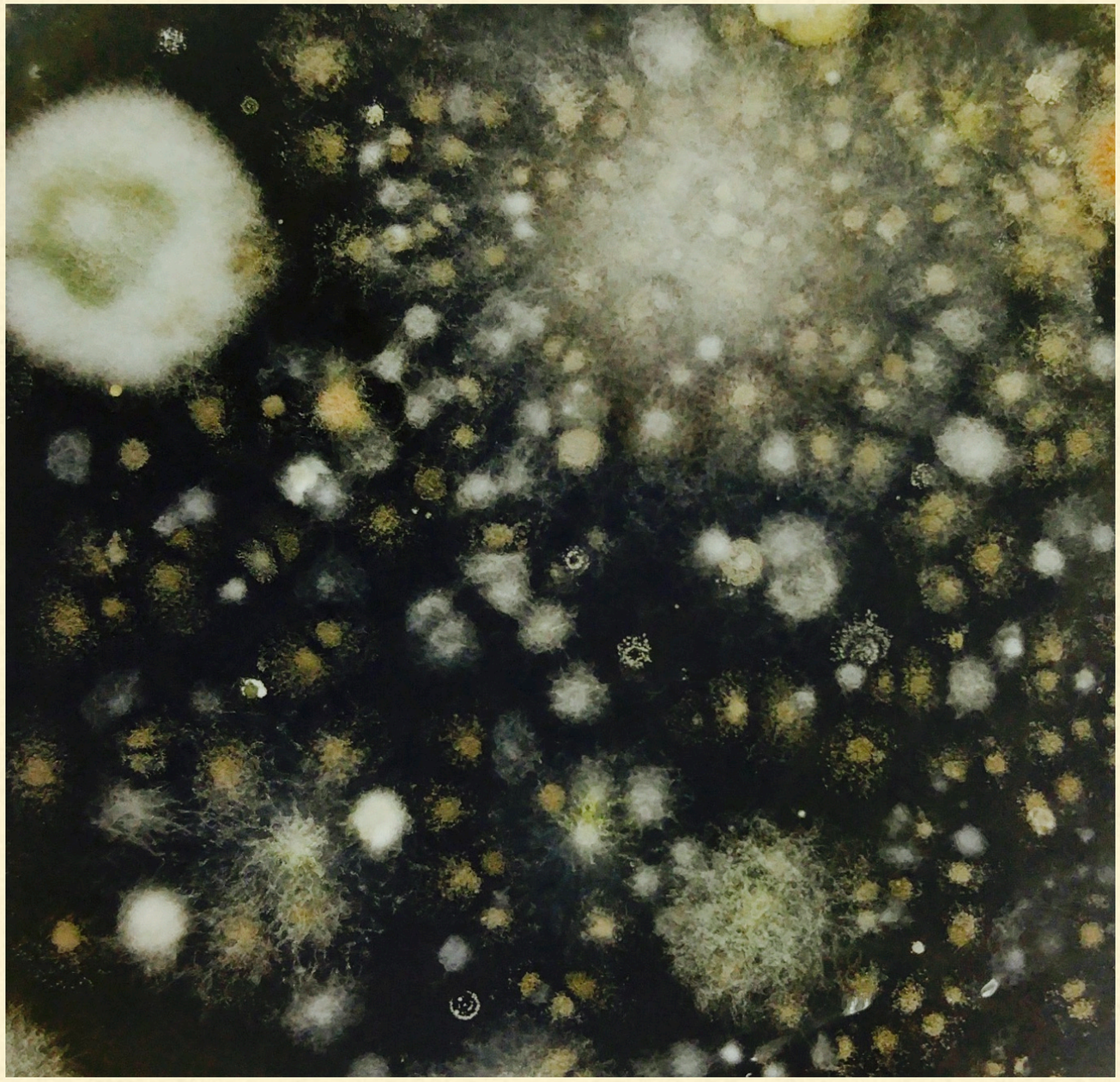
MODELING DIVERSITY



DIVERSITY

- Low dimensional summaries of entire communities
  - α-diversity: one community
    - e.g., species richness, Shannon diversity
  - β-diversity: multiple communities
    - e.g., UniFrac, Bray-Curtis, Jaccard
    - Usually based on distances
- Direct import from macroecology

DIVERSITY & PARAMETERS

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

ALPHA DIVERSITY

- Suppose we have C groups in our environment in proportions p_1, p_2, …, p_C

- Any function of
  - p_1, p_2, …, p_C OR
  - p_1, p_2, …, p_C and ~~some info about relationships amongst groups~~ phylogeny

