MB590-012 Microbiome Analysis

Christine V. Hawkes

March 2, 2021

Contents

Load and install R packages

```
library(phyloseq)
library(microbiome)
library(ggplot2)
library(tidyverse)
library(compositions)
library(rmarkdown)
library(knitr)
library(Biostrings)
library(vegan)

#install.packages("RColorBrewer")
library(RColorBrewer)
#install.packages("reshape2")
library(reshape2)
#devtools::install_github("Russel88/MicEco")
library(MicEco)
```

CODING EXERCISES

Please submit as a knitted html or pdf markdown to GitHub due on 3/9

1. Subset to Vaccinium unique OTUs and clr transform

- Goal is to retain only fungal OTUs found uniquely associated with Vaccinium by removing Pinus OTUs
- p1 venn diagram can help you to confirm expected numbers
- use `phylosmith::unique_taxa` to identify taxa associated only with Pinus in ps_nosing
  - https://schuyler-smith.github.io/phylosmith/analytics.html#unique_taxa
  - `devtools::install_github("schuyler-smith/phylosmith")`
  - `library(phylosmith)`
    library(phylosmith)
- convert list to vector using base::unlist
  - https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/unlist
  - note that this gives you unique Pinus OTUs + OTUs shared with Pinus
- export list of all taxa with `phyloseq::taxa_names` from ps_nosing
  - make new ps object ps_vacc by subsetting the list by removing taxa from Pinus
  - hint: look back at code from lulu
- in new ps object, ps_vacc
  - use remaining taxa list to retain only truly unique taxa with `phyloseq::prune_taxa`
  - use `phyloseq::subset_samples` to limit to Species==“Vaccinium”
  - check for and remove new singletons
- create ps_vacc_clr with clr transformed otu_table using `microbiome::transform`
- include new Vaccinium venn diagram by EcoType
- optional: if you have time and want to practice more, repeat for Pinus

```
# Subset the ps object for Vacc
library(phylosmith)
pine_uniq <- phylosmith::unique_taxa(ps_nosing, "Species", subset="Pinus")
pine_uniq <- unlist(pine_uniq) # convert from list to vector
str(pine_uniq) # confirm conversion
```

```
##  Named chr [1:493] "OTU4" "OTU5" "OTU1" "OTU3" "OTU16" "OTU2" "OTU7" ...
##  - attr(*, "names")= chr [1:493] "Pinus1" "Pinus2" "Pinus3" "Pinus4" ...
```

```
length(pine_uniq) # count num of taxa = 493 (pine + pine shared with vacc)
```

```
## [1] 493
```

```
allTaxa <- phyloseq::taxa_names(ps_nosing) # make vector of taxa names
vaccTaxa <- allTaxa[!(allTaxa %in% pine_uniq)] # remove pine taxa
ps_vacc <- phyloseq::prune_taxa(vaccTaxa, ps_nosing) # retain vacc only taxa in ps obj
ps_vacc # confirm 699 taxa
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 699 taxa and 85 samples ]
## sample_data() Sample Data:       [ 85 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 699 taxa by 5 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 699 reference sequences ]
```

```
ps_vacc <- phyloseq::subset_samples(ps_vacc, Species == "Vaccinium")
ps_vacc # confirm that # of samples is reduced from 85 to 38
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 699 taxa and 38 samples ]
## sample_data() Sample Data:       [ 38 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 699 taxa by 5 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 699 reference sequences ]
```

```
phyloseq::sample_names(ps_vacc) #check that these are only "TVxxx" samples
```

```
##  [1] "T2V10" "T2V1"  "T2V2"  "T2V3"  "T2V4"  "T2V5"  "T2V6"  "T2V7"  "T2V8" 
## [10] "T2V9"  "T3V10" "T3V1"  "T3V2"  "T3V3"  "T3V4"  "T3V5"  "T3V6"  "T3V7" 
## [19] "T3V8"  "T3V9"  "T4V10" "T4V1"  "T4V2"  "T4V3"  "T4V4"  "T4V6"  "T4V8" 
## [28] "T4V9"  "T5V10" "T5V1"  "T5V2"  "T5V3"  "T5V4"  "T5V5"  "T5V6"  "T5V7" 
## [37] "T5V8"  "T5V9"
```

```
# check for and remove singletons if needed
ps_vacc <- phyloseq::prune_taxa(phyloseq::taxa_sums(ps_vacc) > 1, ps_vacc)
phyloseq::ntaxa(ps_vacc) # no losses
```

```
## [1] 699
```

