

MODELING MICROBIOME DATA

Statistical Diversity Lab @ University of Washington

Amy Willis — [@AmyDWillis](#) — Assistant Professor

David Clausen — [@davidandacat](#) — PhD Candidate

Sarah Teichman — [@sarah_teichman](#) — PhD Candidate

|

“How do I rigorously analyze my data?”

—Everyone, all the time

“It depends.”

—*Stat Div Lab, all the time*

DECIDING ON A MODEL

- Your scientific questions should guide you in choosing your model
 - Many studies look at multiple models
 - These can answer *the same or different* questions
- What type of data you have may also constrain you

There is not **one** way to model/analyse your data!
You need to decide what is important to you!

DECIDING ON A MODEL

- Different models make different assumptions...
 - ... and will lead you to different conclusions

LEARNING OBJECTIVES

- Learning objectives
 - 1. ~~Learn all the models~~
 - 2. ~~Understand all their assumptions~~
 - 3. ~~Resolve all confusion about statistical analysis of microbiome data~~

LEARNING OBJECTIVES

- Learning objectives
 - I. Learn *more* about some models
 - 2. Understand *some* of the *most important* assumptions and limitations of some methods
 - 3. Develop some facility using software to fit models
 - 4. Leave with more questions than ever

THE PLAN

- Modeling with microbiome data
 - Abundance
 - Multiple lecture + labs
 - Diversity:
 - Multiple lectures + lab
 - Presence/absence, trees, etc... next week!
 - Misc
 - Questions
-

THE PEP TALK

MODELING ABUNDANCE

ANALYSIS

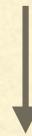
- There are many different data/sequencing types that can be used to model “abundance”
 - amplicon - count tables
 - shotgun - coverage, proportion data...
 - qPCR / ddPCR - counts/concentrations...
- The *type of data* you have impacts the *approach* you need

SCENARIO

ABSOLUTE ABUNDANCE DATA

e.g. qPCR (16S or taxon-specific); ddPCR

ABSOLUTE ABUNDANCE	MICROBE A	MICROBE B	MICROBE C
ENVIRO I	5	5	20
ENVIRO 2	10	10	40



observe

# COPIES OBS'D/UL	MICROBE A	MICROBE B	MICROBE C	TOTAL
ENVIRO I	4	5	18	27
ENVIRO 2	9	11	37	57

Can compare rows to rows, and columns to columns

SCENARIO

MOST HIGH-THROUGHPUT
SEQUENCING DATA
~~compositional data~~
e.g. 16S counts
e.g. genome/MAG coverages

ABSOLUTE ABUNDANCE	MICROBE A	MICROBE B	MICROBE C
ENVIRO 1	5	5	20
ENVIRO 2	10	10	40



observe

# OBSERVED	MICROBE A	MICROBE B	MICROBE C	TOTAL
ENVIRO 1	499	500	2001	3000
ENVIRO 2	250	251	1010	1511

SCENARIO

PROPORTION DATA

e.g., shotgun data processed w metaphlan

ABSOLUTE ABUNDANCE	MICROBE A	MICROBE B	MICROBE C
ENVIRO 1	5	5	20
ENVIRO 2	10	10	40



observe

# OBSERVED	MICROBE A	MICROBE B	MICROBE C	TOTAL
ENVIRO 1	$1.01 / 6 = \mathbf{0.168}$	$1/6 = \mathbf{0.167}$	$3.99 / 6 = \mathbf{0.665}$	1
ENVIRO 2	$0.99 / 6 = \mathbf{0.165}$	$0.99 / 6 = \mathbf{0.165}$	$4.02 / 6 = \mathbf{0.67}$	1

HOW DO YOU KNOW?

- You can't tell your data type from the tables alone
- You need some understanding of
 - what your technology is doing
 - and how it works

COMPARISONS WITH HTS DATA

- Can compare counts within a sample*, e.g. In Sample 1
 - 500 counts from Taxon B
 - 2001 counts from Taxon C
- Cannot compare counts across sample, e.g. for Taxon A
 - 499 from Enviro 1
 - 250 from Enviro 2

COMPARISONS WITH HTS DATA

- I recommend against transforming sequences to proportions, and avoiding bioinformatics pipelines that only give you proportions
 - This is not necessary!
 - It loses information about precision!
 - Good statistical methods model precision

MODELING ABUNDANCE

- **Modeling absolute abundances**
- Modeling counts (via relative abundance parameters)
- Modeling absolute abundance parameters from relative

ABSOLUTE ABUNDANCE DATA

- qPCR and ddPCR data can usually be modeled with techniques you can learn in Applied Stats 101
- **Linear models** most generally look like

$$\text{mean outcome}_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

ABSOLUTE ABUNDANCE DATA

- Examples:

$$\text{mean bacterial load}_i = \beta_0 + \beta_1 \times \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}}$$

- $\hat{\beta}_0$ is an estimate of the mean/average/expected bacterial load for people not on antibiotics
- $\hat{\beta}_1$ is an estimate of the difference in mean bacterial load between people who are versus aren't on antibiotics

ABSOLUTE ABUNDANCE DATA

$$\text{mean bacterial load}_i = \beta_0 + \beta_1 \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}} + \beta_2 (\text{age}_i - 40)$$

- $\hat{\beta}_0$ is an estimate of the mean bacterial load for 40 y.o.'s people *not on antibiotics*
- $\hat{\beta}_1$ is an estimate of the difference in mean bacterial load between people of the same age who *are* versus *aren't* on antibiotics
- $\hat{\beta}_2$ is an estimate of the difference in mean bacterial load between people who differ in age by 1 year who have the same antibiotics use

ABSOLUTE ABUNDANCE DATA

$$\begin{aligned}\text{mean bacterial load}_i = & \beta_0 + \beta_1 \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}} + \beta_2 (\text{age}_i - 40) \\ & + \beta_3 \times \mathbf{1}_{\{\text{person } i \text{ is on antibiotics}\}} \times (\text{age}_i - 40)\end{aligned}$$

- $\hat{\beta}_0$ is an estimate of the mean bacterial load for 40 y.o.'s people *not* on antibiotics
- $\hat{\beta}_1$ is an estimate of the difference in mean bacterial load between 40 y.o.'s who *are* versus *aren't* on antibiotics
- $\hat{\beta}_2$ is an estimate of the difference in mean bacterial load between people who differ in age by 1 year who *aren't* on antibiotics
- $\hat{\beta}_3$ is an estimate of the difference in mean bacterial load between people who differ in age by 1 year who *are* on antibiotics, compared to between people who differ in age by 1 year who *aren't* on antibiotics

ABSOLUTE ABUNDANCE DATA

- Step 1: decide on your model
- Step 2: figure out how to fit it

ABSOLUTE ABUNDANCE DATA

- Step 1: decide on your model

mean bacterial load_i = $\beta_0 + \beta_1 \mathbf{1}_{\text{person } i \text{ is on antibiotics}} + \beta_2 \mathbf{1}_{\text{person } i \text{'s sample is from sputum}}$

- Step 2: figure out how to fit it

```
> my_data %>%  
+   lm(ddpcr ~ Treatment + `Sample Type`, data = .)
```

Call:

```
lm(formula = ddpcr ~ Treatment + `Sample Type`, data = .)
```

Coefficients:

	TreatmentON
(Intercept)	996530

996530	-409238
--------	---------

`Sample Type`Sputum	1006955
---------------------	---------

ABSOLUTE ABUNDANCE DATA

- Step 1: decide on your model

mean bacterial load_i = $\beta_0 + \beta_1 \mathbf{1}_{\text{person } i \text{ is on antibiotics}} + \beta_2 \mathbf{1}_{\text{person } i \text{'s sample is from sputum}}$

+ $\beta_3 \mathbf{1}_{\text{person } i \text{ is on antibiotics}} \mathbf{1}_{\text{person } i \text{'s sample is from sputum}}$

- Step 2: figure out how to fit it

```
> my_data %>%
+   lm(ddpcr ~ Treatment * `Sample Type`, data = .)
```

Call:

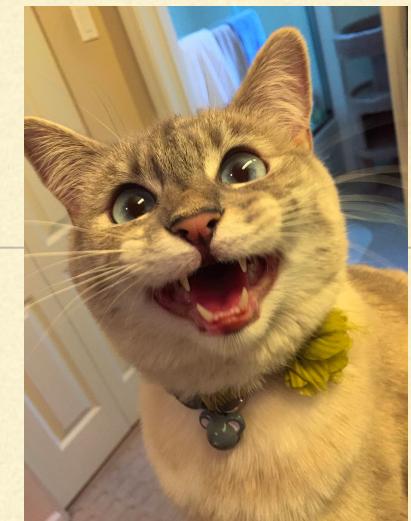
```
lm(formula = ddpcr ~ Treatment * `Sample Type`, data = .)
```

Coefficients:

(Intercept)		TreatmentON
968202		-234550
`Sample Type`Sputum	TreatmentON:`Sample Type`Sputum	
1063610		-349375

Get pumped!

ACCESSING `LM` LAB



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

```
lm lab:
```

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-afternoon/labs/lm_lab/lm-ggplot-lab.R", "lm-ggplot-lab.R")
```

3. Run this command in your RStudio Server console

The screenshot shows the RStudio Server interface with the 'Console' tab selected. The console window displays an R session. The R logo icon is visible, followed by 'R 4.2.1 · ~/'. Below the logo, a command is shown in blue: `> download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-afternoon/labs/lm_lab/lm-ggplot-lab.R", "lm-ggplot-lab.R")`. The text is highlighted in blue, suggesting it is selected for copying.

MODELING ABUNDANCE

- Modeling absolute abundances
- **Modeling counts (via relative abundance parameters)**
- Modeling absolute abundance parameters from relative

MODELING COUNTS

- Common goal:
 - Determine which taxa are present in greater abundance in one group compared to another
 - “*Differential abundance [is] a category subject to some controversy in part on account of the fact that no unambiguous definitions of ‘differential’ or ‘abundance’ are widely agreed upon.*”

MODELING COUNTS

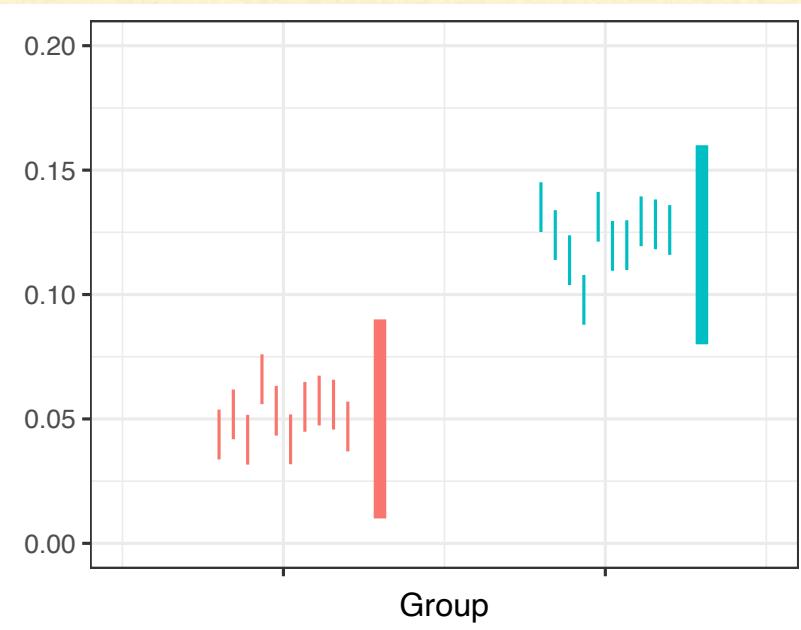
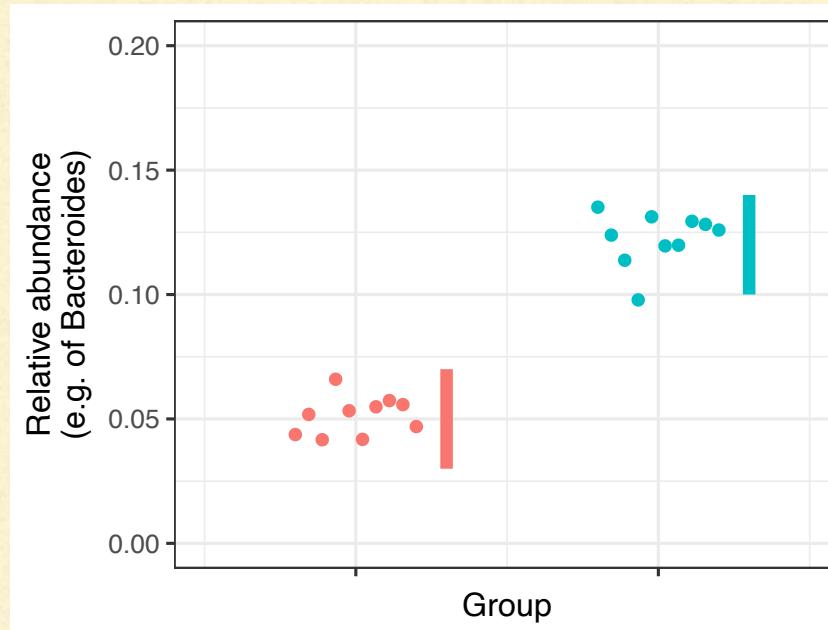
- Many methods exist
 - ALDEx
 - ANCOM
 - corncob
 - DESeq2
 - edgeR
 - metagenomeSeq
 - MaAsLin
 - LEfSE
 - limma voom
 - Wilcoxon on proportions
 - t-tests on proportions
 - multiple versions of almost all methods; multiple options for almost all methods
-

MODELING COUNTS

- Many methods exist
 - ALDEEx
 - ANCOM
 - **corncob**
 - DESeq2
 - edgeR
 - metagenomeSeq
 - MaAsLin
 - LEfSE
 - limma voom
 - Wilcoxon on proportions
 - t-tests on proportions
- multiple versions of almost all methods; multiple options for almost all methods

SAMPLE \neq POPULATION

- Observed relative abundance \neq true relative abundance
- Any statistical test for the microbiome needs to account for this *measurement error*



CORNCOB

COmpositional RegressioN for Correlated Observations with the Beta-binomial



- Models **relative abundance**
 - Connects count data to relative abundances
 - Models relative abundance with covariates
- Hypothesis testing for changes in
 - relative abundance
 - variance in abundance



Bryan Martin, UW Statistics



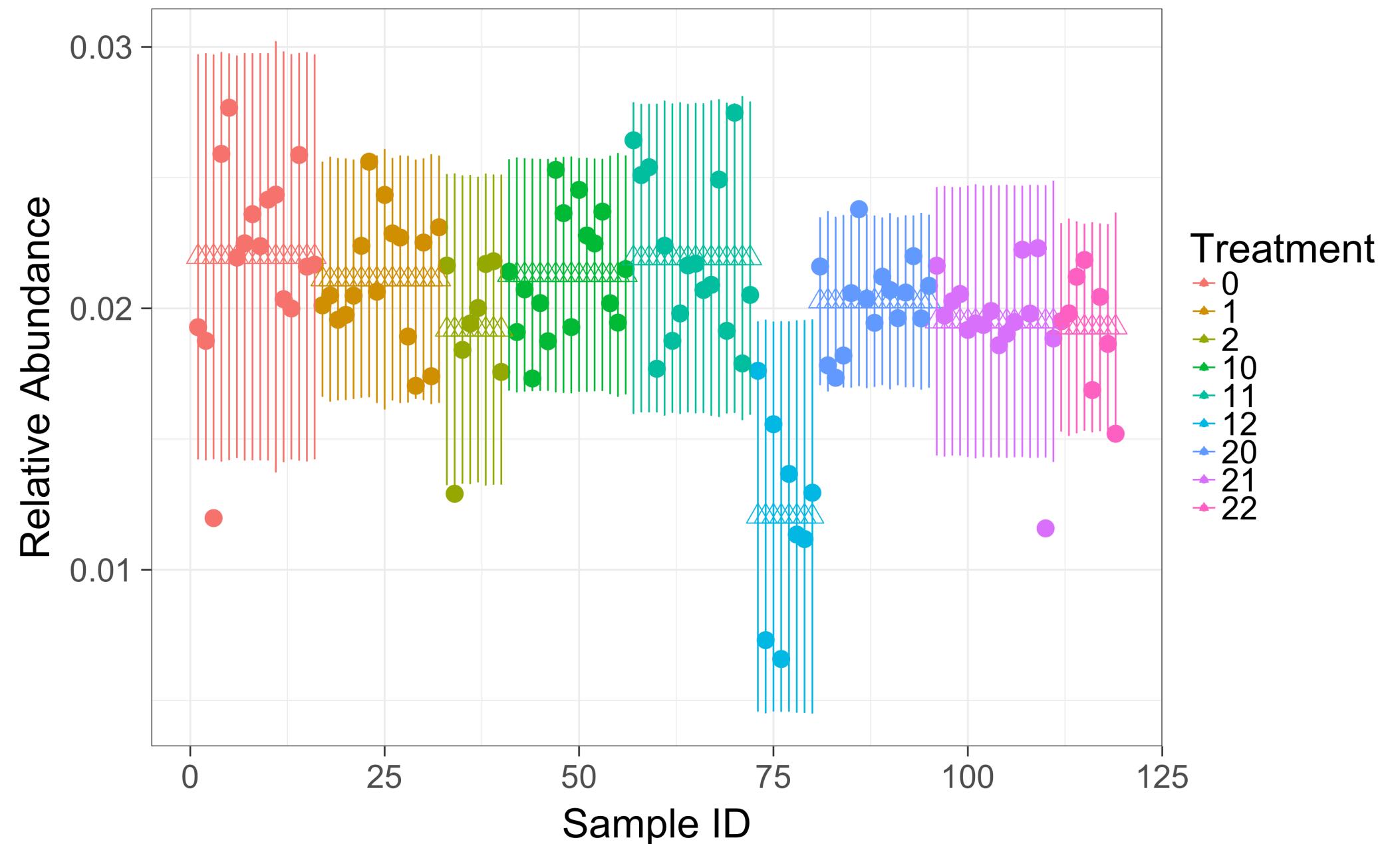
Daniela Witten, UW Statistics

CORNCOB

COmpositional RegressioN for Correlated Observations with the Beta-binomial



- “The relative abundance of *S. aureus* is significantly lower in the treatment group (95% CI for $\beta_{treatment}$: (-4.74, -2.91), $q = 0.003$, see Methods).”
- “Methods: All phylum-level relative abundances were modeled using corncob* with a logit-link for mean and dispersion. Differential abundance (DA) was modeled as a linear function of age and treatment group. Differential variability was modeled as a function of treatment group. The parametric Wald test was used to test DA hypotheses. . . .”



Data from Whitman et. al, (2015) ISME

CORNCOB

COmpositional RegressioN for Correlated Observations with the Beta-binomial



- Addresses measurement error issue
- Adjusts for different sequencing depths
- Suitable with multiple covariates, various experimental designs
- Mean and variance testing
- All taxa/genes; inbuilt FDR control
- No need to choose a pseudocount

BETA-BINOMIAL DISTRIBUTION

$$W_i | Z_i, M_i \sim \text{Binomial}(M_i, Z_i)$$

$$Z_i \sim \text{Beta}(a_{1i}, a_{2i})$$

- n = samples, indexed by $i = 1, \dots, n$
- W_i = # of individuals observed in the taxon/gene of interest
- M_i = total # of counts observed
- Z_i = the (latent) relative abundance in sample i

BETA-BINOMIAL DISTRIBUTION

$$W_i \mid Z_i, M_i$$

$$M_i \sim$$

$$\text{Binomial}(M_i, Z_i)$$

what you need

$$Z_i \sim \text{Beta}(a_{1i}, a_{2i})$$

- n = samples, indexed by $i = 1, \dots, n$
- W_i = # of individuals observed in the taxon/gene of interest
- M_i = total # of counts observed
- Z_i = the (latent) relative abundance in sample i

LINKING ABUNDANCE TO COVARIATES

1. Parameters

$$\mu_i = \frac{a_{1,i}}{a_{1,i} + a_{2,i}}, \quad \text{"(latent) relative abundance"}$$

$$\phi_i = \frac{1}{a_{1,i} + a_{2,i} + 1} \quad \begin{array}{l} \text{"within sample correlation"} \\ \text{"absolute abundance overdispersion"} \end{array}$$

2. Link to covariates

μ_i is a function of $\mathbf{X}_i, \boldsymbol{\beta}$

ϕ_i is a function of $\mathbf{X}_i^*, \boldsymbol{\beta}^*$

LINKING ABUNDANCE TO COVARIATES

1. Parameters

$$\mu_i = \frac{a_{1,i}}{a_{1,i} + a_{2,i}}, \quad \text{"(latent) relative abundance"}$$

$$\phi_i = \frac{1}{a_{1,i} + a_{2,i} + 1} \quad \begin{array}{l} \text{"within sample correlation"} \\ \text{"absolute abundance overdispersion"} \end{array}$$

2. Link to covariates

what you need

μ_i is a function of \mathbf{X}_i , β

ϕ_i is a function of \mathbf{X}_i^* , β^*

Call:

