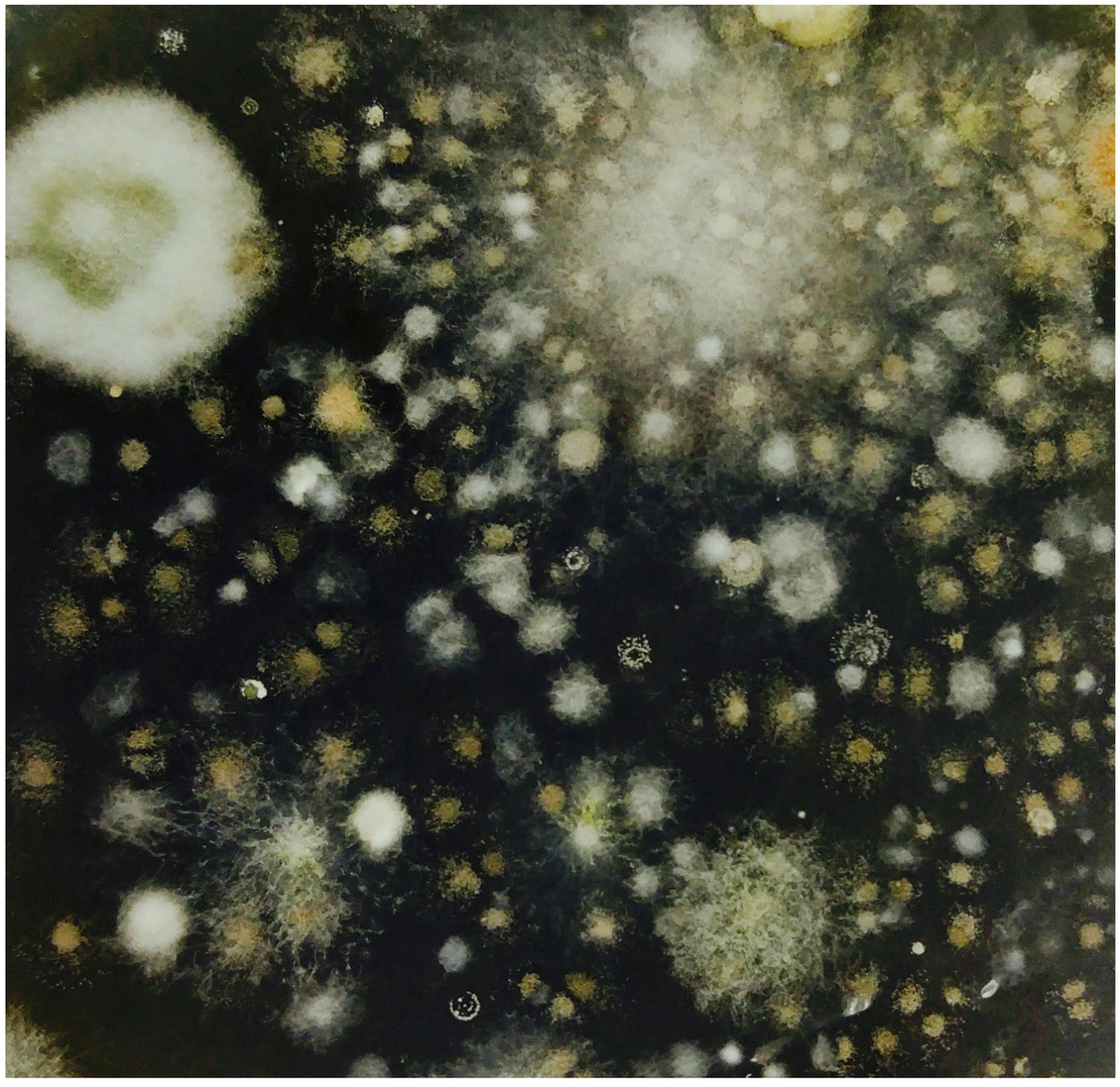
Microbial diversity:

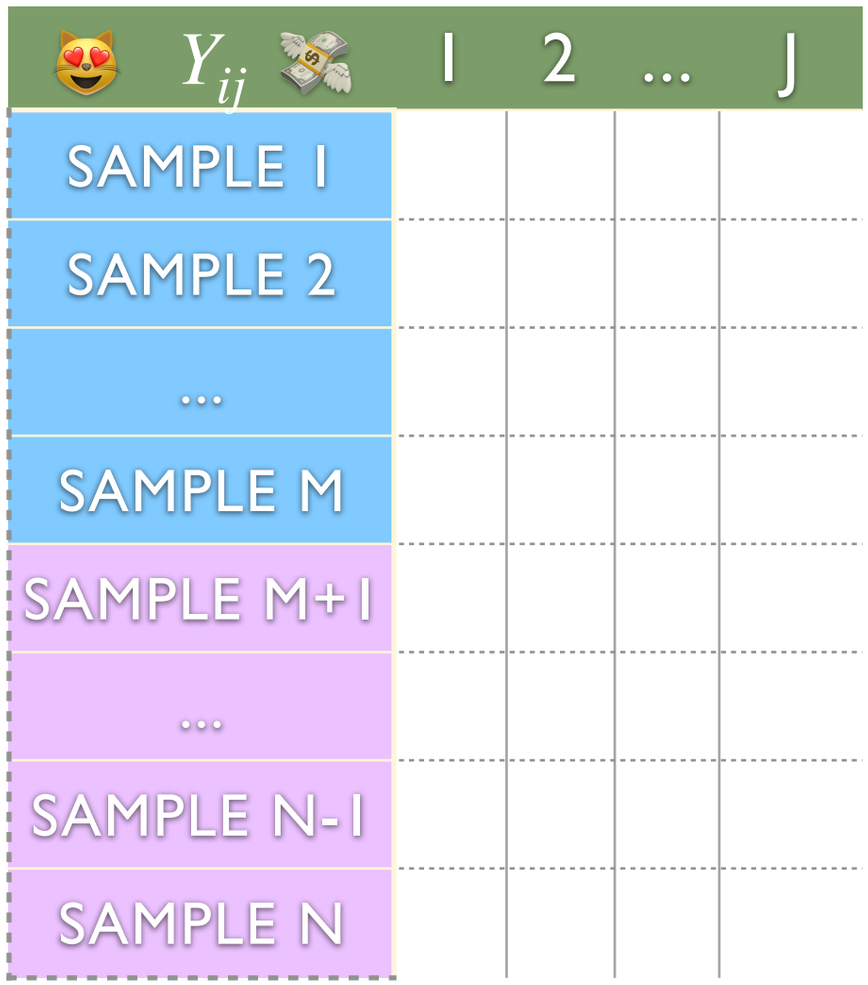
Estimation & comparison



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:
  - Which taxonomic level? (strain/species/genus...)
  - Which diversity parameter?
  - Which *estimator*?