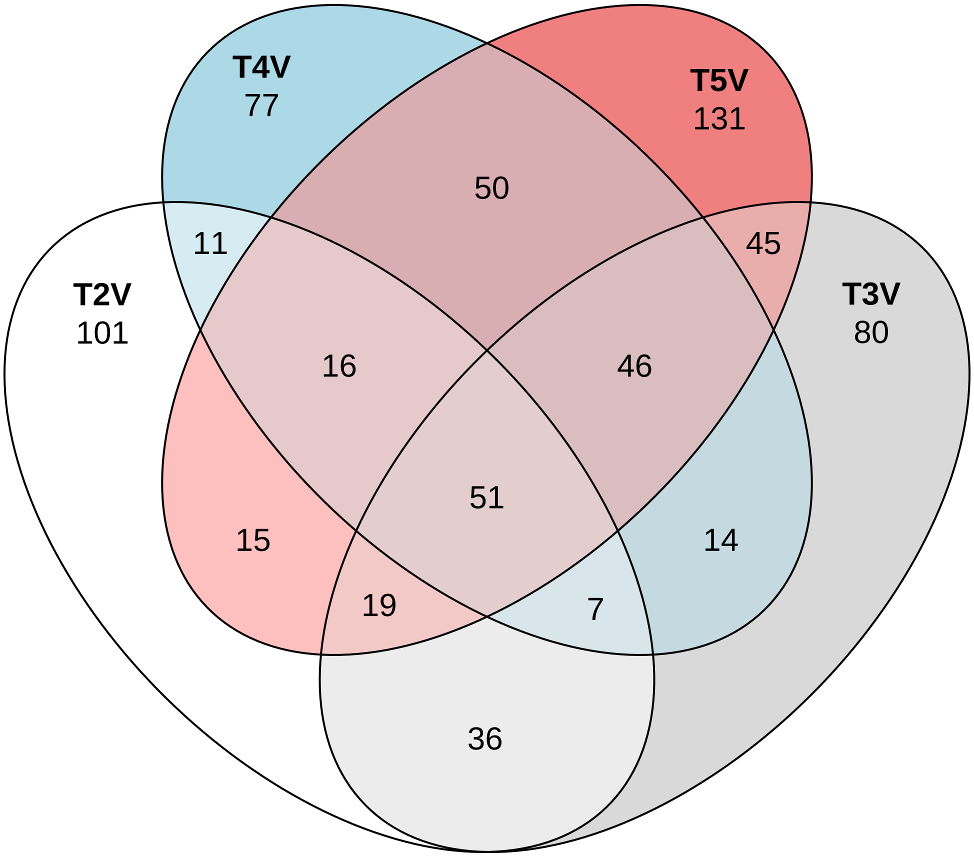
```
# check for and remove samples with zero row sums if needed
phyloseq::nsamples(ps_vacc)
```

```
## [1] 38
```

```
ps_vacc <- phyloseq::prune_samples(phyloseq::sample_sums(ps_vacc)>0, ps_vacc)
phyloseq::nsamples(ps_vacc) # no change 38 samples
```

```
## [1] 38
```

```
# Venn diagram
MicEco::ps_venn(ps_vacc, "EcoType", fraction=0, weight=FALSE, type="counts", relative=FALSE, plot=TRUE)
```



```
# 51 taxa found in all EcoTypes

# transforms
ps_vacc_ra <- microbiome::transform(ps_vacc, "compositional")
ps_vacc_clr <- microbiome::transform(ps_vacc, transform="clr")


# Optional
# Subset the ps object for Pine

vacc_uniq <- phylosmith::unique_taxa(ps_nosing, "Species", subset="Vaccinium")
vacc_uniq  <- unlist(vacc_uniq)
str(vacc_uniq)
```

```
##  Named chr [1:1052] "OTU4" "OTU5" "OTU1" "OTU3" "OTU16" "OTU2" "OTU7" ...
##  - attr(*, "names")= chr [1:1052] "Vaccinium1" "Vaccinium2" "Vaccinium3" "Vaccinium4" ...
```

```
length(vacc_uniq) #1052 (vacc + vacc shared with pine)
```

```
## [1] 1052
```

```
pineTaxa <- allTaxa[!(allTaxa %in% vacc_uniq)] # remove vacc taxa
ps_pine <- phyloseq::prune_taxa(pineTaxa, ps_nosing) # retain pine only taxa in ps obj
ps_pine # confirm 140 taxa
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 140 taxa and 85 samples ]
## sample_data() Sample Data:       [ 85 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 140 taxa by 5 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 140 reference sequences ]
```

```
ps_pine <-phyloseq::subset_samples(ps_pine, Species == "Pinus")
ps_pine # confirm that # of samples is reduced from 85 to 47
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 140 taxa and 47 samples ]
## sample_data() Sample Data:       [ 47 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 140 taxa by 5 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 140 reference sequences ]
```

```
phyloseq::sample_names(ps_pine) #check that "TVxxx" samples are gone
```

```
##  [1] "T1P10" "T1P1"  "T1P2"  "T1P3"  "T1P4"  "T1P5"  "T1P6"  "T1P7"  "T1P8" 
## [10] "T1P9"  "T2P10" "T2P1"  "T2P2"  "T2P3"  "T2P5"  "T2P8"  "T2P9"  "T3P10"
## [19] "T3P1"  "T3P2"  "T3P3"  "T3P4"  "T3P5"  "T3P6"  "T3P7"  "T3P8"  "T3P9" 
## [28] "T4P10" "T4P1"  "T4P2"  "T4P3"  "T4P4"  "T4P5"  "T4P6"  "T4P7"  "T4P8" 
## [37] "T4P9"  "T5P10" "T5P1"  "T5P2"  "T5P3"  "T5P4"  "T5P5"  "T5P6"  "T5P7" 
## [46] "T5P8"  "T5P9"
```

```
# check for and remove singletons if needed
ps_pine <- phyloseq::prune_taxa(phyloseq::taxa_sums(ps_pine) > 1, ps_pine)
phyloseq::ntaxa(ps_pine) # no losses
```

```
## [1] 140
```

```
# check for and remove samples with zero row sums if needed
nsamples(ps_pine)
```

```
## [1] 47
```

```
ps_pine <- phyloseq::prune_samples(phyloseq::sample_sums(ps_pine)>0, ps_pine)
nsamples(ps_pine) # no change 47 samples
```

```
## [1] 47
```

2. **Examine core microbiome for Vaccinium only**

- for one detection and prevalence level, compare clr and rel abund data transforms
- vary detection and prevalence for clr data
  - adjust only detection up and down (at least 3 levels)
  - adjust only prevalence up and down (at least 3 levels)
- describe the effects on the size and characteristics of the core community
- optional: if you want more practice, repeat for Pinus

```
# Repeat the following code set with different detection and prevalence settings
ps_vacc_ra_core <- microbiome::core(ps_vacc_ra, detection = 0.001, prevalence = 50/100)
ps_vacc_ra_core
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 8 taxa and 38 samples ]
## sample_data() Sample Data:       [ 38 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 8 taxa by 5 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 8 reference sequences ]
```

```
microbiome::taxa(ps_vacc_ra_core)
```

```
## [1] "OTU8"    "OTU3130" "OTU35"   "OTU42"   "OTU110"  "OTU4625" "OTU366" 
## [8] "OTU4549"
```

```
ps_vacc_clr_core <- microbiome::core(ps_vacc_clr, detection = 0.001, prevalence = 50/100)
ps_vacc_clr_core
```

