

# INTRODUCTION TO ESTIMATION

**Statistical Diversity Lab @ University of Washington**

@AmyDWillis — Assistant Professor

@BryanDMartin\_ — PhD Candidate

@paulinetrinh — PhD Candidate

@ReiterTaylor — Honorary member

|

---

“How do I rigorously analyse my data?”

*—Everyone, all the time*

---

“It depends.”

—*Stat Div Lab, all the time*

---

# THE PLAN

---

1. What can we estimate with compositional data?
2. Lecture: Relative abundance
3. Lab: Relative abundance for amplicon and shotgun data
4. Lecture: Diversity
5. Lab: Diversity (Pauline)
6. Lecture: Bias and calibration in relative abundance

# ANALYSIS

---

- The questions that you have affect how you do your analysis
- Do you care about...
  - broad scale community structure?
  - granular detail?
  - Both/not sure?

# ANALYSIS

---

- The questions that you have affect how you do your analysis
- Do you care about...
  - broad scale community structure? diversity analyses
  - granular detail? taxon abundance
  - Both/not sure?

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

# ANALYSIS

---

- Type of data you have changes the questions you *can* answer
  - 16S: **taxonomy, function, concentration/abundance**
  - Shotgun: **taxonomy, function, concentration/abundance**
  - qPCR: **taxonomy, function, concentration/abundance**

# ANALYSIS

---

- Type of data you have changes the approach you need
  - Absolute abundances, proportions, compositional counts...

# SCENARIO

ABSOLUTE  
DATA

e.g. from taxon-specific qPCR primers

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	4	5	18	27
ENVIRO 2	9	11	37	57

9

Can compare rows to rows, and columns to columns

# 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 = <b>0.168</b>	1/6 = <b>0.167</b>	3.99 / 6 = <b>0.665</b>	
ENVIRO 2	0.99 / 6 = <b>0.165</b>	0.99 / 6 = <b>0.165</b>	4.02 / 6 = <b>0.67</b>	10

Can compare rows to rows, and columns to columns

# SCENARIO

COMPOSITIONAL  
COUNTS  
e.g. 16S  
e.g. shotgun

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

Can compare rows

---

# 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

# 16S & SHOTGUN DATA ARE COMPOSITIONAL

---

- 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

# I6S & WGS DATA ARE COMPOSITIONAL

---

- We analyse compositional data differently than non-compositional data
- Common (users/software): convert to proportions
  - This is not necessary!
  - It loses information about precision!
  - Good statistical methods model precision

---

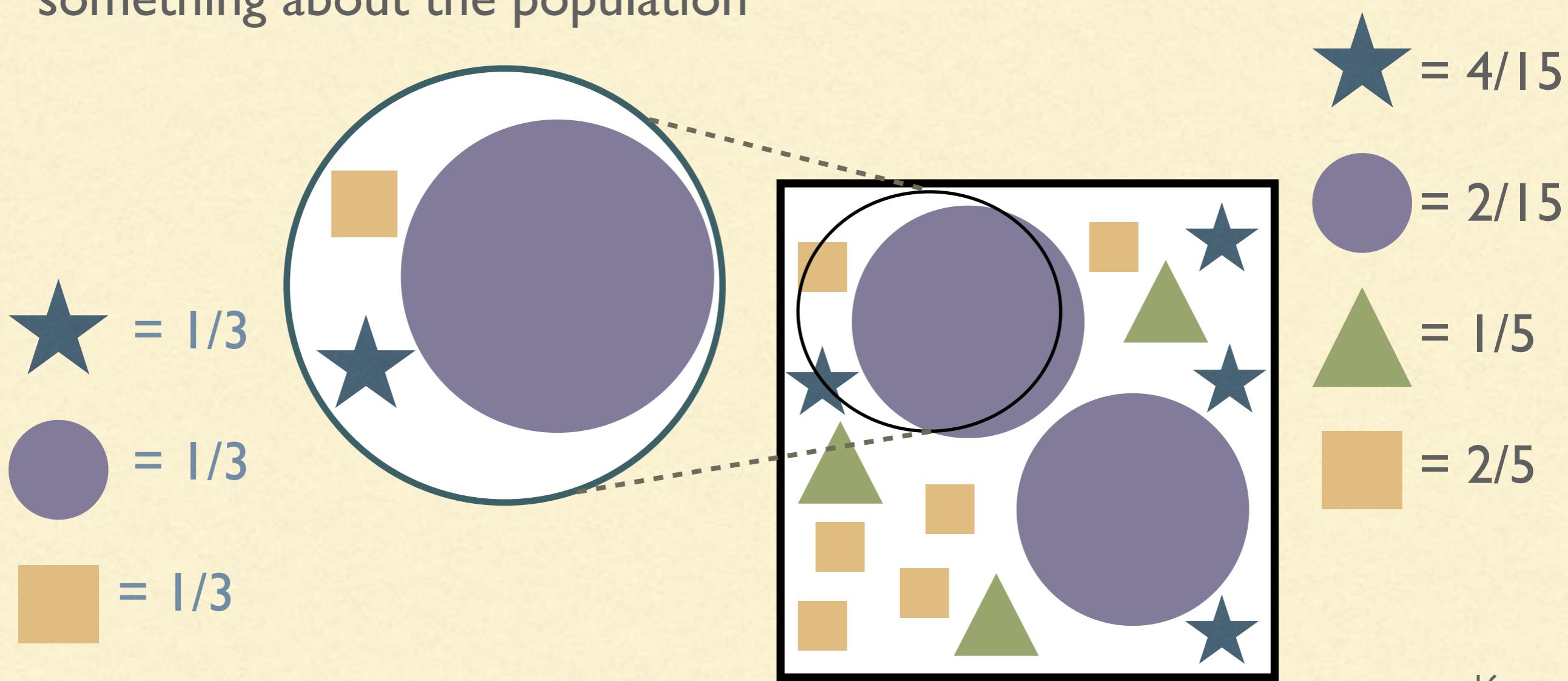
# I6S & WGS DATA ARE COMPOSITIONAL

---

- What are some interesting parameters when we have compositional data?

# PARAMETERS

- Estimation: using information about the sample to estimate something about the population



# COMPOSITIONAL DATA

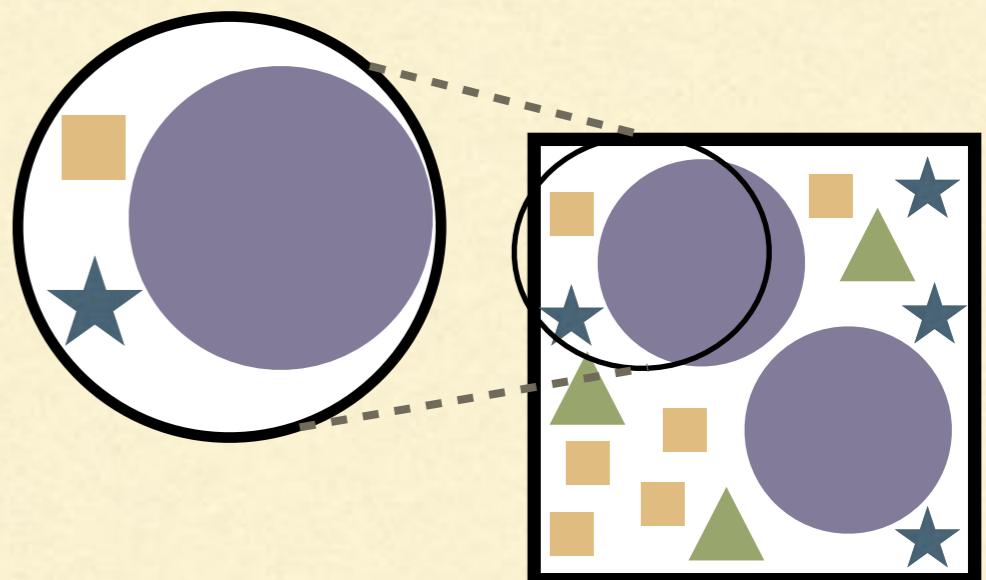
---

- A framework for compositional data:
- We have  $C$  groups in our environment
- Each group has some relative abundance  $p_1, p_2, \dots, p_c$
- $p_1 + p_2 + \dots + p_c = 1$

# THE PROBLEM

---

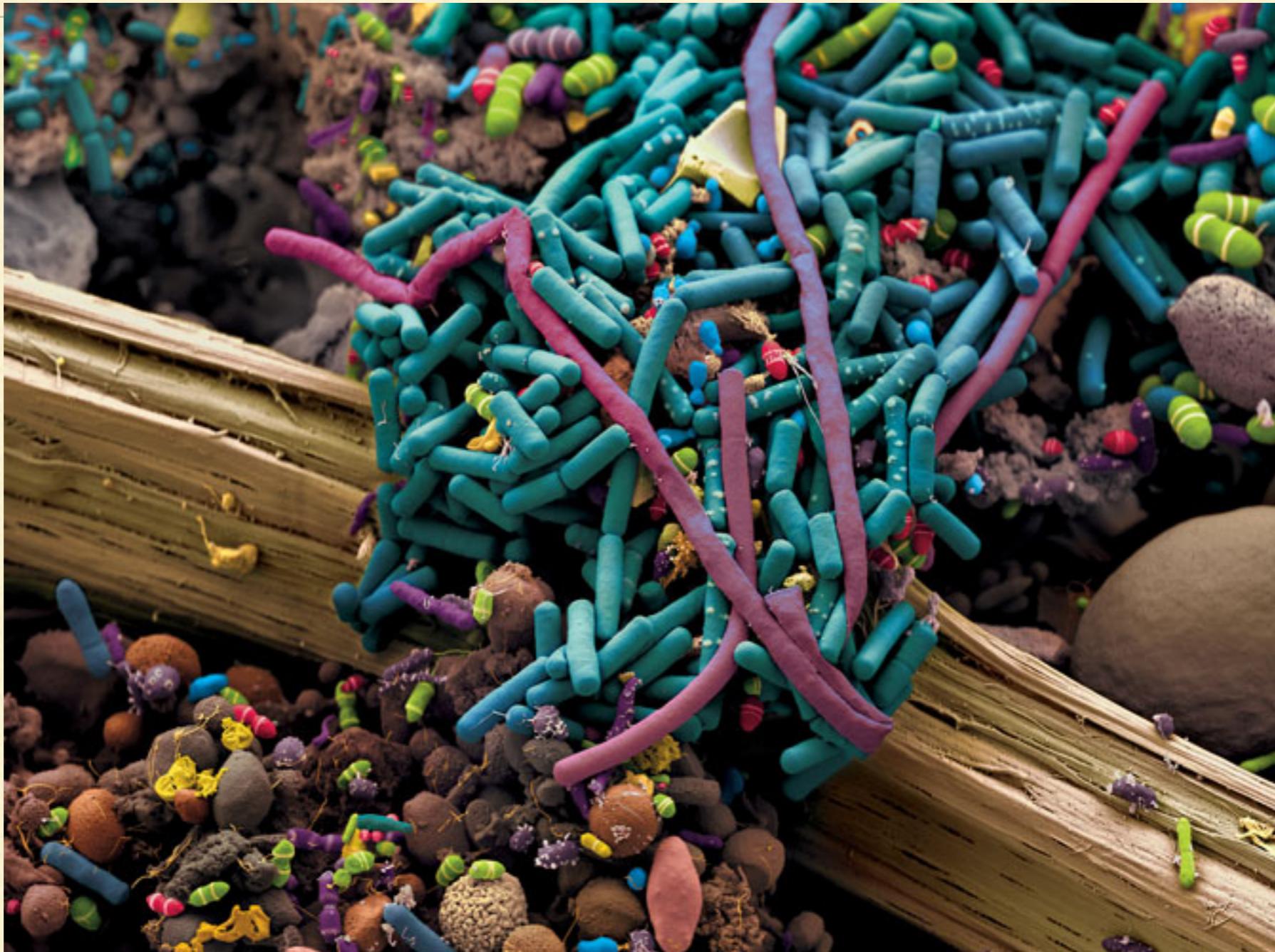
- In practice, we don't observe the entire community, just a sample from it
  - we don't know  $C$
  - We don't know  $p_1, p_2, \dots, p_c$
- **We need to estimate them using the data we collected**



# PARAMETERS FOR COMPOSITIONAL DATA

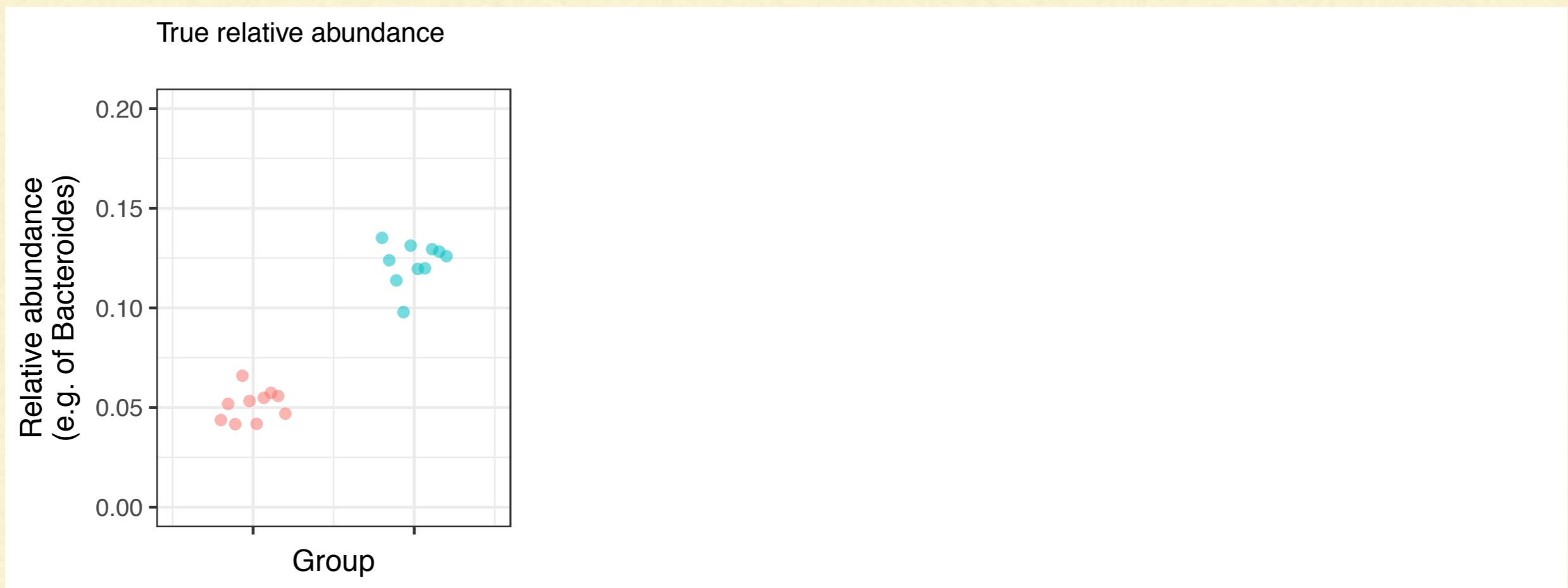
---

- Relative abundance of taxa/genes
- Diversity parameters:
  - $\alpha$ -diversity
  - $\beta$ -diversity
- Presence/absence of taxa/genes
- Abundance of taxon 1 divided by abundance of taxon 2...

