Regression for Microbiome Data: Multinomial Mixture Models

Brendan J. Kelly, MD, MS

Updated: 23 June 2020

Dirichlet Multinomial Mixtures

Implementating DMM in R

ICU Community Types

DMM & Regression

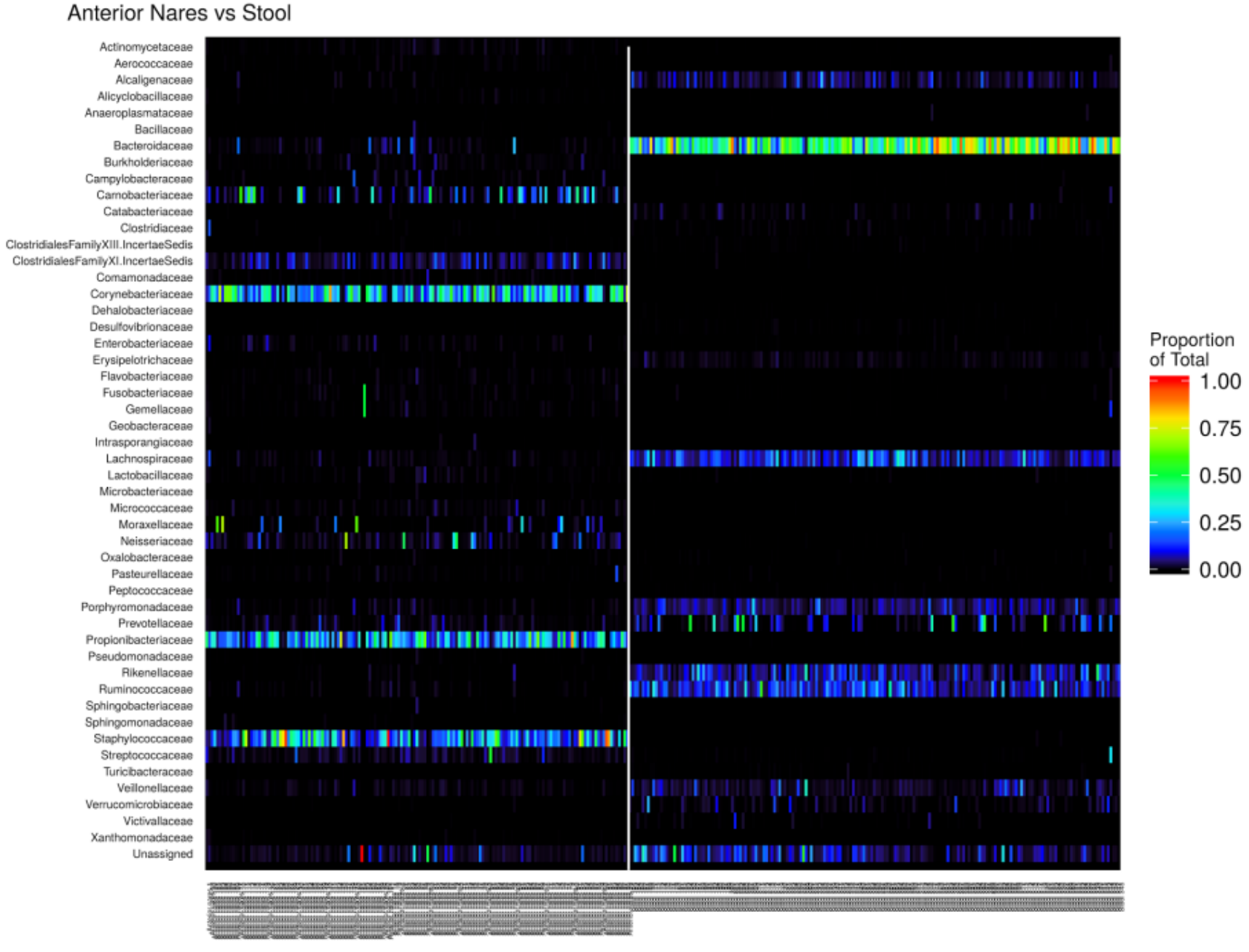
Dirichlet Multinomial Mixtures

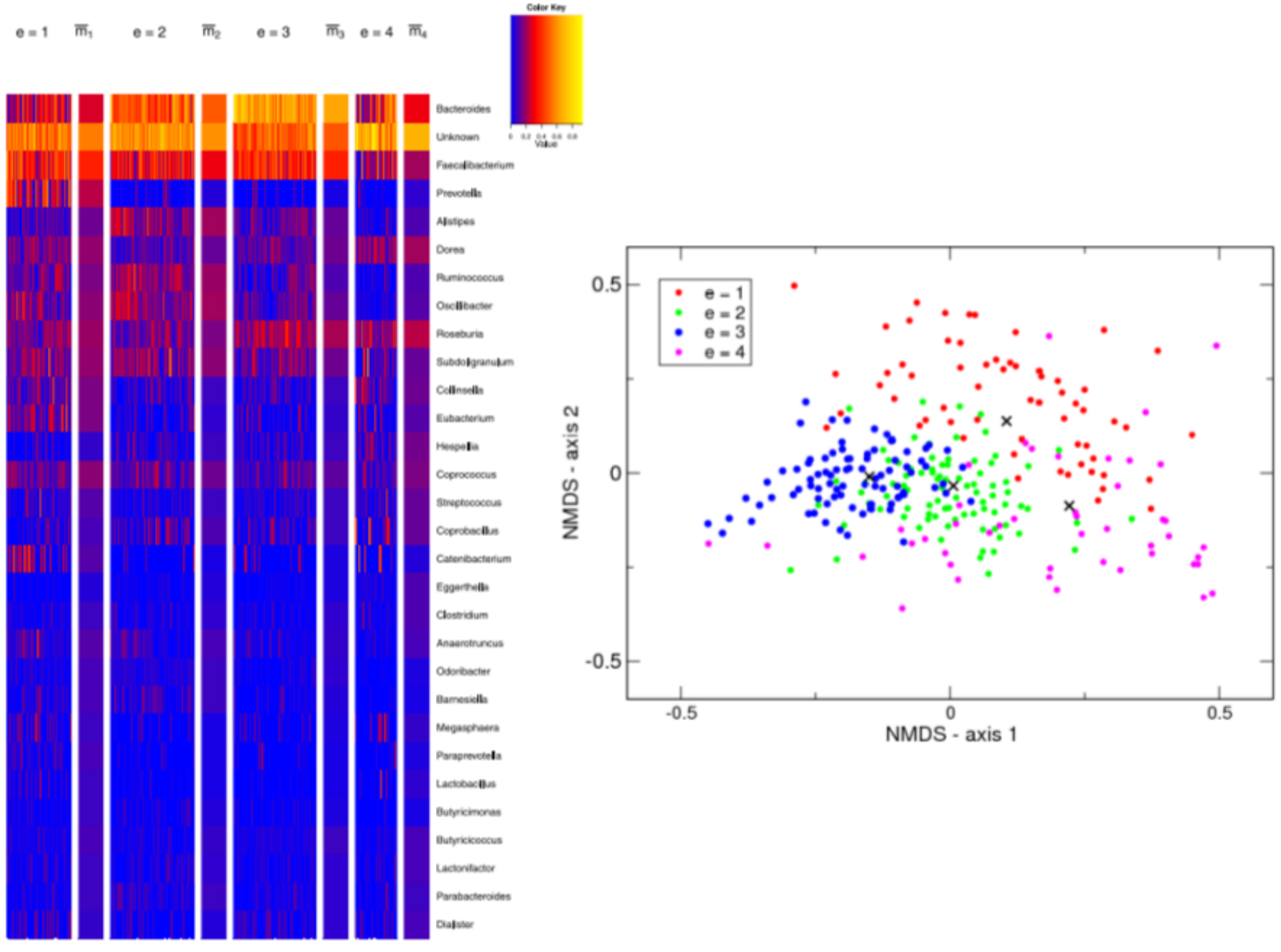
High Dimensional Microbiome Data

```
##              700013549 700014386 700014403 700014409 700014412 700014415
## OTU_97.1             0         0         0         0         0         0
## OTU_97.10            0         0         6         4         1         5
## OTU_97.100           0         0       133         7         1         4
## OTU_97.1000          0         0         0         0         0         0
## OTU_97.10000         0         0         0         0         0         0
## OTU_97.10001         0         0         0         0         0         1
## OTU_97.10002         0         0         0         0         0         0
## OTU_97.10003         0         0         0         0         0         0
## OTU_97.10004         0         0         0         0         0         0
## OTU_97.10005         0         0         0         0         0         0
## OTU_97.10006         0         0         0         0         0         0
## OTU_97.10007         0         0         0         0         0         0
## OTU_97.10008         0         1         0         0         0         0
## OTU_97.10009         0         0         1         0         0         0
## OTU_97.1001          0         0         0         0         0         0
## OTU_97.10010         0         0         0         0         0         0
```

High Dimensional Microbiome Data

- How to deal with high-dimensional microbiome data?
- Descriptive (e.g., heatmaps and stacked barplots)
- Test a priori hypotheses regarding specific OTUs/taxa
- **Reduce dimensions**:
  - single summary statistic (alpha diversity)
  - pairwise distances (beta diversity) with PCoA or PERMANOVA
  - **community types (mixture modeling)**

