Analysis set2 updated

Marko Suokas

#### Libraries

library(mia);packageVersion("mia")

[1] '1.12.0'

library(vegan);packageVersion("vegan")

[1] '2.6.6.1'

library(scater);packageVersion("scater")

[1] '1.32.1'

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(ggthemes);packageVersion("ggthemes")

[1] '5.1.0'

#### Import

#Rds files are easy way to save ja load data objects  
tse\_dada <- readRDS("set2/tse\_dada.rds")  
tse\_vs97 <- readRDS("set2/tse\_vs97.rds")  
tse\_vs99 <- readRDS("set2/tse\_vs99.rds")  
tse\_emu <- readRDS("set2/tse\_emu.rds")

Agglomeration of objects to genus level.

#Agglomerate  
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] 39

nrow(tse\_vs97)

[1] 51

nrow(tse\_vs99)

[1] 49

nrow(tse\_emu)

[1] 39

Calculate relative abundance for assay data

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")

Getting top10 features

top10\_dada <- getTopFeatures(tse\_dada, top = 10, method = "mean",  
 assay.type = "relabundance")  
#Fetch assay table and filter it  
dada\_table <- data.frame(assays(tse\_dada)$relabundance)  
dada\_table <- dada\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top10\_dada)  
kable(dada\_table, digits=2, caption = "Dada2 denoiser") %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

Dada2 denoiser

| Genus | barcode07 | barcode08 | barcode09 | barcode10 | barcode11 | barcode12 |
| --- | --- | --- | --- | --- | --- | --- |
| Acidocella | 0.09 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 |
| Acidithiobacillus | 0.05 | 0.04 | 0.05 | 0.04 | 0.03 | 0.05 |
| Microbacter | 0.00 | 0.11 | 0.00 | 0.21 | 0.12 | 0.00 |
| Fonticella | 0.09 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 |
| Thiomonas | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.10 |
| Lachnoclostridium | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.14 |
| Pelosinus | 0.16 | 0.18 | 0.28 | 0.00 | 0.00 | 0.12 |
| Desulfosporosinus | 0.32 | 0.47 | 0.15 | 0.21 | 0.37 | 0.44 |
| Mobilitalea | 0.06 | 0.00 | 0.03 | 0.00 | 0.10 | 0.00 |
| Herbinix | 0.17 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 |

top10\_vs97 <- getTopFeatures(tse\_vs97, top = 10, method = "mean",  
 assay.type = "relabundance")  
#Fetch assay table and filter it  
vs97\_table <- data.frame(assays(tse\_vs97)$relabundance)  
vs97\_table <- vs97\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top10\_vs97)  
kable(vs97\_table, digits=2, caption = "Vsearch 97") %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

Vsearch 97

| Genus | barcode07 | barcode08 | barcode09 | barcode10 | barcode11 | barcode12 |
| --- | --- | --- | --- | --- | --- | --- |
| Acidocella | 0.03 | 0.01 | 0.00 | 0.52 | 0.00 | 0.01 |
| Microbacter | 0.00 | 0.13 | 0.00 | 0.25 | 0.14 | 0.00 |
| Acidithiobacillus | 0.06 | 0.05 | 0.06 | 0.04 | 0.03 | 0.05 |
| Fonticella | 0.09 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 |
| Thiomonas | 0.01 | 0.00 | 0.02 | 0.00 | 0.17 | 0.10 |
| Desulfosporosinus | 0.27 | 0.35 | 0.24 | 0.16 | 0.30 | 0.48 |
| Pelosinus | 0.21 | 0.24 | 0.26 | 0.02 | 0.01 | 0.08 |
| Lachnoclostridium | 0.00 | 0.00 | 0.24 | 0.00 | 0.00 | 0.15 |
| Herbinix | 0.23 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Mobilitalea | 0.01 | 0.01 | 0.00 | 0.00 | 0.11 | 0.00 |

top10\_vs99 <- getTopFeatures(tse\_vs99, top = 10, method = "mean",  
 assay.type = "relabundance")  
#Fetch assay table and filter it  
vs99\_table <- data.frame(assays(tse\_vs99)$relabundance)  
vs99\_table <- vs99\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top10\_vs99)  
kable(vs99\_table, digits=2, caption = "Vsearch 99") %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