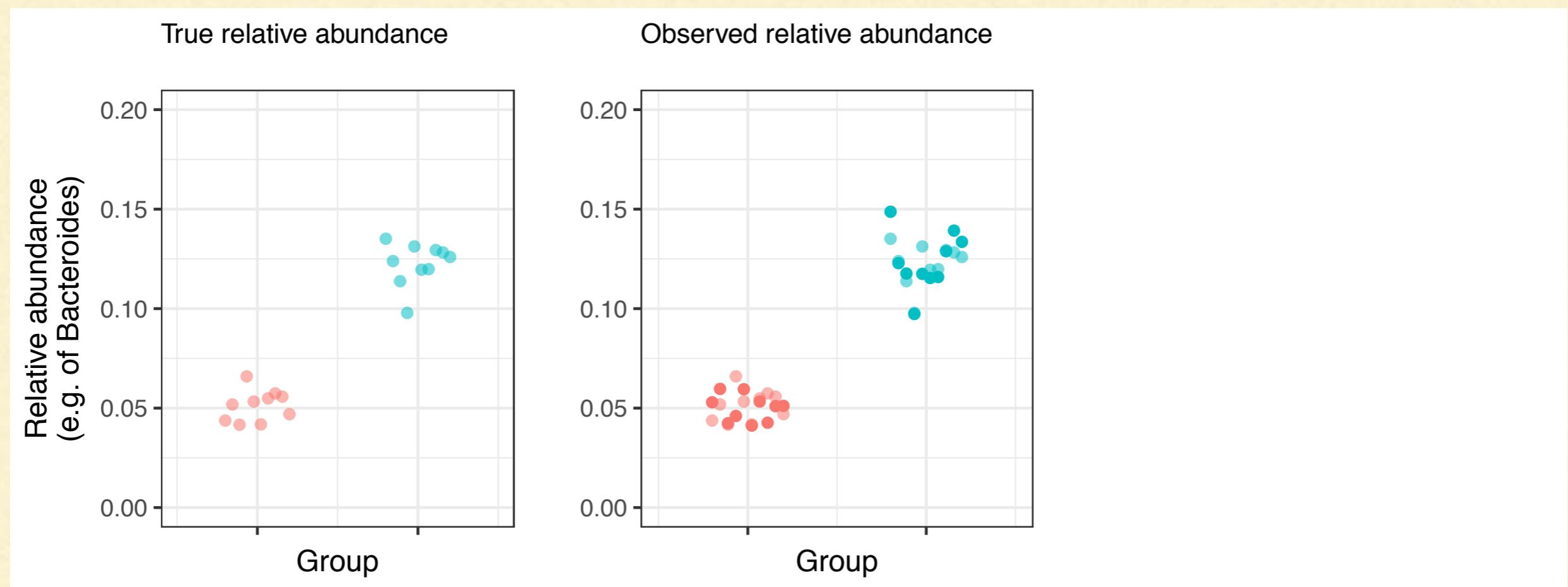


# MODELING RELATIVE ABUNDANCE

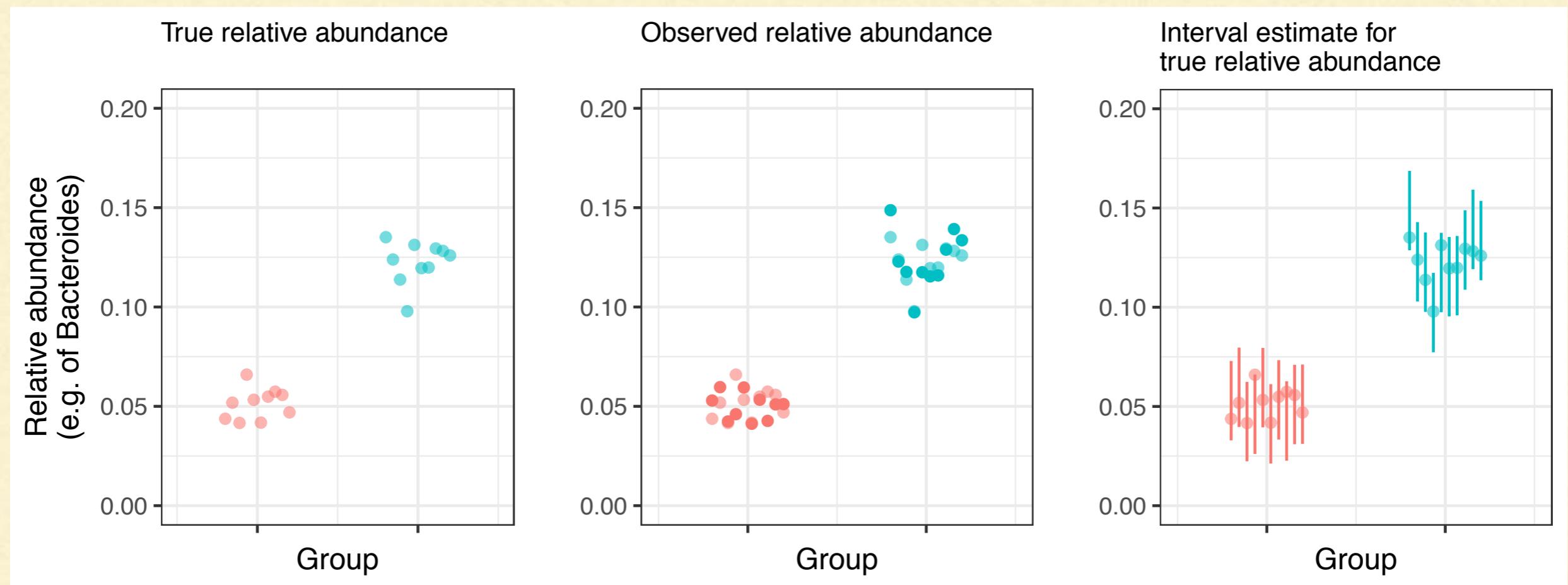
# SAMPLE VS POPULATION



# SAMPLE VS POPULATION

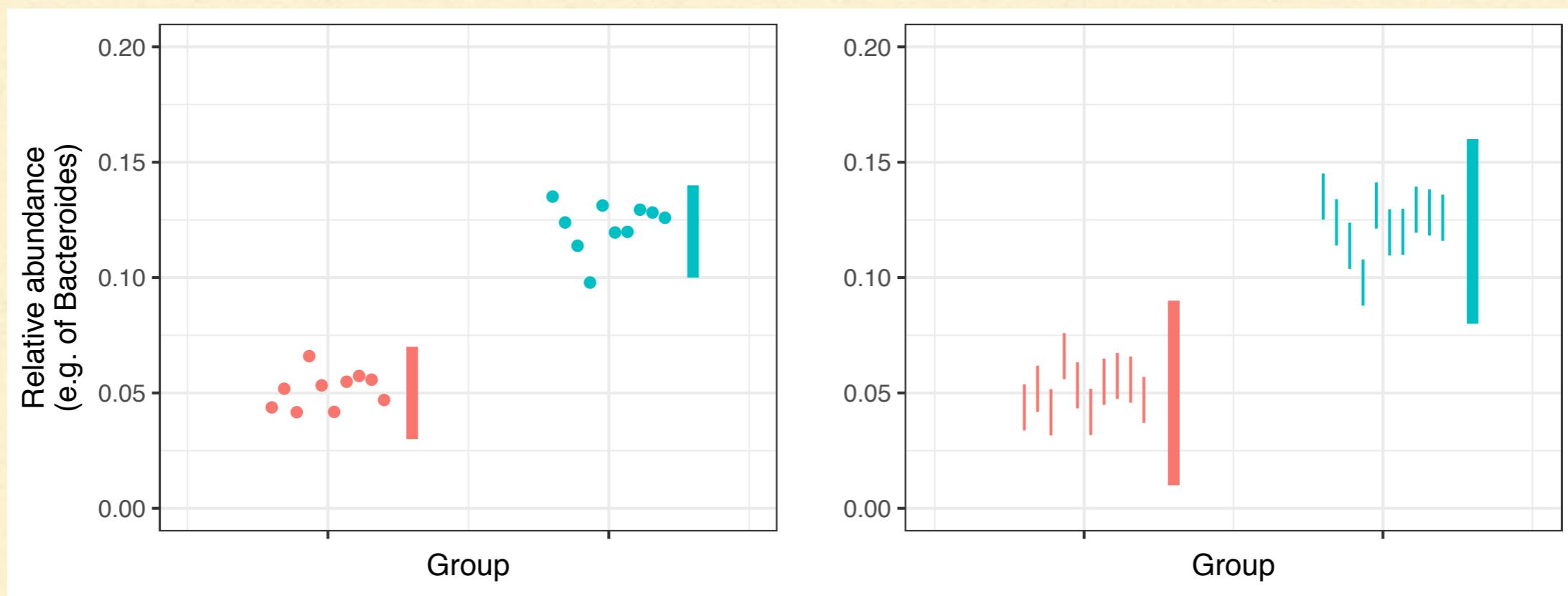


# SAMPLE VS POPULATION



# SAMPLE ≠ POPULATION

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



# CORNCOB

COmpositional RegressionN for Correlated Observations with the Beta-binomial



- Latent variable model for **relative abundance**
- Hypothesis testing for changes in
  - relative abundance
  - variance in abundance
- All taxa/genes!
- Multiple testing corrections with FDR control



Bryan Martin, UW Statistics



Daniela Witten, UW Statistics

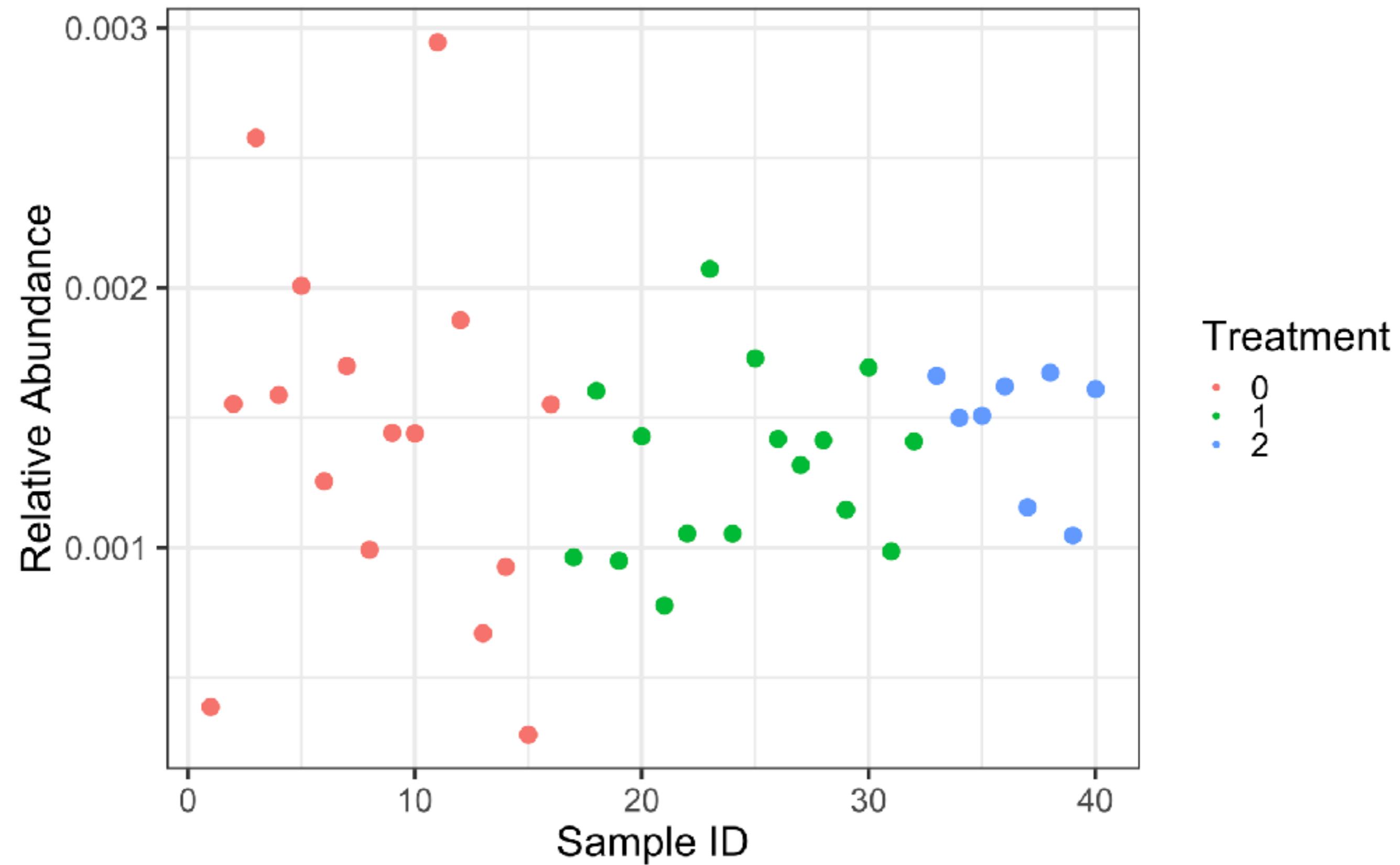
# CORNCOB

COmpositional RegressionN for Correlated Observations with the Beta-binomial

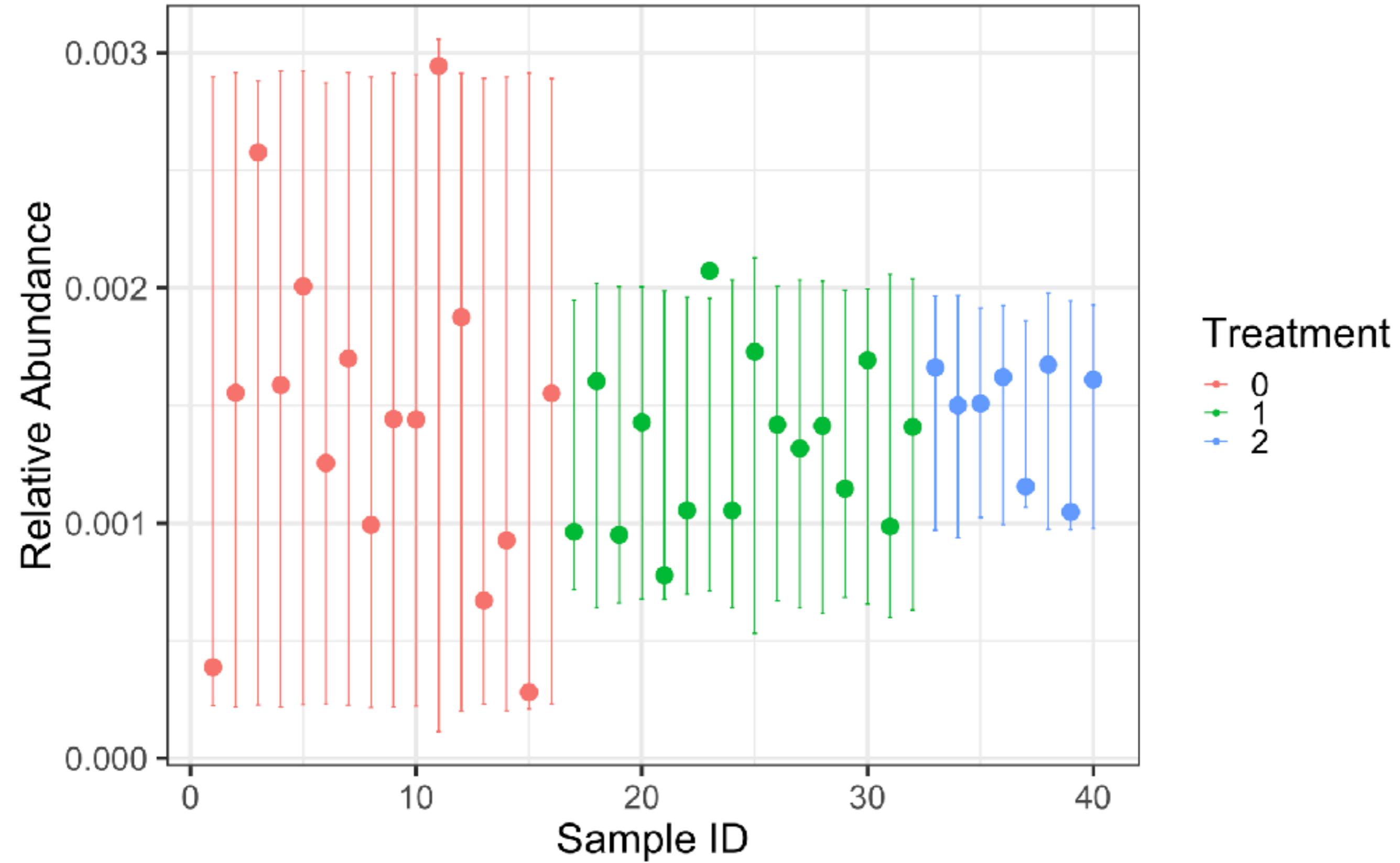


- “The relative abundance of *S. aureus* significantly decreased in the treatment group (95% CI for  $\beta_{treatment}$ : (-4.74, -2.91), FDR-adjusted p = 0.003, see Methods), controlling for the effect of treatment group on the dispersion”
- “Methods: All phylum-level relative abundances were modeled using 🌽corncob🌽 (emojis required) with a logit-links 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. ...”

# Soil by Fertilizer Treatment



# Soil by Fertilizer Treatment



# CORNCOB

COmpositional RegressionN for Correlated Observations with the Beta-binomial



- Addresses measurement error issue
- Adjusts for different library sizes (sequencing depth)
- Suitable for longitudinal/case-control/cross-sectional studies
- Suitable with multiple covariates
- Mean and variance testing

# CORNCOB

COmpositional RegressionN for Correlated Observations with the Beta-binomial



- Easy diagnosis of model misspecification
- R package, friendly with phyloseq
- Fast
- Documentation, tutorials, active support!

[github.com/bryandmartin/corncob](https://github.com/bryandmartin/corncob)

# POSSIBLE USES: WE WANT TO MODEL....

---

- The number of amplicon reads that you give a specific taxonomic label (e.g. *L. iners*), out of the total number of amplicon reads
- Shotgun reads that are recruited to a gene in a MAG, out of all shotgun reads that recruit to the MAG
- Reads that map to a gene in a metagenome out of the total number of reads

# POSSIBLE USES: WE WANT TO MODEL....

---

- The number of amplicon reads that you give a specific taxonomic label (e.g. *L. iners*), out of the total number of amplicon reads
- Shotgun reads that are recruited to a gene in a MAG, out of all shotgun reads that recruit to the MAG
- Reads that map to a gene in a metagenome out of the total number of reads

**You need to decide what numerator and denominator you care about**

# INPUT DATA

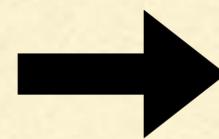
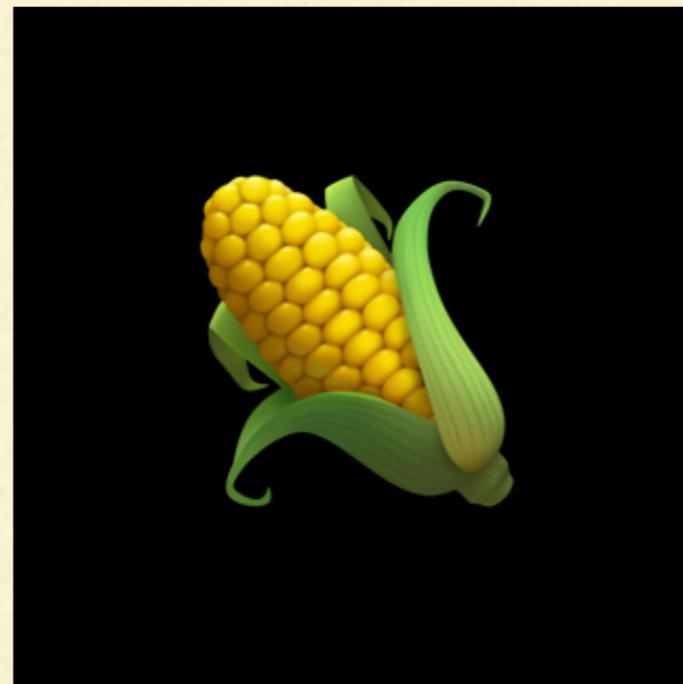
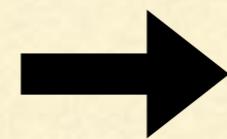
---

- Something out of some total
- For many samples...
  - The total number of “things” you saw
  - The number that “come from” “what you care about”
  - Information about the samples that may change relative abundance

# CORNCOB



Abundance table  
+  
Sample data  
(e.g. disease status, diet,  
treatment, BMI, age, ...)



1. List of differentially abundant taxa (with p-values)
2. List of differentially variable taxa (with p-values)

---

HOW DOES  WORK?

# BETA-BINOMIAL DISTRIBUTION

---

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

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

- $\mathbf{n}$  = samples, indexed by  $i = 1, \dots, n$
- $\mathbf{W}_i$  = # of individuals observed in the taxon/gene of interest
- $\mathbf{M}_i$  = total # of individuals observed
- $\mathbf{Z}_i$  = the (latent) relative abundance in sample  $i$

# BETA-BINOMIAL DISTRIBUTION

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

what you need

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

- $n$  = samples, indexed by  $i = 1, \dots, n$
- $\mathbf{W}_i$  = # of individuals observed in the taxon/gene of interest
- $M_i$  = total # of individuals 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

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

$\phi_i$  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

$\mu_i$  is a function of  $\boxed{\mathbf{X}_i}$   $\beta$

$\phi_i$  is a function of  $\boxed{\mathbf{X}_i^*}$   $\beta^*$

# DETAILS

## Modeling microbial abundances and dysbiosis with beta-binomial regression

Bryan D. Martin

*Department of Statistics  
University of Washington  
Seattle, Washington 98195  
e-mail: [bmartin6@uw.edu](mailto:bmartin6@uw.edu)*

Daniela Witten

*Departments of Statistics and Biostatistics  
University of Washington  
Seattle, Washington 98195  
e-mail: [dwitten@uw.edu](mailto:dwitten@uw.edu)*

Amy D. Willis

*Department of Biostatistics  
University of Washington  
Seattle, Washington 98195  
e-mail: [adwillis@uw.edu](mailto:adwillis@uw.edu)*

**Abstract:** Using a sample from a population to estimate the proportion of the population with a certain category label is a broadly important problem. In the context of microbiome studies, this problem arises when researchers wish to use a sample from a population of microbes to estimate the population proportion of a particular taxon, known as the taxon's *relative abundance*. In this paper, we propose a beta-binomial model for this

# Hypothesis Testing

$$H_0 : \beta = 0$$

- Does the healthy group have a different mean relative abundance of *L. iners*?
- Is a high-fat diet associated with changes in the relative abundance of *Firmicutes*?

$$H_0 : \beta^* = 0$$

- Is disease associated with a change in the variability of *L. iners*?
- Is a high-fat diet associated with changes in the stability of *Firmicutes*?