```
bbdml(formula = OTU.1 ~ Day + Amdmt, phi.formula = ~Day, data = soil_full)
```

Coefficients associated with abundance:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.78562	0.02507	-31.336	< 2e-16 ***
Day1	0.35077	0.03807	9.214	2.31e-15 ***
Day2	0.14267	0.02389	5.971	2.88e-08 ***
Amdmt1	0.03466	0.02436	1.423	0.158
Amdmt2	0.19374	0.04120	4.702	7.44e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Coefficients associated with dispersion:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.6957	0.2971	-19.173	<2e-16 ***
Day1	1.1279	0.4366	2.584	0.0111 *
Day2	-0.7231	0.4292	-1.685	0.0948 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-likelihood: -1056.1

CORNCOB

COmpositional RegressioN for Correlated Observations with the Beta-binomial



- Coefficients are on the logit relative abundance scale
 - Negative coefficients indicate decreased estimated abundance
- Usually we are interested in testing for differential abundance, controlling for the effect of the covariate on variability

CORNCOB AND DESEQ2

■ Similarities

- Easy to use
- Use counts to assess precision of estimates (no rarefying)
- Benjamini-Hochberg adjustment for multiple comparisons
- Dispersion parameter for overdispersion
- Tests for differential abundance

CORNCOB AND DESEQ2

- Designed for microbiome count data
- Models relative abundance & overdispersion
- Different structure for different taxa
- Models zeros without pseudocounts
- Easy to diagnose model misspecification
- Designed for RNAseq data
- Models relative abundance only
- Constrained dispersion
- Individual microbes are assumed independent
- Geometric mean can't handle zeroes without pseudocounts
- Not so easy to diagnose model misspecification problems

SUMMARY: CORNCOB



- Modeling and hypothesis testing for relative abundances using count data
- Adjusts for different sequencing depth
- Valid hypothesis testing with small sample sizes (use Beta-binomial bootstrap for hypothesis test)
- False discovery rate control for testing many taxa

🌽🌽 github.com/bryandmartin/corncob 🌽🌽

🎉🎉 Martin, Witten & Willis, 2020, *Annals of Applied Statistics* 🎉🎉

ACCESSING ‘CORNCOB’ LAB



Get pumped!

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

on

```
corncob lab:  
download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-  
afternoon/labs/corncob_tutorial/corncob_tutorial.R", "corncob-lab.R")
```

3. Run this command in your RStudio Server console

