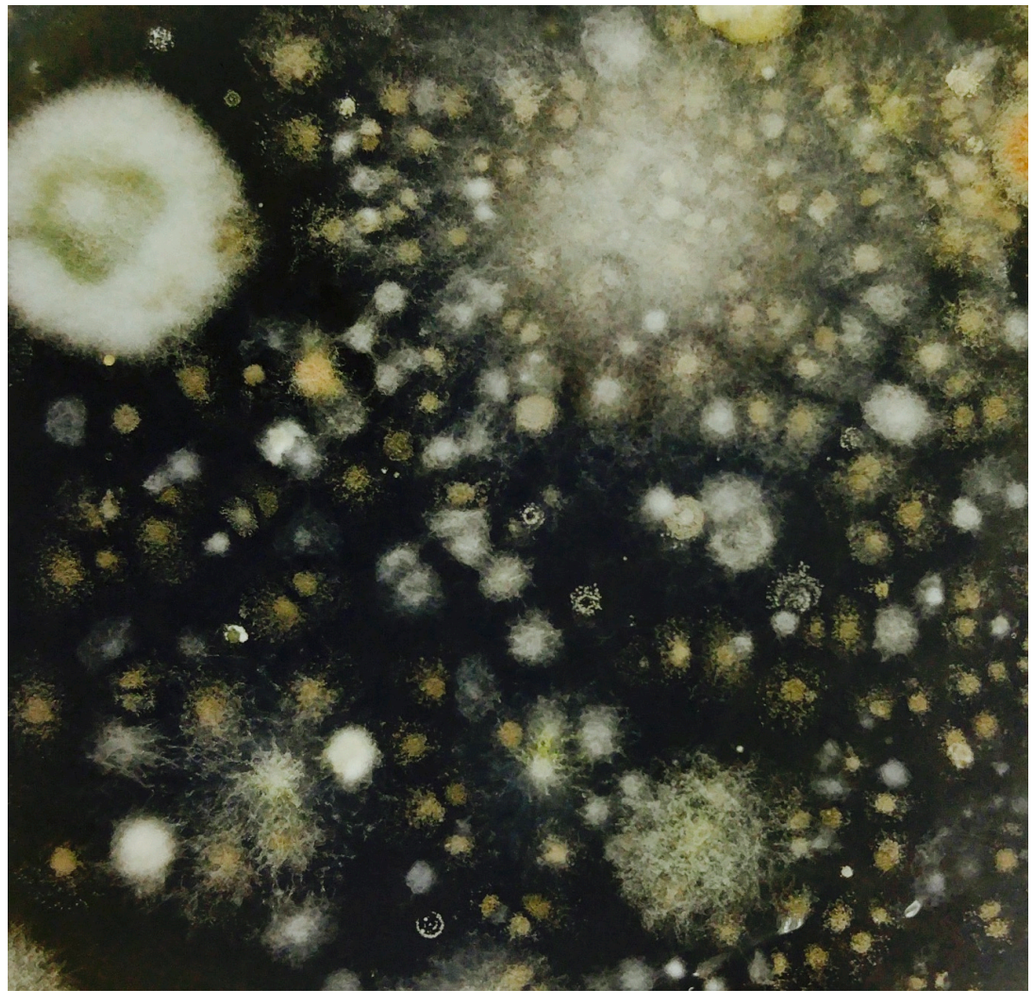
A rigorous & rational approach to

Microbial differential abundance



Amy D Willis PhD
Associate Professor
Department of Biostatistics
University of Washington
Pronouns: she/her
🦋🐦 @AmyDWwillis
📫 adwillis@uw.edu

📷 ✅

📫 ✅

🦋🐦 ✅

Slides: github.com/statdivlab/presentations

Paradigms for microbiome data analysis

- Two statistical paradigms:

1. “How do we *model the data?*”

2. “How do we *learn about biology?*”

Paradigm 1: Model the data

- “Modeling the data” conversations sound like
  - compositional? overdispersed? zero-inflated? Negative binomial? Multinomial-Dirichlet? …
- rarefy?
- best α- and β-diversity metrics?
- best transformation / “normalization”?

Paradigm 2:
Learn about biology

- “Learning about biology” conversations sound like
  - What exists in the environment?
  - What would we like to learn?
  - How does our data reflect that environment?
  - What can we learn from our data? Under what assumptions?

Today: Paradigm 2 for differential abundance

arXiv > stat > arXiv:2402.05231

Search...

Help | Adv

Statistics > Methodology

[Submitted on 7 Feb 2024]

Estimating Fold Changes from Partially Observed Outcomes with Applications in Microbial Metagenomics

David S Clausen, Amy D Willis

We consider the problem of estimating fold-changes in the expected value of a multivariate outcome that is observed subject to unknown sample-specific and category-specific perturbations. We are motivated by high-throughput sequencing studies of the abundance of microbial taxa, in which microbes are systematically over- and under-detected relative to their true abundances. Our log-linear model admits a partially identifiable estimand, and we

Software: github.com/**statdivlab/radEmu**

What exists in the environment?

- "There is some number of a given biological quantity in every environment"
- "There are 54,601 *S. epidermidis* cells on my index finger"
- "There are 0 transcripts of the gene *Core RC1 subunit PsaA* on this podium”
- "There are 874,455,469 genomes circulating in 100 mL seawater with the 16S variant CGGAGGGTGCA…”

What exists in the environment?

Y_{ij} = true number of unit j in sample i

$X_i \in \mathbb{R}^p$ covariate information
(treatment vs control, pH, temp, diet…)

😻 Y_{ij} 💸	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

What would we like to know?

Y_{ij} = true number of unit j in sample i

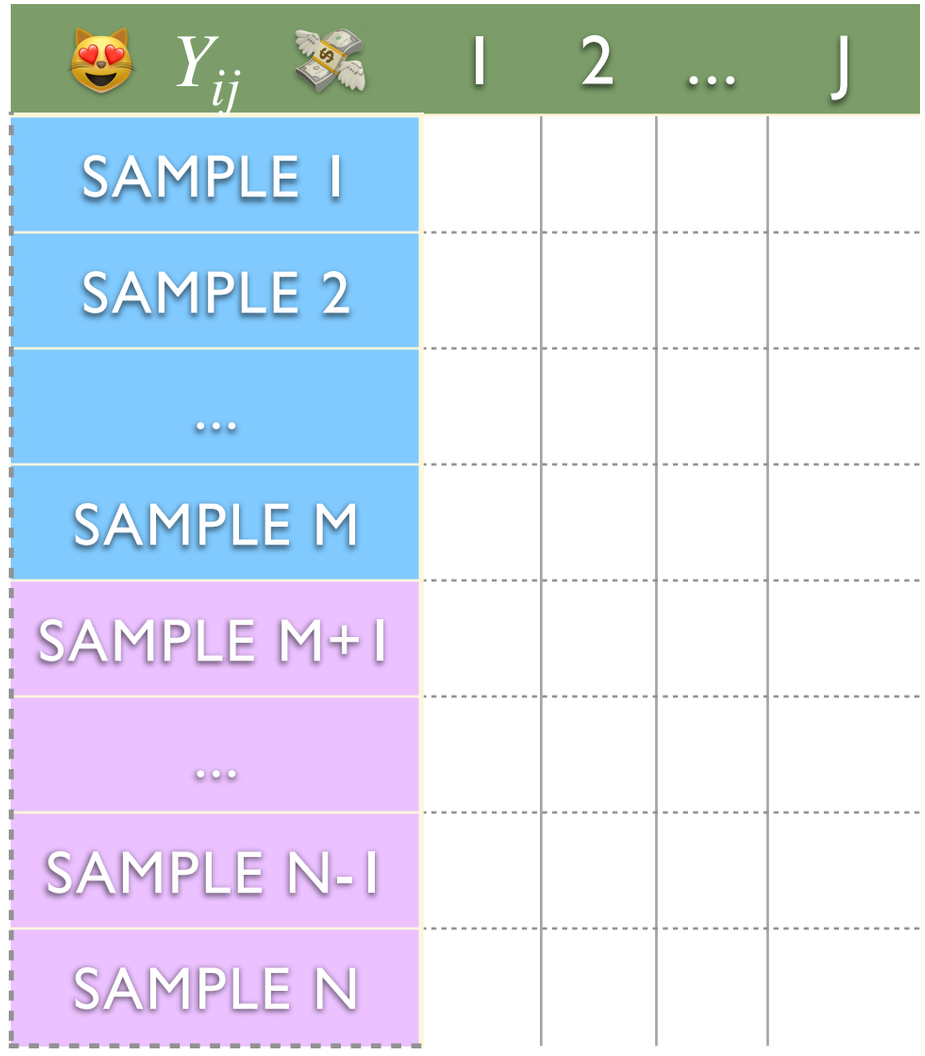
We do **not** observe $\{Y_{ij}\}$…

…but if we did, what would we do?

😻 Y_{ij} 💸	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

What would we like to know?

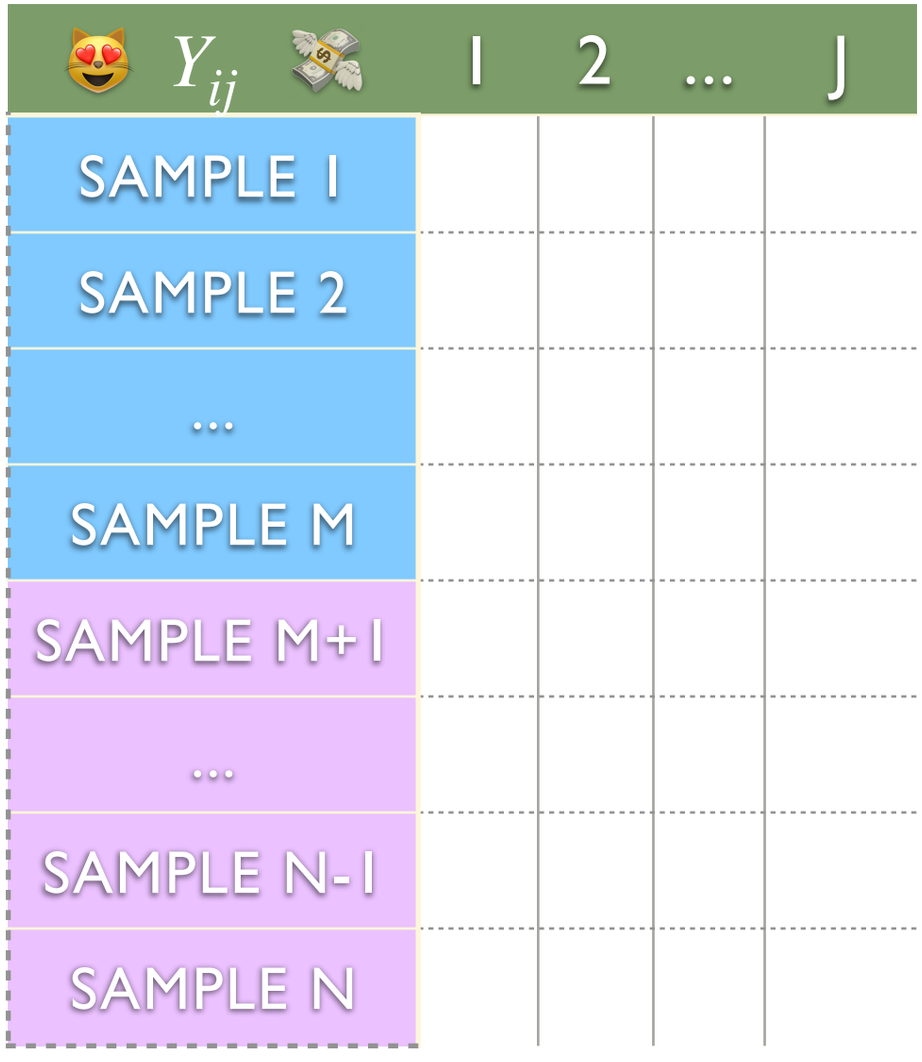
- Average of Y_{i4} across environments
- % of environments in which $Y_{i2} > 0$
- $\#\{j : Y_{ij} > 0\}$
- $-\sum_{j=1}^{J} p_{ij} \log p_{ij}$ for $p_{ij} := \frac{Y_{ij}}{\sum_j Y_{ij}}$
- …



What would we like to know?

- Differential abundance:
  Find j such that … are large

- $\mathbb{E}Y_{\{X_i=1\}j} - \mathbb{E}Y_{\{X_i=0\}j}$
- $\mathbb{E}Y_{\{X_i=1\}j} / \mathbb{E}Y_{\{X_i=0\}j}$
- $\mathbb{E}\left[\frac{Y_{\{X_i=1\}j}}{\sum_{j'} Y_{\{X_i=1\}j'}}\right] - \mathbb{E}\left[\frac{Y_{\{X_i=0\}j}}{\sum_{j'} Y_{\{X_i=0\}j'}}\right]$
- …



What data do we have?

Y_{ij} = true number of unit j in sample i

W_{ij} = number of times unit j observed in sample i from HTS

🌧️ W_{ij} 😿	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

How do we connect the Y_{ij}'s and the W_{ij}'s?

Connecting data to reality

- Traditionally, DA methods assume

$$\mathbb{E}[W_{ij}] = c_i Y_{ij}$$

- Is this reasonable?

Connecting data to reality

- Mock community: An artificially constructed community of known composition

	L.crispatus	L.iners	A.vaginae	S.agalactiae	G.vaginalis	S.amnii	P.bivia
1	0	0	0	1.00	0	0	0
2	0	0	0.5	0	0	0	0.5
3	0.33	0.33	0	0	0	0	0.33
4	0.33	0.33	0	0.33	0	0	0

Connecting data to reality

	L.crispatus	L.iners	A.vaginae	S.agalactiae	G.vaginalis	S.amnii	P.bivia
1	19	4	2	51332	1	14	1
2	0	1	1424	0	0	7	21708
3	4775	11234	0	0	0	1	3249
4	1644	5497	1	4521	0	7	0

	L.crispatus	L.iners	A.vaginae	S.agalactiae	G.vaginalis	S.amnii	P.bivia
1	0	0	0	1.00	0	0	0
2	0	0	0.5	0	0	0	0.5
3	0.33	0.33	0	0	0	0	0.33
4	0.33	0.33	0	0.33	0	0	0

Data from Brooks *et al* (2015) BMC Microbiology

Connecting data to reality

1. Despite equal mixing fractions, some taxa are observed many more times

2. Despite being purportedly absent, taxa are observed

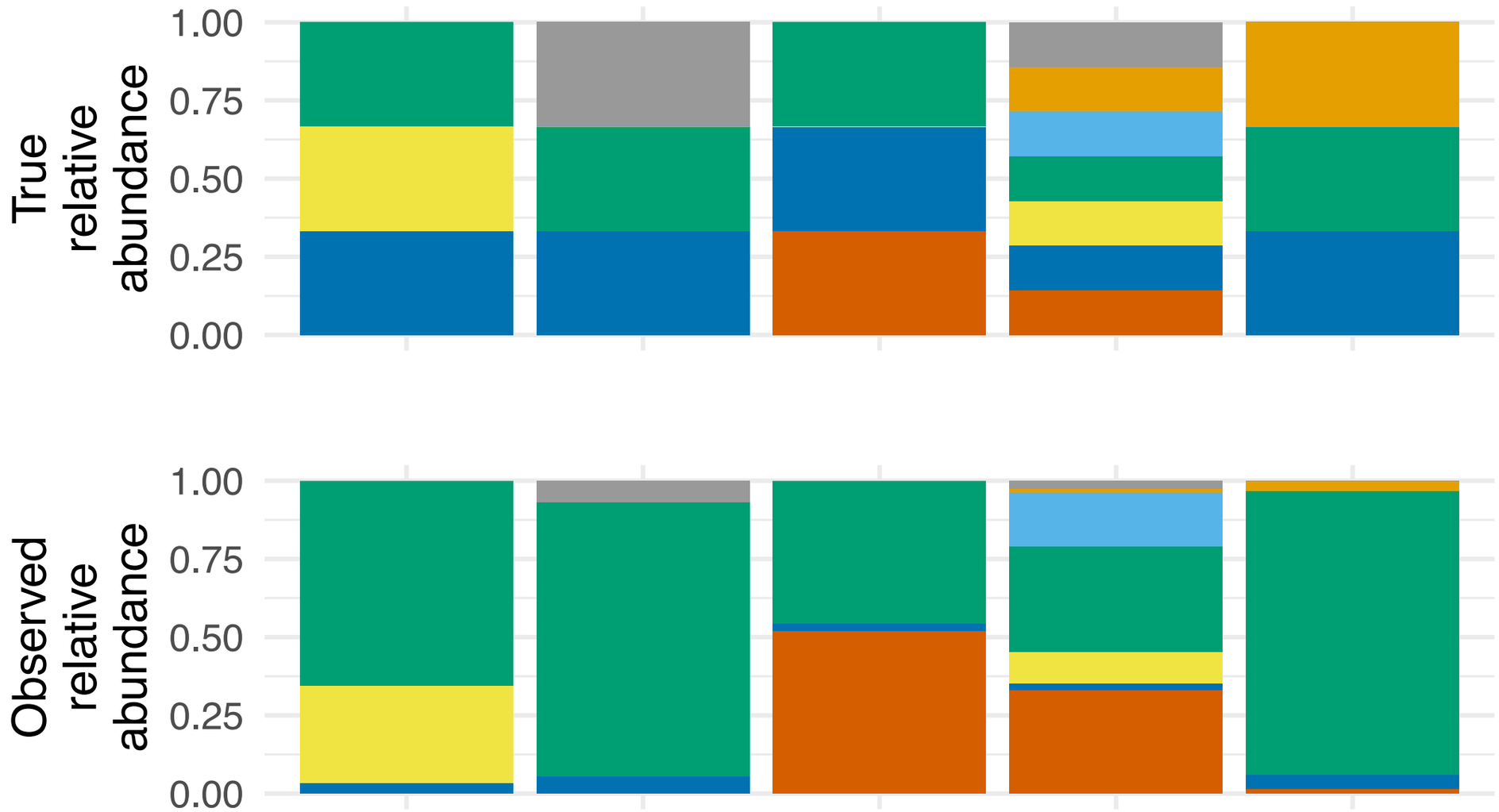
Connecting data to reality

- Despite the common assumption that

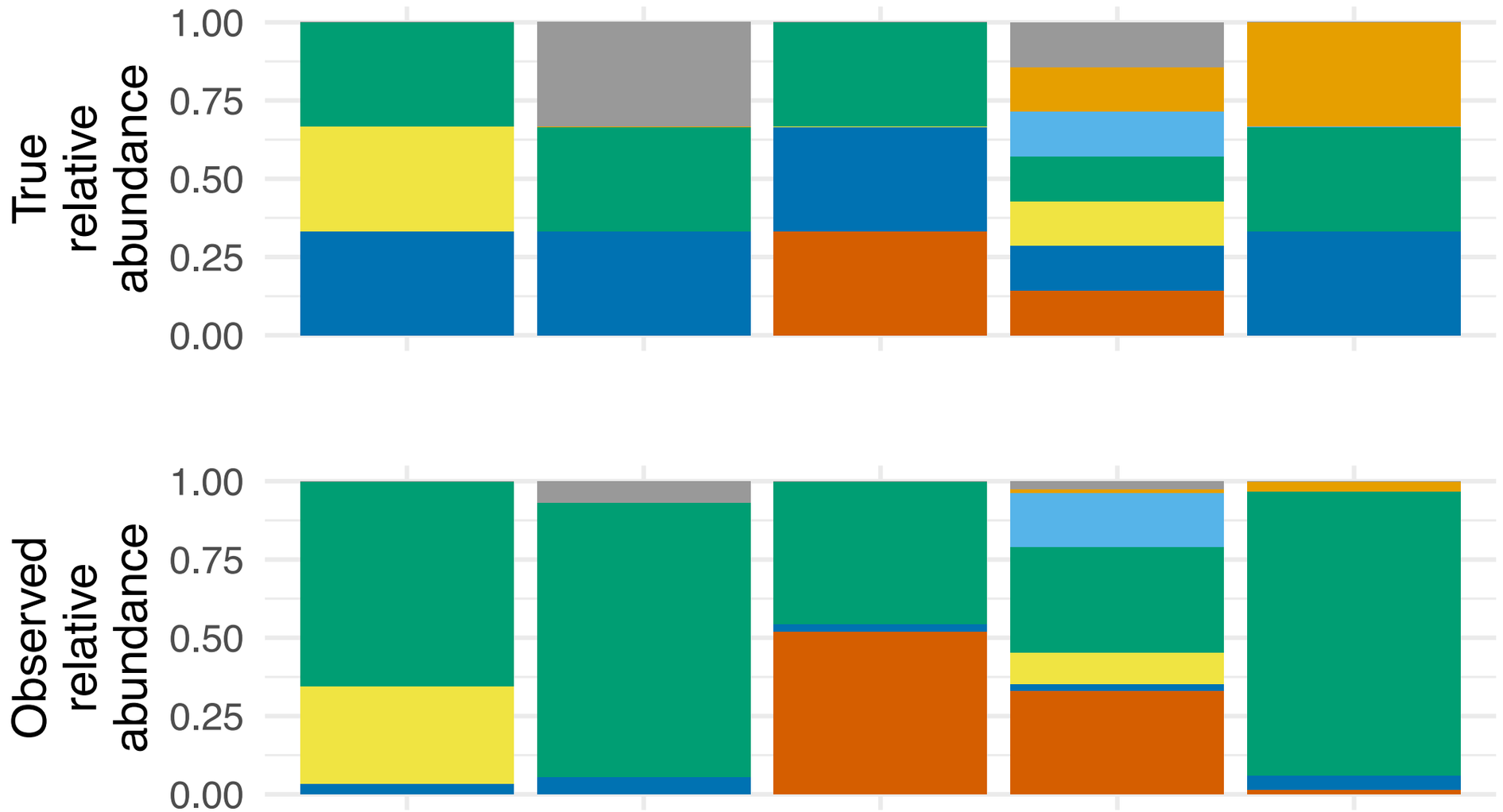
$$\mathbb{E}[W_{ij}] = c_i Y_{ij}$$

some taxa are *over-observed* for equal c_i and Y_{ij}

- What model better explains this observation?



Data from Brooks *et al* (2015) BMC Microbiology



L.iners/S.agalactiae						
	19.5	16.1	18.9	17.4	19.9	**18.3 (1.6)**

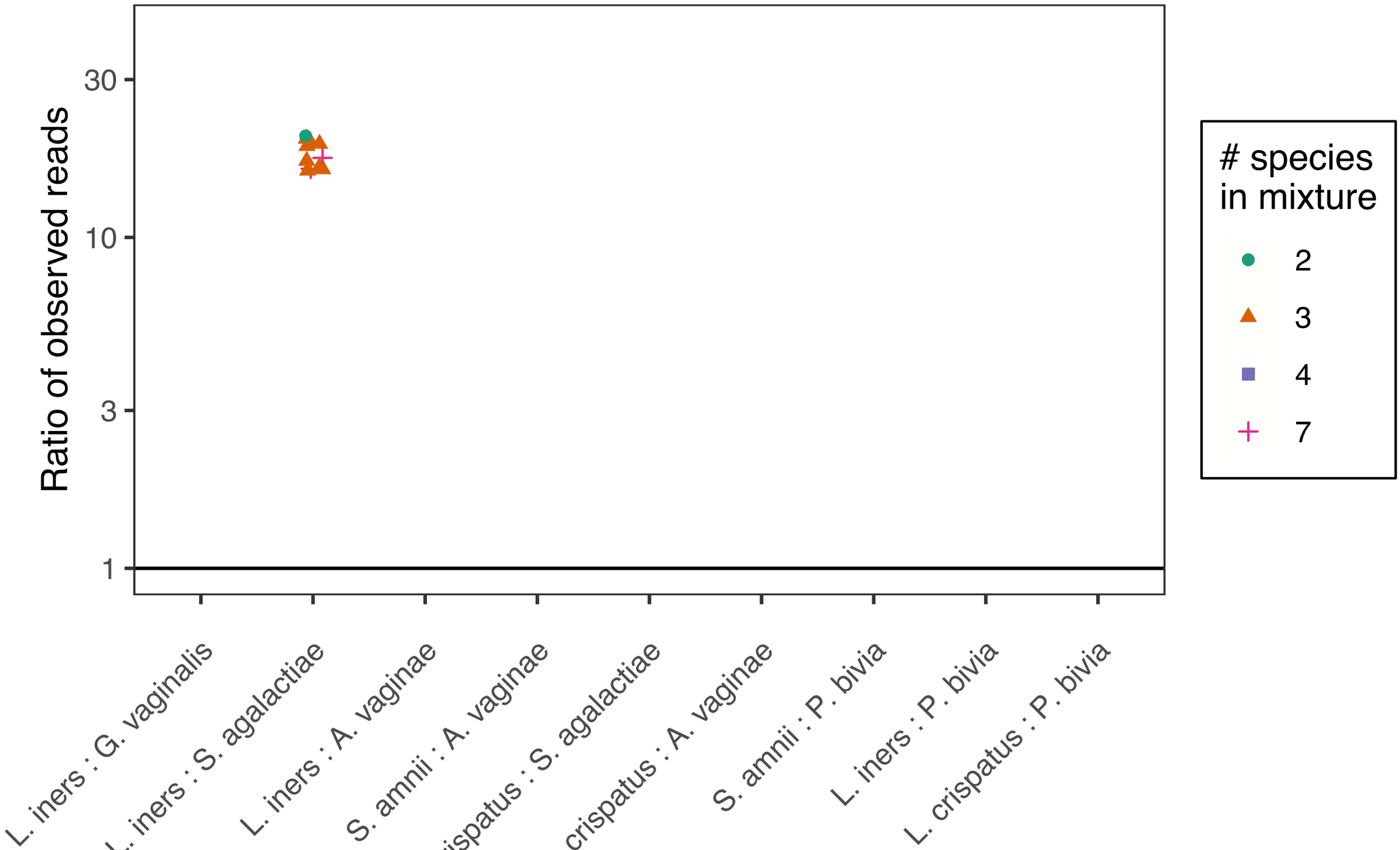


 Figure adapted from McLaren, Willis & Callahan, 2019, *eLife*

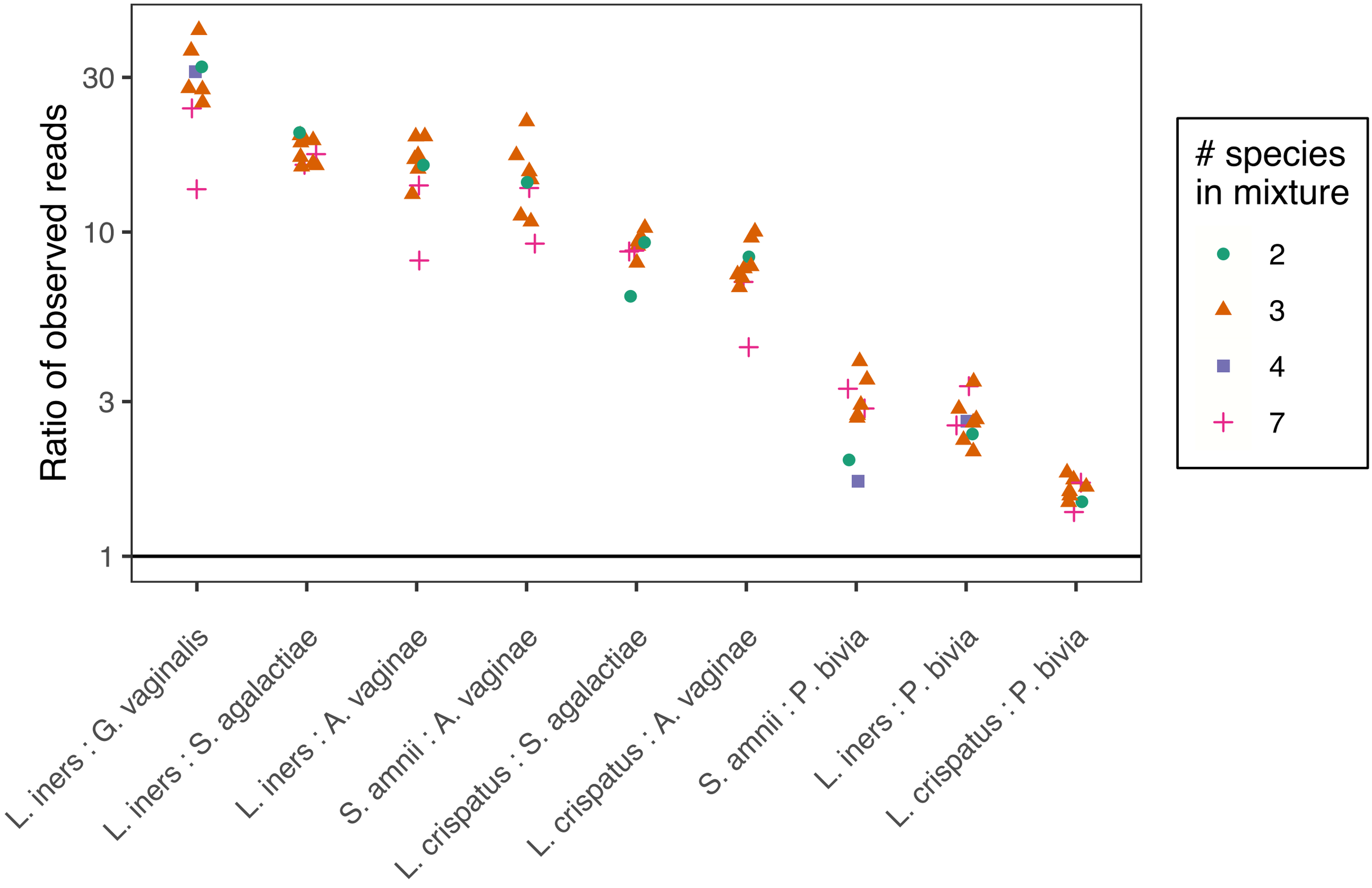


 Figure adapted from McLaren, Willis & Callahan, 2019, *eLife*

Connecting data to reality

- Evidence *against*

$$\mathbb{E}[W_{ij}] = c_i \times Y_{ij}$$

- Better support for

$$\mathbb{E}[W_{ij}] = c_i \times e_j \times Y_{ij}$$

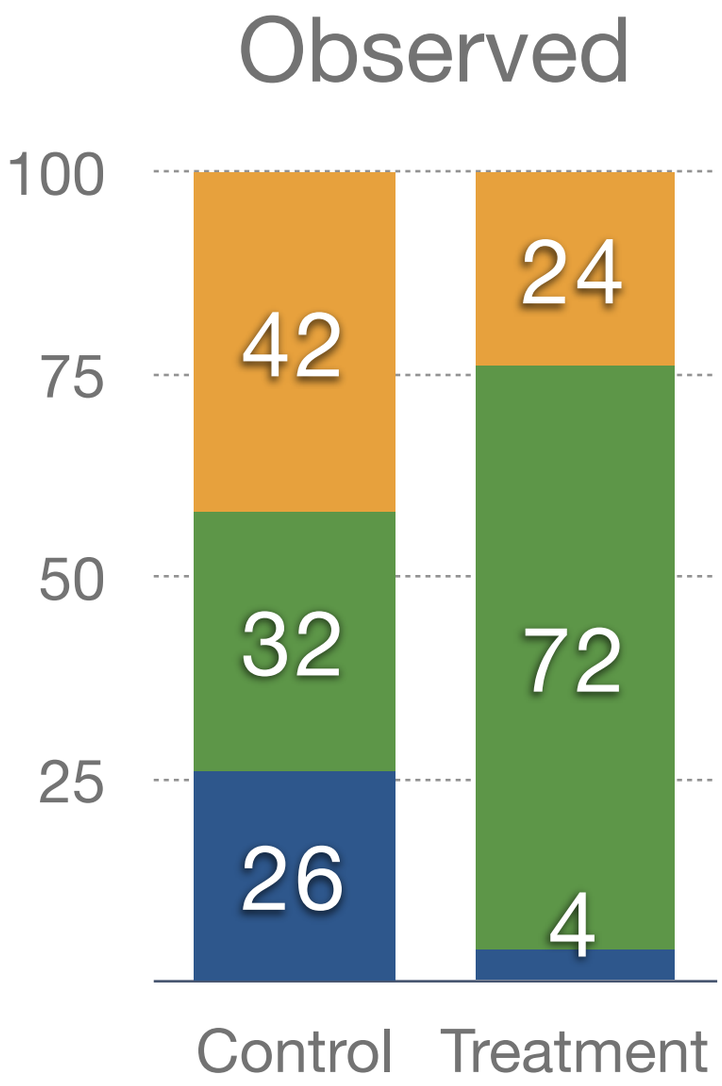
Why is this so important for data analysis?

Connecting data to reality

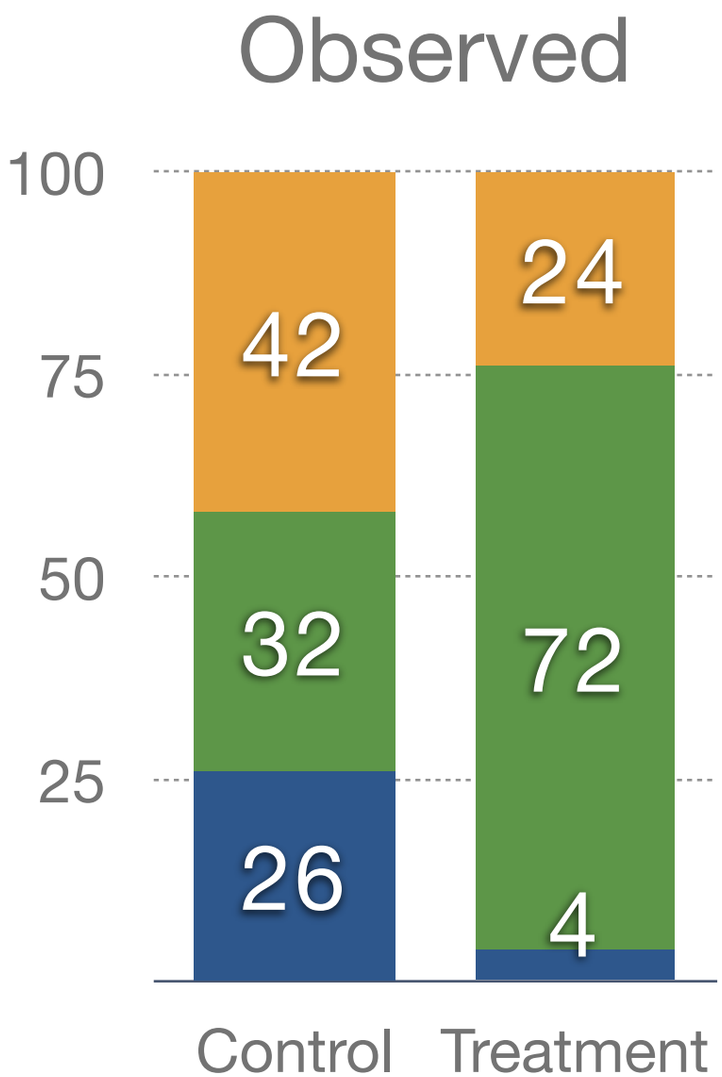
- Stated differently,

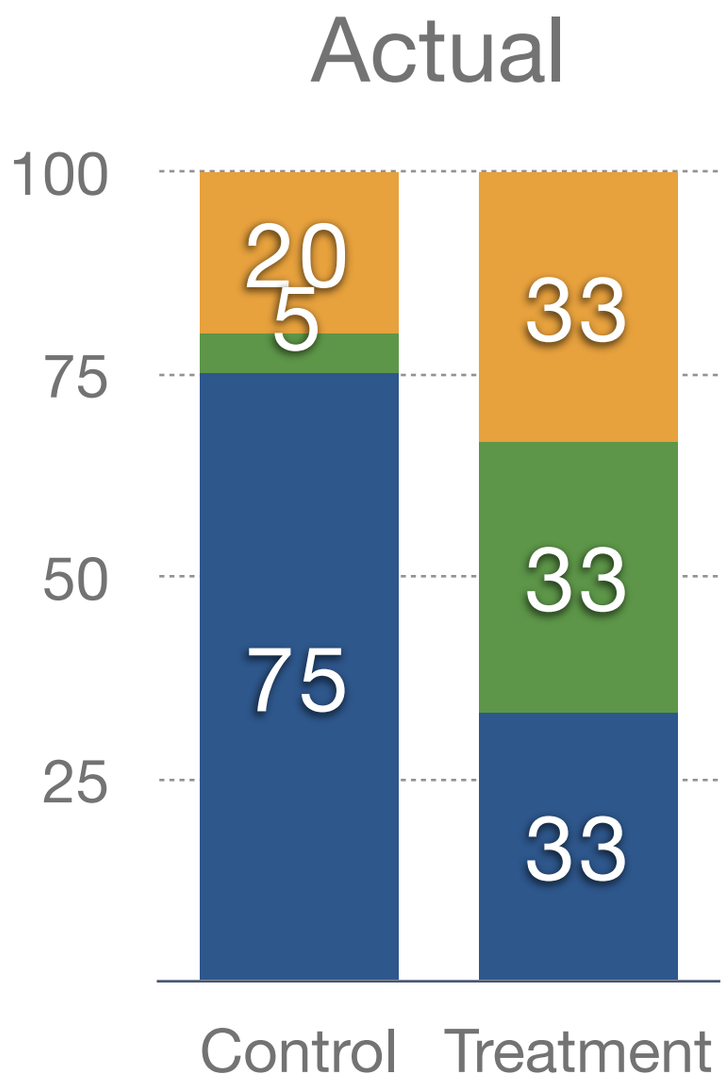
Observed relative abundance $\propto$ True relative abundance $\times$ Taxon-specific efficiencies

Expected value of $\frac{W_{ij}}{\sum_{j'} W_{ij'}} = \frac{p_{ij} e_j}{\sum_{j'} p_{ij'} e_{j'}}$



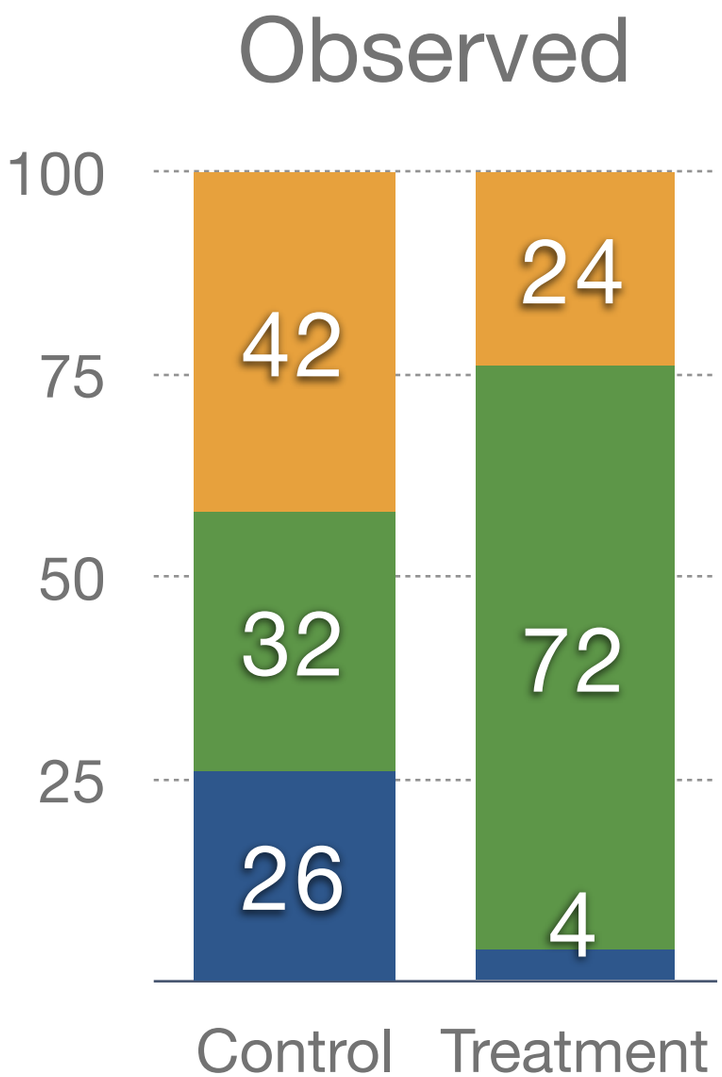
- A tempting conclusion:
  - The relative abundance of **orange** *<u>decreased</u>* in the Treatment sample (right) compared to the Control sample (left)

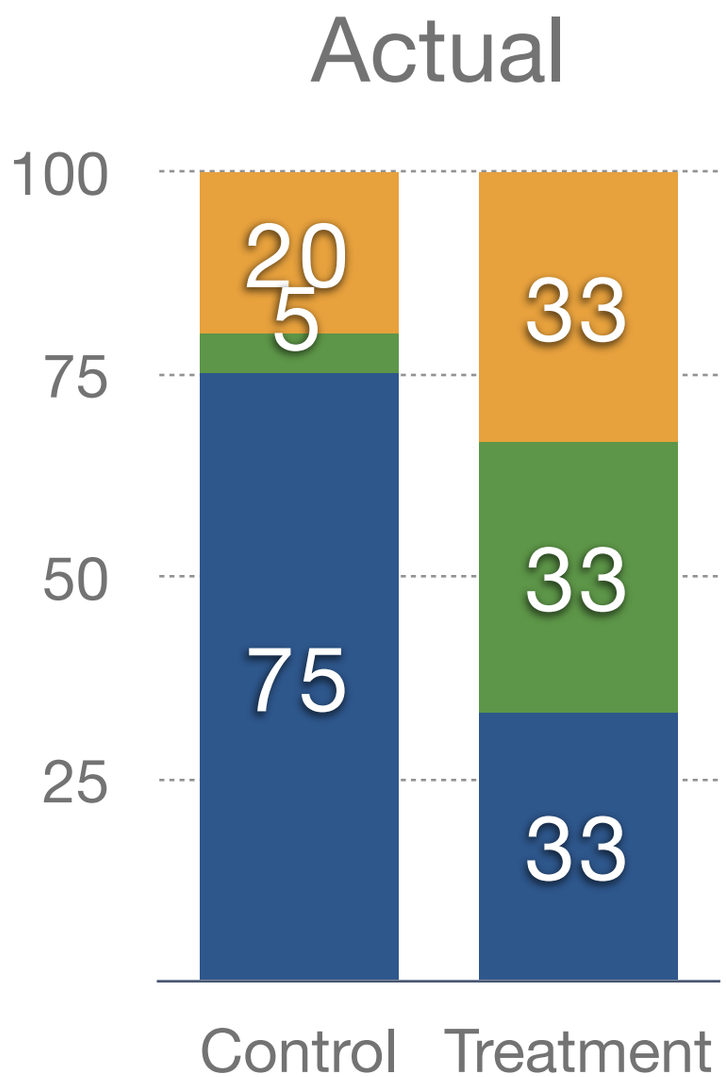


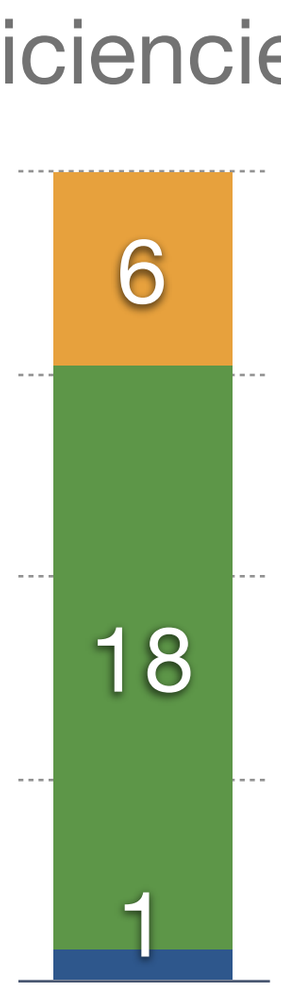


- In fact, the relative abundance of **orange** *increased* in the Treatment sample compared to the Control sample

For more details, see McLaren *et al* (2022+) "Implications of taxonomic bias..."

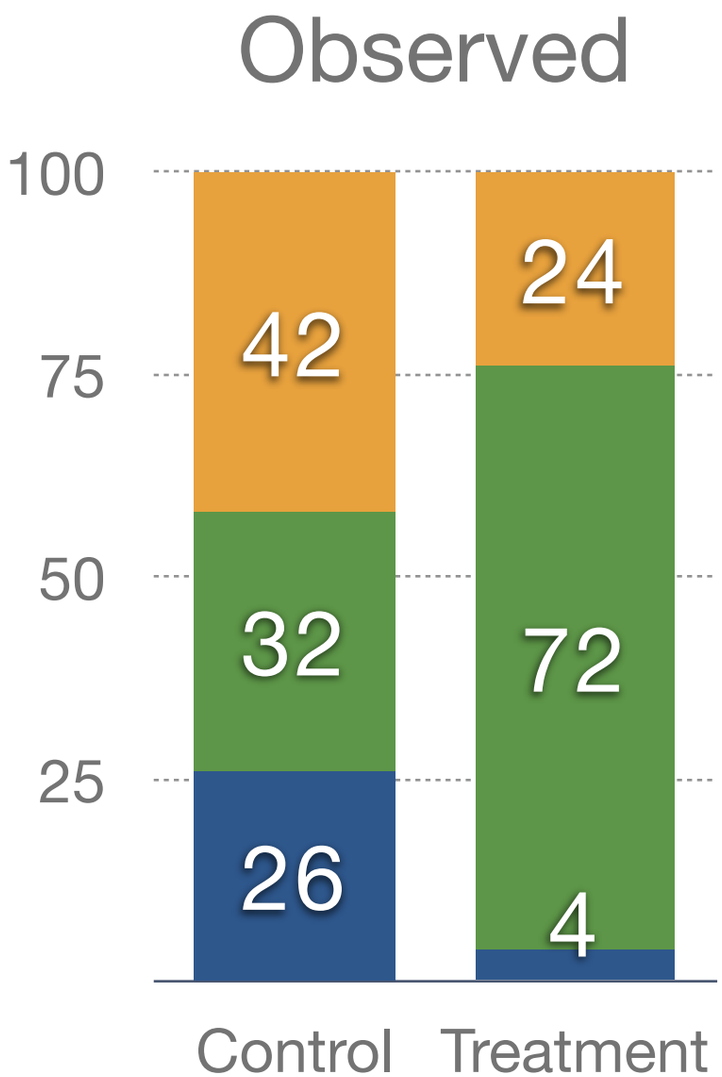


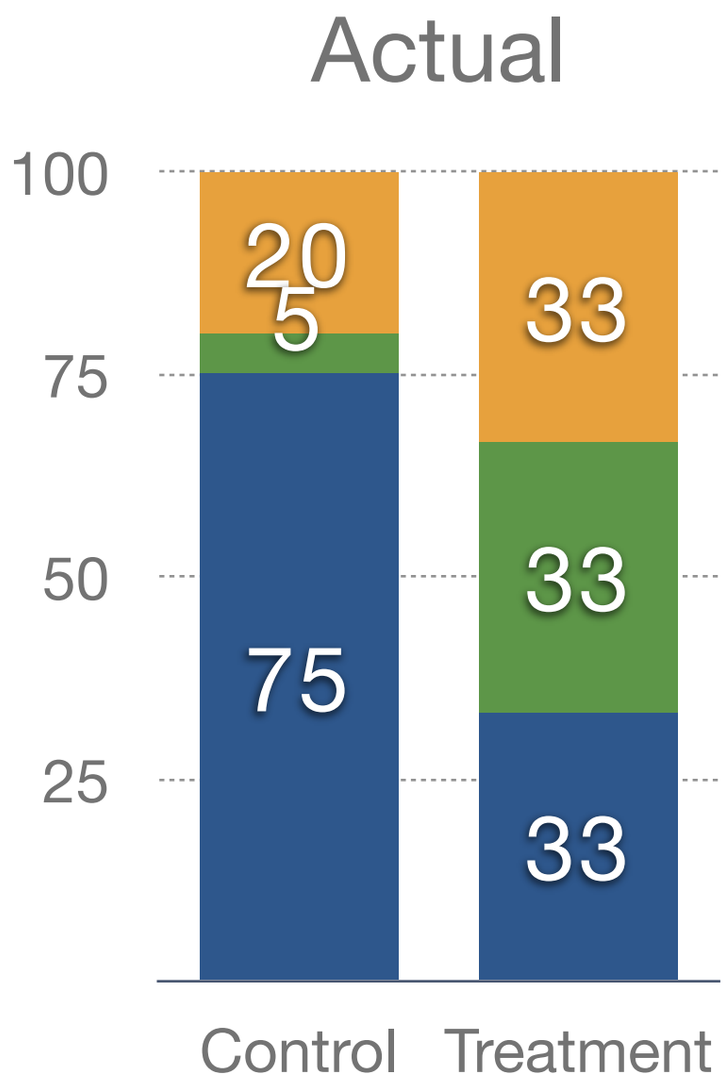


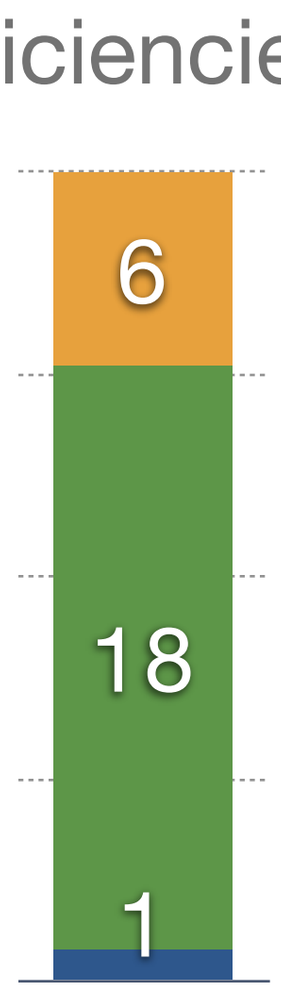


- In fact, the relative abundance of **orange** *increased* in the Treatment sample compared to the Control sample

For more details, see McLaren *et al* (2022+) "Implications of taxonomic bias..."







- **Green** is high efficiency; its abundance increased.
  Blue is low efficiency, and its abundance decreased.

- **Orange**'s abundance depends on the abundance of **the other taxa**.

For more details, see McLaren *et al* (2022+) "Implications of taxonomic bias..."

What can't we learn?

- Result: Under the model

$$\mathbb{E}W_{ij} = c_i \times e_j \times Y_{ij}$$

we *cannot* learn about:

- $\mathbb{E}Y_{\{X_i=1\}j} - \mathbb{E}Y_{\{X_i=0\}j}$
- $\mathbb{E}\left[\frac{Y_{\{X_i=1\}j}}{\sum_{j'} Y_{\{X_i=1\}j'}}\right] - \mathbb{E}\left[\frac{Y_{\{X_i=0\}j}}{\sum_{j'} Y_{\{X_i=0\}j'}}\right]$
- $\frac{\mathbb{E}Y_{\{X_i=1\}j}}{\mathbb{E}Y_{\{X_i=0\}j}}$ and $\log\left(\frac{\mathbb{E}Y_{\{X_i=1\}j}}{\mathbb{E}Y_{\{X_i=0\}j}}\right)$

What can we learn?

- Result: Under the model

$$\mathbb{E}W_{ij} = c_i \times e_j \times Y_{ij}$$

we *can* learn about:

- $$\log\left(\frac{\mathbb{E}Y_{\{X_i=1\}j}}{\mathbb{E}Y_{\{X_i=0\}j}}\right) - \log\left(\frac{\mathbb{E}Y_{\{X_i=1\}j'}}{\mathbb{E}Y_{\{X_i=0\}j'}}\right)$$

- $$\log\left(\frac{\mathbb{E}Y_{\{X_i=1\}j}}{\mathbb{E}Y_{\{X_i=0\}j}}\right) - \text{average}_{j'} \log\left(\frac{\mathbb{E}Y_{\{X_i=1\}j'}}{\mathbb{E}Y_{\{X_i=0\}j'}}\right)$$

What can we learn?

- Result: Under the model

$$\mathbb{E}W_{ij} = c_i \times e_j \times Y_{ij}$$

$$\log Y_{ij} = X_i^T \beta_j$$

we *can* learn about

- $\beta_{kj} - \beta_{kj'}$ log ratios of ratios
- $\beta_{kj} -$ average$(\beta_{k.})$ log ratios relative to average log ratios

What can we learn?

- Res

We can identify
groups (taxa, genes, etc.)
that are
changing the **most**
in abundance
from HTS

we

- p_{kj} — average($p_{k.}$) log ratios relative to average log ratios

🦤 radEmu 🦤

- We propose an estimator of β_{kj} – average($\beta_{k.}$) under the model

$$\mathbb{E}W_{ij} = c_i \times e_j \times Y_{ij}$$

$$\log Y_{ij} = X_i^T \beta_j$$

- Estimator is *consistent* under weak conditions, *efficient* under stronger conditions

🦤 radEmu 🦤

✅ Estimates a ecologically-relevant, model-agnostic, interpretable parameter

✅ Robust to differential detection

✅ Adjusts for differential sequencing depth

✅ Model-robust inference controls Type 1 error unlike DESeq2, ANCOM-BC2, t-tests…

✅ Handles zeroes without pseudocounts $\mathbb{E}(\log Y_{ij})$ vs $\log \mathbb{E} Y_{ij}$

✅ Simulation: Smallest estimation error out of methods that control T1E beats ALDEx2

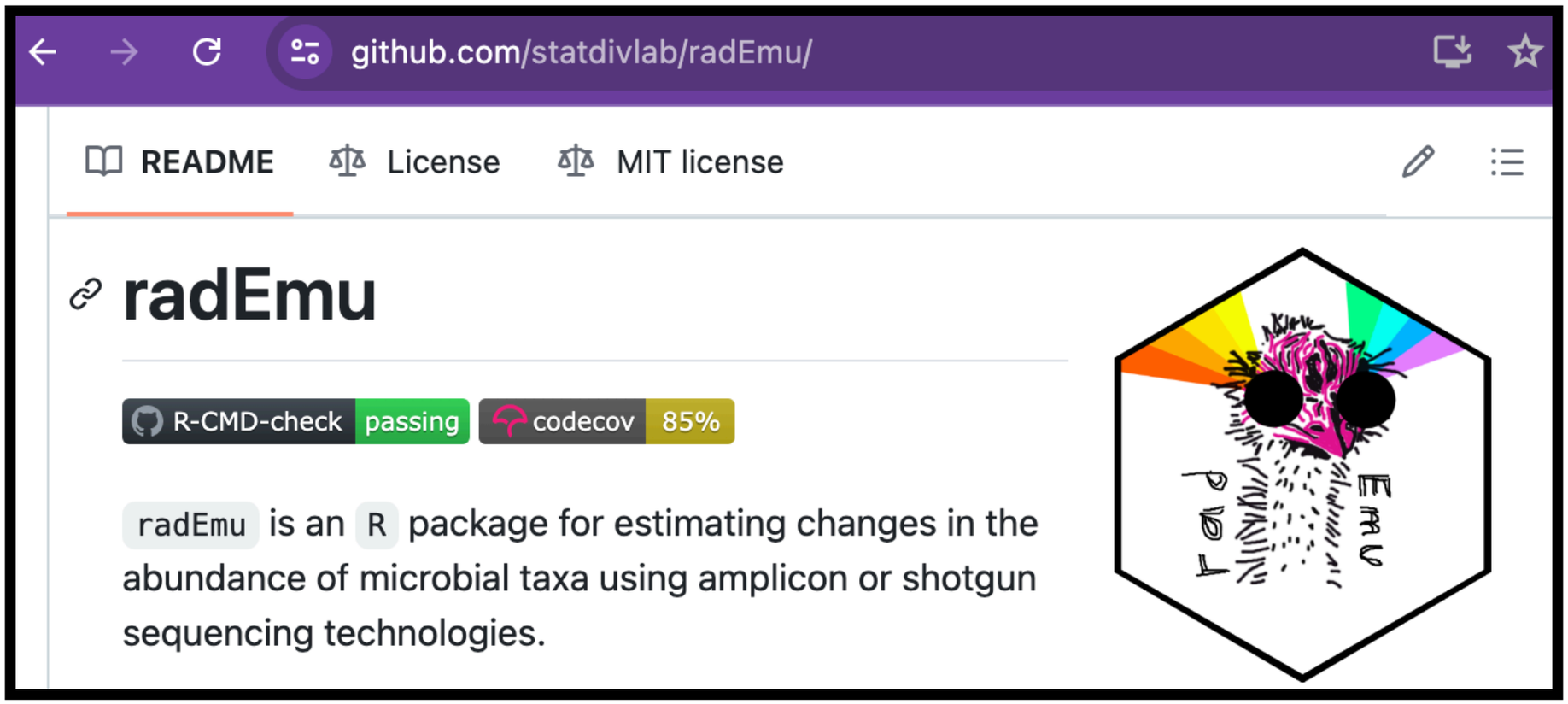
✅ Covariate adjustment; inference under independence & cluster correlation

❌ Slower than other methods

Summary

- ~~"How do we *model the data?*"~~
- "How do we *learn about biology?*"
- Want: estimate β in $\log Y_{ij} = X_i^T \beta_j$
- Have: distorted data $W_{ij} \approx c_i \times e_j \times Y_{ij}$
- Result: $\beta_{kj} -$ average$(\beta_{k.})$ is identifiable
- Method: model-robust estimation & inference with 🦤radEmu🦤

Software



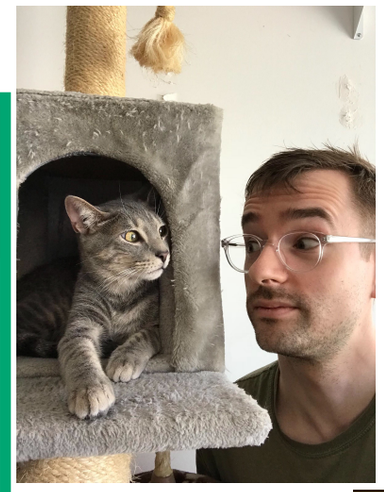

```
emuFit(formula = ~ cases + age + sex,
       data = my_metadata,
       Y = my_counts)
```

arXiv > stat > arXiv:2402.05231

Statistics > Methodology

Estimating Fold Changes from Partially Observed Outcomes with Applications in Microbial Metagenomics

David S Clausen, Amy D Willis



David Clausen

Sarah Teichman

eLIFE elifesciences.org RESEARCH ARTICLE

Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren[1], Amy D Willis[2], Benjamin J Callahan[1,3]*

Michael McLaren (MIT, SecureBio)

NIH NIGMS

Implications of taxonomic bias for microbial differential-abundance analysis

Michael R. McLaren, Jacob T. Nearing, Amy D. Willis, Karen G. Lloyd, Benjamin J. Callahan

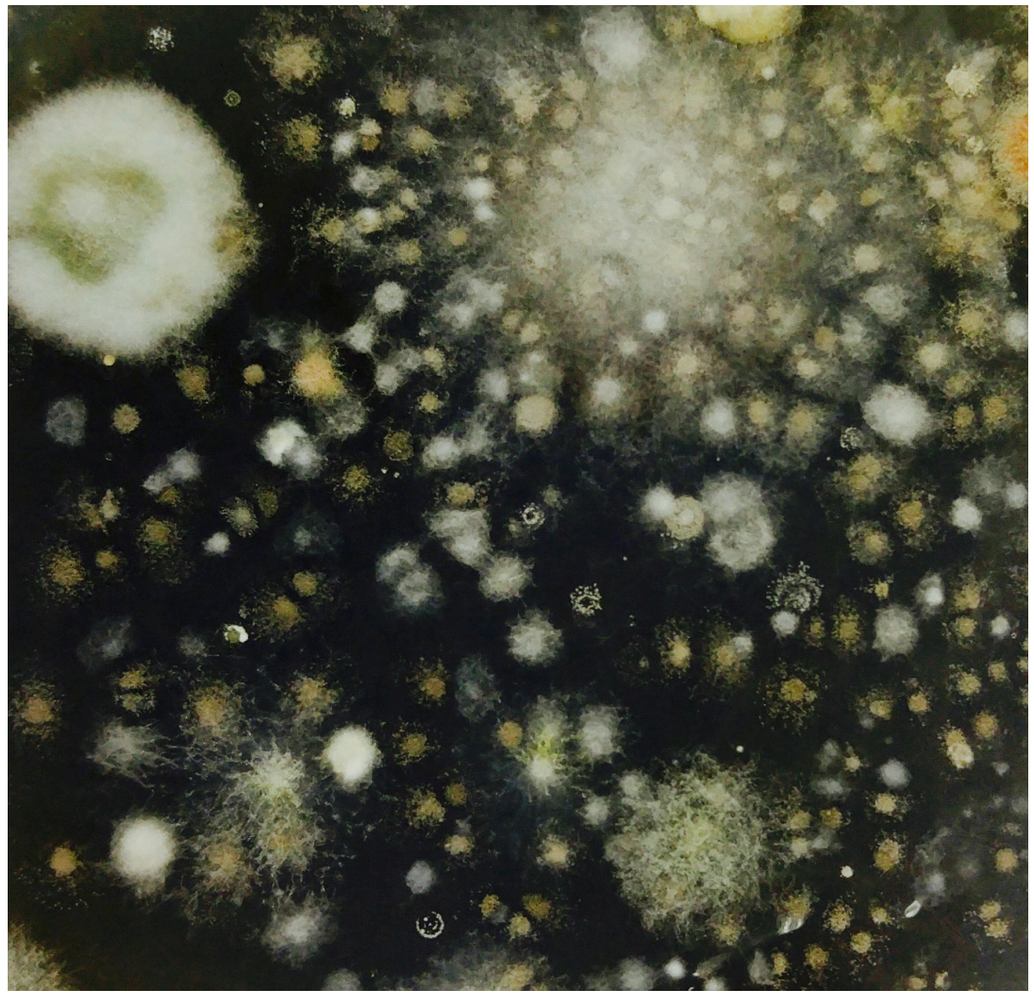
doi: https://doi.org/10.1101/2022.08.19.504330

Ben Callahan (NCSU)

R packages: radEmu from github.com/statdivlab

A rigorous & rational approach to

Microbial differential abundance



Amy D Willis PhD
Associate Professor
Department of Biostatistics
University of Washington
Pronouns: she/her
🦋🐦 @AmyDWwillis
📫 adwillis@uw.edu

📷 ✅

📫 ✅

🦋🐦 ✅

Slides: github.com/statdivlab/presentations