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

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	-31.4e-16	***
Day1	0.35077	0.03807	9.2e-15	***
Day2	0.1456	0.02389	6.08	***
Amdmt1	0.0354	0.02436	1.58	
Amdmt2	0.19374	0.0412	4.66e-06	***
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Coefficients associated with phi:

	Estimate	S.E.	t value	Pr(> t )
(Intercept)	-5.6957	0.001	-5.6957	*
Day1	1.1279	0.001	1.1279	
Day2	-0.7231	0.001	-0.7231	
---				
Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’
	0.1 ‘ ’	1		

Log-likelihood: -1056.1

# CORNCOB

COmpositional RegressionN for Correlated Observations with the Beta-binomial



- Coefficients are (by default) on the logit relative abundance scale
- Negative coefficients => decreased abundance
- Usually we are interested in testing for differential abundance, controlling for the effect of the covariate on variability

# CORNCOB

COmpositional RegressionN for Correlated Observations with the Beta-binomial



- CORNCOB does not assume microbes behave independently
- Parameter  $\phi$  controls cooccurrence of taxa of the same group
- CORNCOB reflects structure in microbial communities
  - Urn model interpretation of microbial reproduction... ask Bryan later if you want to know more. *Very cool!*

# CORNCOB AND DESEQ2

---

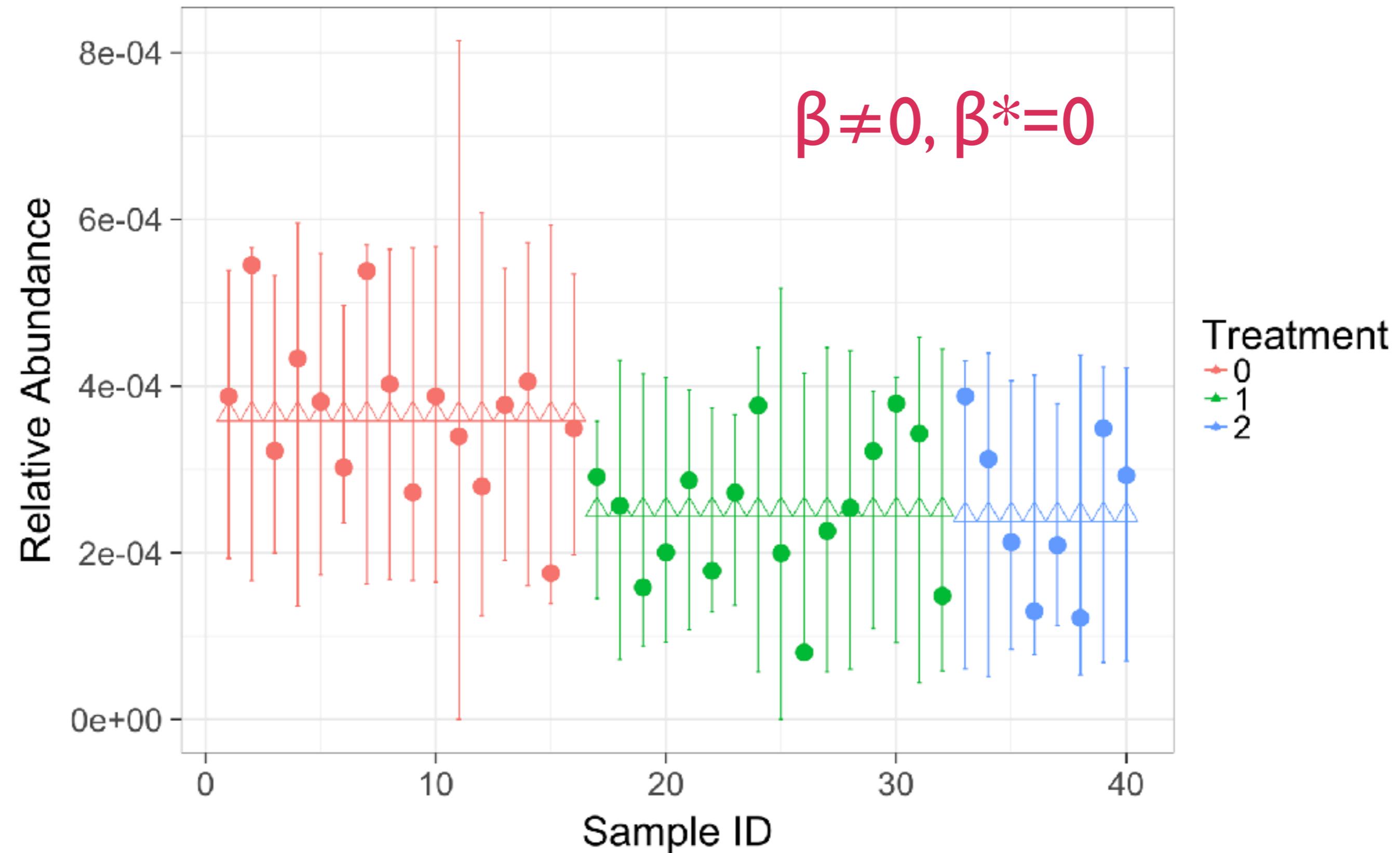
## ■ Similarities

- Easy to use with phyloseq (easier, in fact)
- Use un-normalized counts to assess precision of estimates
- Benjamini-Hochberg adjustment for multiple comparisons
- Dispersion parameter for overdispersion
- Tests for differential abundance

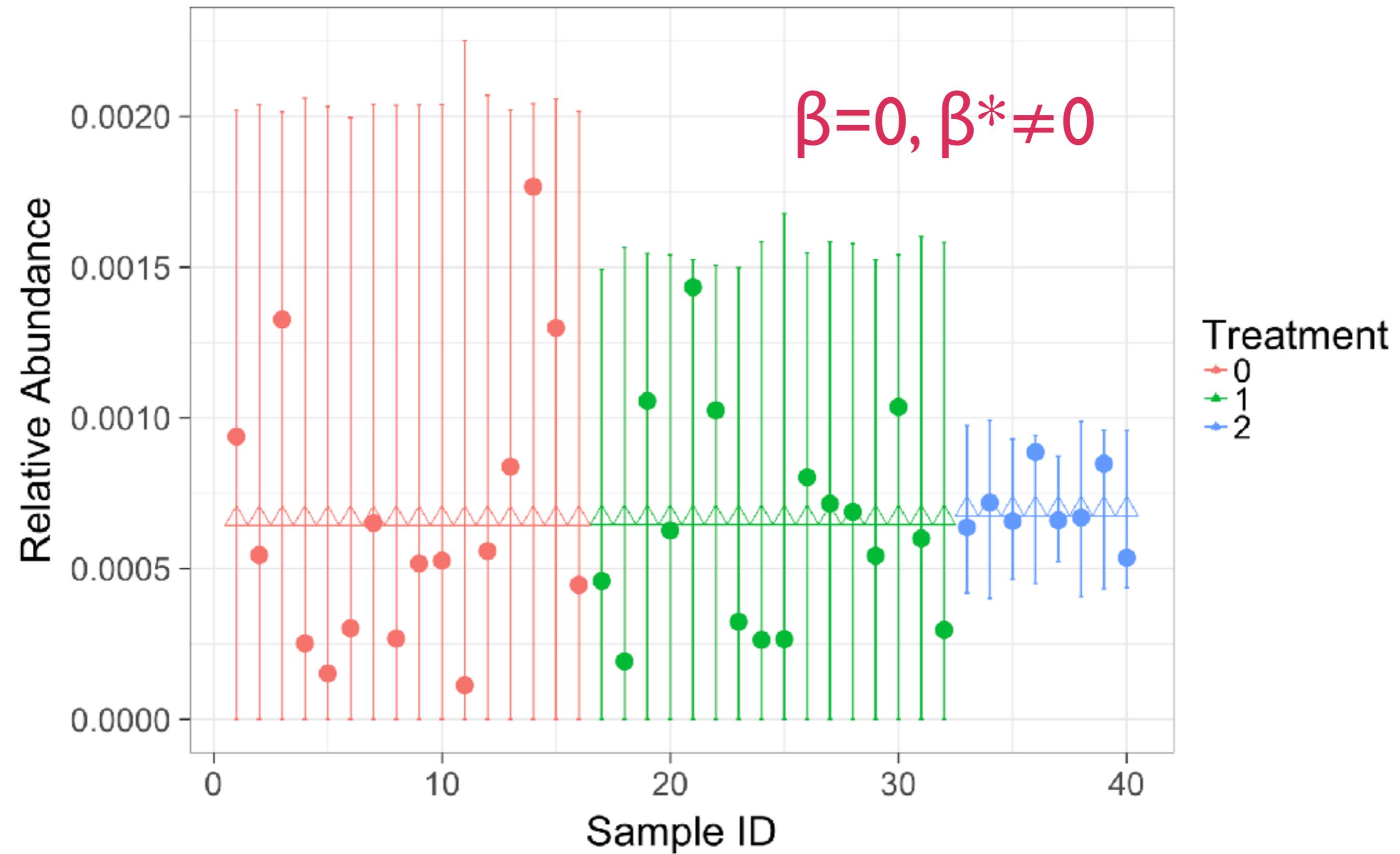
# CORNCOB AND DESEQ2

- Designed for marker gene (compositional) data
- Models relative abundance & overdispersion
- Different structure for different taxa
- Uses within-taxon correlation to model zeros
- Easy to diagnose model misspecification
- Designed for RNAseq (different data structure)
- Tests changes in abundance
- Constrained dispersion
- Individual microbes are assumed independent
- Not so easy to diagnose problems

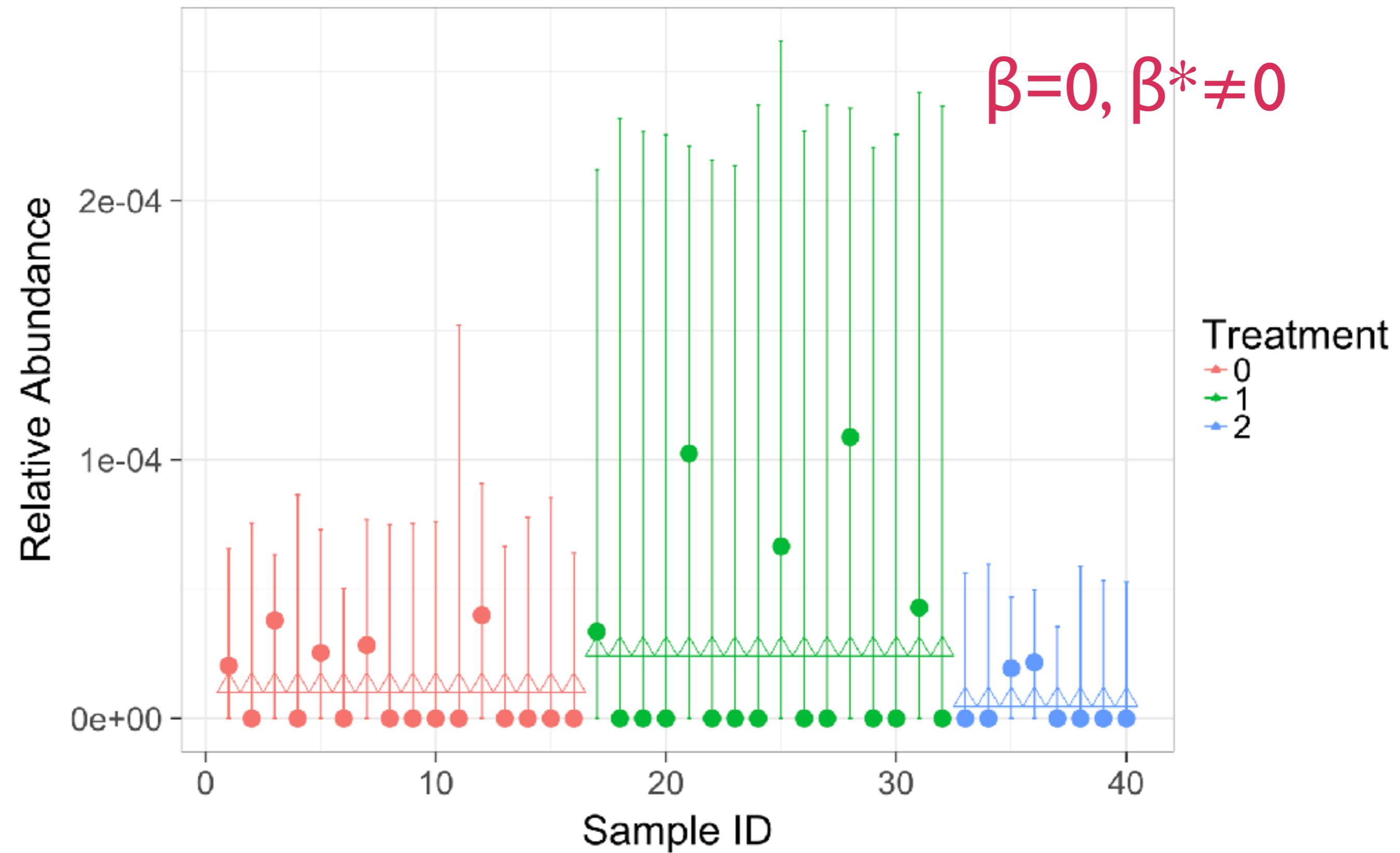
# Rare taxon, different means, same variance



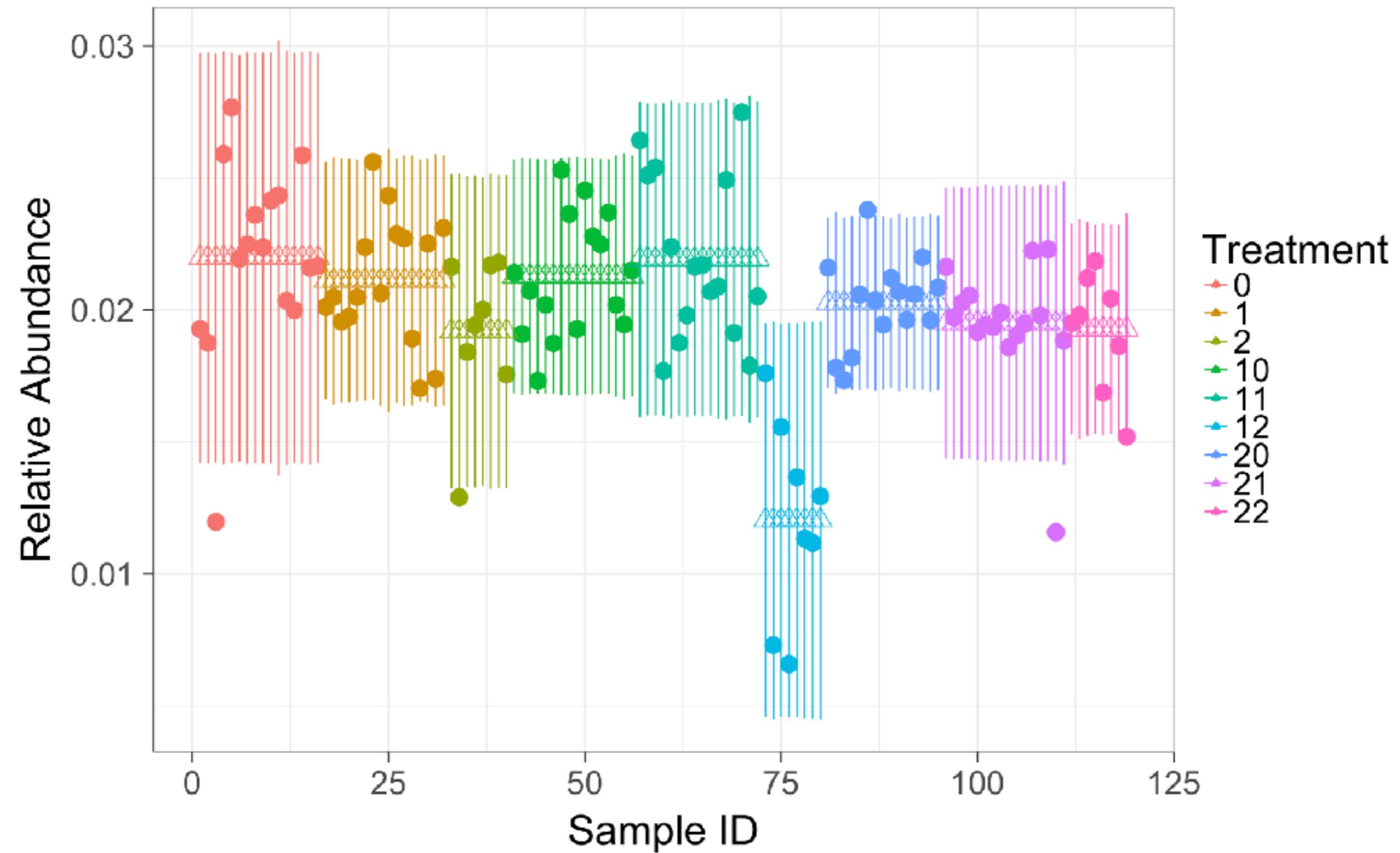
# Different variance, same mean abundance