```
Console Terminal × Jobs ×  
R 4.2.1 · ~/  
> download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-afternoon/labs  
/corncob_tutorial/corncob_tutorial.R", "corncob-lab.R")|
```

MODELING ABUNDANCE

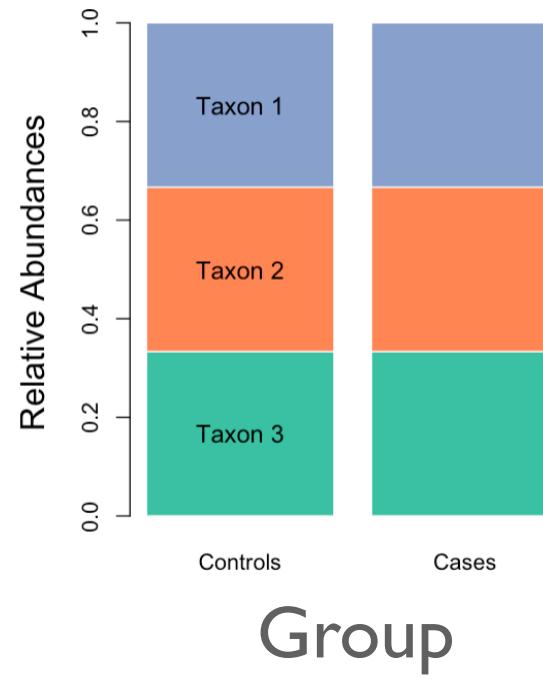
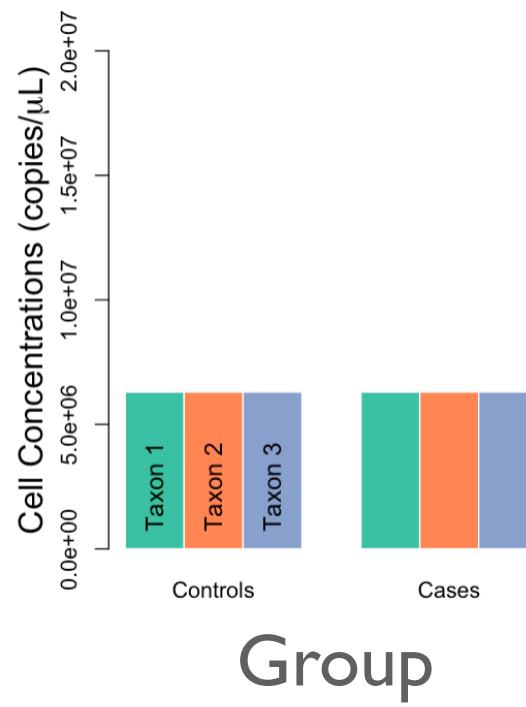
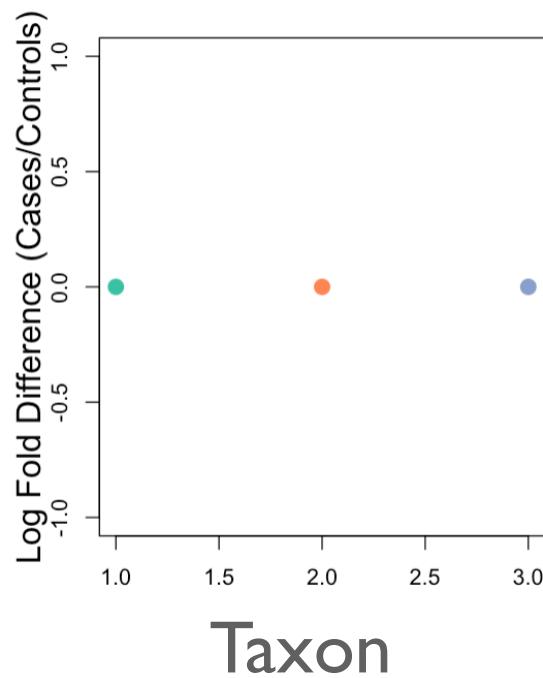
- Modeling absolute abundances
- Modeling counts (via relative abundance parameters)
- **Modeling absolute abundance parameters from relative**

RADEMU

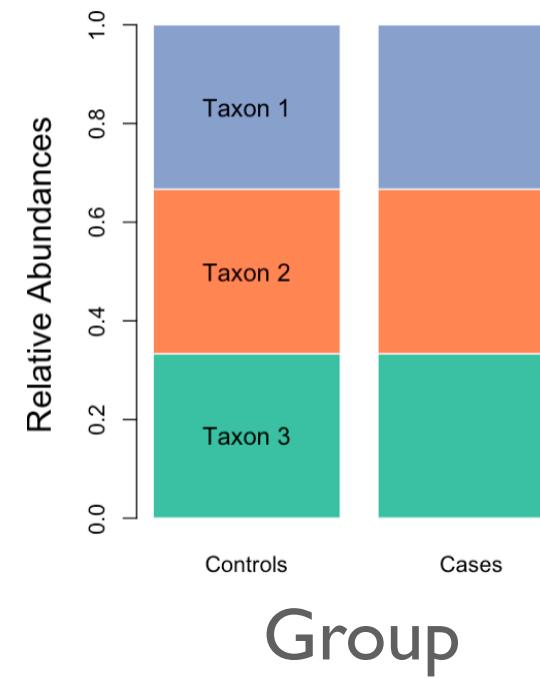
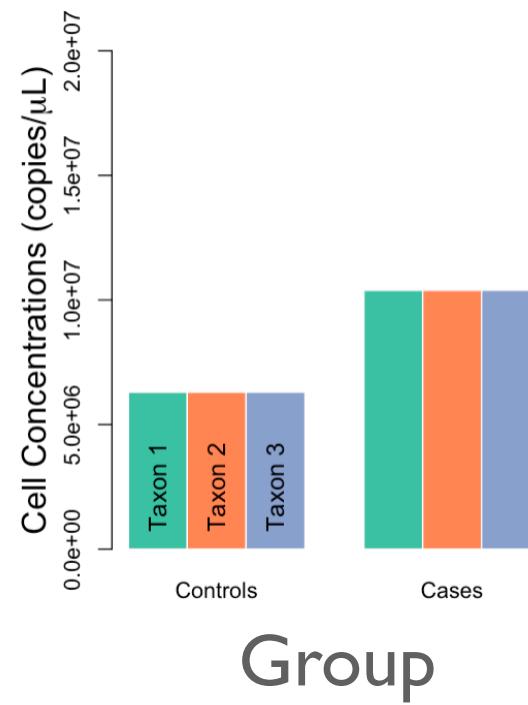
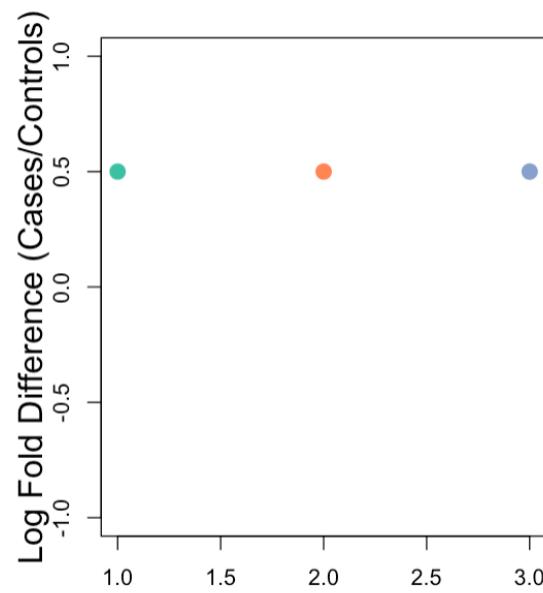
- radEmuAbPill
 - Using **relative abundance** data
 - to estimate **multiplicative** differences in **absolute** abundances
 - with **partially identified log-linear models**
 - Basic idea
 - We observe “relative abundance” data (e.g., #reads by taxon)
 - We’d like to say something about an “absolute” quantity (like cell concentration)
 - Can we?



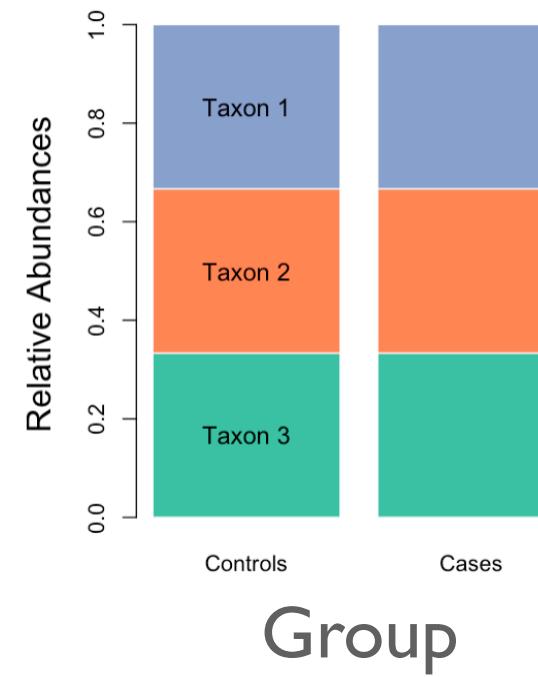
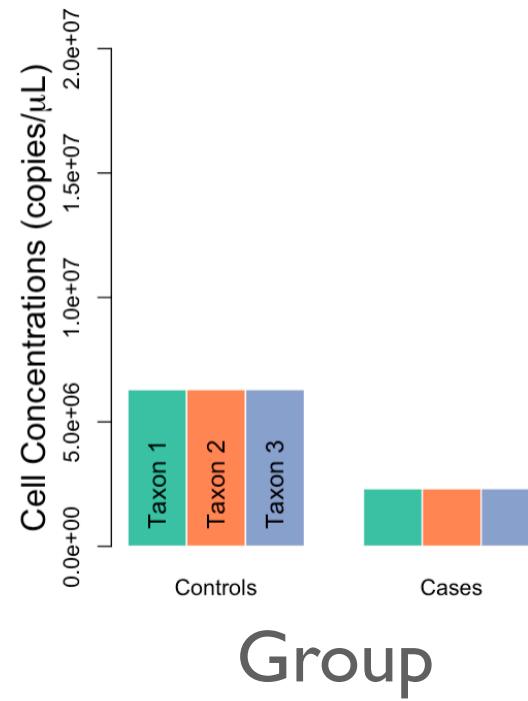
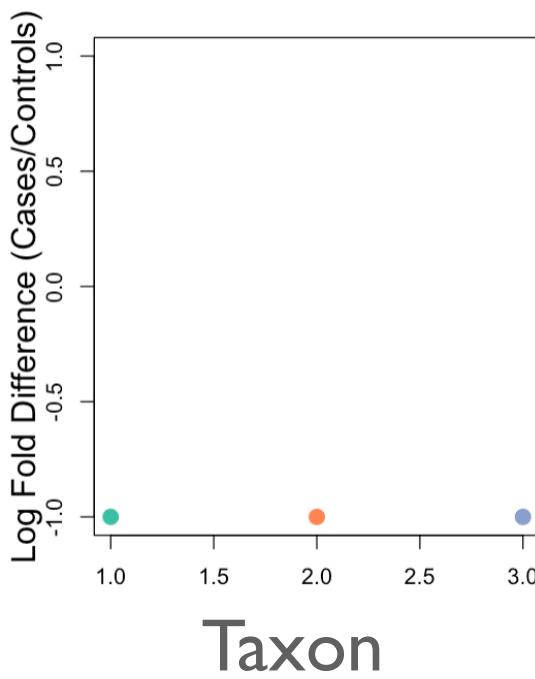
FOLD DIFFERENCES



FOLD DIFFERENCES

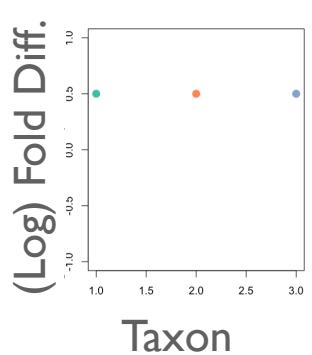
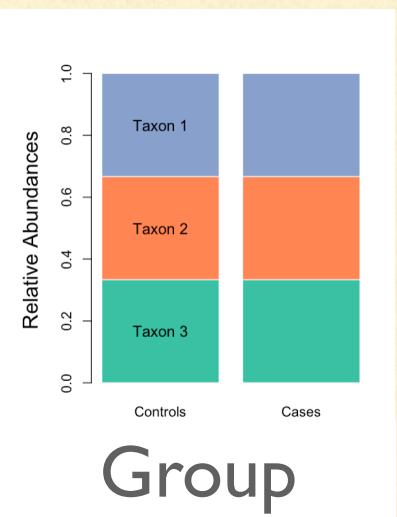


FOLD DIFFERENCES



STARTING FROM PROPORTIONS

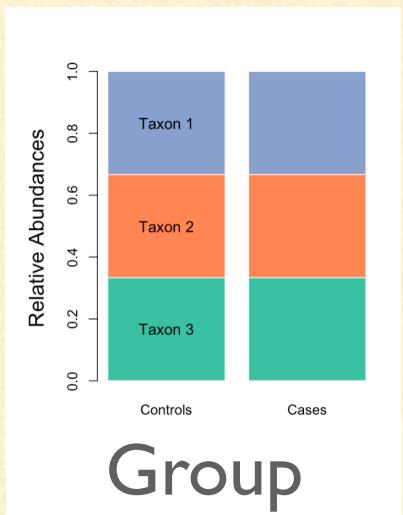
We observe



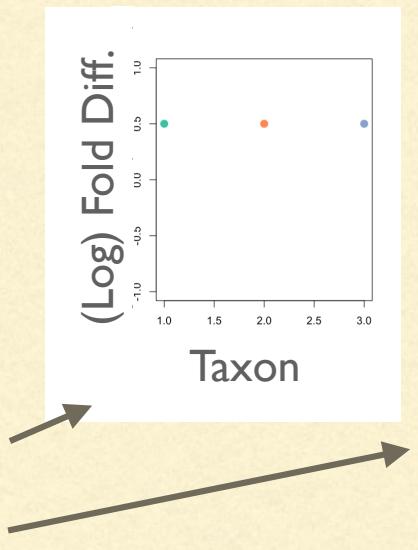
Consistent with this

STARTING FROM PROPORTIONS

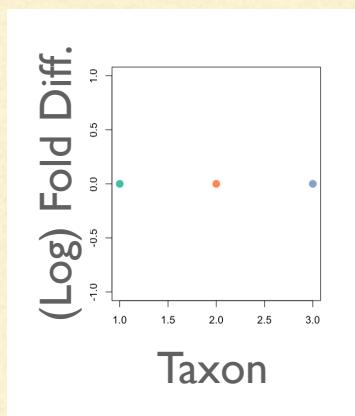
We observe



Group



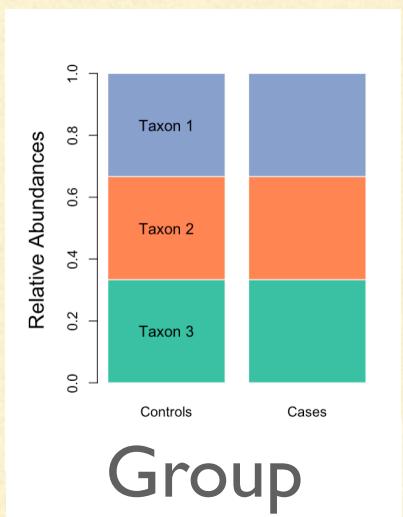
Consistent with this



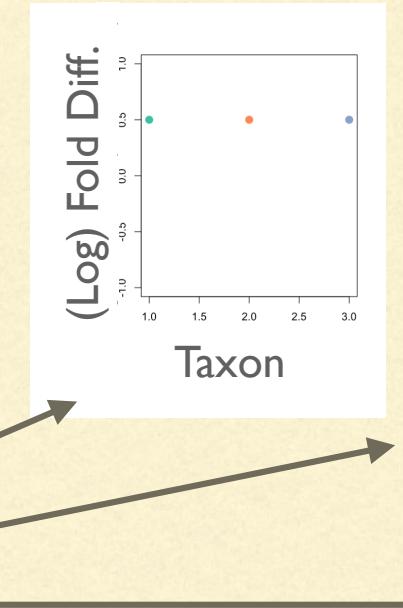
Or this

STARTING FROM PROPORTIONS

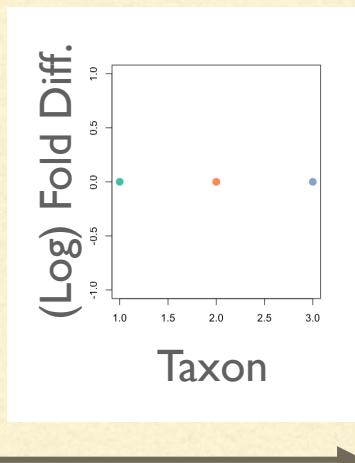
We observe



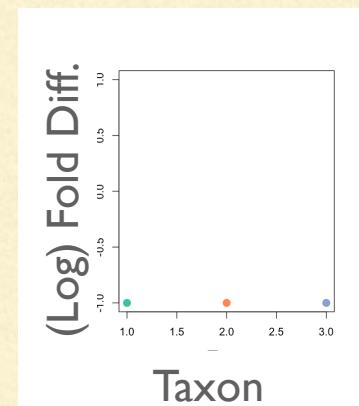
Group



Consistent with this



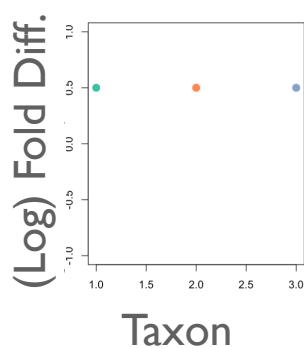
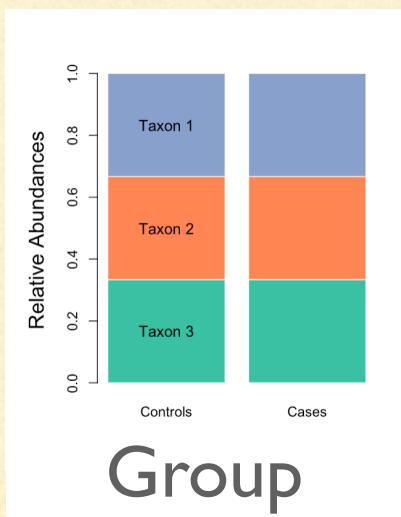
Or this



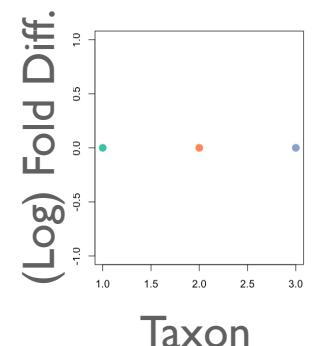
Or this

STARTING FROM PROPORTIONS

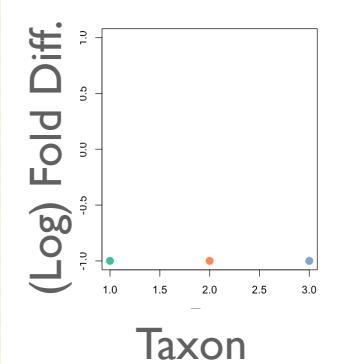
We observe



Consistent with this



Or this

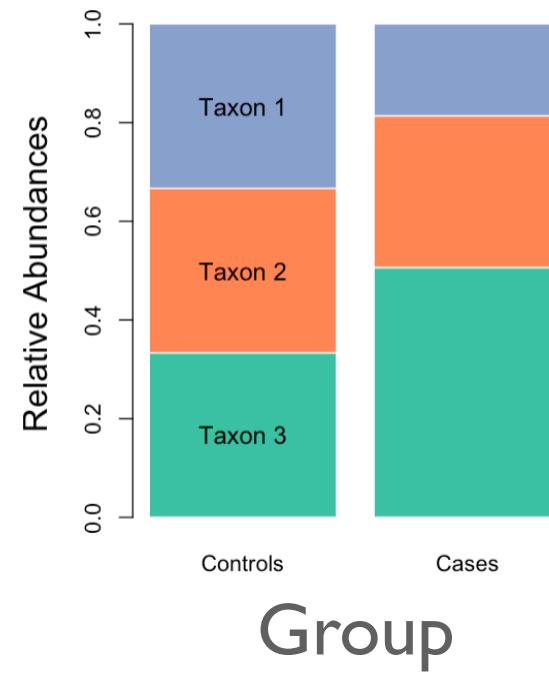
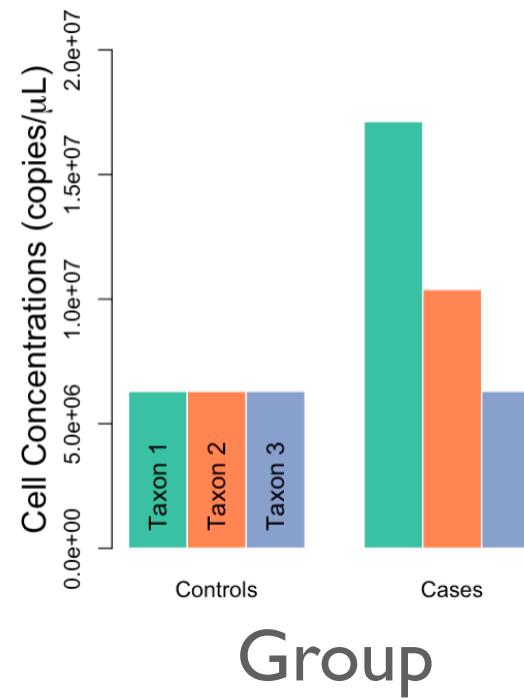
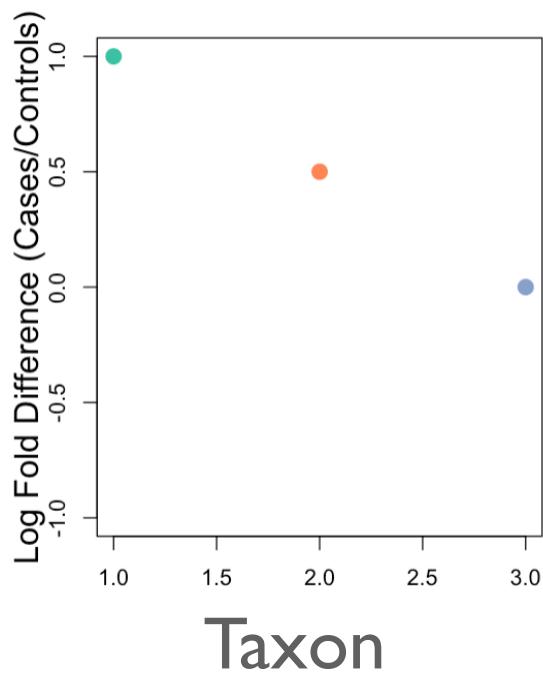


Or this

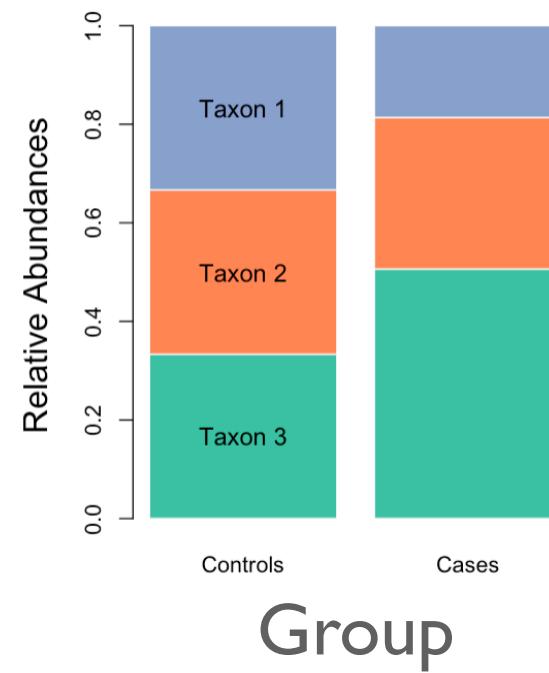
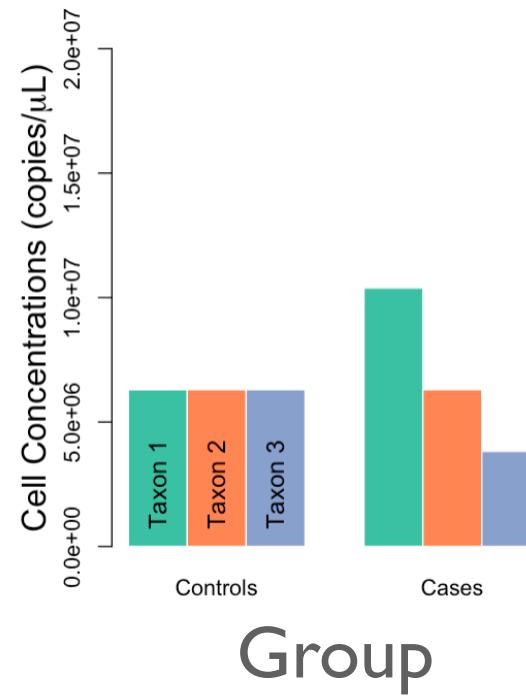
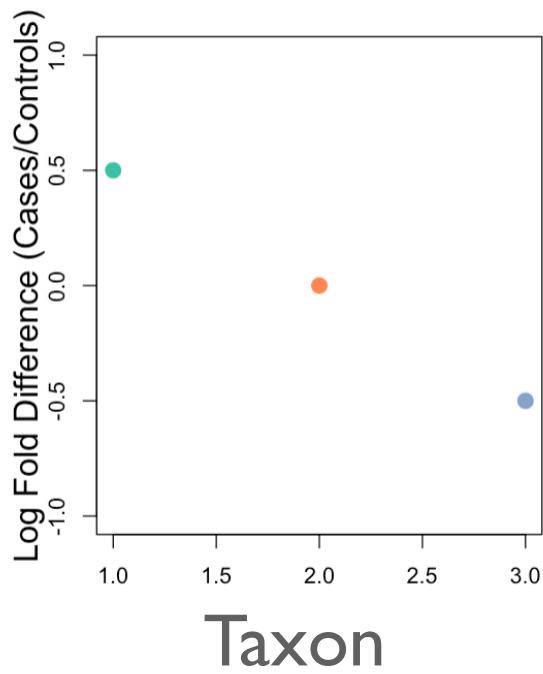
Or...

• • •

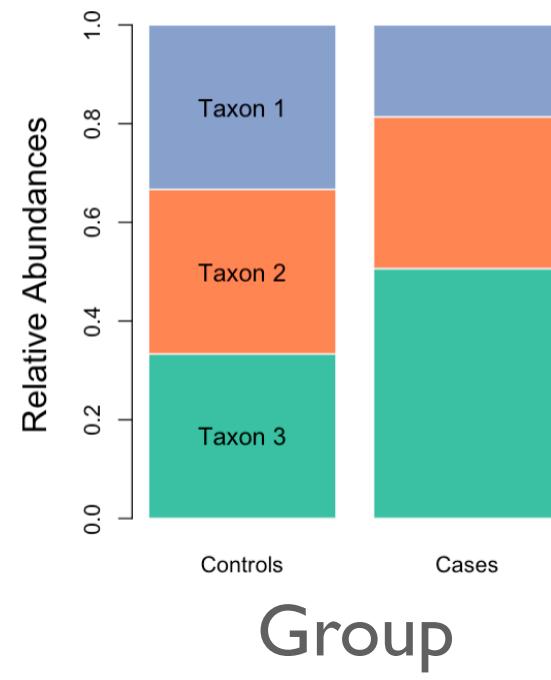
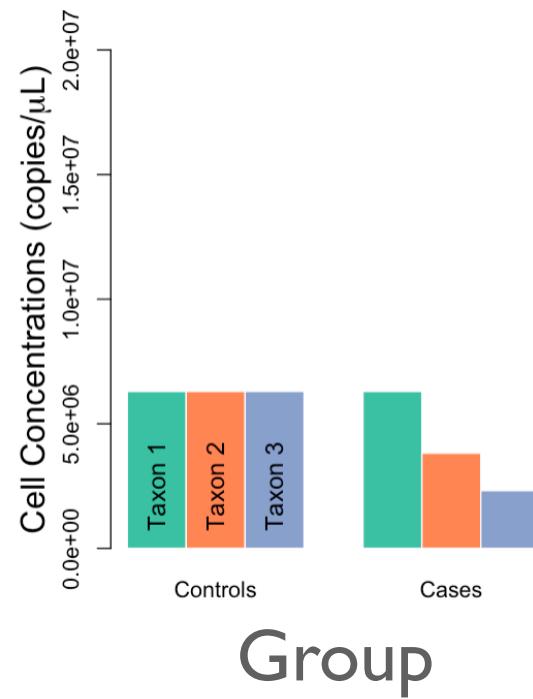
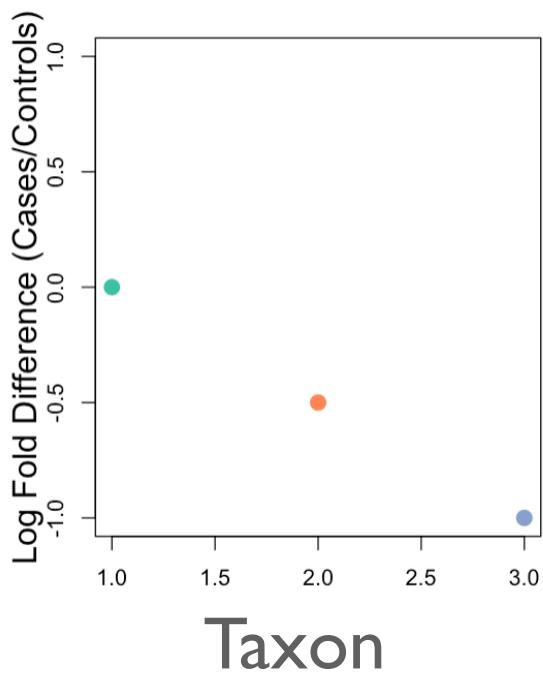
FOLD DIFFERENCES



FOLD DIFFERENCES

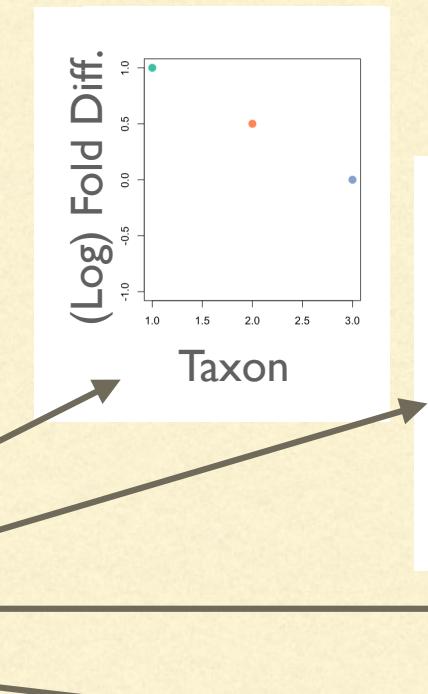
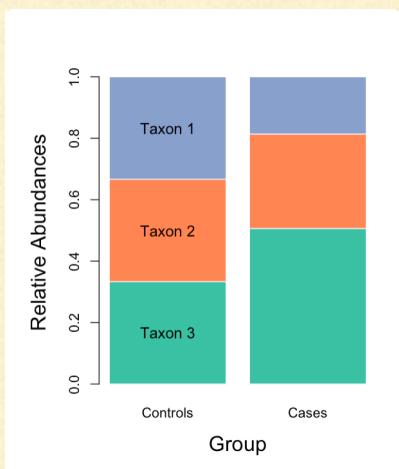


FOLD DIFFERENCES

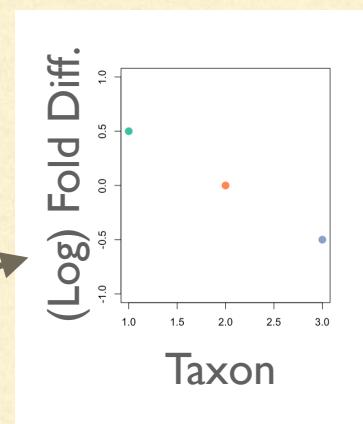


STARTING FROM PROPORTIONS

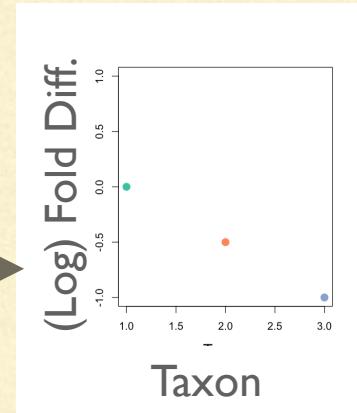
We observe



Consistent with this



Or this



Or this

Or...

• • •

RADEMU AGAIN

MATH FACT!

- This pattern holds in general
- We can use this to estimate fold differences* in absolute abundance from sequencing data
 - *Up to a location shift

RADEMU AGAIN

MATH FACT!

- This pattern holds in general
- We can use this to estimate fold differences* in absolute abundance from sequencing data
- *Up to a location shift



RADEMU AGAIN

MATH FACT!

$$\text{e.g., } \frac{\text{mean cell conc. } F. \text{ prauznitzii in cases}}{\text{mean cell conc. } F. \text{ prauznitzii in controls}}$$

- This pattern holds in general
- We can use this to estimate fold differences* in absolute abundance from sequencing data
- *Up to a location shift



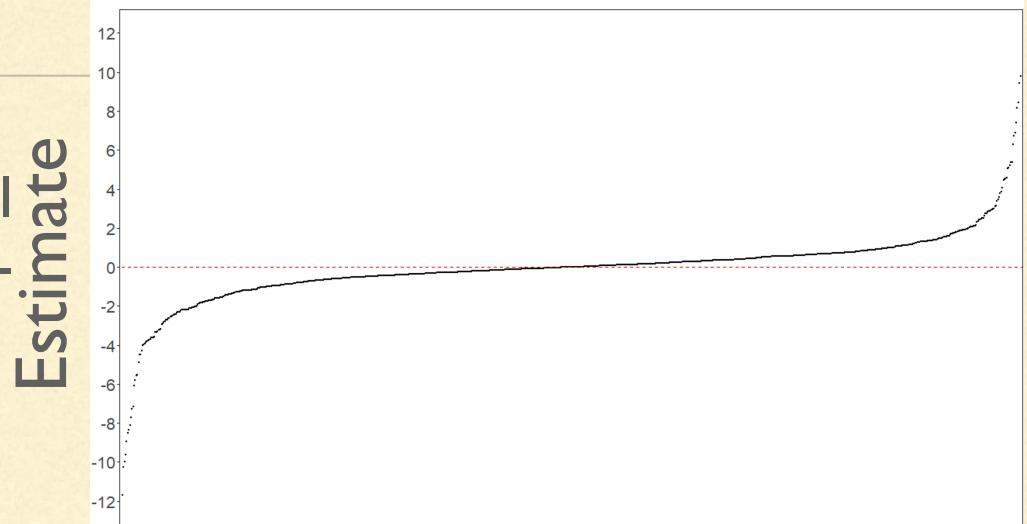
RADEMU: EXAMPLE

- Wirbel et al. (2019): meta-analysis of 4 case-control studies of association between gut microbiome and colorectal cancer
 - Cases = participants diagnosed with colorectal cancer
 - Controls = participants without a colorectal cancer diagnosis
- Published data: mOTU tables
 - mOTU = “metagenomic OTU”
 - Clustering based on a set of highly conserved marker genes
- Our goal: characterize which mOTUs unusually over/under-represented in cases

Wirbel, Jakob, et al. "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer." *Nature Medicine* 25.4 (2019): 679-689.

RADEMU: EXAMPLE

- Here's some output from a model fitted on a colorectal cancer case-control dataset
 - What's on y-axis?
 - $\log \frac{\text{mean cell conc. taxon j in cases}}{\text{mean cell conc. taxon j in controls}}$
 - but constrained: we force median to be zero
 - **This changes interpretation!**



RADEMU: EXAMPLE

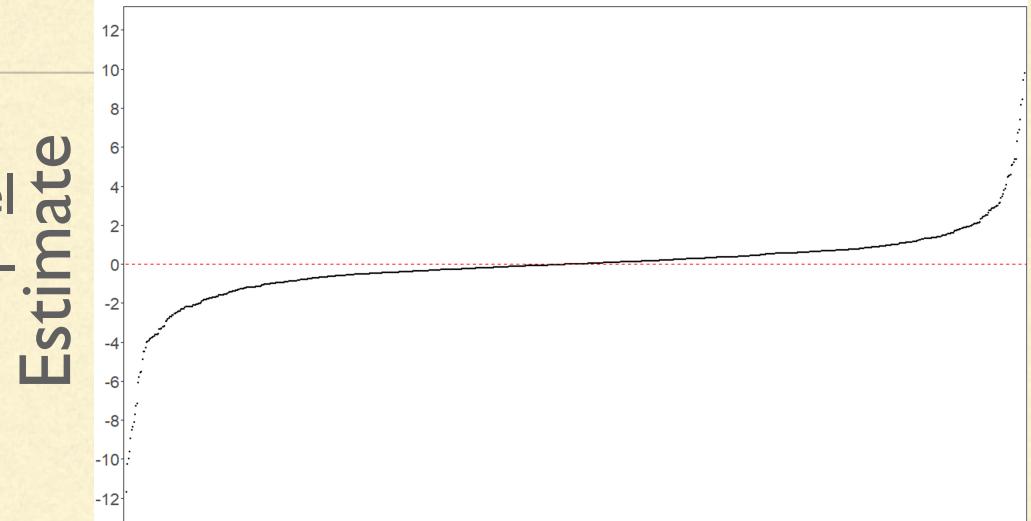
- Here's some output from a model fitted on a colorectal cancer case-control dataset

- What's on y-axis?

- $\log \frac{\text{mean cell conc. taxon } j \text{ in cases}}{\text{mean cell conc. taxon } j \text{ in controls}}$

- but constrained: we force median to be zero

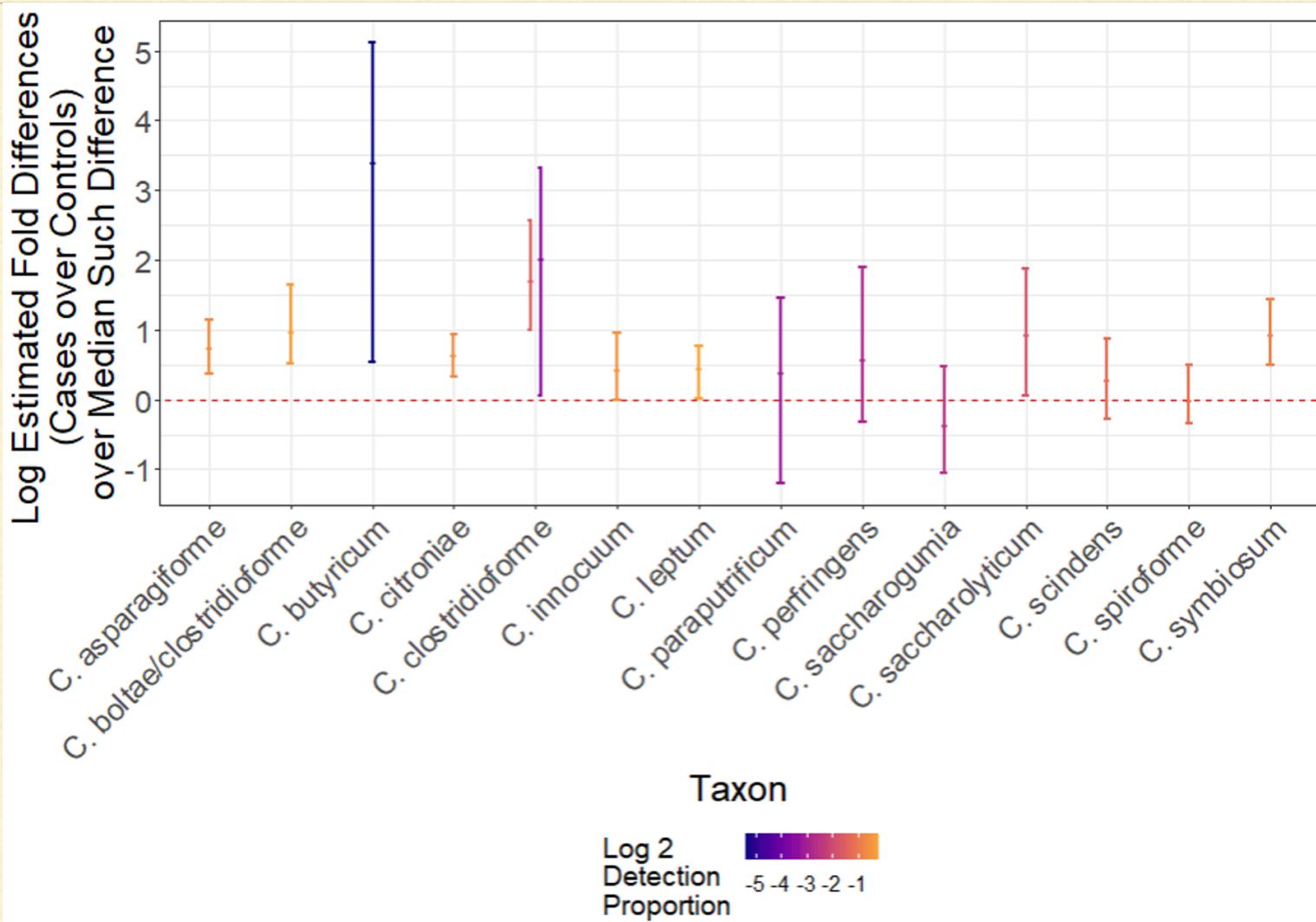
- **This changes interpretation!**



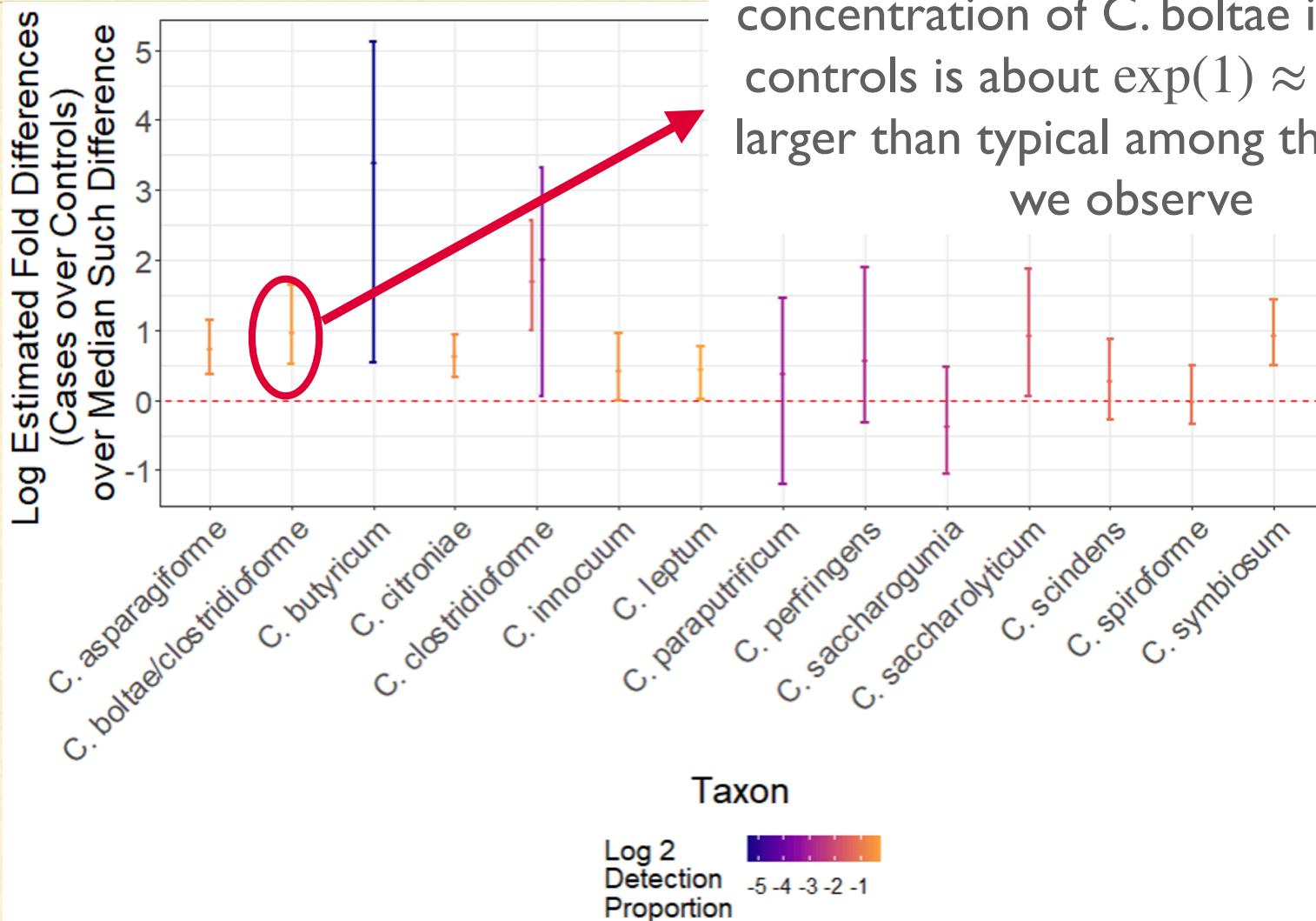
Taxon

How different are cell concentrations across groups... relative to the *typical* difference?

RADEMU: EXAMPLES



RADEMU: EXAMPLES



We estimate that the ratio of mean concentration of *C. boltae* in cases to controls is about $\exp(1) \approx 2.7$ times larger than typical among the mOTUs we observe

RADEMU: ASSUMPTIONS

- What assumptions does radEmu make (that you might consider caring about)?
 - Most important:
 - expected counts* are proportional to mean \times efficiencies
 - Multiplicative over/under-detection of taxa not a problem
 - Contamination: could be a problem!

*not actually necessary for the outcome to be a count (coverage would also work, for example)

ACCESSING ‘RADEMU’ LAB



Get pumped!

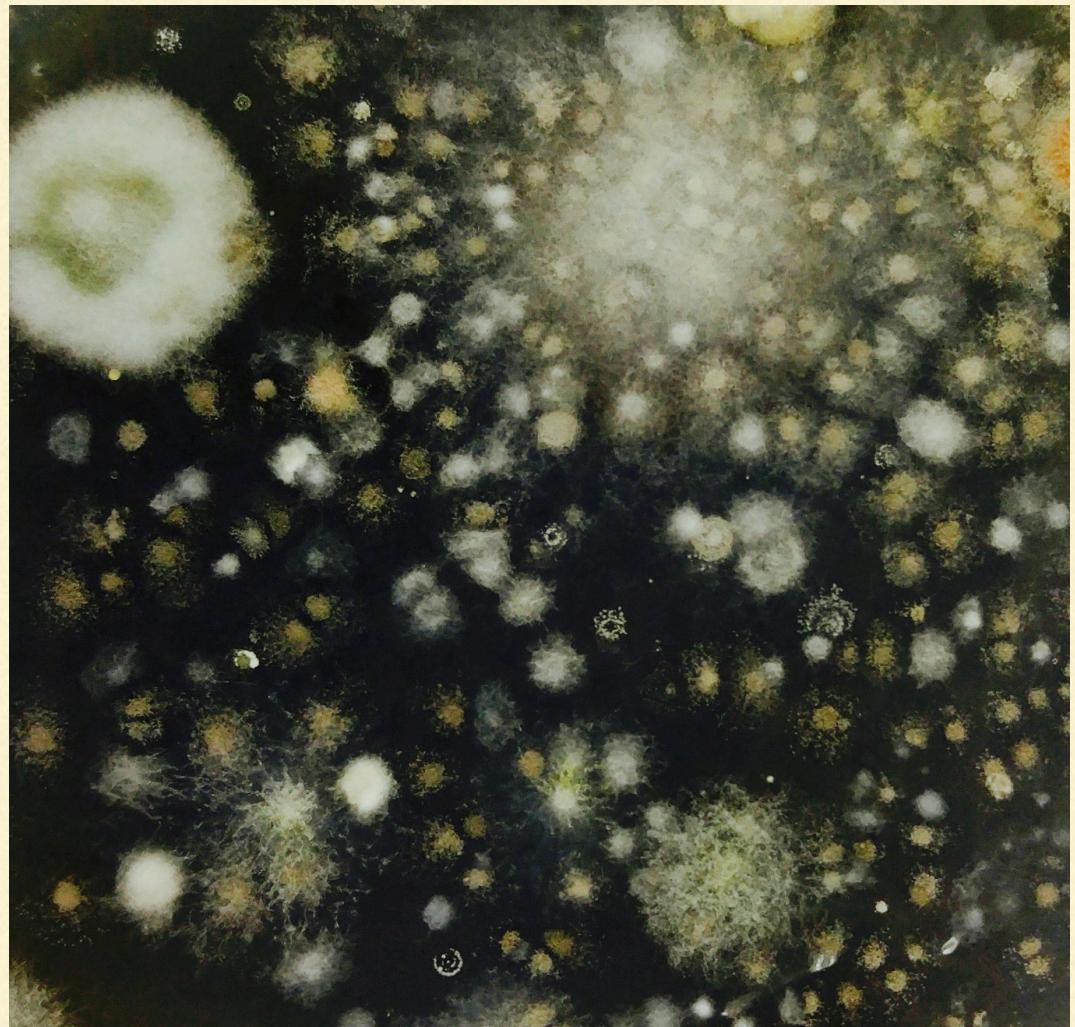
- I. Go to schedule on Wiki to Sunday afternoon, click on “Labs”
2. Copy the command under the lab we’re working on