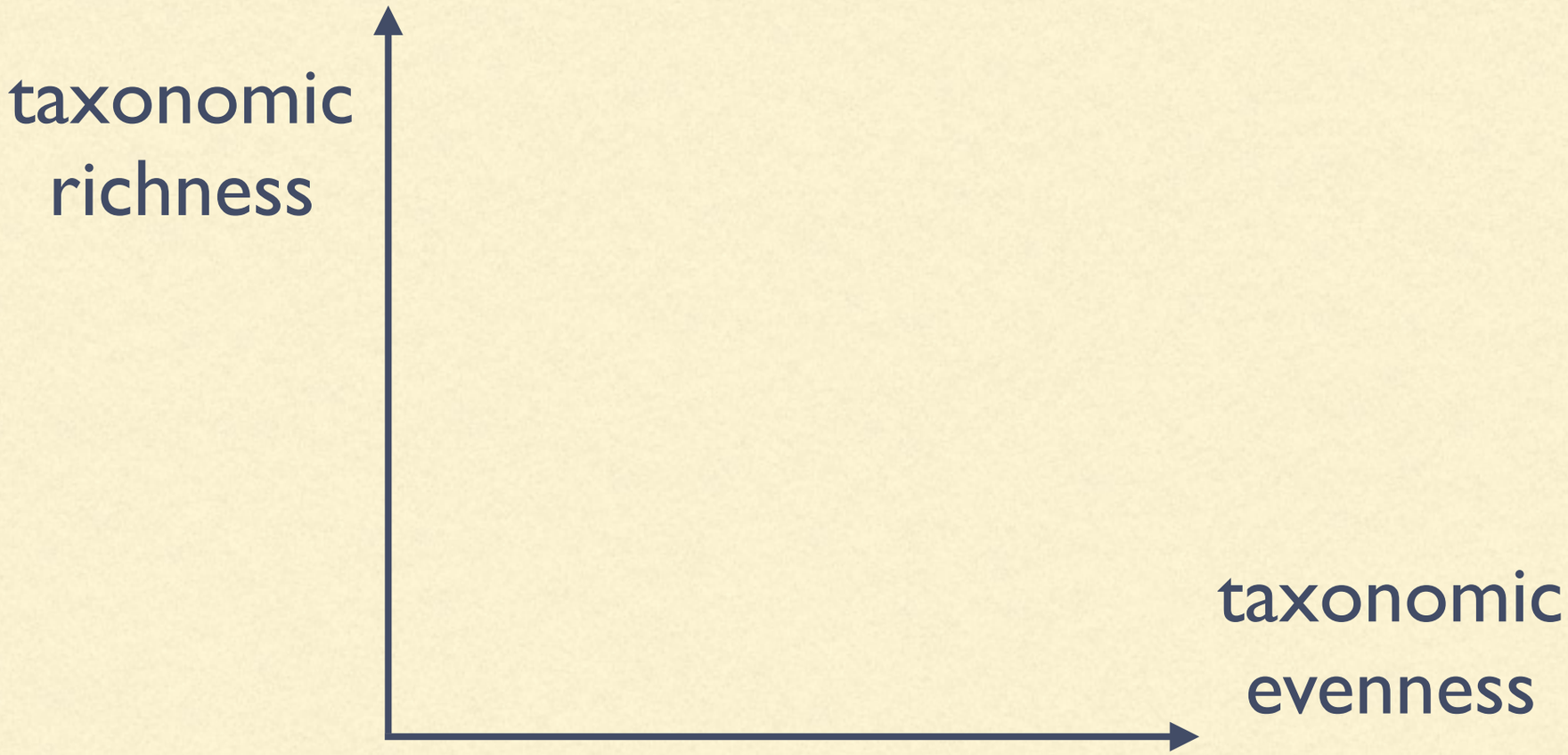
is a valid α-diversity parameter

ALPHA DIVERSITY

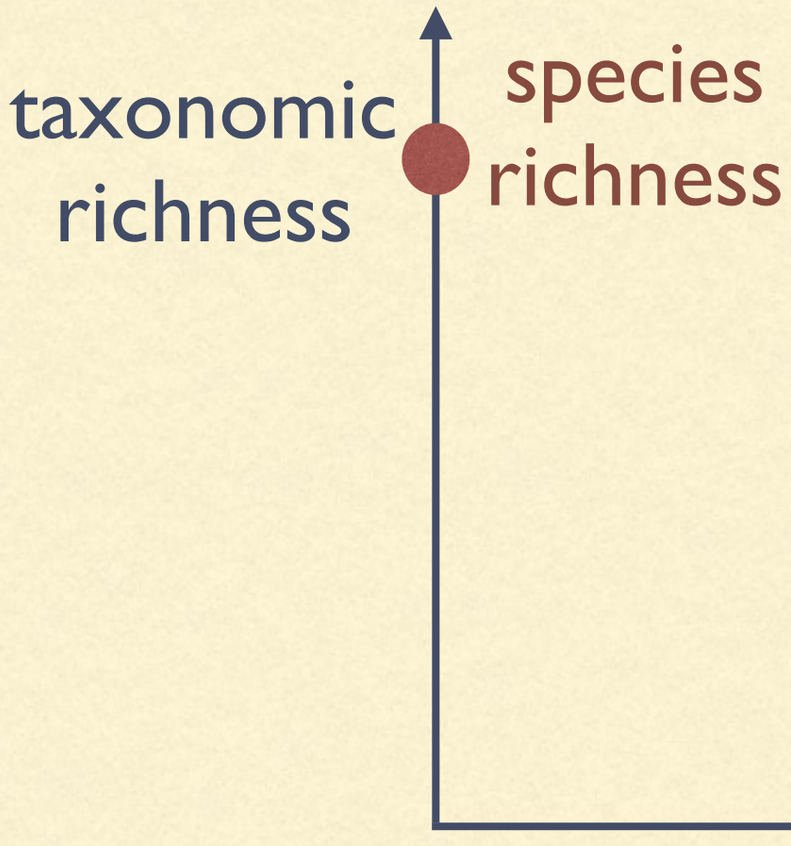
- Some examples of α-diversity measures include
  - Species richness: C
  - Simpson's index: $\sum_{i=1}^{C} p_i^2$
  - Shannon diversity: $-\sum_{i=1}^{C} p_i \ln p_i$
  - Shannon's E: $\frac{-\sum_{i=1}^{C} p_i \ln p_i}{\ln C}$

YOUR CHOICE

- Think: What difference do you want to highlight?