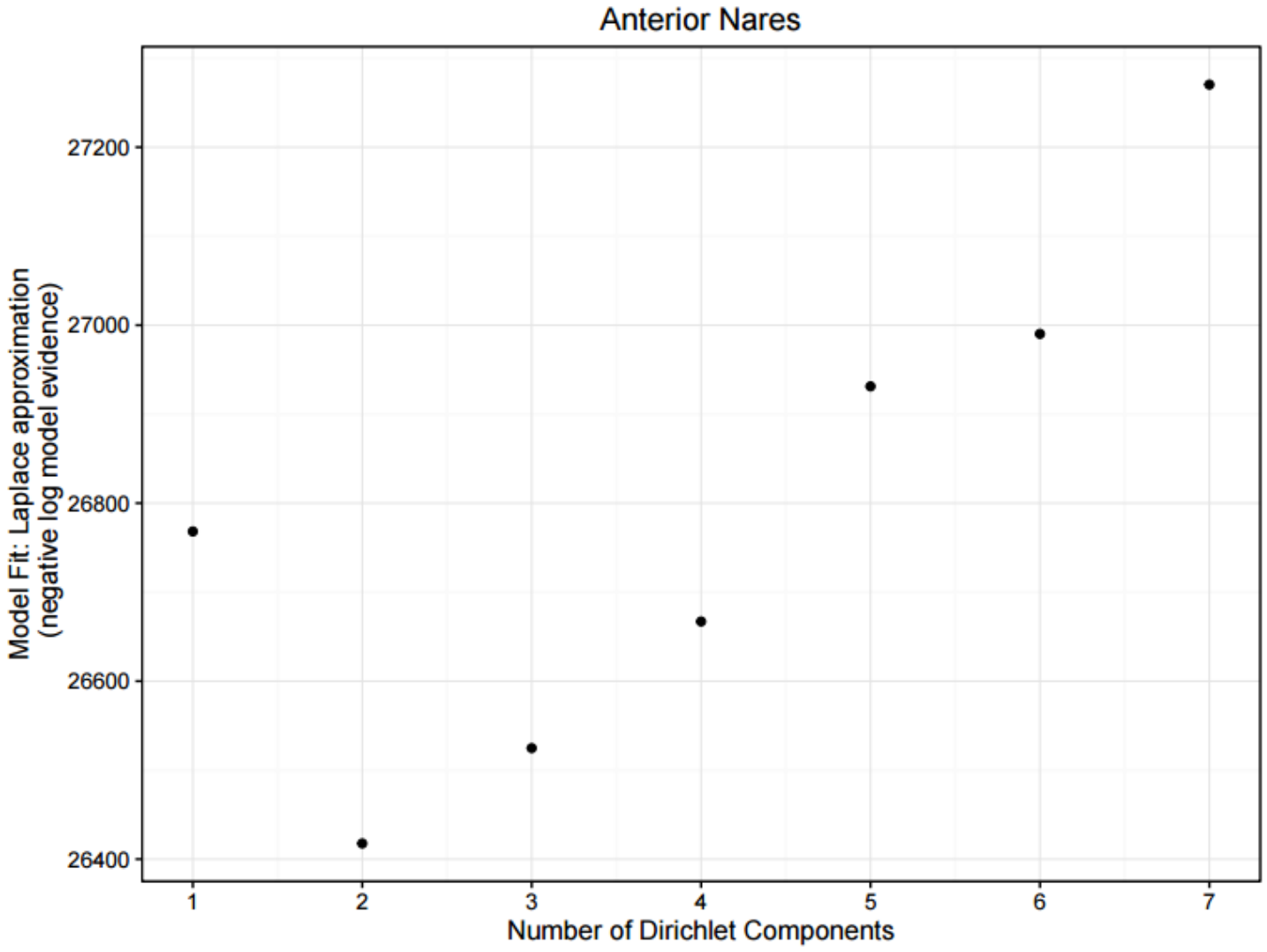

Anterior Nares vs Stool
Actinomycetaceae
Aerococcaceae
Alcaligenaceae
Alicyclobacillaceae
Anaeroplasmataceae
Bacillaceae
Bacteroidaceae
Burkholderiaceae
Campylobacteraceae
Carnobacteriaceae
Catabacteriaceae
Clostridiaceae
ClostridialesFamilyXIII.IncertaeSedis
ClostridialesFamilyXI.IncertaeSedis
Comamonadaceae
Corynebacteriaceae
Dehalobacteriaceae
Desulfovibrionaceae
Enterobacteriaceae
Erysipelotrichaceae
Flavobacteriaceae
Fusobacteriaceae
Gemellaceae
Geobacteraceae
Intrasporangiaceae
Lachnospiraceae
Lactobacillaceae
Microbacteriaceae
Micrococcaceae
Moraxellaceae
Neisseriaceae
Oxalobacteraceae
Pasteurellaceae
Peptococcaceae
Porphyromonadaceae
Prevotellaceae
Propionibacteriaceae
Pseudomonadaceae
Rikenellaceae
Ruminococcaceae
Sphingobacteriaceae
Sphingomonadaceae
Staphylococcaceae
Streptococcaceae
Turicibacteraceae
Veillonellaceae
Verrucomicrobiaceae
Victivallaceae
Xanthomonadaceae
Unassigned
Proportion of Total
1.00
0.75
0.50
0.25
0.00



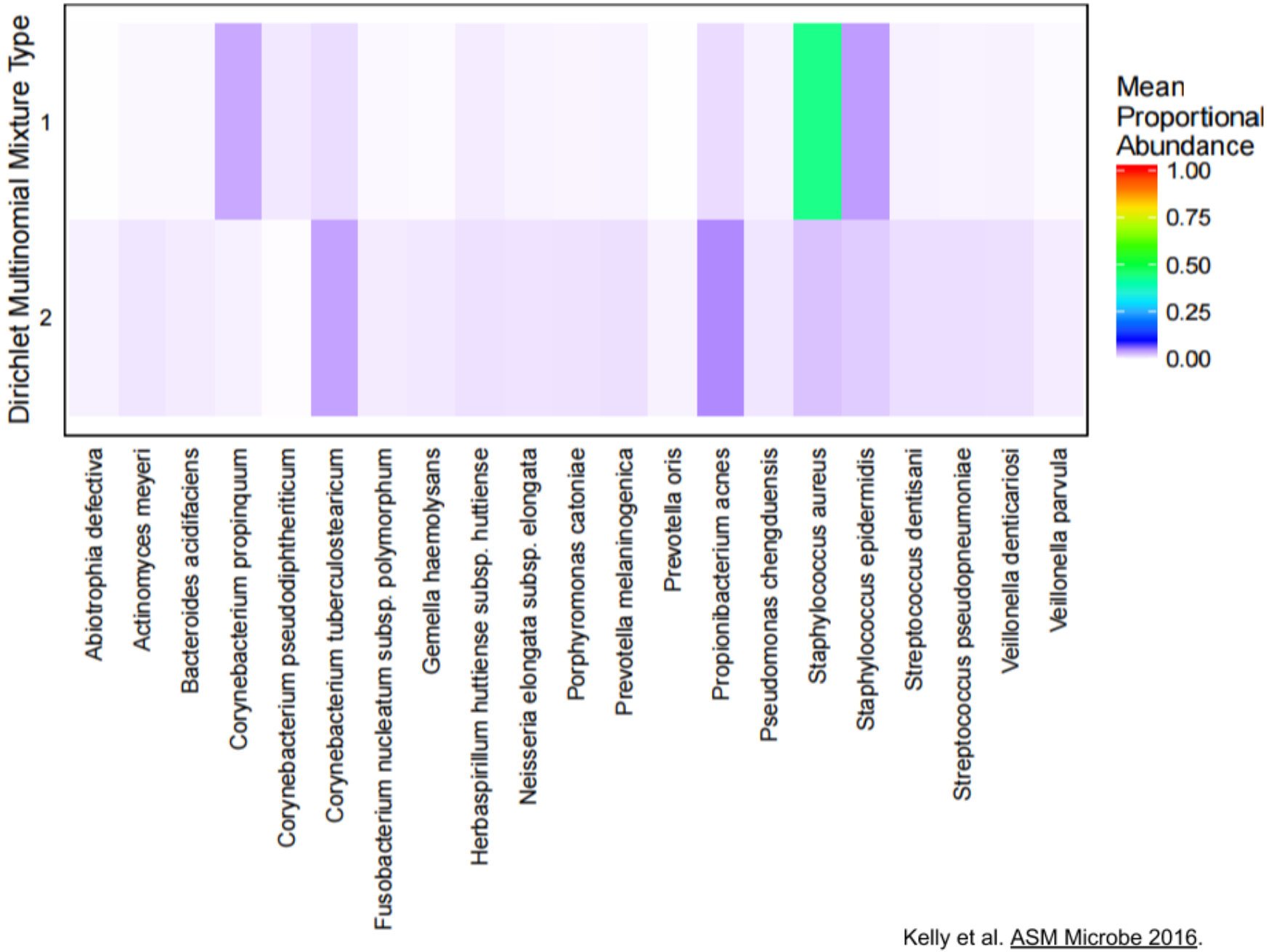
Holmes I et al. PLoS One 2012.