```
rad emu lab:  
download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-  
afternoon/labs/rademu_lab/rademu_lab.R", "rad-emu-lab.R")
```

3. Run this command in your RStudio Server console

A screenshot of an RStudio Server interface. At the top, there are tabs for 'Console' (which is selected), 'Terminal x', and 'Jobs x'. Below the tabs, the R logo and 'R 4.2.1 · ~/`' are displayed. In the main area, a blue command line shows the code: '> download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-afternoon/labs/rademu_lab/rademu_lab.R", "rad-emu-lab.R")|'. The cursor is at the end of the command.

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

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, \dots, p_c
- Any function of
 - p_1, p_2, \dots, p_c OR
phylogeny
 - p_1, p_2, \dots, p_c and ~~some info about relationships amongst groups~~

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



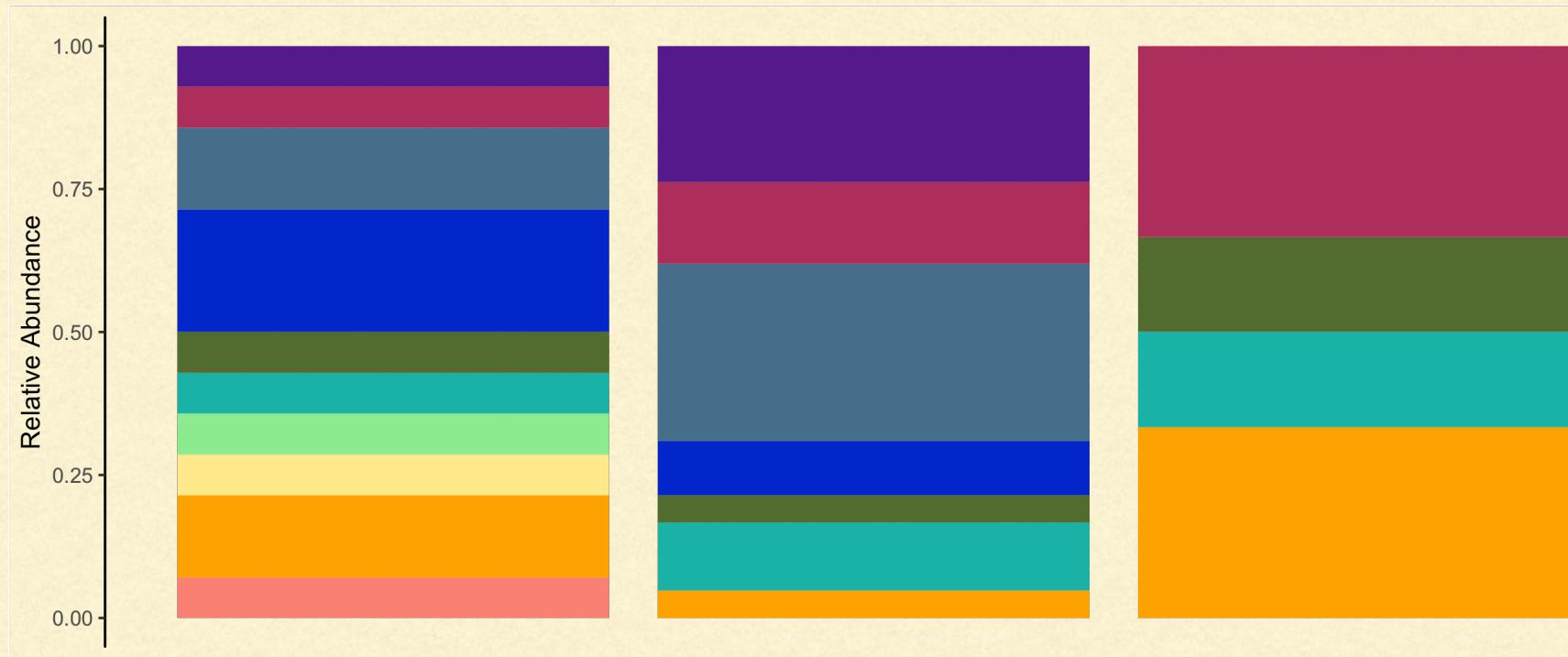
YOUR CHOICE



YOUR CHOICE



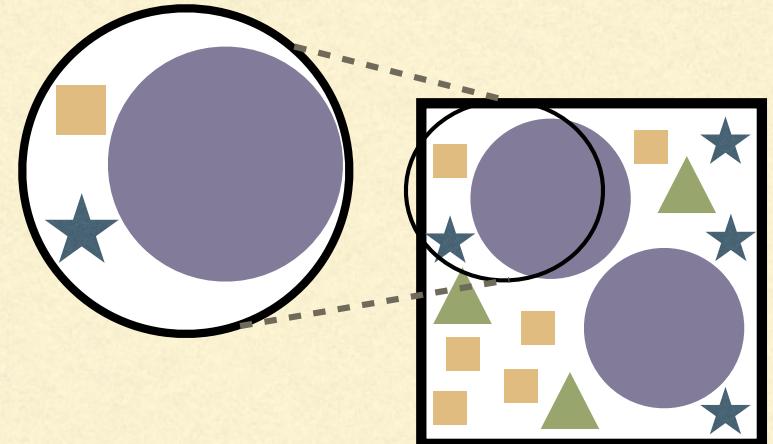
This is a question of *parameter choice*:
Which parameter highlights the differences I care about?



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, \dots, p_c
- We need to *estimate* them using the data we collected



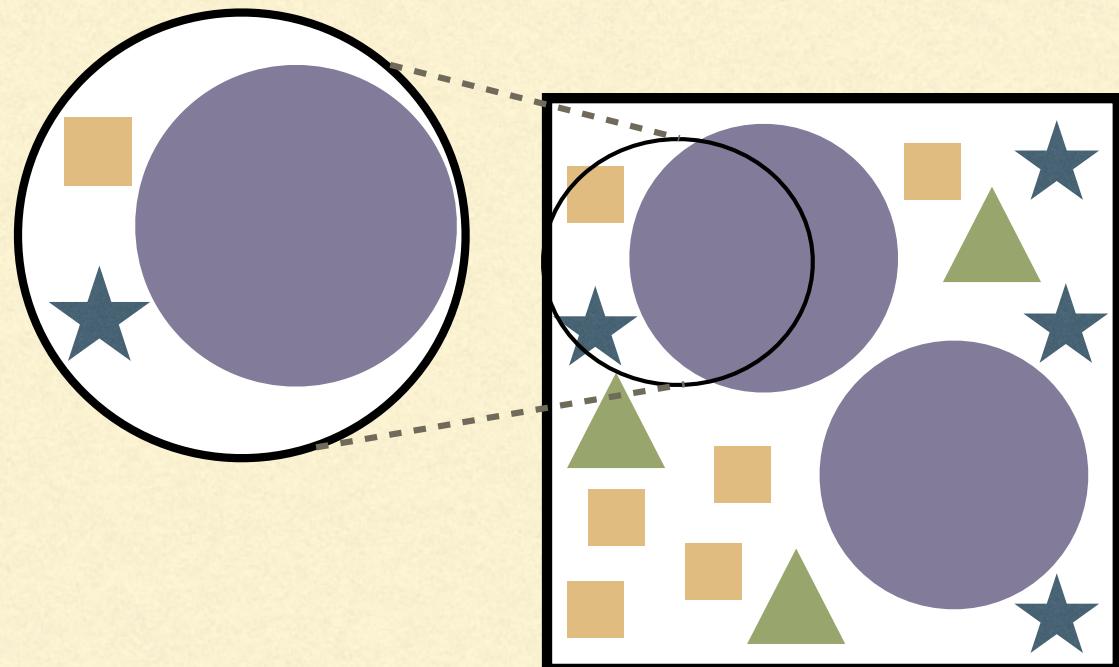
naive

THE "~~CLASSICAL~~" APPROACH

- Substitute the observed abundances $\hat{p}_1, \dots, \hat{p}_c$ for the unknown, true abundances p_1, p_2, \dots, 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



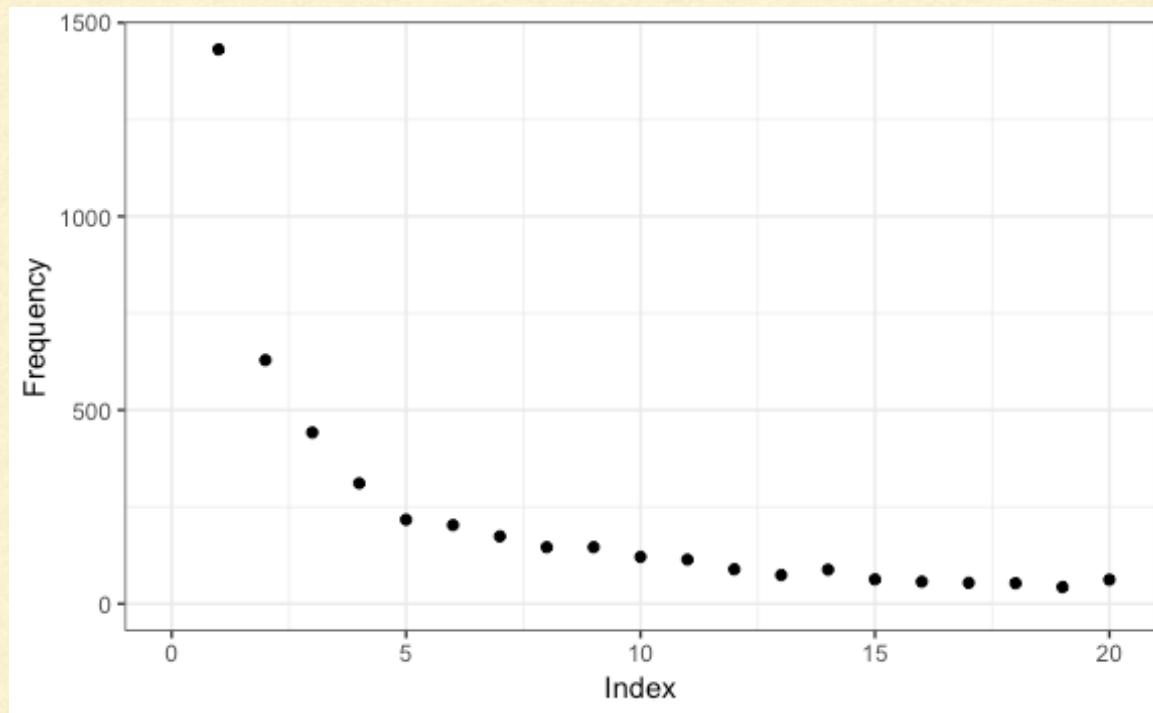
SPECIES RICHNESS ESTIMATION

- 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        60
```

