Microbial community analysis in R

Marko Suokas

#### Libraries

library(tidyverse);packageVersion("tidyverse")

[1] '2.0.0'

library(kableExtra);packageVersion("kableExtra")

[1] '1.4.0'

library(patchwork);packageVersion("patchwork")

[1] '1.2.0'

library(mia);packageVersion("mia")

[1] '1.12.0'

library(ggplot2);packageVersion("ggplot2")

[1] '3.5.1'

library(ggthemes);packageVersion("ggthemes")

[1] '5.1.0'

Reload tse objects

tse\_dada <- readRDS("set1/tse\_dada.rds")  
tse\_vs97 <- readRDS("set1/tse\_vs97.rds")  
tse\_vs99 <- readRDS("set1/tse\_vs99.rds")  
tse\_emu <- readRDS("set1/tse\_emu.rds")

Agglomerate data to genus level

#agglomeration to genus level  
tse\_dada<- agglomerateByRank(tse\_dada, rank = "Genus", onRankOnly = T,  
 na.rm = F)  
tse\_vs97 <- agglomerateByRank(tse\_vs97, rank = "Genus", onRankOnly = T,  
 na.rm = F)  
tse\_vs99 <- agglomerateByRank(tse\_vs99, rank = "Genus", onRankOnly = T,  
 na.rm = F)  
tse\_emu <- agglomerateByRank(tse\_emu, rank = "Genus", onRankOnly = T,  
 na.rm = F)  
#check number of variants  
nrow(tse\_dada)

[1] 19

nrow(tse\_vs97)

[1] 38

nrow(tse\_vs99)

[1] 34

nrow(tse\_emu)

[1] 22

Next, we convert counts to relative abundance values

#relabundance  
tse\_dada <- transformAssay(tse\_dada, assay.type = "counts",  
 method = "relabundance")  
tse\_vs97 <- transformAssay(tse\_vs97, assay.type = "counts",  
 method = "relabundance")  
tse\_vs99 <- transformAssay(tse\_vs99, assay.type = "counts",  
 method = "relabundance")  
tse\_emu <- transformAssay(tse\_emu, assay.type = "counts",  
 method = "relabundance")

Pick five most abundant features

#get top5 features  
top5\_dada <- getTopFeatures(tse\_dada, top = 5, method = "sum",  
 assay.type = "relabundance")  
dada\_table <- data.frame(assays(tse\_dada)$relabundance)  
dada\_table <- dada\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top5\_dada)  
kable(dada\_table, digits=2) %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

| Genus | barcode01 | barcode02 | barcode03 | barcode04 | barcode05 | barcode06 |
| --- | --- | --- | --- | --- | --- | --- |
| Stenotrophomonas | 0 | 0 | 1 | 0 | 0 | 1 |
| Delftia | 0 | 0 | 0 | 0 | 1 | 0 |
| Aeromonas | 0 | 1 | 0 | 0 | 0 | 0 |
| Pseudomonas | 1 | 0 | 0 | 0 | 0 | 0 |
| Providencia | 0 | 0 | 0 | 1 | 0 | 0 |

#get top5 features  
top5\_vs97 <- getTopFeatures(tse\_vs97, top = 5, method = "sum",  
 assay.type = "relabundance")  
vs97\_table <- data.frame(assays(tse\_vs97)$relabundance)  
vs97\_table <- vs97\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top5\_vs97)  
kable(vs97\_table, digits=2) %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

| Genus | barcode01 | barcode02 | barcode03 | barcode04 | barcode05 | barcode06 |
| --- | --- | --- | --- | --- | --- | --- |
| Stenotrophomonas | 0 | 0 | 1 | 0 | 0 | 0.98 |
| Pseudomonas | 1 | 0 | 0 | 0 | 0 | 0.01 |
| Delftia | 0 | 0 | 0 | 0 | 1 | 0.00 |
| Providencia | 0 | 0 | 0 | 1 | 0 | 0.00 |
| Aeromonas | 0 | 1 | 0 | 0 | 0 | 0.00 |