Dirichlet-Multinomial Mixtures

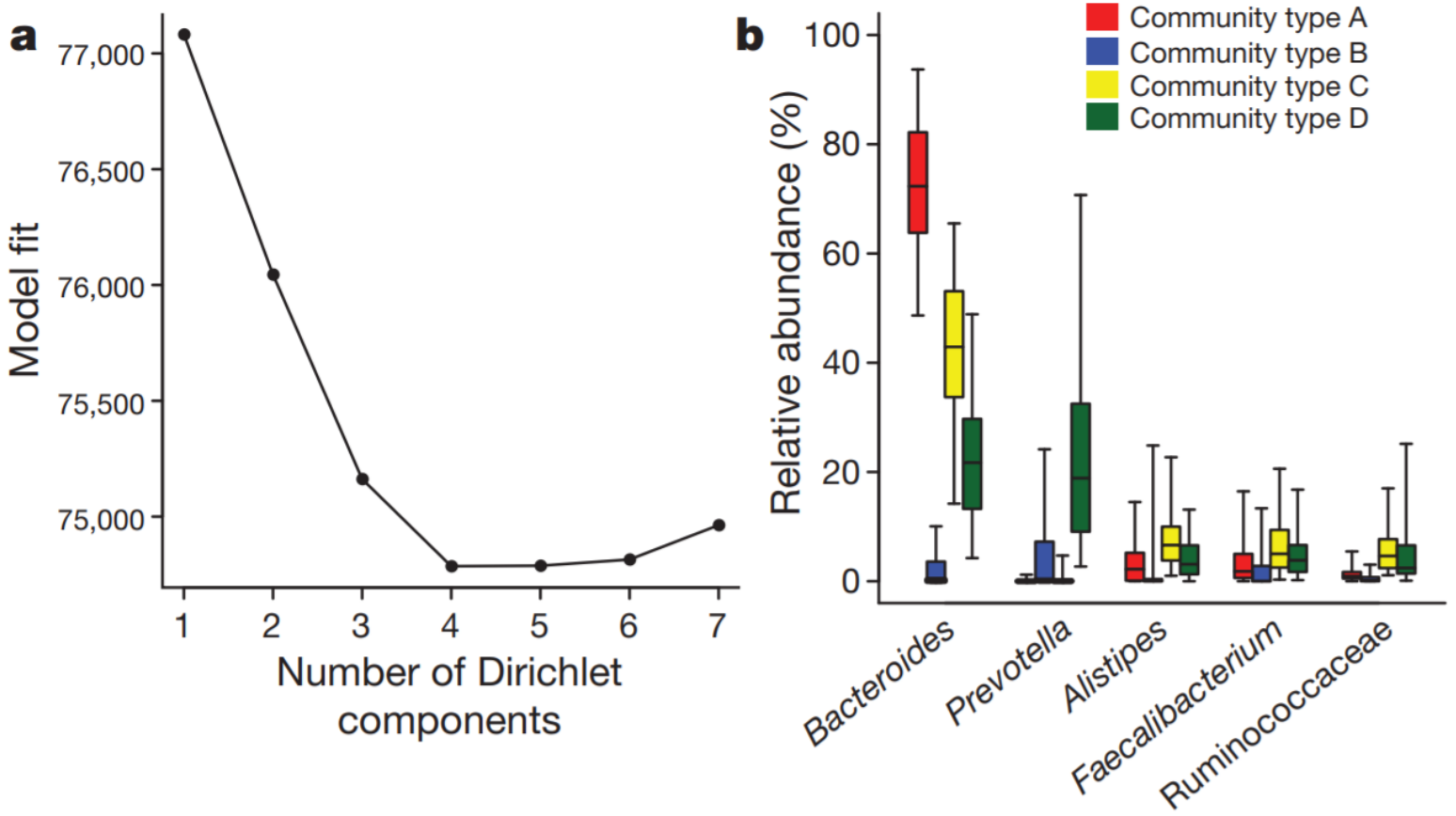
- Dirichlet-multinomial distribution:
  - compound probability distribution
  - probability vector drawn from Dirichlet distribution (generalized beta)
  - observation drawn from multinomial distribution (generalized binomial)
- D-M mixture modelling:
  - each sample ~ multinomial from one Dirichlet vector
  - vector number: minimize -log(model evidence, Laplace approx)
  - Dirichlet probability vectors = “community types”



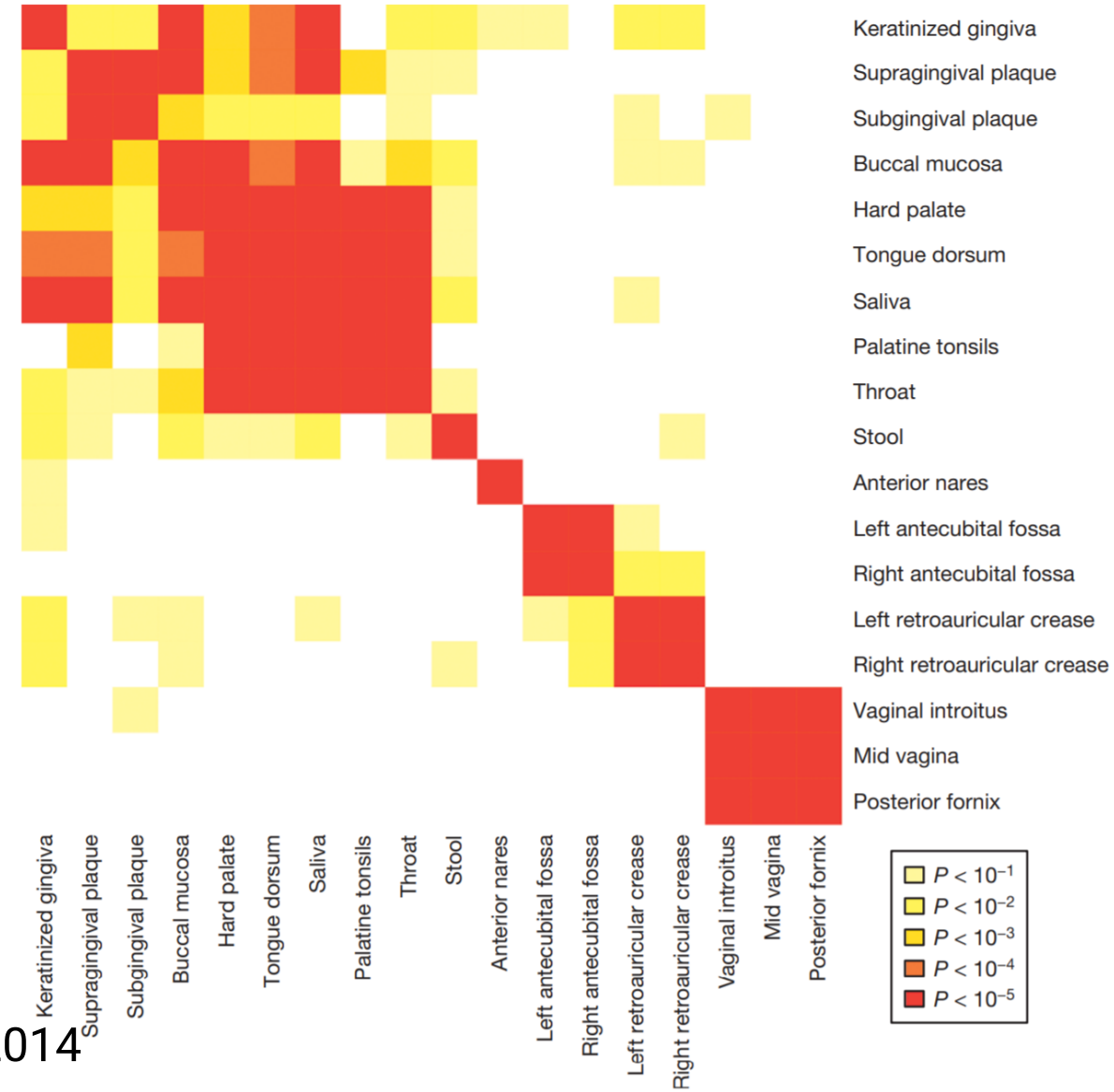
Kelly et al. ASM Microbe 2016.



Kelly et al. ASM Microbe 2016.



Ding & Schloss *Nature* 2014



Ding & Schloss *Nature* 2014

Implementating DMM in R

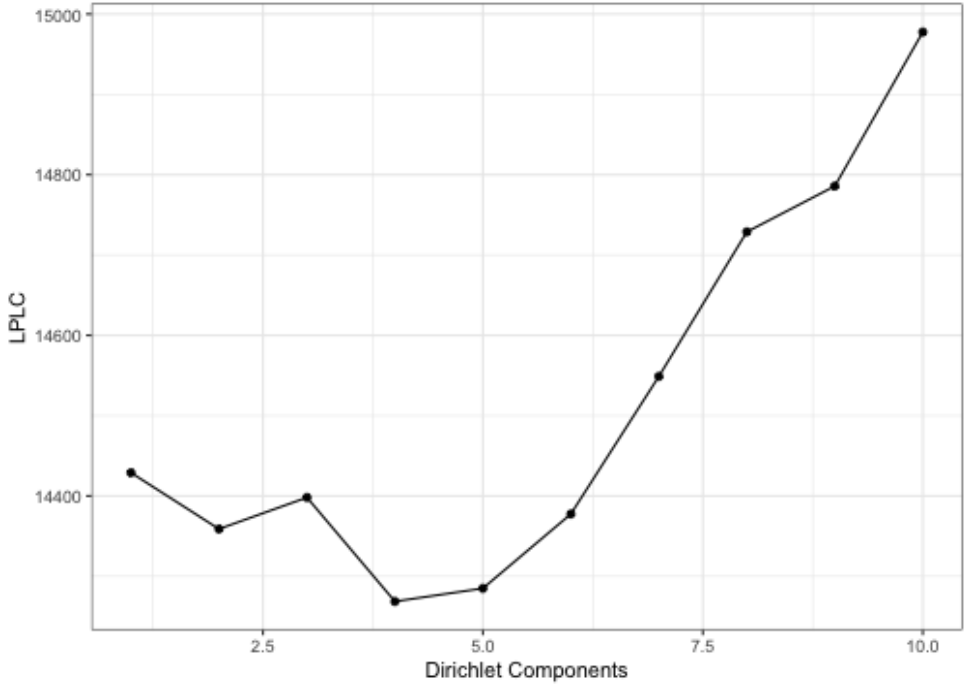
Preparation for DMM

```
# install tidyverse ...
# install.packages("tidyverse")
library(tidyverse)

# new package for heatmap color schemes...
# install.packages("viridis")
library(viridis)

# install package from Bioconductor...
# install.packages("BiocManager")
# BiocManager::install("DirichletMultinomial")
library(DirichletMultinomial)

set.seed(16) # for consistent DMM results

icu_matrix_et <- read_rds(
  path = "./data/icu_ET_specimen_otu_table.rds"
  )

icu_matrix_et[1:16,1:2]
```

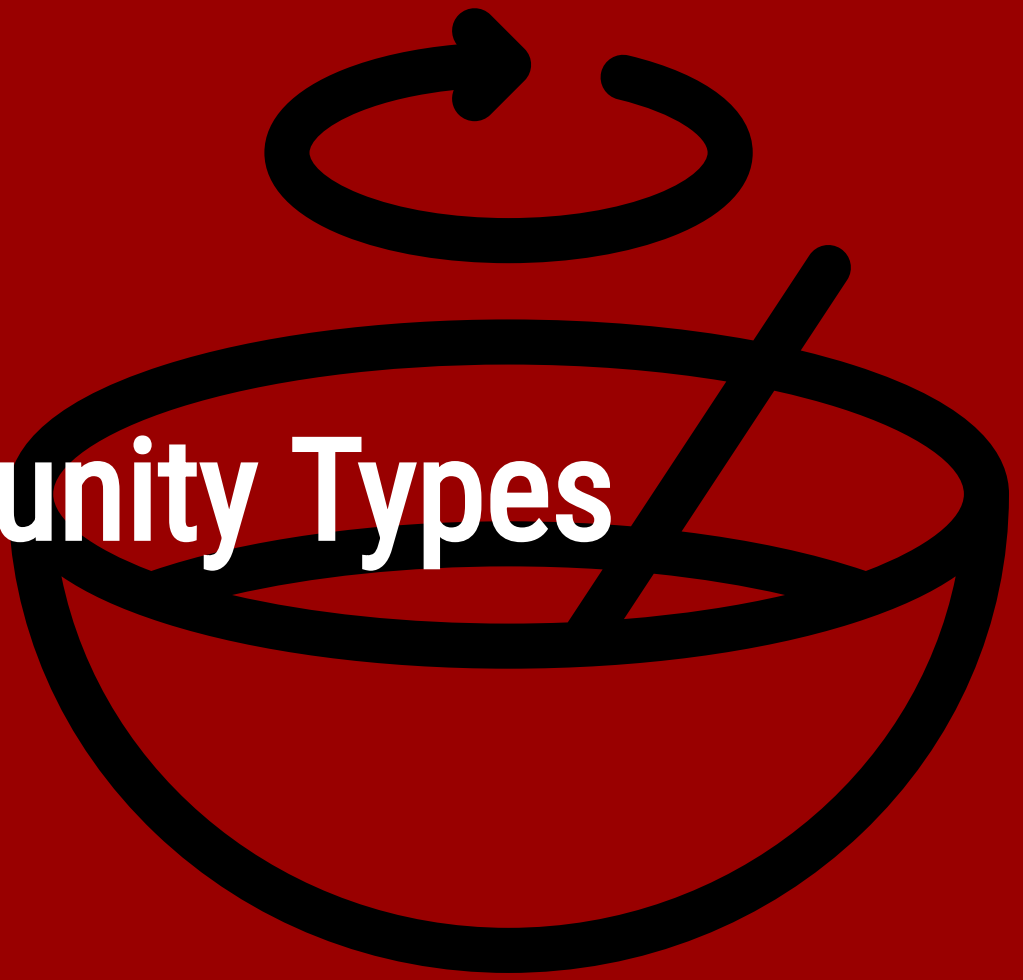
```
##             VAP.001.ET.20130726 VAP.001.ET.20130729
## denovo1                       0                   0
## denovo10004                   0                   0
## denovo10011                   0                   0
## denovo10015                   0                   0
## denovo10018                   0                   0
## denovo10022                   0                   0
## denovo10039                   0                   0
## denovo1004                    0                   0
## denovo10042                   0                   0
## denovo10049                   0                   0
## denovo10065                   0                   0
## denovo10078                   3                   2
## denovo10080                   0                   0
## denovo10089                   0                   3
## denovo10091                   0                   0
## denovo10092                   2                   0
```

DirichletMultinomial

```
#filter to speed DMM model fitting
icu_matrix_et[rowSums(icu_matrix_et) > 500,] ->
  small_icu_matrix_et

dmm <- lapply(1:10,
              dmn,
              count = t(small_icu_matrix_et),
              verbose = FALSE)

lplc <- sapply(dmm, laplace)

qplot(x = seq_along(lplc),
      y = lplc,
      geom = c("point","line")) +
  theme_bw() +
  labs(x = "Dirichlet Components",
       y = "LPLC")
```



ICU Community Types



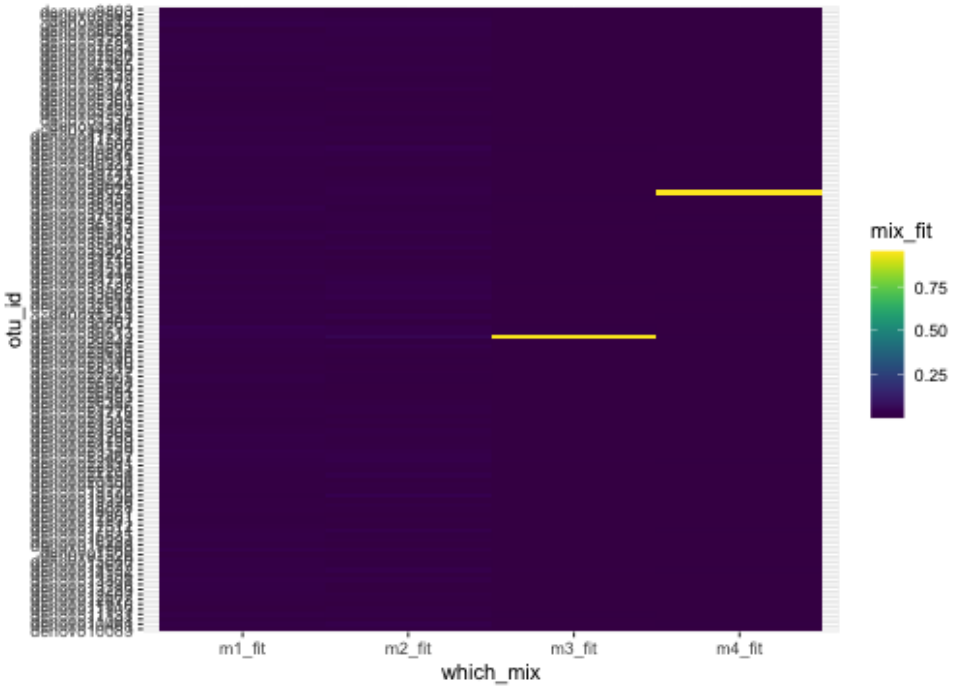
DMM Assignments

```
best_dmm <- dmm[[which.min(lplc)]]

mixture(best_dmm) %>%
  as_tibble(rownames = "specimen") %>%
  rename_at(.vars = vars(contains("V")),
            .funs = function(x)
              paste0(
                gsub("V","m",x),"_prob")
            ) %>%
  mutate(assignment =
           mixture(best_dmm,
                   assign = TRUE)) ->
  icu_et_dmm_assignments

icu_et_dmm_assignments
```

```
## # A tibble: 42 x 6
##    specimen             m1_prob  m2_prob   m3_prob m4_prob assignme
##    <chr>                  <dbl>    <dbl>     <dbl>   <dbl>      <in
##  1 VAP.001.ET.20130726 3.44e-13 1.00e+ 0 0.              0
##  2 VAP.001.ET.20130729 2.68e-24 1.00e+ 0 0.              0
##  3 VAP.001.ET.20130731 1.37e-15 1.00e+ 0 0.              0
##  4 VAP.002.ET.20130729 1.00e+ 0 2.30e-31 0.              0
##  5 VAP.002.ET.20130731 1.00e+ 0 4.43e-36 0.              0
##  6 VAP.002.ET.20130802 1.00e+ 0 3.42e-31 0.              0
##  7 VAP.002.ET.20130805 3.09e-19 1.00e+ 0 0.              0
##  8 VAP.003.ET.20130730 1.52e-22 1.00e+ 0 5.23e-310       0
##  9 VAP.004.ET.20130808 1.00e+ 0 1.65e-69 0.              0
## 10 VAP.005.ET.20130814 9.92e-28 4.01e-12 1.00e+  0       0
## # … with 32 more rows
```

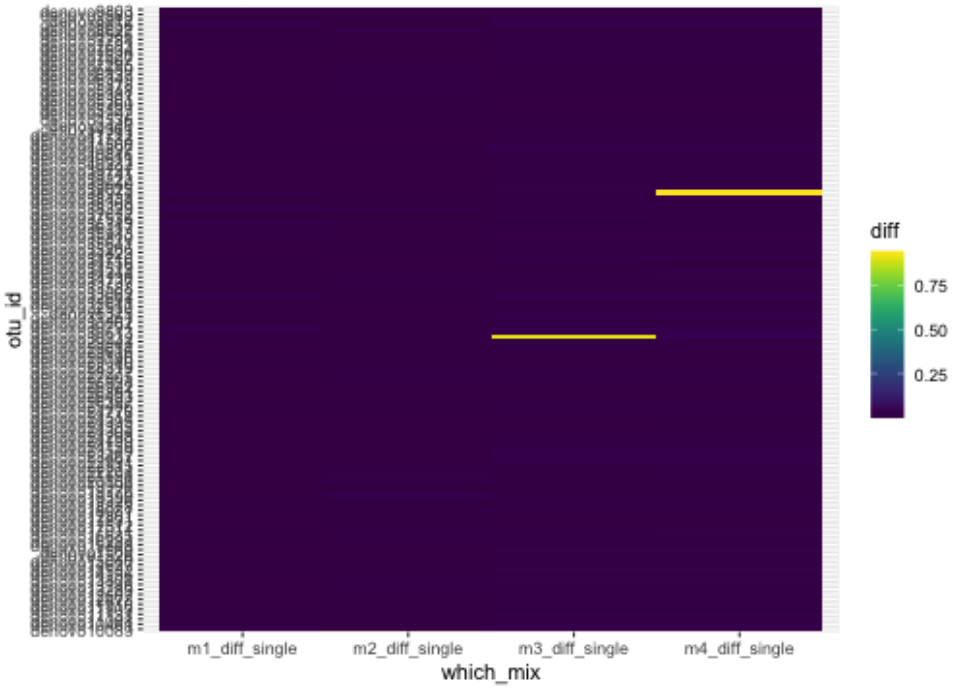
DMM Mixture Fits

```
fitted(best_dmm, scale=TRUE) %>%
# scale indicates whether fits scaled by the...
# ... variability of mixturewt parameter theta
  as_tibble(rownames = "otu_id") %>%
  rename_at(.vars = vars(contains("V")),
            .funs = function(x)
              paste0(gsub("V","m",x),"_fit")) ->
  icu_et_dmm_otu_fits

icu_et_dmm_otu_fits %>%
  gather(key = which_mix, value = mix_fit, -otu_id) %>%
  ggplot(data = .) +
  geom_tile(mapping = aes(x = which_mix,
                          y = otu_id,
                          fill = mix_fit)) +
  scale_fill_viridis()
```



