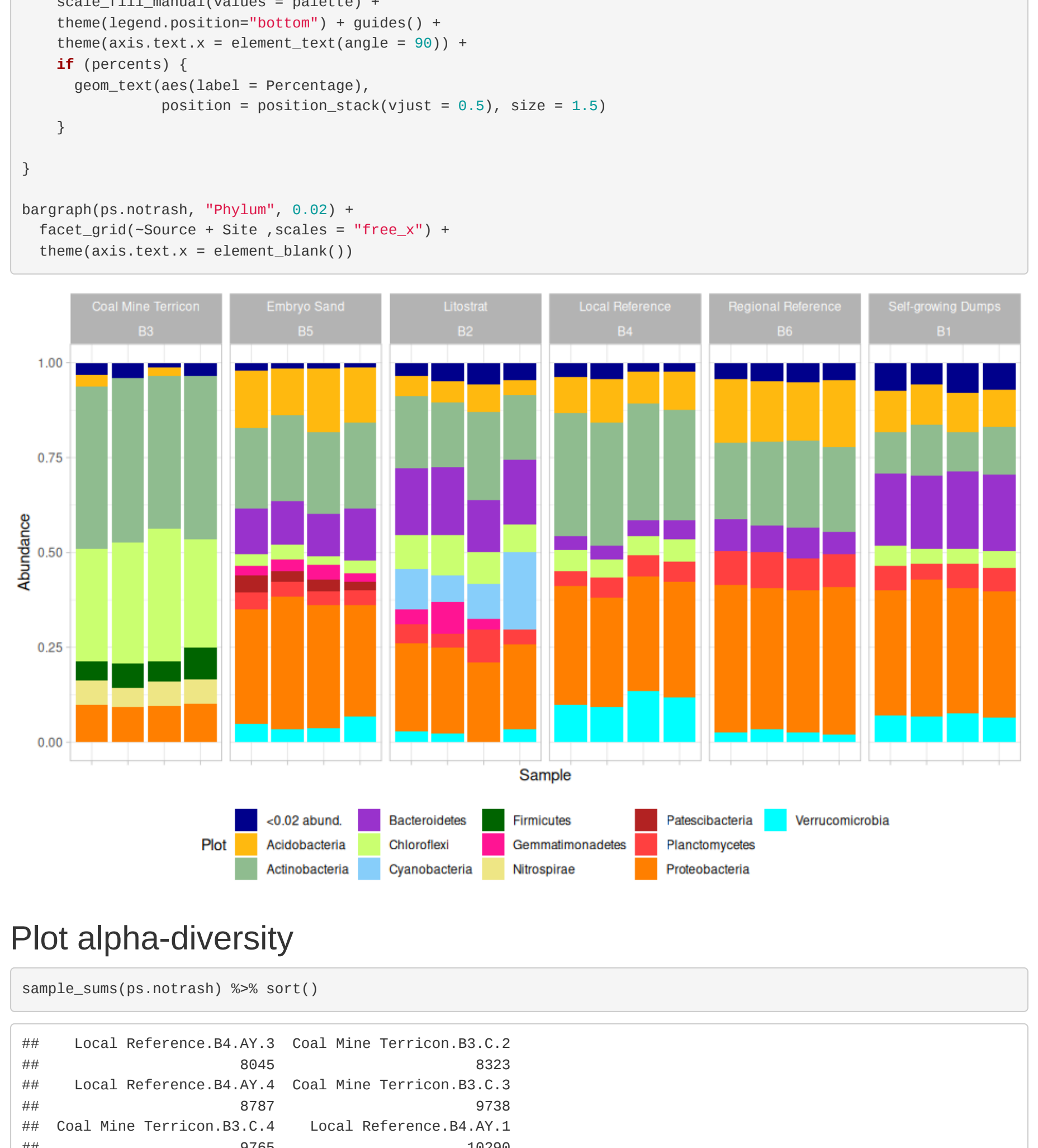


Remove chloroplast and mitochondria

```
# ps@tax_table %>% View()

ps.filtered <- subset_taxa(ps, Phylum != "NA")

asvs.keep <-ps@tax_table %>%
      data.frame() %>%
      filter((Family != "Mitochondria" & Order != "Chloroplast") %>%
              replace_na(TRUE)) %>%
      rownames()
ps.notrash <- prune_taxa(asvs.keep, ps.filtered)

# ps.notrash@tax_table %>% View()

saveRDS(ps.notrash, "ps.no.organells.RData")
```

Plot barplots

```
bargraph <- function(ps, rank, threshold=0.05, percents=FALSE){
  require(dplyr)
  require(ggplot2)
  require(phyloseq)

  ps <- prune_taxa(taxa_sums(ps) > 0, ps)
  ps2 <- tax_glom(ps, taxrank = rank)
  ps3 = transform_sample_counts(ps2, function(x) x / sum(x) )
  data <- psmelt(ps3) # create dataframe from phyloseq object
  data$Plot <- as.character(data[,rank]) # convert to character
  data$Plot[data$Abundance < threshold] <- paste0("<", threshold, " abund.")
  medians <- data %>% group_by(Plot) %>% mutate(median=median(data$Abundance))
  remainder <- medians[medians$median <= threshold,]$Plot
  data$Percentage = ifelse(data$Plot != paste0("<", threshold, " abund."),
                           round(data$Abundance, 3)*100, NA)

  # create palette long enough for our data
  base.palette <- c("darkblue", "darkgoldenrod1", "darkseagreen", "darkorchid", "darkolivegreen1", "lightskyblu
e",
                    "darkgreen", "deeppink", "khaki2", "firebrick", "brown1", "darkorange1", "cyan1", "royalblue
4",
                    "darksalmon", "dodgerblue3", "steelblue1", "darkgoldenrod1", "brown1", "cyan1", "darkgrey")
  required.colors <- nlevels(factor(data$Plot))
  repeats = required.colors %/% length(base.palette) + 1
  palette <- rep(base.palette, length.out = repeats * length(base.palette))

  p <- ggplot(data=data, aes(x=Sample, y=Abundance, fill=Plot))
  p + geom_bar(aes(), stat="identity", position="stack") + theme_light() +
    scale_fill_manual(values = palette) +
    theme(legend.position="bottom") + guides() +
    theme(axis.text.x = element_text(angle = 90)) +
    if (percents) {
      geom_text(aes(label = Percentage),
                position = position_stack(vjust = 0.5), size = 1.5)
    }

}

bargraph(ps.notrash, "Phylum", 0.02) +
  facet_grid(~Source + Site ,scales = "free_x") +
  theme(axis.text.x = element_blank())
```



Plot alpha-diversity

```
sample_sums(ps.notrash) %>% sort()
```

```
##    Local Reference.B4.AY.3  Coal Mine Terricon.B3.C.2
##                       8045                       8323
##    Local Reference.B4.AY.4  Coal Mine Terricon.B3.C.3
##                       8787                       9738
##  Coal Mine Terricon.B3.C.4    Local Reference.B4.AY.1
##                       9765                      10290
##    Local Reference.B4.AY.2 Regional Reference.B6.AY.4
##                      10668                      11550
##  Coal Mine Terricon.B3.C.1        Embryo Sand.B5.AY.4
##                      11604                      11908
## Regional Reference.B6.AY.3 Regional Reference.B6.AY.2
##                      12129                      12393
## Self-growing Dumps.B1.AY.4 Self-growing Dumps.B1.AY.2
##                      13071                      13831
##           Litostrat.B2.C.2 Self-growing Dumps.B1.AY.1
##                      14025                      14051
##        Embryo Sand.B5.AY.2 Regional Reference.B6.AY.1
##                      14194                      15441
##           Litostrat.B2.C.1        Embryo Sand.B5.AY.3
##                      16150                      16378
##           Litostrat.B2.C.3        Embryo Sand.B5.AY.1
##                      17593                      17657
##           Litostrat.B2.C.4 Self-growing Dumps.B1.AY.3
##                      18268                      19589
```

```
ps.raref <- rarefy_even_depth(ps.notrash, sample.size = 8000)
```