#get top5 features  
top5\_vs99 <- getTopFeatures(tse\_vs99, top = 5, method = "sum",  
 assay.type = "relabundance")  
vs99\_table <- data.frame(assays(tse\_vs99)$relabundance)  
vs99\_table <- vs99\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top5\_vs99)  
kable(vs99\_table, digits=2) %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

| Genus | barcode01 | barcode02 | barcode03 | barcode04 | barcode05 | barcode06 |
| --- | --- | --- | --- | --- | --- | --- |
| Stenotrophomonas | 0.01 | 0.06 | 0.99 | 0.01 | 0.00 | 0.92 |
| Pseudomonas | 0.95 | 0.03 | 0.00 | 0.01 | 0.00 | 0.02 |
| Delftia | 0.01 | 0.02 | 0.00 | 0.00 | 0.98 | 0.01 |
| Providencia | 0.01 | 0.03 | 0.00 | 0.96 | 0.00 | 0.02 |
| Aeromonas | 0.02 | 0.76 | 0.00 | 0.01 | 0.00 | 0.03 |

#get top5 features  
top5\_emu <- getTopFeatures(tse\_emu, top = 5, method = "sum",  
 assay.type = "relabundance")  
emu\_table <- data.frame(assays(tse\_emu)$relabundance)  
emu\_table <- emu\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top5\_emu)  
kable(emu\_table, digits=2) %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

| Genus | barcode01 | barcode02 | barcode03 | barcode04 | barcode05 | barcode06 |
| --- | --- | --- | --- | --- | --- | --- |
| Pseudomonas | 1 | 0 | 0 | 0 | 0 | 0 |
| Aeromonas | 0 | 1 | 0 | 0 | 0 | 0 |
| Stenotrophomonas | 0 | 0 | 1 | 0 | 0 | 1 |
| Delftia | 0 | 0 | 0 | 0 | 1 | 0 |
| Providencia | 0 | 0 | 0 | 1 | 0 | 0 |

For stacked barplots, we create long table, i.e. single column contains all samples

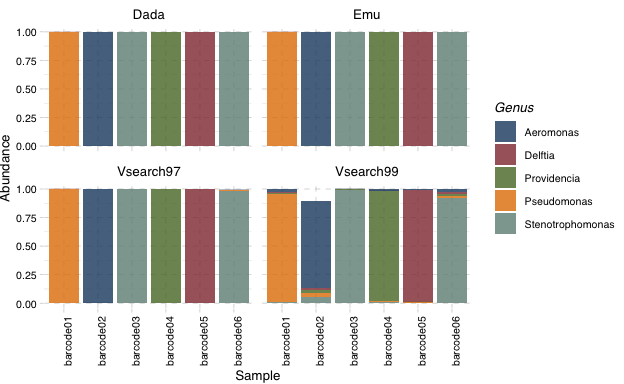
#transform data to ggplot  
dada\_long <- dada\_table %>% pivot\_longer(cols = starts\_with("barcode"),  
 names\_to = "Sample",  
 values\_to = "Abundance") %>%  
 mutate(Method = "Dada")  
#transform data to ggplot  
vs97\_long <- vs97\_table %>% pivot\_longer(cols = starts\_with("barcode"),  
 names\_to = "Sample",  
 values\_to = "Abundance") %>%  
 mutate(Method = "Vsearch97")  
#transform data to ggplot  
vs99\_long <- vs99\_table %>% pivot\_longer(cols = starts\_with("barcode"),  
 names\_to = "Sample",  
 values\_to = "Abundance") %>%  
 mutate(Method = "Vsearch99")  
#transform data to ggplot  
emu\_long <- emu\_table %>% pivot\_longer(cols = starts\_with("barcode"),  
 names\_to = "Sample",  
 values\_to = "Abundance") %>%  
 mutate(Method = "Emu")  
#combine  
long\_table <- bind\_rows(dada\_long, vs97\_long, vs99\_long, emu\_long)

Plot objects

#Create stacked barplot  
ab\_plot <- ggplot(long\_table, aes(x = Sample, y = Abundance, fill = Genus)) +   
 geom\_bar(stat = "identity", alpha=0.8) + facet\_wrap(~Method) +  
 theme\_pander(base\_size = 10) + scale\_fill\_stata() + theme(axis.text.x =  
 element\_text(angle = 90))

Results side by side

#show plots side by side  
ab\_plot



#### Observations

In this dataset, it is likely that we have pure microbial cultures. However, increased noise is observed with vsearch at 99%, particularly in the barcode02 sample. This discrepancy was initially obscured because phyloseq did not correctly import the taxonomy results, leading to *Microbacter* being mistakenly labeled as *Aeromonas*.