# Rare taxon with high variability

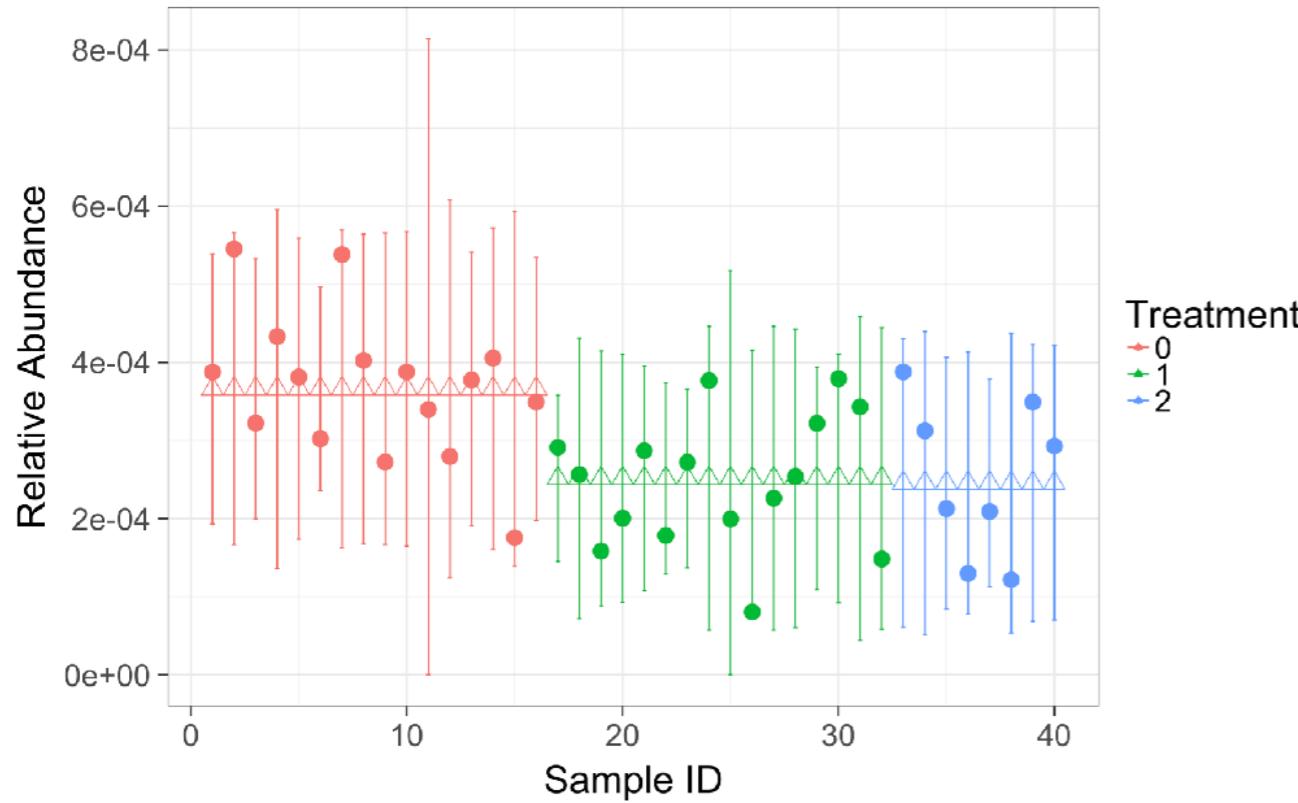


# Many classes

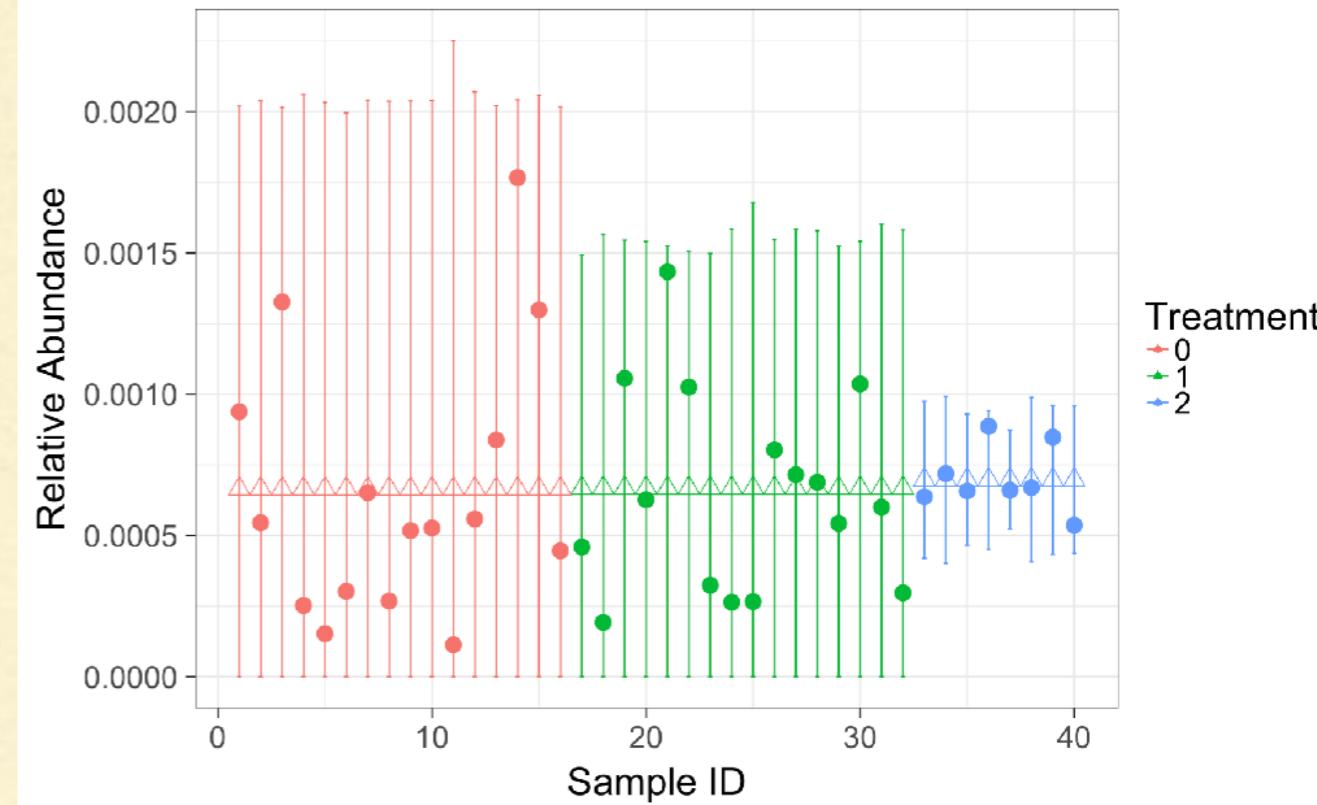


# Model Fit

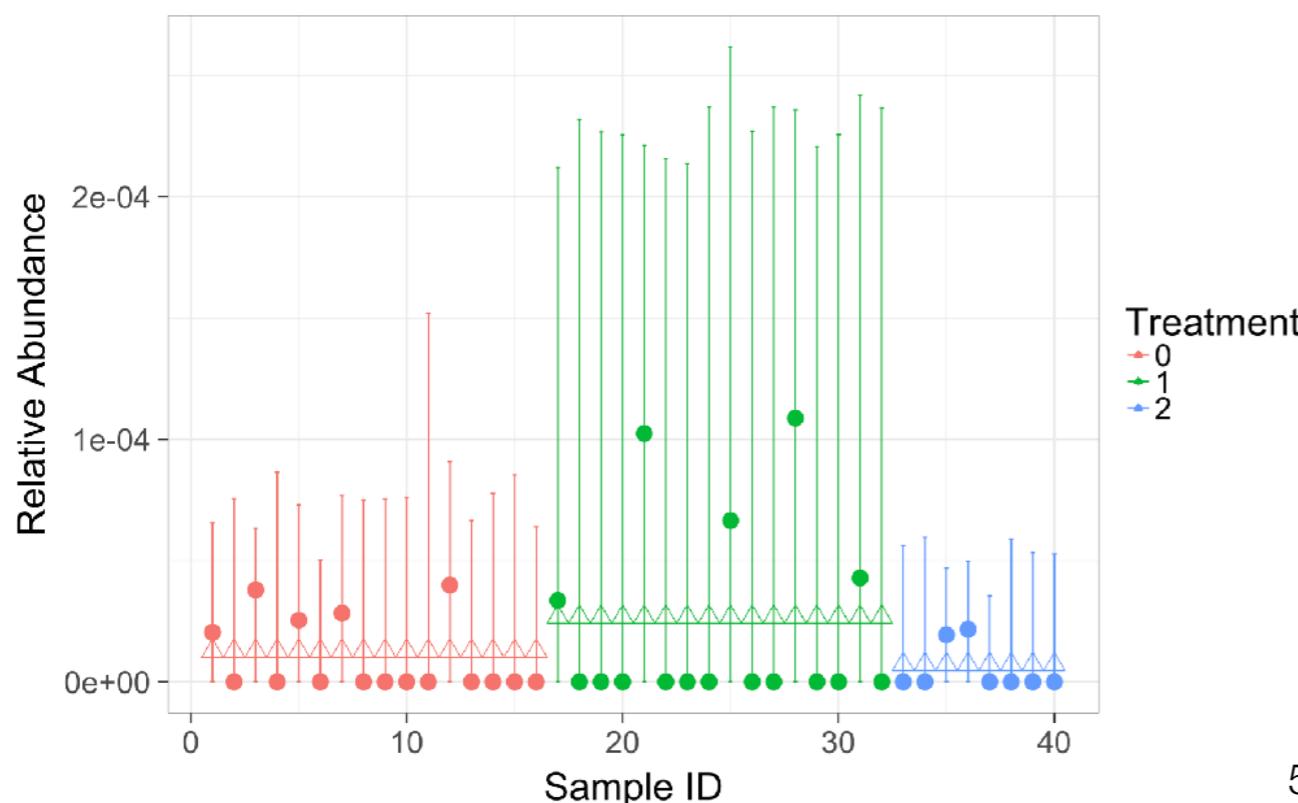
Rare taxon, different means, same variance



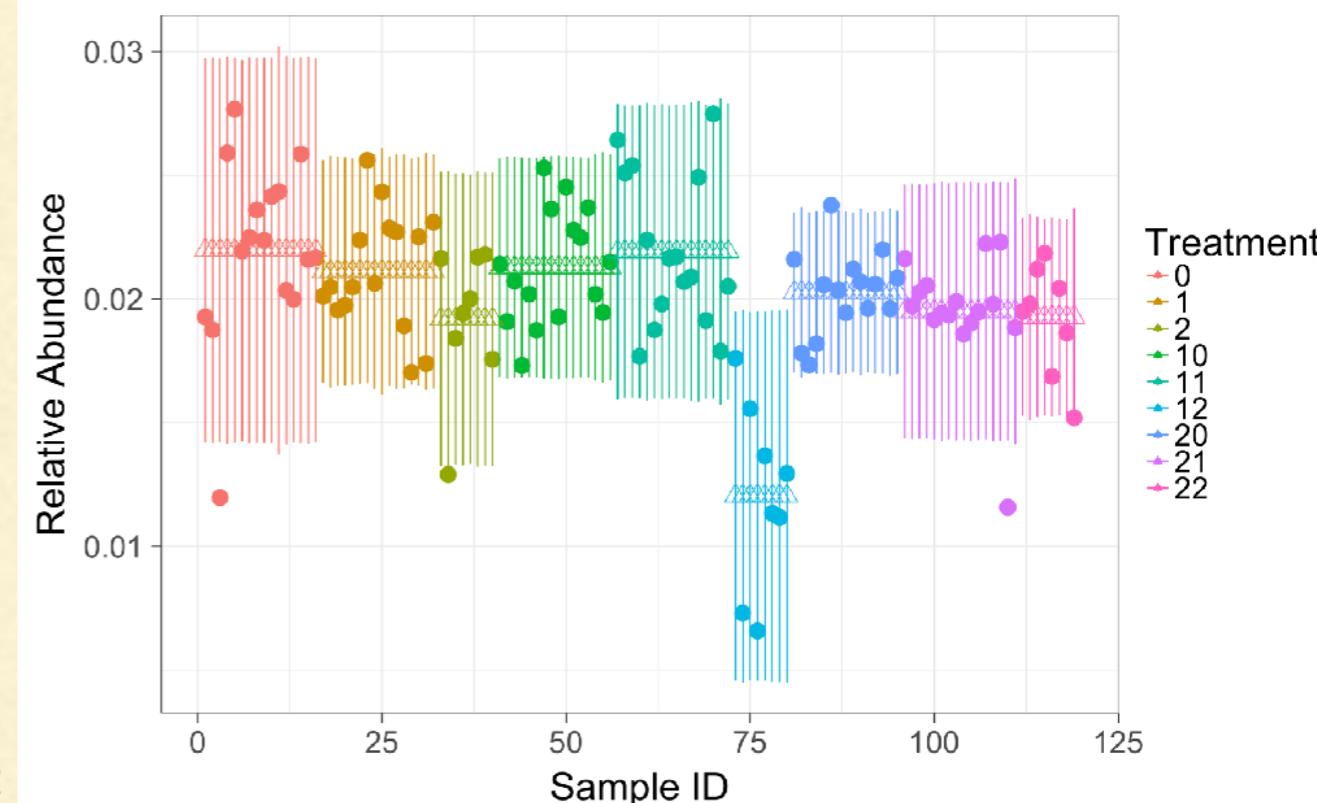
Different variance, same mean abundance



## Rare taxon with high variability



## Many classes



# DETAILS

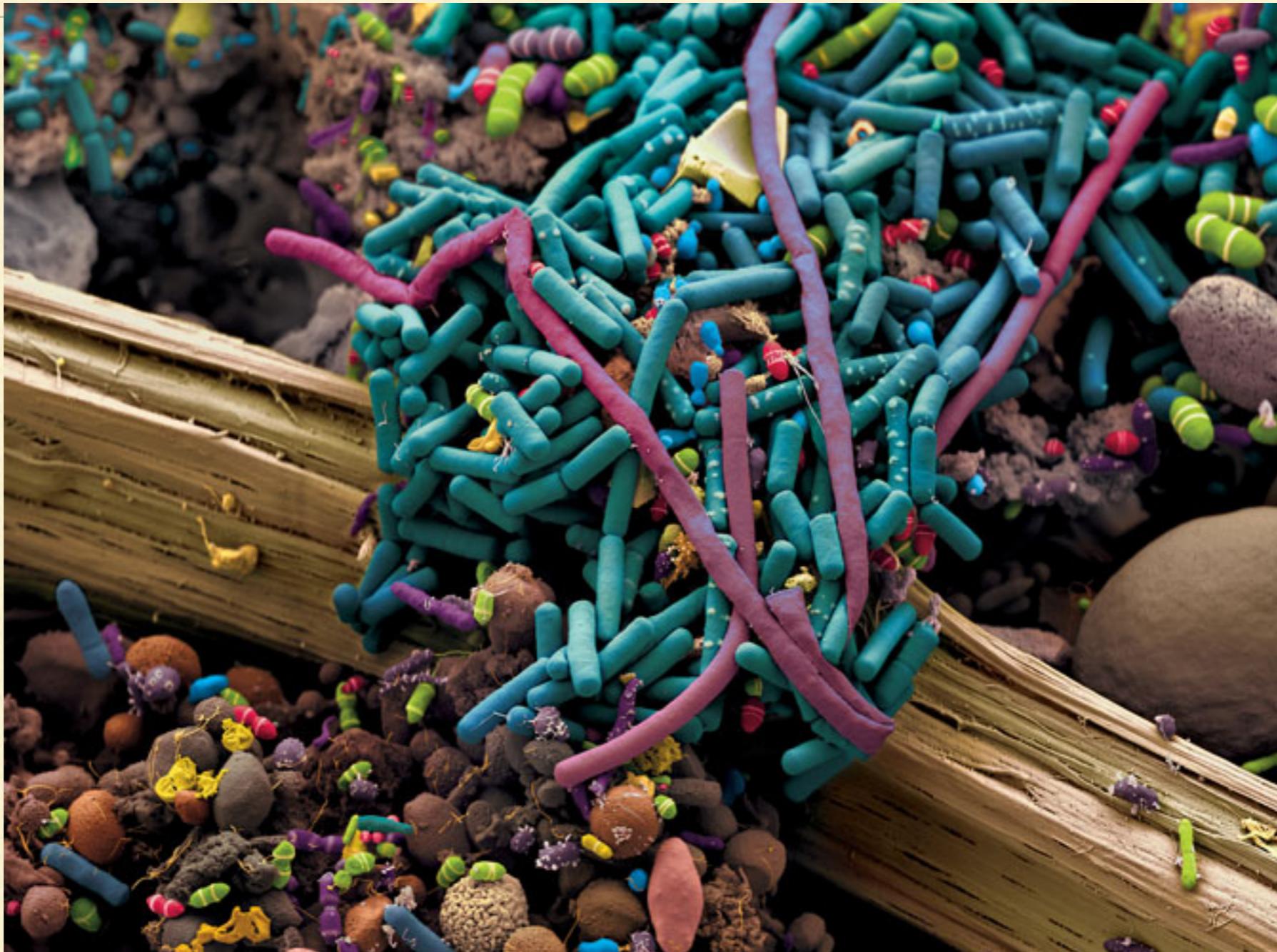


- Analytic gradient and Hessian to optimize likelihood (Fast!)
- Parametric bootstrap hypothesis testing framework
  - More samples better, but works with few samples
- **Handles zero counts well**
- Available, published
  - [github.com/bryandmartin/corncob](https://github.com/bryandmartin/corncob)
  - Martin, Witten & Willis, 2019, *Annals of Applied Statistics* 

# SUMMARY: CORNCOB



- Modeling and testing relative abundances
- Adjusts for sequencing depth
- Hypothesis testing for mean and overdispersion
- Allows for valid hypothesis testing with small sample sizes
- False discovery rate control for testing many taxa



# ABUNDANCES LABS

# LAB PLAN

---

- Lab: Relative abundance
  - 1. intro: 🌽corncob🌽 using 16S (Bryan)
  - 2. How to process shotgun data to use with 🌽corncob🌽 (Taylor)
  - 3. 🌽corncob🌽 using shotgun data (Bryan)
  - 4. 🌽corncob🌽 vs 😢DESeq2😢: a comparison (Bryan)

# PAULINE AND BRYAN...

Instructions

- Navigate to <https://github.com/statdivlab/stamps2019>
- Scroll down to Tutorials and copy and paste the download commands for the corncob and corncob/DESeq2 R-markdown files

## Tutorials

### Corncob Tutorial

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2019/master/labs/corncob_tutorial/cor
```

### Corncob and DESeq2 Tutorial

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2019/master/labs/corncobDESeq2/cornc
```

### Diversity Estimation Tutorial

```
download.file("https://raw.githubusercontent.com/statdivlab/stamps2019/master/labs/diversity-lab.Rmd",
```

```
corn <- bbdml(formula = OTU.1 ~ Day * Amdmt,  
               phi.formula = ~ Day * Amdmt,  
               data = soil)  
  
corn
```

Call:

```
bbdml(formula = OTU.1 ~ Day * Amdmt, phi.formula = ~Day * Amdmt,  
       data = soil)
```

Coefficients associated with abundance:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.71523	0.02664	-26.848	< 2e-16 ***
Day1	0.17110	0.04209	4.065	0.000127 ***
Amdmt1	-0.01036	0.02947	-0.352	0.726175
Amdmt2	-0.06366	0.03283	-1.939	0.056674 .
Day1:Amdmt1	0.10855	0.05683	1.910	0.060316 .
Day1:Amdmt2	0.65613	0.07330	8.951	0.000000000000413 ***
---				
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Coefficients associated with dispersion:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-5.9998	0.3563	-16.838	< 2e-16 ***
Day1	0.4659	0.5023	0.927	0.35699
Amdmt1	-1.5434	0.5131	-3.008	0.00368 **
Amdmt2	-1.3993	0.6332	-2.210	0.03049 *
Day1:Amdmt1	1.7696	0.7167	2.469	0.01607 *
Day1:Amdmt2	1.8951	0.8811	2.151	0.03505 *
---				
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