YOUR CHOICE



YOUR CHOICE

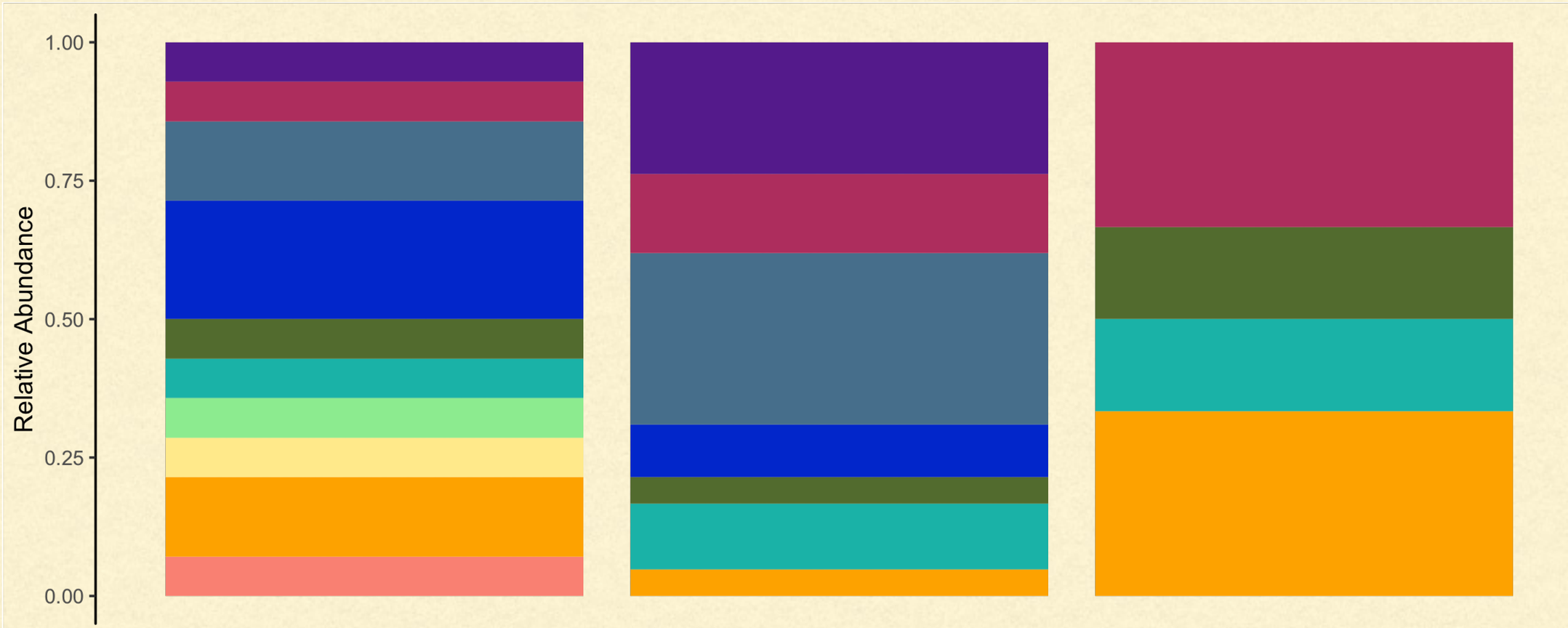

taxonomic
richness
species
richness
Shannon's E
taxonomic
evenness

YOUR CHOICE


taxonomic
richness
species
richness
Shannon
taxonomic
evenness
Shannon's E

YOUR CHOICE

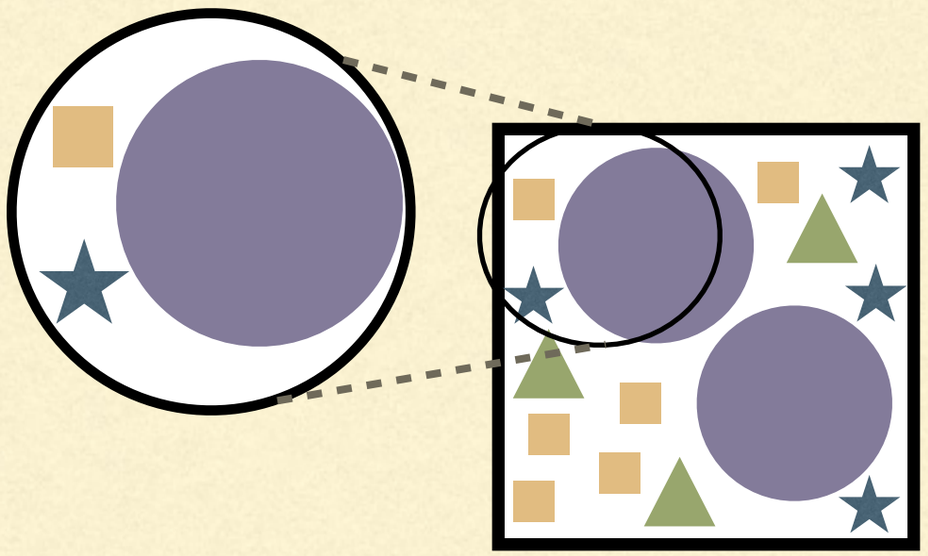
This is a question of *parameter choice:*
Which parameter highlights the differences I care about?



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

THE PROBLEM

- In practice, we don't observe the entire community, just a sample from it
  - we don't know C or $p_1, p_2, \ldots, p_C$



