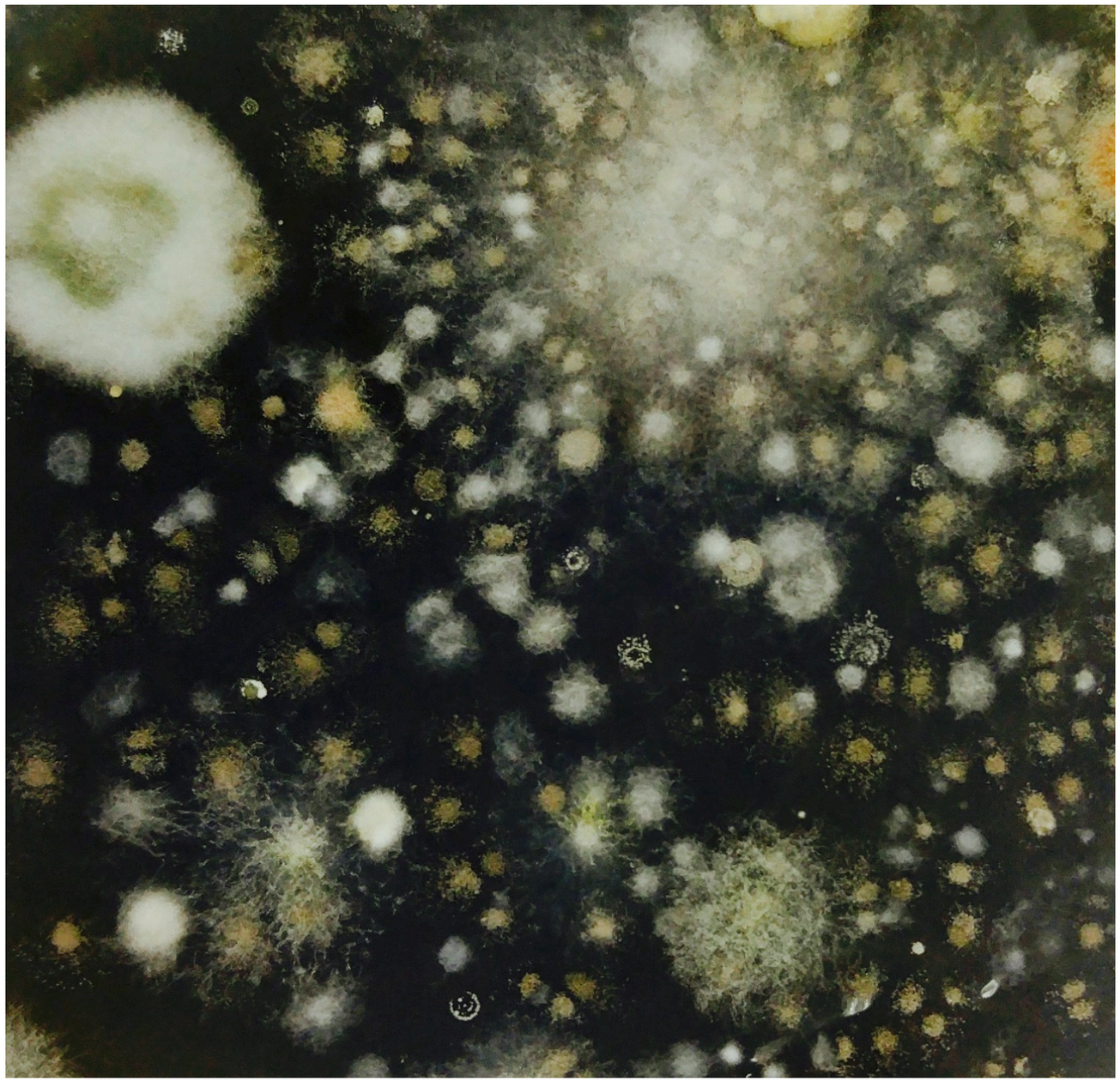
Statistical miscellanea

Diversity, ordination, prediction…

Statistical Diversity Lab @ University of Washington
Amy Willis — @AmyDWillis — Associate Professor
Shirley Mathur — PhD Candidate
Sarah Teichman — PhD Candidate
María Valdez — PhD Candidate

Microbial diversity:

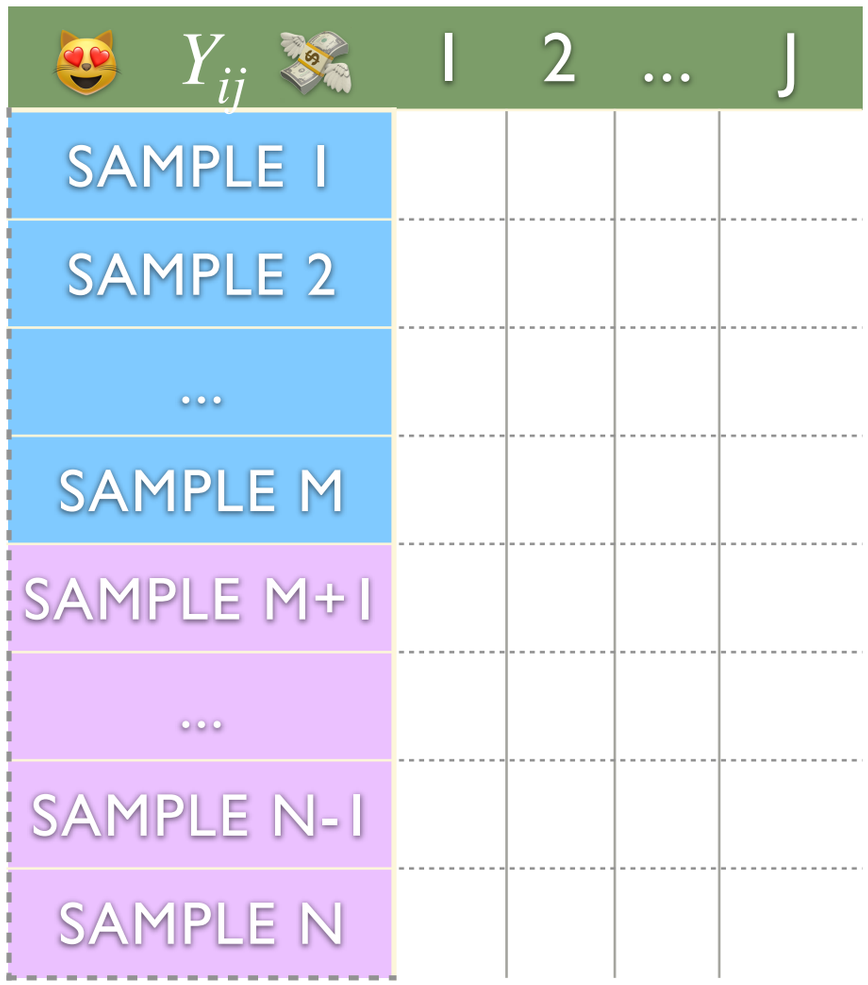
Estimation & comparison



Tools for testing null hypotheses that are false

Diversity

- Low dimensional summaries of entire communities
- α-diversity: one community
- β-diversity: multiple communities



Diversity & parameters

- There are multiple choices to make when analyzing microbial diversity
  - Which taxonomic level? (strain/species/genus...)
  - Which diversity parameter?
  - Which estimate of the diversity parameter?

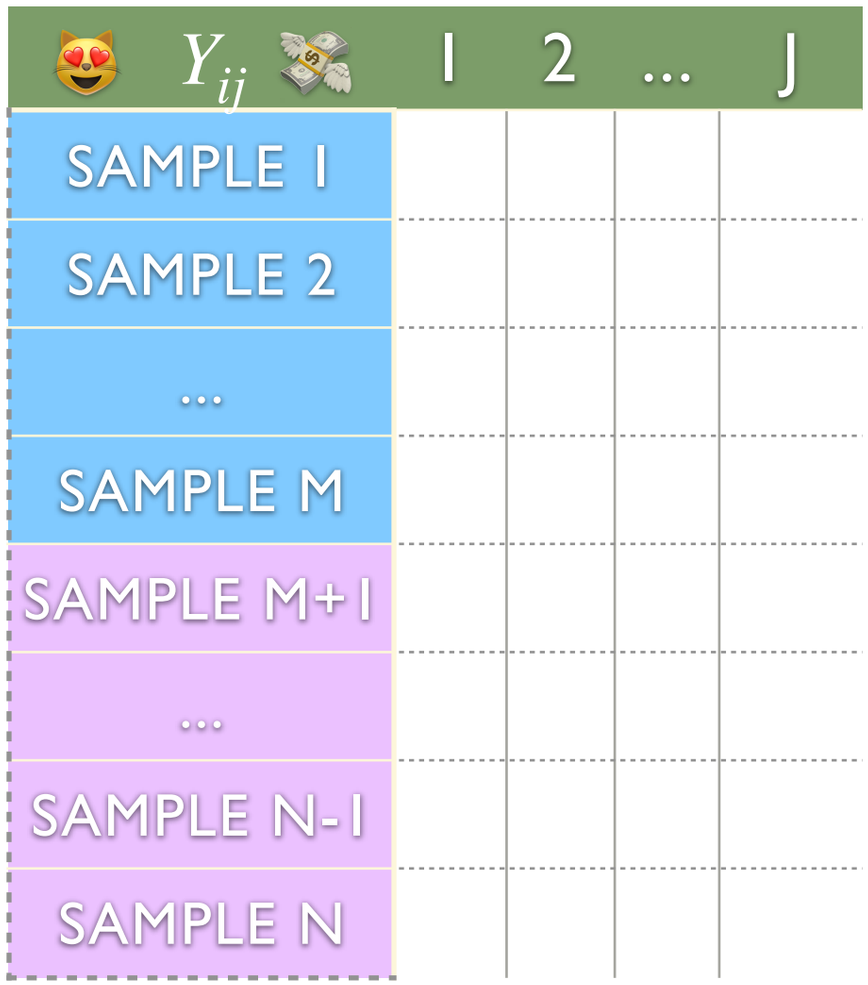
Diversity & parameters

- There are multiple choices to make when analyzing microbial diversity
- Which taxonomic level? (strain/species/genus...)
- **Which diversity parameter?**
- Which estimate of the diversity parameter?

Microbial universe

- Y_{ij} = true number of unit j in sample i

- $$p_{ij} = \frac{Y_{ij}}{\sum_{j'=1}^{J} Y_{ij'}}$$



α-diversity

- Amy: Any function of
- $p_{i1}, \ldots, p_{iJ}$ OR

  phylogeny
- $p_{i1}, \ldots, p_{iJ}$ and ~~some info about relationships amongst groups~~

is a valid α-diversity parameter

α-diversity

- Some examples of α-diversity parameters include
  - Species richness: $C_i = \#\{j : p_{ij} > 0\}$
  - Simpson's index: $\sum_{j:p_{ij}>0} p_{ij}^2$
  - Shannon diversity: $-\sum_{j:p_{ij}>0} p_{ij} \log p_{ij}$
  - Shannon's E: $\frac{-\sum_{j:p_{ij}>0} p_{ij} \log p_{ij}}{C_i}$

α-diversity

- My wish list remains unchanged
- You choose a meaningful parameter to estimate
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

α-diversity

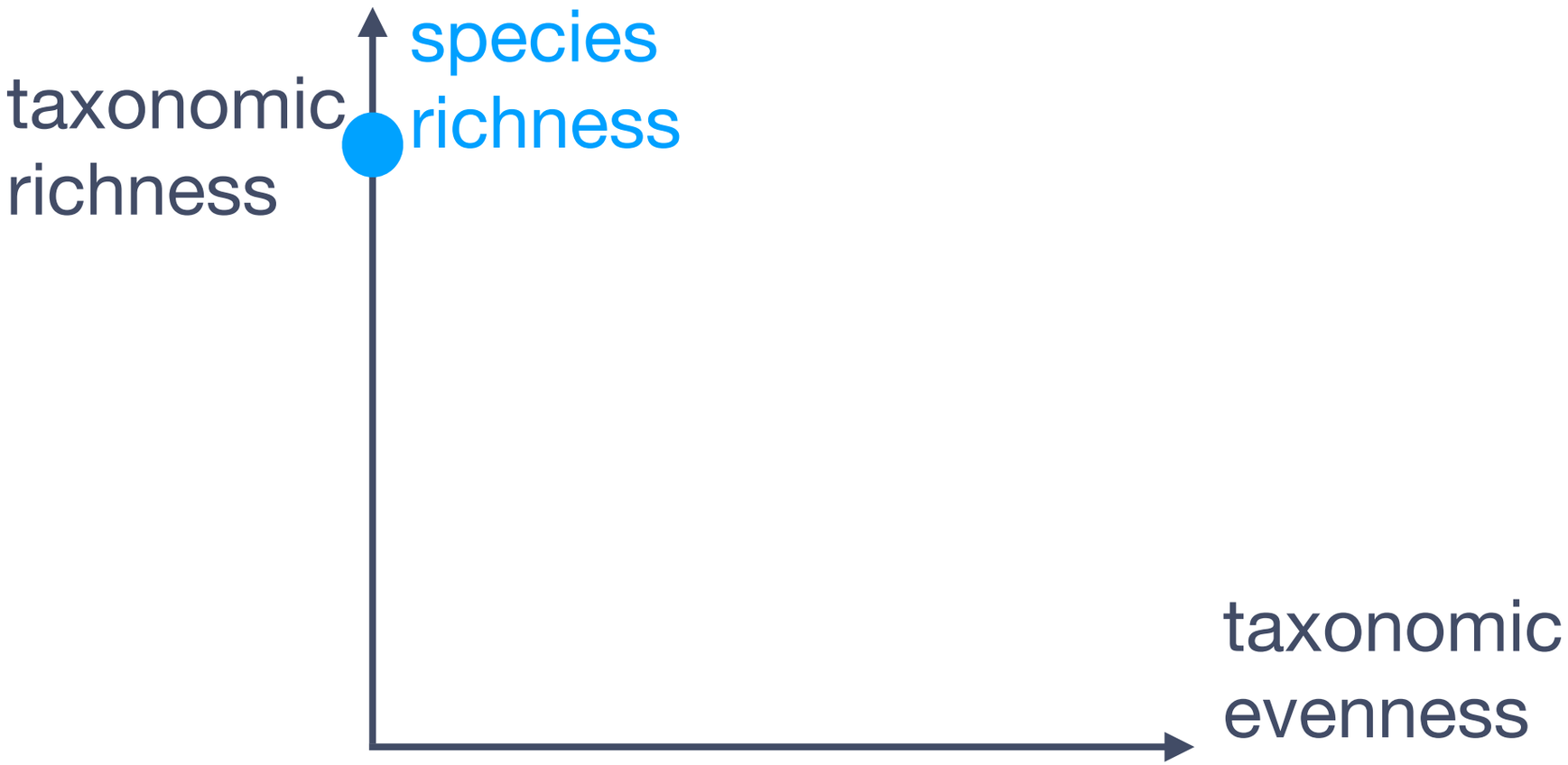
- My wish list remains unchanged
- **You choose a meaningful parameter to estimate**
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

What α-diversity parameter? You decide

- Think: What difference do you want to highlight?

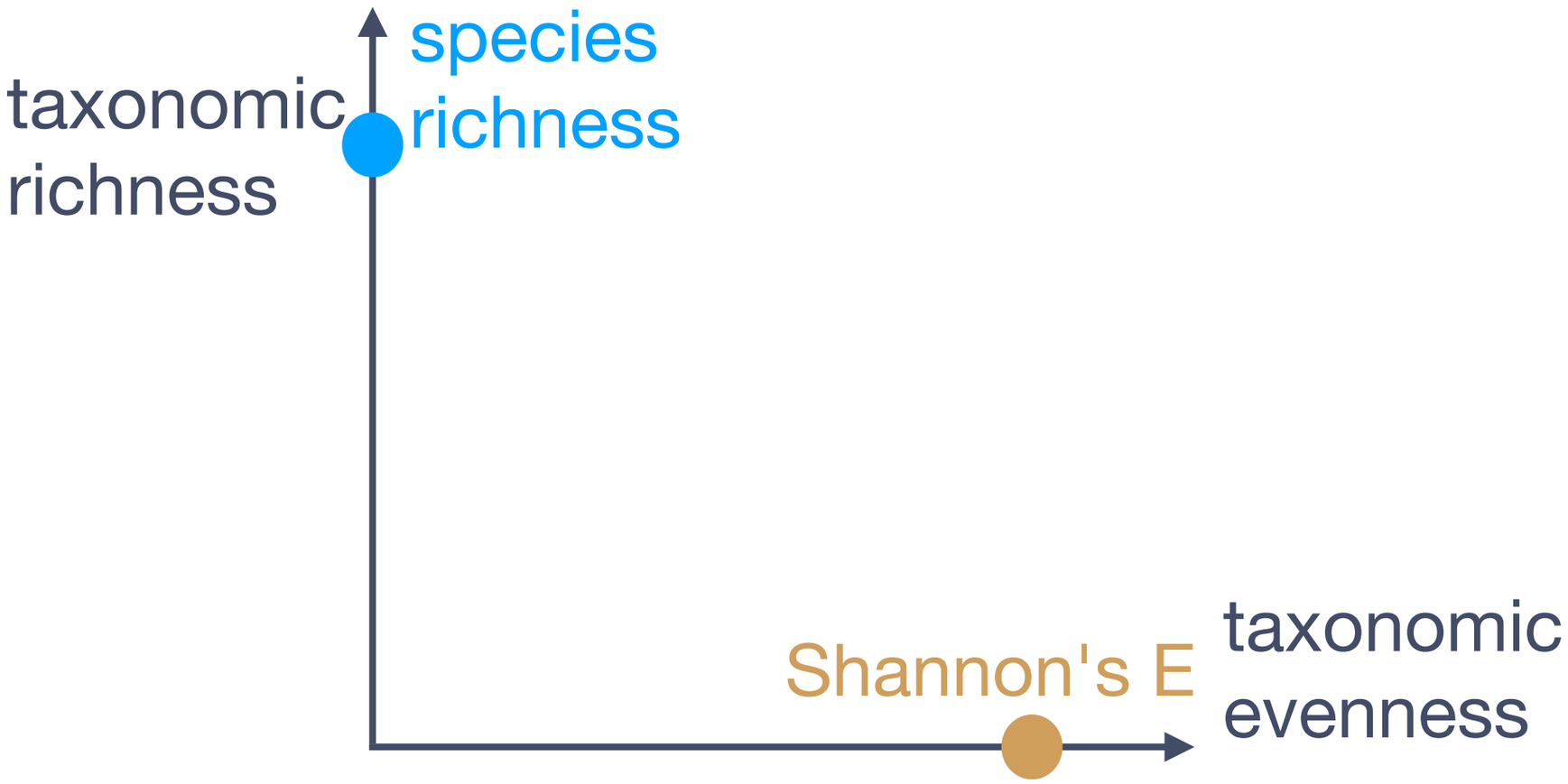
A heuristic diagram I find helpful!

What α-diversity parameter? You decide



A heuristic diagram I find helpful!

What α-diversity parameter? You decide



A heuristic diagram I find helpful!

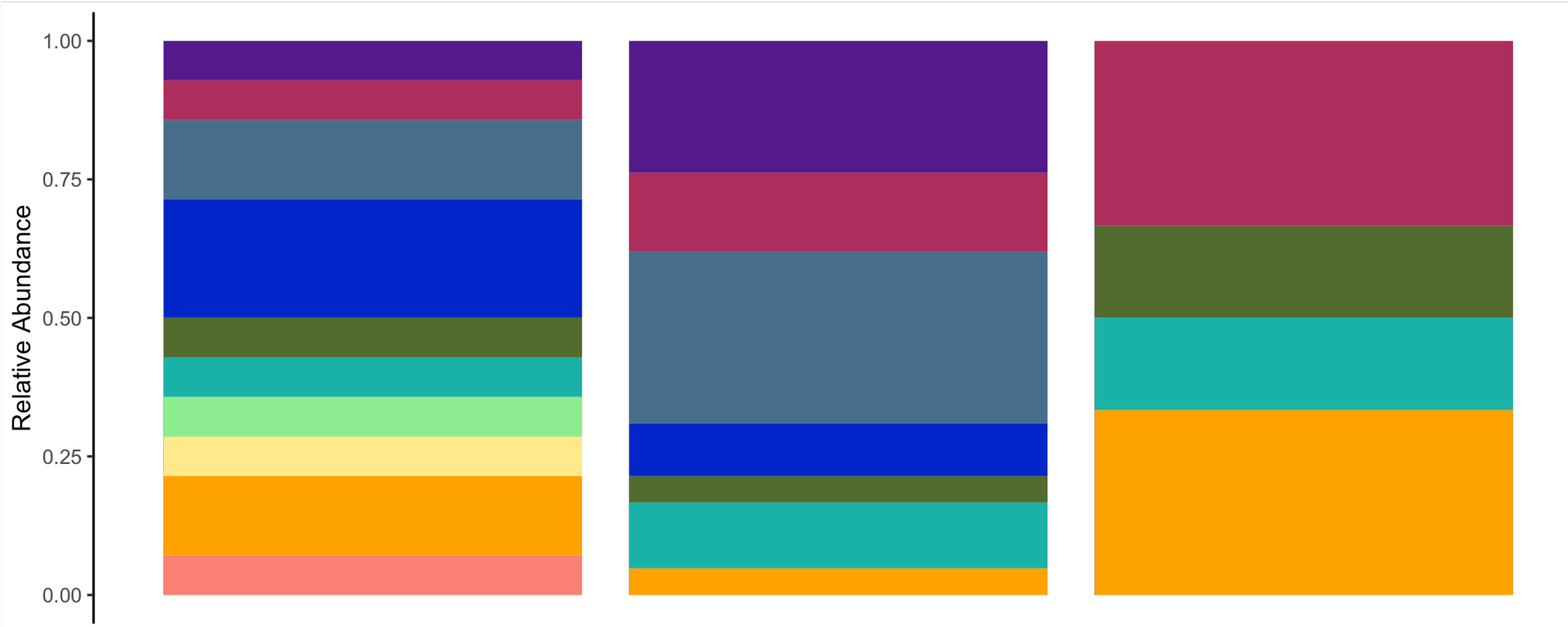
What α-diversity parameter? You decide

A heuristic diagram I find helpful!

What α-diversity parameter? You decide

This is a question of *parameter choice:*
Which parameter highlights the differences I care about?

A heuristic diagram I find helpful!



Richness	10	7	4
Shannon	2.21	1.75	1.33
Evenness	0.96	0.90	0.96
Simpson's	0.88	0.80	0.72

α-diversity estimators

- As always, we don't know the true abundances Y_{ij}
    - So we don't know the true relative abundances p_{ij}

- We can't know/"calculate" our α-diversity parameter of choice

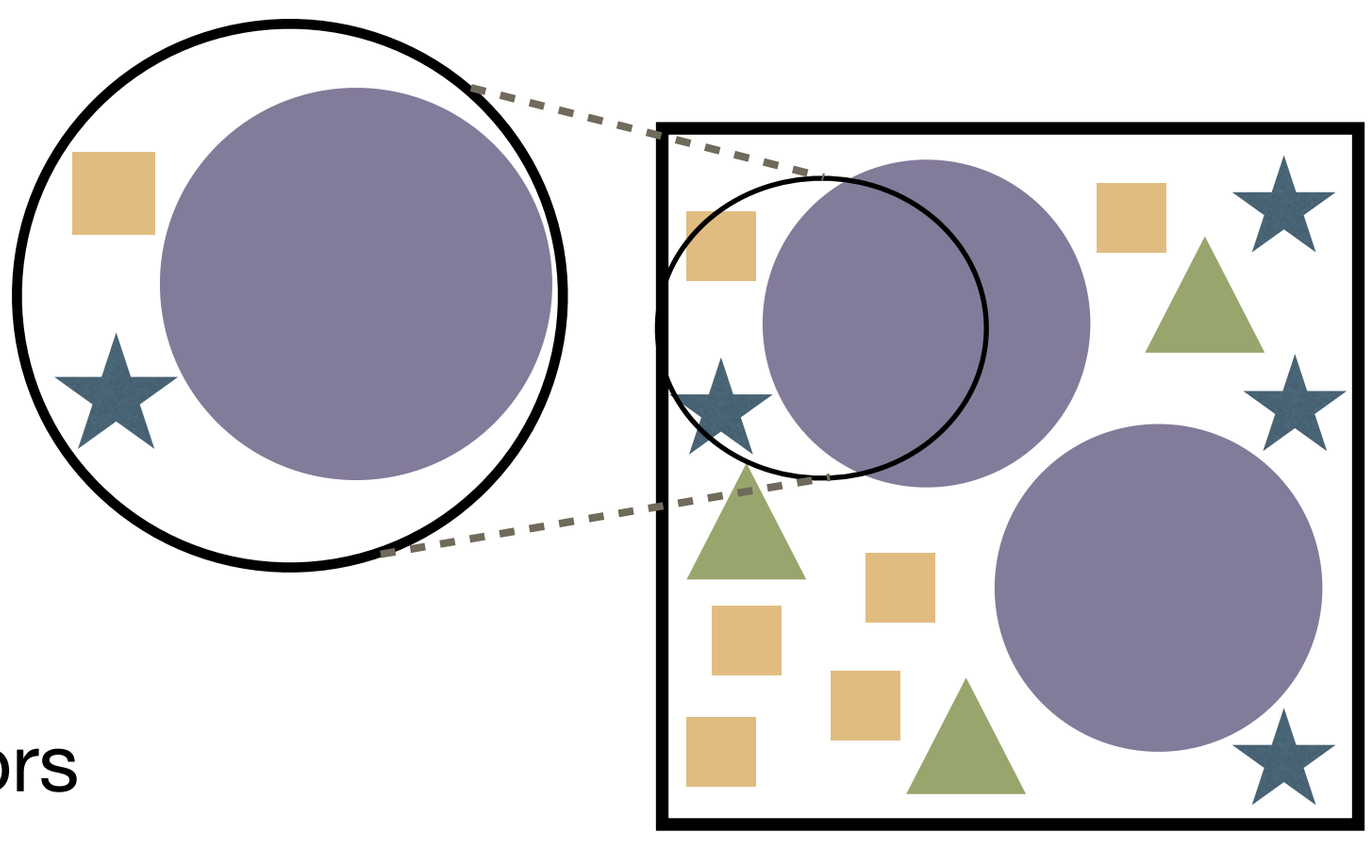
- We need to *estimate* it!

The "classical" approach

- Substitute the observed relative abundances $\hat{p}_{i1}, \ldots, \hat{p}_{iJ}$ for the unknown, true abundances $p_{i1}, \ldots, p_{iJ}$ and pretend nothing happened
  - e.g. Estimate the richness with: $c_i = \{\#j : \hat{p}_{ij} > 0\}$
  - e.g. Estimate the Simpsons index: $\sum_{j:\hat{p}_{ij}>0} \hat{p}_{ij}^2$

Unobserved species are *one* source of bias

- Plug-in estimates
  - Species richness: *underestimates*
  - Simpson: *overestimates*
- ~~Need new indices~~
- Need new estimators




>r32
atcttcgatcggctggagacactaatacta
>r37
ggctccctctgtatcagaatggaccaggcg
>r40
aaggcaaacttcagtgtgatcttcgatcgg
>r9
>r12
acactaatactactaagagctttcactgaa
>r46
cccaagcaaaaagtagcaggctccctctgt
>r10
gactggttcatgctcatgcccaagcaaaaa
>r8
atggaccaggcgattatggataagaacatc
>r35
gataagaacatcatactgaaggcaaacttc
>r22
>r11
gacatcgaaacggccactcatgctggaaag
r41
tggagaggattctggaggaagaatccgac
>r33
gaagaatccgacgaggcactcaaaatgact
>r27
ctcaaaatgactattgcctctgtgcctgct
>r1
tctgtgcctgcttcacgctacctaactgat
>r23
tacctaactgatatgactcttgaagagatg
Friendly (principled) neighbourhood statisticians

α-diversity estimators

- We evaluate estimators with respect to two criteria
- bias = under/overestimation
- variance = how variable/uncertain they are

Species richness

- The "species problem": how many species were missing from the sample

- Idea

  - If many rare species in sample, likely there are many missing species

  - If few rare species in sample, likely there are few missing species

  - Use data on rare species to predict # missing species

Species richness estimation

- The necessary data for richness is the **frequency counts**

- f_k = number of species observed k times

- f_1 = singletons,

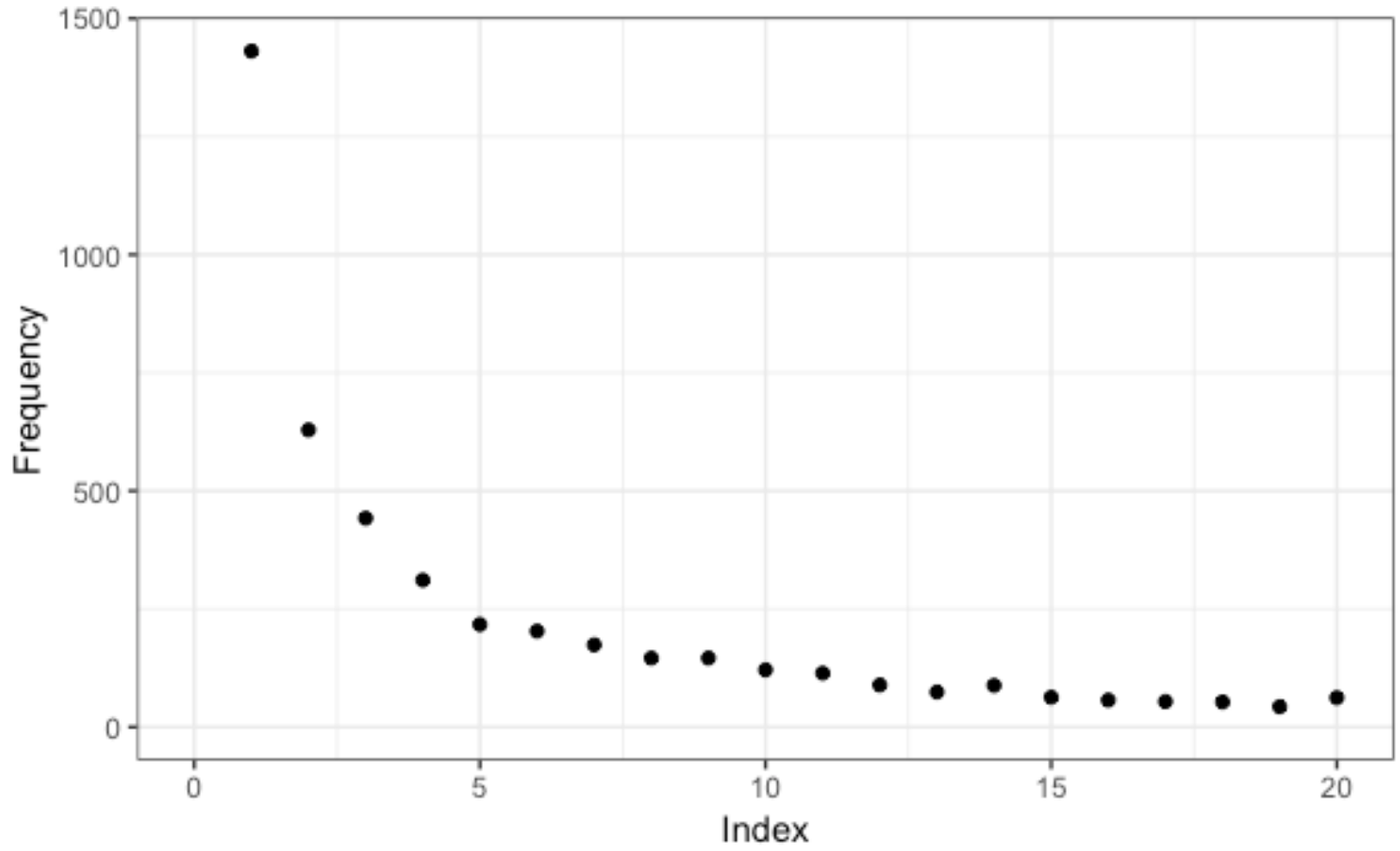
- f_2 = doubletons, …

- e.g. 1431 strains observed once

```
> library(phyloseq)
> library(magrittr)
> library(breakaway)
> data("GlobalPatterns")
> GlobalPatterns %>%
+   otu_table %>%
+   build_frequency_count_tables %>%
+   head(1)
$CL3
      Index Frequency
 [1,]     1      1431
 [2,]     2       629
 [3,]     3       442
 [4,]     4       311
 [5,]     5       217
 [6,]     6       203
 [7,]     7       174
 [8,]     8       146
 [9,]     9       146
[10,]    10       121
[11,]    11       114
[12,]    12        89
[13,]    13        74
[14,]    14        88
```

Species richness estimation

- Idea: extend the pattern in f_1, f_2, f_3… to f_0



- Rare taxa are most informative for missing taxa

Species richness estimation

- Good options
  - breakaway::breakaway() - Kemp models
  - breakaway::chao_bunge() - Negative binomial model
  - breakaway:: objective_bayes_*() - mixed Poisson
  - CatchAll - mixed Poisson

- Bad options
  - anything involving rarefaction
  - QIIME2: chao1; scikitbio…
  - R:vegan::…

Species richness estimation

- "Chao1 diversity index" $c_i + \frac{f_{i1}^2}{2f_{i2}}$ is *not* an index
  - It's an *estimate* of species richness, and it's based on the questionable assumption that

    all species have the same abundance

  - Large negative bias; very high variance
- **Should not be used**

Species richness estimation

- My perspective
    - Species richness is a parameter
    - Three estimators of many
        - sample species richness: assumes everything seen
        - Chao1: assumes everything equally abundant
        - *breakaway*: flexible parametric models for frequency counts

Species richness estimation

- Species richness estimation is *hard*

- Recall: You are trying to predict f_0 from $f_1, f_2, f_3, \ldots$

- Where are we the least confident in our data?

Bias and diversity

- Alternative approach ~~that I loathe~~ that I just can’t get upset about any more: rarefaction

- Idea:
  - Discover more diversity with more sequencing
  - Can’t directly compare samples with different depths
  - Randomly throw away reads until all samples have same depth

Bias and diversity

- Alternative approach ~~that I loathe~~ that I just can't get upset about any more: rarefaction

- Idea:
  - Discover more diversity with more sequencing
  - Can't directly compare samples with different depths
  - Randomly throw away reads until all samples have same depth
- Better idea: **Statistical estimation that accounts for different sequencing depths!**

Bias and diversity

- Alternative approach ~~that I loathe~~ that I just can't get upset about


PLOS | COMPUTATIONAL BIOLOGY

BROWSE PUBLISH ABOUT

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

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

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.1371/journal.pcbi.1003531

h

- Better idea: **Statistical estimation that accounts for different sequencing depths!**

Bias and diversity

- Alternative approa

PLOS | COMPUTA BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Waste Not, Want Not:

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org

Microbiome

Home About Articles Submission Guidelines

Research | Open Access

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde and Rob Knight

Microbiome 2017 **5**:27

https://doi.org/10.1186/s40168-017-0237-y | © The Author(s). 2017

Received: 9 October 2015 | **Accepted:** 27 January 2017 | **Published:** 3 March 2017

- Better idea: **Statis**
  sequencing dept

Bias and diversity

- Alternative approa

Bias and diversity

- Alternative approa

PLOS COMPUTA BIOLOGY

Microbiome

Home About Articles Submission Guidelines

frontiers in Microbiology

| Editor's Pick | Human Microbiome | Research Article | 22 January 2024

Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses

Author: Patrick D. Schloss | AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/msphere.00354-23 • Check for updates

4,413 CITE PDF/EPUB

Amy D. Willis*

Department of Biostatistics, University of Washington, Seattle, WA, United States

Understanding the drivers of diversity is a fundamental question in ecology. Extensive literature discusses different methods for describing diversity and documenting its effects on ecosystem health and function. However, it is widely believed that diversity depends on the intensity of sampling. I discuss a statistical perspective on diversity, framing the

7

d: 3 March 2017

Other challenges with estimating diversity

- Unobserved species are *one source of bias* in estimating some α-diversity parameters
- What other problems are there with estimating

$$-\sum_{j:p_{ij}>0} p_{ij} \log p_{ij}$$

?

Other challenges with estimating diversity

Other challenges with estimating diversity

- We have *uncorrectable* bias in Shannon, Simpson diversity
- “Given that normally we are *comparing average* [Shannon] diversity across samples, what problems does this cause?”
    - Special thanks to Shirley for studying this for us! 🤩

Estimating differences in diversity

- We are interested in estimating

$\beta_1 =$ mean Shannon diversity in group 1 samples

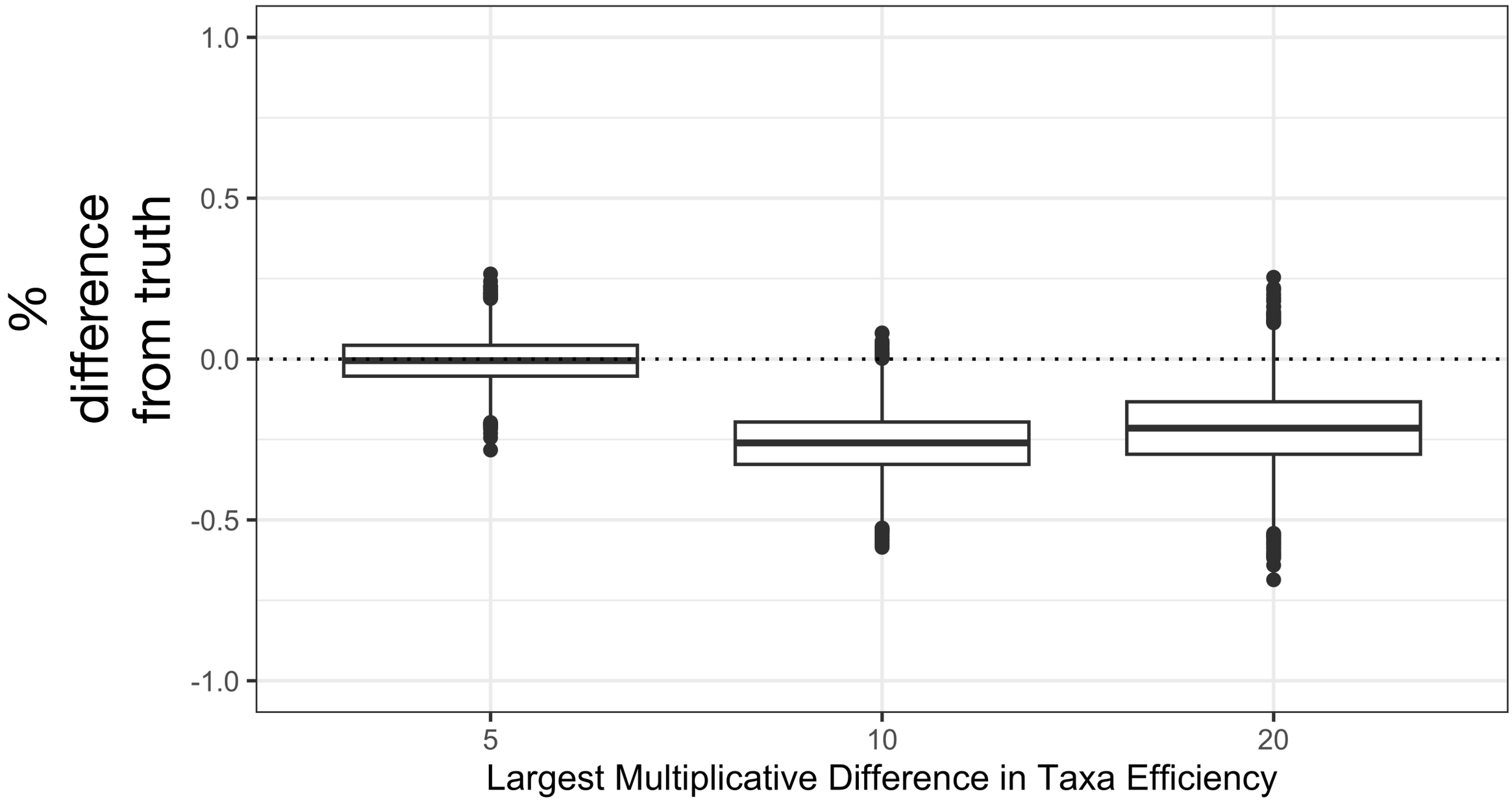
minus

mean Shannon diversity in group 2 samples

via a linear regression on plug-in Shannon diversities

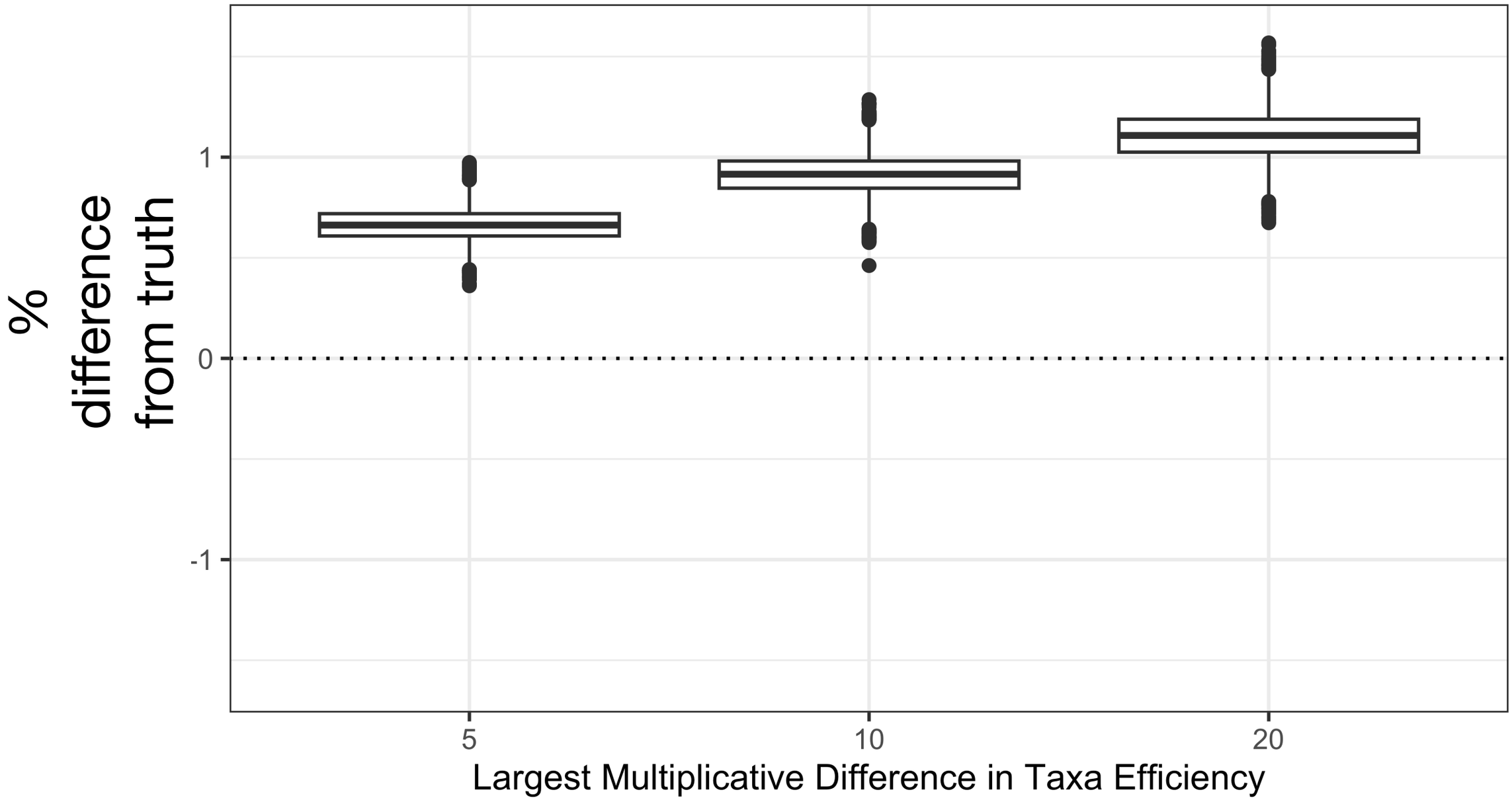
in the presence of differential detection

- Our estimates of β_1 are biased
- Unsurprisingly, more and more so as detectability is more and more variable
    - ~25% underestimation for 10x differences in detection



Mathur et al, 2024+

- When detectability is correlated with abundance, ~100% overestimation for 10x differences



Mathur et al, 2024+

Estimating differences in diversity

- In the presence of differential detection
  - We have biased estimates of Shannon diversity plug-in
  - We have biased estimates of differences in Shannon diversity eg via linear regression on plug-in'ed
  - More samples don't fix the problem more precision in your bad estimator

Mathur et al, 2024+

Bias and diversity

- My perspective
- Statisticians weren’t paying attention until ~2014
- Statisticians know how to address problems like “this estimator is biased”

Bias and diversity

- My perspective
- Species richness estimation is hard
    - Prediction off the range of the data
    - Most useful info is the lowest quality

Bias and diversity

- My perspective
    - If you believe that throwing away data is the right thing to do in general or for your science, go for it

α-diversity

- α-diversity analyses as typically practiced
    - Choose one/two of: species richness, Shannon diversity, (Inverse) Simpson
        - Estimate sample-specific α-diversity: α_i
        - Estimate the average estimated parameters across populations (t-test or linear regression) betta
- SGTM!

α-diversity

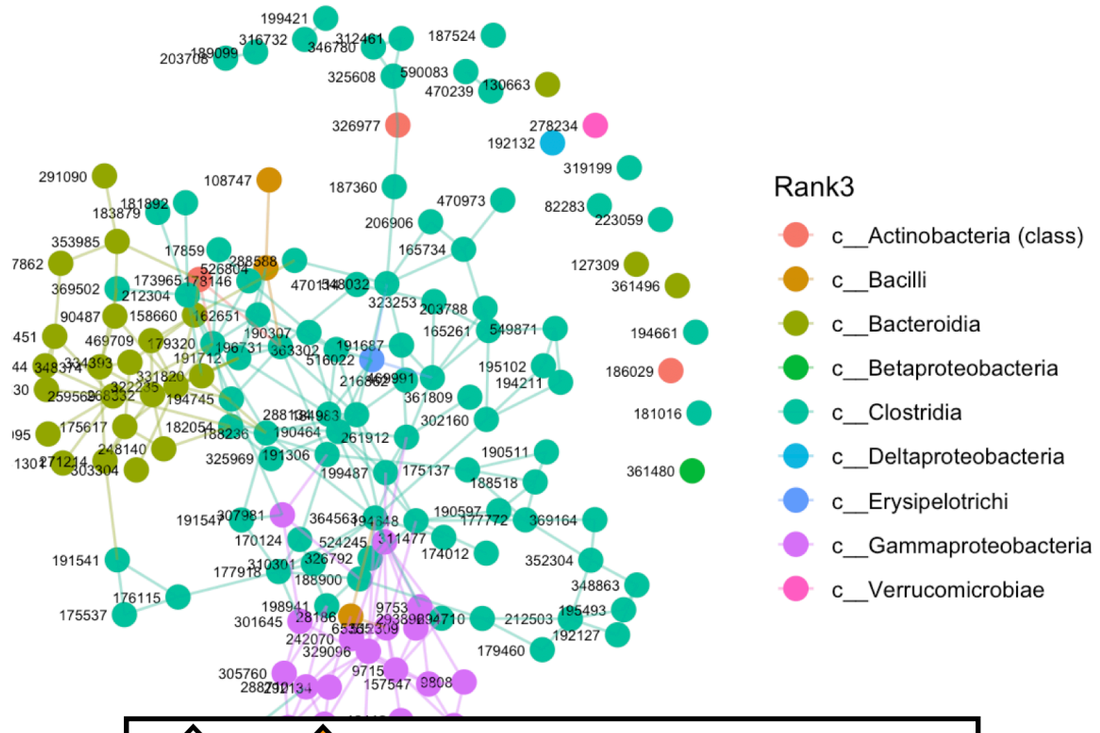
- Rejecting H_0 that "diversity is equal across groups" tells you *nothing* about *what's* different between the groups

- α-diversity reduces your rich, fascinating data into one number

- Personally, I want to know *what's* different which is why 2024 Amy is more excited about differential abundance

α-diversity: Shannon & simpson

- Slightly different approach:
- Share strength across multiple samples to estimate C and p_1, p_2, …, p_C, then use network models to get variance



github.com/adw96/DivNet

Accessing `Diversity` Lab

1. Go to schedule on Wiki to Thursday morning, click on “Labs”
2. Copy the command under the lab we’re working on

`breakaway` and `DivNet` lab:

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2024/main/stats-labs/diversity-lab/d
```

3. Run this command in your RStudio Server console

```
> download.file("https://raw.githubusercontent.com/statdivlab/stamps2024/main/stats
  -labs/diversity-lab/diversity-lab.Rmd", "diversity-lab.Rmd")
```

β-diversity

Tools for testing null hypotheses that are both false and uninterpretable

β-diversity

- Consider the rows of relative abundances: $p_{i\cdot} = (p_{i1}, p_{i2}, \ldots, p_{iJ})$

- β-diversity parameters are usually distances between relative abundances vectors

- Bray-Curtis: $\beta_{BC}(i, i') = 1 - \sum_{j=1}^{J} \min(p_{ij}, p_{i'j})$

- Jaccard: $\beta_J(i, i') =$ % taxa not shared

- UniFrac: Weights phylogeny

β-diversity

- Typically, we estimate these… using plug-in estimators
    - i.e.., Substitute
- Bray-Curtis: $\hat{\beta}_{BC}(i, i') = 1 - \frac{2\sum_j \min(W_{ij}, W_{ij'})}{\sum_j W_{ij} + W_{ij'}}$
- Jaccard: $\hat{\beta}_J(i, i') =$ % taxa not observed in both

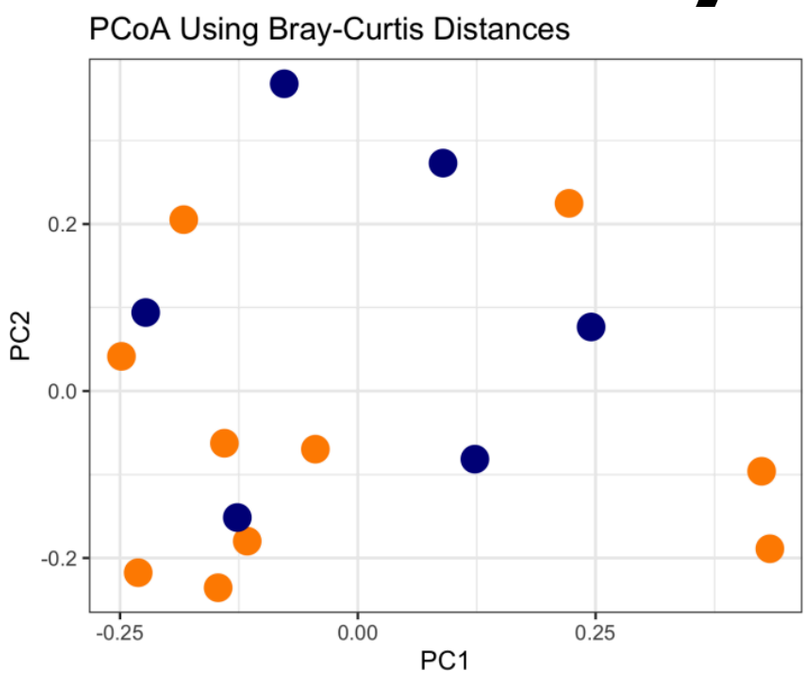
β-diversity

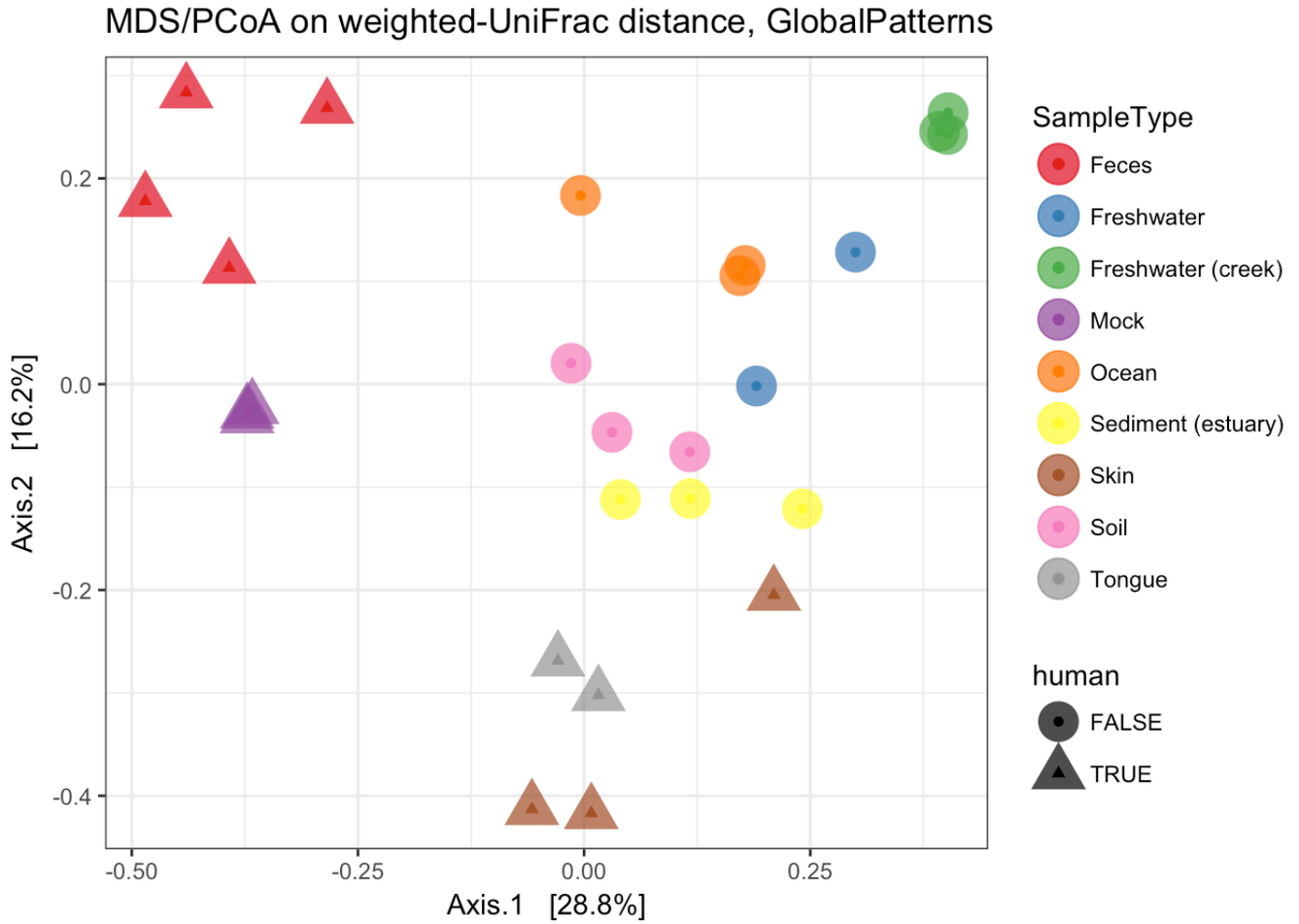
Distances	Sample 1	Sample 2	…	Sample n
Sample 1	0	d-hat(1,2)	…	d-hat(1,n)
Sample 2	d-hat(1,2)	0	…	d-hat(2,n)
…	…	…	…	…
Sample n	d-hat(1,n)	d-hat(2,n)	…	0

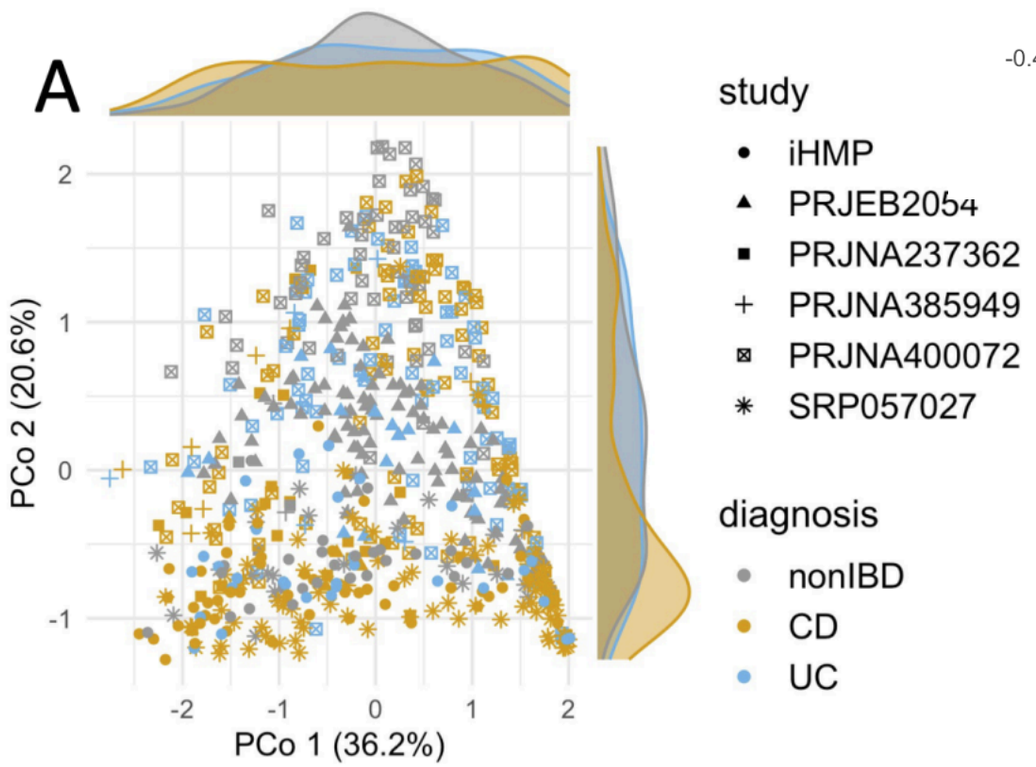
β-diversity

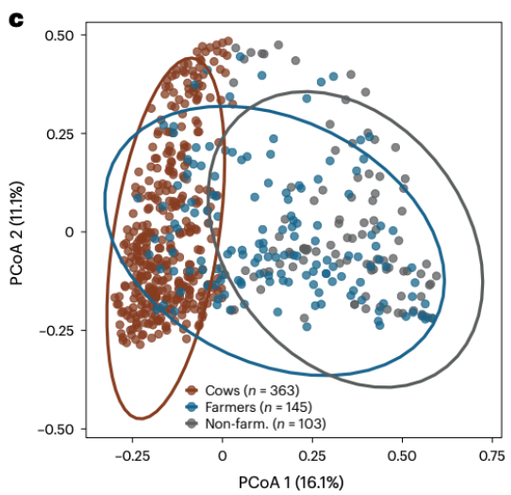
- Challenge: These $n \times n$ tables are hard to learn from
- Approach: Make it easier?
- Put n points on a scatterplot, one per sample
- PCoA/MDS: “Find the best arrangement for the points such that $\text{distance}_{\text{scatterplot}}(i, i')$ is as close to $\hat{\beta}(i, i')$ for all pairs i, i'

β-diversity









β-diversity

- This is all fine
- Things that I don’t love
  - Putting rings around them 💍 ambiguous
  - Throwing p-values on this with PERMANOVA 🫛

β-diversity

- PERMANOVA is just a tool to get a p-value 🫛
- PERMANOVA does not estimate a parameter
- Most generously, it models

Centroid for sample i using distance d

$$= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

and tests $\beta_1 = 0$ or $\beta_1 = \ldots = \beta_p = 0$

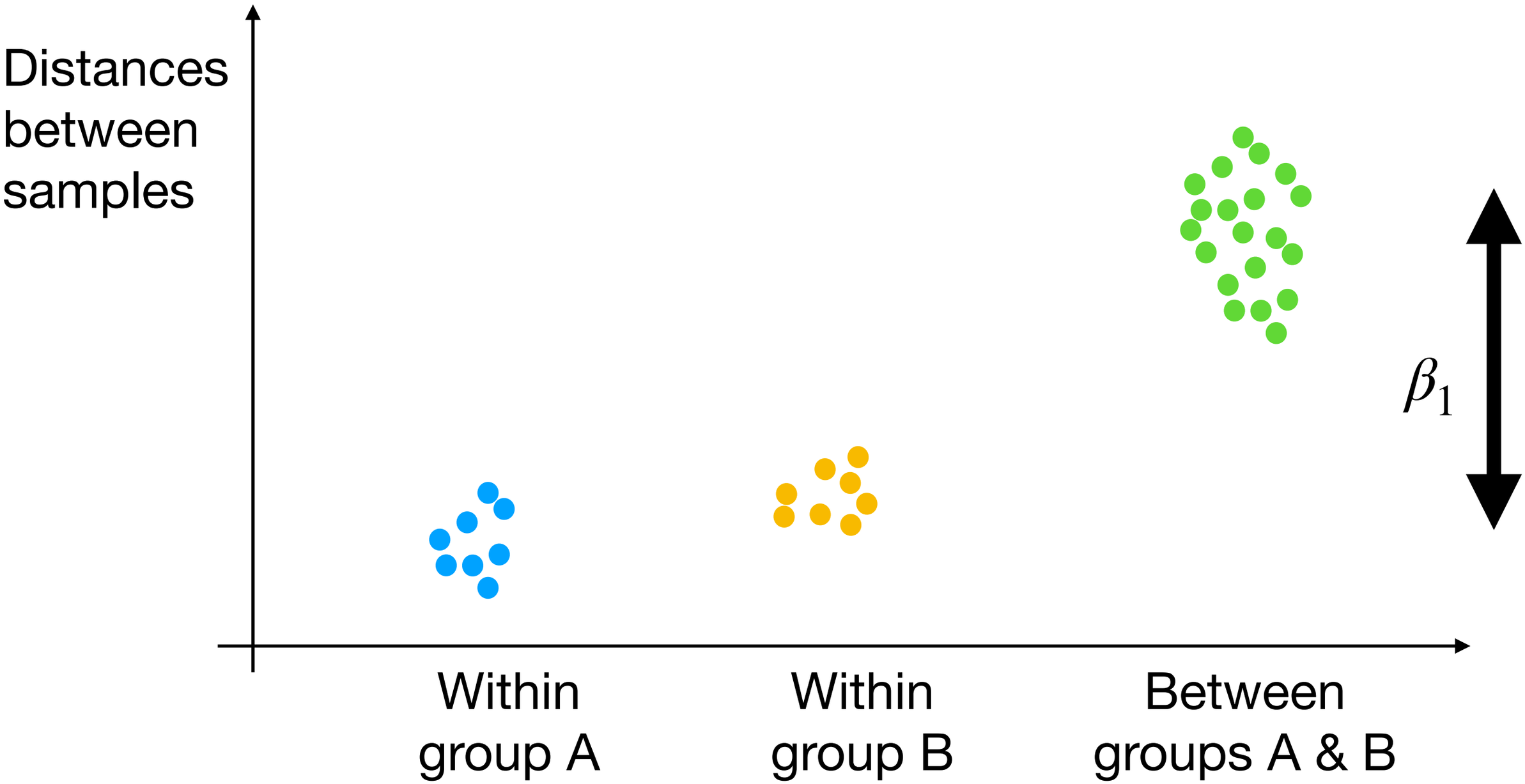
β-diversity

- *Centroids* are hard to interpret middle in distance-land
- If you care about comparing distances within- and across-groups, consider

$$\text{average distance}(i, i') = \beta_0 + \beta_1 \mathbf{1}_{\{i,i' \text{ from different groups}\}}$$

- β_1 is the difference in average distances between samples from different groups compared to the average distances between samples from the same group

Alternative viz



β-diversity

- Plotting your data = great
- Quantifying uncertainty = great
- Testing meaningful & interpretable null hypotheses = great
- Testing uninterpretable null hypotheses = not great

β-diversity

- It's up to you what β-diversity parameter you want to estimate/plot
- We can estimate some β-diversity parameters better than others

β-diversity

- If you made me make a β-diversity parameter, I'd choose Aitchison distance
  - Implicit parameter I think

$$\beta_{\text{Aitchison}}(i, i') = \sum_{j=1}^{J} \left(\log(p_{ij}) - \log(p_{i'j}) - \left(\text{mean} \log(p_{ij}) - \text{mean} \log(p_{i'j}) \right) \right)^2$$

- Meaningful only when all $p_{ij} > 0$

β-diversity

- Personally, if you made me make a β-diversity parameter, I'd choose Aitchison distance

  - easier to estimate than others

$$\text{clr}\left(W_{ij}\right) = \log W_{ij} - \frac{1}{J}\sum_{j'=1}^{J} \log W_{ij'}$$

$$\hat{\beta}_{\text{Aitchison}}(i, i') = \sum_{j=1}^{J} \left(\text{clr}(W_{ij}) - \text{clr}(W_{i'j})\right)^2$$

What we've been talking about

- microbial outcome $\sim$ information about environment's type
- species richness $\sim$ information about environment's type
- abundance of taxon j $\sim$ information about environment's type

Prediction

- In contrast, there is a different set of questions that look like

environment type ~ information about microbiome

- e.g., “predict cancer”, “predict diarrhea”, “predict itchiness”

- A totally different question!

Prediction

- If I were to do this…
  - I'd tranform my data to clr($W_{i.}$)'s
  - Fitting a model:
    - Lots of observations: Random forest? Boosted tree? Neural net? Large language model?
      - Challenges: Batches!! Include sample labels as predictors?
    - Fewer observations: L_1 regularised regression?

I'd just look at Willis & Clausen (2022), Evaluating replicability in microbiome data 🤩 and just do that again

Amy’s wish list

- You choose a meaningful parameter to estimate
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

Statistical miscellanea

Diversity, ordination, prediction…

Statistical Diversity Lab @ University of Washington
Amy Willis — @AmyDWillis — Associate Professor
Shirley Mathur — PhD Candidate
Sarah Teichman — PhD Candidate
María Valdez — PhD Candidate