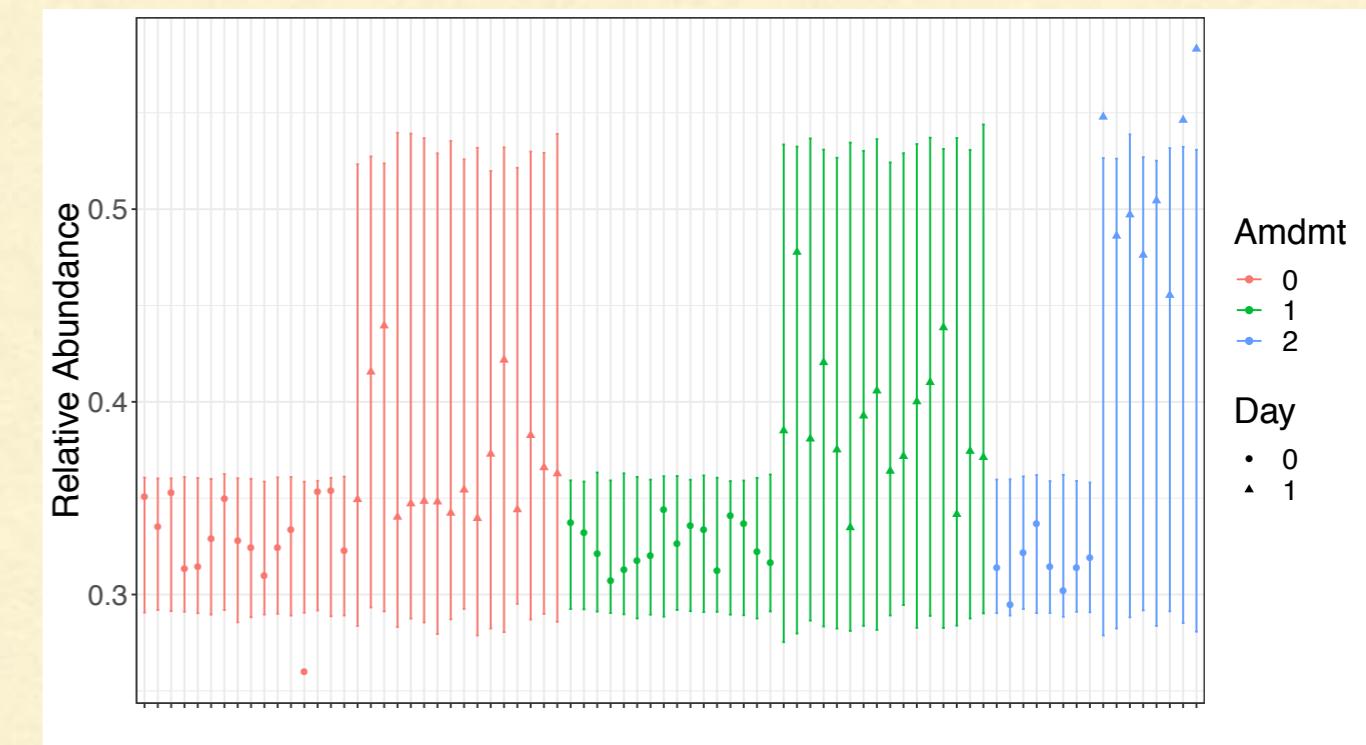
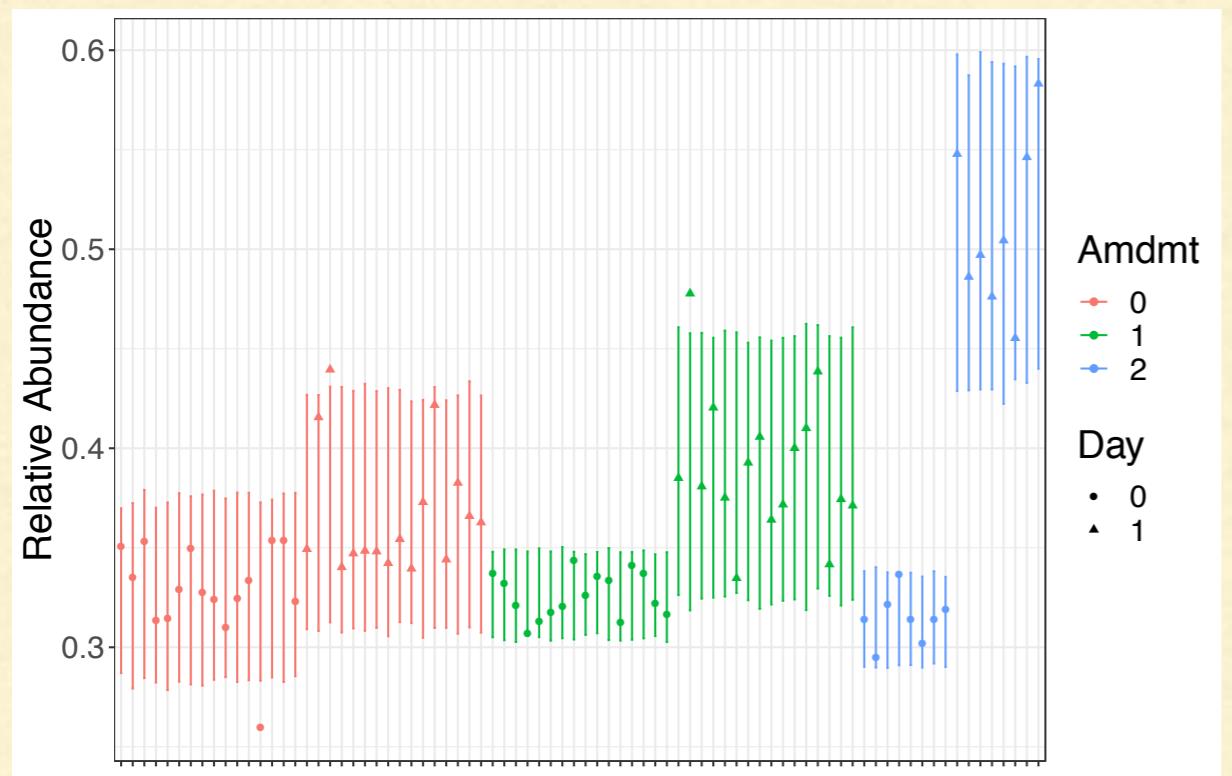
Log-likelihood: -681.79

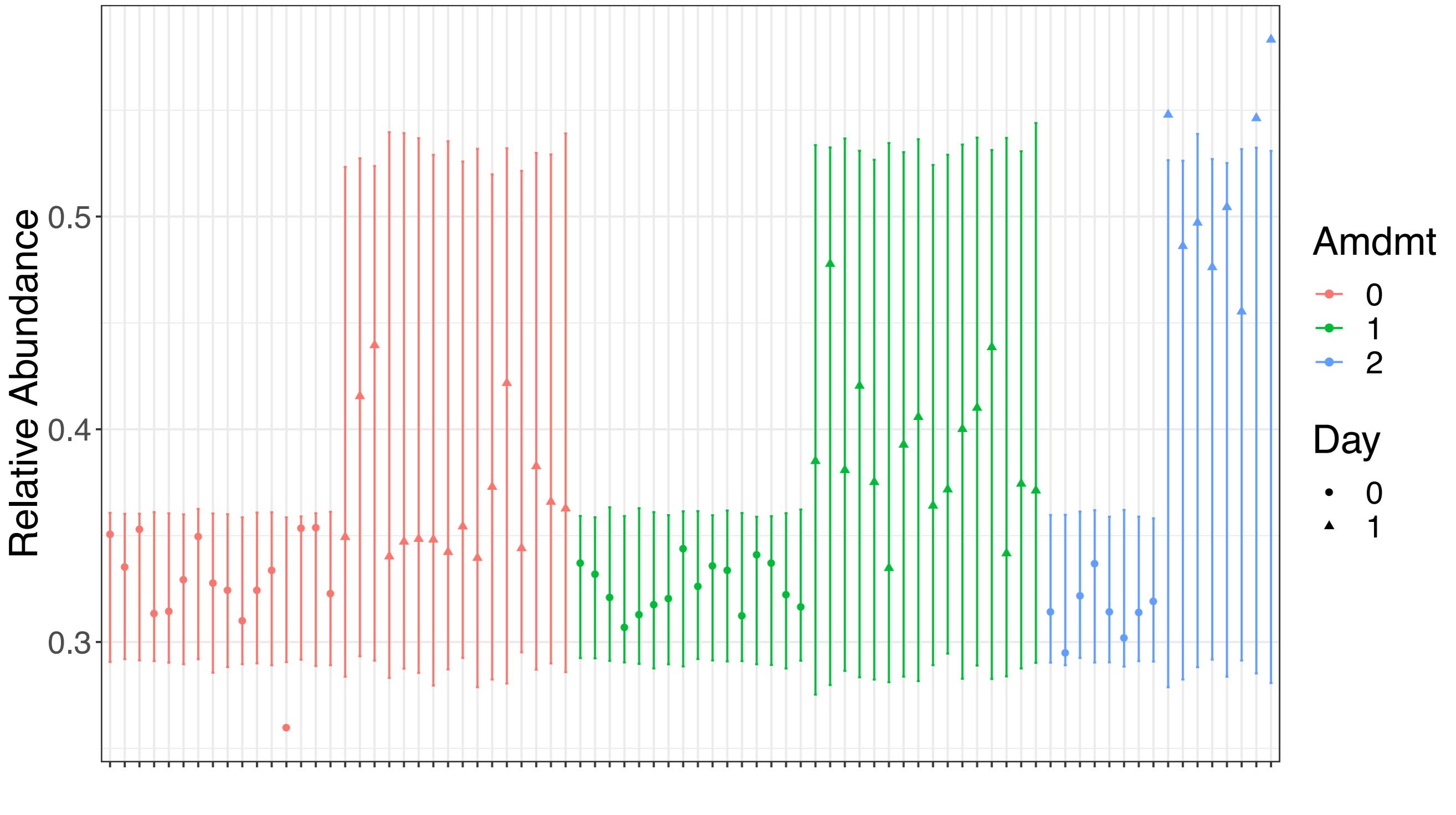
---

```
plot(corn, color = "Amdmt", shape = "Day") +  
  theme(axis.text.x = element_blank(), text = element_text(size = 20))  
ggsave("fullplot.pdf")
```



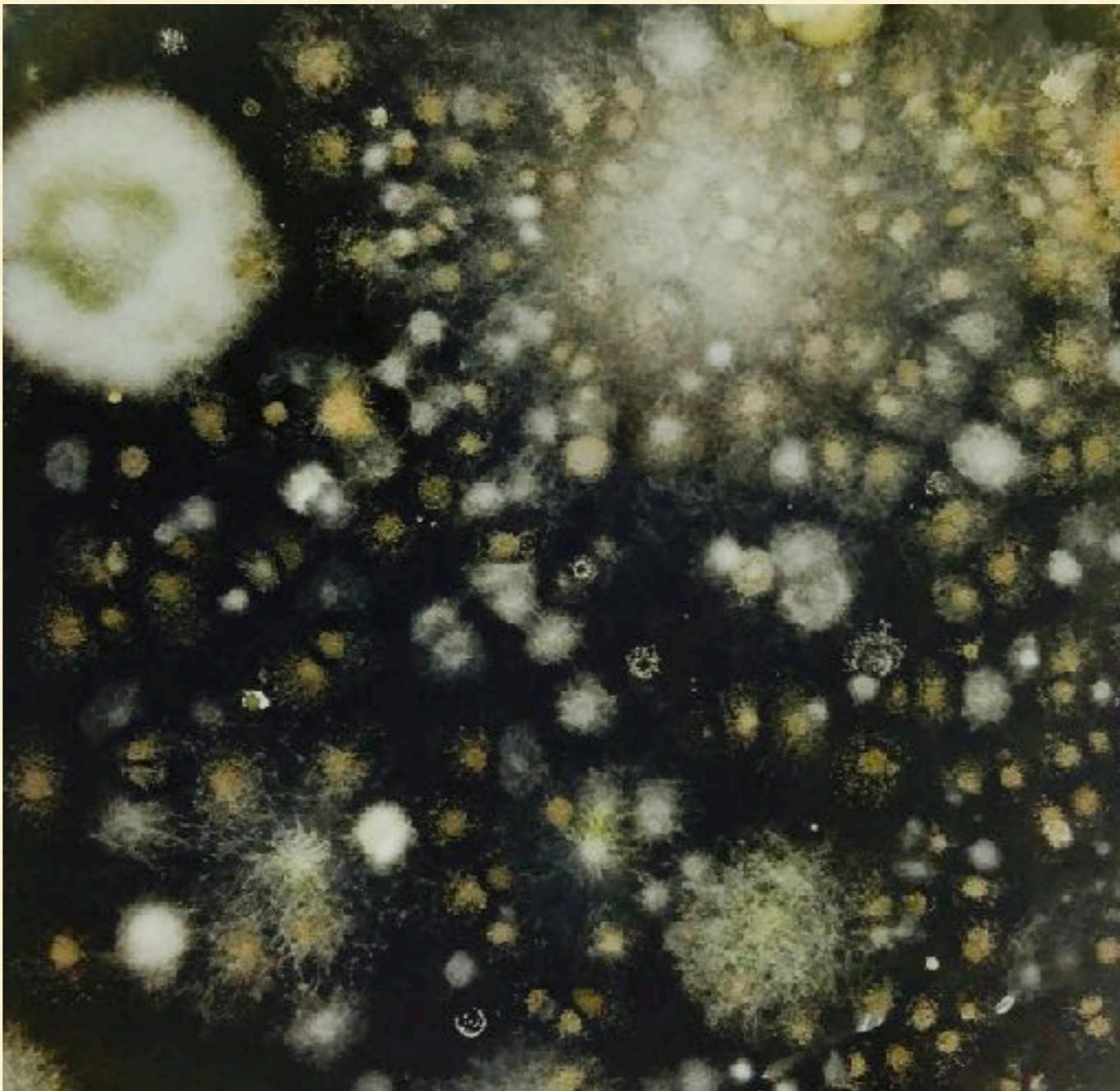
```
> waldchisq(corn, restrictions = c("Day", "Day:Amdmt"))
[1] 4.044134e-56
```





# TAYLOR

Notes on  
what  
she did with  
the data



# DIVERSITY: CONTEXT



# Microbial diversity in the deep sea and the underexplored “rare biosphere”

Mitchell L. Sogin, Hilary G. Morrison, Julie A. Huber, David Mark Welch, Susan M. Huse, Phillip R. Neal, Jesus M. Arrieta, and Gerhard J. Herndl

PNAS August 8, 2006 103 (32) 12115-12120; <https://doi.org/10.1073/pnas.0605127103>

Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

Article

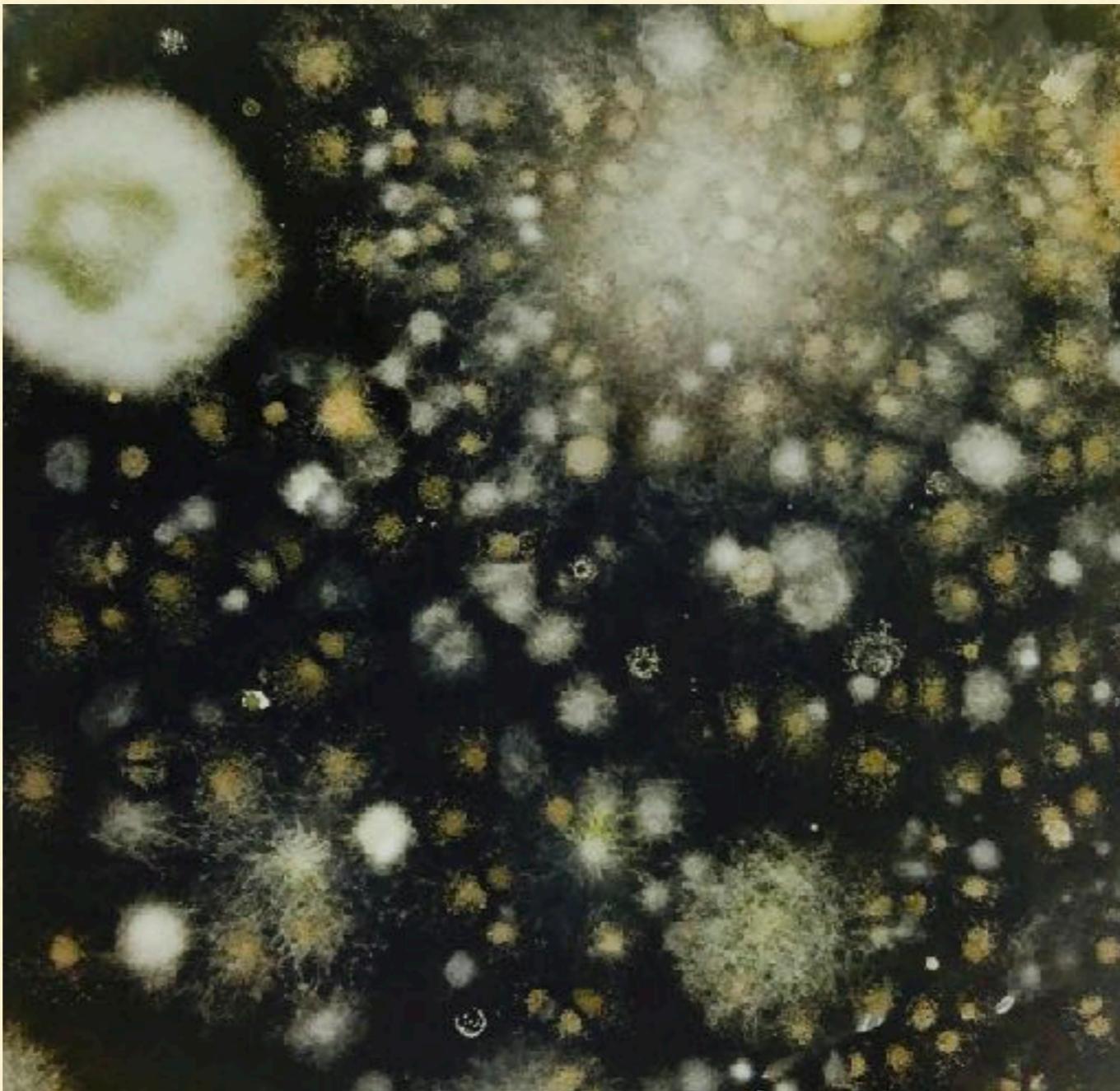
Figures & SI

Info & Metrics

PDF

## Abstract

The evolution of marine microbes over billions of years predicts that the composition of microbial communities should be much greater than the published estimates of a few thousand distinct kinds of microbes per liter of seawater. By adopting a massively parallel tag sequencing strategy, we show that bacterial communities of deep water masses of the North Atlantic and diffuse flow hydrothermal vents are one to two orders of magnitude more complex than previously reported for any microbial environment. A relatively small number of different populations dominate all samples, but thousands of low-abundance populations account for most of the observed phylogenetic diversity. This “rare biosphere” is very ancient and may represent a nearly inexhaustible source of genomic innovation. Members of the rare biosphere are highly divergent from each other and, at different times in earth's history, may have had a profound impact on shaping planetary processes.



# DIVERSITY: PRACTICE

# DIVERSITY

---

- Low dimensional summaries of entire communities
  - $\alpha$ -diversity: one community
    - e.g., species richness, Shannon diversity
  - $\beta$ -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  $\alpha$ -diversity parameter

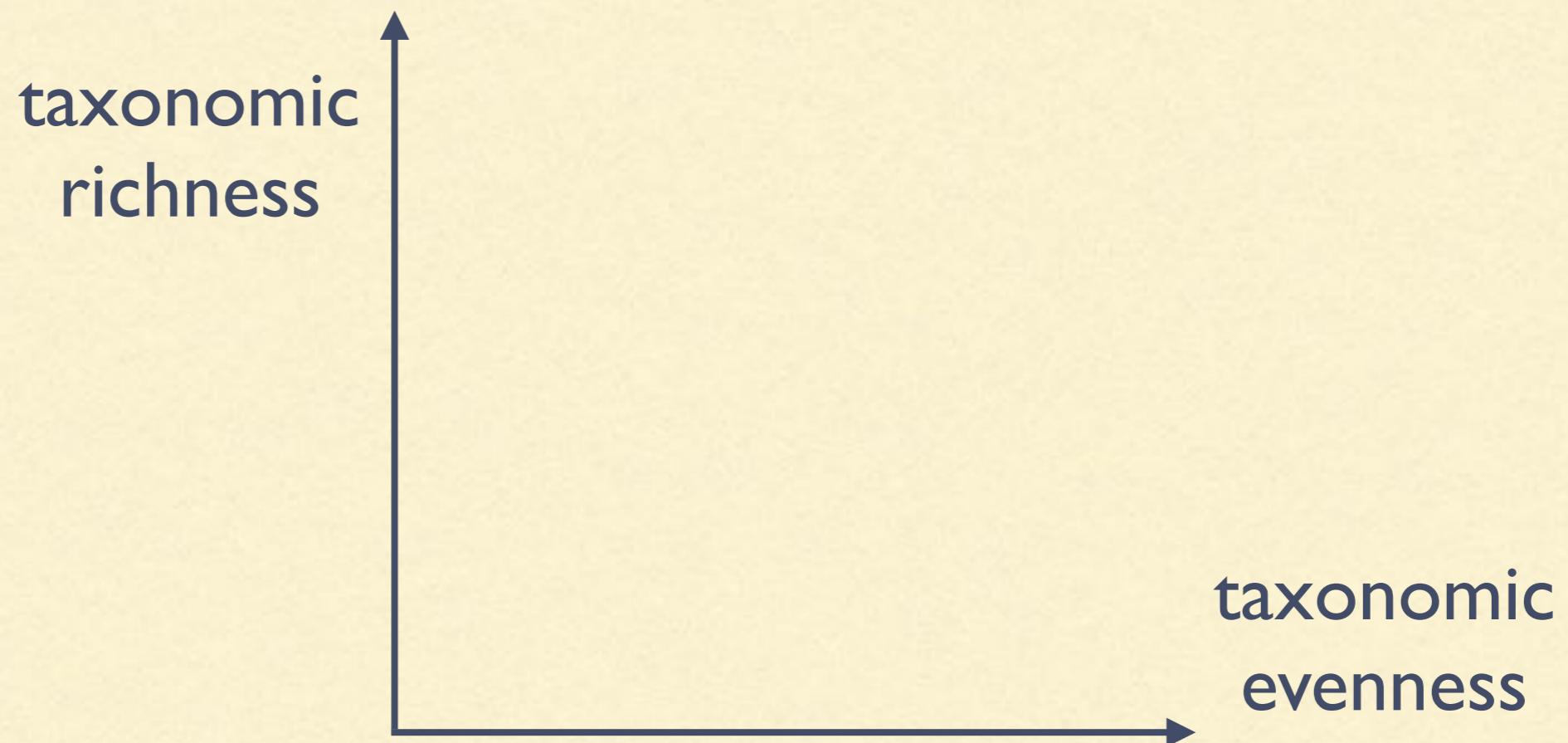
# ALPHA DIVERSITY

---

- Some examples of  $\alpha$ -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:  $\frac{-\sum_{i=1}^C p_i \ln p_i}{\ln C}$