- **We need to *estimate* them using the data we collected**

naive

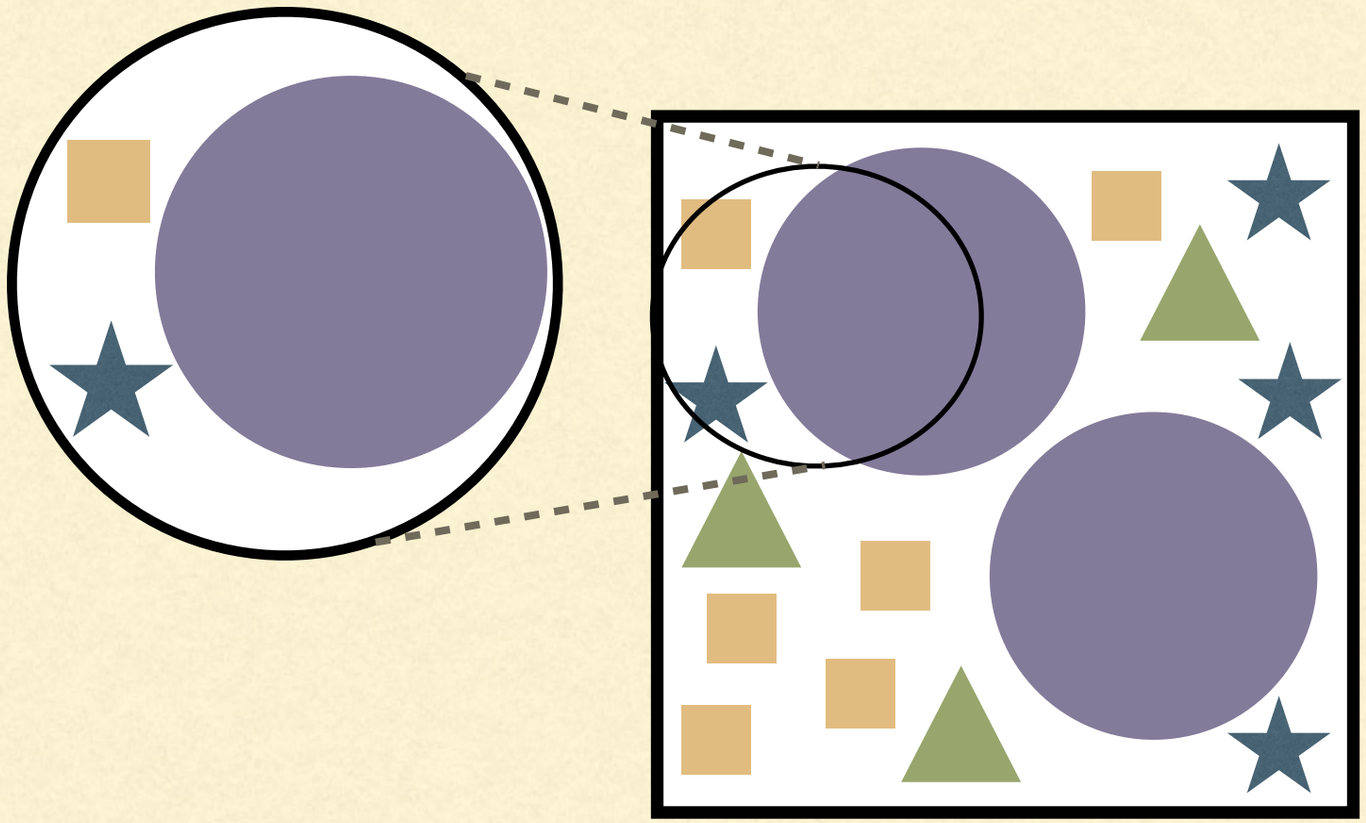
THE "~~CLASSICAL~~" APPROACH

- Substitute the observed abundances $\hat{p}_1, \ldots, \hat{p}_c$ for the unknown, true abundances $p_1, p_2, \ldots, p_c$ and pretend nothing happened
  - e.g. Estimate the richness with: $c = \#\{i : \hat{p}_i \neq 0\}$
  - e.g. Estimate the Simpsons index: $\sum_{i=1}^{c} \hat{p}_i^2$

ONE PROBLEM (OF MANY)

- Species richness: plug-in estimate *underestimates*
- Simpson: estimate *overestimates*

- ~~Need new indices~~
- Need new estimators



HOW TO FIX

- Two things are wrong here:
  - bias (under/overestimation)
  - variance (how big are the error bars — you'll never be exactly right)

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

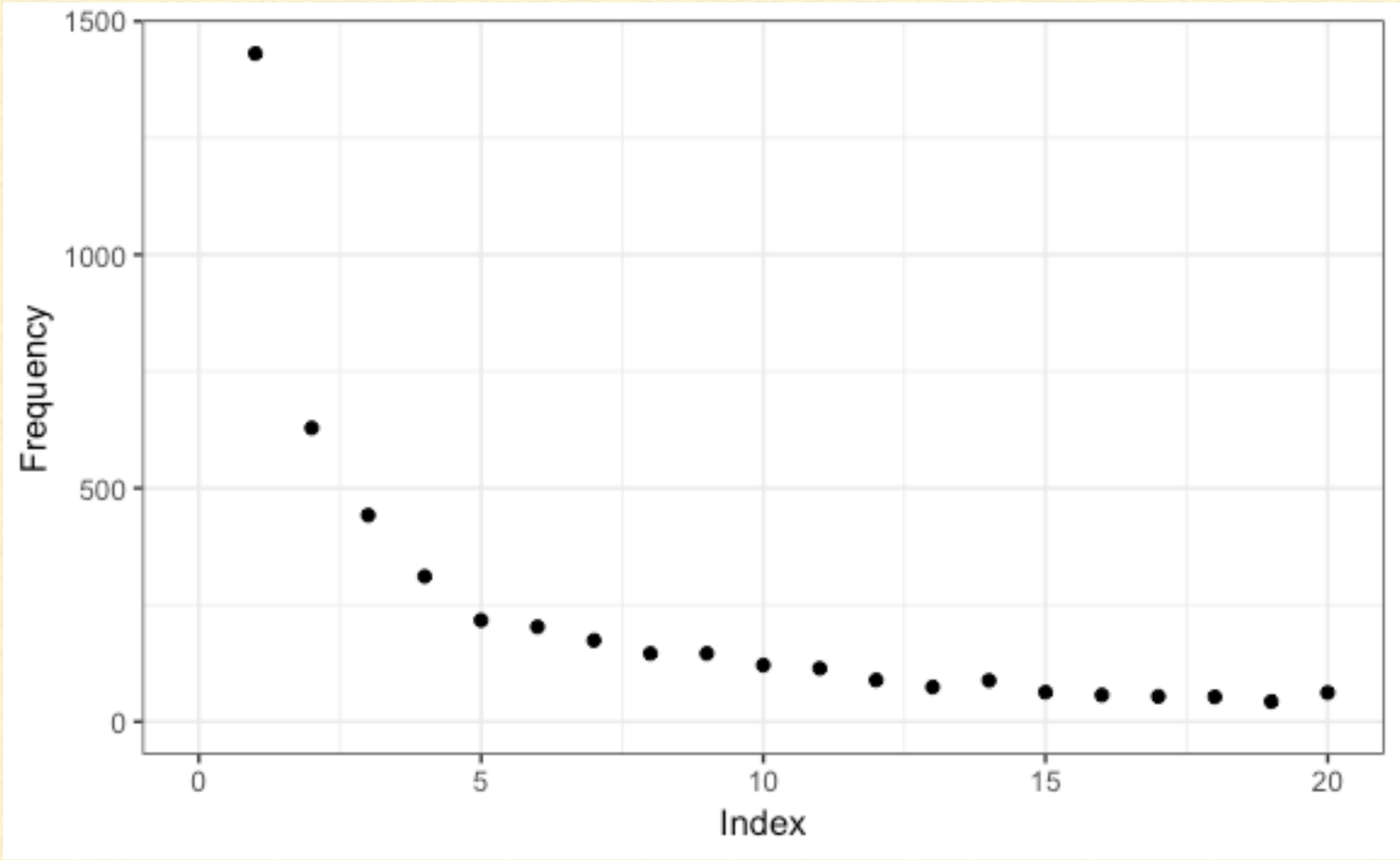
SPECIES RICHNESS ESTIMATION

- The necessary data for richness is the **frequency counts**
- f_j = number of species observed j 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 $f_1, f_2, f_3 \ldots$ 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