Vsearch 99

| Genus | barcode07 | barcode08 | barcode09 | barcode10 | barcode11 | barcode12 |
| --- | --- | --- | --- | --- | --- | --- |
| Acidocella | 0.09 | 0.00 | 0.00 | 0.53 | 0.00 | 0.01 |
| Microbacter | 0.02 | 0.11 | 0.01 | 0.17 | 0.13 | 0.01 |
| Acidithiobacillus | 0.06 | 0.05 | 0.06 | 0.04 | 0.03 | 0.06 |
| Fonticella | 0.09 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 |
| Thiomonas | 0.00 | 0.00 | 0.00 | 0.01 | 0.19 | 0.06 |
| Desulfosporosinus | 0.27 | 0.37 | 0.17 | 0.18 | 0.30 | 0.41 |
| Pelosinus | 0.14 | 0.24 | 0.28 | 0.04 | 0.01 | 0.14 |
| Lachnoclostridium | 0.01 | 0.01 | 0.23 | 0.01 | 0.01 | 0.15 |
| Herbinix | 0.20 | 0.02 | 0.01 | 0.00 | 0.00 | 0.01 |
| Mobilitalea | 0.03 | 0.01 | 0.03 | 0.00 | 0.10 | 0.03 |

top10\_emu <- getTopFeatures(tse\_emu, top = 10, method = "mean",  
 assay.type = "relabundance")  
#Fetch assay table and filter it  
emu\_table <- data.frame(assays(tse\_dada)$relabundance)  
emu\_table <- emu\_table %>% rownames\_to\_column(var = "Genus") %>%  
 filter(Genus %in% top10\_emu)  
kable(emu\_table, digits=2, caption = "Emu") %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

Emu

| Genus | barcode07 | barcode08 | barcode09 | barcode10 | barcode11 | barcode12 |
| --- | --- | --- | --- | --- | --- | --- |
| Acidocella | 0.09 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 |
| Acidithiobacillus | 0.05 | 0.04 | 0.05 | 0.04 | 0.03 | 0.05 |
| Microbacter | 0.00 | 0.11 | 0.00 | 0.21 | 0.12 | 0.00 |
| Fonticella | 0.09 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 |
| Thiomonas | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.10 |
| Lachnoclostridium | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.14 |
| Pelosinus | 0.16 | 0.18 | 0.28 | 0.00 | 0.00 | 0.12 |
| Desulfosporosinus | 0.32 | 0.47 | 0.15 | 0.21 | 0.37 | 0.44 |
| Herbinix | 0.17 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 |
| Anaerovorax | 0.02 | 0.04 | 0.00 | 0.00 | 0.03 | 0.01 |

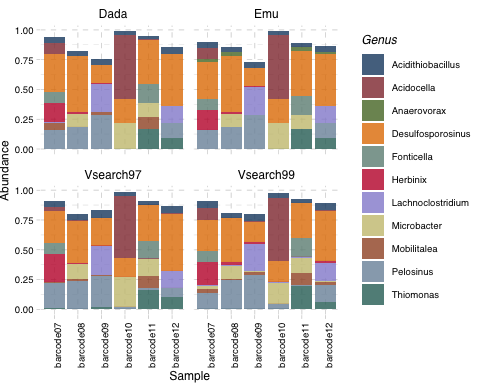
For barplot it is necessary to create long table

#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 object

#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 = 9) + scale\_fill\_stata() +  
 theme(axis.text.x = element\_text(angle = 90))

ab\_plot



Alpha diversity measured by Shannon index

#dada  
tse\_dada <- readRDS("set2/tse\_dada.rds")  
tse\_dada<- transformAssay(tse\_dada, assay.type = "counts", method = "relabundance",  
 name = "relabundance")  
tse\_dada <- estimateDiversity(tse\_dada, assay.type="relabundance", index="shannon")  
#vsearch97  
tse\_vs97 <- readRDS("set2/tse\_vs97.rds")  
tse\_vs97 <- transformAssay(tse\_vs97, assay.type = "counts", method = "relabundance",  
 name = "relabundance")  
tse\_vs97 <- estimateDiversity(tse\_vs97, assay.type="counts", index="shannon")  
#vsearch99  
tse\_vs99 <- readRDS("set2/tse\_vs99.rds")  
tse\_vs99 <- transformAssay(tse\_vs99, assay.type = "counts", method = "relabundance",  
 name = "relabundance")  
tse\_vs99 <- estimateDiversity(tse\_vs99, assay.type="counts",index="shannon")  
#emu  
tse\_emu <- readRDS("set2/tse\_emu.rds")  
tse\_emu <- transformAssay(tse\_emu, assay.type = "counts", method = "relabundance",  
 name = "relabundance")  
tse\_emu <- estimateDiversity(tse\_emu, assay.type="counts",index="shannon")  
  
#combine results  
alpha <- data.frame(dada = colData(tse\_dada)$shannon,  
 vsearch97 = colData(tse\_vs97)$shannon,  
 vsearch99 = colData(tse\_vs99)$shannon,  
 emu = colData(tse\_emu)$shannon)

#table  
kable(alpha, digits=2, caption = "Shannon index") %>%  
 kable\_styling(latex\_options = c("HOLD\_position", "striped"), font\_size = 11) %>%  
 row\_spec(0, background = "teal", color = "white")

Shannon index

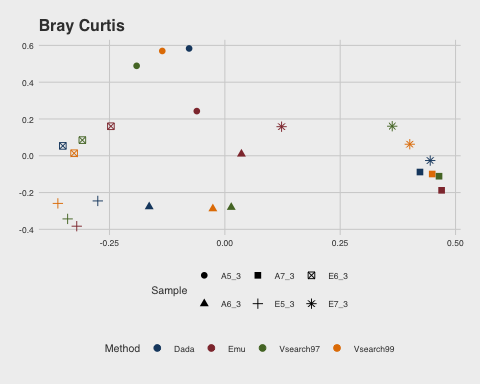
|  | dada | vsearch97 | vsearch99 | emu |
| --- | --- | --- | --- | --- |
| barcode07 | 3.91 | 3.19 | 4.12 | 2.13 |
| barcode08 | 4.31 | 3.31 | 4.04 | 1.98 |
| barcode09 | 4.05 | 2.45 | 3.72 | 2.20 |
| barcode10 | 2.58 | 1.87 | 2.17 | 1.24 |
| barcode11 | 3.76 | 2.70 | 3.00 | 1.83 |
| barcode12 | 4.05 | 2.47 | 4.57 | 2.12 |

Beta diversity measured by Bray-Curtis dissimilarity

#create bray-curtis distance matrix  
tse\_dada <- runMDS(tse\_dada, FUN = vegan::vegdist, method = "bray",  
 name="PCoA\_BC", exprs\_values = "relabundance")  
dada\_bray <- plotReducedDim(tse\_dada, "PCoA\_BC")  
#create dataframe for plot  
bray\_dada\_df <- data.frame(pcoa1 = dada\_bray$data[,1],  
 pcoa2 = dada\_bray$data[,2],  
 Sample = colData(tse\_dada)$Name)  
tse\_vs99 <- runMDS(tse\_vs99, FUN = vegan::vegdist, method = "bray",  
 name="PCoA\_BC", exprs\_values = "relabundance")  
vsearch99\_bray <- plotReducedDim(tse\_vs99, "PCoA\_BC")  
#create dataframe for plot  
bray\_vsearch99\_df <- data.frame(pcoa1 = vsearch99\_bray$data[,1],  
 pcoa2 = vsearch99\_bray$data[,2],  
 Sample = colData(tse\_vs99)$Name)  
tse\_vs97 <- runMDS(tse\_vs97, FUN = vegan::vegdist, method = "bray",  
 name="PCoA\_BC", exprs\_values = "relabundance")  
vsearch97\_bray <- plotReducedDim(tse\_vs97, "PCoA\_BC")  
#create dataframe for plot  
bray\_vsearch97\_df <- data.frame(pcoa1 = vsearch97\_bray$data[,1],  
 pcoa2 = vsearch97\_bray$data[,2],  
 Sample = colData(tse\_vs97)$Name)  
tse\_emu <- runMDS(tse\_emu, FUN = vegan::vegdist, method = "bray",  
 name="PCoA\_BC", exprs\_values = "relabundance")  
emu\_bray <- plotReducedDim(tse\_emu, "PCoA\_BC")  
#create dataframe for plot  
bray\_emu\_df <- data.frame(pcoa1 = emu\_bray$data[,1],  
 pcoa2 = emu\_bray$data[,2],  
 Sample = colData(tse\_emu)$Name)  
#add method for each dataframe and combine  
bray\_dada\_df$Method <- "Dada"  
bray\_vsearch99\_df$Method <- "Vsearch99"  
bray\_vsearch97\_df$Method <- "Vsearch97"  
bray\_emu\_df$Method <- "Emu"  
combined\_pcoa <- bind\_rows(bray\_dada\_df, bray\_vsearch99\_df, bray\_vsearch97\_df, bray\_emu\_df)  
#create plot object  
plot\_pcoa <- ggplot(data = combined\_pcoa, aes(x=pcoa1, y=pcoa2,  
 color = Method, shape = Sample)) +  
 labs(title = "Bray Curtis",   
 x = "pcoa1", y = "pcoa2") + geom\_point(size=2) +  
 theme\_fivethirtyeight(base\_size=8) + scale\_color\_stata()

Plot pcoa

plot\_pcoa



#### Observations

The overall results are generally consistent across methods, though it is important to note that the diversity in the analyzed samples is lower than what is typically found in most biological samples.

Clustering methods generate a higher number of variants compared to denoising, but the most abundant microbes remain consistent across methods. Interestingly, denoising yields a slightly higher Shannon index in 4 out of 6 samples. In the Bray-Curtis dissimilarity analysis, only one out of six samples shows a notable difference when denoising is applied.

The Emu profiler produced the smallest number of taxa, with only 92 identified. This lower diversity is also reflected in the Shannon index. In the Bray-Curtis plot, Emu displays noticeable differences compared to other methods, although the extent of this variation depends on the sample.

Given the current data, it may be prudent to use vsearch clustering for long amplicons until more comprehensive data is available on the performance of other methods in more complex datasets. A 99% identity level is recommended, along with filtering the least abundant variants to manage data size effectively. In this data set, there are 2800 “OTUs” with more than 10 counts globally.