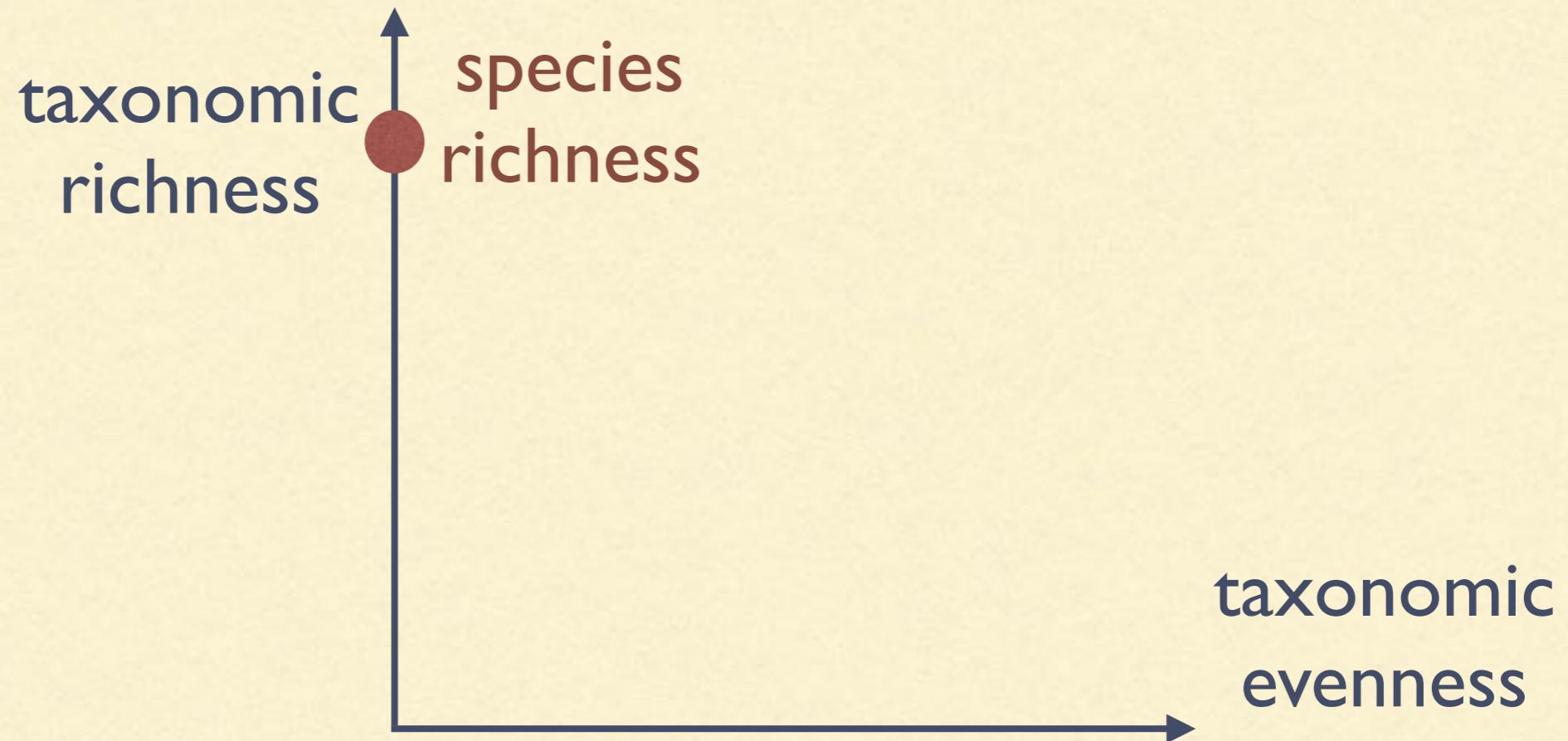
# YOUR CHOICE

- Think: What difference do you want to highlight?



# YOUR CHOICE

---



# YOUR CHOICE



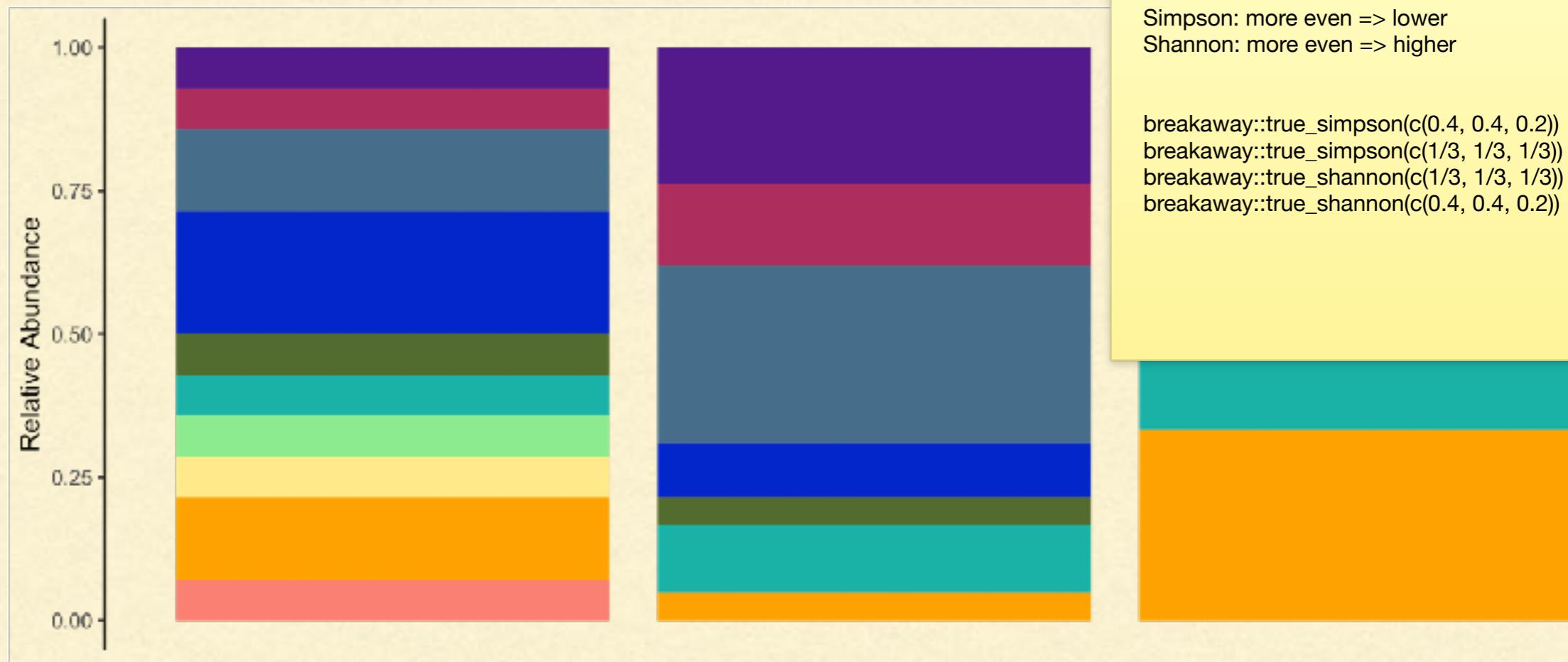
# YOUR CHOICE



# YOUR CHOICE



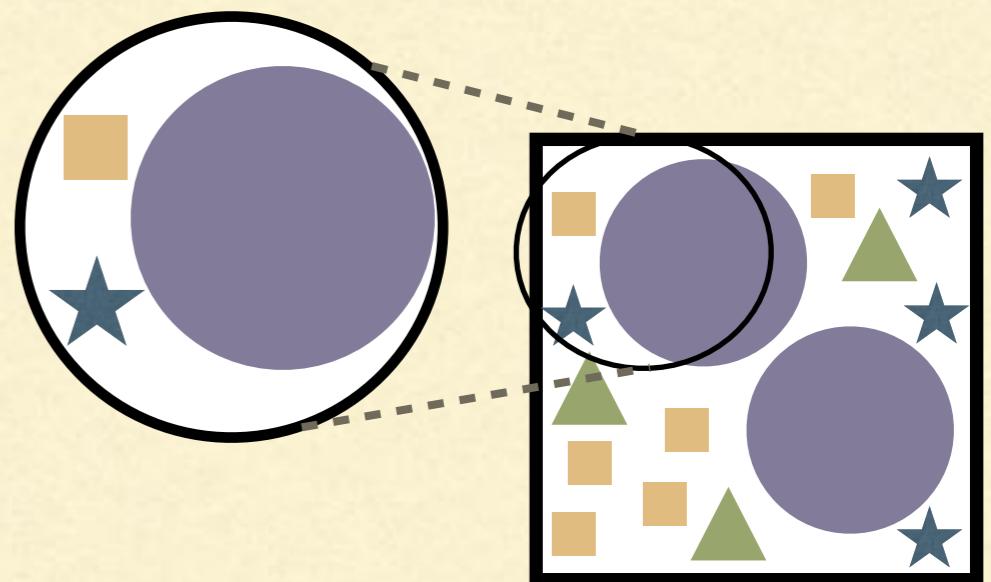
This is a question of *parameter choice*:  
Which parameter highlights the differences I care about?



<b>Richness</b>	10	7	4
<b>Shannon</b>	2.21	1.75	1.33
<b>Evenness</b>	0.96	0.90	0.96
<b>Simpson's</b>	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**



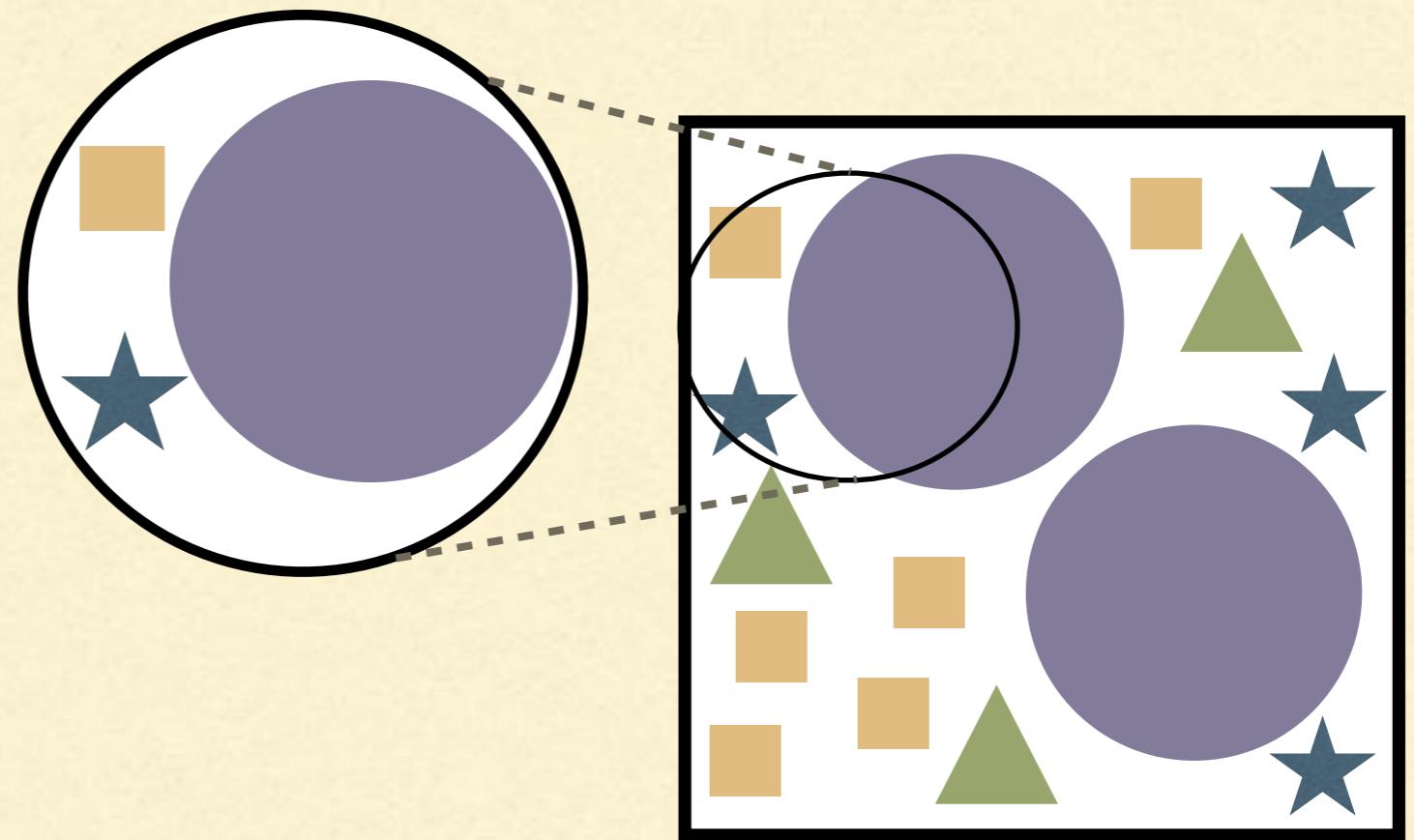
naive

# 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

---

- 2 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



Kendrick Li

Alex Paynter



- CatchAll: mixed Poisson models
- **stable, restrictive, hard to use**
- breakaway: non-mixed Poisson models
- **Higher variance, flexible models, in R**



# SPECIES RICHNESS ESTIMATION

- Good options



- `breakaway::breakaway()`; QIIME2 breakaway plug-in

- `breakaway::chao_bunge()`

- `breakaway:: objective_bayes_*`()

- CatchAll

- Bad options

- QIIME2: `chao1`; scikitbio...

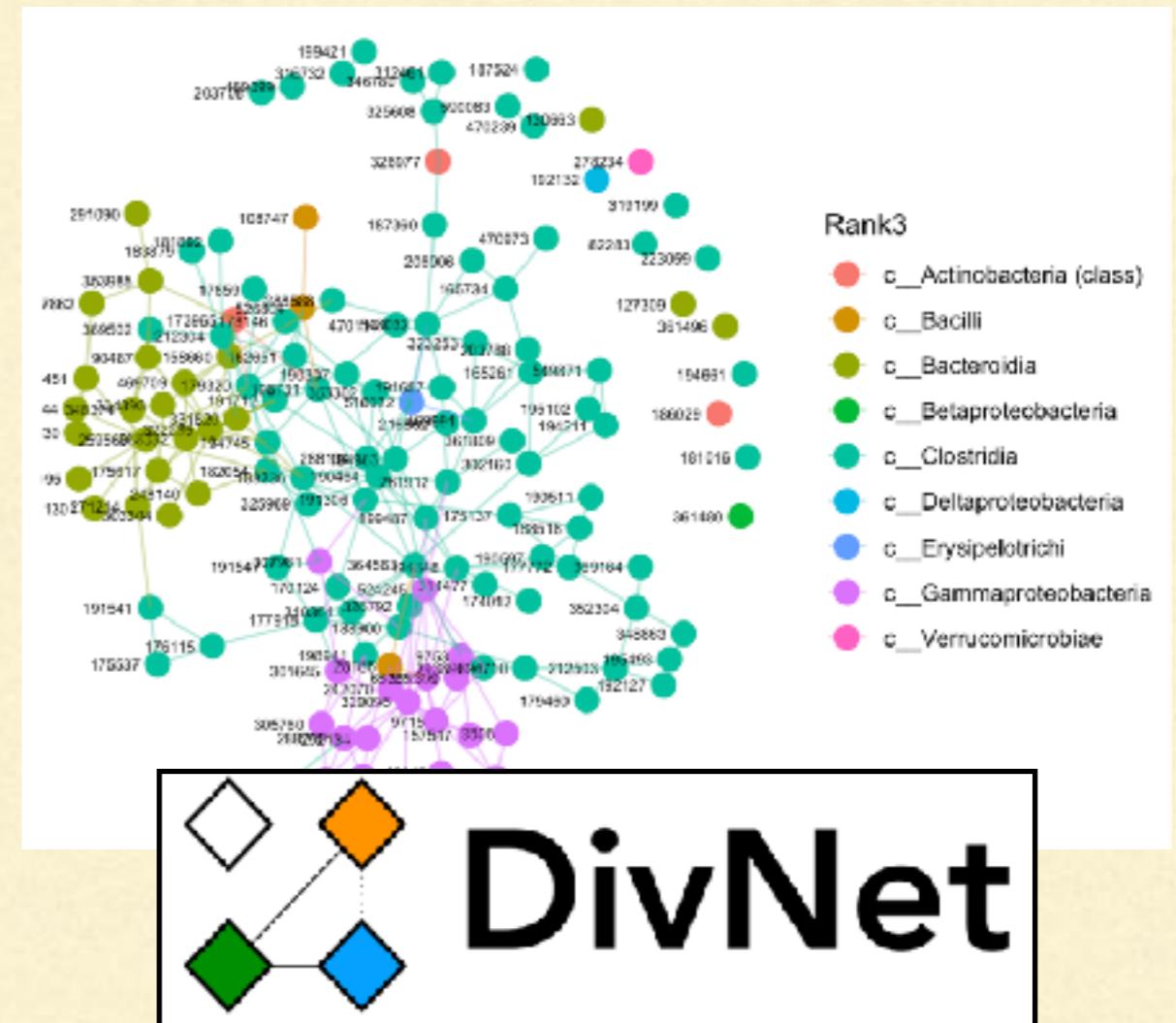
- R:`vegan::...`



**Pauline Trinh**  
**(Q2 wizard)**

# ALPHA DIVERSITY: SHANNON DIVERSITY, SIMPSON, ETC.

- 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



# DIVNET



Bryan Martin



Pauline Trinh

- This idea works for estimating any diversity index ( $\alpha$  or  $\beta$ ) that is a function of relative abundances
- It can also be used to estimate any diversity index that is a function of the tree

[github.com/adw96/DivNet](https://github.com/adw96/DivNet)

Soon to include...