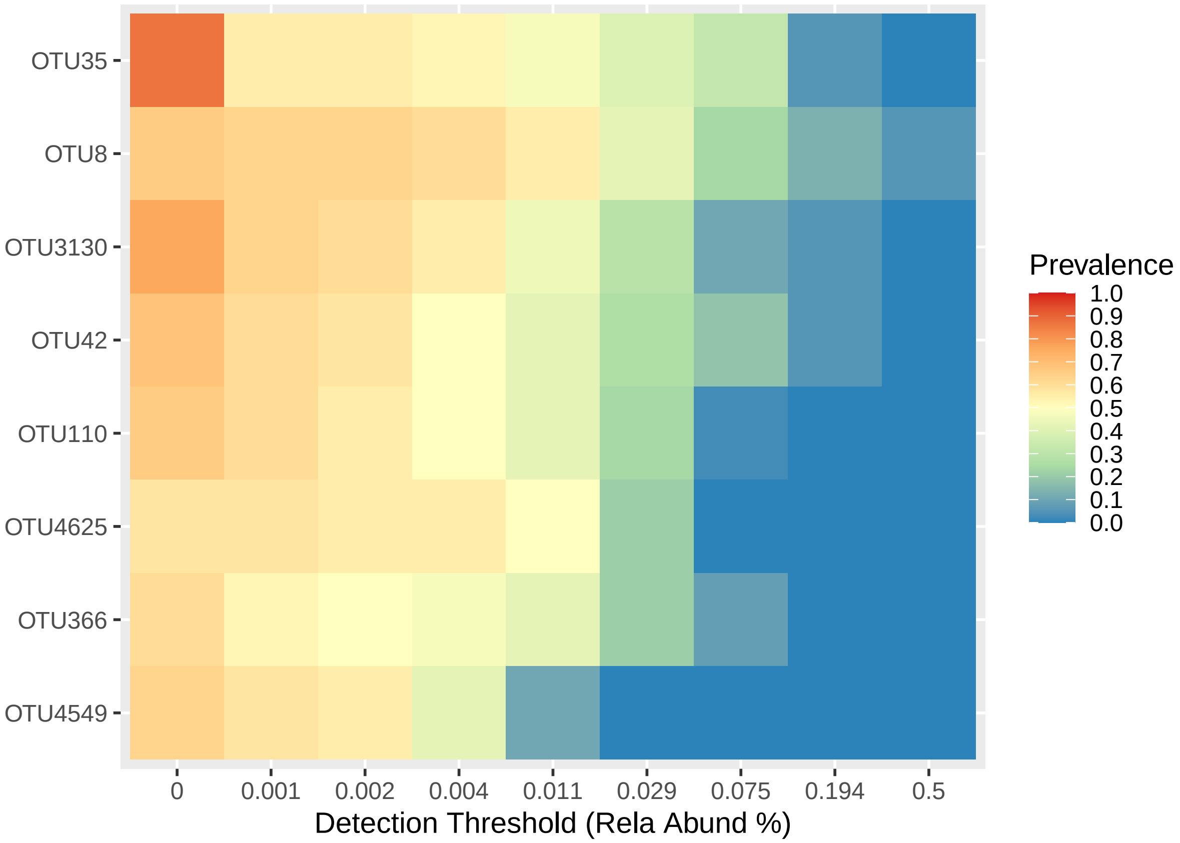
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 18 taxa and 38 samples ]
## sample_data() Sample Data:       [ 38 samples by 31 sample variables ]
## tax_table()   Taxonomy Table:    [ 18 taxa by 5 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 18 reference sequences ]
```

```
microbiome::taxa(ps_vacc_clr_core)
```

```
##  [1] "OTU8"    "OTU3130" "OTU35"   "OTU26"   "OTU42"   "OTU107"  "OTU110" 
##  [8] "OTU4032" "OTU4625" "OTU89"   "OTU366"  "OTU4518" "OTU4549" "OTU356" 
## [15] "OTU1638" "OTU2502" "OTU4573" "OTU2728"
```

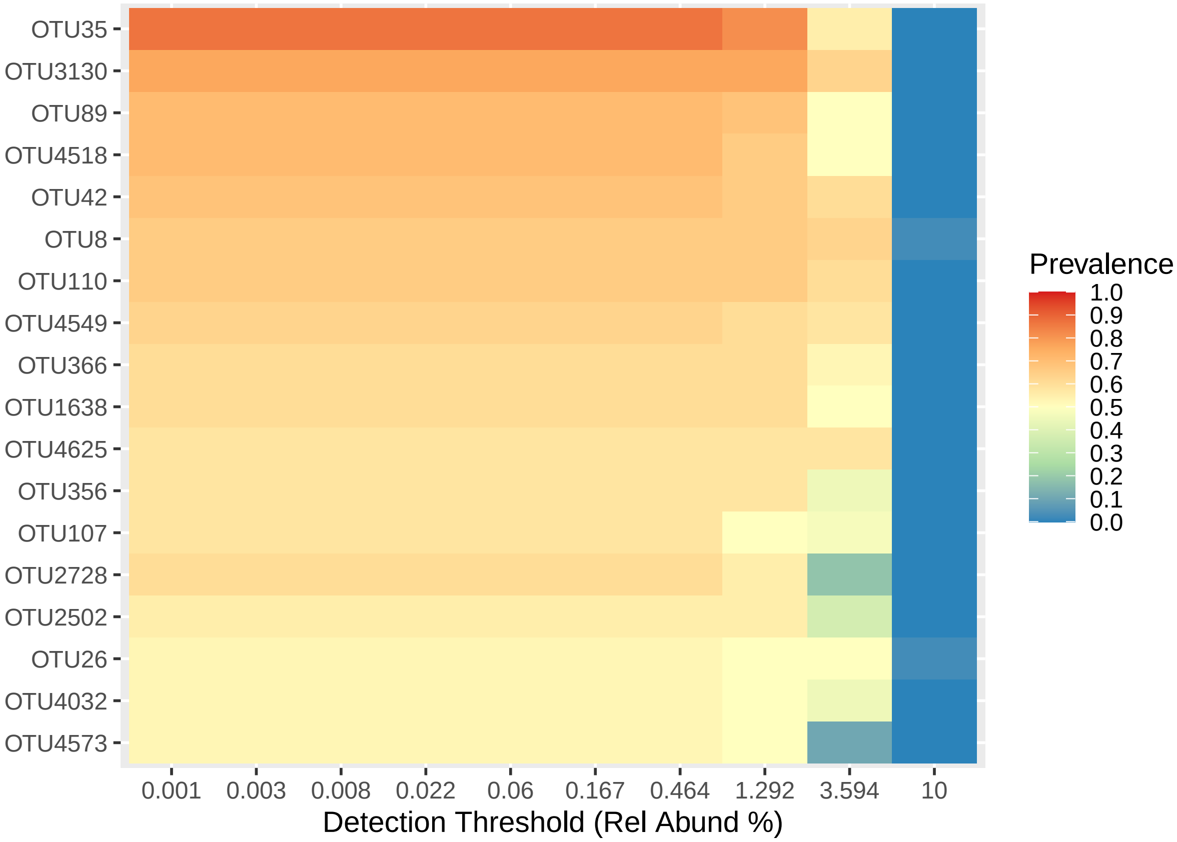
```
# Heatmap of core - note that clr shifts everything left to lower rel abund
prevalences <- seq(0.05, 1, 0.05)
detections <-  round(10^seq(log10(1e-4), log10(0.5), length = 10), 3)

microbiome::plot_core(ps_vacc_ra_core, plot.type = "heatmap",
            colours = rev(RColorBrewer::brewer.pal(5, "Spectral")),
            prevalences = prevalences,
            detections = detections) +
            ggplot2::labs(x = "Detection Threshold (Rela Abund %)")
```

```
## Warning in microbiome::plot_core(ps_vacc_ra_core, plot.type = "heatmap", : The plot_core function is
##                   data. The data is not compositional. Make sure that you
##                   intend to operate on non-compositional data.
```

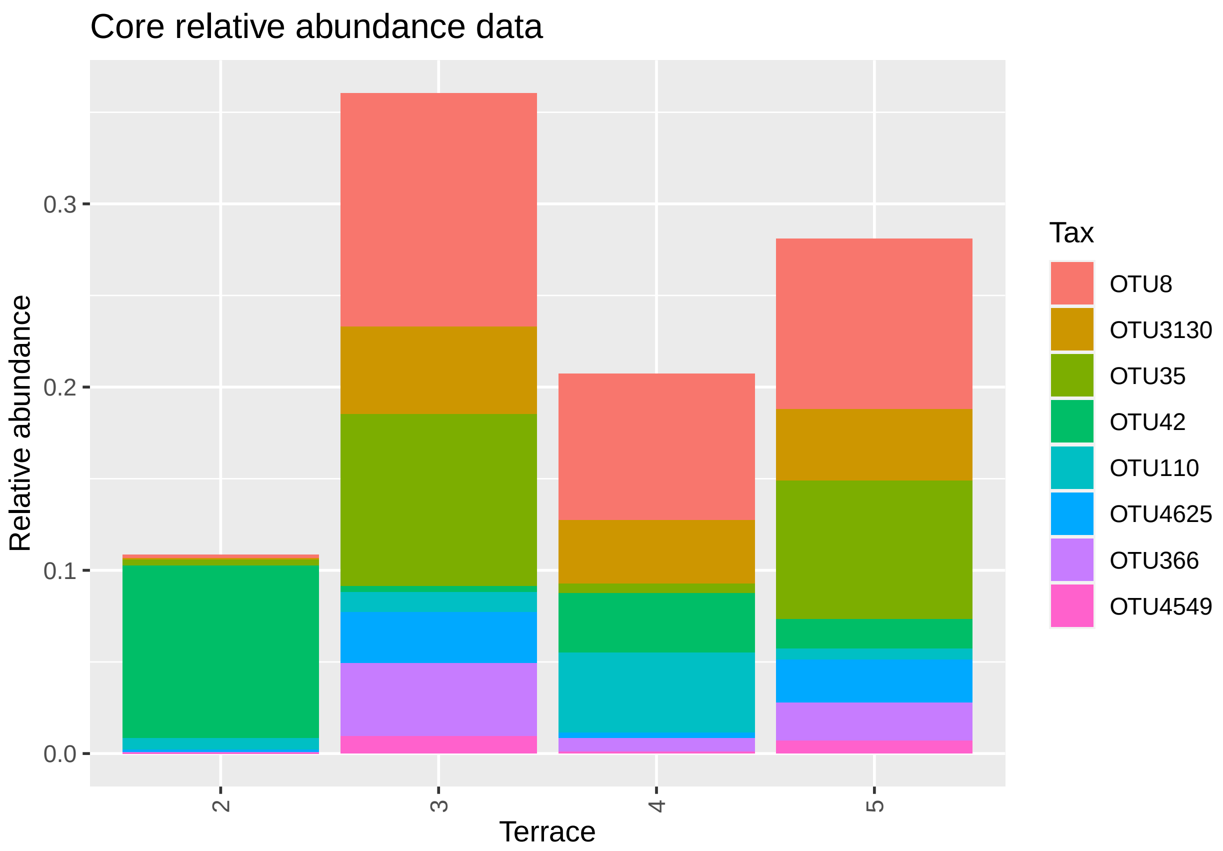


```
prevalences <- seq(0.05, 1, 0.05)
detections <-  round(10^seq(log10(1e-3), log10(10), length = 10), 3)

microbiome::plot_core(ps_vacc_clr_core, plot.type = "heatmap",
            colours = rev(RColorBrewer::brewer.pal(5, "Spectral")),
            prevalences = prevalences,
            detections = detections) +
            ggplot2::labs(x = "Detection Threshold (Rel Abund %)")
```

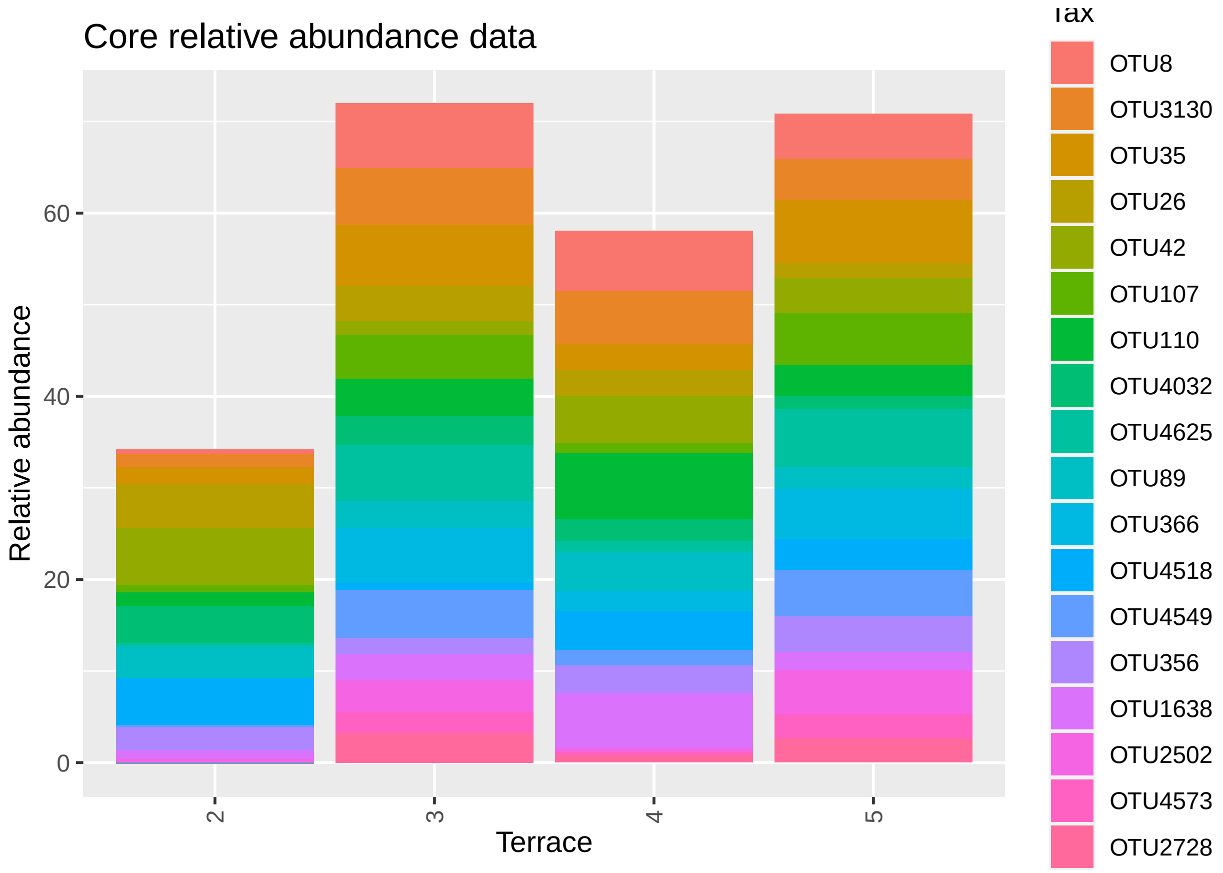
```
## Warning in microbiome::plot_core(ps_vacc_clr_core, plot.type = "heatmap", : The plot_core function i
##                   data. The data is not compositional. Make sure that you
##                   intend to operate on non-compositional data.
```



```
# Barplot of core taxa by terrace

microbiome::plot_composition(ps_vacc_ra_core,
                             average_by="Terrace",
                             plot.type = "barplot",
                             sample.sort="Terrace") +
                 guides(fill = guide_legend(ncol = 1)) +
                 labs(x = "Terrace",
                 y = "Relative abundance",
                title = "Core relative abundance data")
```



```
microbiome::plot_composition(ps_vacc_clr_core,
                             average_by="Terrace",
                             plot.type = "barplot",
                             sample.sort="Terrace") +
                 guides(fill = guide_legend(ncol = 1)) +
                 labs(x = "Terrace",
                 y = "Relative abundance",
                title = "Core relative abundance data")
```


Core relative abundance data
Relative abundance
0
20
40
60
2
3
4
5
Terrace
Tax
OTU8
OTU3130
OTU35
OTU26
OTU42
OTU107
OTU110
OTU4032
OTU4625
OTU89
OTU366
OTU4518
OTU4549
OTU356
OTU1638
OTU2502
OTU4573
OTU2728

3. **Identify the core microbiota of built-in `soilrep` data**

- load built-in `soilrep` data and examine
- remove clipped samples with `phyloseq::subset_samples`
- remove singletons with `phyloseq::prune_taxa`
- identify core with `microbiome::core`
  - indicate why you selected your specific prevalence and detection settings
- produce a table of core ASVs using `kable` and specify column name
- plot results as heatmap, barplot, or other plot of your choice

```
library(phyloseq)
data(soilrep)
ps_sr <- soilrep
sample_data(ps_sr)
```

```
##        Treatment warmed clipped Sample
## a_C026        UC     no     yes    6CC
## a_C066        UU     no      no    3UC
## a_C070        WU    yes      no    5UW
## a_C074        UU     no      no    2UC
## a_C075        WC    yes     yes    5CW
## a_C077        WU    yes      no    4UW
## a_C079        UU     no      no    6UC
## a_C081        UC     no     yes    3CC
## a_C082        UC     no     yes    1CC
## a_C083        UU     no      no    6UC
## a_C084        WC    yes     yes    4CW
## a_C085        UC     no     yes    1CC
## a_C086        UC     no     yes    3CC
## a_C088        WC    yes     yes    5CW
## a_C089        WC    yes     yes    1CW
## a_C090        WC    yes     yes    3CW
## a_C091        UU     no      no    2UC
## a_C093        UU     no      no    2UC
## a_C095        WU    yes      no    6UW
## a_C096        WC    yes     yes    1CW
## a_C098        WU    yes      no    5UW
## a_C099        UU     no      no    4UC
## a_C100        WU    yes      no    2UW
## a_C101        WU    yes      no    3UW
## a_C102        WC    yes     yes    3CW
## a_C116        UC     no     yes    3CC
## a_C125        UC     no     yes    4CC
## a_C126        WU    yes      no    2UW
## a_C127        WU    yes      no    6UW
## a_C128        WU    yes      no    4UW
## a_C130        WC    yes     yes    4CW
## a_C131        UU     no      no    6UC
## a_C132         WC    yes     yes    5CW
## a_C134         UC     no     yes    6CC
## a_C136         WC    yes     yes    1CW
## a_C137         WC    yes     yes    6CW
## a_C139         UC     no     yes    2CC
## a_C140         UU     no      no    3UC
## a_C141         UU     no      no    1UC
## a_C143         WU    yes      no    3UW
## a_C144         WU    yes      no    4UW
## a_C145         WC    yes     yes    6CW
## a_C146         UU     no      no    5UC
## a_C147         WU    yes      no    2UW
## a_C149         UU     no      no    1UC
## a_C150         UC     no     yes    4CC
## a_C151         WC    yes     yes    3CW
## a_C153         WU    yes      no    1UW
## a_C154         UC     no     yes    2CC
## a_C156         UC     no     yes    1CC
## a_C157         UC     no     yes    5CC
## a_C158         WU    yes      no    6UW
## a_C159         UU     no      no    4UC
## a_C160         WC    yes     yes    2CW
## a_C161         UC     no     yes    5CC
## a_C162         UU     no      no    5UC
```

```
# remove clipped samples (goes from 56 to 28 samples)
ps_sr <- phyloseq::subset_samples(ps_sr, clipped == "no")
ps_sr
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 16825 taxa and 28 samples ]
## sample_data() Sample Data:       [ 28 samples by 4 sample variables ]
```

```
# remove singletons (reduces from 16825 to 7250 taxa)
ps_sr <- phyloseq::prune_taxa(taxa_sums(ps_sr) > 1, ps_sr)
phyloseq::ntaxa(ps_sr)
```

```
## [1] 7250
```

```
# check for any samples with zero counts
phyloseq::sample_sums(ps_sr) #none
```

```
## a_C066 a_C070 a_C074 a_C077 a_C079 a_C083 a_C091 a_C093 a_C095 a_C098 a_C099
##   1587   1858   3868   1986   1368   1991   1128   3852   3321   1506   1572
## a_C100 a_C101 a_C126 a_C127 a_C128 a_C131 a_C140 a_C141 a_C143 a_C144 a_C146
##   1543   1833    832    939   1445   1139   1241    862   1217   1166   1038
## a_C147 a_C149 a_C153 a_C158 a_C159 a_C162
##   1065   1218   2104   1104   1192   1393
```

```
# identify core - here using stringent prevalence criteria and moderate detection
ps_sr_core <- microbiome::core(ps_sr, detection = 50/100, prevalence = 90/100)
ps_sr_core
```

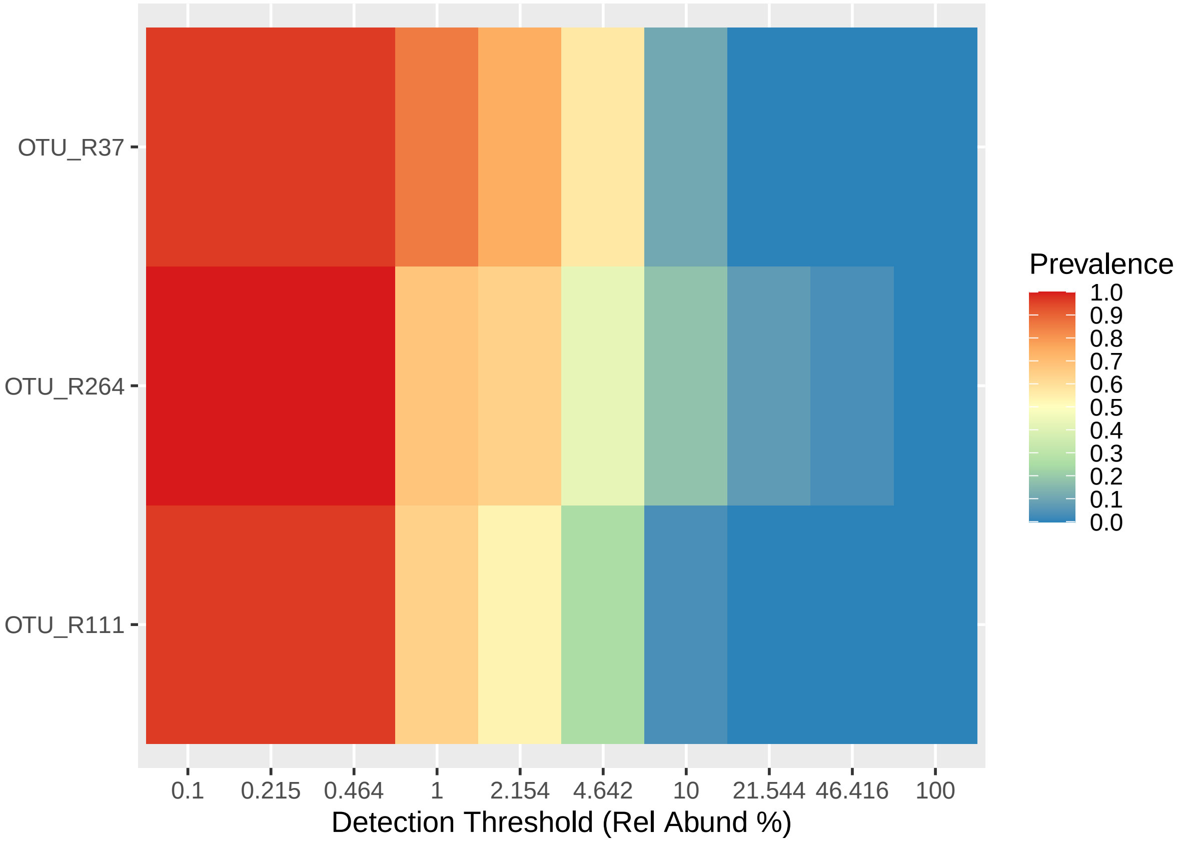
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 3 taxa and 28 samples ]
## sample_data() Sample Data:       [ 28 samples by 4 sample variables ]
```

```
# core table
knitr::kable(taxa(ps_sr_core), col.names=c("core ASVs"))
```

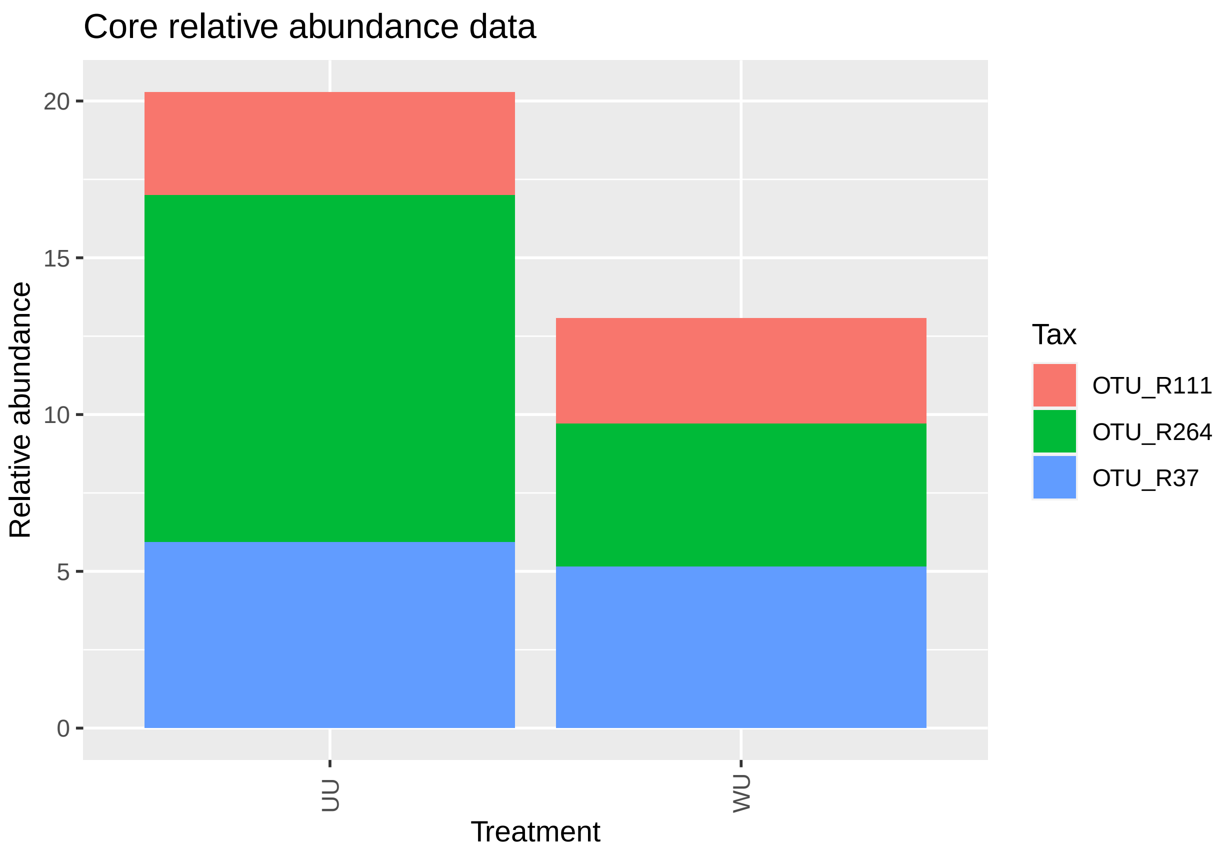
core ASVs
OTU_R111
OTU_R264
OTU_R37

```
# plot options
prevalences <- seq(0.05, 1, 0.05)
detections <- round(10^seq(log10(1e-1), log10(100), length = 10), 3)

microbiome::plot_core(ps_sr_core, plot.type = "heatmap",
          colours = rev(RColorBrewer::brewer.pal(5, "Spectral")),
          prevalences = prevalences,
          detections = detections) +
          ggplot2::labs(x = "Detection Threshold (Rel Abund %)")
```

```
## Warning in microbiome::plot_core(ps_sr_core, plot.type = "heatmap", colours = rev(RColorBrewer::brew
##                   data. The data is not compositional. Make sure that you
##                   intend to operate on non-compositional data.
```



```
microbiome::plot_composition(ps_sr_core,
                             average_by="Treatment",
                             plot.type = "barplot",
                             sample.sort="Treatment") +
                 guides(fill = guide_legend(ncol = 1)) +
                 labs(x = "Treatment",
                 y = "Relative abundance",
                title = "Core relative abundance data")
```


