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

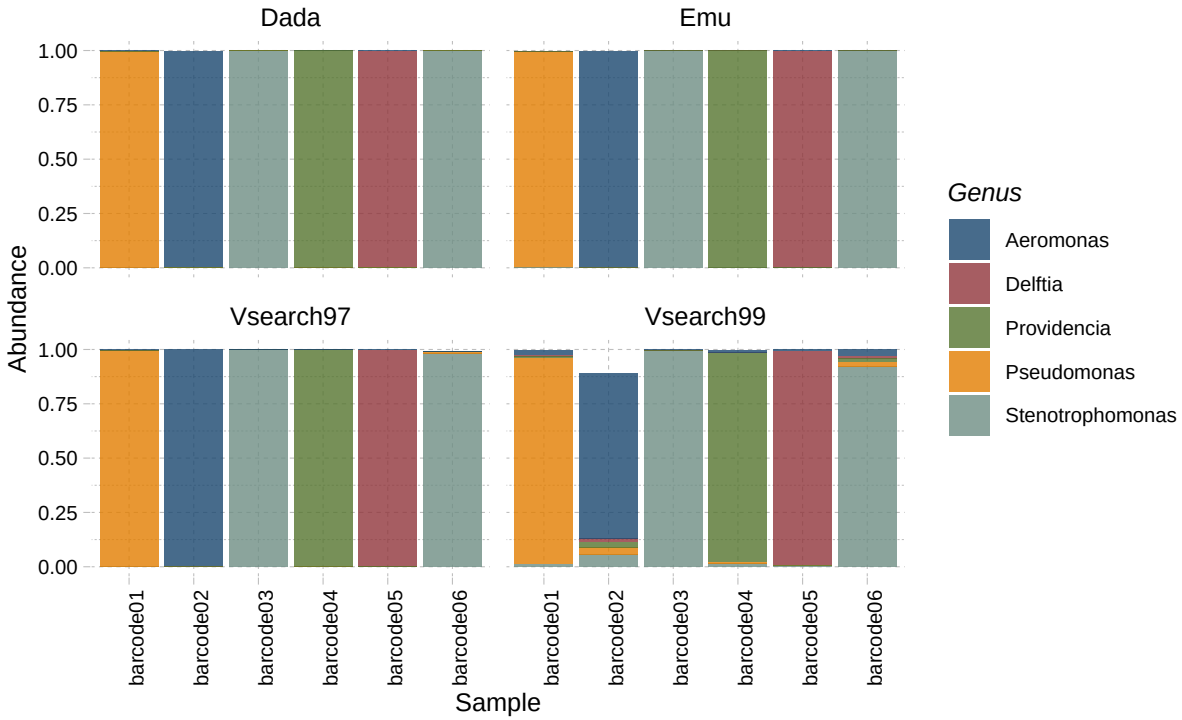
```
#transform data to ggplot
dada_long <- dada_table %>% pivot_longer(cols = starts_with("barcode"),
                                         names_to = "Sample",
                                         values_to = "Abundance") %>%
  mutate(Method = "Dada")
#transform data to ggplot
vs97_long <- vs97_table %>% 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*.