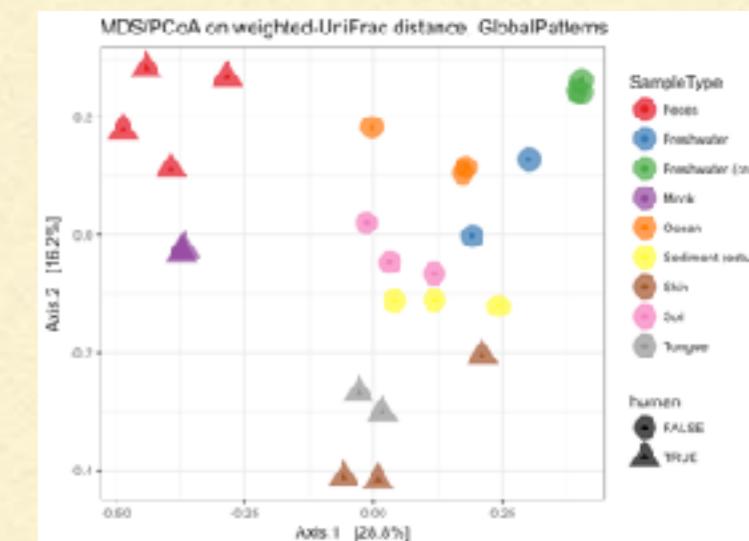
# BETA DIVERSITY

---

- Community 1:  $p_1^{(1)}, p_2^{(1)}, \dots, p_c^{(1)}$ ; Community 2:  $p_1^{(2)}, p_2^{(2)}, \dots, p_c^{(2)}$
- $\beta$ -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: HYPOTHESIS TESTING

- Sometimes diversity is analysed as an exploratory tool
  - e.g., ordination
- Other times you want to do inference
  - e.g.,  $H_0$ : two communities have zero dissimilarity
  - e.g.,  $H_0$ : communities A & B have same dissimilarity as communities A & C



# HYPOTHESIS TESTING FOR DIVERSITY

---

- Common approach: PERMANOVA
- Critical issue: adjust for different resolution
- Best solution = ask “*do I really want to do this test?*”
- Better solution = use error bars
  - `breakaway::betta(); DivNet::testDiversity`
- (Bad solution = rarefy)

# 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 the journal PLOS Computational Biology. The header includes the PLOS logo, the journal name, and navigation links for BROWSE, PUBLISH, and ABOUT. Below the header, it indicates the article is OPEN ACCESS and PEER-REVIEWED. 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: Whole-Genome Sequencing of the Human Gut Microbiome**

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 bioRxiv preprint page for a study titled "Microbial differential abundance depends upon data bias and diversity".

**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

**Microbiome**

Home About Articles Submission Guidelines

OPEN RESEARCH

CSH Cold Spring Harbor Laboratory

Waste

Paul J. Newland  
Published by

New Results

Rarefaction, alpha diversity, and statistics

Amy Willis

doi: <https://doi.org/10.1101/231878>

This article is a preprint and has not been peer-reviewed [what does this mean?].

2 comments

Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Juez-Baeza, Amanda Birmingham,

Author(s). 2017

017 | Published: 3 March 2017

# DIVERSITY

---

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

# DIVERSITY

---

- Good news
  - Diversity *is* different in the environments you care about
  - You can safely reject the null...
  - ... and publish a paper
- Bad news
  - You didn't advance science in any way

# 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



# DIVERSITY LAB

---



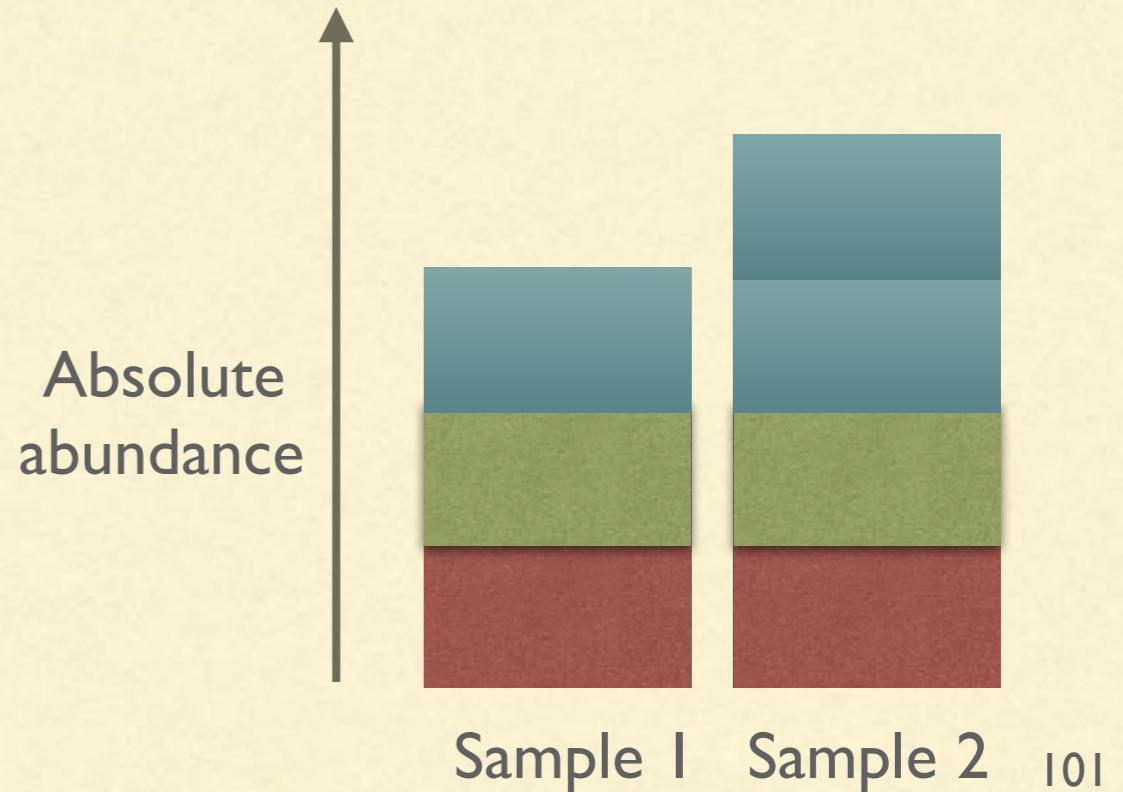
- Pauline to fill in

# 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

---

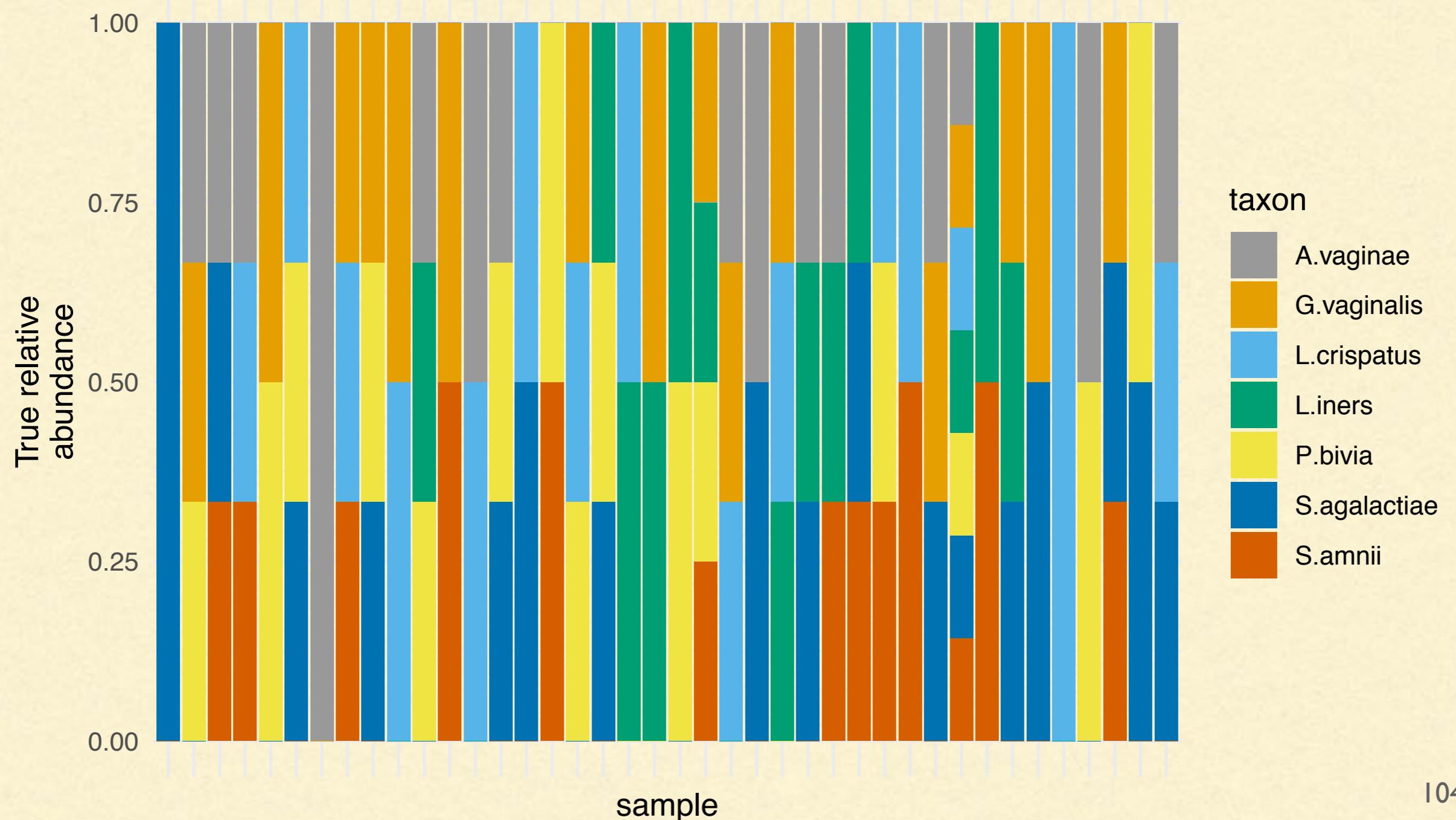
- Can we even estimate true relative abundance from sequencing data?
- Does our sequencing technology give us information about relative abundance?
- Is observed relative abundance proportional to actual relative abundance?

# DECONVOLUTION OF BIOLOGY AND SEQUENCING

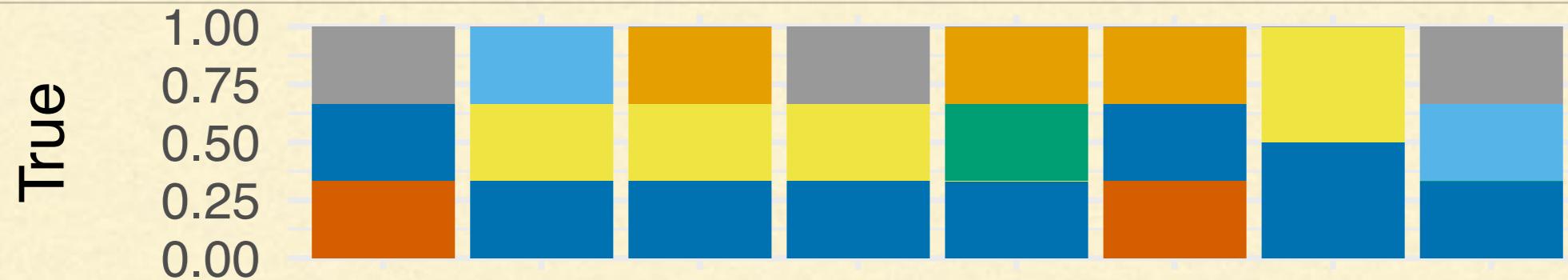
---

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

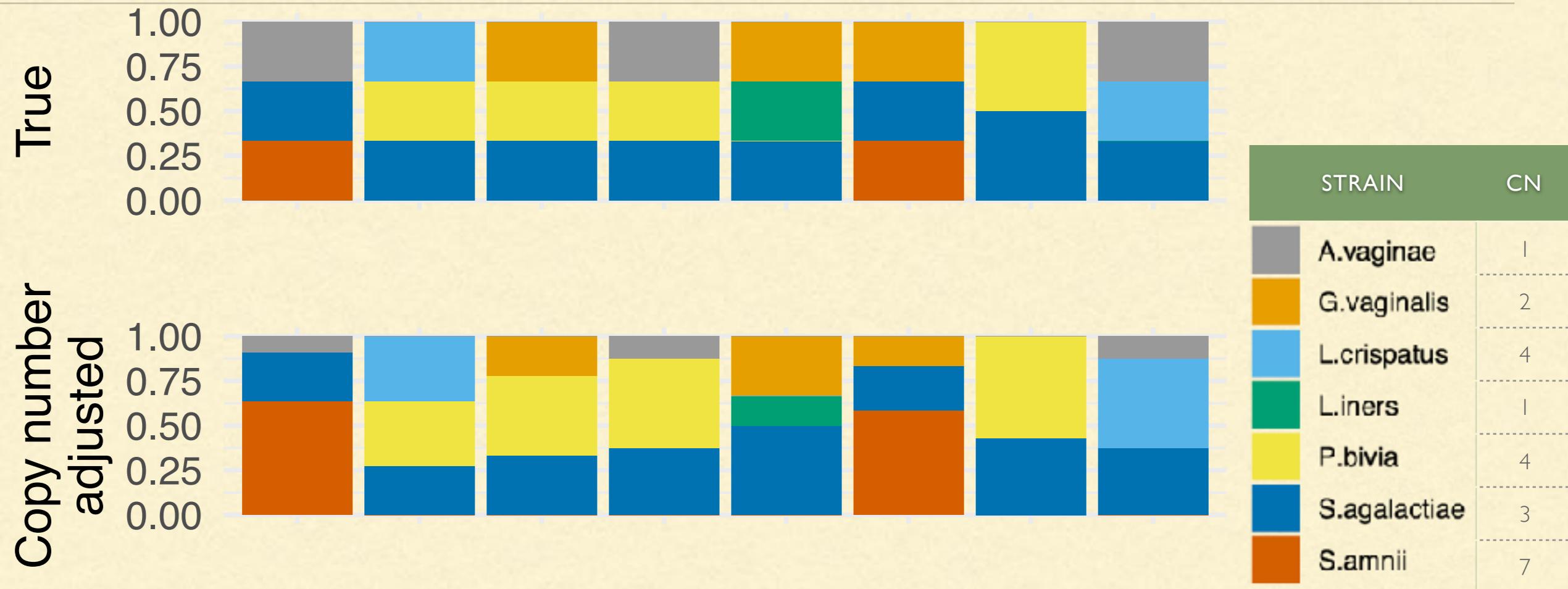
# MOCK COMMUNITY DATA



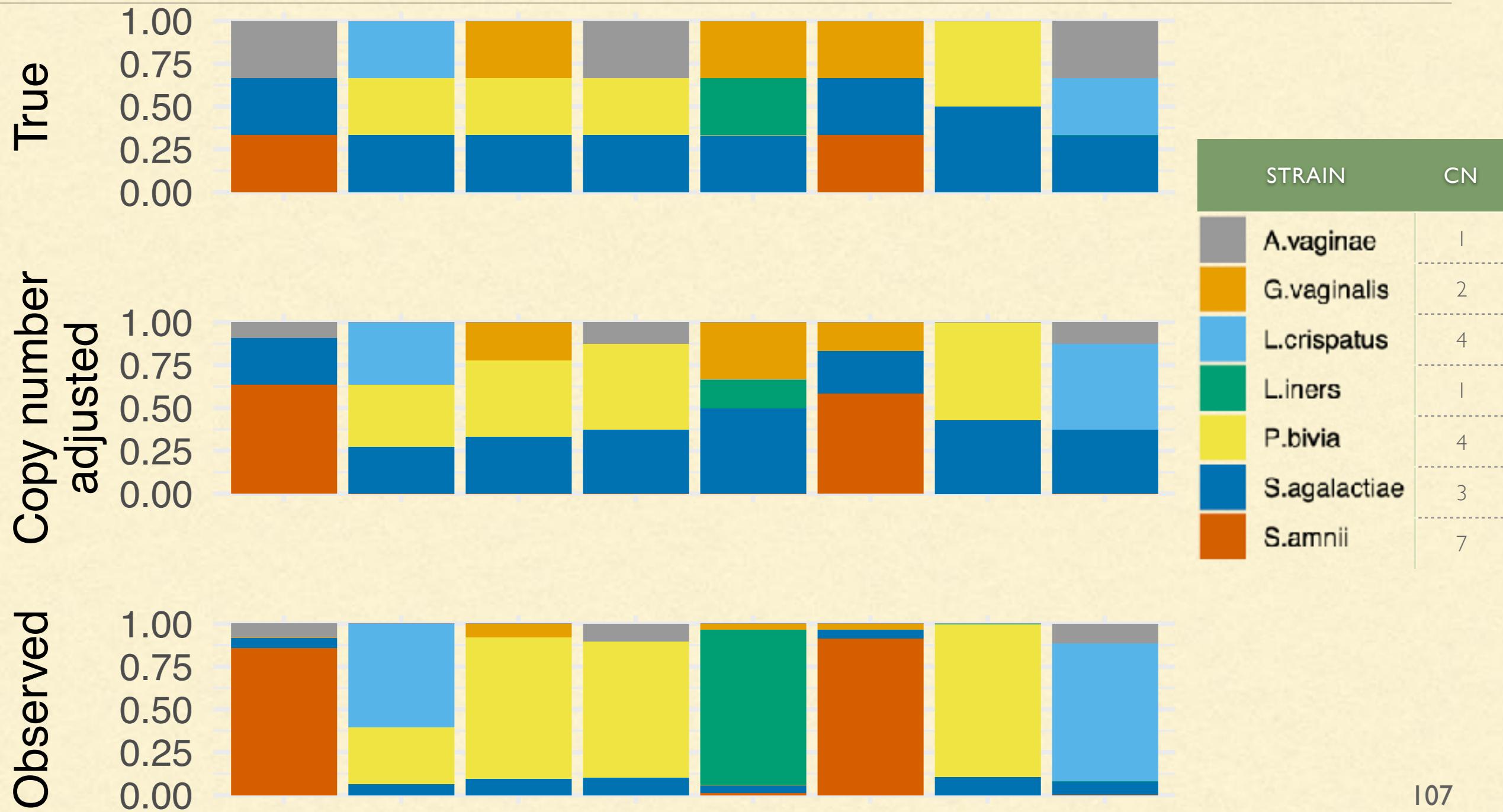
# WHAT DO WE OBSERVE?



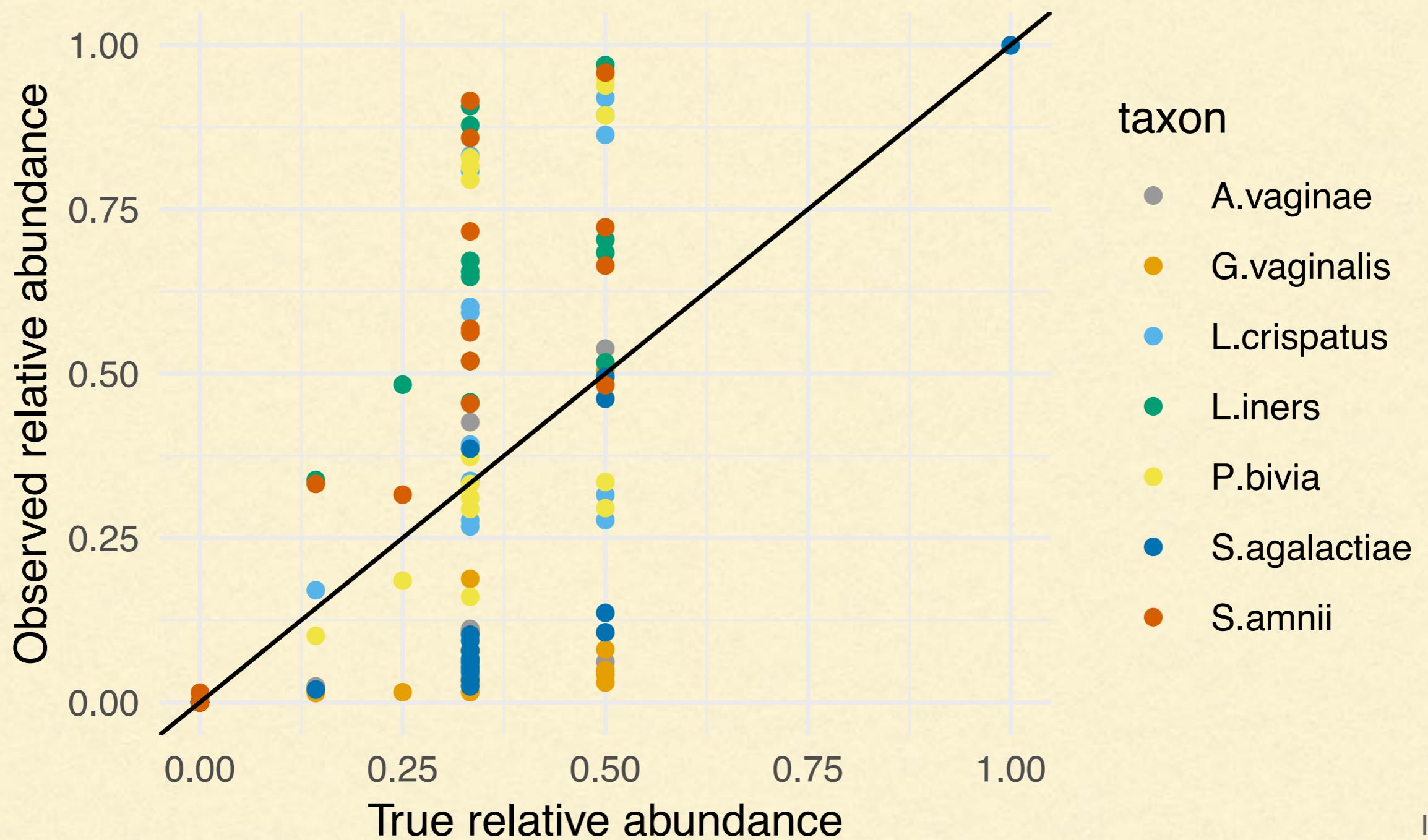
# WHAT DO WE OBSERVE?



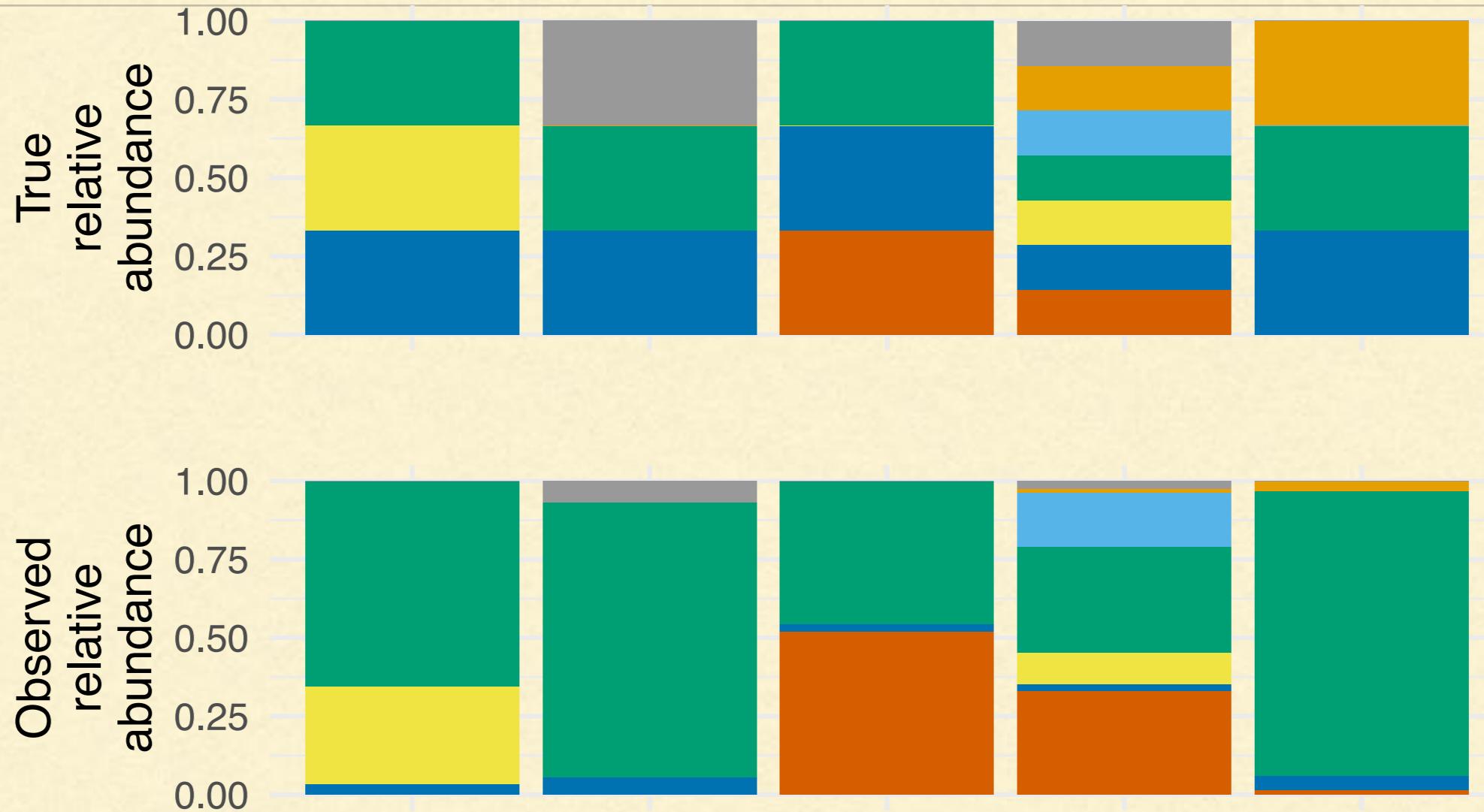
# WHAT DO WE OBSERVE?



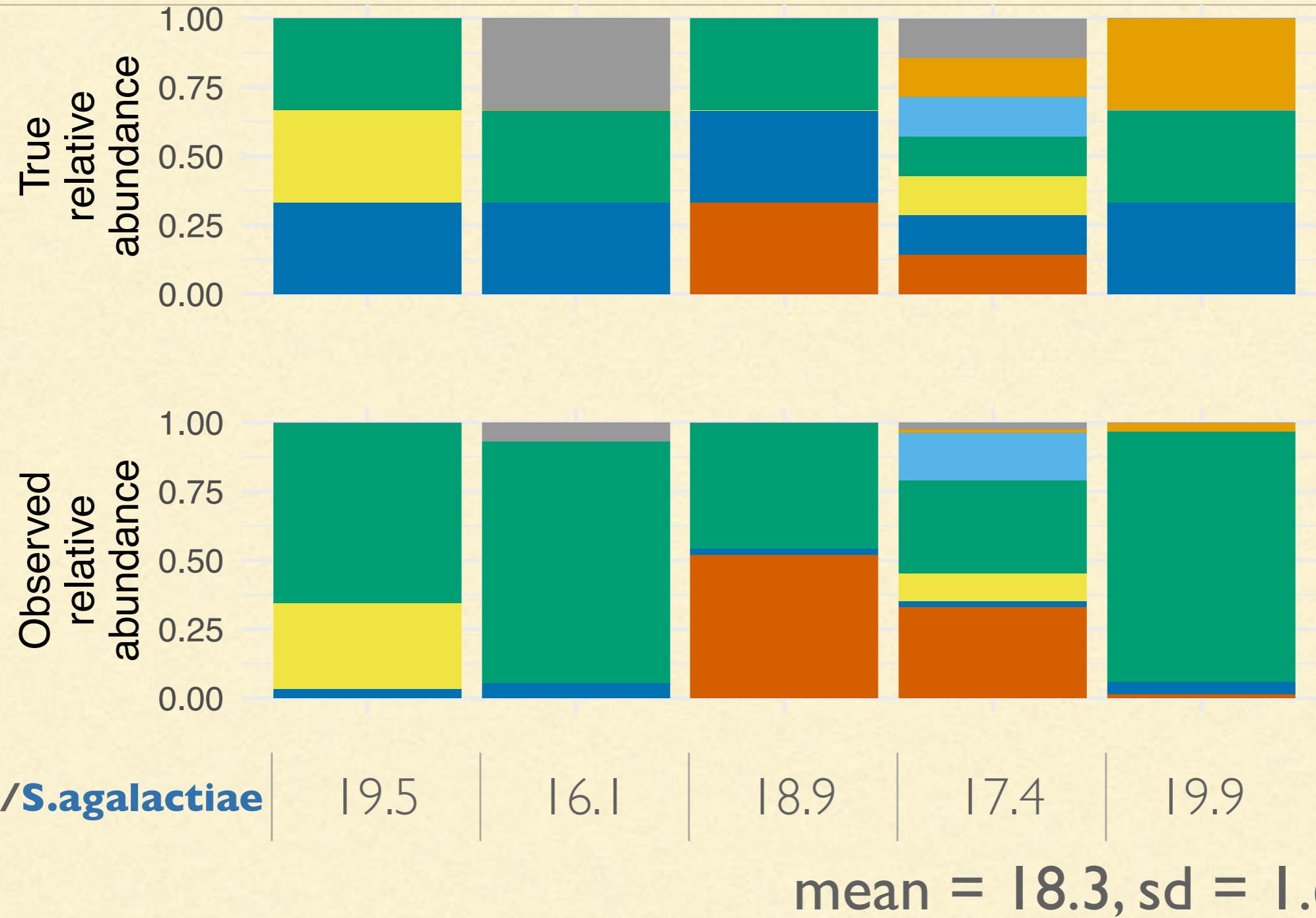
# WHAT PATTERNS CAN WE FIND?



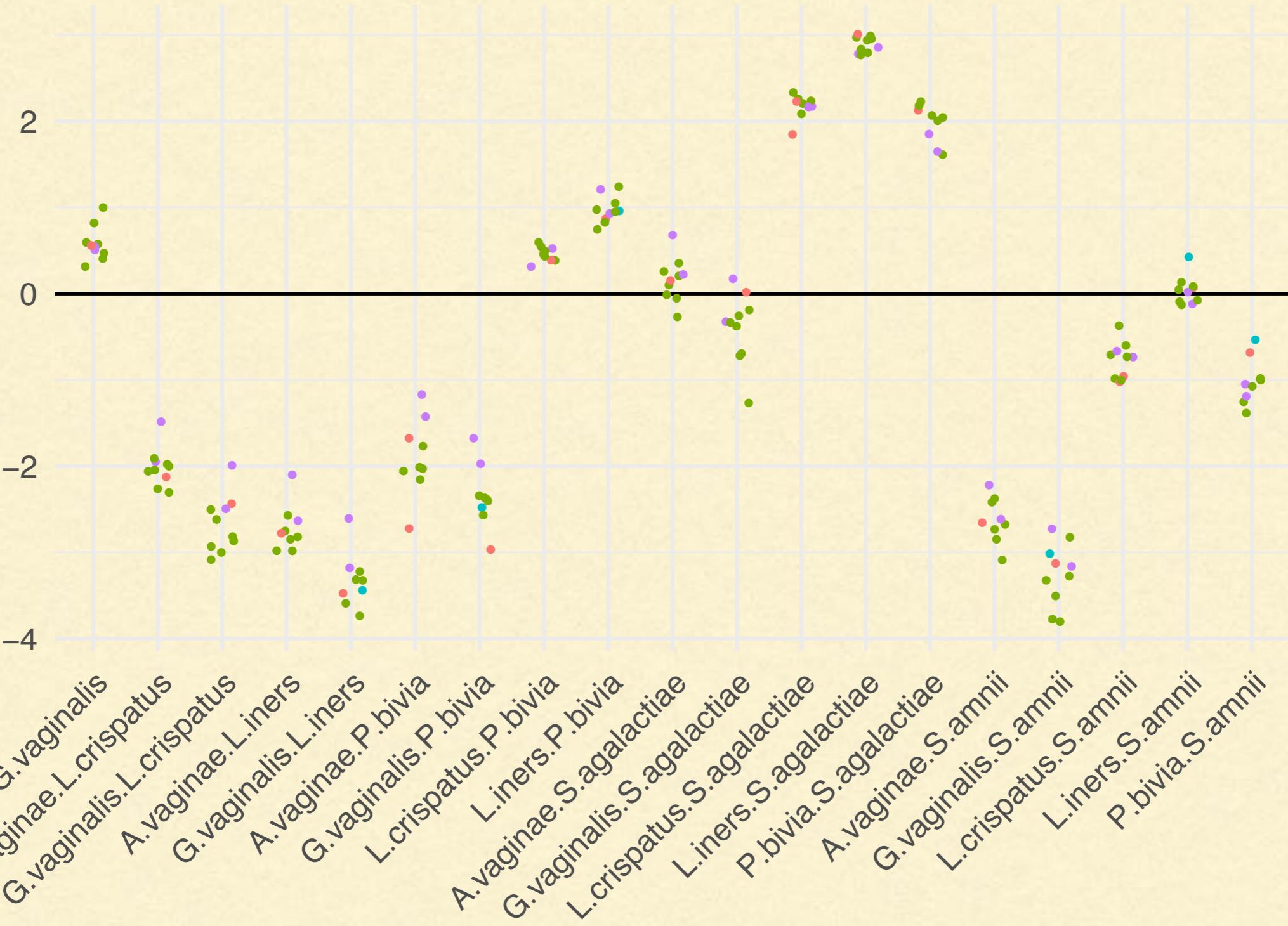
# WHAT PATTERNS CAN WE FIND?



# WHAT PATTERNS CAN WE FIND?



Difference:  
sample log-ratio and  
true log-ratio



**Very consistent!**

# A DETERMINISTIC MODEL

---

- Let  $\mathcal{O}$  be the observed abundances
- Let  $A$  be the actual abundances
- Our model is

$$\mathcal{O} \sim A \cdot B^{(P)}$$

- in expectation
- $B^{(P)}$  is the efficiency of protocol P for each taxon
- $\sim$  = "is proportional to"

# KEY RESULTS

---

- There is a single efficiency value for each taxon
- Comparing observed and actual abundances allows us to estimate B

**What drives efficiency?**

# A MULTIPLICATIVE MODEL

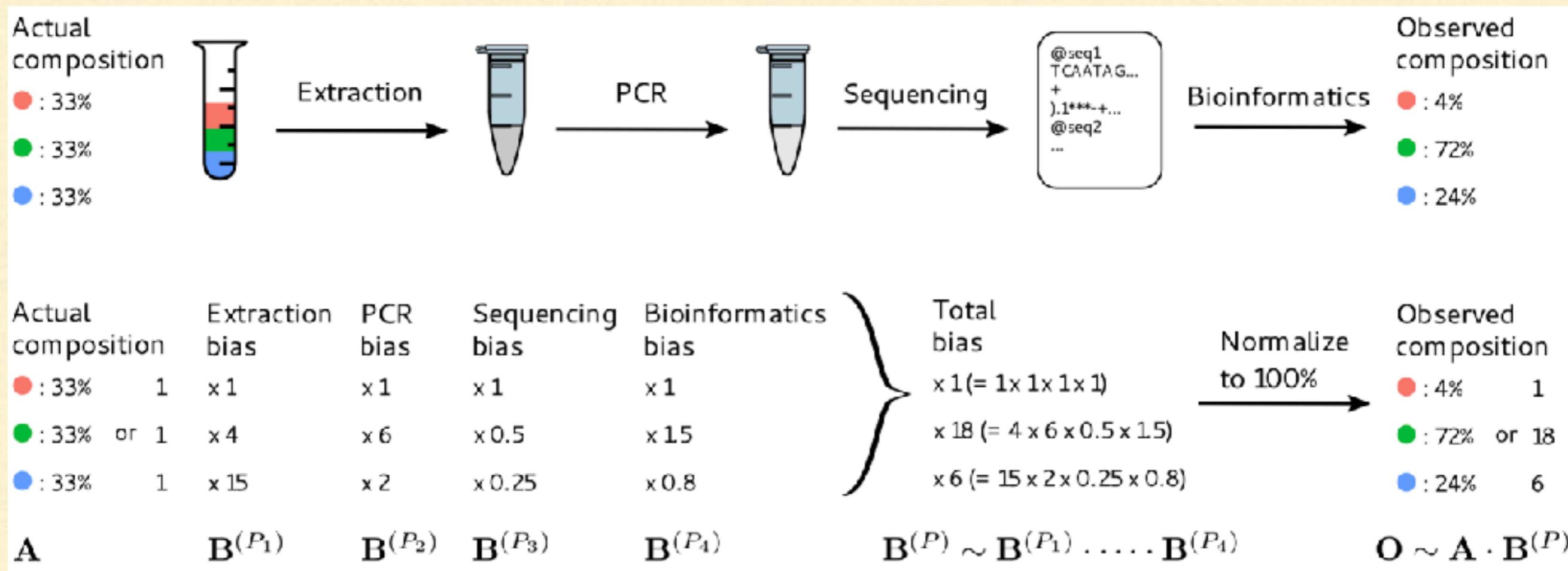
---

$$\mathcal{O} \sim A \cdot B^{(P)}$$

- Furthermore, if  $P$  has  $k$  steps  $P_1, P_2, \dots, P_k$

$$B^{(P)} \sim B^{(P_1)} \cdot B^{(P_2)} \cdot \dots \cdot B^{(P_k)}$$

- *total efficiency is the product of step-wise efficiencies*



**Consistent and correctable bias in metagenomic sequencing measurements** Michael R McLaren,  Amy D Willis,  Benjamin J Callahan**doi:** <https://doi.org/10.1101/559831>

This article is a preprint and has not been peer-reviewed [what does this mean?].

- Model for bias in observed abundances
  - Validated on 16S and shotgun metagenomic data
- Critical evaluation of model specification
- Method for bias correction *via mock communities*

# ONGOING WORK

---

- Now we know this exists, what can we do?
  - Cautious: model relative abundance, but understand limitations
  - Cynical: For taxa you really care about, only rely on qPCR to discuss abundance
  - Progressive: don't model relative abundance, model ratios
  - Aspirational: For taxa you really care about, have live cell mixtures for calibration...

# CALIBRATION WITH Q. PCR

---

- It can be done... talk with me if you have taxon-specific qPCR abundances and want to determine if efficiency is causing problems!
- Preprint coming soon...
- Williamson, Hughes & Willis (2019+) *In Prep*

# CONCLUSIONS

---

- High throughput sequencing generates measurement error
- 16S and shotgun data is biased for true microbial abundance
- This bias is consistent and multiplicative
- Statistical methods for calibration permit correct biological conclusions to be drawn from data

**Data generation process confounds  
biology and sequencing**

# ACKNOWLEDGEMENTS



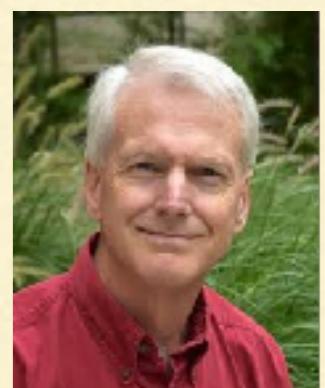
David  
Clausen  
(UW)



Ben  
Callahan  
(NCSU)



Michael  
McLaren  
(NCSU)



Jim  
Hughes  
(UW)



Brian  
Williamson  
(UW)

- McLaren, Willis & Callahan, 2019+, *eLife*
  - Model specification for bias in microbiome data
- Clausen & Willis, 2019+, *In Prep*
  - Estimating the reproducibility of microbiome data
- Williamson, Hughes & Willis, 2019+, *In Prep*
  - Calibration of abundance using qPCR controls