SPECIES RICHNESS ESTIMATION

- Idea: extend the pattern in $f_1, f_2, f_3 \dots$ to f_0

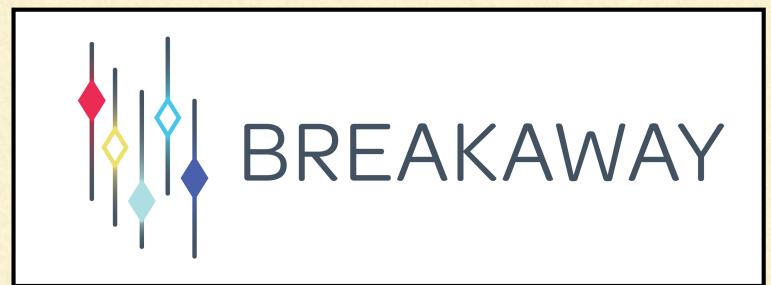


- Rare taxa are most informative for missing taxa

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

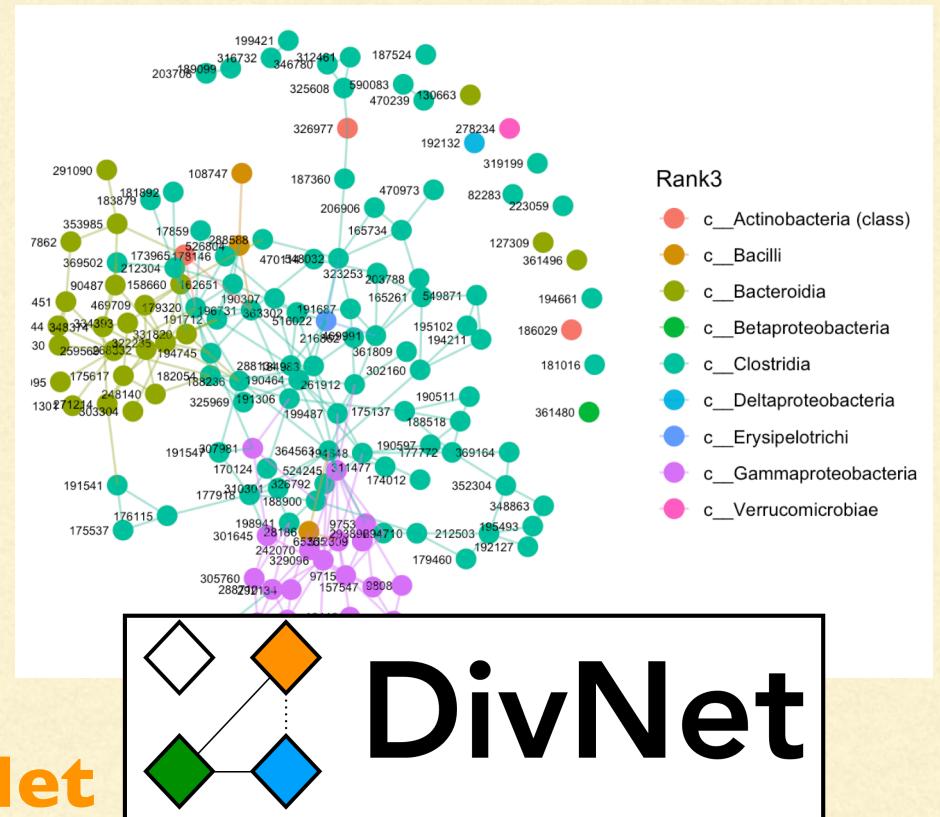
- “Chao I diversity index” is not an index — it's an estimate of species richness, and it's based on the questionable assumption that

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, \dots, p_C , then use network models to get variance



github.com/adw96/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
- Better idea: **Statistical estimation that accounts for different sequencing depths!**

BIAS AND DIVERSITY

- Alternative approach that I loathe: rarefaction

The screenshot shows a research article from PLOS Computational Biology. The header includes the PLOS logo, the journal name "COMPUTATIONAL BIOLOGY", and navigation links for "BROWSE", "PUBLISH", and "ABOUT". Below the header, the article is identified as an "OPEN ACCESS" and "PEER-REVIEWED" research article. The title of the article is "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible". The authors listed are Paul J. McMurdie and Susan Holmes. The article was published on April 3, 2014, with the DOI <https://doi.org/10.1371/journal.pcbi.1003531>.

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>

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

BIAS AND DIVERSITY

■ Alternative approach



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

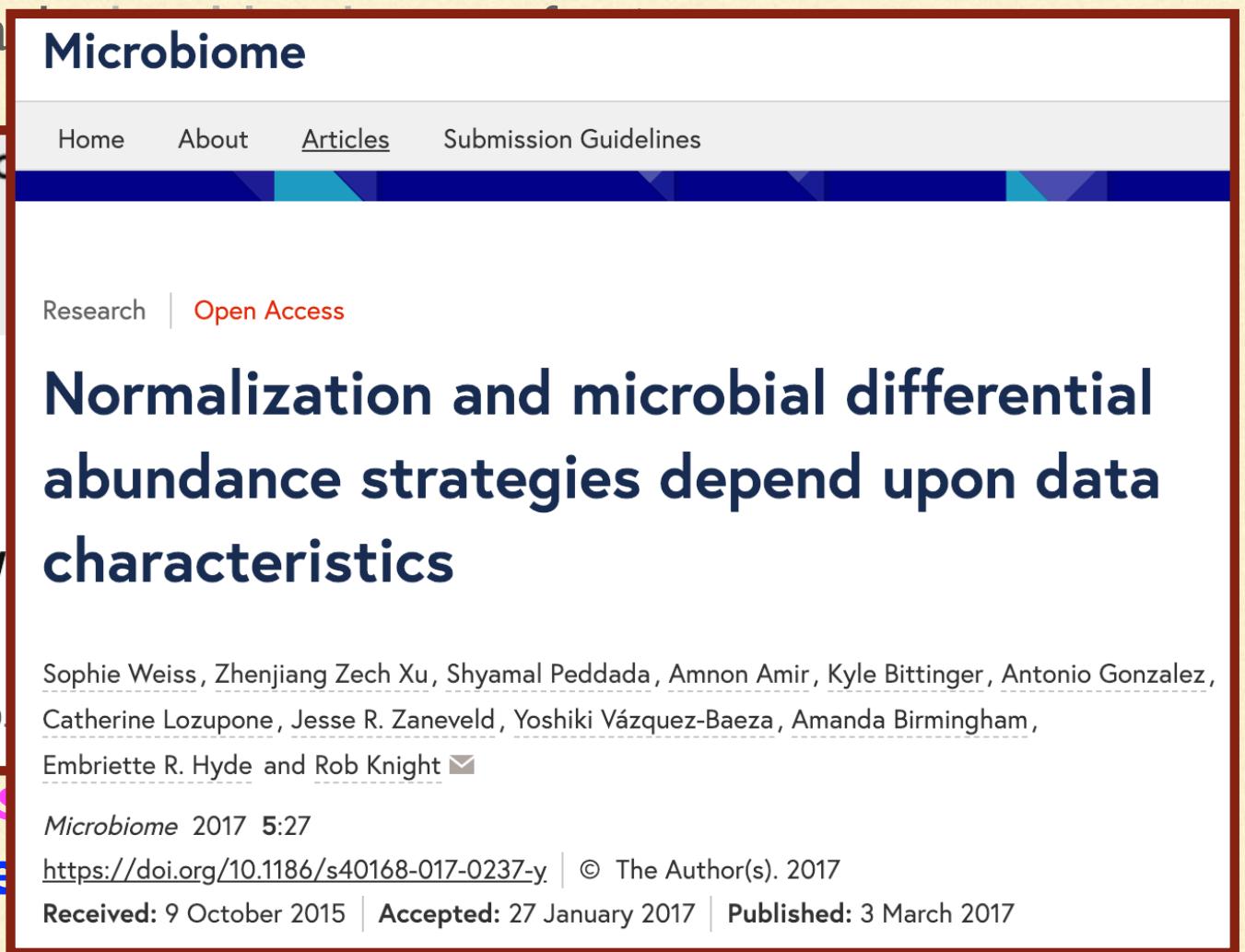
RESEARCH ARTICLE

Waste Not, Want Not: W

Paul J. McMurdie, Susan Holmes 

Published: April 3, 2014 • <https://doi.org/10.1371/journal.pcbi.100168>

■ Better idea: **Statistically different sequences**



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

BIAS AND DIVERSITY

■ Alternative approaches

The screenshot shows a research article page from the journal **PLOS COMPUTATIONAL BIOLOGY**. The article is titled **Microbiome** and is categorized as a **PERSPECTIVE** published on 23 October 2019 with doi: 10.3389/fmicb.2019.02407. The article is labeled as **Open Access**. The main title of the article is **Rarefaction, Alpha Diversity, and Statistics**, authored by **Amy D. Willis*** from the **Department of Biostatistics, University of Washington, Seattle, WA, United States**. The abstract discusses the statistical perspective on diversity, mentioning that diversity depends on sampling intensity. The page also features a sidebar for **frontiers in Microbiology** and a right sidebar with author names and a check-for-updates button.

frontiers
in Microbiology

PERSPECTIVE
published: 23 October 2019
doi: 10.3389/fmicb.2019.02407

Check for updates

Rarefaction, Alpha Diversity, and Statistics

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

differential
d upon data

DIVERSITY

- Very useful summary of (high-dimensional) compositional data... in many settings!
- A change in diversity: a useful *first question*



ACCESSING ‘DIVERSITY’ LAB

1. Go to schedule on Wiki to Sunday afternoon, click on “Labs”

2. Copy the command under the lab we’re working on

```
diversity lab:  
download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-  
afternoon/labs/diversity_lab/diversity-lab.R", "diversity-lab.R")
```

3. Run this command in your RStudio Server console



Get pumped!

```
Console Terminal × Jobs ×  
R 4.2.1 · ~/ ↗  
> download.file("https://raw.githubusercontent.com/statdivlab/stamps2022/main/Sunday-afternoon/labs/  
/diversity_lab/diversity-lab.R", "diversity-lab.R")
```

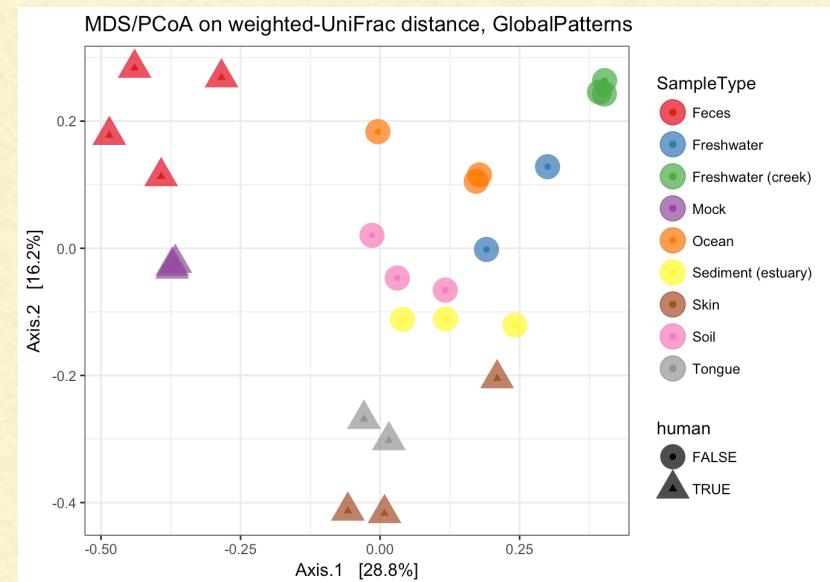
THOUGHTS ON BETA DIVERSITY

BETA DIVERSITY

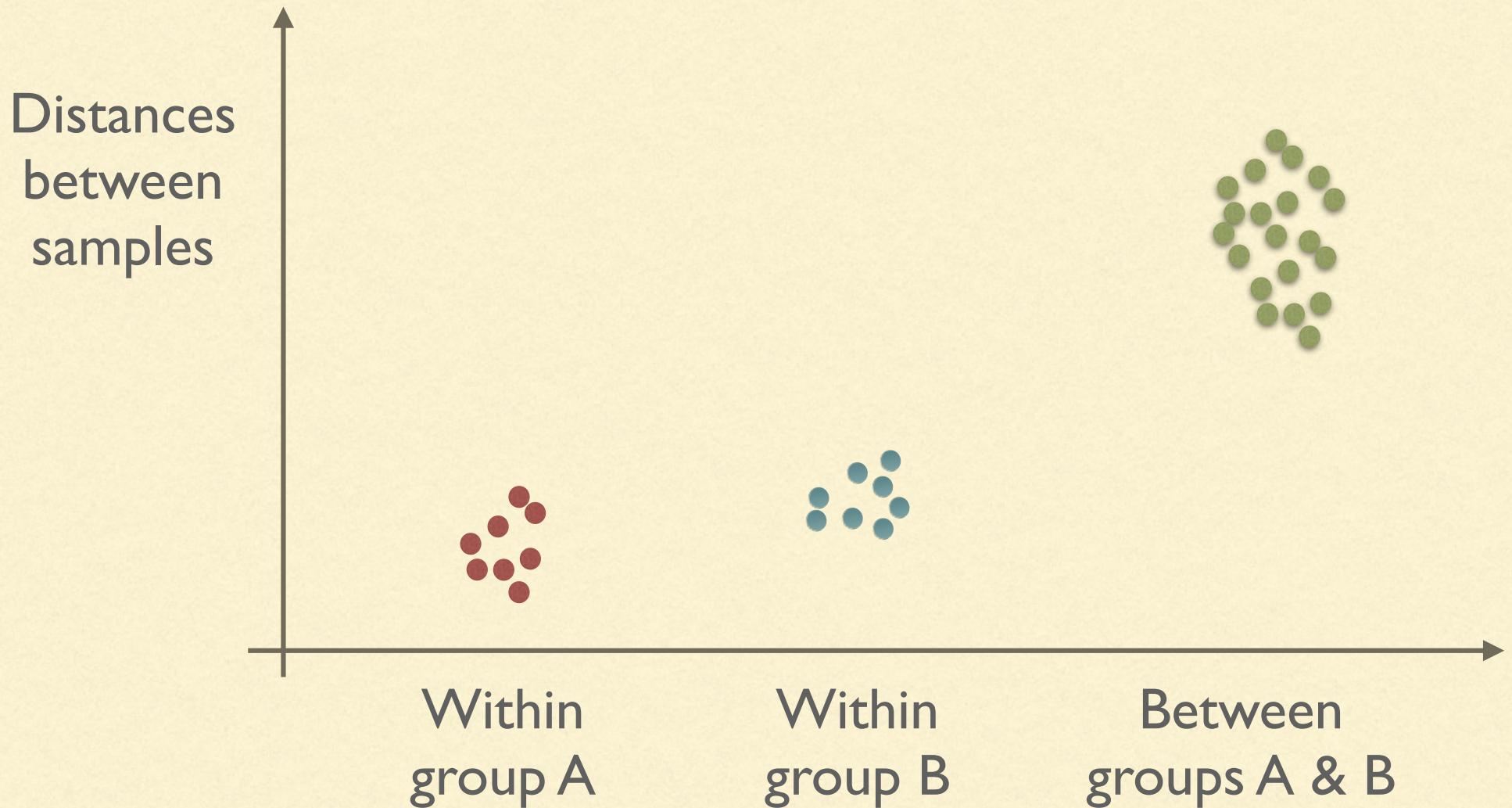
- Community 1: $p_1^{(1)}, p_2^{(1)}, \dots, p_c^{(1)}$; Community 2: $p_1^{(2)}, p_2^{(2)}, \dots, p_c^{(2)}$
- β -diversity parameters are usually distances between compositional vectors
- Bray-Curtis: $\beta_{BC} = 1 - \sum_{i=1}^C \min(p_i^{(1)}, p_i^{(2)})$
- Jaccard: $\beta_J = \% \text{ taxa not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



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)

CLOSING THOUGHTS

DIVERSITY

- If you really care about diversity, we 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

PEOPLE WORRY ABOUT THE WRONG THINGS

- Examples
 - Is data compositional?
 - Where to rarefy to?
 - Which beta diversity metric?

PEOPLE DON'T WORRY ABOUT THE RIGHT THINGS

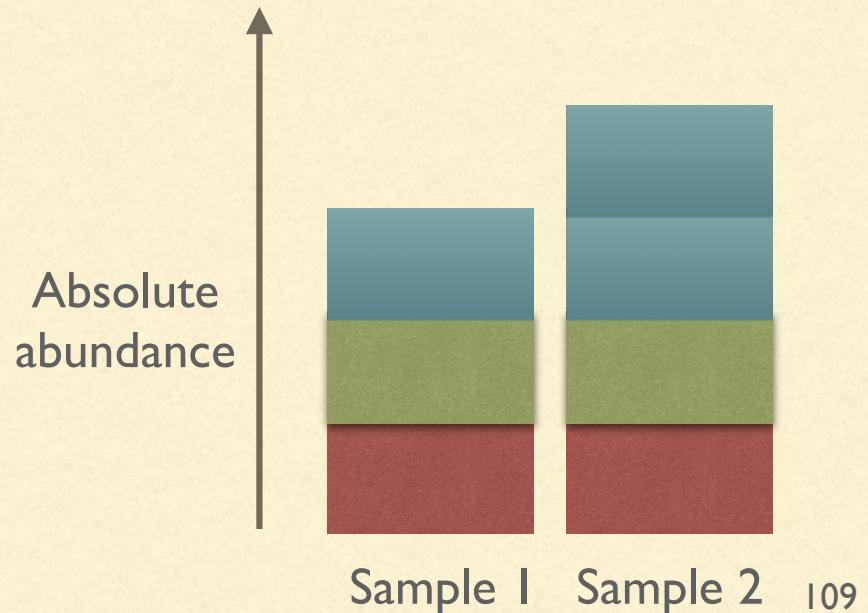
- Examples
 - Am I analyzing something I care about?
 - Can I estimate what I care about?

RELATIVE ABUNDANCE

- Limitations of relative abundance
 - Relative abundance of all taxa change when only one relative abundance changes

LIMITATIONS OF RELATIVE ABUNDANCE

- Relative abundance of all taxa change even when only one relative abundance changes
- Not “spurious” but misleading
- **0.33 / 0.33 / 0.33**
- **0.25 / 0.25 / 0.50**
- This is an inherent limitation of this type of analysis



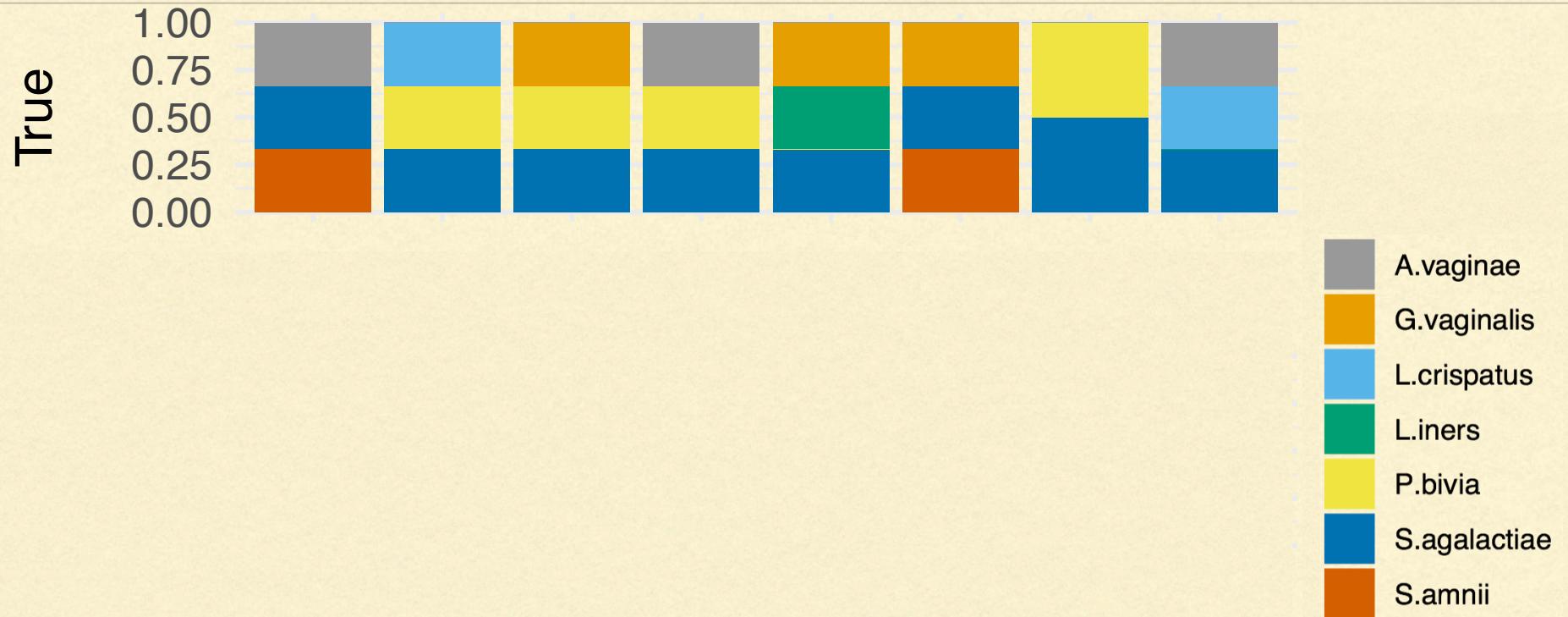
LIMITATIONS OF RELATIVE ABUNDANCE

- Can we even estimate true relative abundance from sequencing data?
- Does our sequencing technology give us information about relative abundance?
 - Is observed relative abundance proportional to actual relative abundance?

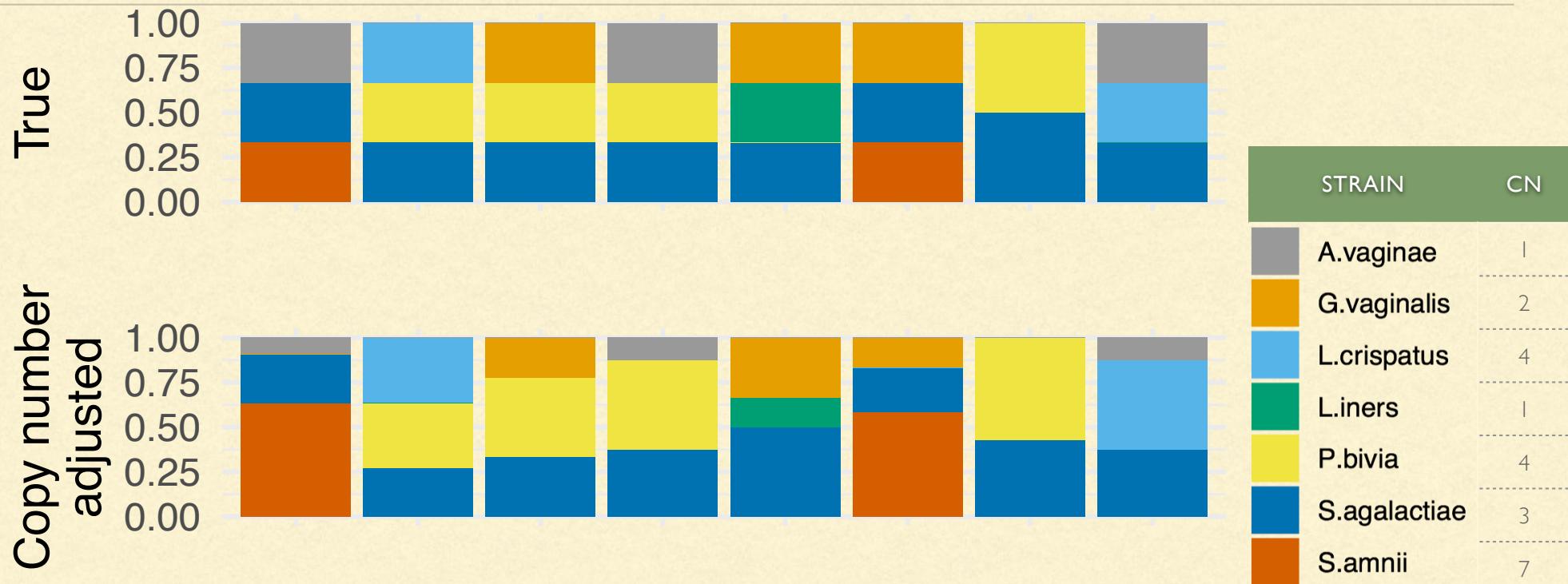
DECONVOLUTION OF BIOLOGY AND SEQUENCING

- Mock communities
 - Artificially constructed communities of known composition
 - Commonly used to benchmark sequencing and bioinformatics pipelines

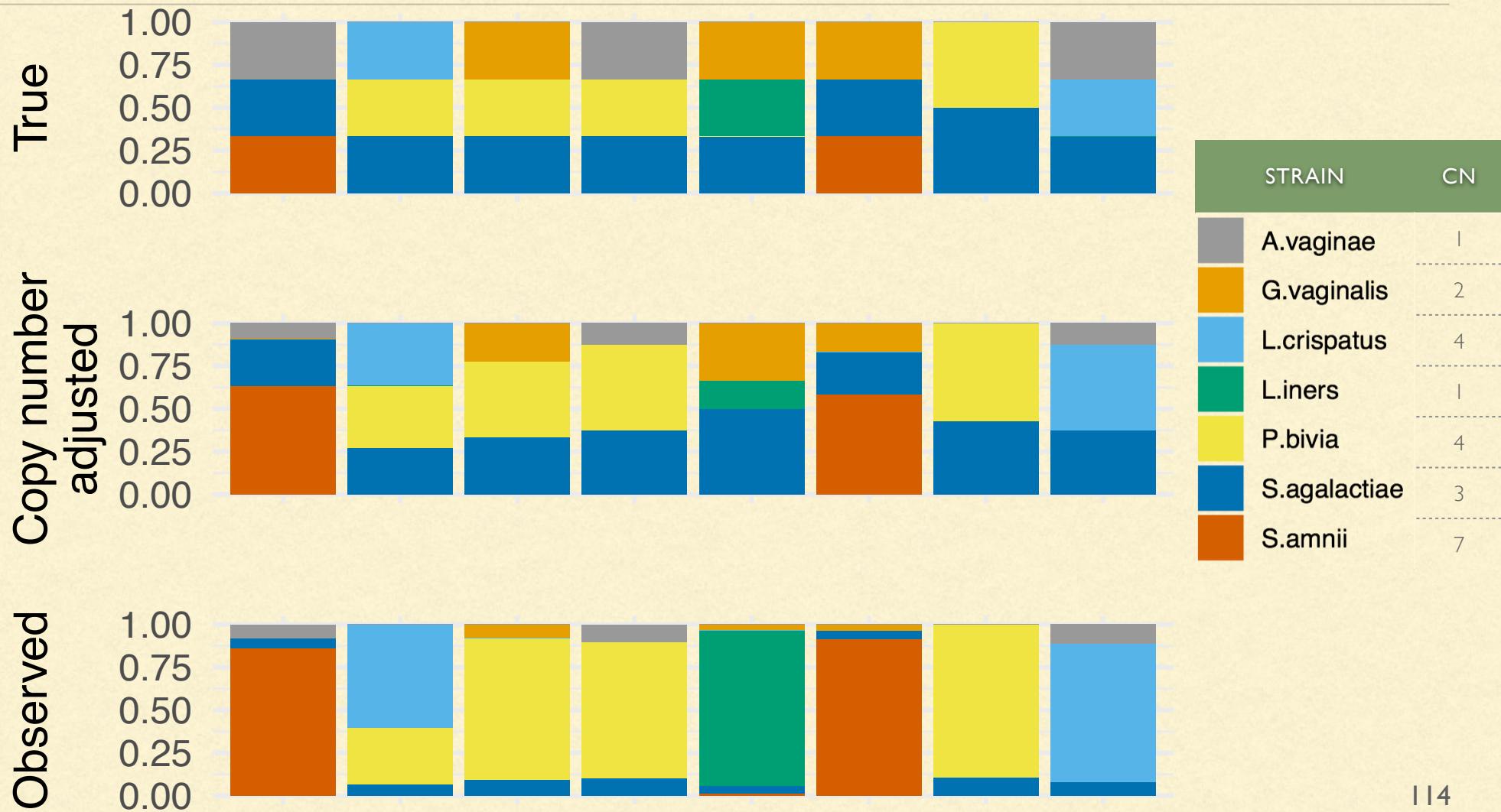
WHAT DO WE OBSERVE?



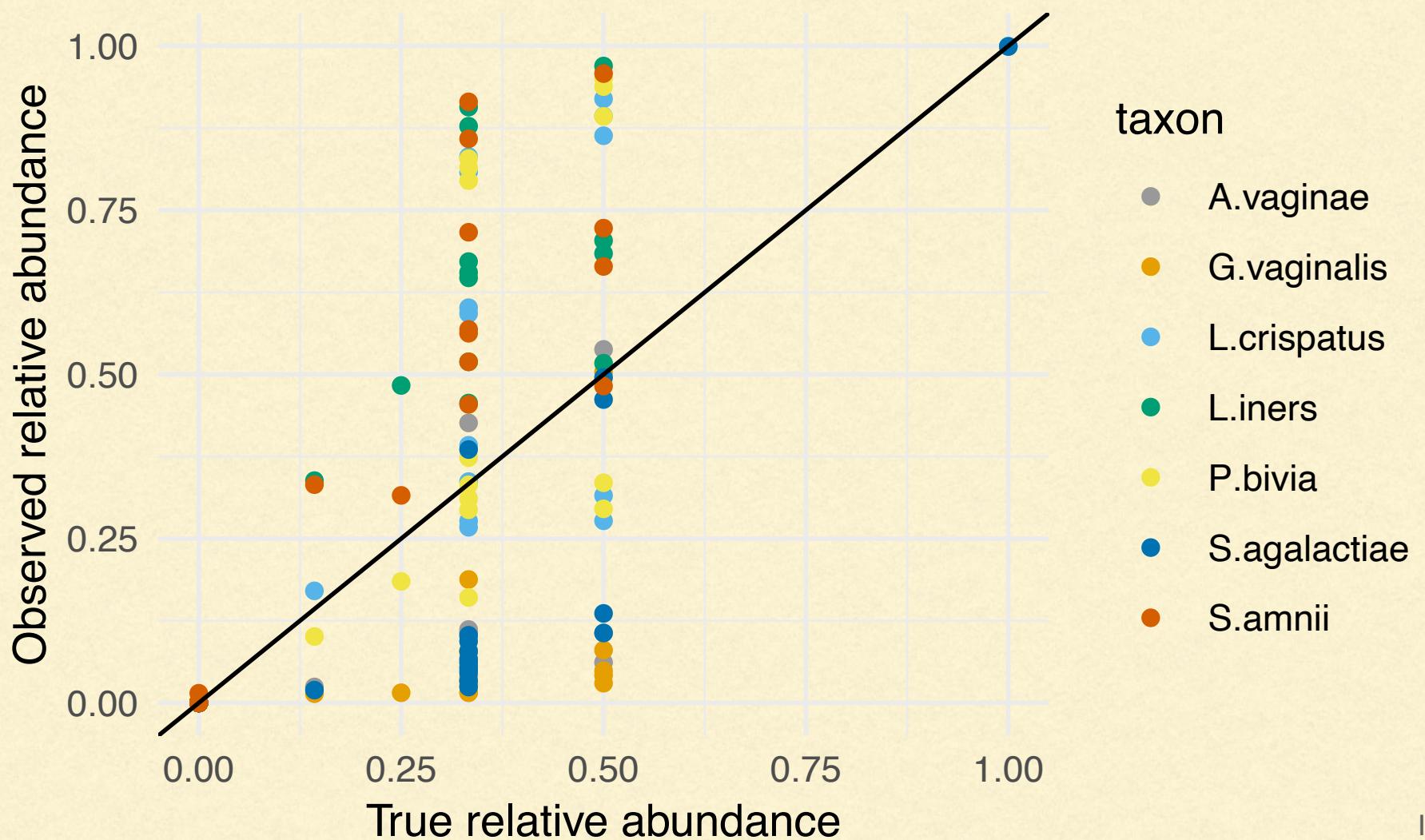
WHAT DO WE OBSERVE?



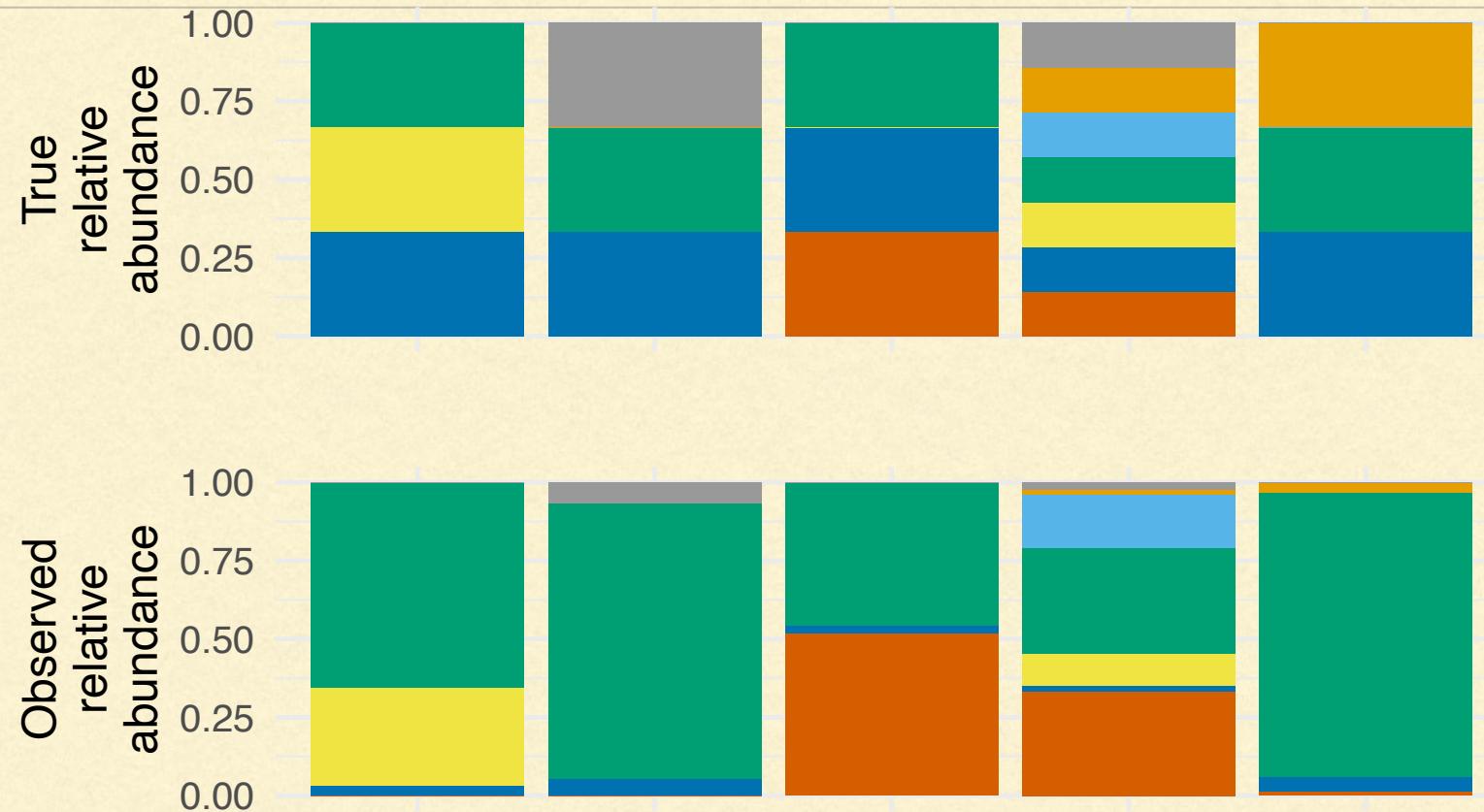
WHAT DO WE OBSERVE?



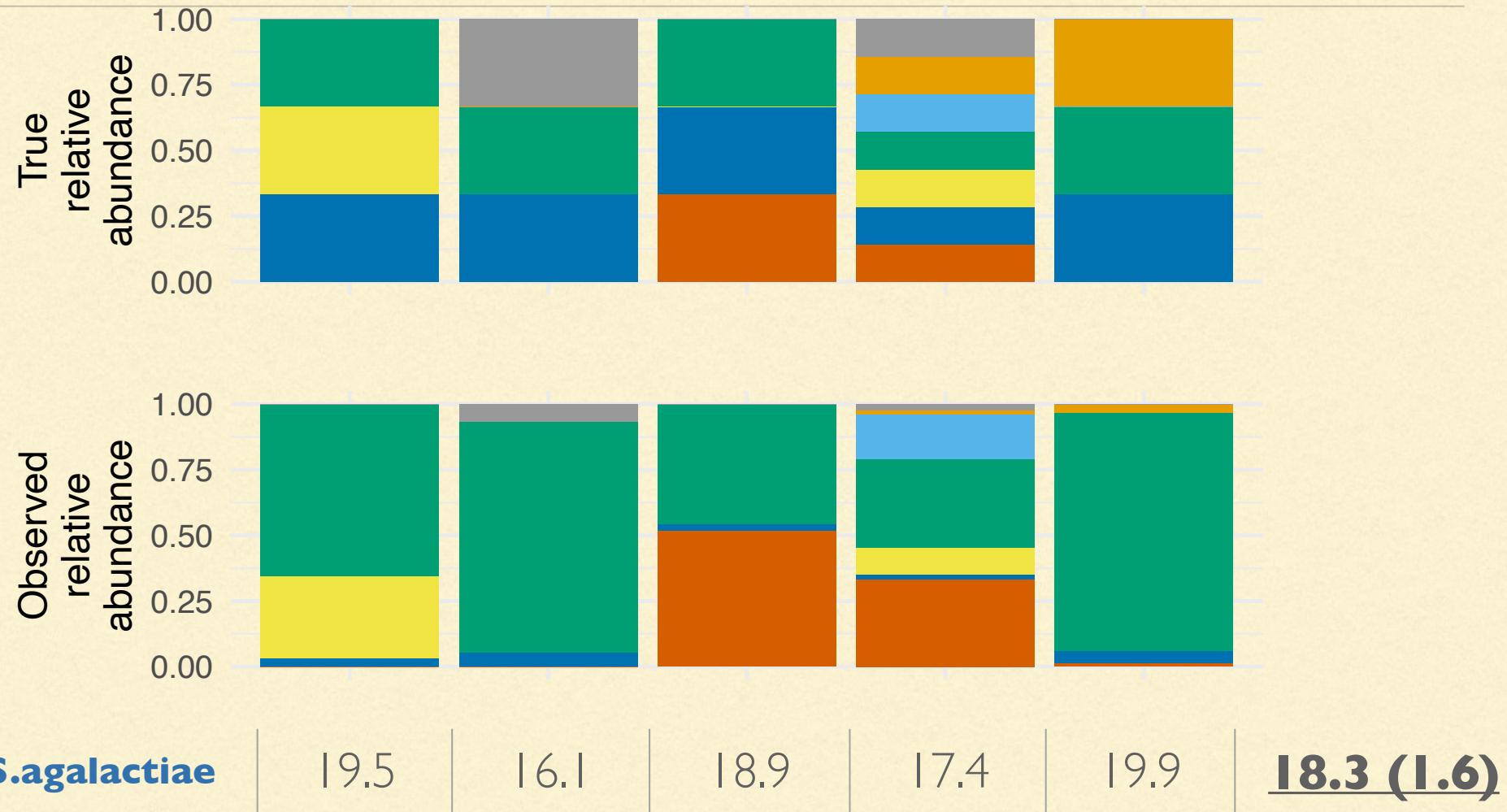
WHAT PATTERNS CAN WE FIND?

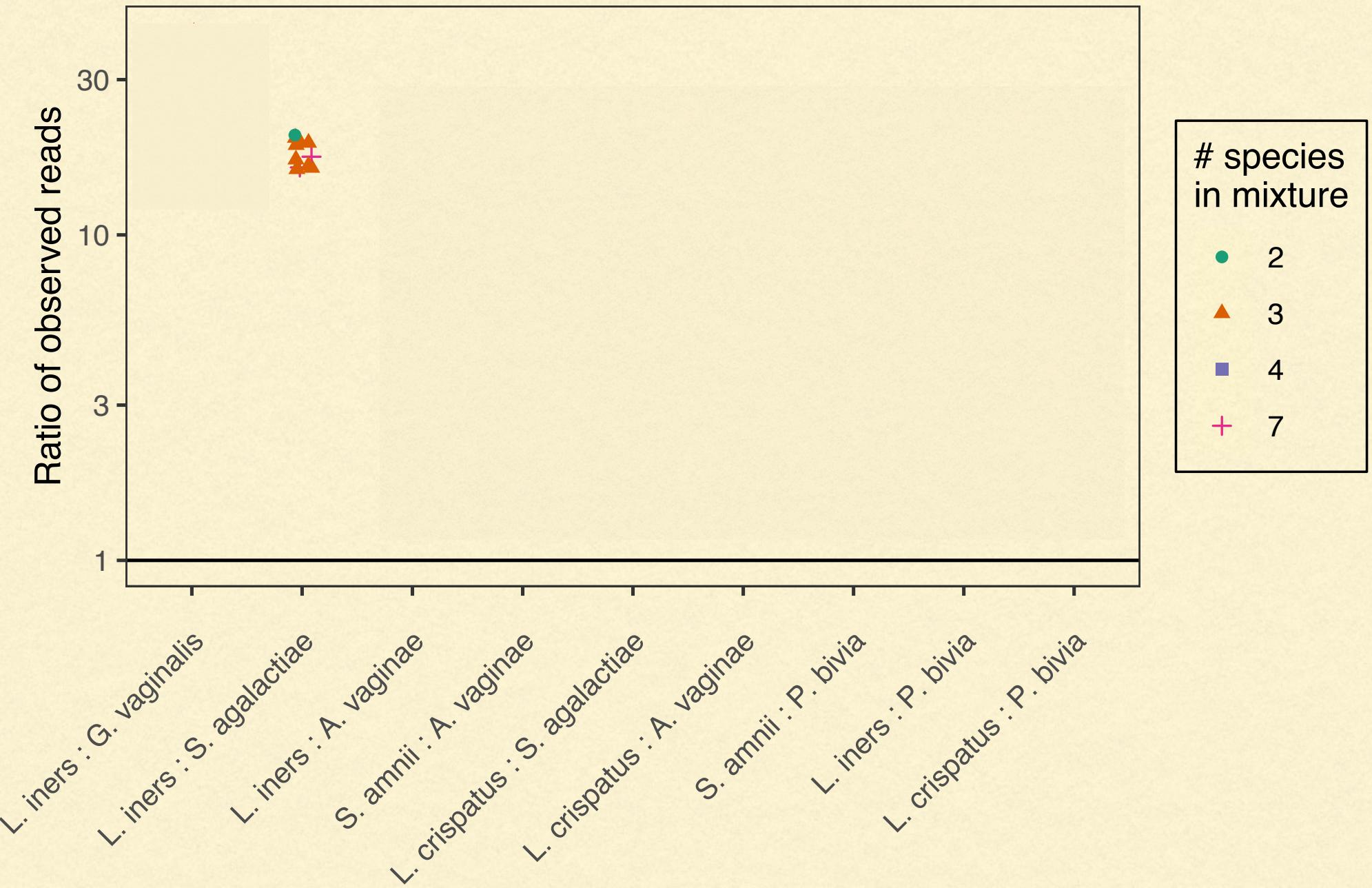


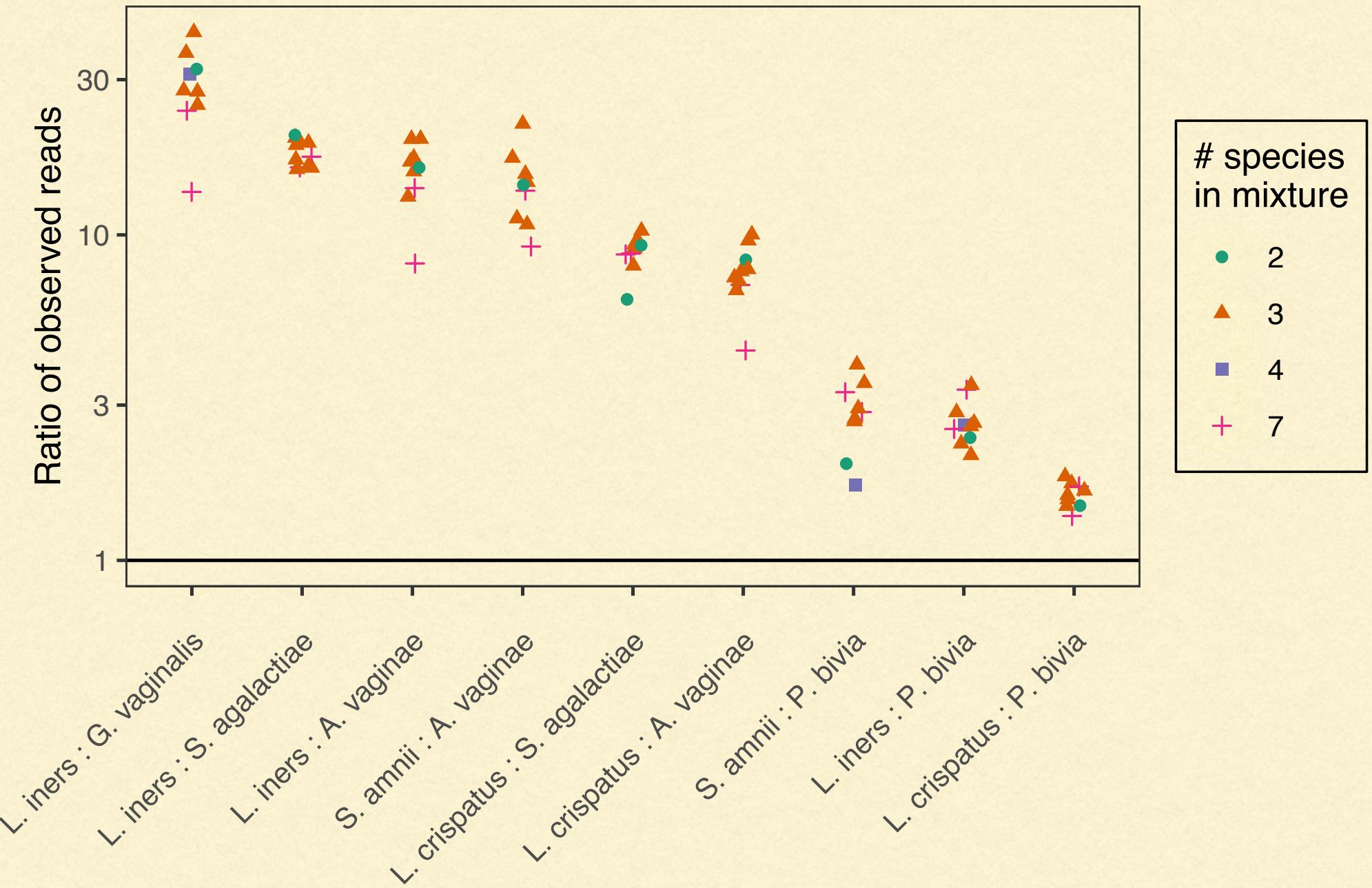
WHAT PATTERNS CAN WE FIND?



WHAT PATTERNS CAN WE FIND?







MODEL SPECIFICATION

- Strong support against

expected count_{*ij*} = true proportion_{*ij*} × sampling intensity_{*i*}

MODEL SPECIFICATION

- Strong support against

expected count_{ij} = true proportion_{ij} × sampling intensity_i

- Better support for

expected count_{ij} = true proportion_{ij} × sampling intensity_i
 × detection efficiency_j



eLIFE

elifesciences.org

RESEARCH ARTICLE



Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren¹, Amy D Willis², Benjamin J Callahan^{1,3*}

- Model for bias in observed abundances
 - Validated on 16S and shotgun metagenomic data
- Critical evaluation of model specification

Modeling complex measurement error in microbiome experiments

David S Clausen, Amy D Willis

DOI: 10.1111/biom.13503

BIOMETRIC PRACTICE

Biometrics WILEY
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

A multiview model for relative and absolute microbial abundances

Brian D. Williamson  | James P. Hughes  | Amy D. Willis 



Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren¹, Amy D Willis², Benjamin J Callahan^{1,3*}



David
Clausen
(UW)

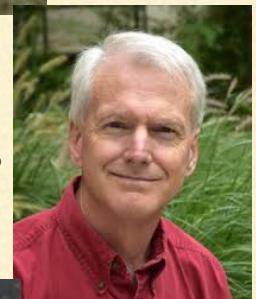
Ben
Callahan
(NCSU)



Michael
McLaren
(MIT)



Jim
Hughes
(UW)



Brian
Williamson
(Kaiser)



Evaluating replicability in microbiome data

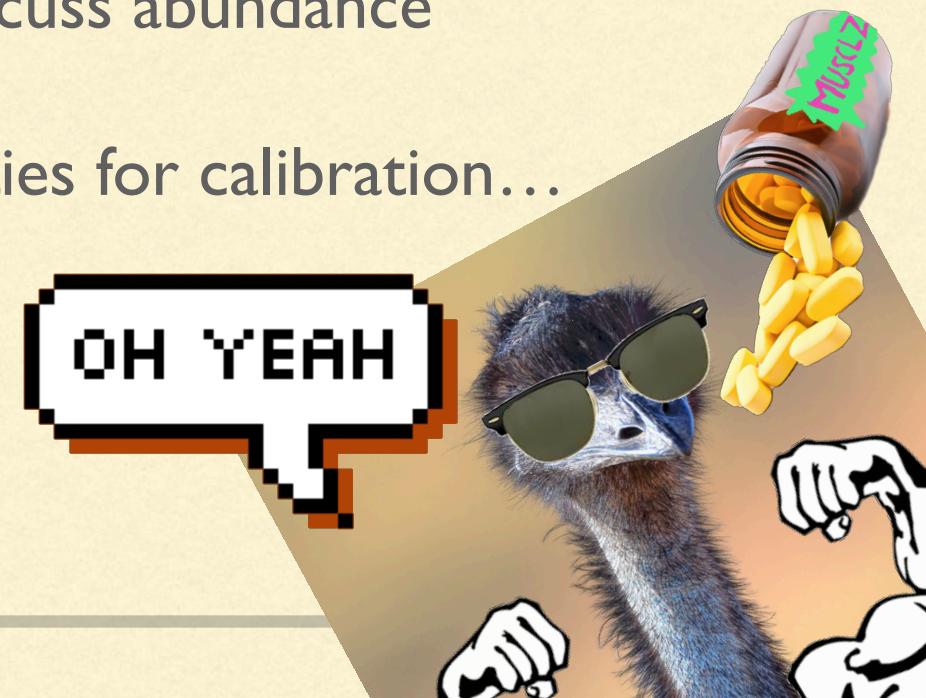
David S Clausen, Amy D Willis 

Biostatistics, kxab048, <https://doi.org/10.1093/biostatistics/kxab048>



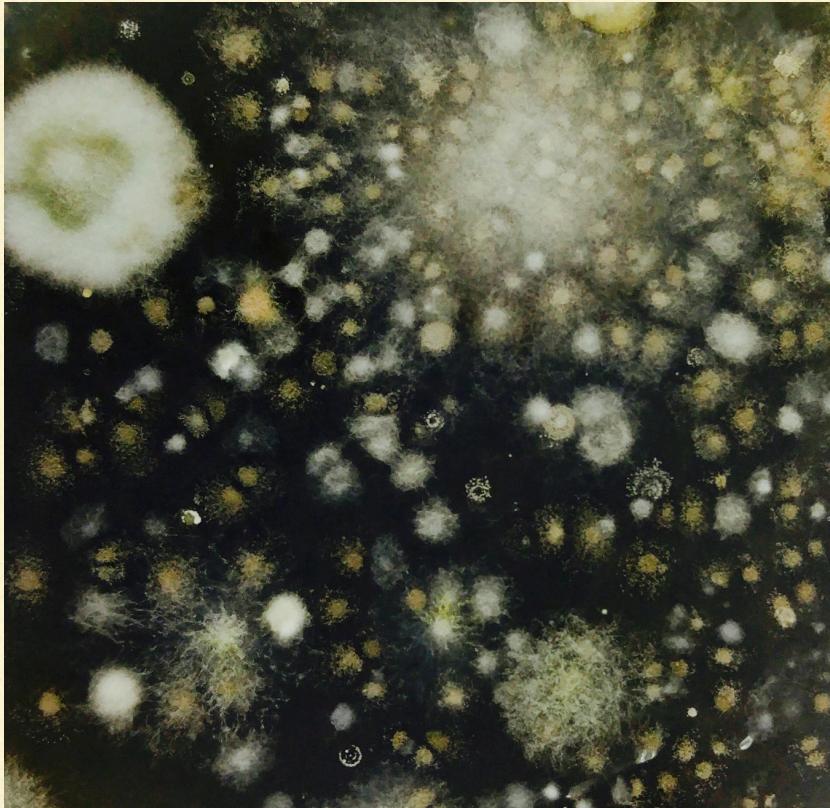
IMPLICATIONS

- Now we know this phenomenon exists, what can we do?
 - Cautious: model relative abundance, but understand limitations
 - Cynical: Only rely on qPCR to discuss abundance
 - Aspirational: Use mock communities for calibration...
 - Aware: model ratios
 - Progressive: radEmu



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



MODELING MICROBIOME DATA

Statistical Diversity Lab @ University of Washington

Amy Willis — [@AmyDWillis](#) — Assistant Professor

David Clausen — [@davidandacat](#) — PhD Candidate

Sarah Teichman — [@sarah_teichman](#) — PhD Candidate

126

CONSISTENCY OF EFFICIENCIES

STRAIN	GENOME SIZE (MBP)	COPY NUMBER	ESTIMATED EFFICIENCY
L CRISPATUS	2.04	4	2.03
L INERS	1.30	1	6.83

16S COPY NUMBER IS PREDICTIVE OF PCR BIAS BUT NOT OF TOTAL BIAS