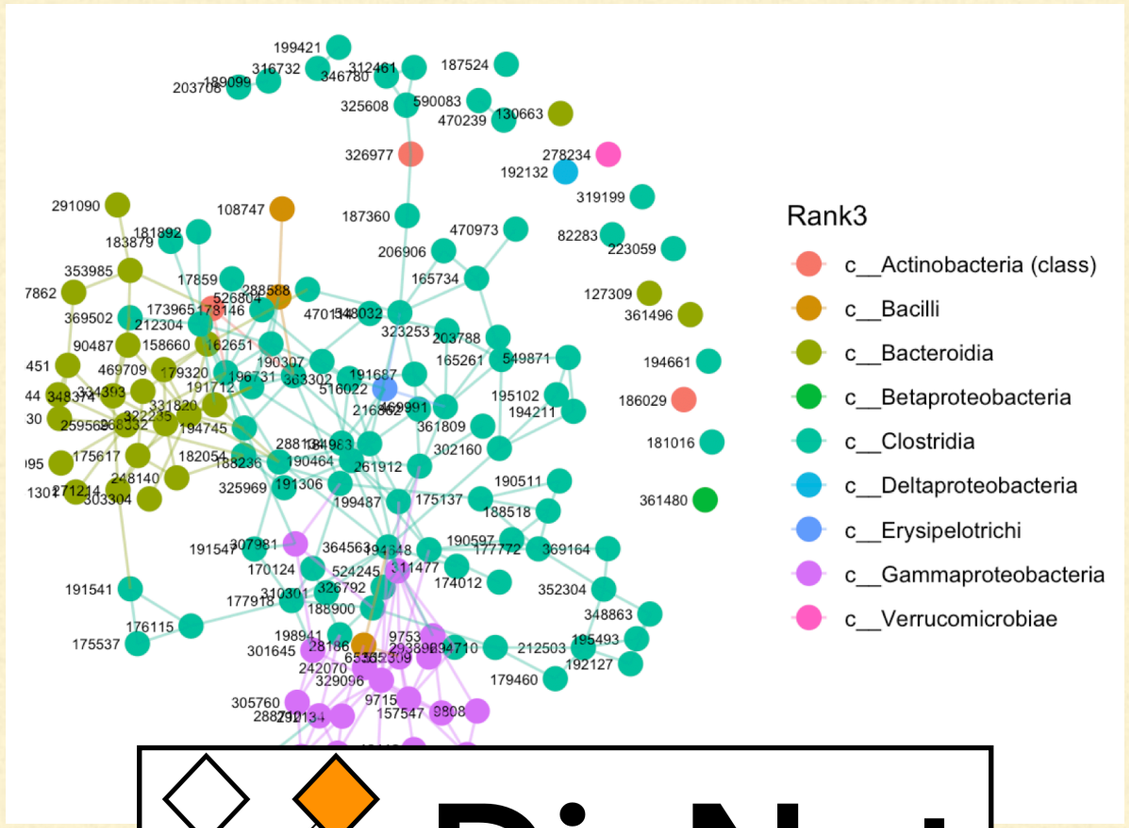
- "Chao1 diversity index" is not an index — it's an estimate of species richness, and it's based on the questionable assumption that

all species have the same abundance

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

ALPHA DIVERSITY: SHANNON & SIMPSON

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



github.com/adw96/DivNet

DivNet

BIAS AND DIVERSITY

- Alternative approach that I loathe: 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: 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: rarefaction

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

epth

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

BIAS AND DIVERSITY

- Alternative approa

PLOS | COMPUTATIO BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Waste Not, Want Not: W

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.

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**
  different seque

BIAS AND DIVERSITY

- Alternative approa

DIVERSITY

- Useful summary of (high-dimensional) compositional data… in many settings!