```
## You set `rngseed` to FALSE. Make sure you've set & recorded
##  the random seed of your session for reproducibility.
## See `?set.seed`
```

```
## ...
```

```
## 56OTUs were removed because they are no longer
## present in any sample after random subsampling
```

```
## ...
```

```
plot_richness(ps.raref, x = "Source", measures=c("Observed", "Shannon"), color = "Site")
```

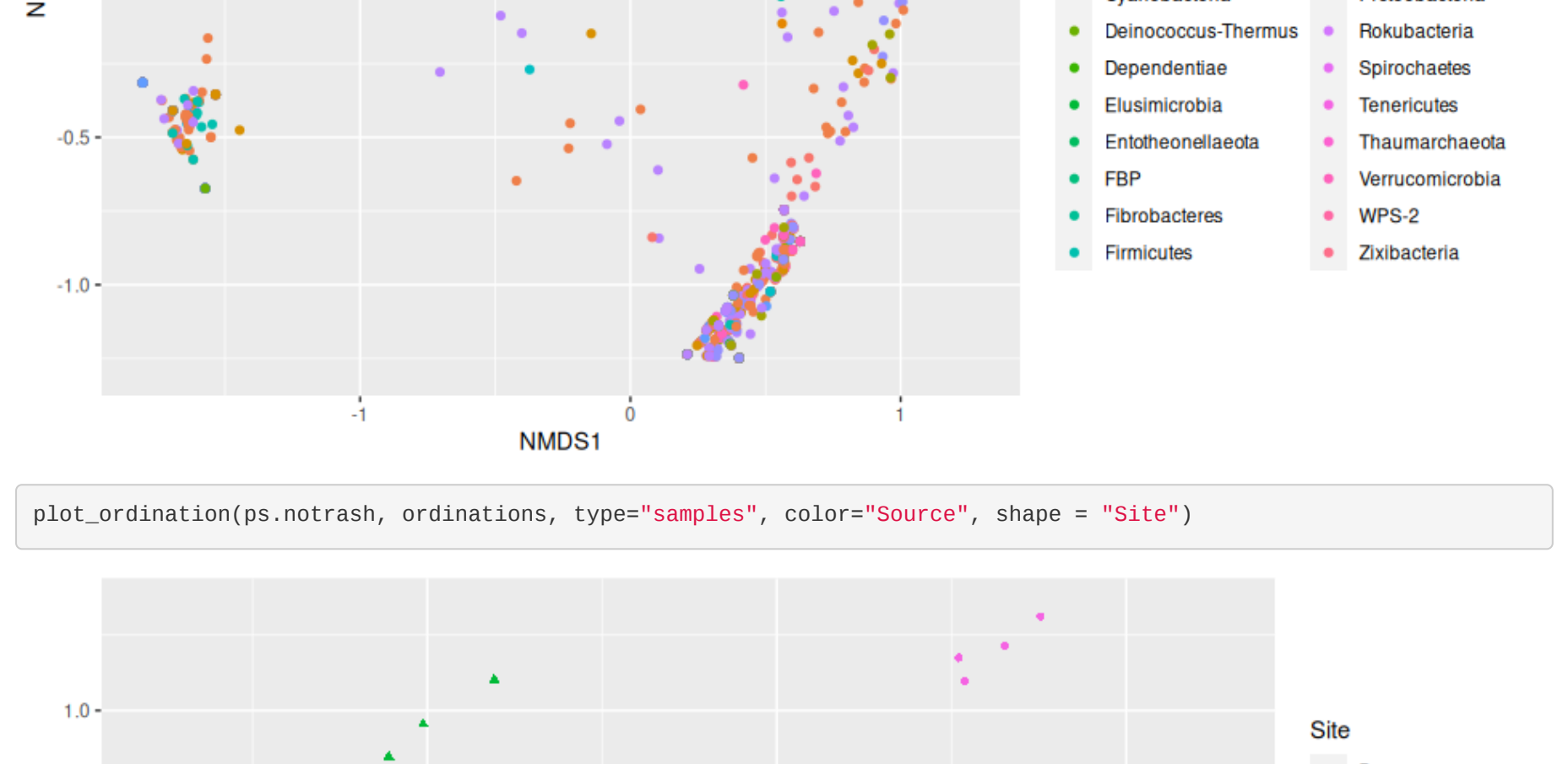
Beta-diversity

матрица попарных расстояний и поиск координат в алгоритме снижения размерностей по методу NMDS

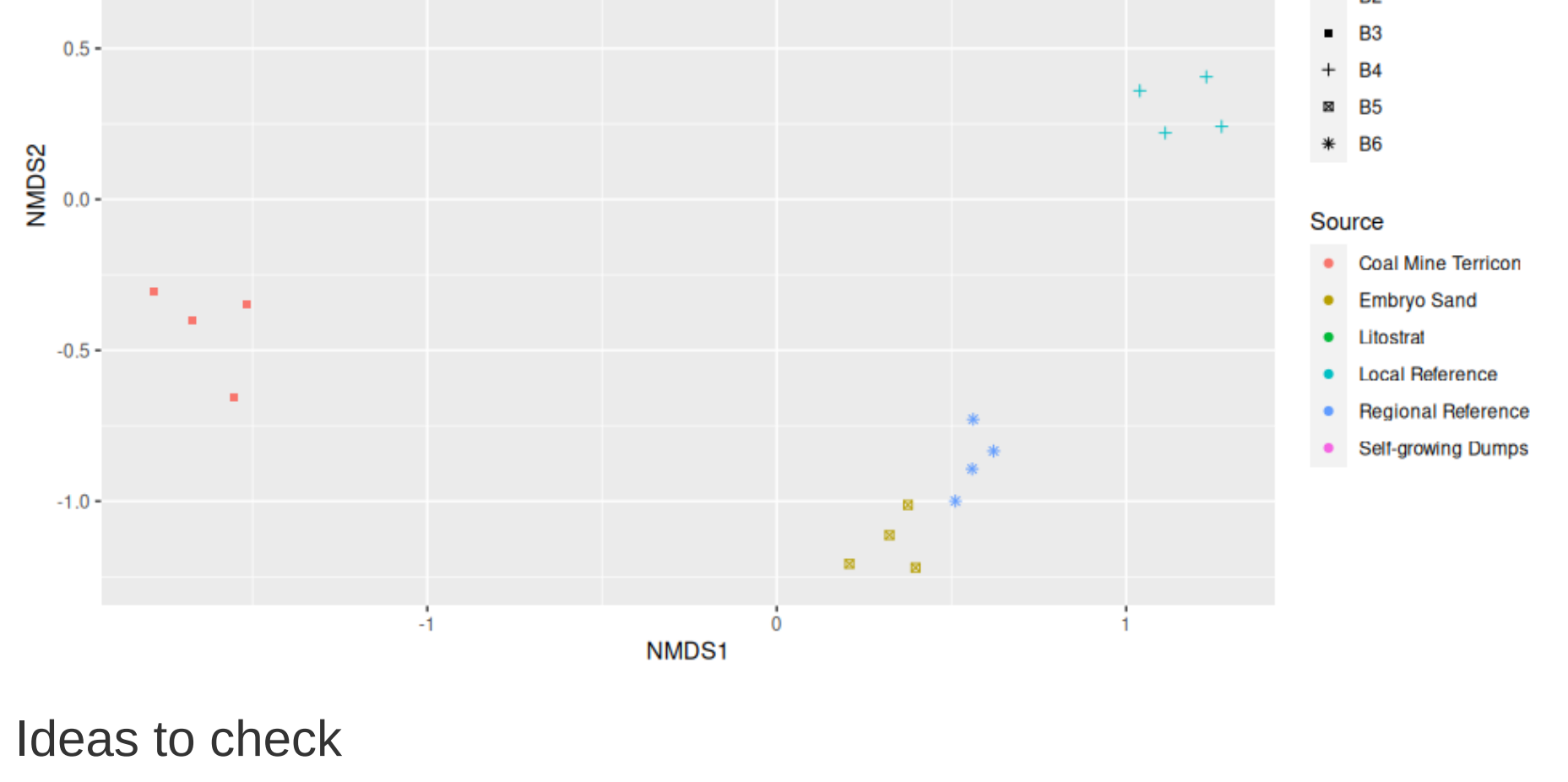
```
ordinations <- ordinate(ps.notrash, "NMDS", "bray")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.1237723
## Run 1 stress 0.1365955
## Run 2 stress 0.1237723
## ... New best solution
## ... Procrustes: rmse 3.761098e-06  max resid 8.491447e-06
## ... Similar to previous best
## Run 3 stress 0.1237723
## ... Procrustes: rmse 2.13041e-06  max resid 7.720301e-06
## ... Similar to previous best
## Run 4 stress 0.1237723
## ... Procrustes: rmse 4.295536e-06  max resid 7.97864e-06