Core relative abundance data
Relative abundance
20
15
10
5
0
UU
WU
Treatment
Tax
OTU_R111
OTU_R264
OTU_R37

Session Info

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] phylosmith_1.0.6    MicEco_0.9.17       reshape2_1.4.4
##  [4] RColorBrewer_1.1-2  vegan_2.5-7         lattice_0.20-44
##  [7] permute_0.9-7       Biostrings_2.60.2   GenomeInfoDb_1.28.4
## [10] XVector_0.32.0      IRanges_2.26.0      S4Vectors_0.30.2
## [13] BiocGenerics_0.38.0 knitr_1.37          rmarkdown_2.12
## [16] compositions_2.0-4  forcats_0.5.1       stringr_1.4.0
## [19] dplyr_1.0.8         purrr_0.3.4         readr_2.1.2
## [22] tidyr_1.2.0         tibble_3.1.6        tidyverse_1.3.1
## [25] microbiome_1.14.0   ggplot2_3.3.5       phyloseq_1.36.0
##
## loaded via a namespace (and not attached):
##   [1] readxl_1.3.1           snow_0.4-4             backports_1.4.1
##   [4] Hmisc_4.6-0            plyr_1.8.6             igraph_1.2.11
##   [7] polylabelr_0.2.0       splines_4.1.1          digest_0.6.29
##  [10] foreach_1.5.2          htmltools_0.5.2        viridis_0.6.2
##  [13] fansi_1.0.2            magrittr_2.0.2         checkmate_2.0.0
##  [16] cluster_2.1.2          tzdb_0.2.0             graphlayouts_0.8.0
##  [19] modelr_0.1.8           RcppParallel_5.1.5     bayesm_3.1-4
##  [22] bdsmatrix_1.3-4        jpeg_0.1-9             colorspace_2.0-3
##  [25] ggrepel_0.9.1          rvest_1.0.2            haven_2.4.3
##  [28] xfun_0.29              crayon_1.5.0           RCurl_1.98-1.6
##  [31] jsonlite_1.8.0         survival_3.2-11        iterators_1.0.14
##  [34] ape_5.6-2              glue_1.6.2             polyclip_1.10-0
##  [37] gtable_0.3.0           zlibbioc_1.38.0        Rhdf5lib_1.14.2
##  [40] DEoptimR_1.0-10        abind_1.4-5            scales_1.1.1
##  [43] pheatmap_1.0.12        mvtnorm_1.1-3          DBI_1.1.2
##  [46] Rcpp_1.0.8             viridisLite_0.4.0      htmlTable_2.4.0
##  [49] units_0.8-0            proxy_0.4-26           foreign_0.8-81
##  [52] Formula_1.2-4          htmlwidgets_1.5.4      httr_1.4.2
##  [55] ellipsis_0.3.2         pkgconfig_2.0.3        farver_2.1.0
##  [58] nnet_7.3-16            dbplyr_2.1.1           utf8_1.2.2
##  [61] tidyselect_1.1.2       labeling_0.4.2         rlang_1.0.1
##  [64] munsell_0.5.0          cellranger_1.1.0       tools_4.1.1
##  [67] cli_3.2.0              generics_0.1.2         ade4_1.7-18
##  [70] broom_0.7.12           evaluate_0.15          biomformat_1.20.0
##  [73] fastmap_1.1.0          yaml_2.3.5             fs_1.5.2
##  [76] tidygraph_1.2.0        robustbase_0.93-9      ggraph_2.0.5
##  [79] nlme_3.1-152           xml2_1.3.3             compiler_4.1.1
##  [82] rstudioapi_0.13        png_0.1-7              e1071_1.7-9
##  [85] reprex_2.0.1           tweenr_1.0.2           stringi_1.7.6
##  [88] highr_0.9              Matrix_1.3-4           classInt_0.4-3
##  [91] tensorA_0.36.2         multtest_2.48.0        vctrs_0.3.8
##  [94] pillar_1.7.0           lifecycle_1.0.1        rhdf5filters_1.4.0
##  [97] eulerr_6.1.1           data.table_1.14.2      bitops_1.0-7
## [100] R6_2.5.1               latticeExtra_0.6-29    KernSmooth_2.23-20
## [103] gridExtra_2.3          codetools_0.2-18       MASS_7.3-54
## [106] assertthat_0.2.1       picante_1.8.2          rhdf5_2.36.0
## [109] withr_2.5.0            GenomeInfoDbData_1.2.6 mgcv_1.8-36
## [112] hms_1.1.1              doSNOW_1.0.20          grid_4.1.1
## [115] rpart_4.1-15           class_7.3-19           Rtsne_0.15
## [118] sf_1.0-7               ggforce_0.3.3          bbmle_1.0.24
## [121] numDeriv_2016.8-1.1    Biobase_2.52.0         lubridate_1.8.0
## [124] base64enc_0.1-3
```