- A change in diversity: a useful *first question*

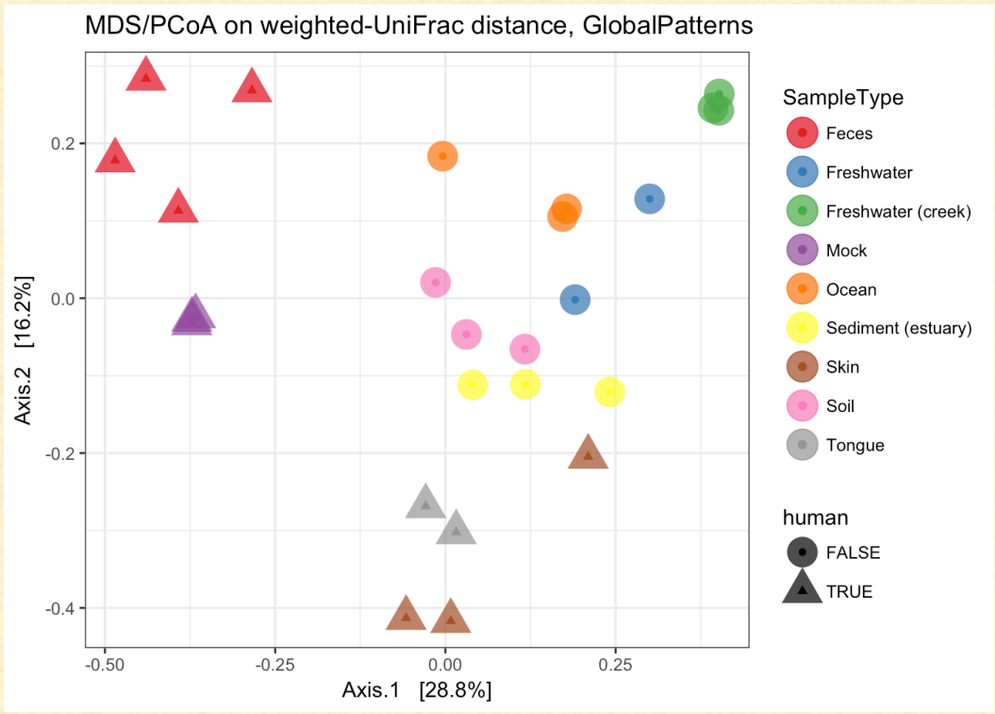
THOUGHTS ON BETA DIVERSITY

BETA DIVERSITY

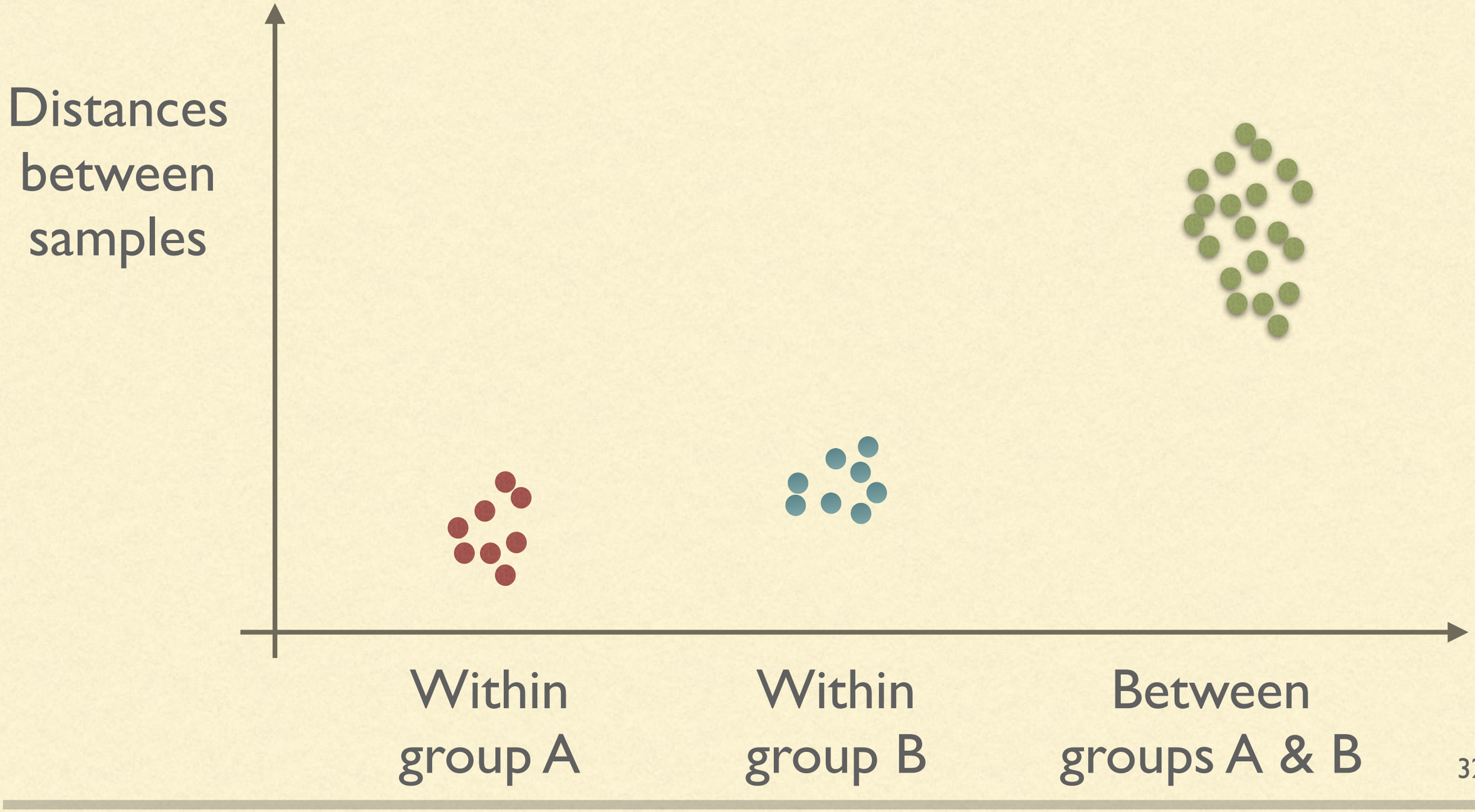
- Community 1: $p_1^{(1)}, p_2^{(1)}, \ldots, p_C^{(1)}$; Community 2: $p_1^{(2)}, p_2^{(2)}, \ldots, p_C^{(2)}$
  - β-diversity parameters are usually distances between compositional vectors
- Bray-Curtis: $\beta_{BC} = 1 - \sum_{i=1}^{C} \min(p_i^{(1)}, p_i^{(2)})$
- Jaccard: $\beta_J =$ % taxa not shared
- UniFrac: Weights phylogeny

DIVERSITY: EXPLORATORY

- Sometimes diversity is analysed as an exploratory tool
  - e.g., ordination
  - *However:* consider not doing an ordination
    - Difficult to interpret (& often over-interpreted)
    - Is there a more informative way to plot your data?



ALTERNATIVE VIZ



A COMPROMISE

- Supplementary Figure 1, not Figure 1?

DIVERSITY: HYPOTHESIS TESTING

- Other times you want to do inference on beta diversities
  - e.g., H_0: dissimilarity within two communities is same as dissimilarity across
  - e.g., H_0: communities A & B have same dissimilarity as communities A & C
- But (yet again): *before running a test, ask "am I sure testing this hypothesis answers a meaningful scientific question?"*

HYPOTHESIS TESTING FOR DIVERSITY

- Common approach: PERMANOVA
- Best solution = ask *"do I really want to do this test?"*
- Better solution = use error bars
  - breakaway::betta(); DivNet::testDiversity
- (Bad solution = rarefy)

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

diversity lab:

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2023/main/labs/diversity-lab/diversity-lab
```

3. Run this command in your RStudio Server console

```
> download.file("https://raw.githubusercontent.com/statdivlab/stamps2023/main/labs
  /diversity-lab/diversity-lab.R", "diversity-lab.R")
```

EXPERIMENTAL DESIGN

The first rule of experimental design is there are no rules.
The second rule of experimental design is that you probably want more data.

~~PRINCIPLES OF EXPERIMENTAL DESIGN~~

- A random smattering of topics in experimental design
  - Randomization
  - Power
  - Observational studies
  - Control data
- Happy to discuss!

POWER CALCULATION

The power calculation to rule them all…

A StatDivLab exclusive…

POWER CALCULATION

$$\#\ \text{samples} = \frac{\text{budget (\$)}}{\text{cost per sample (\$ / sample)}}$$

SHOTGUN METAGENOMICS

- A recent shotgun quote (150 bp PE reads):
  - Extraction: $35 per sample
  - Library prep: $36 per sample
  - Sequencing
    - 20M reads: $250 per sample
    - 80M reads: $730 per sample
    - 100M reads: $904 per sample
  - Annotation: $25 per sample
- ~$10k for 30 samples at 20M reads

POWER

- Recall that failing to reject a null hypothesis (large p-value) may be because
  - the null is true, or
  - the null is false, but because we didn't have enough evidence to reject

POWER

- Power = real
  - Probability of rejecting a false null
    - “Finding a true signal”
- Only meaningful for tests that control Type I error

POWER

- Depends on..
  - True effect size (e.g., β_1)
  - Level of significance α
  - Sample size
    - Number of samples
  - Strong assumptions about the data generating process
    - Depending on setting, possibly also sequencing depth
  - The estimator and test you're using
  - …

POWER CURVES

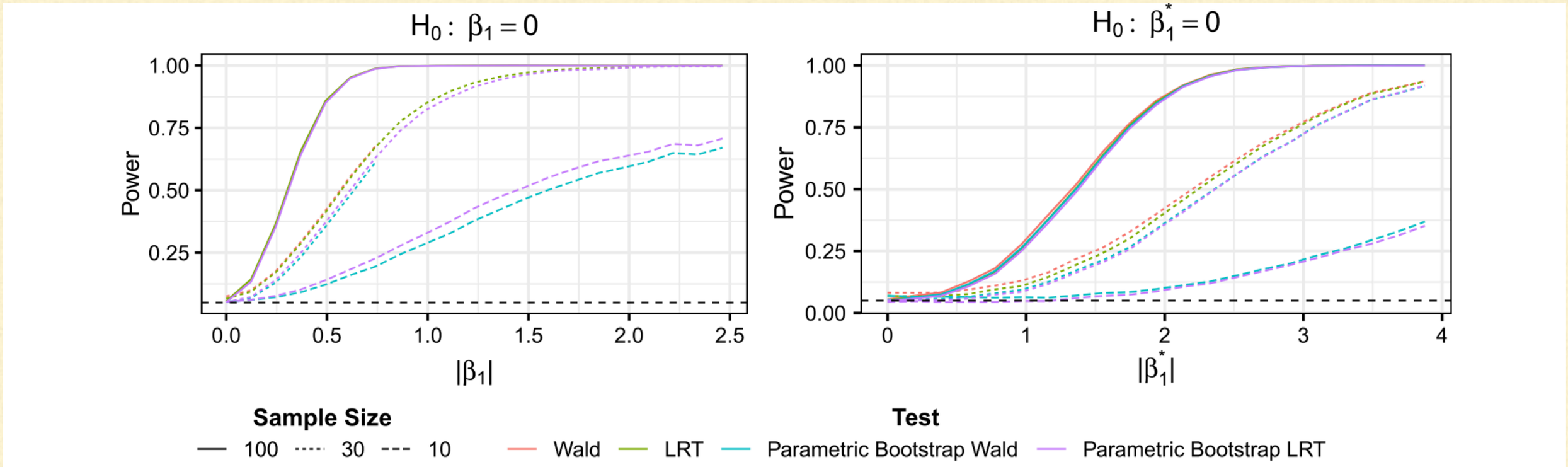


FIG. 3. *Power curves of p-values obtained from the power simulations. Setting* 4 (*left*) *tests* $H_0 : \beta_1 = 0$. *Setting* 5 (*right*) *tests* $H_0 : \beta_1^* = 0$. *A horizontal dashed line is shown at* 0.05. *p-values were obtained using Wald* (*red*), *likelihood ratio* (*green*), *parametric bootstrap Wald* (*blue*) *and parametric bootstrap likelihood ratio* (*purple*) *tests. Sample sizes used were* 10 (*dashed*), 30 (*dotted*) *and* 100 (*solid*).

POWER CALCULATIONS

- Power calculations = fiction
- Often expected in grant applications
- In general, meaningless in ‘omics studies
  - Unless you have a single parameter of interest, power calculations are a furfy*

*A furphy is Australian slang for an erroneous or improbable story that is claimed to be factual.

DESIGNS

- Yesterday predictor of interest
- Ideal: randomize your predictor of interest to your ‘biomes
  - Breaks any possible association between outcome and other covariates
  - Confounding is impossible
- Not always possible
  - e.g., observational studies, practical impossibilities…

OBSERVATIONAL STUDIES

- Cross-sectional
  - Allows you to make comparisons *across groups*
    - outcome ~ group
- Longitudinal
  - Allows you to make comparisons within individuals *across groups*
  - outcome ~ group | subject (with structured errors)

OBSERVATIONAL STUDIES

- There exists unhelpful rhetoric in the field about "longitudinal being better"
  - Effect sizes *within* individuals may be larger than effect sizes *across* groups…
  - … but these are *different* parameters 🍎🍊◼️🧀
