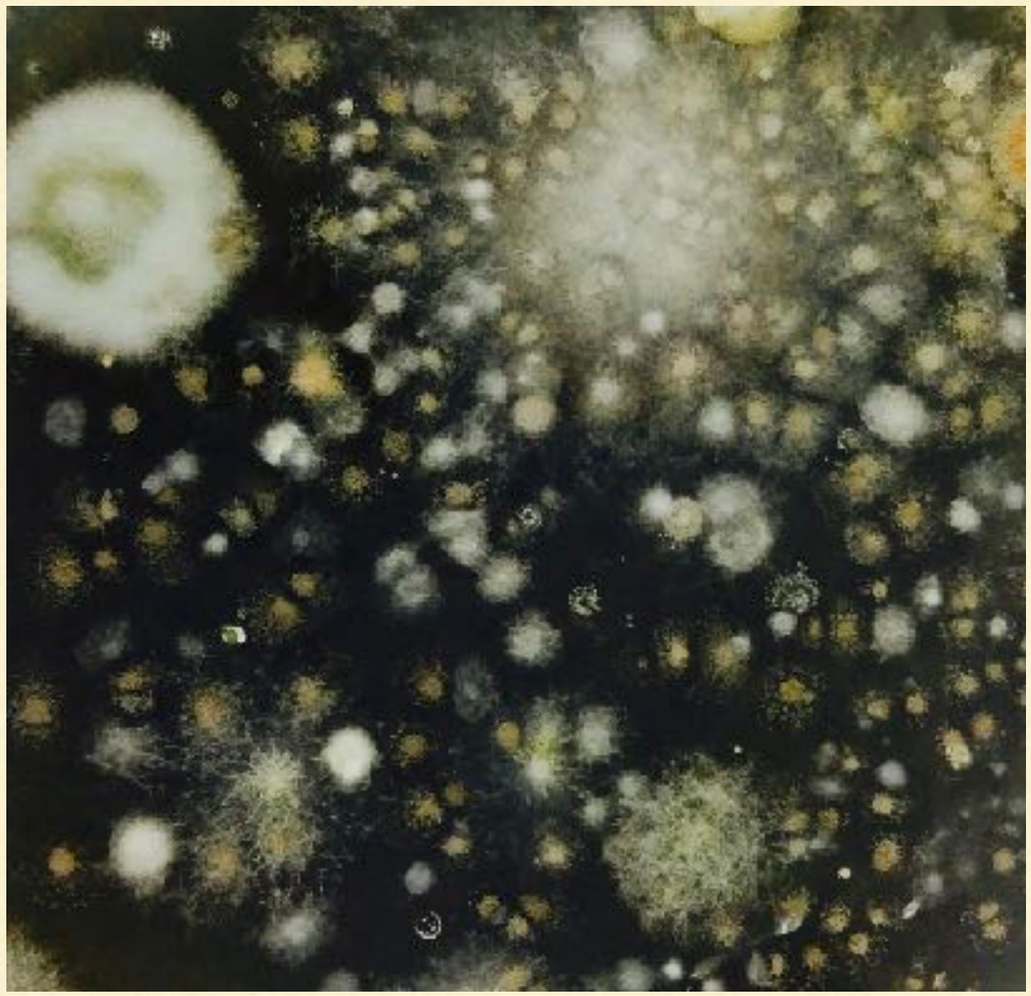


INTRODUCTION TO ESTIMATION

Research Group: Statistical Diversity Lab

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWwillis adwillis@uw.edu

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

"How do I rigorously analyse my data?"

–Everyone, all the time

Joke credit: 🌽 Bryan Martin 🌽

"It depends."

–Stat Div Lab, all the time

Note: Not a joke

THE PLAN

1. What can we estimate with compositional data?
2. Lecture: Relative abundance
3. Lab: Relative abundance
    1. intro: 🌽corncob🌽 using 16S (Bryan)
    2. 🌽corncob🌽 using shotgun data (Bryan)
    3. How to process shotgun data to use with 🌽corncob🌽 (Taylor)
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**

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 = **0.168**	1/6 = **0.167**	3.99 / 6 = **0.665**	1
ENVIRO 2	0.99 / 6 = **0.165**	0.99 / 6 = **0.165**	4.02 / 6 = **0.67**	1

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

*with some caveats, like always — more later

16S & WGS DATA ARE COMPOSITIONAL

- We analyze compositional data differently than non-compositional data

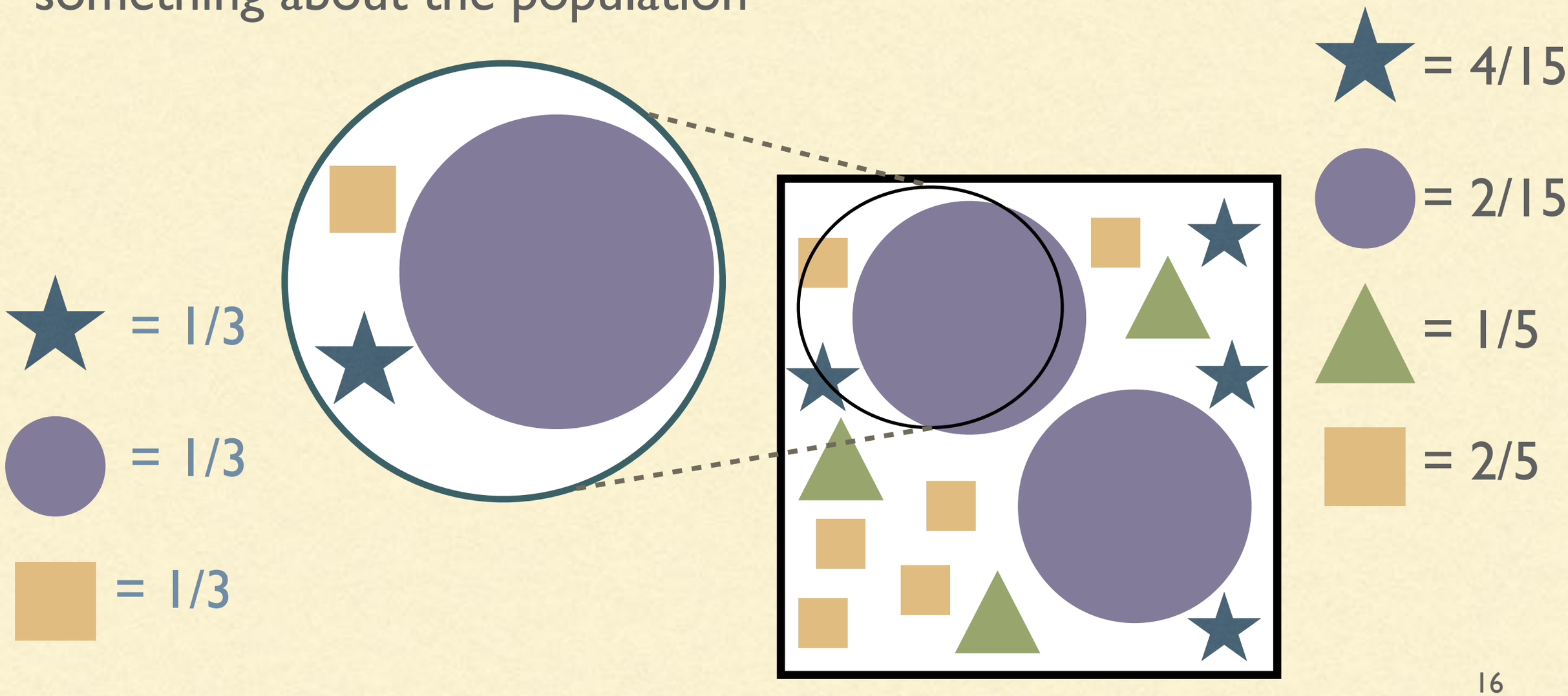
- Common (users/software): convert to proportions
  - This is not necessary!
  - It loses information about precision!
  - Good statistical methods model precision

16S & 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

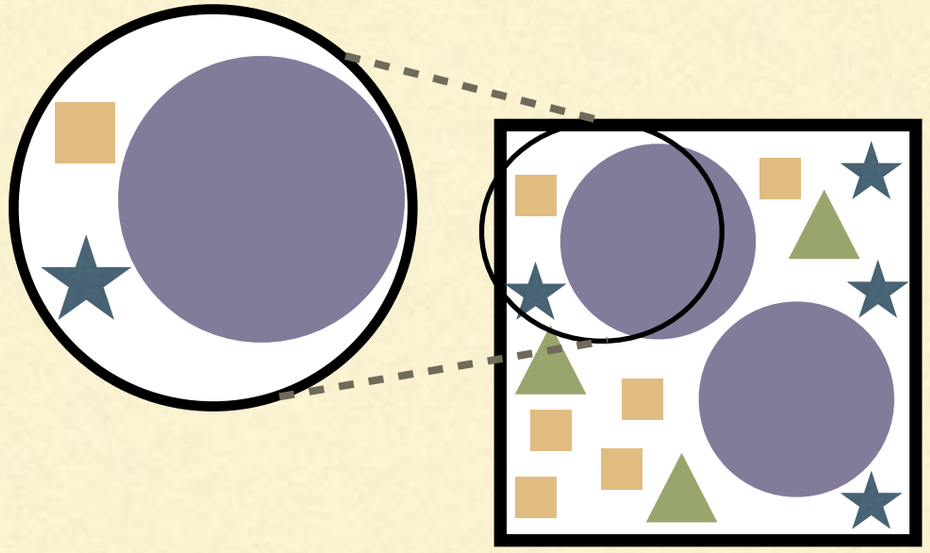


COMPOSITIONAL DATA

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

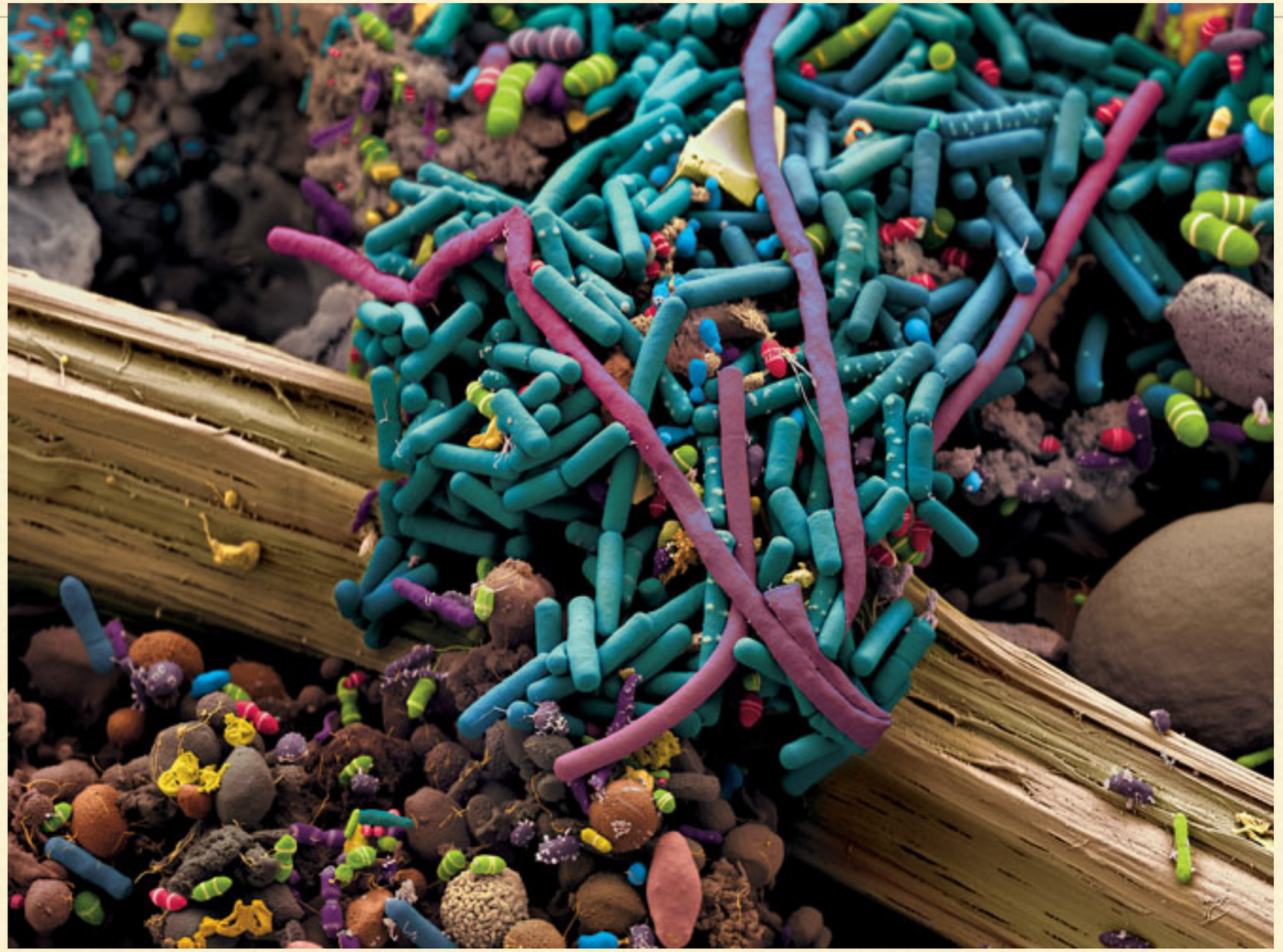
THE PROBLEM

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



PARAMETERS FOR COMPOSITIONAL DATA

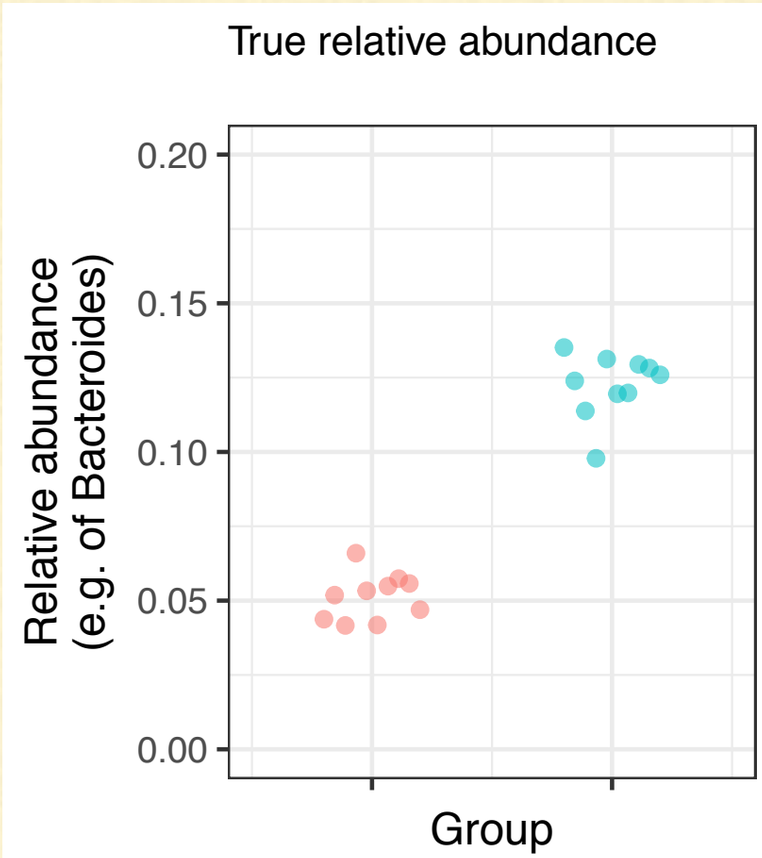
- Relative abundance of taxa/genes
- Diversity parameters:
  - α-diversity
  - β-diversity
- Presence/absence of taxa/genes
- Abundance of taxon 1 divided by abundance of taxon 2…



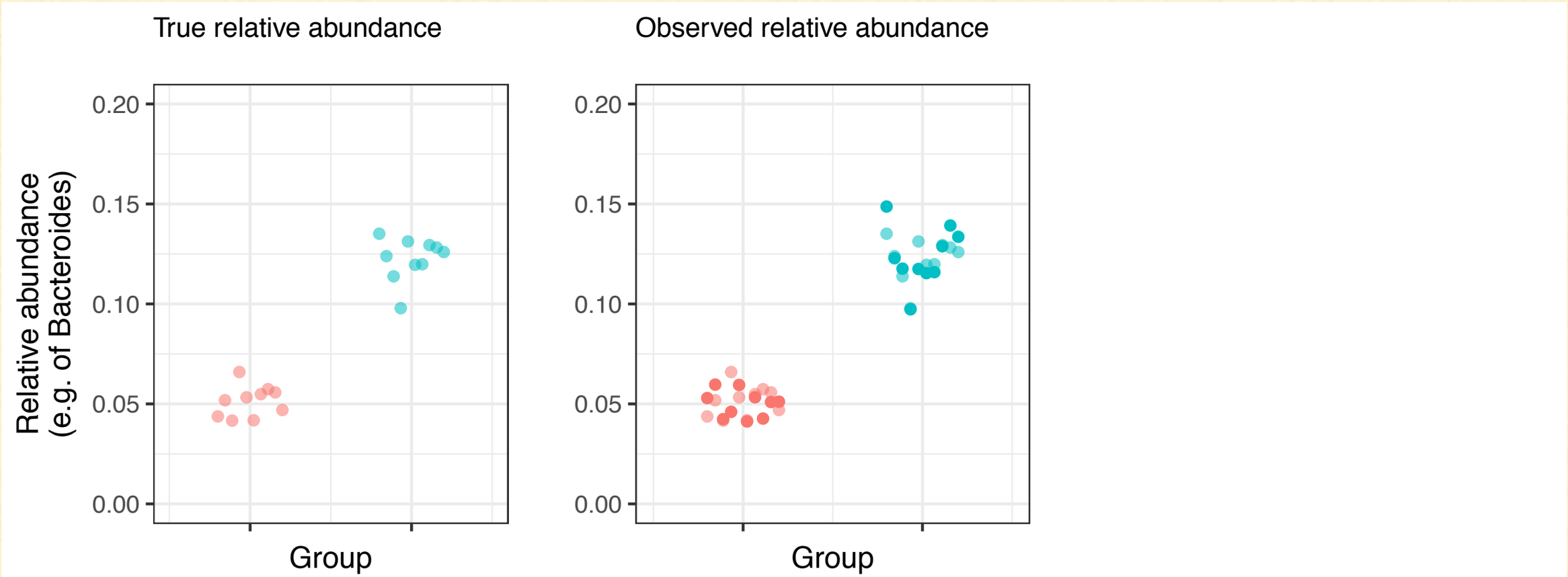
MODELING RELATIVE ABUNDANCE

Photo credit: Martin Oeggerli

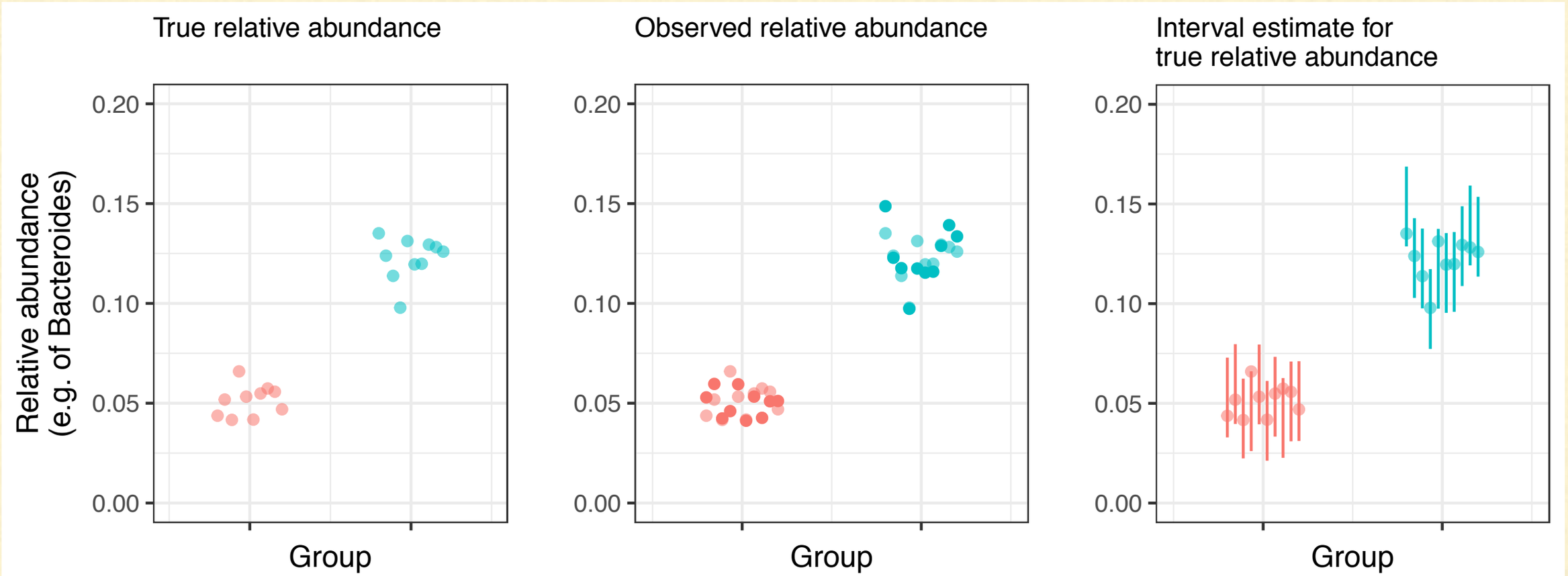
SAMPLE VS POPULATION



SAMPLE VS POPULATION

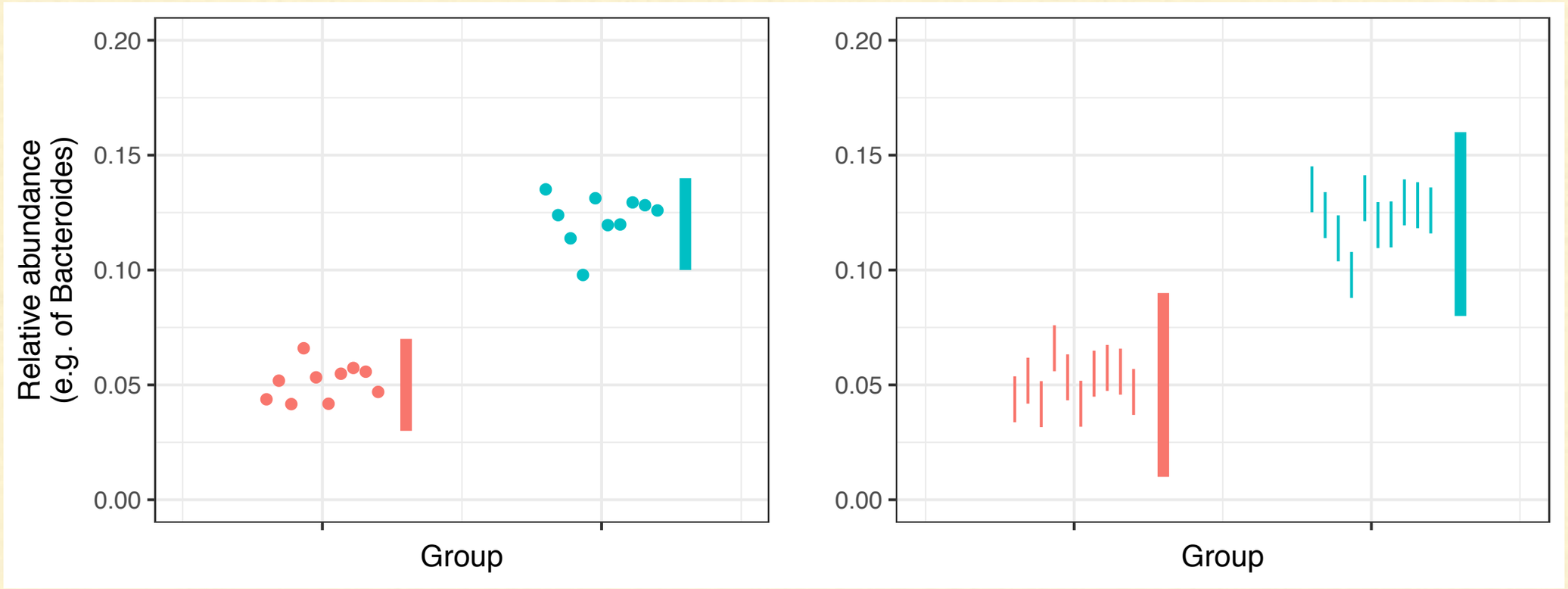


SAMPLE VS POPULATION



SAMPLE ≠ POPULATION

- Observed relative abundance ≠ 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

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

Bryan Martin, UW Statistics

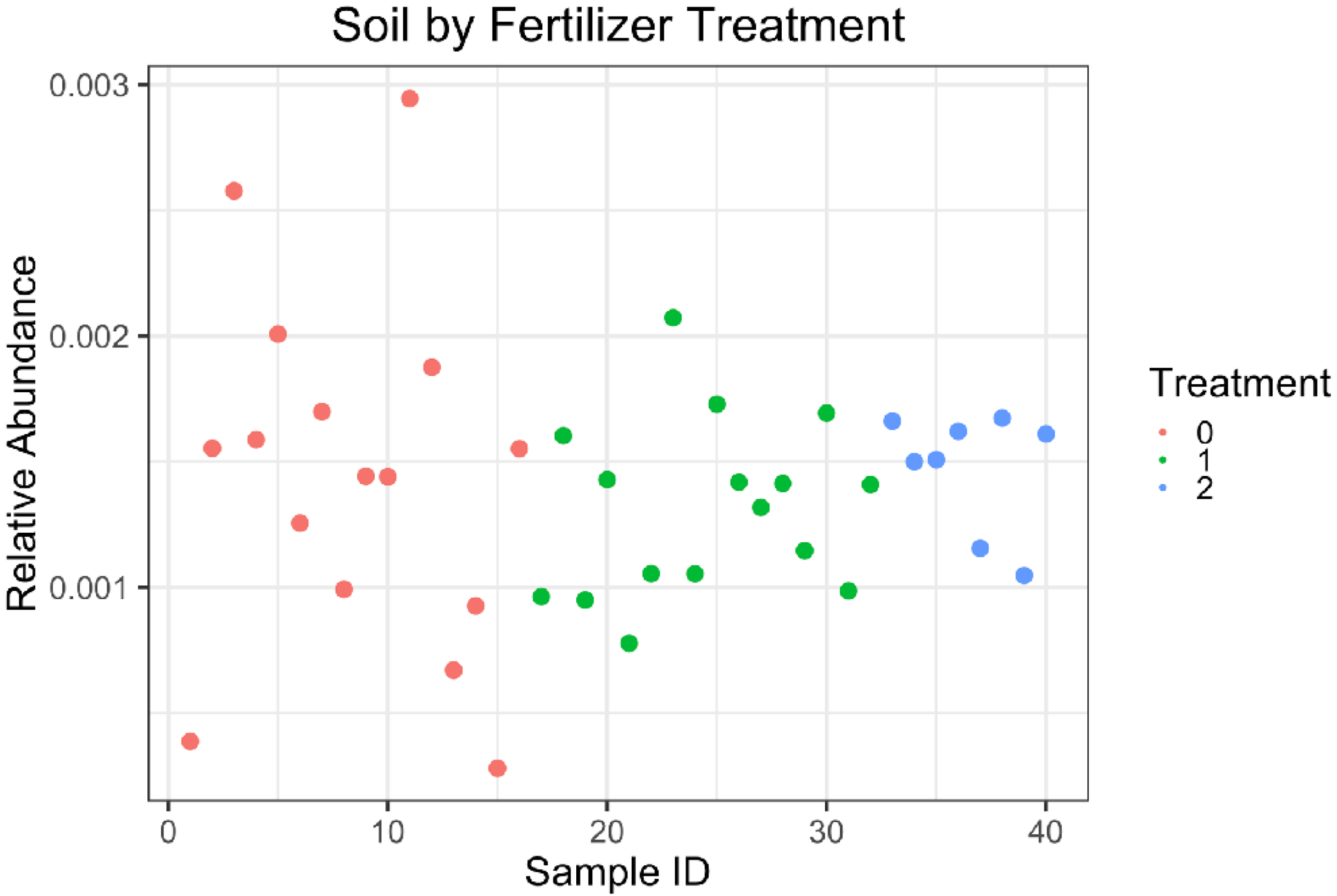
Daniela Witten, UW Statistics

CORNCOB

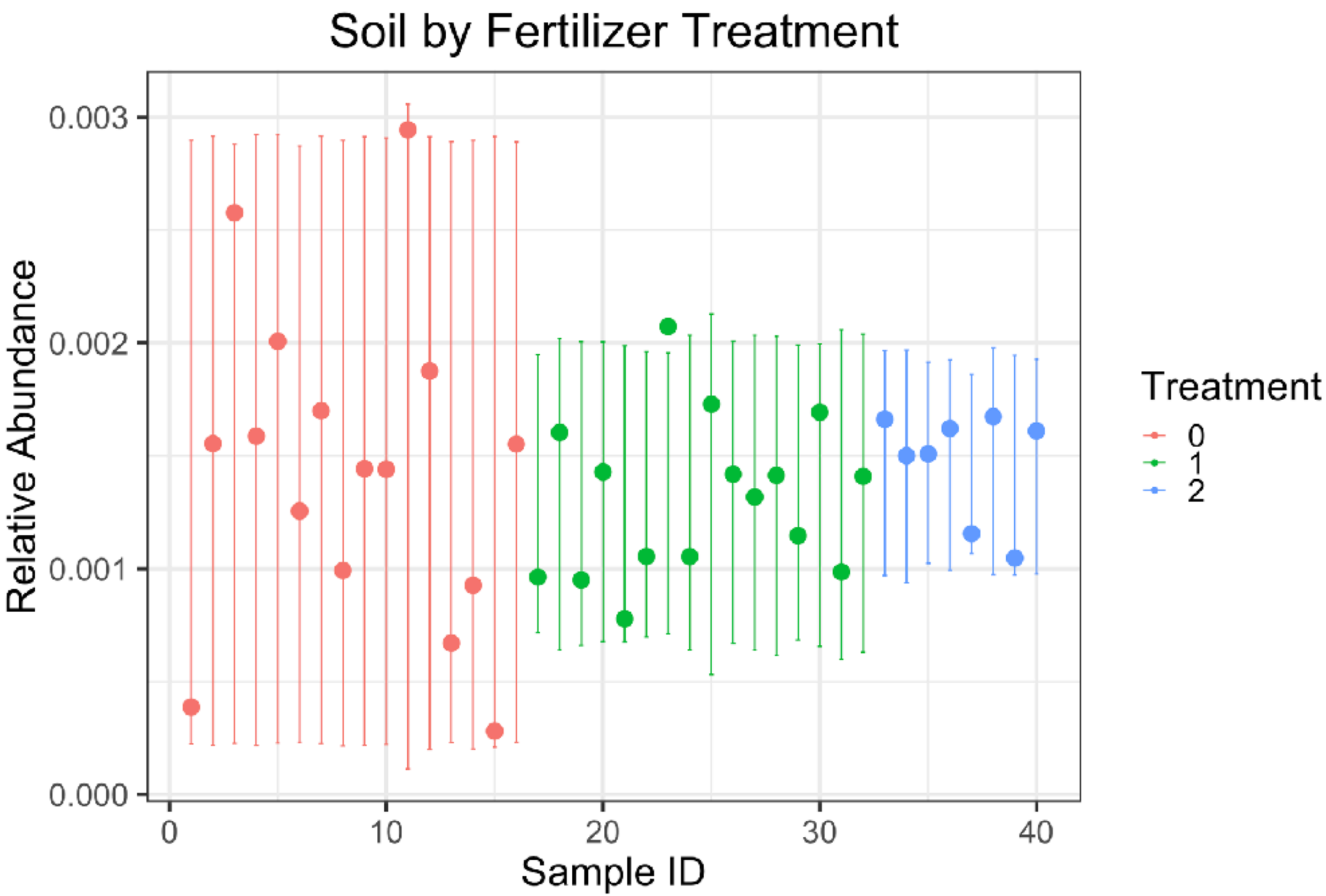
COmpositional RegressioN for Correlated Observations with the Beta-binomial

- "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-link function for covariates associated with the mean and the dispersion. Differential abundance (DA) was modeled as a function of treatment group and age, and 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*



Data from Whitman et. al, (2015) *ISME*

CORNCOB

COmpositional RegressioN 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 RegressioN 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

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)

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”

POSSIBLE USES: WE WANT TO MODEL…

- The number of amplicon reads that you give a taxonomic label (e.g. *L. iners)*, out of the total number of amplicon reads

- Shotgun reads that are recruited by a MAGs, 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 taxonomic label (e.g. *L. iners)*, out of the total number of amplicon reads

- Shotgun reads that are recruited by a MAGs, 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

HOW DOES 🌽🌽🌽 WORK?

BETA-BINOMIAL DISTRIBUTION

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

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

- **n** = samples, indexed by **i** = 1, …, 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

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

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

what you need

- **n** = samples, indexed by **i** = 1, …, 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**

LINKING ABUNDANCE TO COVARIATES

1. Parameters

$$\mu_i = \frac{a_{1,i}}{a_{1,i} + a_{2,i}},$$

"(latent) relative abundance"

$$\phi_i = \frac{1}{a_{1,i} + a_{2,i} + 1}$$

"within sample correlation"
"absolute abundance overdispersion"

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

$$\mu_i = \frac{a_{1,i}}{a_{1,i} + a_{2,i}},$$

"(latent) relative abundance"

$$\phi_i = \frac{1}{a_{1,i} + a_{2,i} + 1}$$

"within sample correlation"
"absolute abundance overdispersion"

2. Link to covariates

what you need

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

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

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
```

CORNCOB

COmpositional RegressioN 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 OR differential variability, controlling for the effect of the covariate on mean abundance

CORNCOB

COmpositional RegressioN for Correlated Observations with the Beta-binomial

- CORNCOB does not assume microbes behave independently
- Parameter ϕ 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!*

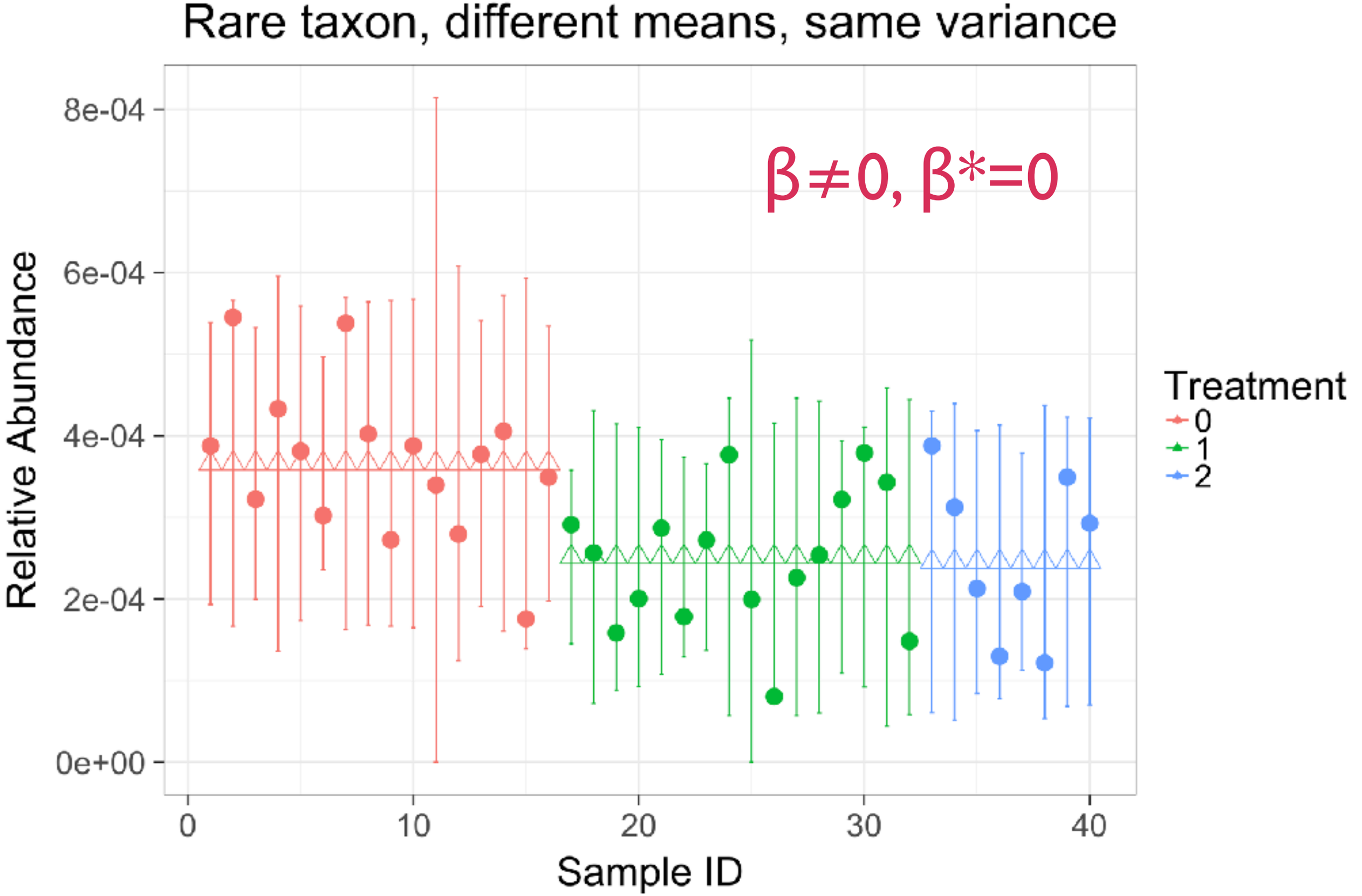
CORNCOB AND DESEQ2

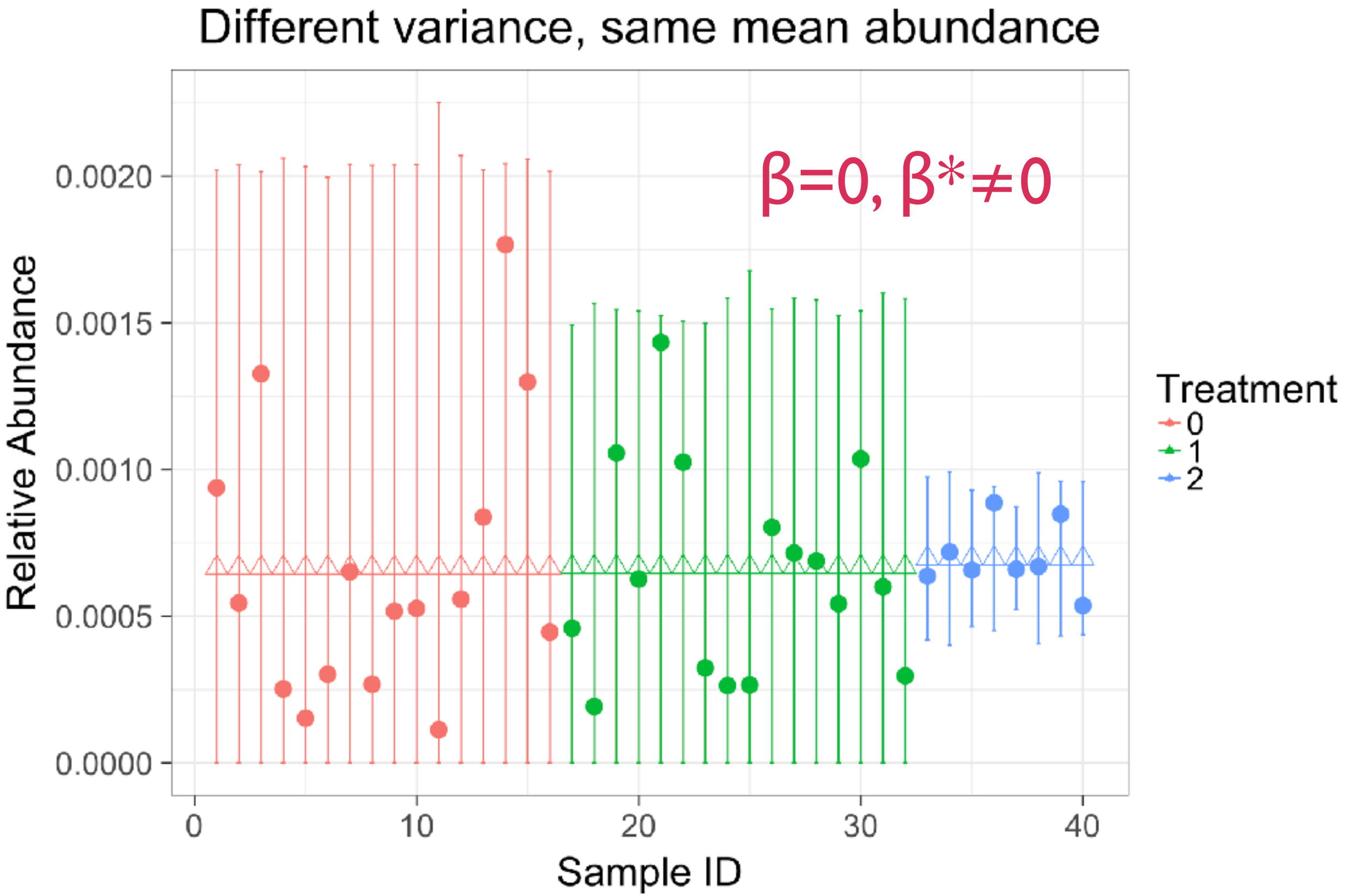
- **Similarities**
  - Easy to use with phyloseq (easier, in fact)
  - Use un-normalized counts to assess precision of estimates
  - Benjamini-Hochberg adjustment for multiple comparisons
  - Dispersion parameter for overdispersion
  - Tests for differential abundance

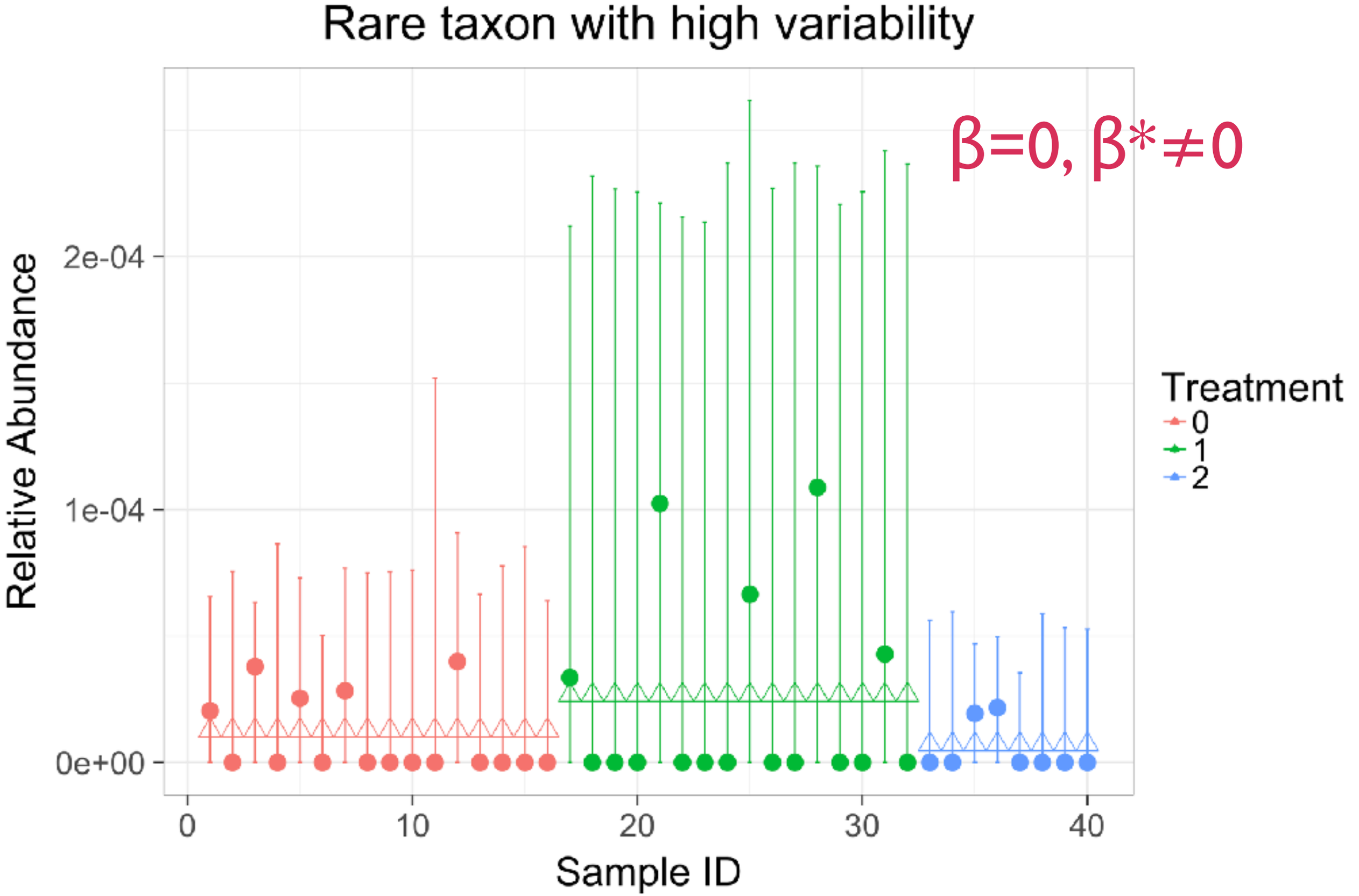
CORNCOB AND DESEQ2

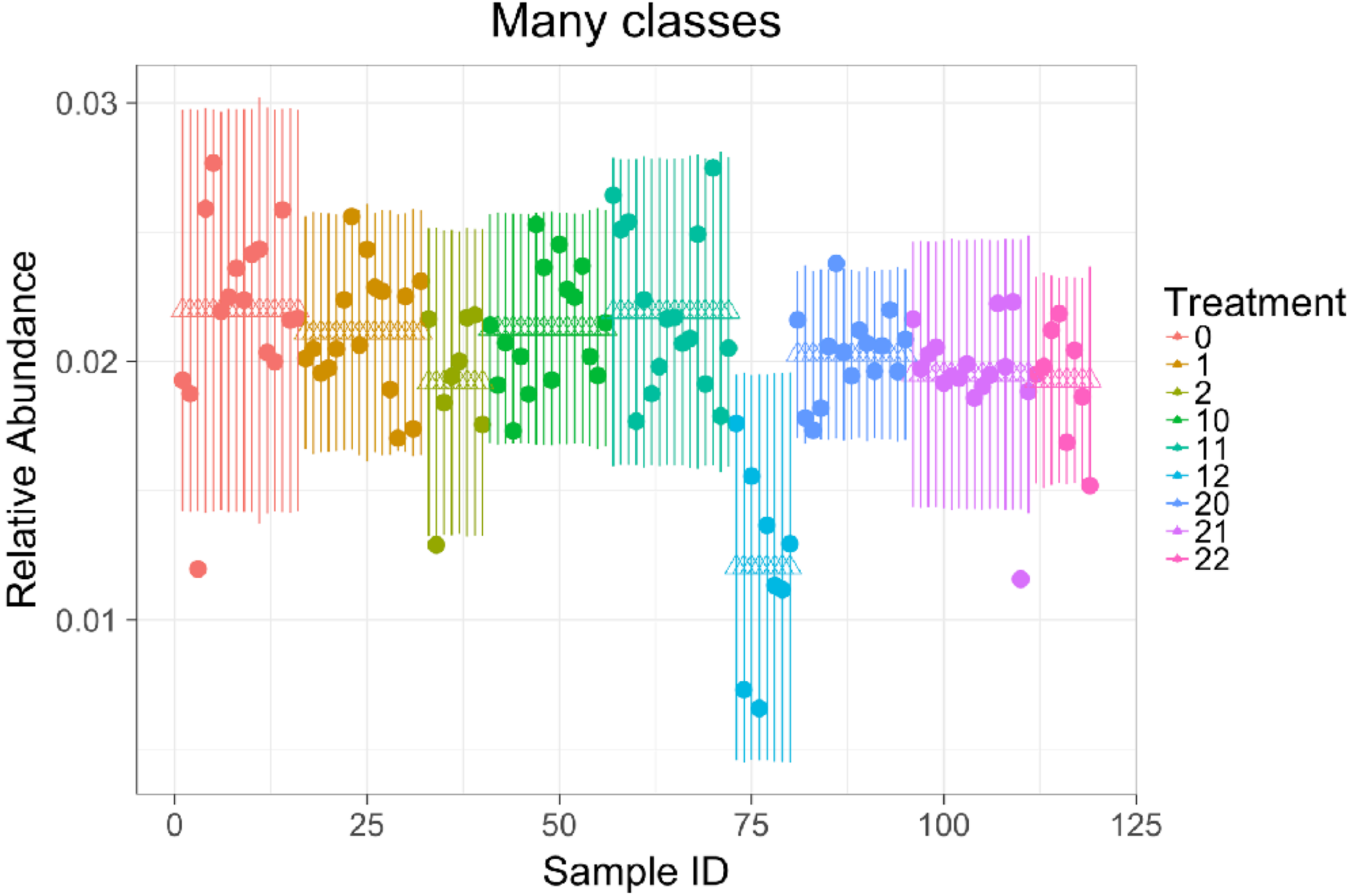
- Designed for marker gene (compositional) data
- Models relative abundance & overdispersion
- Different structure for different taxa
- Uses within-taxon correlation to model zeros
- Easy to diagnose model misspecification

- Designed for RNAseq (different data structure)
- Tests changes in abundance
- Constrained dispersion
- Individual microbes are assumed independent
- Not so easy to diagnose problems

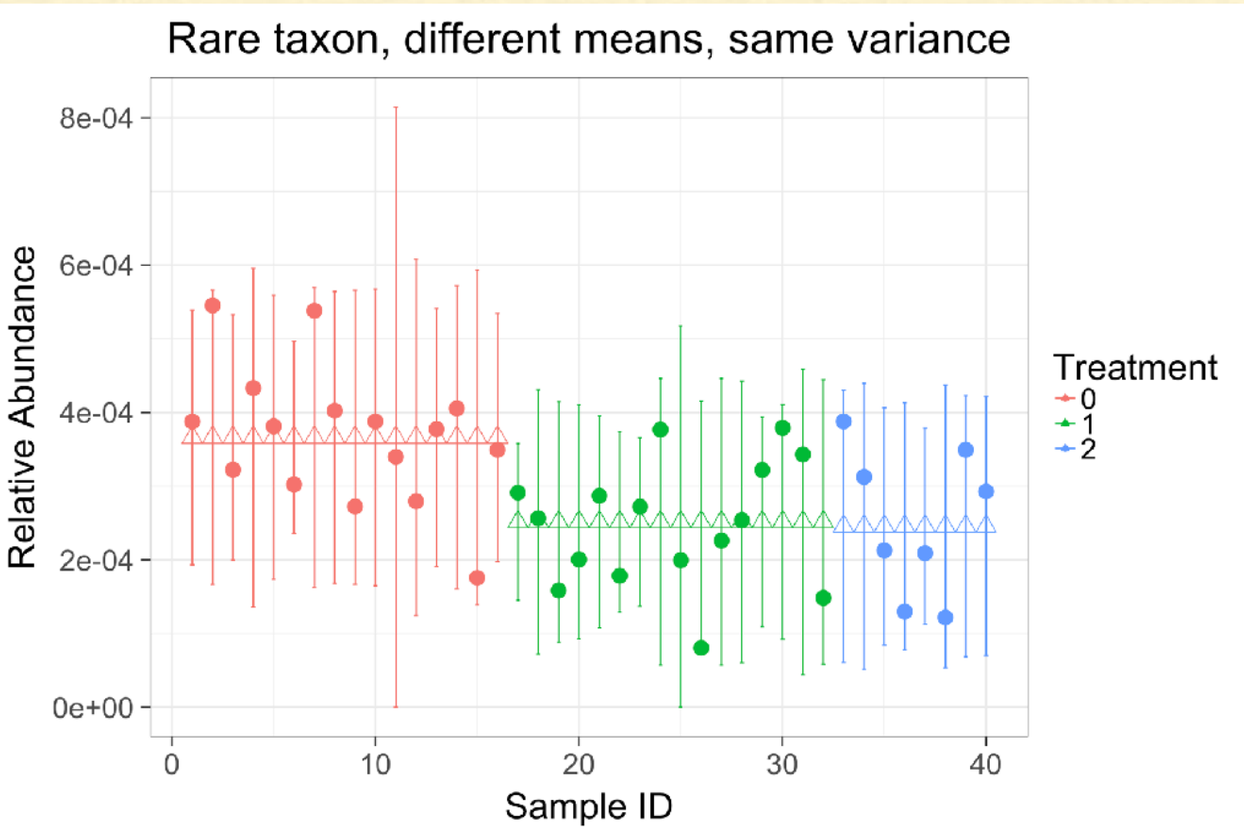

Rare taxon, different means, same variance
β≠0, β*=0
Relative Abundance
8e-04
6e-04
4e-04
2e-04
0e+00
0
10
20
30
40
Sample ID
Treatment
0
1
2

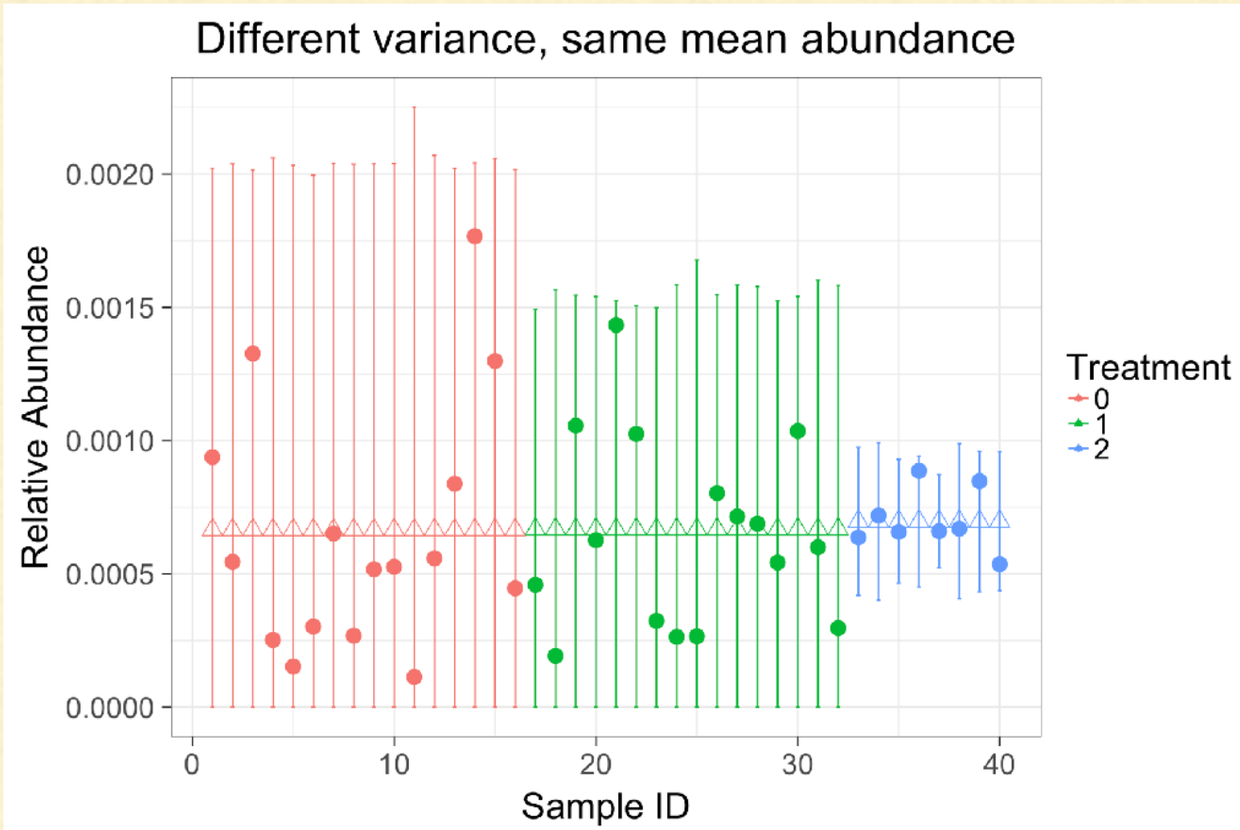

Different variance, same mean abundance
β=0, β*≠0
Relative Abundance
0.0020
0.0015
0.0010
0.0005
0.0000
0
10
20
30
40
Sample ID
Treatment
0
1
2

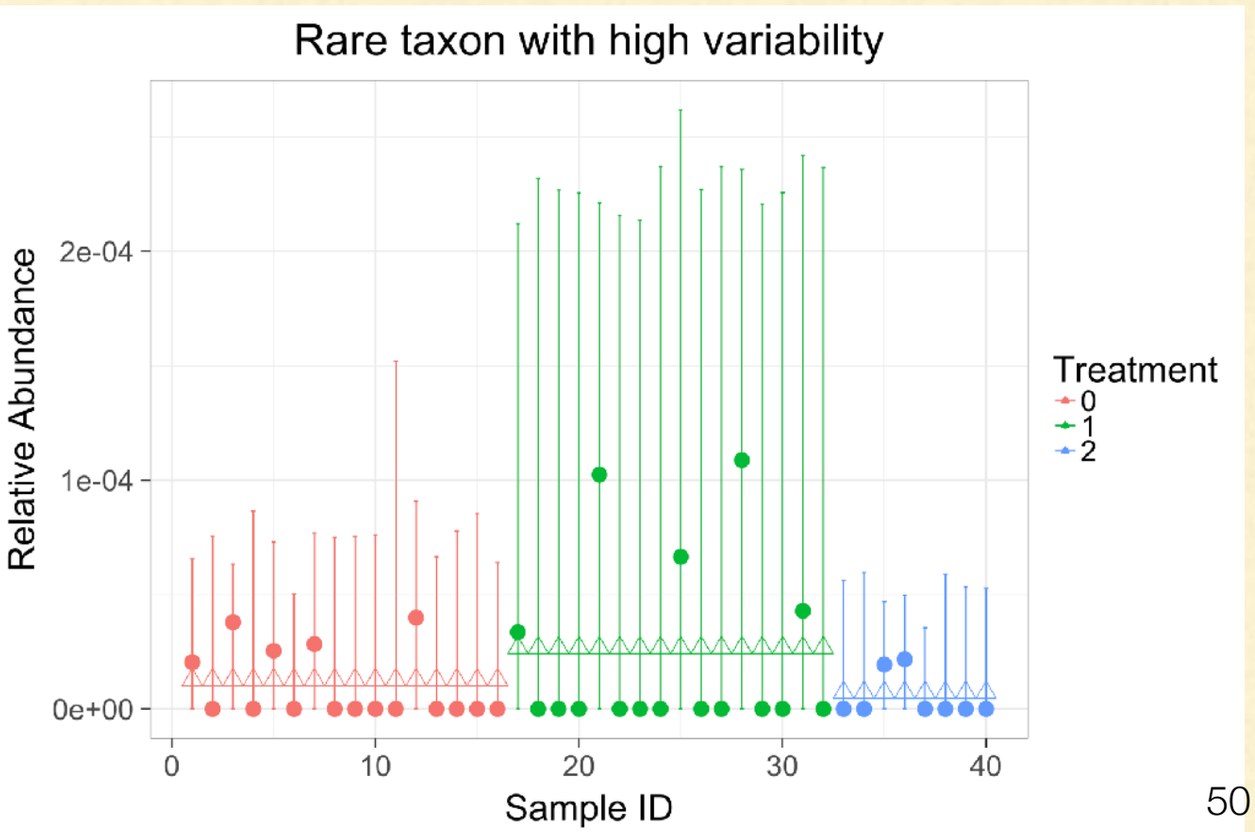

Rare taxon with high variability
β=0, β*≠0
Relative Abundance
2e-04
1e-04
0e+00
0
10
20
30
40
Sample ID
Treatment
0
1
2

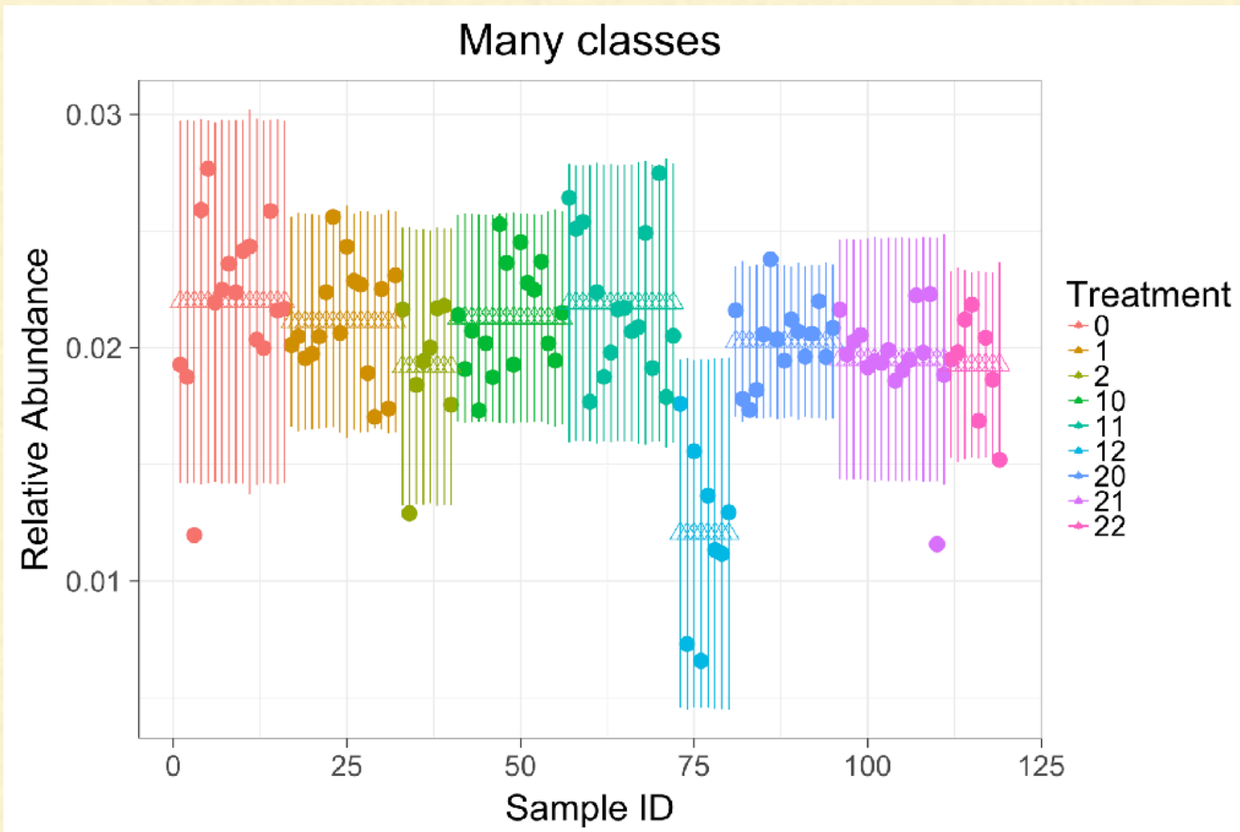

Many classes
0.03
0.02
0.01
Relative Abundance
0
25
50
75
100
125
Sample ID
Treatment
0
1
2
10
11
12
20
21
22

Model Fit


Rare taxon, different means, same variance
Relative Abundance
8e-04
6e-04
4e-04
2e-04
0e+00
0
10
20
30
40
Sample ID
Treatment
0
1
2


Different variance, same mean abundance
Relative Abundance
0.0020
0.0015
0.0010
0.0005
0.0000
0
10
20
30
40
Sample ID
Treatment
0
1
2


Rare taxon with high variability
Relative Abundance
2e-04
1e-04
0e+00
0
10
20
30
40
Sample ID
Treatment
0
1
2

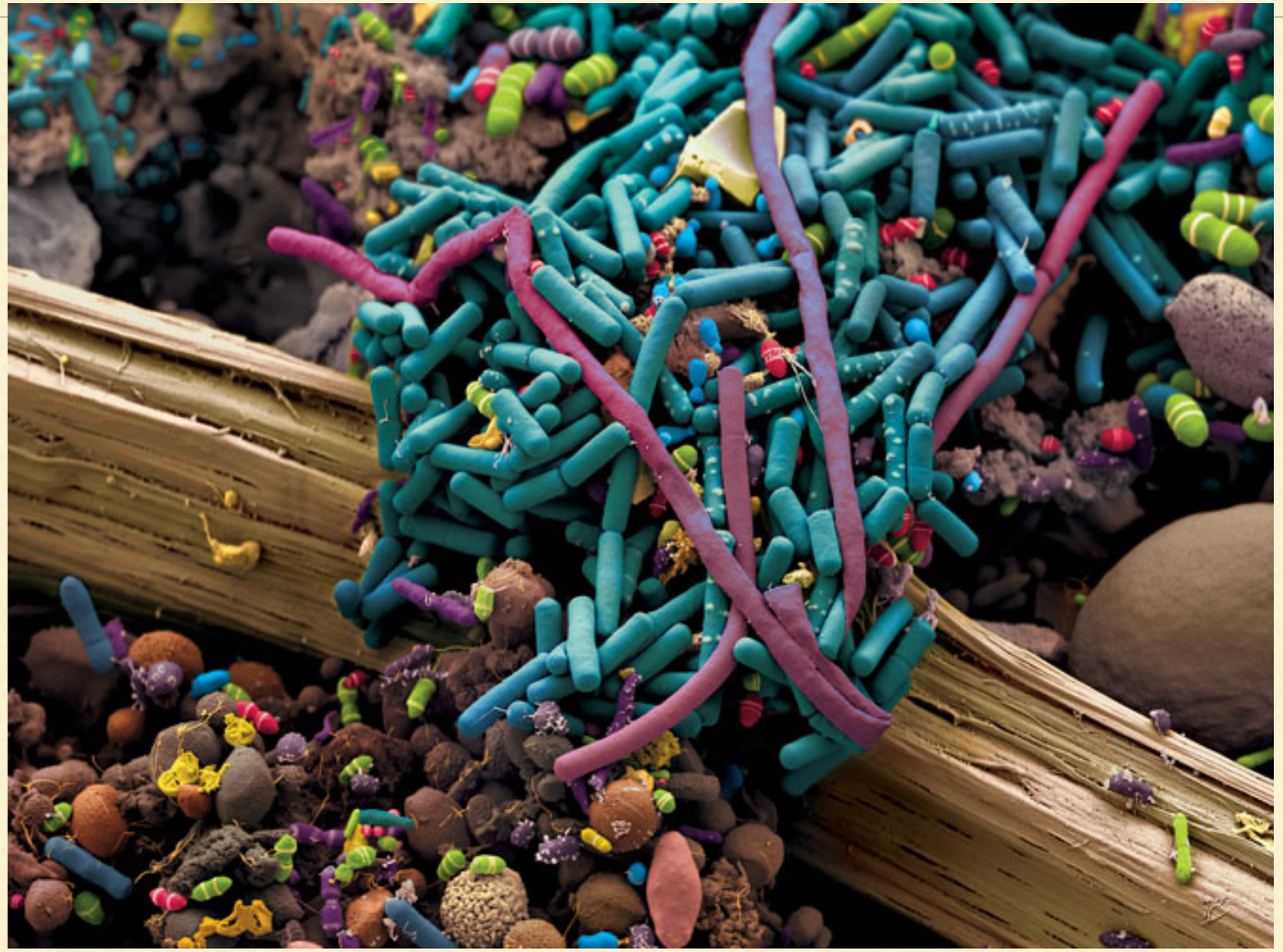

Many classes
Relative Abundance
0.03
0.02
0.01
0
25
50
75
100
125
Sample ID
Treatment
0
1
2
10
11
12
20
21
22

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
  - 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

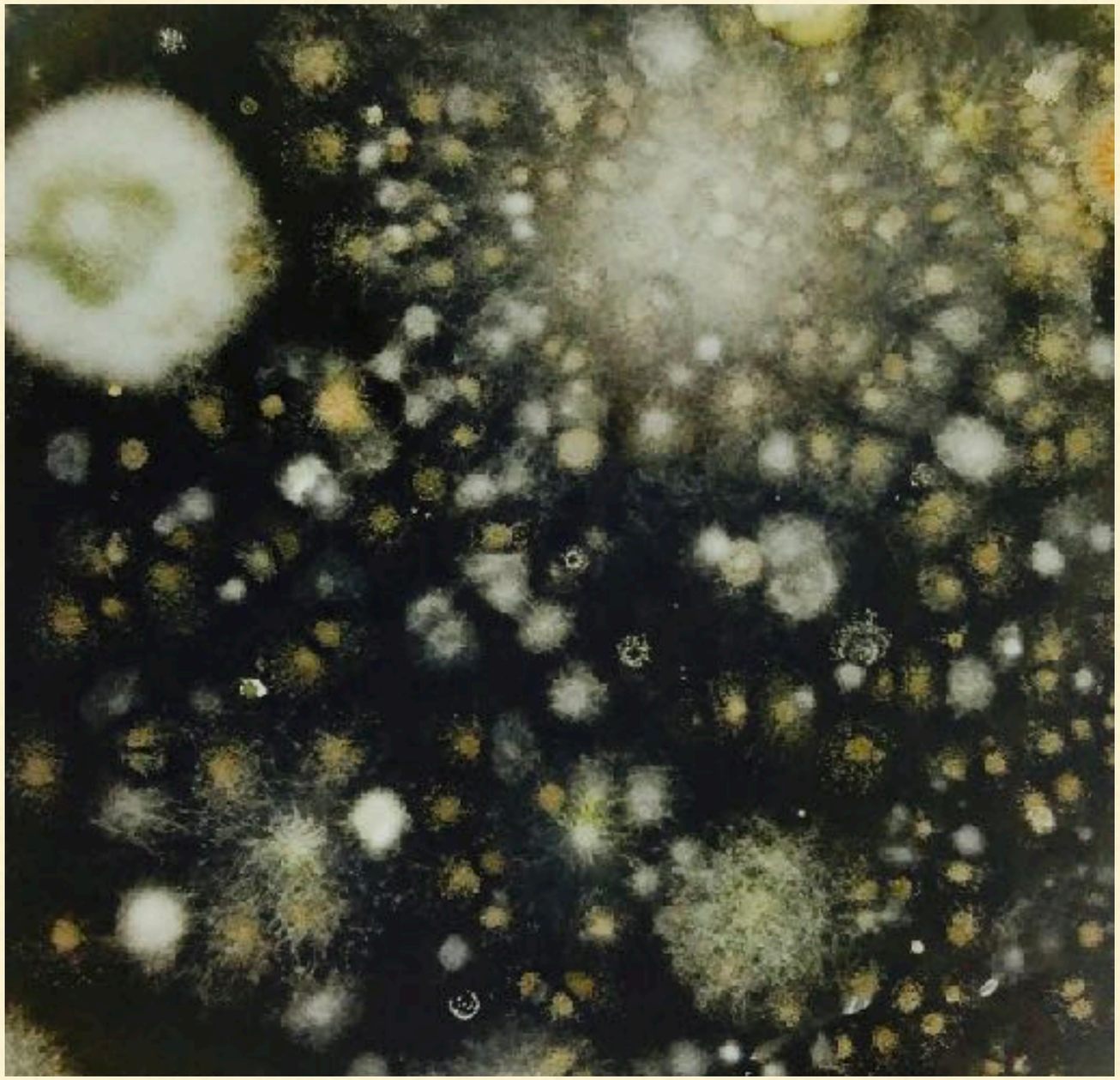
Photo credit: Martin Oeggerli

PAULINE AND BRYAN

- Please make instructions here

OLD ABUNDANCE LABS

- **github.com/statdivlab/stamps2019**
- Click on **labs** then **corncob_tutorial**
- Click on **corncob-tutorial.R**
- Click on **raw**
- Copy and paste into an R script and start reading and working through
  - ~30 minutes



DIVERSITY

Photo credit: Martin Oeggerli

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)

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

- 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
    - 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, …, 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

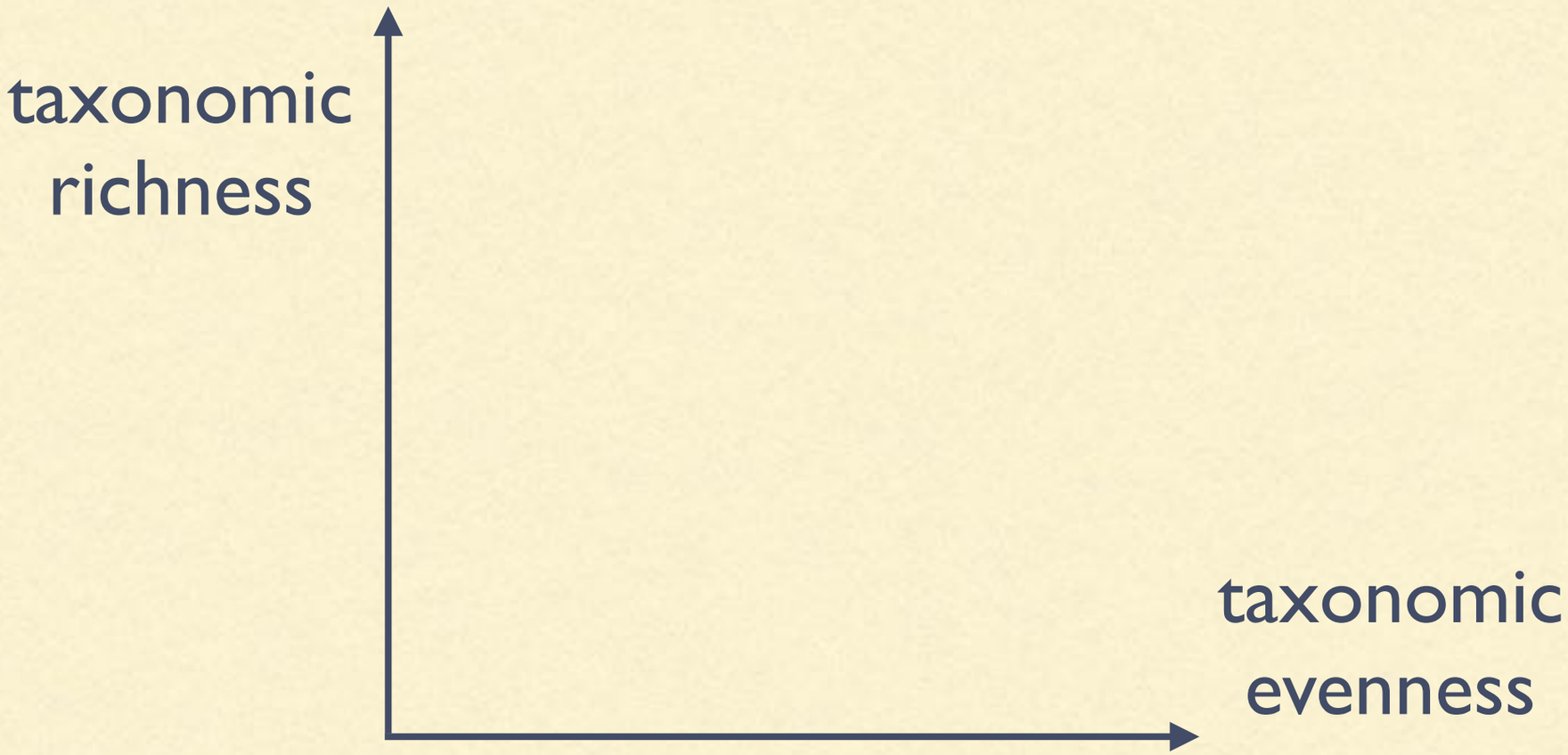
is a valid α-diversity parameter

ALPHA DIVERSITY

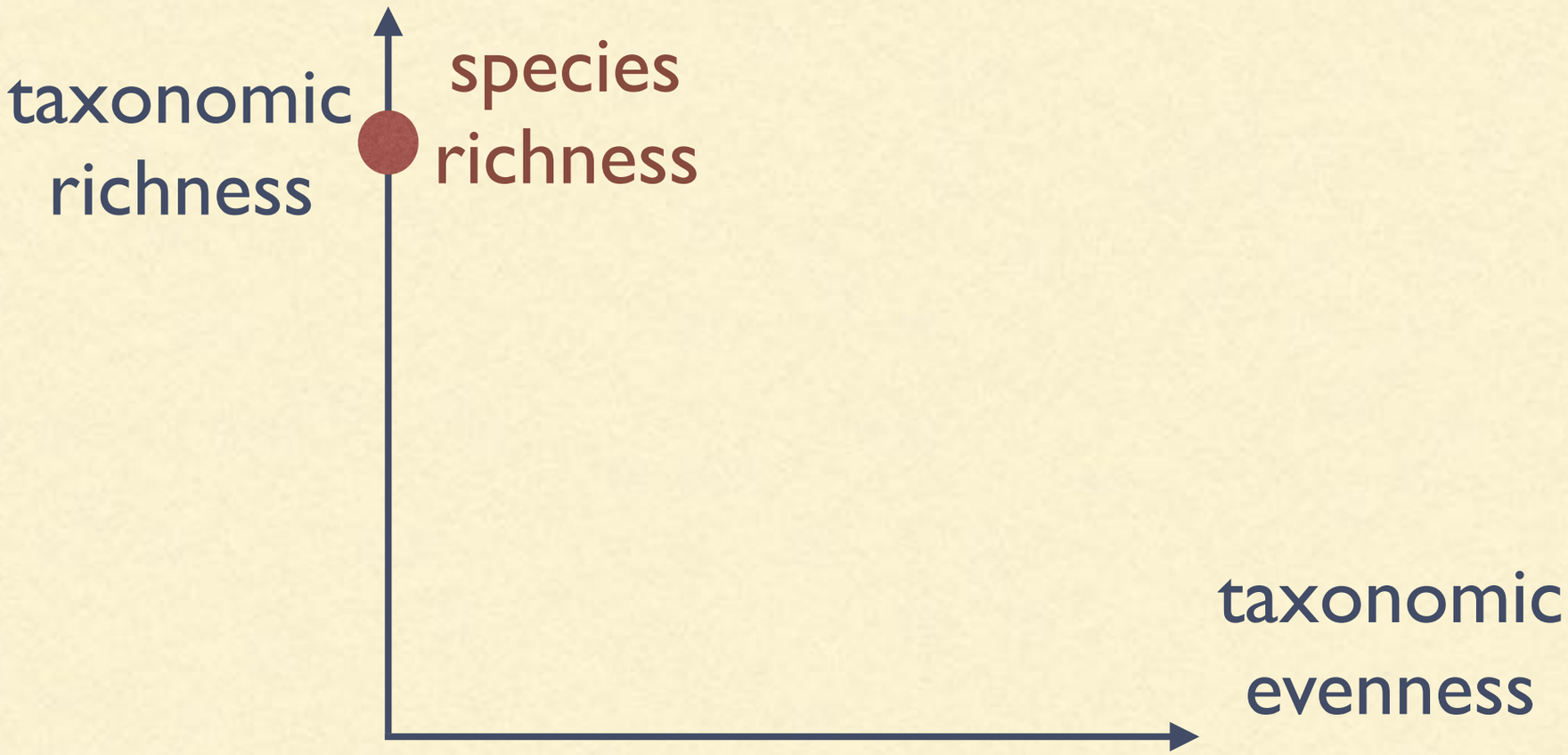
- Some examples of α-diversity measures include
  - Species richness: C
  - Simpson's index: $\sum_{i=1}^{C} p_i^2$
  - Shannon diversity: $-\sum_{i=1}^{C} p_i \ln p_i$
  - Shannon's E: $\frac{-\sum_{i=1}^{C} p_i \ln p_i}{\ln C}$

YOUR CHOICE

- Think: What difference do you want to highlight?



YOUR CHOICE



YOUR CHOICE

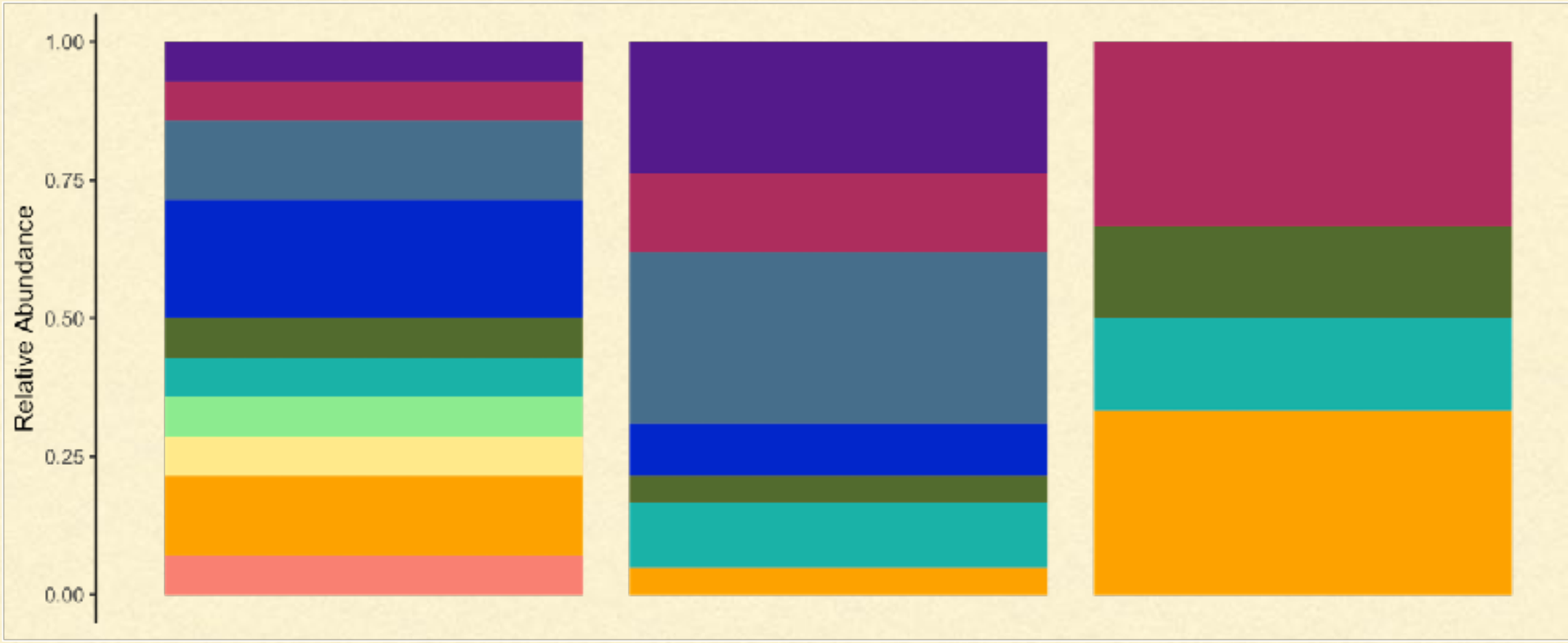

taxonomic
richness
species
richness
Shannon's E
taxonomic
evenness

YOUR CHOICE


taxonomic
richness
species
richness
Shannon
Shannon's E
taxonomic
evenness

YOUR CHOICE

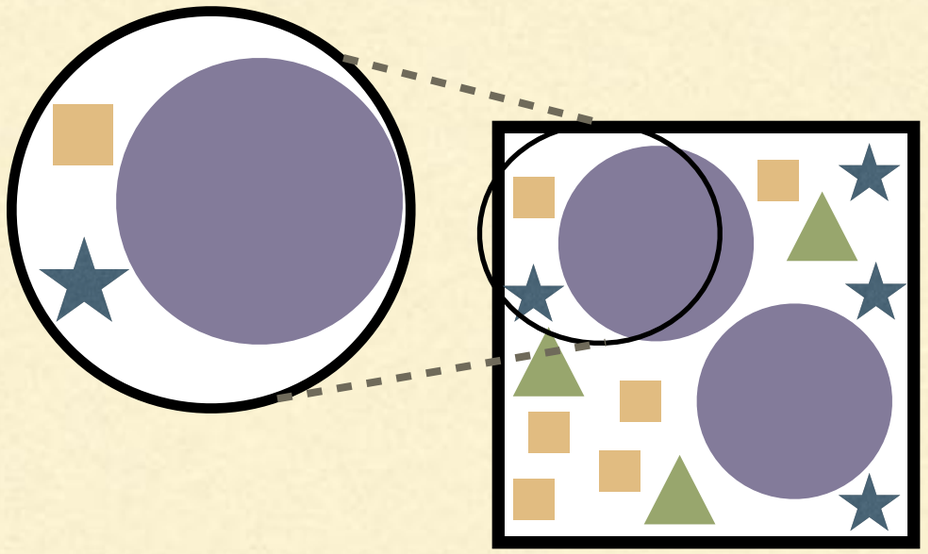
This is a question of *parameter choice:*
Which parameter highlights the differences I care about?



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
Inverse Simpson's	8.17	4.98	3.60

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

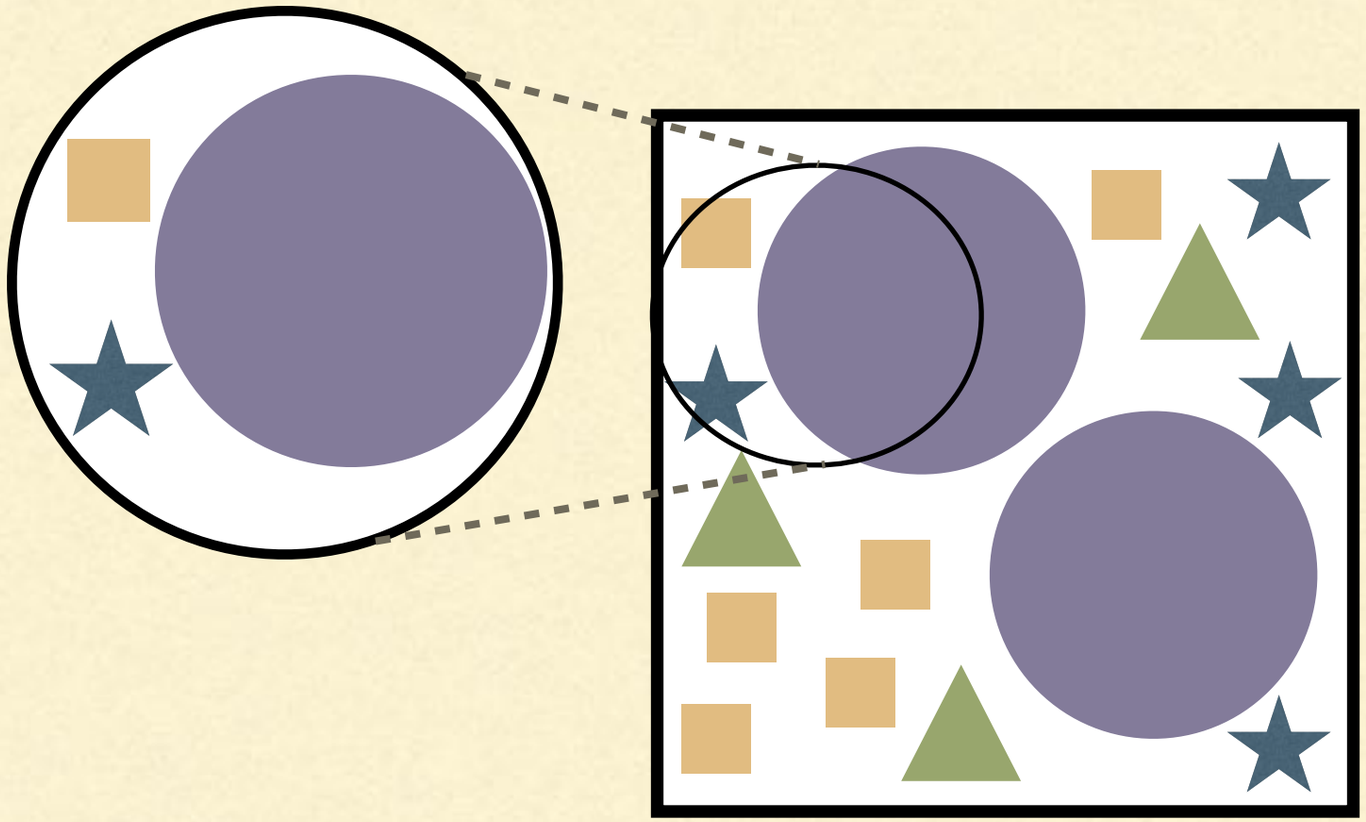
THE "~~CLASSICAL~~" APPROACH

- 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

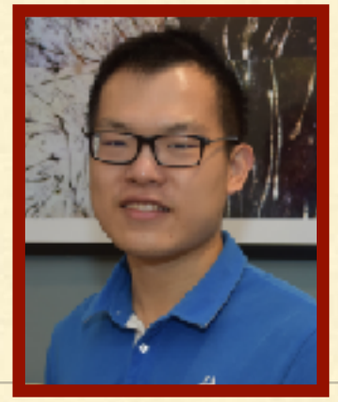
- 2 things are wrong here:
  - The bias (under/overestimation)
  - The variance (how big are the error bars — you'll never be exactly right)

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

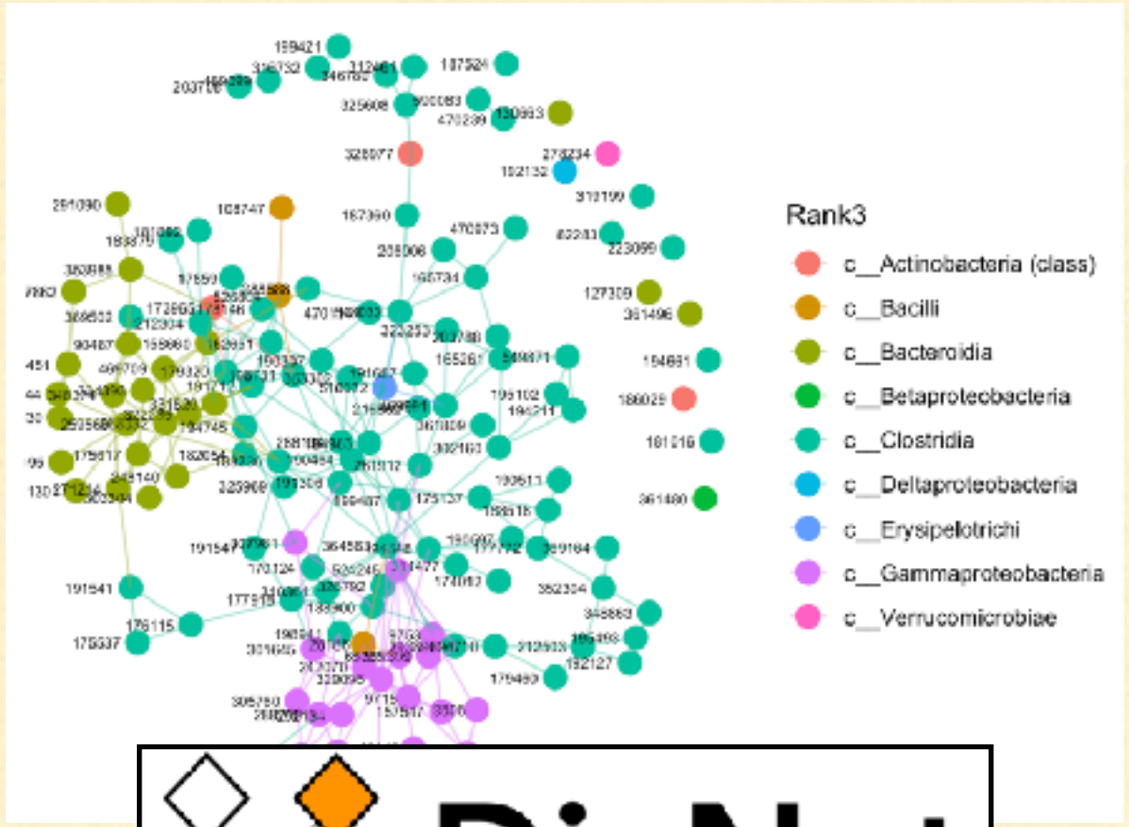
SPECIES RICHNESS ESTIMATION

- 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 $p_1, p_2, \ldots, p_C$, then use network models to get variance



DivNet

DIVNET

Bryan Martin

Pauline Trinh

- This idea works for estimating any diversity index (**α or β**) 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

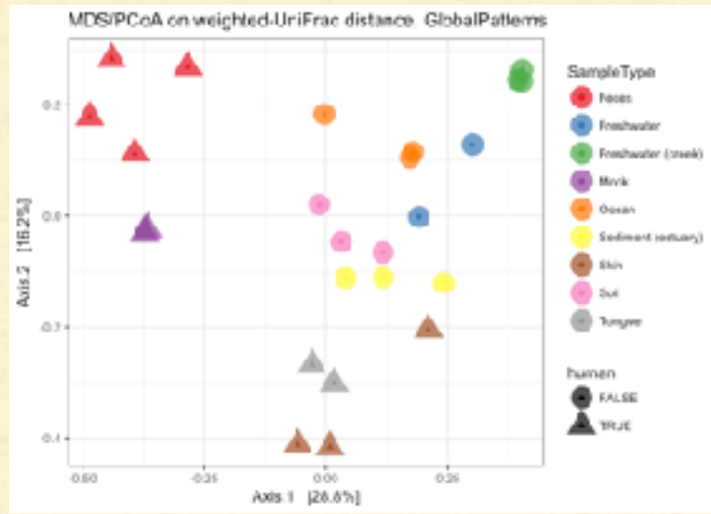
Now including…

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: 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
  - Good solution = use error bars
    - breakaway::betta(); DivNet::testDiversity
  - (Bad solution = rarefy)

VARIANCE AND HYPOTHESIS TESTS

- Why is estimating variance important?
- Hypothesis testing
- Most hypothesis tests take the form

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0,1)$$

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 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

- **Statistical estimation accounts for different sequencing depths!**

BIAS AND DIVERSITY

- Alternative approach

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

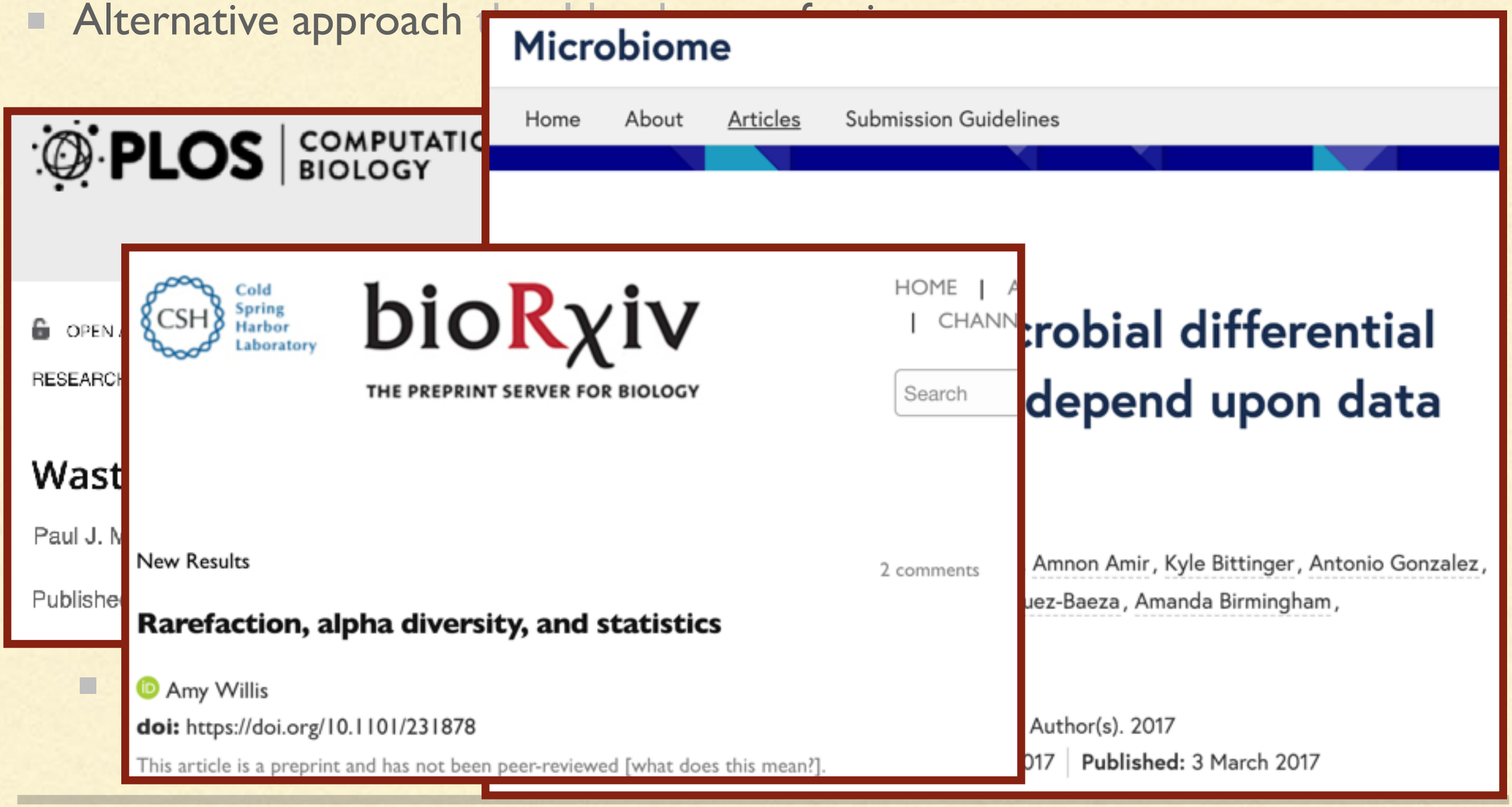
https://doi.org/10.1186/s40168-017-0237-y

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

- **Statistical estim depths!**

BIAS AND DIVERSITY

- Alternative approach



DIVERSITY

- Very useful summary of (high-dimensional) compositional data… in many settings!
- A change in diversity: a useful *first question*
- Diversity gives you limited resolution to understand interesting ecology

DIVERSITY

- Good news
  - diversity is different 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

HUGE THANKS

- My hardworking & brilliant research group: Statistical Diversity Lab:
  - Bryan Martin (@BryanDMartin_), Pauline Trinh (@paulinetrinh), David Clausen, Alex Paynter, Jake Price (@Jake_in_the_Lab)
- Collaborators whose joint work I discuss
  - **Sam Minot** (Fred Hutch), **Alon Shaiber & M Eren** (U Chicago), **Michael McLaren & Ben Callahan** (NC State)
- The **heroic organizers of #STAMPS2019**

HUGE THANKS

YOU!

- For jumping on the  
- For participating, contributing, correcting me throughout

MISCELLANEA

- Bias
- CAGs

CAGs AS BIOLOGICAL UNITS

Work lead by Sam Minot (Fred Hutch)

- Co-abundant gene (CAG) construction algorithm
- No databases
- Reproducibly associated with disease