## ... Similar to previous best
## Run 5 stress 0.1237723
## ... New best solution
## ... Procrustes: rmse 1.625311e-06  max resid 5.656985e-06
## ... Similar to previous best
## Run 6 stress 0.1237723
## ... Procrustes: rmse 3.035809e-06  max resid 8.191659e-06
## ... Similar to previous best
## Run 7 stress 0.1237723
## ... Procrustes: rmse 9.889131e-07  max resid 1.8773e-06
## ... Similar to previous best
## Run 8 stress 0.1365955
## Run 9 stress 0.1237723
## ... Procrustes: rmse 4.127966e-06  max resid 8.00787e-06
## ... Similar to previous best
## Run 10 stress 0.1237723
## ... Procrustes: rmse 2.282423e-06  max resid 4.995196e-06
## ... Similar to previous best
## Run 11 stress 0.1237723
## ... Procrustes: rmse 2.862817e-06  max resid 8.80411e-06
## ... Similar to previous best
## Run 12 stress 0.1365955
## Run 13 stress 0.1237723
## ... Procrustes: rmse 4.921051e-07  max resid 1.329029e-06
## ... Similar to previous best
## Run 14 stress 0.1237723
## ... Procrustes: rmse 1.43957e-06  max resid 2.977306e-06
## ... Similar to previous best
## Run 15 stress 0.1365955
## Run 16 stress 0.1237723
## ... Procrustes: rmse 2.310509e-06  max resid 4.713e-06
## ... Similar to previous best
## Run 17 stress 0.1237723
## ... Procrustes: rmse 3.062411e-06  max resid 5.73378e-06
## ... Similar to previous best
## Run 18 stress 0.1365955
## Run 19 stress 0.1237723
## ... Procrustes: rmse 3.737661e-06  max resid 6.878881e-06
## ... Similar to previous best
## Run 20 stress 0.1237723
## ... Procrustes: rmse 1.445691e-06  max resid 2.903068e-06
## ... Similar to previous best
## *** Best solution repeated 12 times
```

```
plot_ordination(ps.notrash, ordinations, type="taxa", color="Phylum")
```




```
plot_ordination(ps.notrash, ordinations, type="samples", color="Source", shape = "Site")
```



Ideas to check

1. Для сравнения разнообразия между группами можно использовать статастические тесты, например ANOVA
2. Можно сравнить равномерность сообществ между группами
3. Можно сравнить относительную абундантность филумов или таксонов
4. Также необходимо исследовать корреляционные связи между микробными сообществами