- Cost also a consideration

TANK, CAGE EFFECTS

- If working with animal models, want to deconvolve housing & treatment as much as possible
  - Some Group A mice & some Group B mice in all cages is vastly preferable to all treatment mice in the same cage & all control mice in the same cage
  - Lots of cages & fewer mice per cage >> lots of mice in a cage
  - Sometimes this is impossible… ideally seek multiple cages per treatment group (to understand inter-cage variation)

BATCH EFFECTS

- If you are sequencing/sampling in batches, try to *balance* your batches across your other covariates
  - If you can't balance, *randomize*

IF YOU HAVE A CHOICE…

- If you have control over your design (not everyone does!)
- Take the time to simulate data under various designs and see which have the greater power
- *Much* easier to do this in R than IRL!

MOCKS

- If you contract a core for your sequencing…
  - Encourage them to sequence mock communities alongside all samples
- If you work at a core…
  - Please consider adding in system-specific mock communities in every run, and *sharing this data with your clients*

BLANKS

- Useful for detecting contamination
  - Ben says: 4+ useful for *decontam*
  - Cost: 4 additional extractions

REPLICATES

- Historically, variance prioritized over accuracy
- Good to know that your data generating process doesn't *just* produce noise
- If I can add technical replicates for free, yes!
- If I have to decide between biological replicates and technical replicates, *I will choose biological replicates* over technical replicates *almost always*
- BUT there may be settings where technical replicates are helpful
  - e.g., protocol optimization…

QUESTIONS ON EXPERIMENTAL DESIGN

CLOSING THOUGHTS

DIVERSITY

- If you really care about diversity, I recommend using
  - breakaway for species richness
  - DivNet for Shannon/Simpson diversity
  - ~~DivNet for weighted UniFrac/Bray-Curtis/Aitchison asking yourself why you care about β-diversity~~ Thinking creatively about how to make the comparisons that you care about

WELL-PLACING YOUR WORRY

- I discourage you from worrying about…
  - Is data compositional?
  - Where to rarefy to?
  - Which beta diversity metric?
- I encourage you to worry about…
  - Am I analyzing something I care about?
  - Can I estimate what I care about?
  - Is sequencing the technology that I need to answer my question?

AMY'S RECAP ON STATISTICAL PRACTICE

- Many current conventions for statistical analysis in science are inherently unscientific (yes, this is unfair!)
- I encourage you to
  - know what your hypothesis tests are testing: what is the parameter, model, null hypothesis…
  - not consider "p < [favourite number]" a benchmark for publication!

WHAT ELSE CAN WE DO?

- Be scientific
  - "What else could have driven this result?"
  - Corroborate your findings using multiple approaches
    - e.g. multiple HTS technologies, qPCR, explore literature with consideration of their methods, *in silico* modelling…

WHAT ELSE CAN WE DO?

- Report estimates, confidence intervals, and p-/q-values
- Understand your methods
- 🐤 test
- Plot your data & publish your figures, not (just) p-values
- Ask the developers!

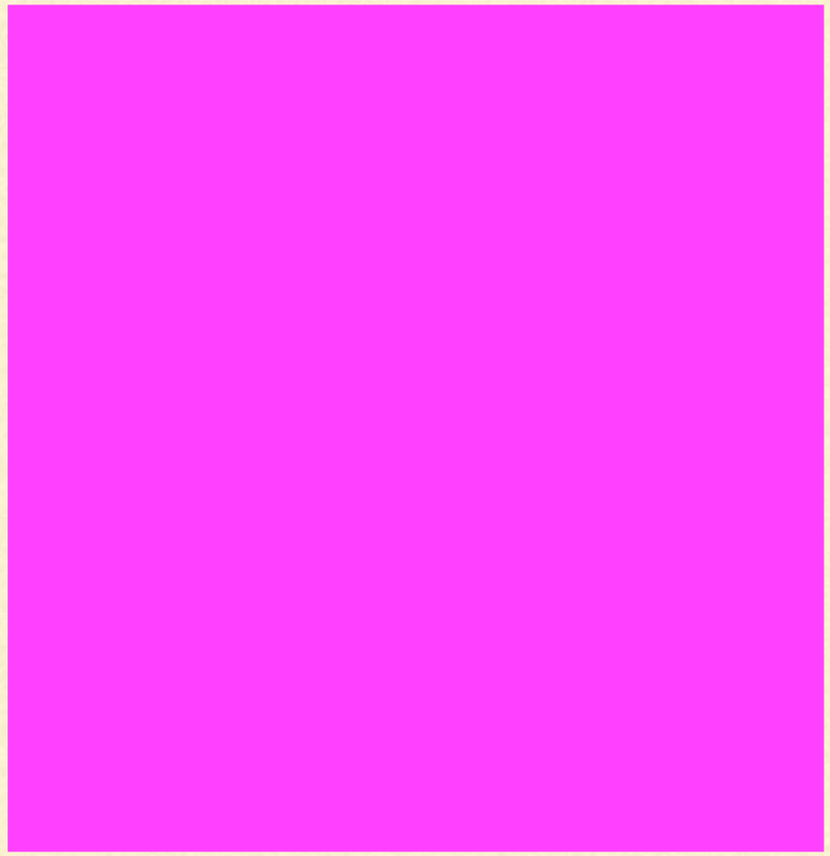
WHAT ELSE CAN WE DO?

- Be honest
  - Keep all analyses that you ran, not just the final one
- Write down all of the hypotheses that you care about
  - Before doing the experiment, before doing the analysis
- Your university might house a statistician; try to involve them...
  - ...in the entire process!

CLOSING THOUGHTS

- Methods for modeling microbiome data is a fast-moving field, and new methods are constantly emerging
- Talk to lots of people
  - "What's the biggest limitation of this?"
- Stay critical but open-minded

A COLLECTION OF KNOWLEDGE… OPINIONS… INSTINCTS… FROM



Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Associate Professor

Sarah Teichman — @sarah_teichman — PhD Candidate

Photo credit: T.D. Berry, Whitman lab, UW Madison

OPEN LAB TIME

- LM
- ABUNDANCE
- DIVERSITY