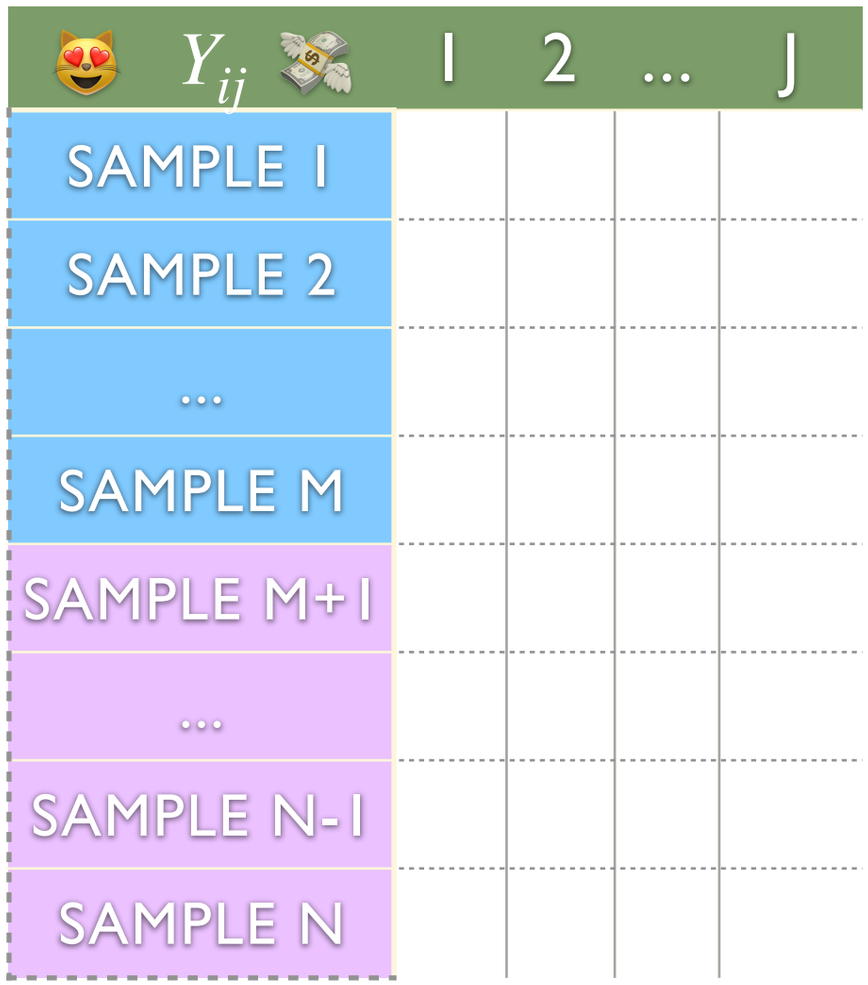
Diversity & parameters

- There are multiple choices to make:
  - Which taxonomic level? (strain/species/genus...)
  - **Which diversity parameter?**
  - Which *estimator*?

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

α-diversity

Some other of α-diversity parameters include

1. Average species richness in group 1
minus
Average species richness in group 2

2. Average species richness in group 1
divided by
Average species richness in group 2

3. Average Shannon diversity when temp is (t+1) and from site s
minus
Average Shannon diversity when temp is t and from site s

4. …

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

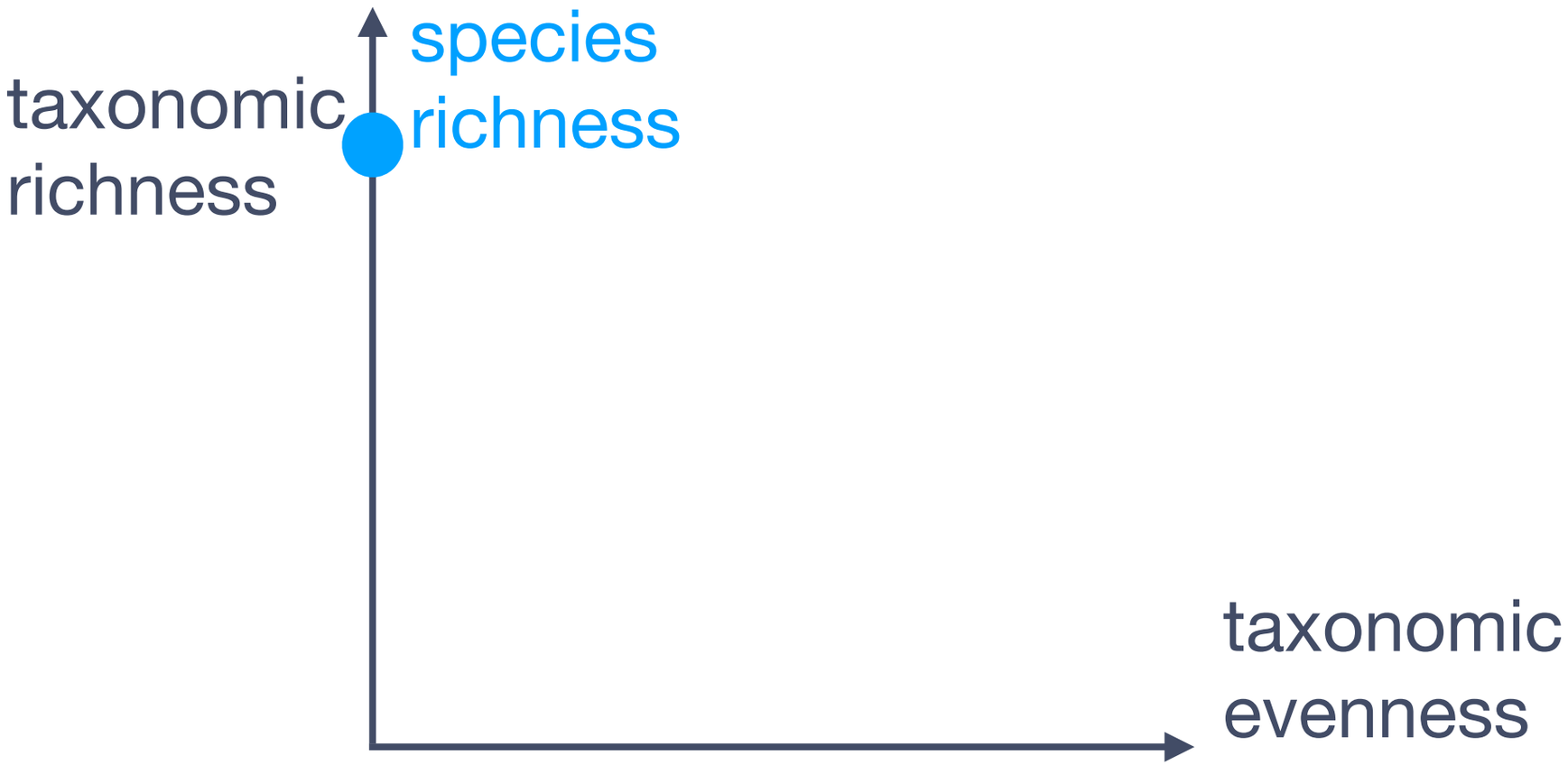
- My wish list remains unchanged
- **You choose a meaningful parameter to estimate**
- You choose a sensible way to estimate the parameter
- You choose tests that control Type 1 error

What α-diversity parameter? You decide

- Think: What community *quality* do you want to highlight?

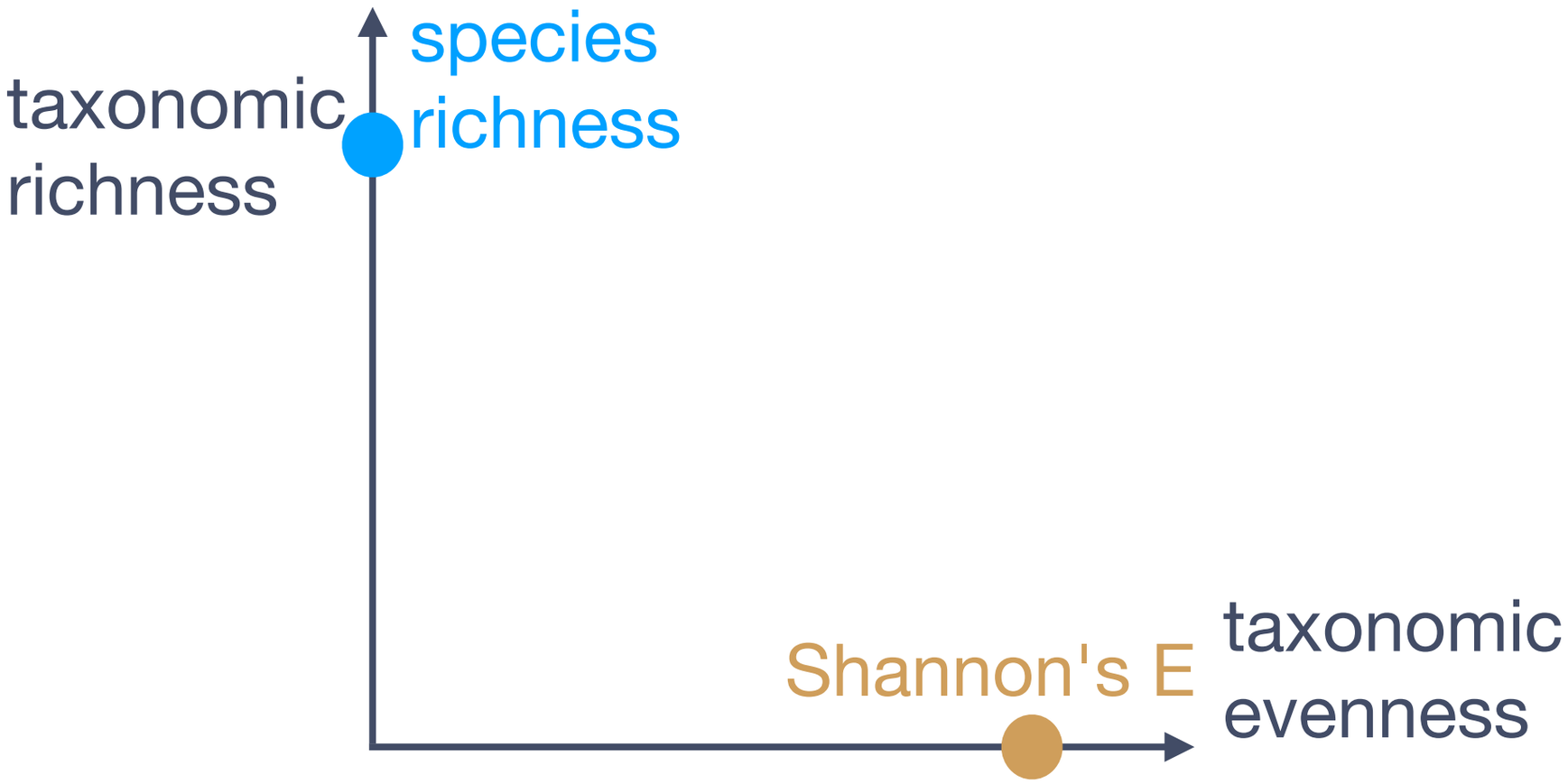
A heuristic diagram I find helpful!

What α-diversity parameter? You decide



A heuristic diagram I find helpful!

What α-diversity parameter? You decide



A heuristic diagram I find helpful!

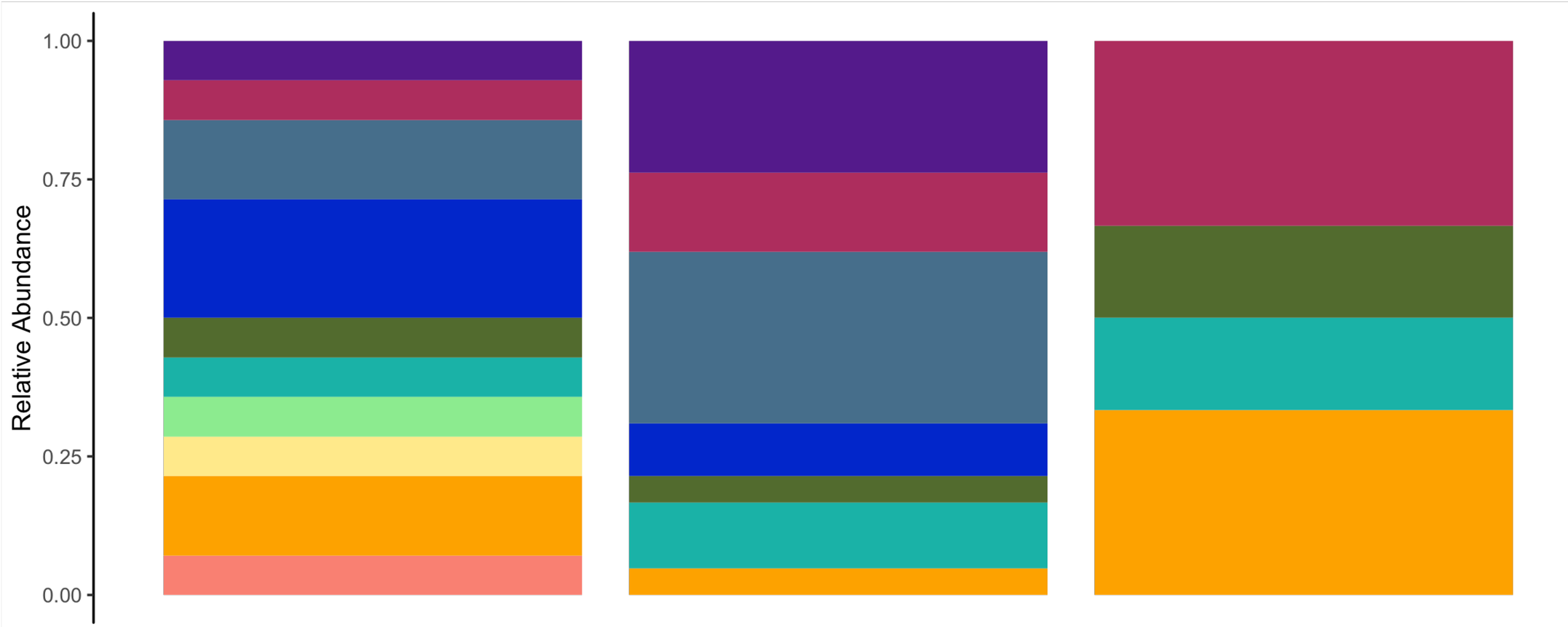
What α-diversity parameter? You decide

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
Shannon's 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

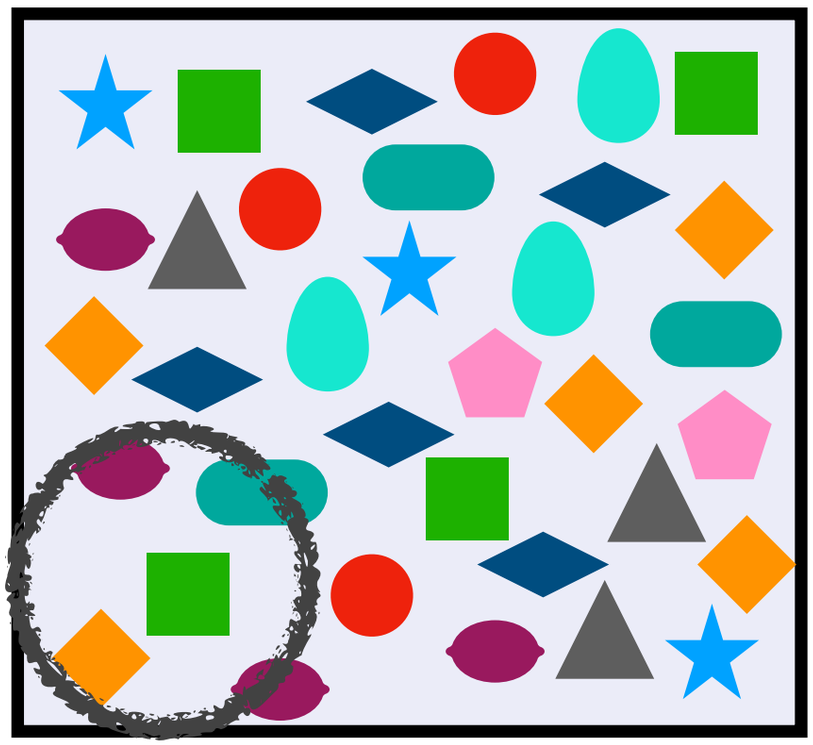
Estimate the richness with: $c_i = \{\#j : \hat{p}_{ij} > 0\}$

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



α-diversity estimators…

- … are like any other estimators!
- We want estimators to be
  - Accurate = correct on average = unbiased / consistent
  - Precise = usually close to their average = low variance

Species richness:

$$C_i = \#\{j : p_{ij} > 0\}$$

- Estimating species richness: core ideas
  - 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,

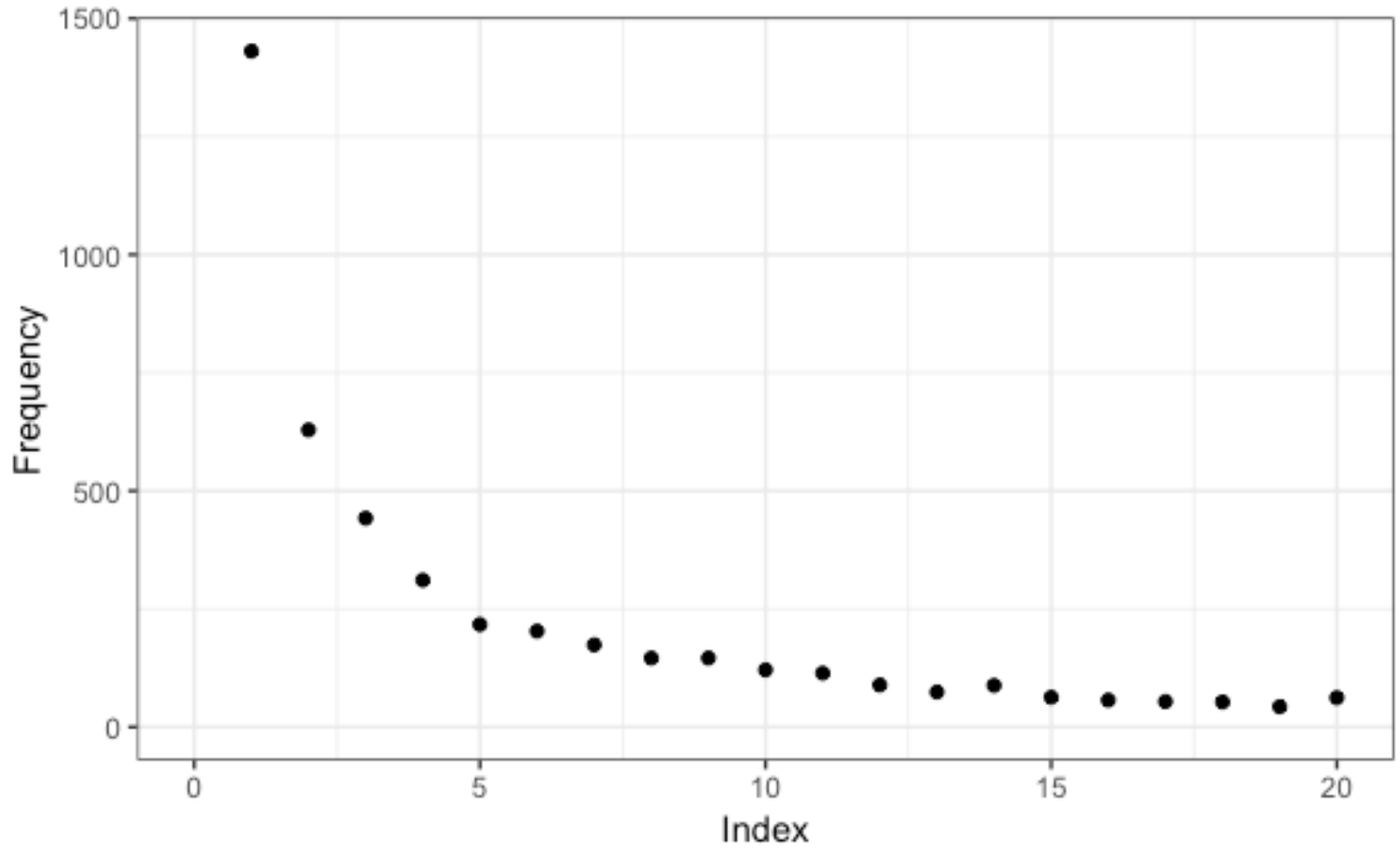
- f_2 = doubletons, …

- 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

- Way less good 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 *estimator* of species richness, and it's based on the assumption that

all species have the same abundance

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

Species richness

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

α-diversity: Species richness

- What I recommend:
    - Estimate species richness for each sample along with their standard deviations: $\hat{C}_i$ and std error $\left(\hat{C}_i\right)$
    - Estimate average differences in true richness across populations using heteroskedastic linear regression
    - Do not rarefy
- Do this only if you actually care about species richness. Resist peer pressure.

α-diversity: non-richness

- What I recommend:
    - Estimate Shannon/Simpson/etc for each sample along with their standard deviations: $\hat{C}_i$ and std error $\left(\hat{C}_i\right)$
    - Estimate average differences in true Shannon/Simpson/etc across populations using heteroskedastic linear regression
    - Do not rarefy
- Do this only if you actually care about α-diversity. Resist peer pressure.

α-diversity: non-richness

- An alternative <u>that I think is fine</u>
    - Estimate Shannon/Simpson/etc for each sample using **plug-in estimators**
    - Estimate differences in average true Shannon/Simpson/etc across populations **studied with the same sequencing effort**
        - Do this by including sequencing depth as an adjustment variable in your regression model
    - Do not rarefy
- Do this only if you actually care about α-diversity. Resist peer pressure.

α-diversity: Species richness

- What I recommend:
  - Estimate species richness for each sample along with their standard deviations: $\hat{C}_i$ and std error $\left(\hat{C}_i\right)$
    - Tool example: breakaway::breakaway()
  - Estimate average differences in true richness across populations using heteroskedastic linear regression
    - Tool: breakaway::betta()
  - Do not rarefy
- Do this only if you actually care about species richness. Resist peer pressure.

α-diversity: non-richness

- What I recommend:
  - Estimate Shannon/Simpson/etc for each sample along with their standard deviations: $\hat{C}_i$ and std error $\left(\hat{C}_i\right)$
    - Tool: DivNet:: divnet(… X = diag(nsamples(phy_obj)), …)
  - Estimate average differences in true Shannon/Simpson/etc across populations using heteroskedastic linear regression
    - Tool: breakaway::betta()
  - Do not rarefy
- Do this only if you actually care about α-diversity. Resist peer pressure.

α-diversity: non-richness

- An alternative <u>that I think is fine</u>
  - Estimate Shannon/Simpson/etc for each sample using plug-in estimators
    - Tools: breakaway::sample_simpson(), breakaway::sample_shannon(), breakaway::sample_shannon_e(), vegan::diversity()…
  - Estimate differences in average true Shannon/Simpson/etc across populations **studied with the same sequencing effort**
    - Tool: lm() for additive distances, glm(family = poisson) for fold-differences…
    - Better: rigr::regress("mean") or rigr::regress("rate") or raoBust::glm_test(family=poisson)
    - Do this by including sequencing depth as an adjustment variable in your regression model
    - formula = estimated_diversity ~ predictor1 + predictor2 + … + seq_depth
- Do not rarefy

α-diversity: non-richness

- An alternative that I recommend against but is very common
    - Rarefy: Randomly discard sequences from more deeply sequenced samples until all samples have the same number of subsampled reads
    - Estimate Shannon/Simpson/etc for each sample using plug-in estimators
    - Run a Wilcoxon/Kruskal-Wallis/Spearman correlation/ Kendall's τ and report a p-value

Bias and diversity

- Justification for rarefaction not mine
  - We discover more diversity with more sequencing
  - So we can’t directly compare samples with different depths
  - Solve this “problem” by throwing away reads until all samples have same depth
  - Then do whatever comparison we want

Bias and diversity

- My perspective
    - Rarefaction deliberately induces bias
    - Analyzing rarefied data cannot target a meaningful parameter
        - At best, knowingly chooses a very, very bad estimator…
- You *should not* do inference when there is no meaningful parameter

Bias and diversity

- My perspective
  - Rarefaction deliberately induces bias
  - Analyzing rarefied data cannot target a meaningful parameter
    - At best, knowingly chooses a very, very bad estimator…
- You *should not* do inference when there is no meaningful parameter
- We can still be friends if you have rarefied, or will rarefy

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 2025 Amy is more excited about differential abundance of taxa or differential presence of genes/pathways or evolutionary pressures or genomic plasticity or genome architecture or codon usage streamlining or translational mechanisms or…

Questions about α-diversity…

…before we move onto β-diversity

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

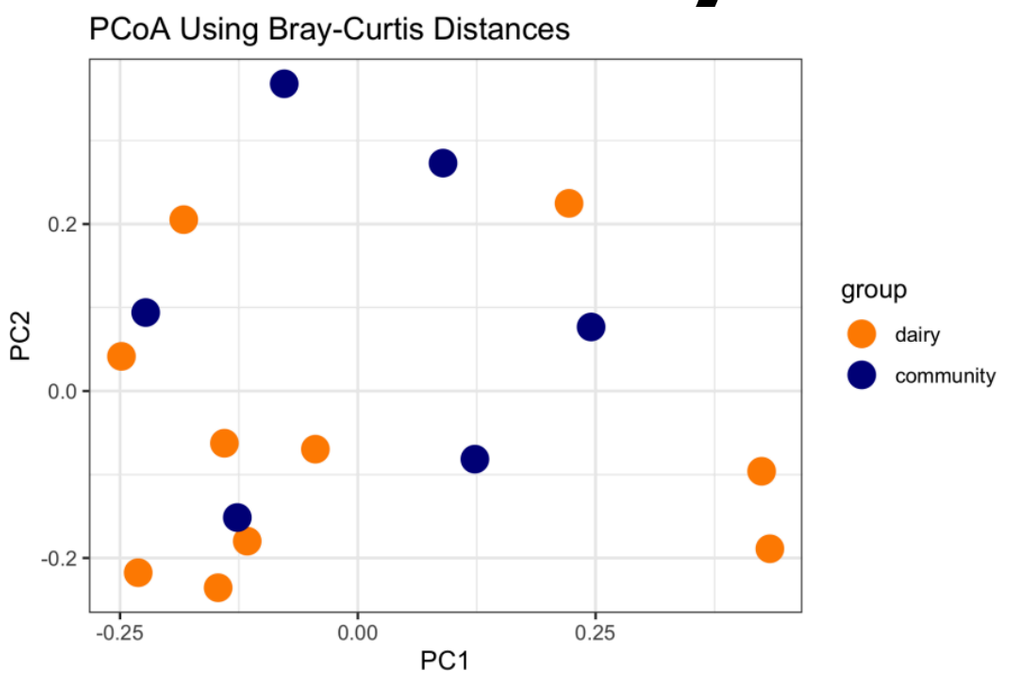
β-diversity

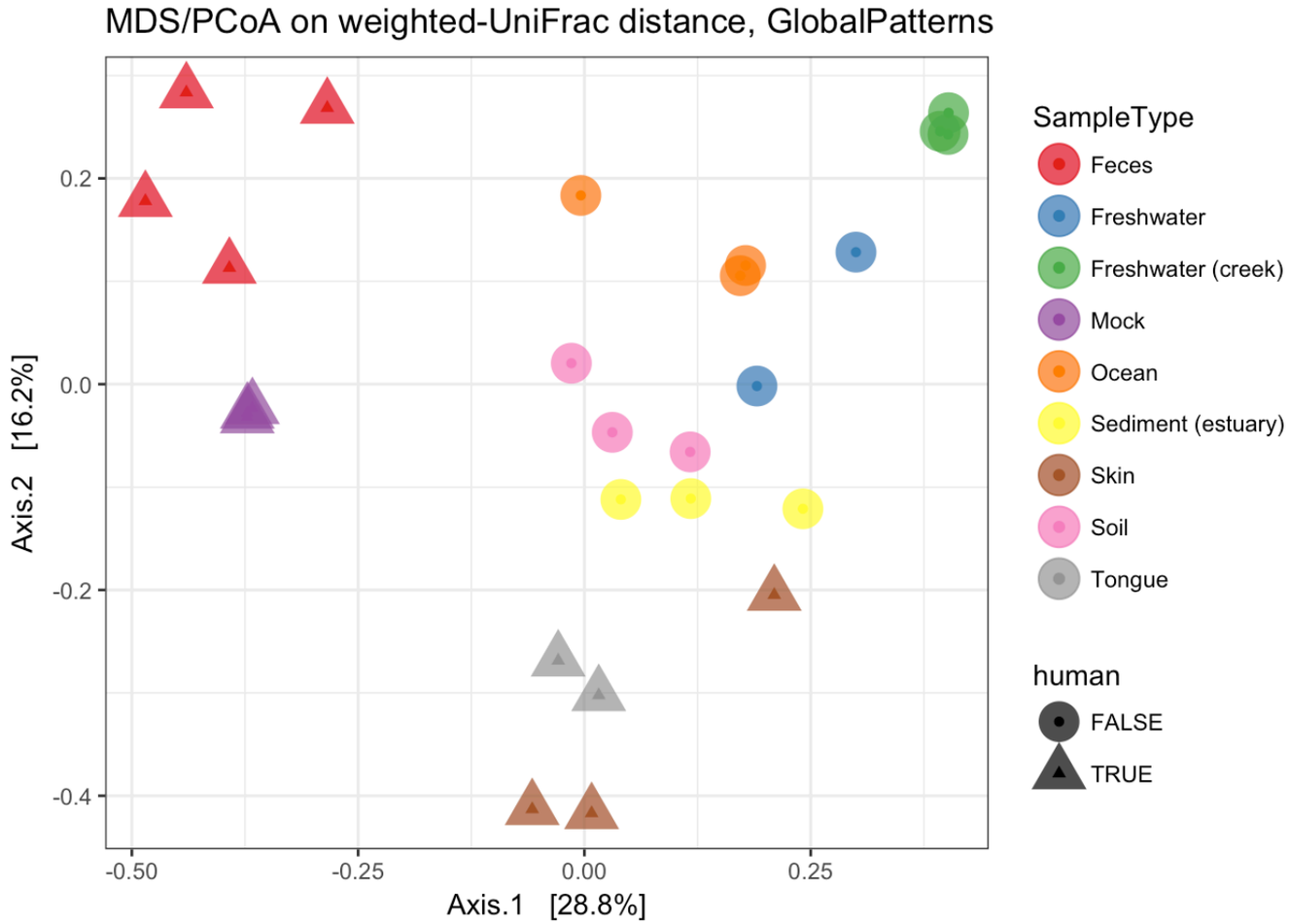
Distances	Sample 1	Sample 2	…	Sample n
Sample 1	0	d-hat(1,2)	…	d-hat(1,n)
Sample 2	d-hat(1,2)	0	…	d-hat(2,n)
…	…	…	…	…
Sample n	d-hat(1,n)	d-hat(2,n)	…	0

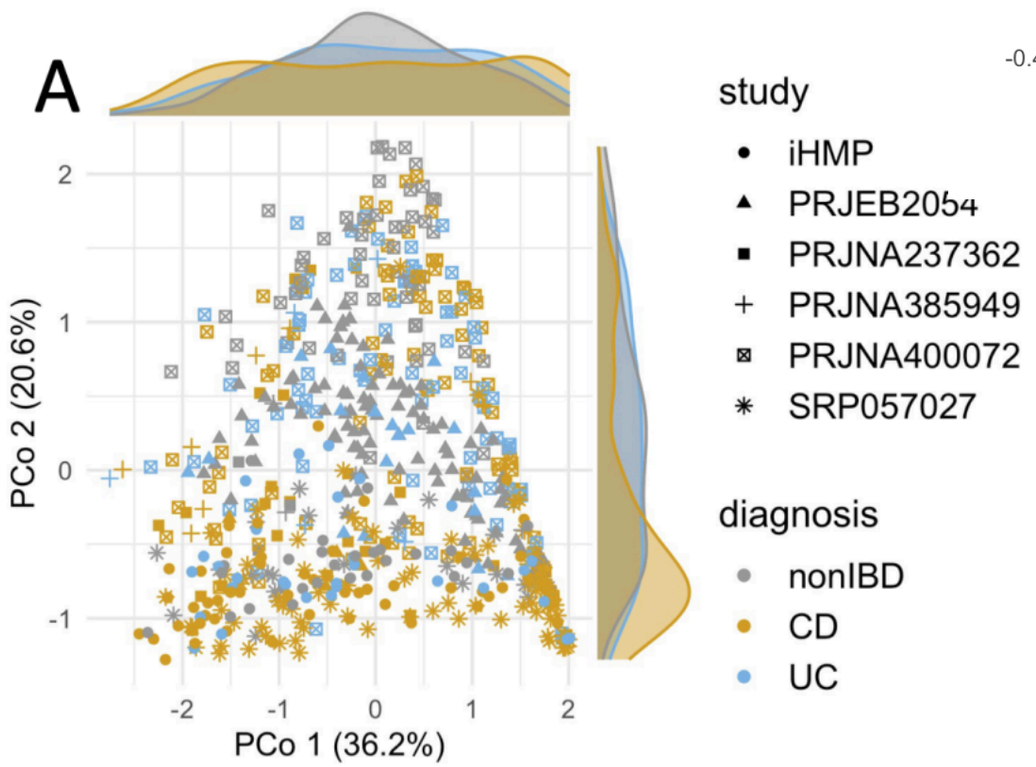
β-diversity

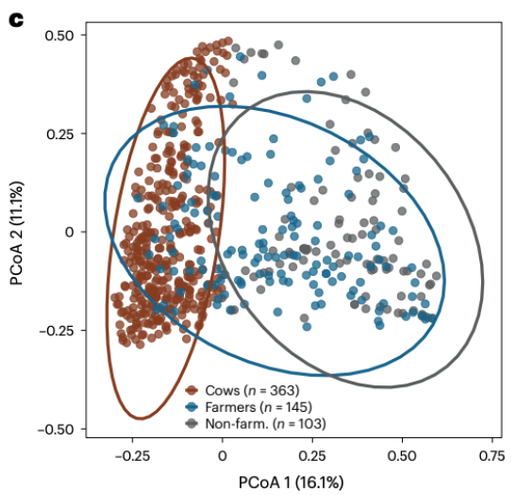
- Challenge: These $n \times n$ tables are hard to learn from
- Approach: Make it easier?
- Put n points on a scatterplot, one per sample
- PCoA/MDS: “Find the best arrangement for the points such that $\text{distance}_{\text{scatterplot}}(i, i')$ is as close to $\hat{\beta}(i, i')$ for all pairs i,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

$$= \theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip}$$

and tests $\theta_1 = 0$ or $\theta_1 = \ldots = \theta_p = 0$

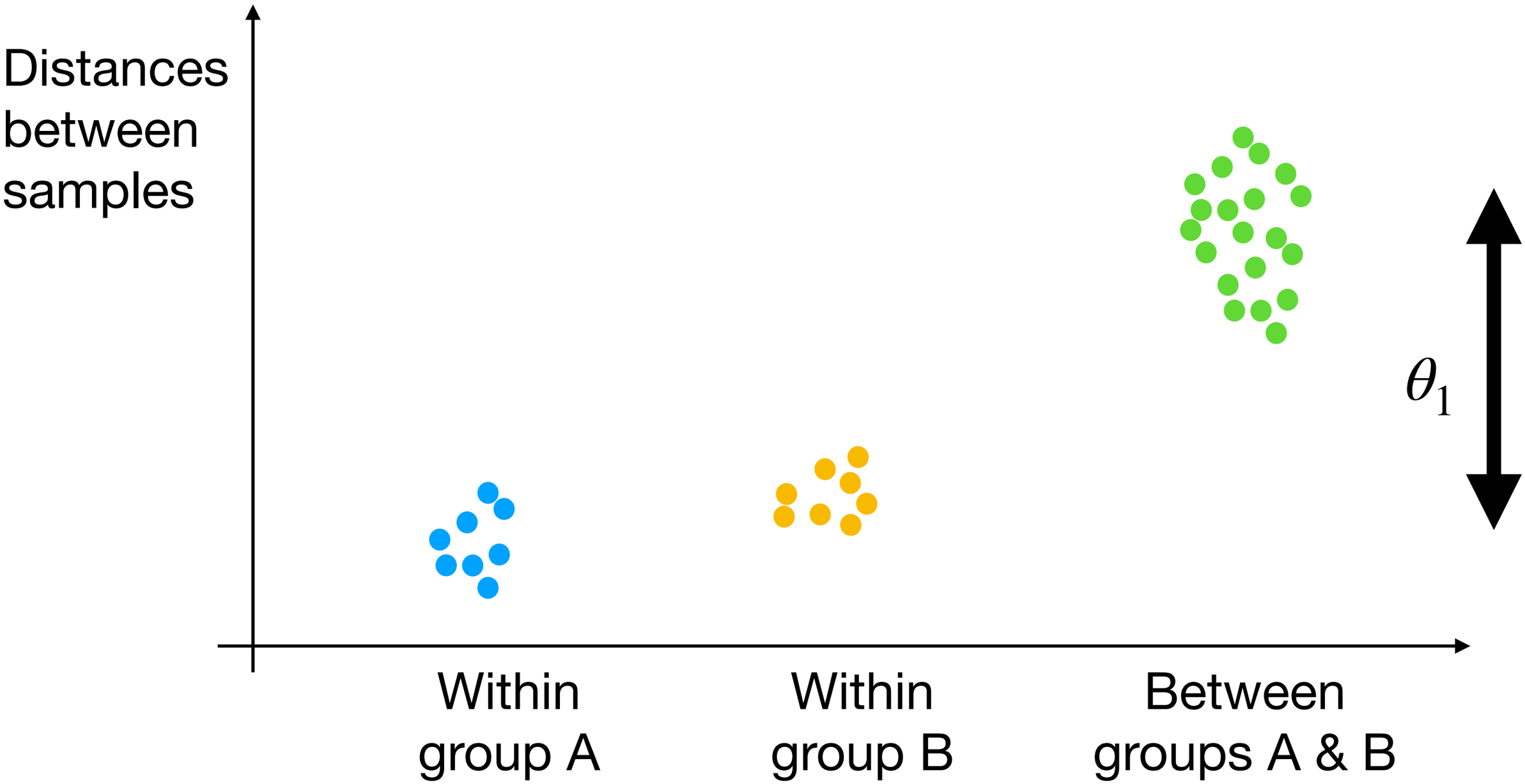
β-diversity

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

$$\text{average distance}(i, i') = \theta_0 + \theta_1 \mathbf{1}_{\{i,i' \text{ from different groups}\}}$$

- θ_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

- If you are just plotting your data, **plot whatever you want**

- There is a culture of throwing p-values on β-diversity

  - Knee-jerk p-values make me sad 😿💔

β-diversity

- If you made me make a β-diversity parameter against my will I'd choose Aitchison distance

    - The most generous interpretation of underpinning parameter is identifiable. Many others are not.

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

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

My current position

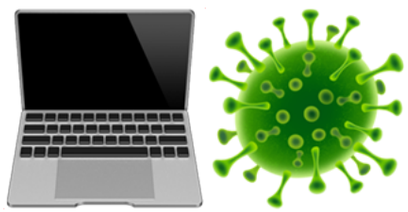
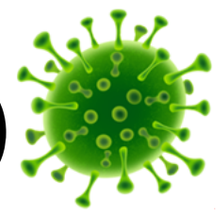
- Analyzing α- and β-diversity is *fine*
    - Sensible background, but generally *not very interesting*
- Real estate in your paper is *very valuable*
    - α- and β-diversity ➡️ Supplementary text & figures
    - Your most exciting discoveries ➡️ Main Text & Figures 🤩😻
- I am not going to die on this hill ☠️🐺

Amy’s wish list

- **You choose a meaningful parameter to estimate**

- You choose a sensible way to estimate the parameter

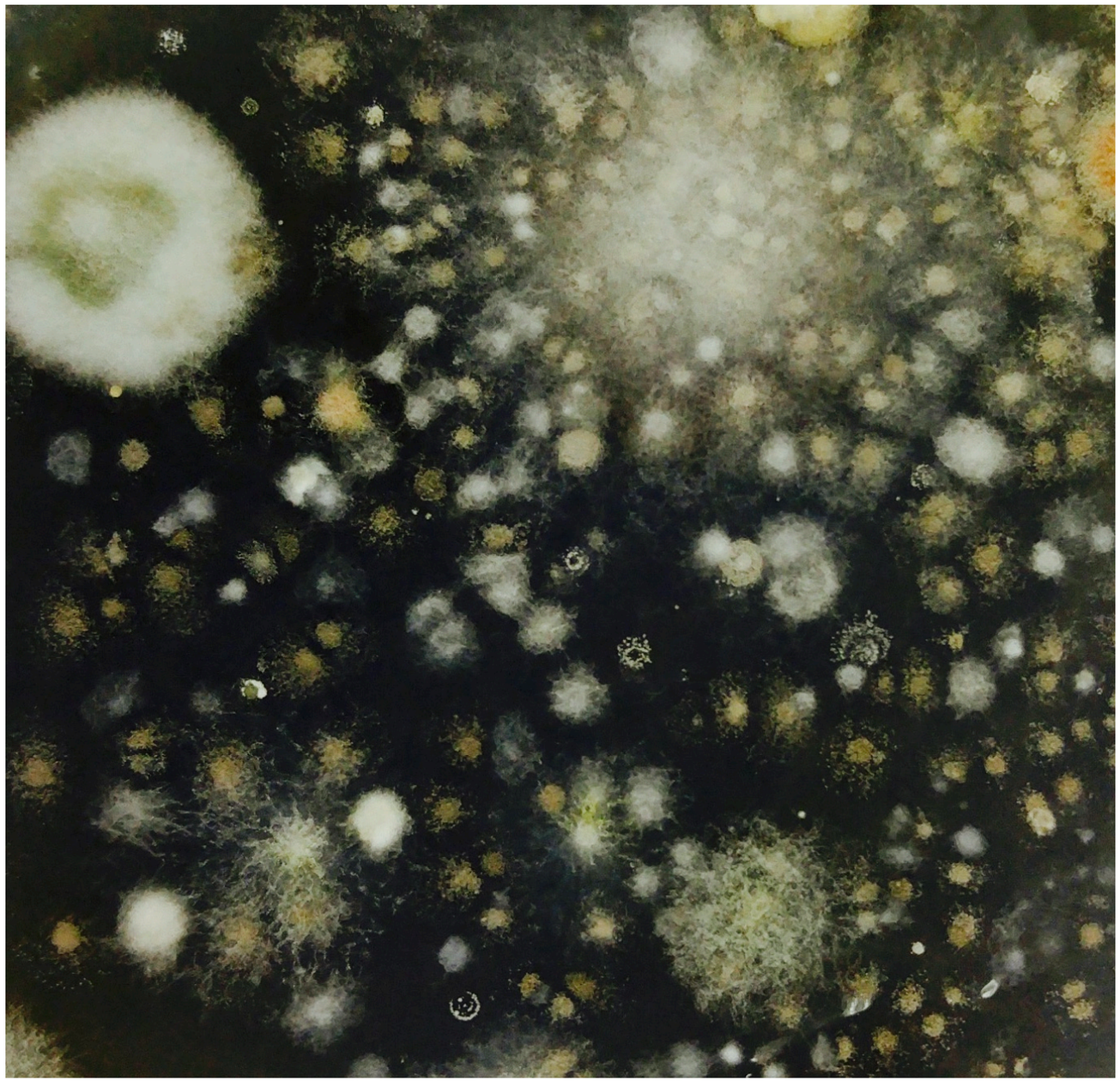
- You choose tests that control Type 1 error

 Diversity lab  

1. Go to schedule on wiki, click on “Statistics labs”
2. Copy the command under “diversity lab”
3. Run the copied command in your R Studio Server console
4. Open the downloaded file diversity-lab.Rmd and work through the code and exercises

Microbial diversity:

Estimation & comparison



Tools for testing null hypotheses that are false

Supplementary crap I moved out or around

Other challenges with estimating diversity

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

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

?

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 $p_{ij} > 0$