Difference From Single-Mixture

```
abs(fitted(best_dmm, scale=TRUE) -
      as.vector(fitted(dmm[[1]],
                       scale=TRUE))) %>%
# scale indicates whether fits scaled by the...
# ... variability of mixturewt parameter theta
  as_tibble(rownames = "otu_id") %>%
  rename_at(.vars = vars(contains("V")),
            .funs = function(x)
              paste0(gsub("V","m",x),"_diff_single")) ->
  icu_et_dmm_otu_diff_single

icu_et_dmm_otu_diff_single %>%
  gather(key = which_mix,
         value = diff,
         -otu_id) %>%
  ggplot(data = .) +
  geom_tile(mapping = aes(x = which_mix,
                          y = otu_id,
                          fill = diff)) +
  scale_fill_viridis()
```



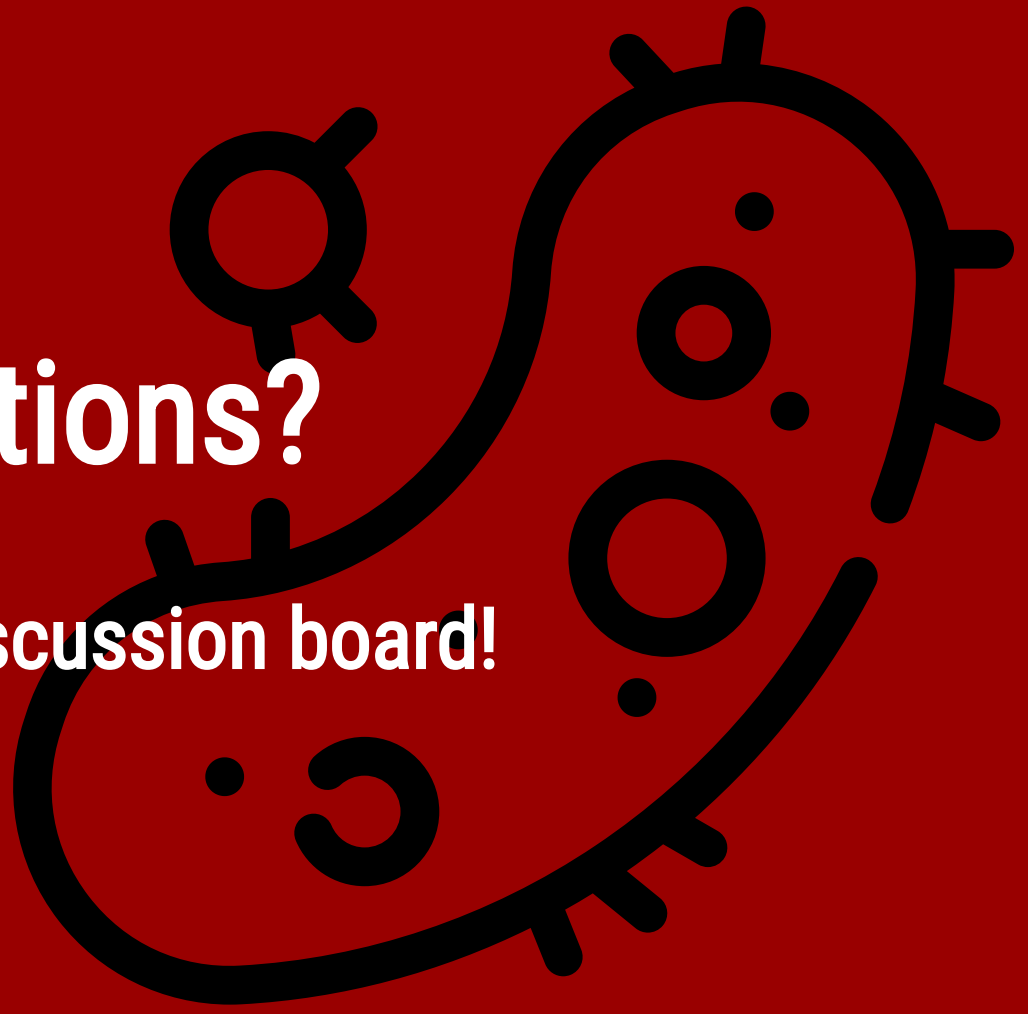
DMM & Regression

DMM & Regression?

- DMM community types as exposure variable:
  - easy → `lm()` or `glm()`
  - (like α-diversity or β-diversity PC1)
- DMM community types as outcome variables:
  - e.g., categorical logistic regression
- Biological validity of DMM community types? Reproducibility?

Questions?

Post to the discussion board!



Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu