Alpha Diversity: Summarizing Microbial Communities

Brendan J. Kelly, MD, MS

Updated: 09 June 2020

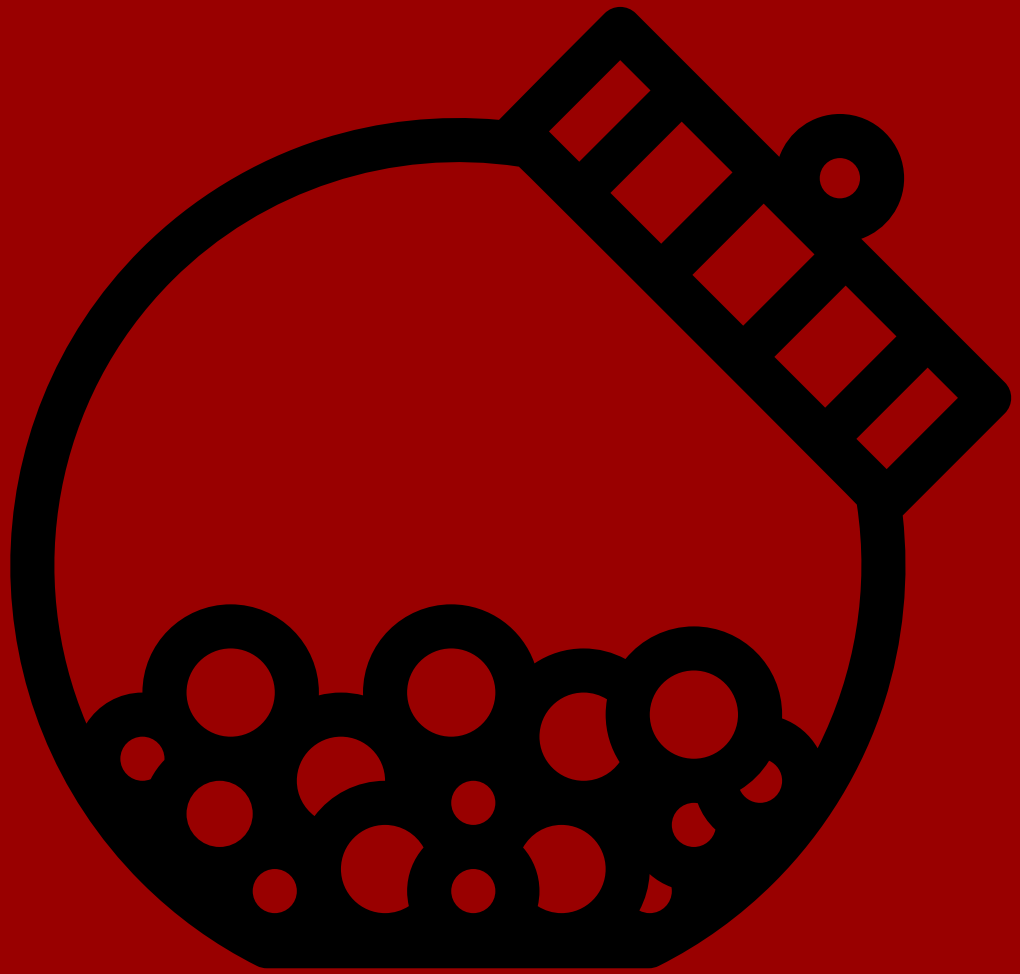
R/tidyverse recap

Summarizing OTU tables

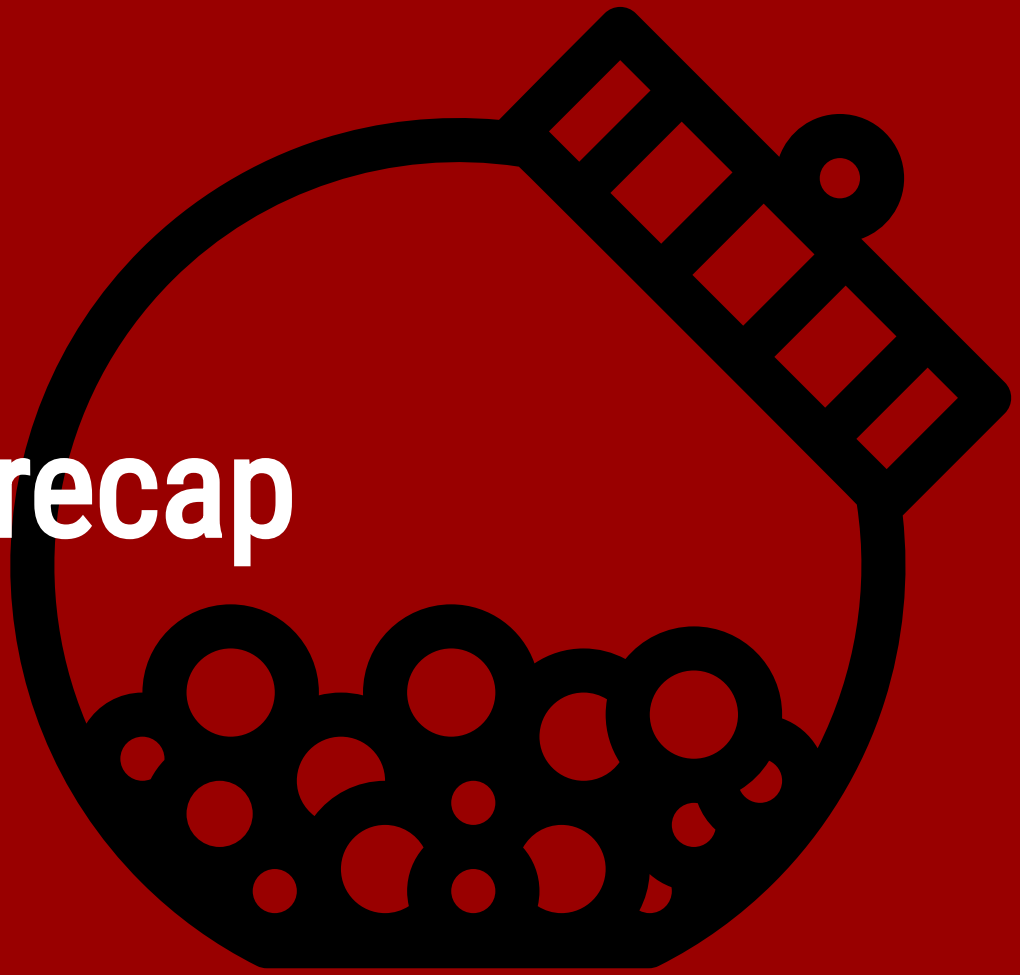
Alpha diversity

Rarefaction vs rarefying

R's vegan package



R/tidyverse recap



R: Making New Variables

- `tidyverse` to create new variables
- Make new variables & keep old variables:
  - `mutate()` function
  - must return same length as input, or length 1 (wll be recycled)
- Make a new variable & lose old variables:
  - `summarise()` function (often follows `group_by()`)
  - simplest use case: return length of 1

R: Making New Variables

```
# install tidyverse once
install.packages('tidyverse')

# load tidyverse functions each time you use
library(tidyverse)
```

- Code at left installs and loads the `tidyverse` package
- The `tidyverse` package includes a set of other packages that permit streamlined data processing
- See Hadley Wickham's *R For Data Science*: https://r4ds.had.co.nz/

mutate()

```
# make sure tidyverse loaded
library(tidyverse)

# load (trimmed) HMP V1-V3 OTU table
otu <- read_csv(
  file =
    "./data/HMP_OTU_table_longformat.csv.gz",
)

otu # show what you've read
```

```
## # A tibble: 1,380,480 x 4
##    HMPbodysubsite specimen_id otu_id       read_count
##    <chr>                <dbl> <chr>             <dbl>
##  1 Anterior_nares   700014445 OTU_97.1              0
##  2 Anterior_nares   700014445 OTU_97.10             0
##  3 Anterior_nares   700014445 OTU_97.100            0
##  4 Anterior_nares   700014445 OTU_97.1000           0
##  5 Anterior_nares   700014445 OTU_97.10000          0
##  6 Anterior_nares   700014445 OTU_97.10001          0
##  7 Anterior_nares   700014445 OTU_97.10002          0
##  8 Anterior_nares   700014445 OTU_97.10003          0
##  9 Anterior_nares   700014445 OTU_97.10004          0
## 10 Anterior_nares   700014445 OTU_97.10005          0
## # … with 1,380,470 more rows
```

mutate()

```
otu %>%
  mutate(
    HMPbodysubsite =
           gsub("_"," ",HMPbodysubsite)
    )
```

```
## # A tibble: 1,380,480 x 4
##    HMPbodysubsite specimen_id otu_id       read_count
##    <chr>                <dbl> <chr>             <dbl>
##  1 Anterior nares   700014445 OTU_97.1              0
##  2 Anterior nares   700014445 OTU_97.10             0
##  3 Anterior nares   700014445 OTU_97.100            0
##  4 Anterior nares   700014445 OTU_97.1000           0
##  5 Anterior nares   700014445 OTU_97.10000          0
##  6 Anterior nares   700014445 OTU_97.10001          0
##  7 Anterior nares   700014445 OTU_97.10002          0
##  8 Anterior nares   700014445 OTU_97.10003          0
##  9 Anterior nares   700014445 OTU_97.10004          0
## 10 Anterior nares   700014445 OTU_97.10005          0
## # … with 1,380,470 more rows
```

mutate()

```
otu %>%
  mutate(
    HMPbodysubsite =
           gsub("_"," ",HMPbodysubsite),
    log_reads = log10(read_count)
    )
```

```
## # A tibble: 1,380,480 x 5
##    HMPbodysubsite specimen_id otu_id       read_count log_reads
##    <chr>                <dbl> <chr>             <dbl>     <dbl>
##  1 Anterior nares   700014445 OTU_97.1              0      -Inf
##  2 Anterior nares   700014445 OTU_97.10             0      -Inf
##  3 Anterior nares   700014445 OTU_97.100            0      -Inf
##  4 Anterior nares   700014445 OTU_97.1000           0      -Inf
##  5 Anterior nares   700014445 OTU_97.10000          0      -Inf
##  6 Anterior nares   700014445 OTU_97.10001          0      -Inf
##  7 Anterior nares   700014445 OTU_97.10002          0      -Inf
##  8 Anterior nares   700014445 OTU_97.10003          0      -Inf
##  9 Anterior nares   700014445 OTU_97.10004          0      -Inf
## 10 Anterior nares   700014445 OTU_97.10005          0      -Inf
## # … with 1,380,470 more rows
```

summarise()

```
otu %>%
  group_by(HMPbodysubsite) %>%
  summarise(
    mean_reads = mean(read_count, na.rm = TRUE)) %>%
  ungroup()
```

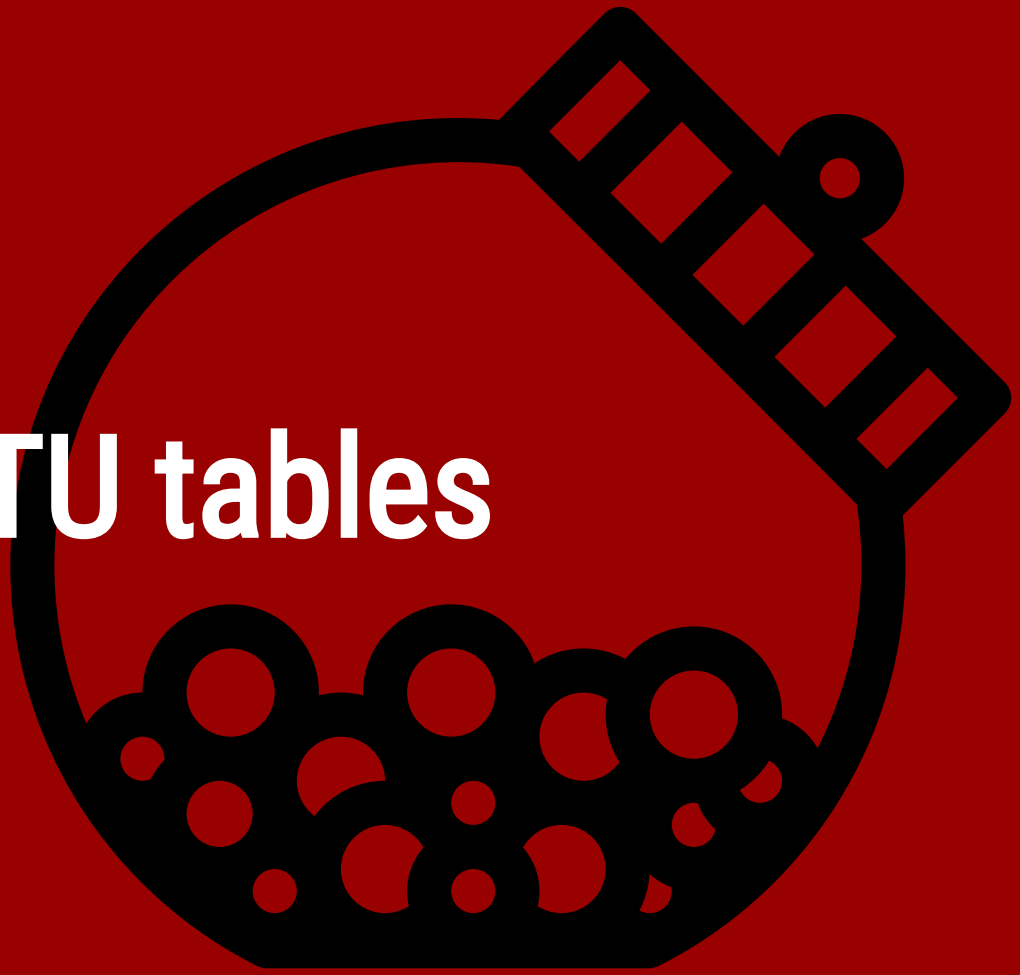
```
## # A tibble: 16 x 2
##    HMPbodysubsite                mean_reads
##    <chr>                              <dbl>
##  1 Anterior_nares                    0.0958
##  2 Attached_Keratinized_gingiva      0.121
##  3 Buccal_mucosa                     0.172
##  4 Hard_palate                       0.116
##  5 Left_Retroauricular_crease        0.0403
##  6 Mid_vagina                        0.267
##  7 Palatine_Tonsils                  0.133
##  8 Posterior_fornix                  0.257
##  9 Right_Retroauricular_crease       0.0835
## 10 Saliva                            0.213
## 11 Stool                             0.221
## 12 Subgingival_plaque                0.262
## 13 Supragingival_plaque              0.260
## 14 Throat                            0.117
## 15 Tongue_dorsum                     0.202
## 16 Vaginal_introitus                 0.247
```

summarise()

```
otu %>%
  group_by(HMPbodysubsite) %>%
  summarise(
    mean_reads = mean(read_count, na.rm = TRUE),
    sum_reads = sum(read_count, na.rm = TRUE)
    ) %>%
  ungroup()
```

```
## # A tibble: 16 x 3
##    HMPbodysubsite                mean_reads sum_reads
##    <chr>                              <dbl>     <dbl>
##  1 Anterior_nares                    0.0958      8262
##  2 Attached_Keratinized_gingiva      0.121      10458
##  3 Buccal_mucosa                     0.172      14831
##  4 Hard_palate                       0.116       9977
##  5 Left_Retroauricular_crease        0.0403      3481
##  6 Mid_vagina                        0.267      23043
##  7 Palatine_Tonsils                  0.133      11445
##  8 Posterior_fornix                  0.257      22189
##  9 Right_Retroauricular_crease       0.0835      7207
## 10 Saliva                            0.213      18336
## 11 Stool                             0.221      19072
## 12 Subgingival_plaque                0.262      22573
## 13 Supragingival_plaque              0.260      22404
## 14 Throat                            0.117      10061
## 15 Tongue_dorsum                     0.202      17431
## 16 Vaginal_introitus                 0.247      21354
```

Summarizing OTU tables



High Dimensional Microbiome Data

- Typical OTU table orientation in microbiome studies:
  - 43140 rows
  - 32 columns

```
# TYPICAL OTU TABLE ORIENTATION IN MICROBIOME STUDIES

otu %>%
  reshape2::acast(otu_id ~ specimen_id,
                  # rows = otu_id, columns = specimen_id
                  value.var = "read_count") %>%
     .[1:10,1:5]

# 43140 ROWS & 32 COLUMNS
```

```
##              700013549 700014386 700014403 700014409 700014412
## OTU_97.1             0         0         0         0         0
## OTU_97.10            0         0         6         4         1
## OTU_97.100           0         0       133         7         1
## OTU_97.1000          0         0         0         0         0
## OTU_97.10000         0         0         0         0         0
## OTU_97.10001         0         0         0         0         0
## OTU_97.10002         0         0         0         0         0
## OTU_97.10003         0         0         0         0         0
## OTU_97.10004         0         0         0         0         0
## OTU_97.10005         0         0         0         0         0
```

High Dimensional Microbiome Data

- Typical species table orientation in ecology studies:
  - 32 rows
  - 43140 columns

```
# TYPICAL SPECIES TABLE ORIENTATION IN ECOLOGY STUDIES

otu %>%
  reshape2::acast(specimen_id ~ otu_id,
                  # rows = specimen_id, columns = otu_id
                  value.var = "read_count") %>%
   .[1:10,1:5]

# 32 ROWS & 43140 COLUMNS
```

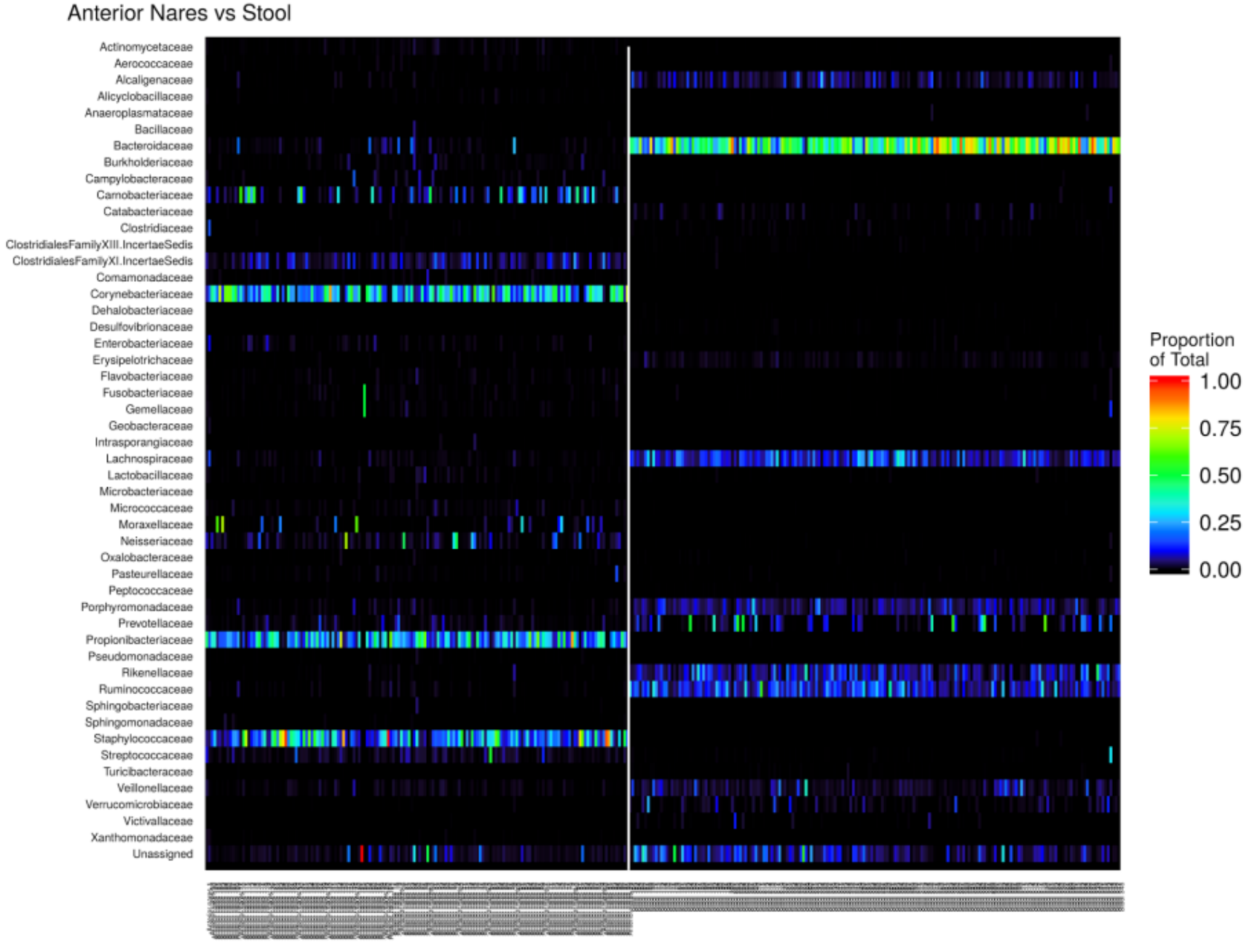
```
##           OTU_97.1 OTU_97.10 OTU_97.100 OTU_97.1000 OTU_97.10000
## 700013549        0         0          0           0            0
## 700014386        0         0          0           0            0
## 700014403        0         6        133           0            0
## 700014409        0         4          7           0            0
## 700014412        0         1          1           0            0
## 700014415        0         5          4           0            0
## 700014418        0         2          0           0            0
## 700014421        0         3         25           0            0
## 700014424        0         1          5           0            0
## 700014427        0         1          0           0            0
```

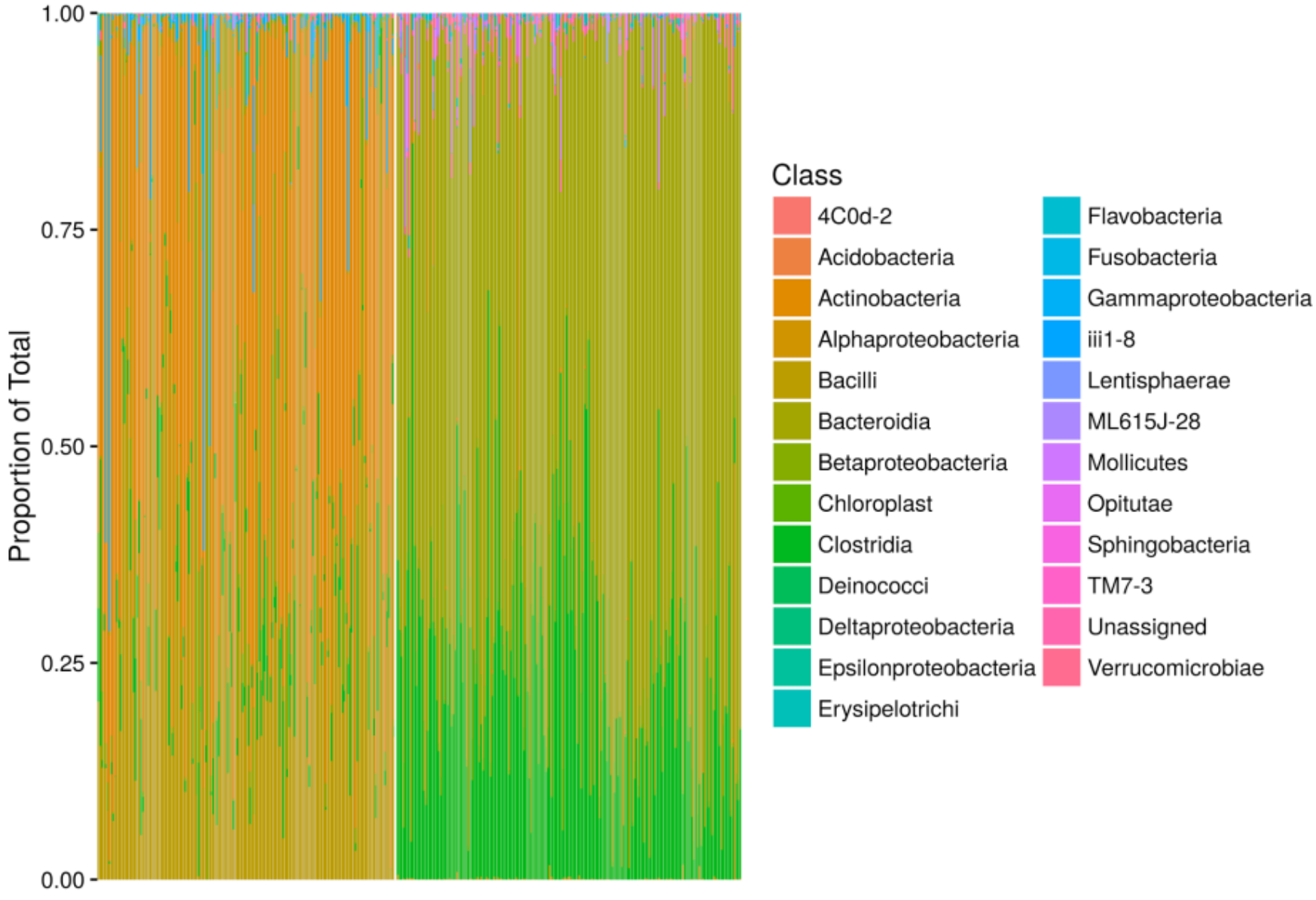
High Dimensional Microbiome Data

- 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)

Descriptive Plots

- Visualization of OTU table:
  - typically present counts as a proportion of sample total
  - choice of sample order can highlight group differences
- Limitations:
  - cannot depict full list of OTUs
  - space dictates taxonomic level presented


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


1.00
0.75
0.50
0.25
0.00
Proportion of Total
Class
4C0d-2
Acidobacteria
Actinobacteria
Alphaproteobacteria
Bacilli
Bacteroidia
Betaproteobacteria
Chloroplast
Clostridia
Deinococci
Deltaproteobacteria
Epsilonproteobacteria
Erysipelotrichi
Flavobacteria
Fusobacteria
Gammaproteobacteria
iii1-8
Lentisphaerae
ML615J-28
Mollicutes
Opitutae
Sphingobacteria
TM7-3
Unassigned
Verrucomicrobiae

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)

Single-Taxon Hypotheses

- You suspect Bacteroides has a relationship with outcome of interest...
  - *Bacteroides* (genus)?
  - *Bacteroidaceae* (family)?
  - *Bacteroidales* (order)?
  - *Bacteroidetes* (class)?
- Hypotheses focusing on specific taxa often fail to account for possibility of selection bias from culture.

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)

Alpha Diversity

Alpha Diversity

- One solution to the p >> n problem:
  - capture entire community with a single numerical value
  - richness: what's there?
  - evenness: how are members distributed?
  - diversity: richness + evenness

Alpha Diversity

- Many alpha diversity metrics (weight richness/evenness):
  - species number, Chao1 (singletons & doubletons)
  - Shannon diversity:

$$H' = -\sum p_i * \log_b (p_i)$$

(note: typically natural log or base 2 are used)

Aside on Information Theory

- Shannon diversity:

$$H' = -\sum p_i * \log_b (p_i)$$

- Claude Shannon & information entropy:

$$H(p) = -\sum p_i * \log_b (p_i)$$

"The uncertainty contained in a probability distribution is the average log-probability of an event." (McElreath *Statistical Rethinking, 2nd* 2020)

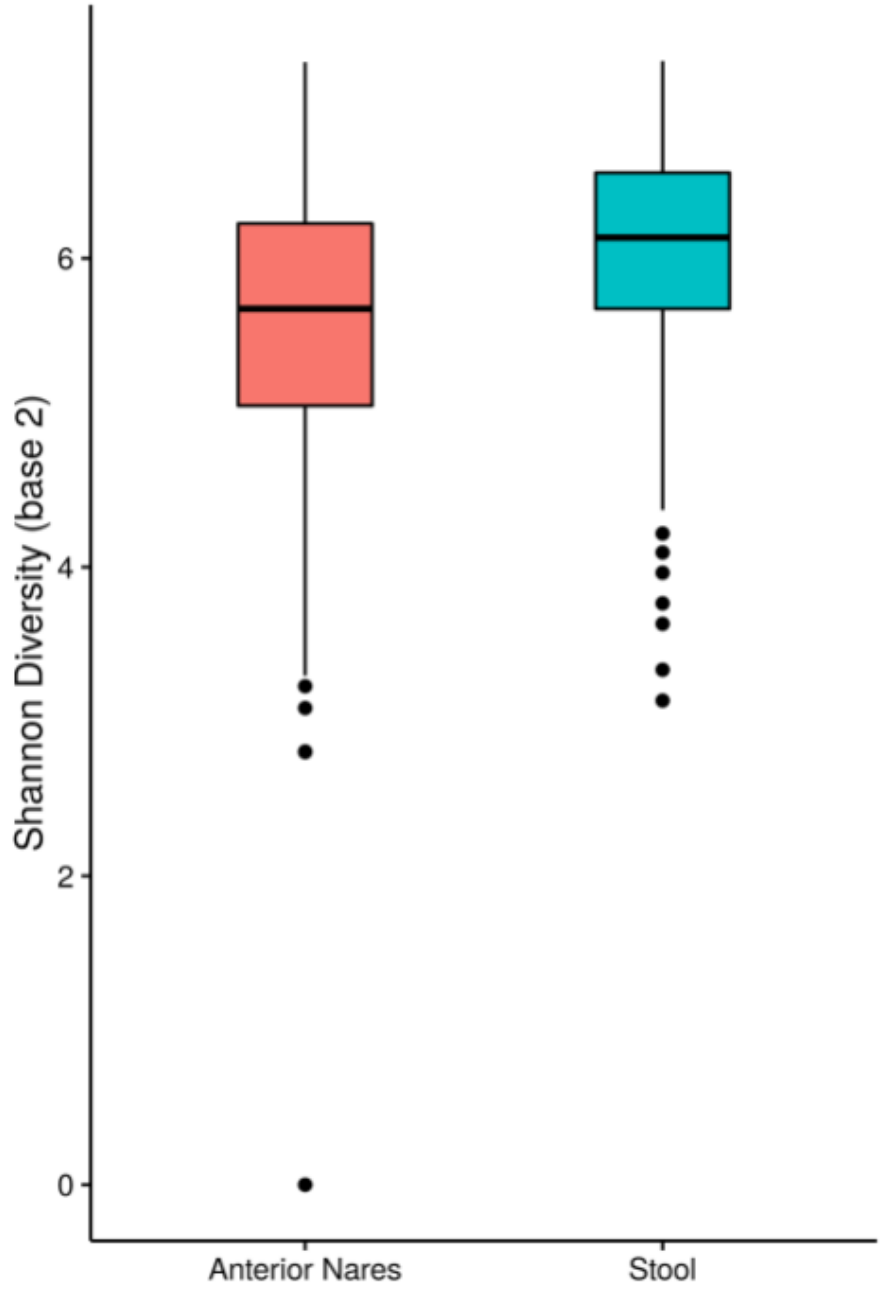
Alpha Diversity: Chao1

- The Chao1 index estimates the total species **richness**:

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2 * n_2}$$

(S = number of species; n1 = number singletons; n2 = number doubletons)

- Chao1 is particularly useful for data sets skewed toward the low-abundance classes, as is likely to be the case with microbes.


Shannon Diversity (base 2)
0
2
4
6
Anterior Nares
Stool

Rarefaction vs Rarefying

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2001, p. 4399–4406 Vol. 67, No. 10
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.10.4399–4406.2001

MINIREVIEW

Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity

JENNIFER B. HUGHES,* JESSICA J. HELLMANN,† TAYLOR H. RICKETTS, AND BRENDAN J. M. BOHANNAN

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020

Microbes Too Diverse to Count?

- “In any community, the number of types of organisms observed increases with sampling effort until all types are observed.”
- “The relationship between the number of types observed and sampling effort gives information about the total diversity of the sampled community.”
- “Pattern can be visualized by plotting an accumulation or a rank-abundance curve.”

Microbes Too Diverse to Count?

- “An accumulation curve is a plot of the cumulative number of types observed versus sampling effort.”
- “Because all communities contain a finite number of species, if the surveyors continued to sample, the curves would eventually reach an asymptote at the actual community richness (number of types).”
- “The curves contain information about how well the communities have been sampled (i.e., what fraction of the species in the community have been detected).”

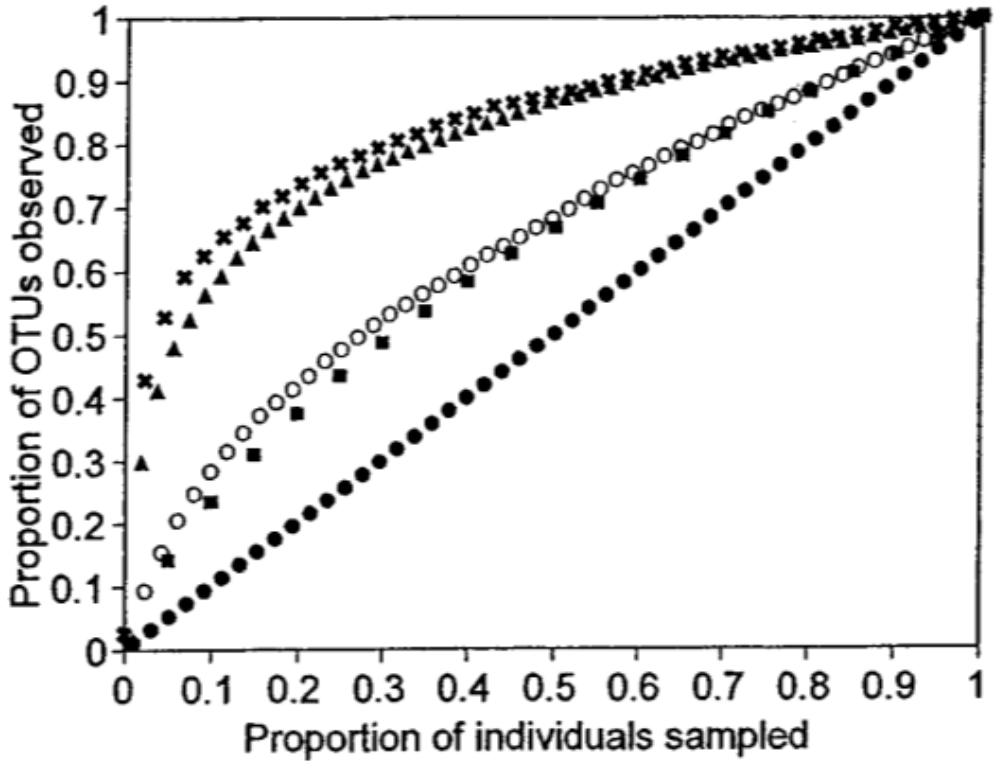
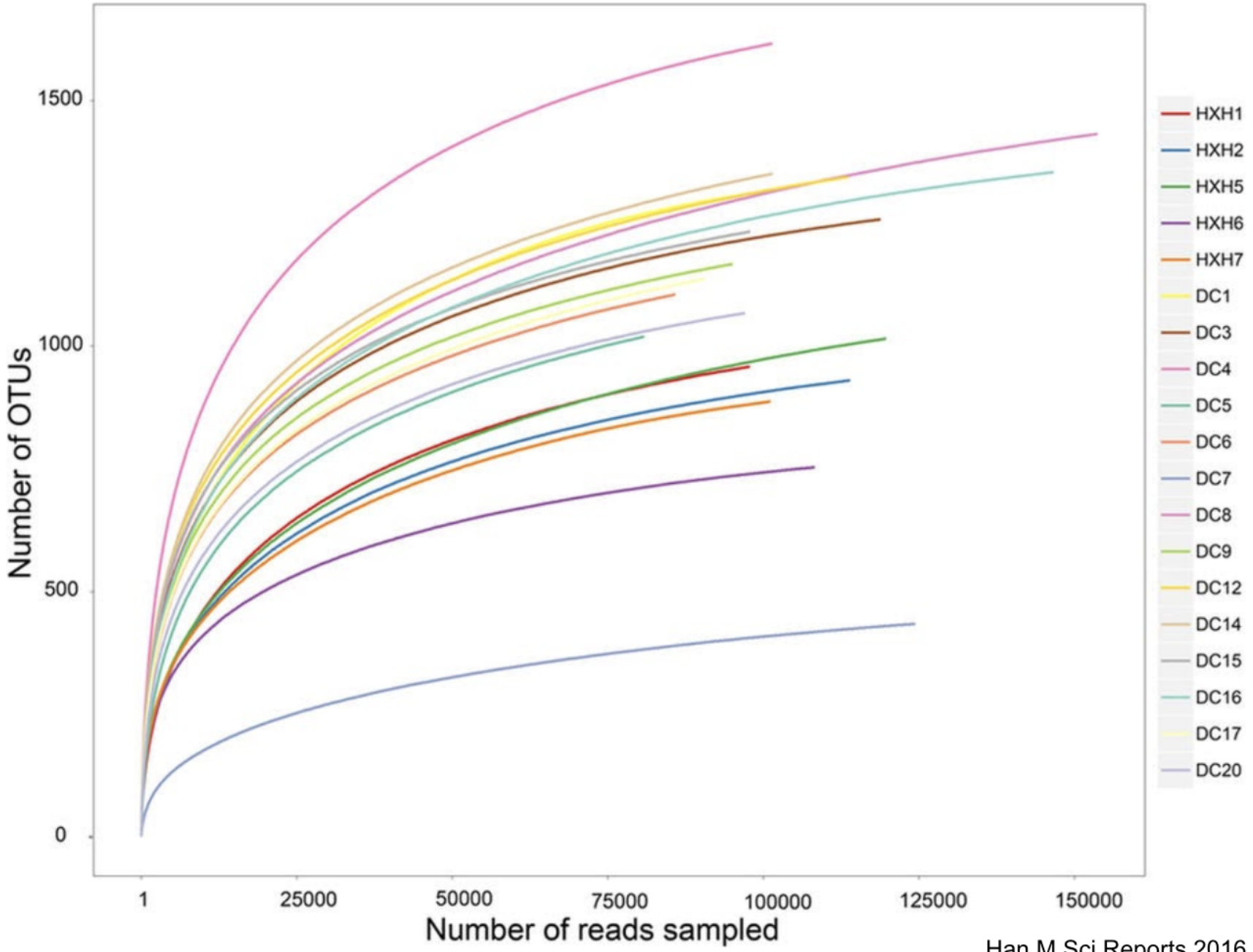


FIG. 1. Accumulation curves for Michigan plants (✕; n = 1,783) (26), Costa Rican birds (▲; n = 5,007) (J. B. Hughes, unpublished data), human oral bacteria (○; n = 264) (33), Costa Rican moths (■; n = 4,538) (56), and East Amazonian soil bacteria (●; n = 98) (6). Curves are averaged over 100 simulations using the computer program EstimateS and are standardized for the number of individuals and species observed.

Microbes Too Diverse to Count?

- “The more concave-downward the curve, the better sampled the community.”
- “The idea that microbial diversity cannot be estimated comes from the fact that many microbial accumulation curves are linear or close to linear because of high diversity, small sample sizes, or both.”
- “Ultimately, microbes—like tropical insects—are too diverse to count exhaustively.”



Han M Sci Reports 2016.

Rarefaction & Rarefying

- “Rarefaction was originally introduced as a method of estimating species richness (i.e., an alternative to parametric or nonparametric estimators like Chao1)"
- “Rarefaction compares observed richness among sites, treatments, or habitats that have been unequally sampled. A rarefied curve results from averaging randomizations of the observed accumulation curve…. The variance around the repeated randomizations allows one to compare the observed richness among samples.”

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes*

Statistics Department, Stanford University, Stanford, California, United States of America

Abstract

Current practice in the normalization of microbiome count data is inefficient in the statistical sense. For apparently historical reasons, the common approach is either to use simple proportions (which does not address heteroscedasticity) or to use *rarefying* of counts, even though both of these approaches are inappropriate for detection of differentially abundant species. Well-established statistical theory is available that simultaneously accounts for library size differences and biological variability using an appropriate mixture model. Moreover, specific implementations for DNA sequencing read count data (based on a Negative Binomial model for instance) are already available in RNA-Seq focused R packages such as edgeR and DESeq. Here we summarize the supporting statistical theory and use simulations and empirical data to demonstrate substantial improvements provided by a relevant mixture model framework over simple proportions or rarefying. We show how both proportions and rarefied counts result in a high rate of false positives in tests for species that are differentially abundant across sample classes. Regarding microbiome sample-wise clustering, we also show that the rarefying procedure often discards samples that can be accurately clustered by alternative methods. We further compare different Negative Binomial methods with a recently-described zero-inflated Gaussian mixture, implemented in a package called *metagenomeSeq*. We find that metagenomeSeq performs well when there is an adequate number of biological replicates, but it nevertheless tends toward a higher false positive rate. Based on these results and well-established statistical theory, we advocate that investigators avoid rarefying altogether. We have provided microbiome-specific extensions to these tools in the R package, phyloseq.

Citation: McMurdie PJ, Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10(4): e1003531. doi:10.1371/journal.pcbi.1003531

Editor: Alice Carolyn McHardy, Heinrich Heine University, Germany

Received October 18, 2013; **Accepted** February 3, 2014; **Published** April 3, 2014

Funding: This work was supported by the NIH (http://www.nih.gov) under grant number NIH R01-GM086884. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: susan@stat.stanford.edu

Rarefaction & Rarefying

- “Microbiome analysis workflows often begin with an ad hoc library size normalization by random subsampling without replacement, or so-called rarefying…. There is confusion in the literature regarding terminology, and sometimes this normalization approach is conflated with a non-parametric resampling technique — called rarefaction, or individual-based taxon resampling curves — that can be justified for coverage analysis or species richness estimation in some settings.”

Rarefaction & Rarefying

- “Here we emphasize the distinction between taxon re-sampling curves and normalization by strictly adhering to the terms rarefying or rarefied counts when referring to the normalization procedure, respecting the original definition for rarefaction.”
- "Rarefying defined by the following steps: (1) Select a minimum library size... (2) Discard libraries (microbiome samples) that have fewer reads than [minimum library size]... (3) Subsample the remaining libraries without replacement such that they all have [same] size"
- “Rarefying is now an exceedingly common precursor to microbiome multivariate workflows that seek to relate sample covariates to sample-wise distance matrices.”

Rarefaction & Rarefying

- “Despite its current popularity in microbiome analyses **rarefying biological count data is statistically inadmissible** because it requires the omission of available valid data. This holds even if repeated rarefying trials are compared for stability as previously suggested.”

R's vegan package

vegan::diversity()

```
library(vegan)

diversity(x = c(100,0,5,10), index = "shannon", base = 2)

diversity(x = c(5,5,5,5), index = "shannon", base = 2)
```

```
## [1] 0.6784071

## [1] 2
```

vegan::rrarefy()

```
library(vegan)

rrarefy(x = c(100,0,5,10), sample = 10)

rrarefy(x = c(1421,170,205,3607), sample = 1000)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    8    0    0    2
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  248   34   41  677
```

Questions?

Post to the discussion board!

Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu