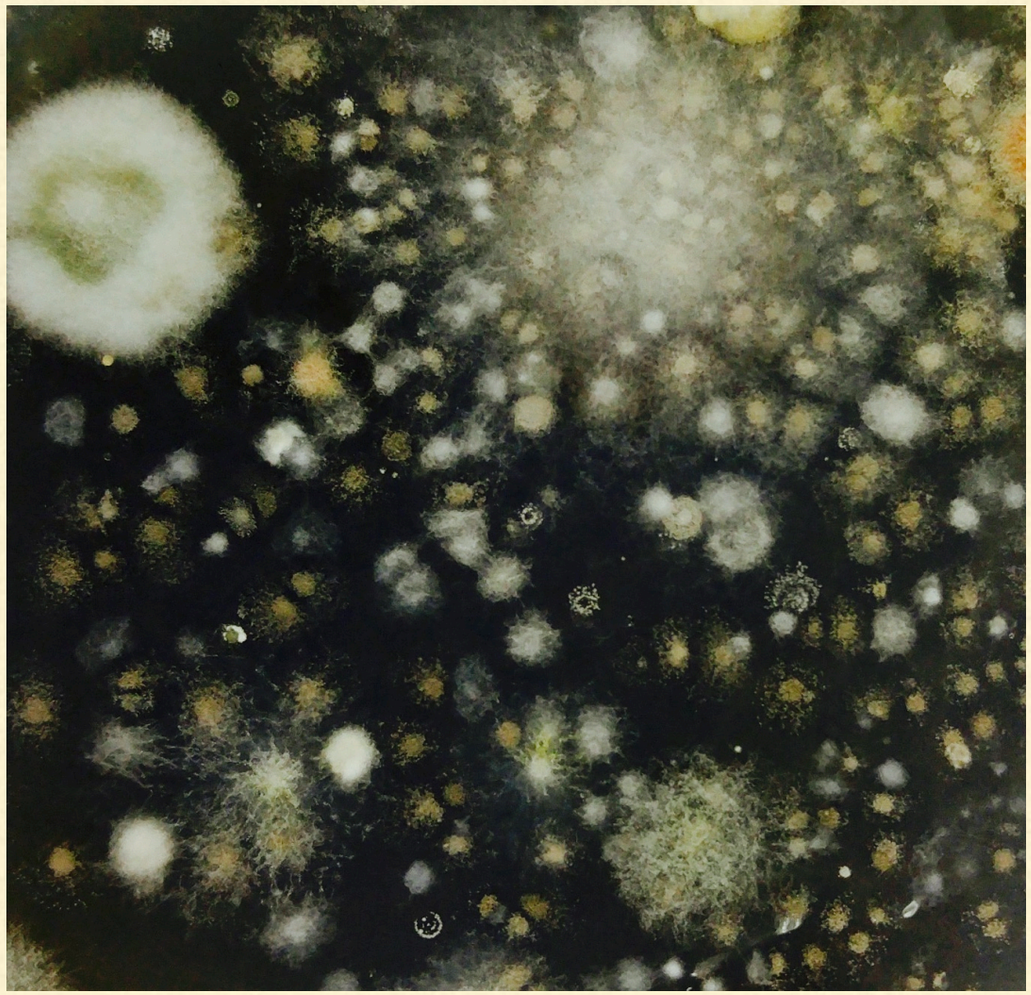
MEASUREMENT ERROR IN MICROBIOME DATA



Amy D Willis PhD

Assistant Professor

Department of Biostatistics

University of Washington, USA

@AmyDWillis

adwillis@uw.edu

📷 ✅

📬 ✅

🐦 ✅

Slides available at github.com/adw96/presentations

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

HIGH-THROUGHPUT SEQUENCING

- High-throughput sequencing (HTS) has given us high-resolution profiles of the genetic/genomic content of microbial communities
- Today's focus:

Are we *interpreting* and *analysing* HTS data correctly?

MICROBIOME DATA FROM HTS

- At the end of the sample preparation + sequencing + bioinformatics pipeline, we often have data as counts or relative abundances

Counts

	G.vaginalis	A.vaginae	L.crispatus	L.iners
s1-1	1	1	13670	0
s1-2	1650	2930	22645	0
s1-3	1514	2073	3	1
s1-4	0	2217	16037	0
s1-5	754	0	3	0
s1-6	1	500	2	9866

Relative abundances

	G.vaginalis	A.vaginae	L.crispatus	L.iners
s1-1	0.0000	0.0000	0.6016	0.0000
s1-2	0.0606	0.1076	0.8316	0.0000
s1-3	0.0782	0.1071	0.0002	0.0001
s1-4	0.0000	0.1117	0.8081	0.0000
s1-5	0.0489	0.0000	0.0002	0.0000
s1-6	0.0001	0.0340	0.0001	0.6718

CURRENT PARADIGM

- *Widely-held belief:*
  - Within a study, we can compare the relative abundance of taxa across samples
    - Directly: DeSeq2, LefSE, corncob…
    - Indirectly: Diversity (Shannon, Simpson, UniFrac), networks…

Michael McLaren (NCSU)

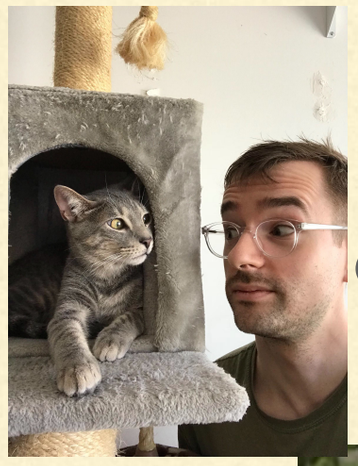
eLIFE
elifesciences.org

RESEARCH ARTICLE

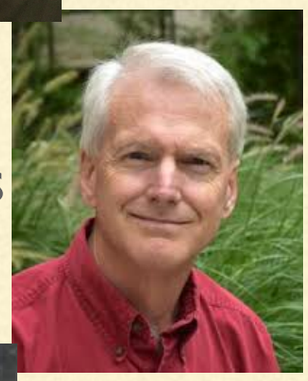
Consistent and correctable bias in metagenomic sequencing experiments

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

Ben Callahan (NCSU)



David Clausen (UW)



Jim Hughes (UW)

Brian Williamson (UW)

CSH Cold Spring Harbor Laboratory

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME | A

Search

New Results

Comment on this paper

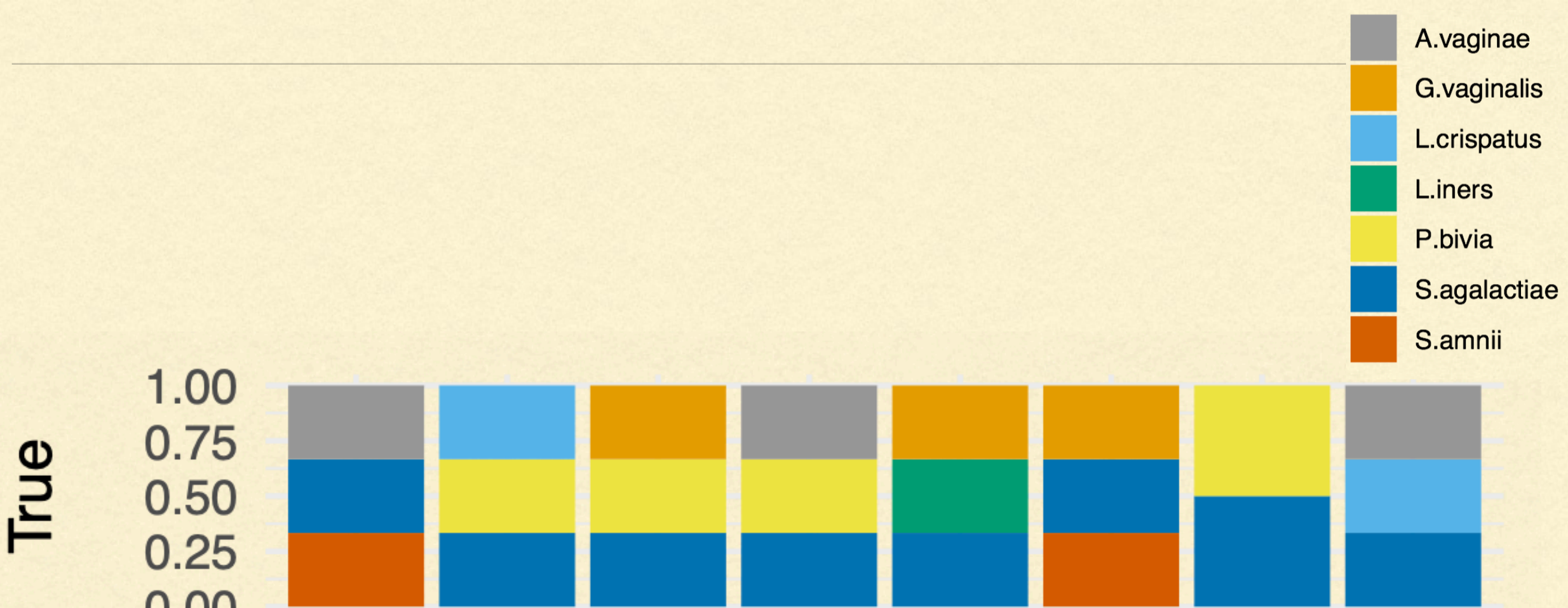
A multi-view model for relative and absolute microbial abundances

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

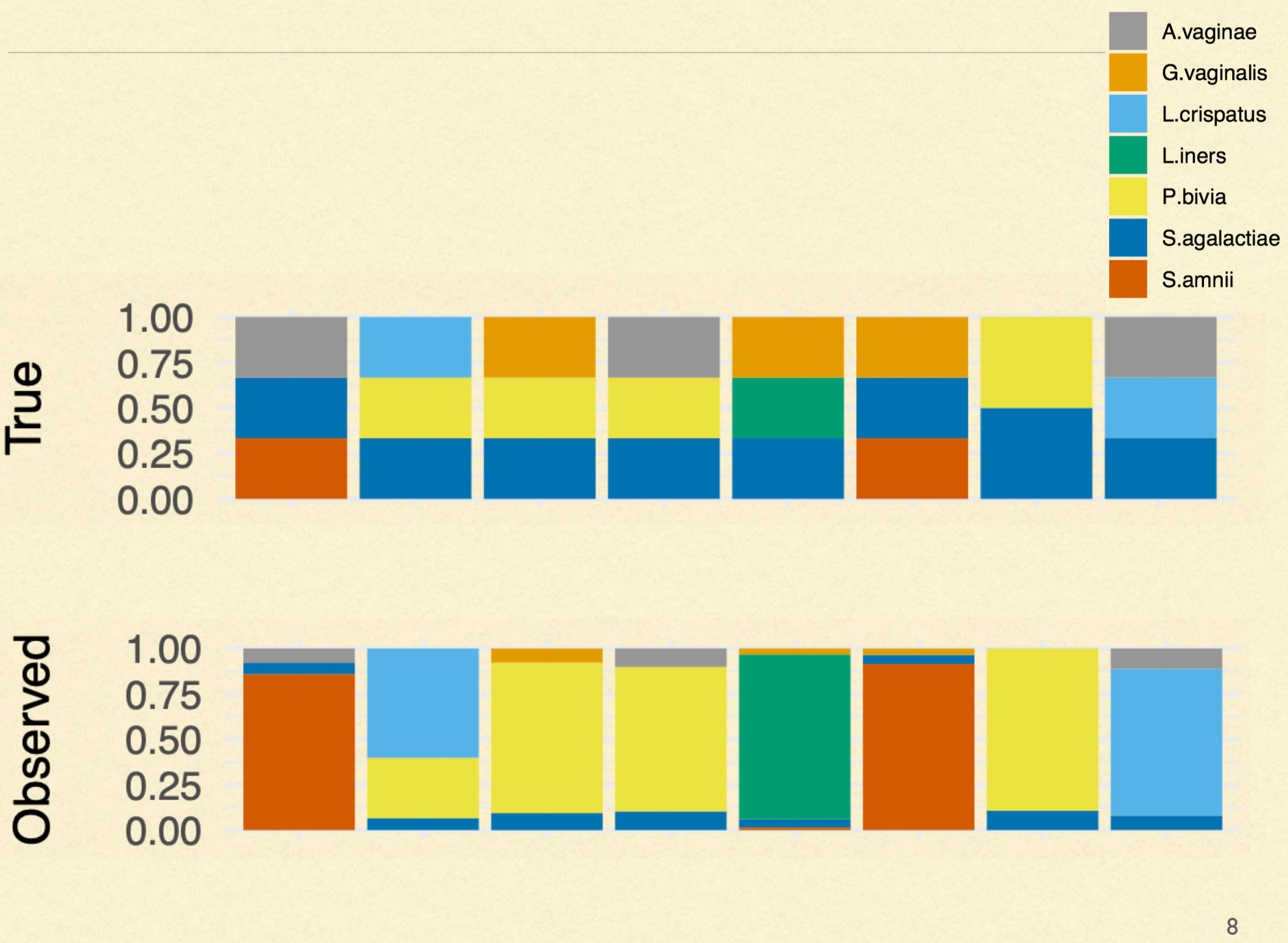
doi: https://doi.org/10.1101/761486

EVALUATING RELATIVE ABUNDANCE DATA

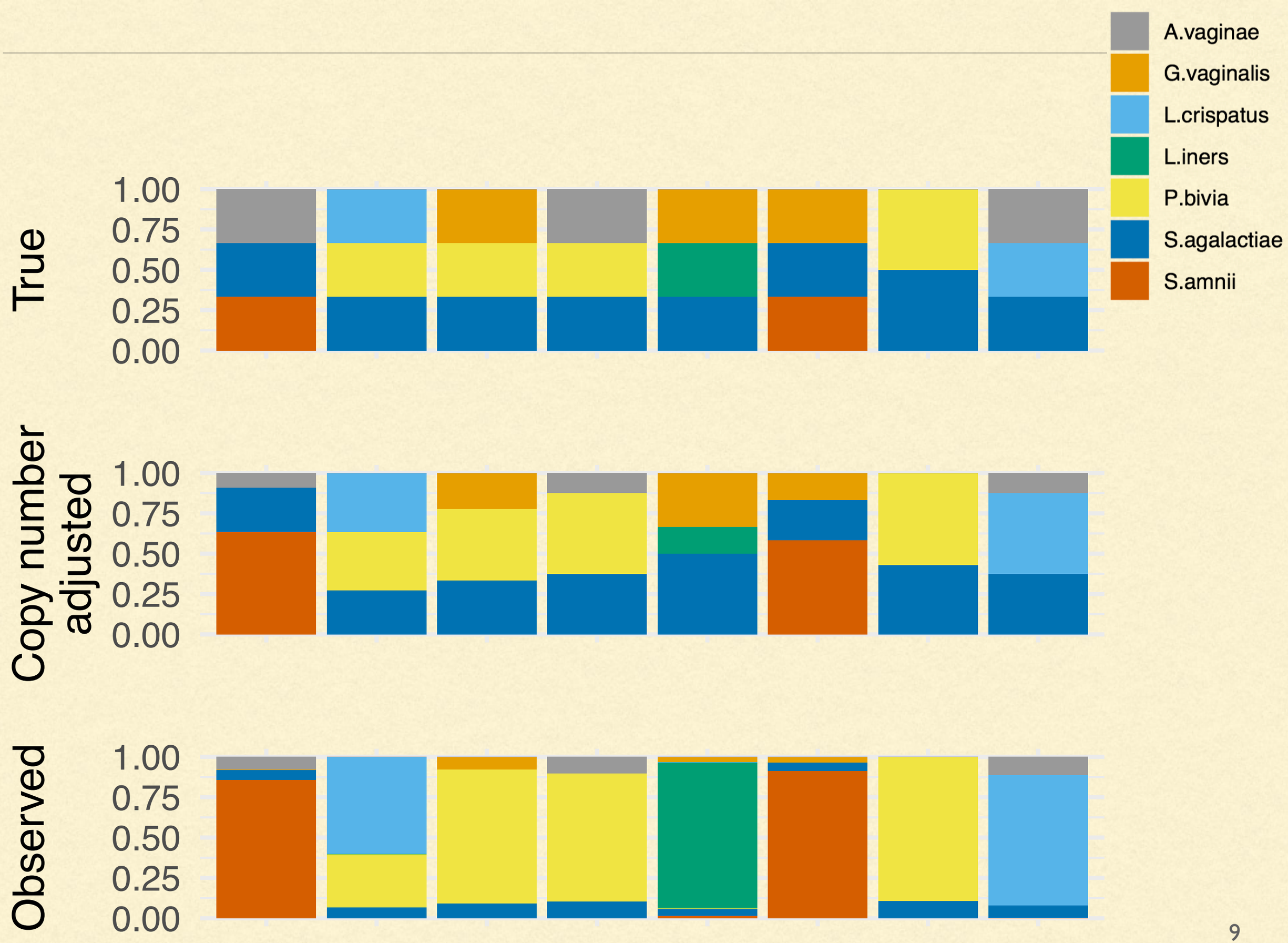
- Mock communities
- Artificially constructed communities of known composition
- Commonly used to benchmark sequencing and bioinformatics pipelines



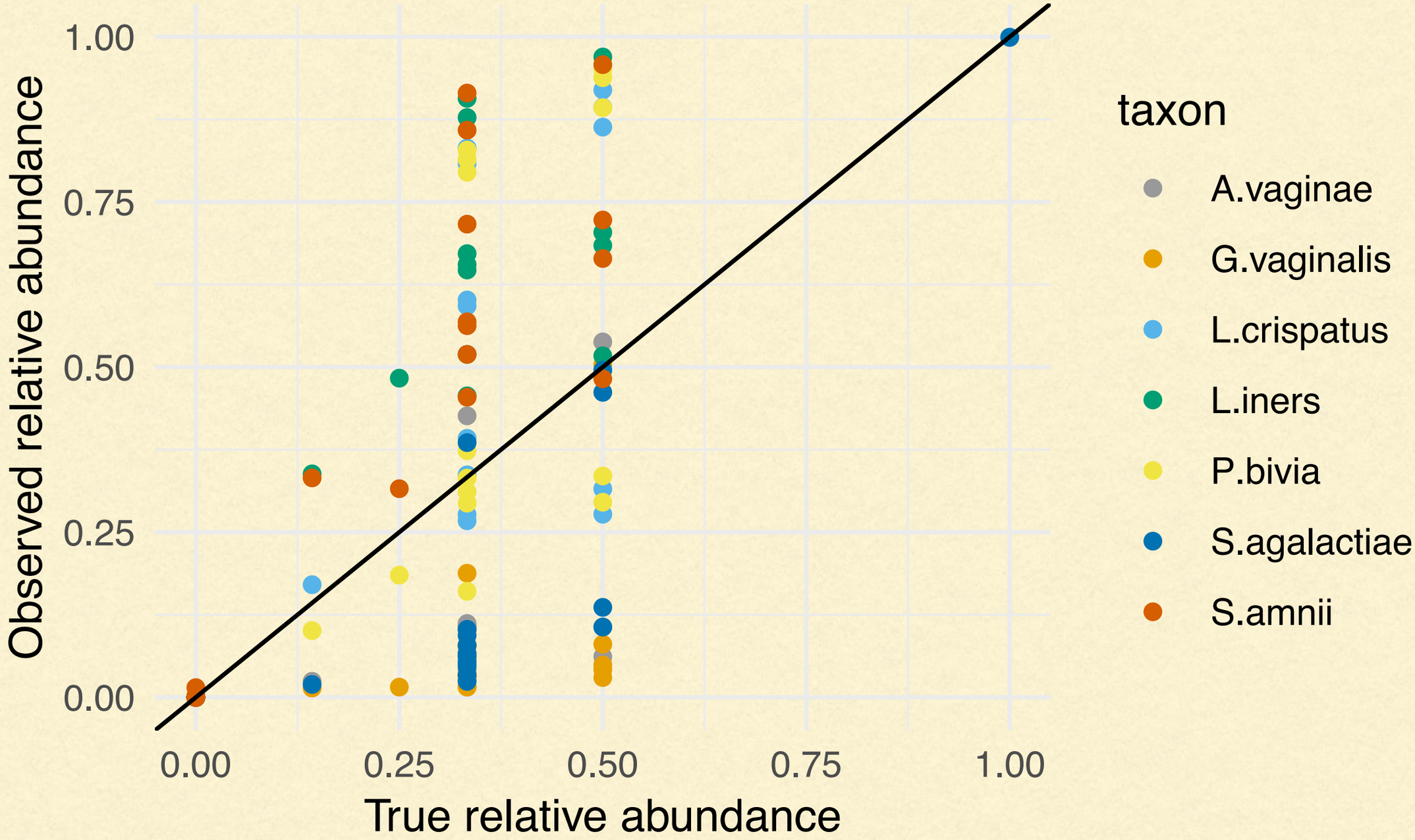
Data from Brooks *et al* (2015) BMC Microbiology



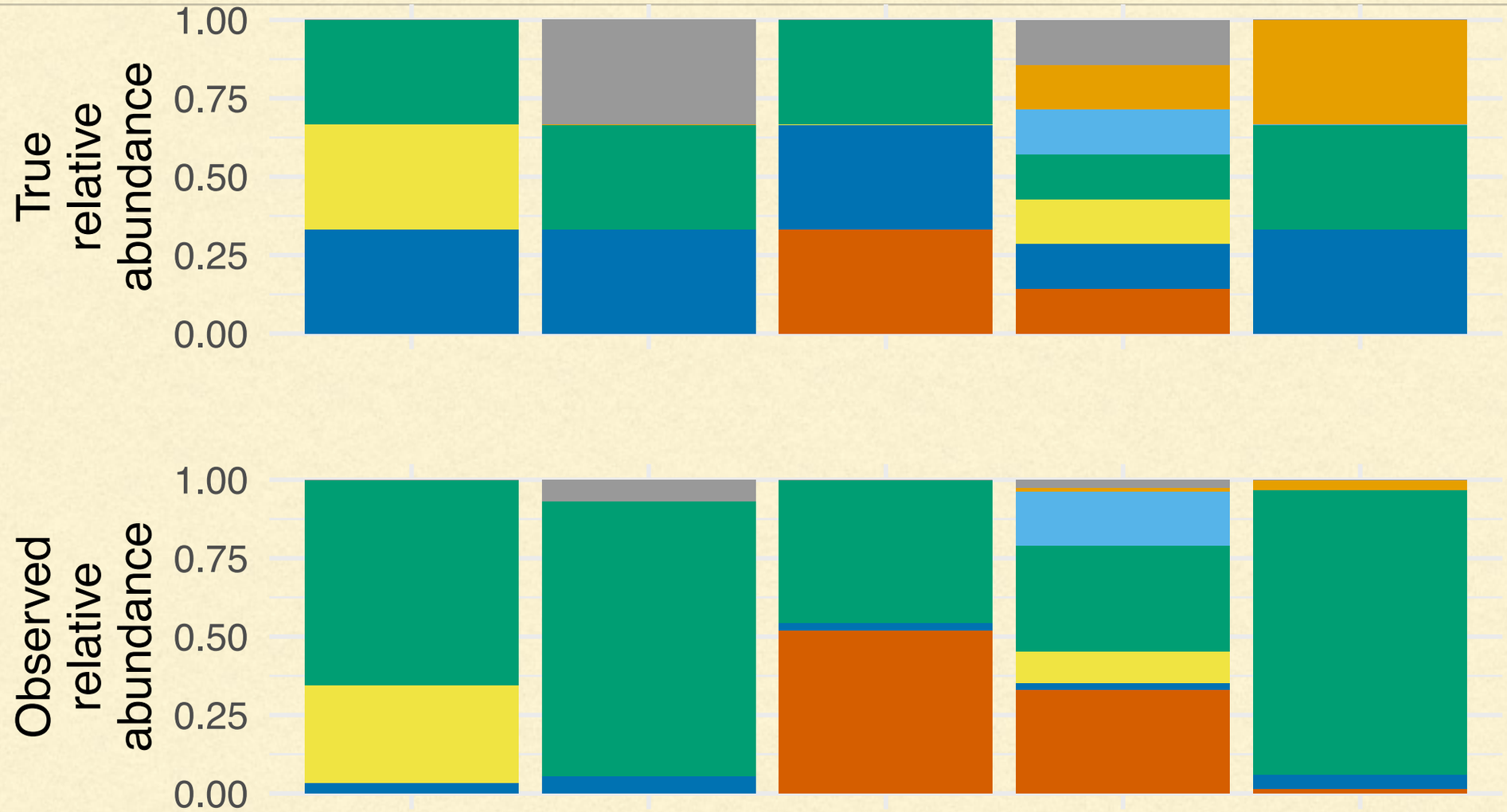
Data from Brooks *et al* (2015) BMC Microbiology



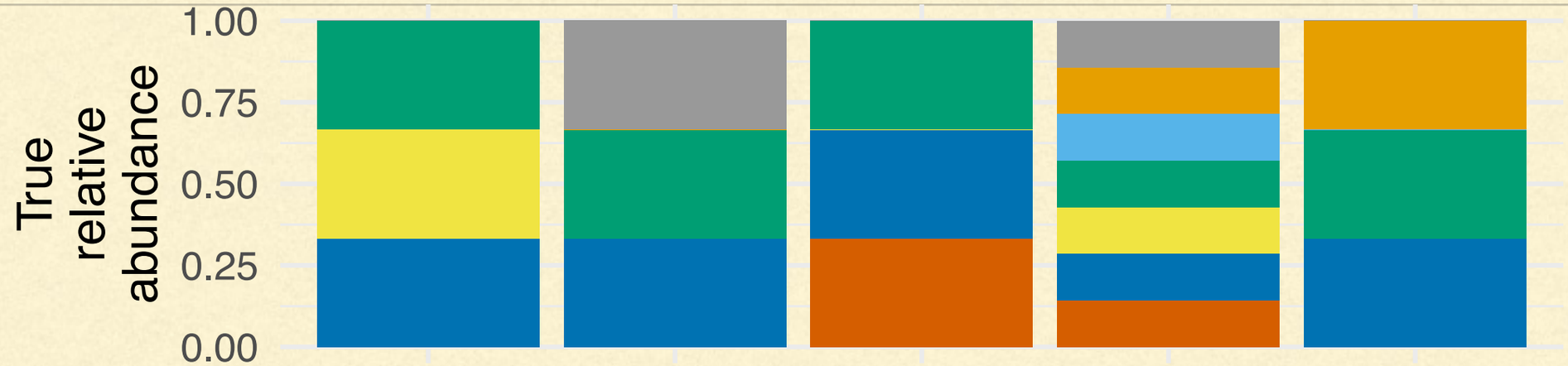
Data from Brooks *et al* (2015) BMC Microbiology

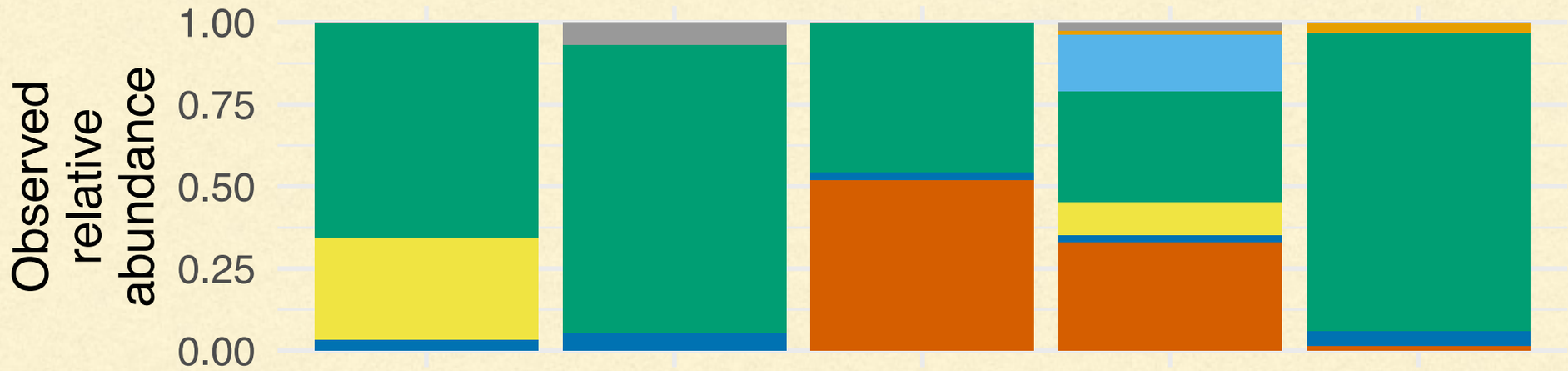

1.00
0.75
0.50
0.25
0.00
Observed relative abundance
0.00
0.25
0.50
0.75
1.00
True relative abundance
taxon
A.vaginae
G.vaginalis
L.crispatus
L.iners
P.bivia
S.agalactiae
S.amnii

WHAT PATTERNS CAN WE FIND?



WHAT PATTERNS CAN WE FIND?





L.iners/S.agalactiae	19.5	16.1	18.9	17.4	19.9

mean = 18.3, std dev = 1.6

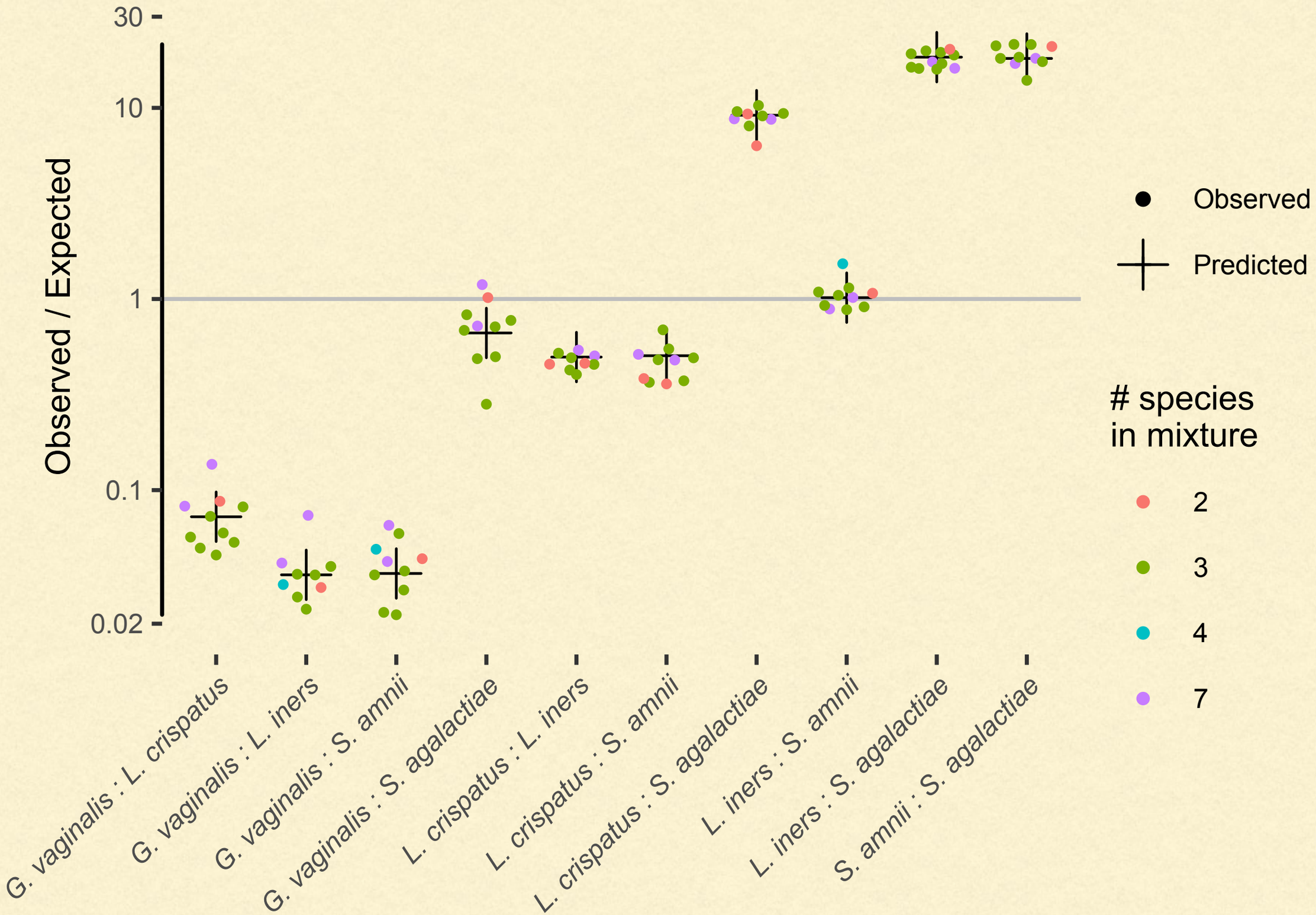


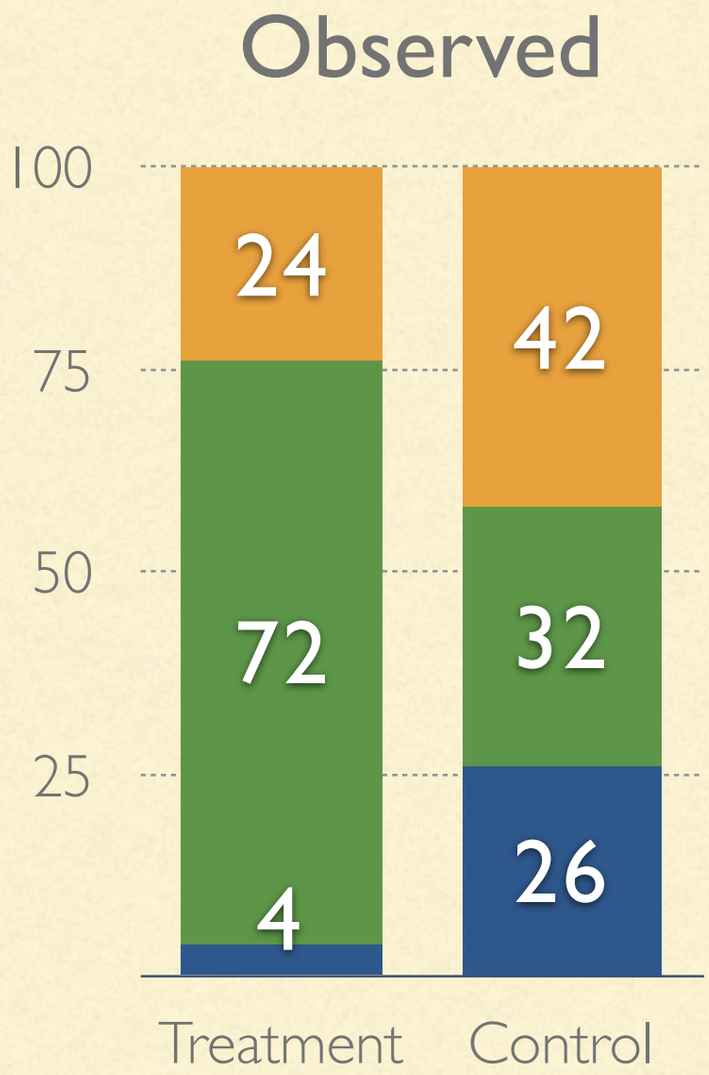
Figure from McLaren et al., 2019, *eLife*

MEASUREMENT ERROR MODEL

- This data suggests the following model:

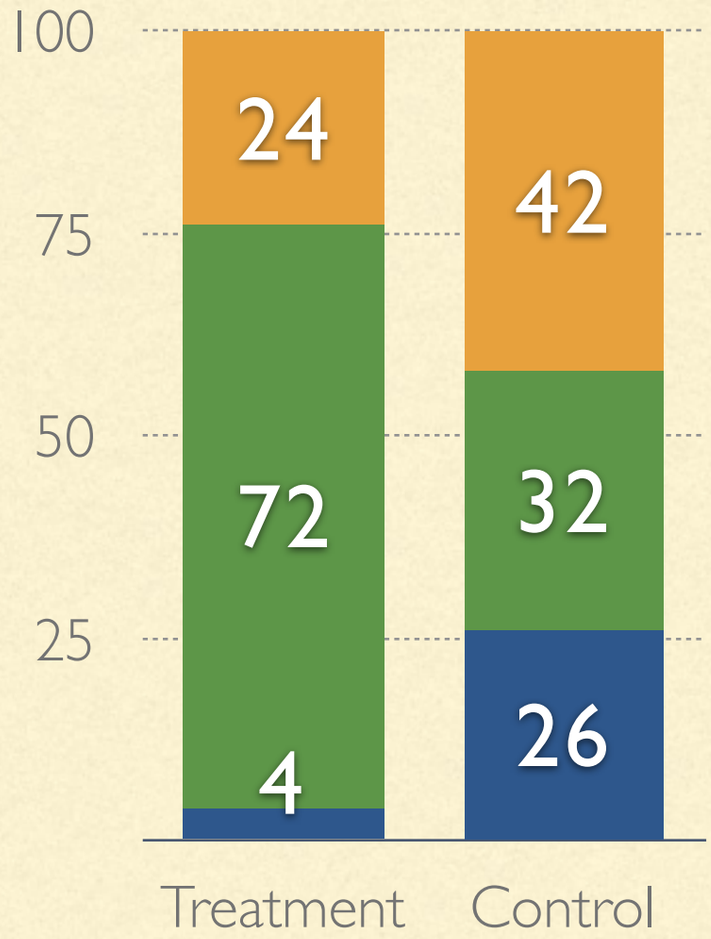
Observed relative abundance $\propto$ True relative abundance $\times$ Taxon-specific efficiencies

Expected value of $\dfrac{W_{ij}}{\sum_k W_{ik}} = \dfrac{p_{ij}e_j}{\sum_k p_{ik}e_k}$

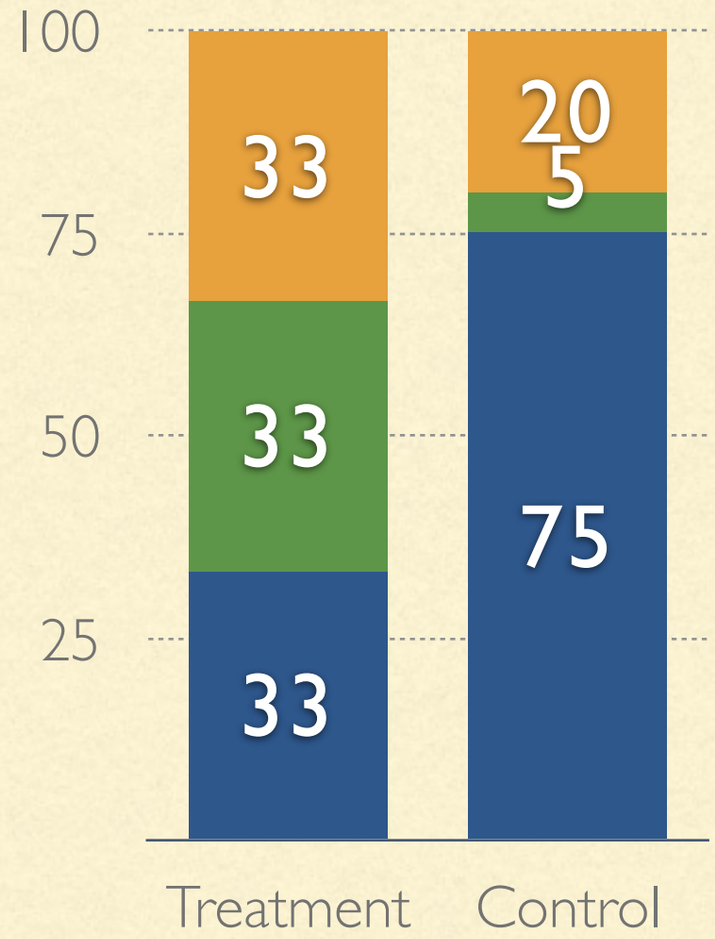


- A tempting conclusion:
  - The relative abundance of **taxon orange** *decreased* in the Treatment sample (left) compared to the Control sample (right)

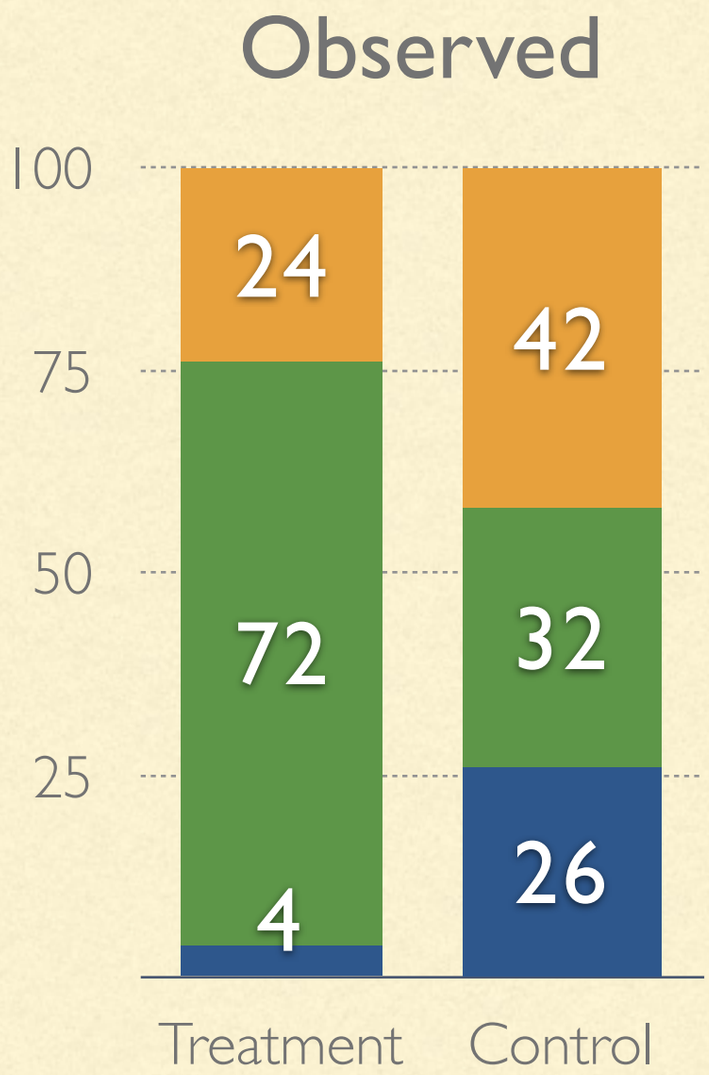
Observed

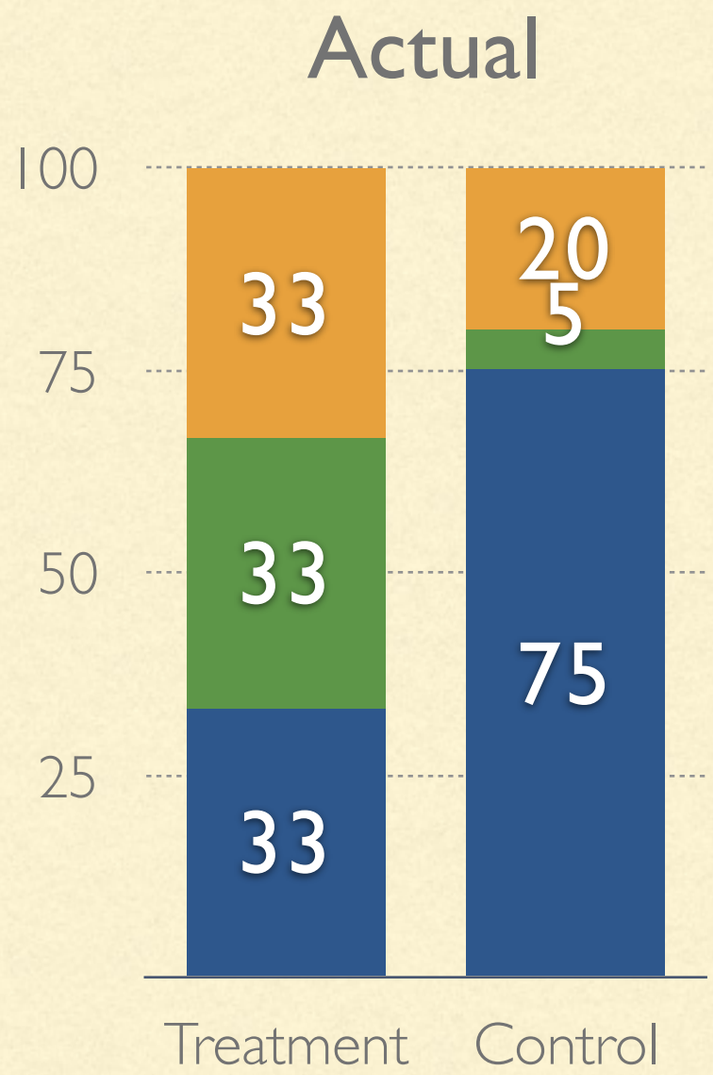


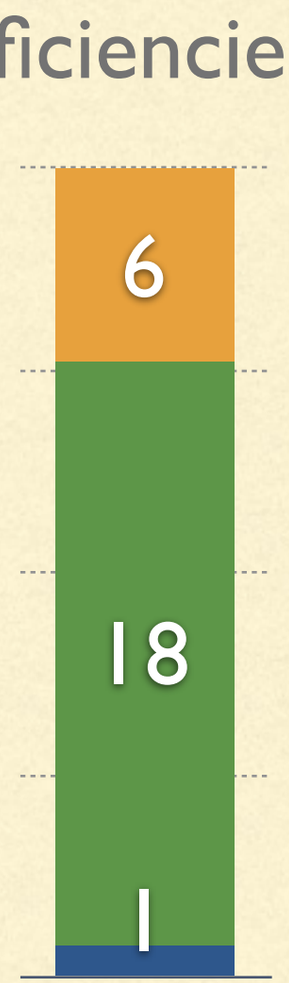
Actual



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

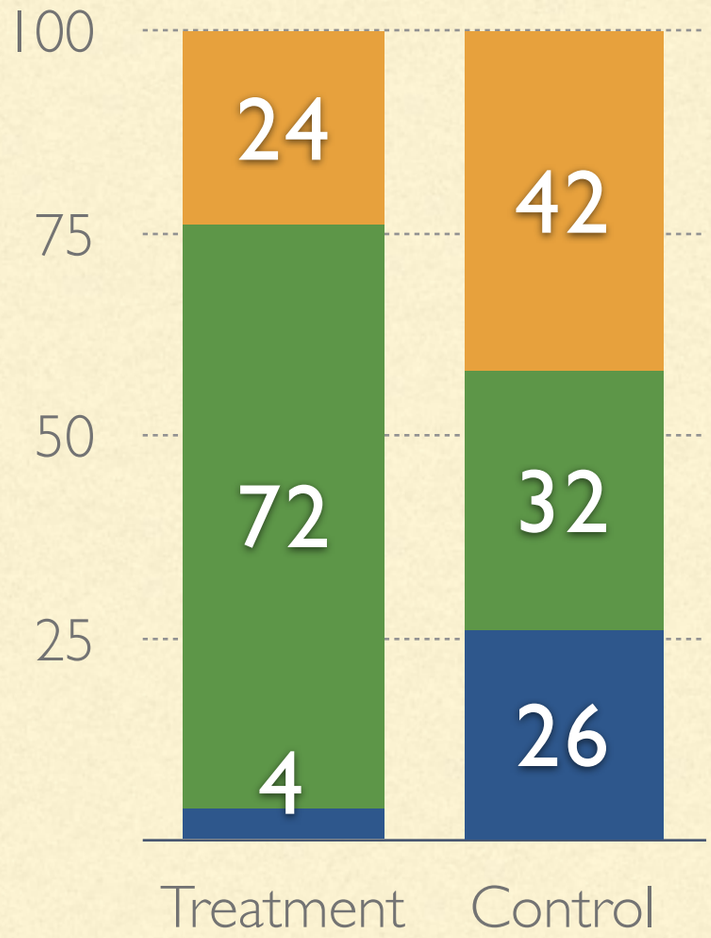




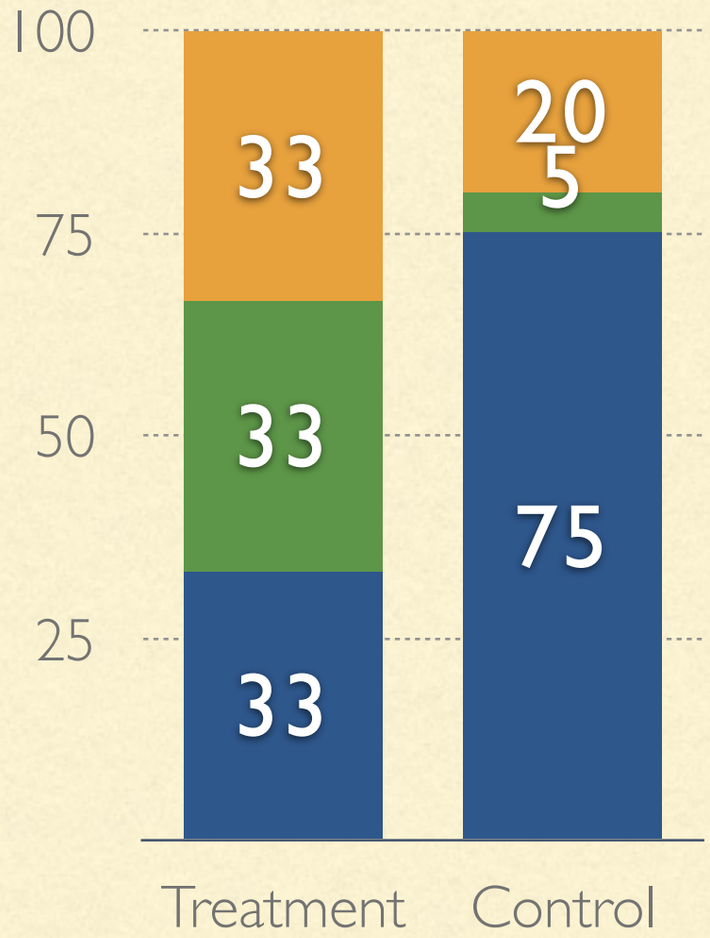


- In fact, the relative abundance of **taxon orange** *increased* in the Treatment sample compared to the Control sample

Observed



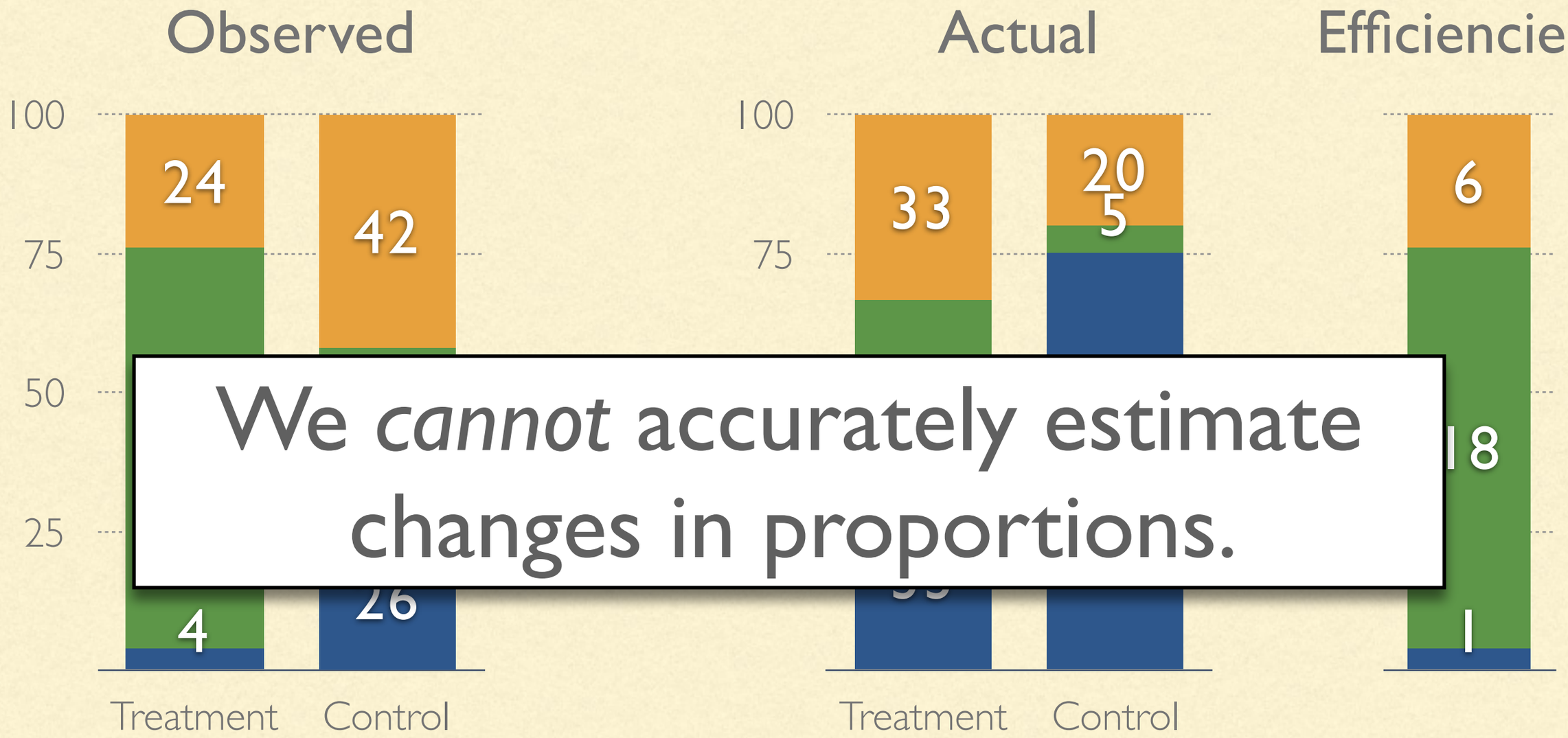
Actual



Efficiencies

- **Taxon green** is high efficiency; its abundance increased (Ctrl vs Tmt). Additionally, **taxon blue** is low efficiency, and its abundance decreased.

- **Taxon orange**'s abundance depends on the abundance of **the other taxa**

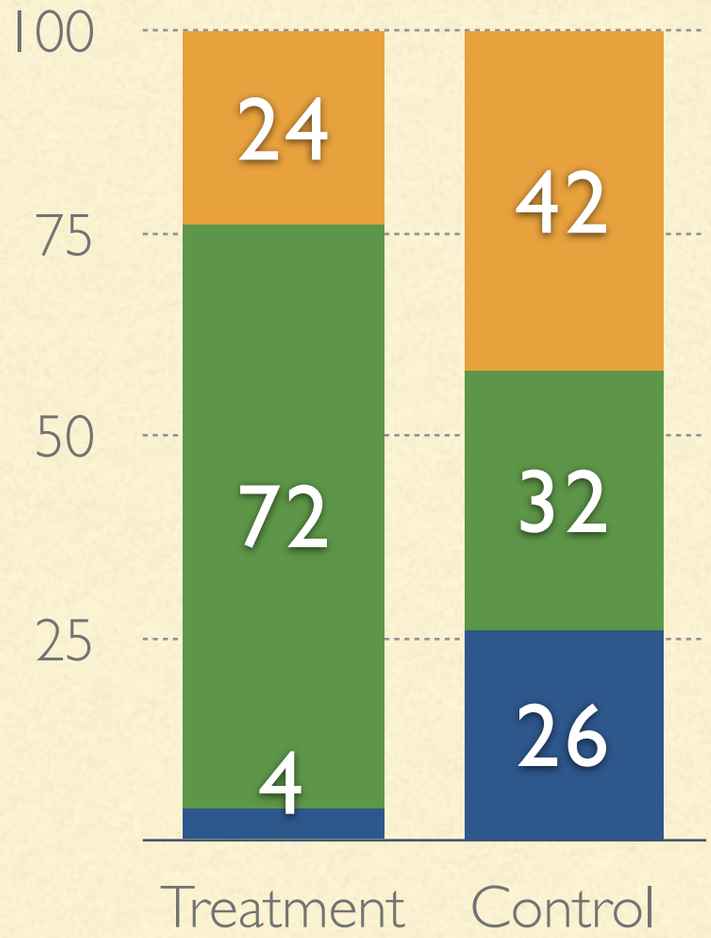


We *cannot* accurately estimate changes in proportions.

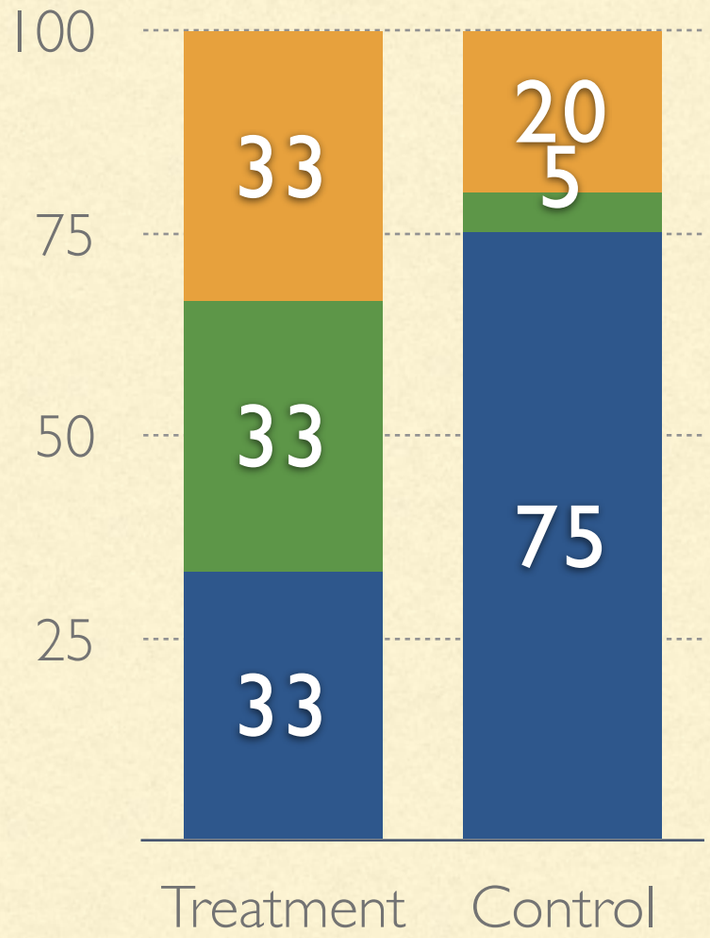
- **Taxon green** is high efficiency; its abundance increased (Ctrl vs Tmt). Additionally, **taxon blue** is low efficiency, and its abundance decreased.

- **Taxon orange**'s abundance depends on the abundance of **the other taxa**

Observed

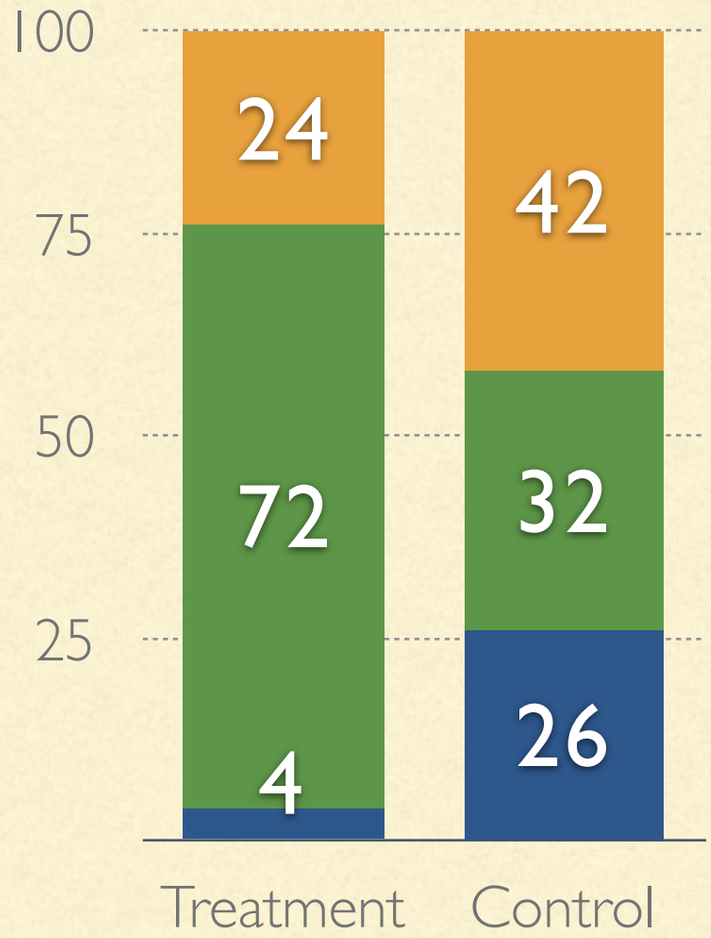


Actual



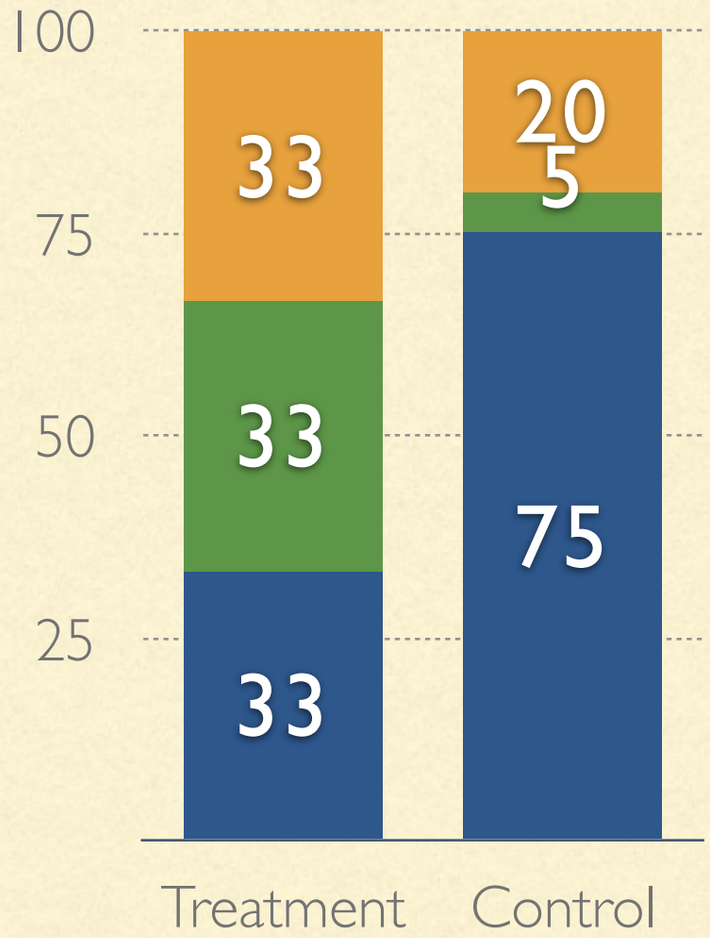
Efficiencies

Observed



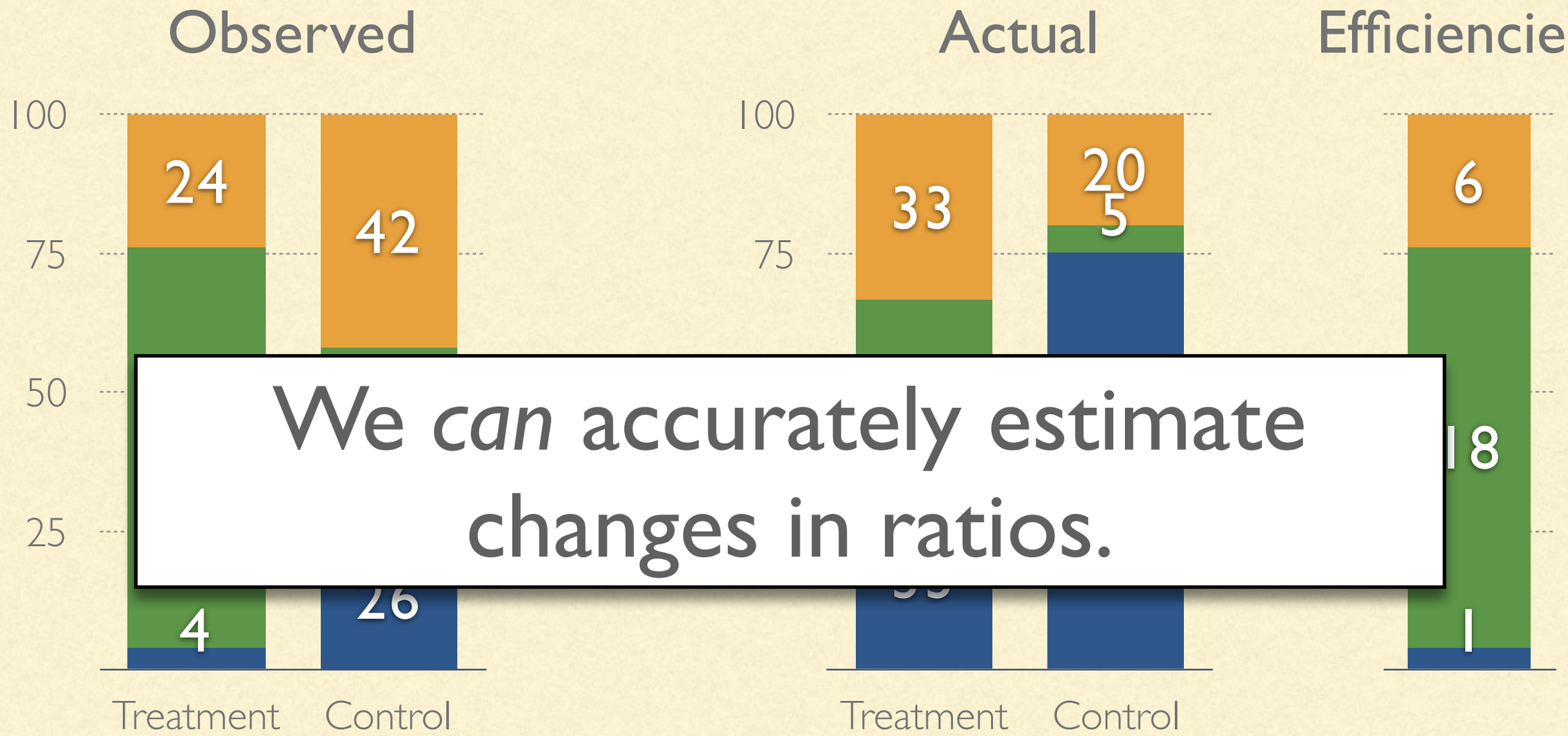
$$\frac{72}{4} \Big/ \frac{32}{26} = 15$$

Actual



$$\frac{33}{33} \Big/ \frac{5}{75} = 15$$

Efficiencies



We *can* accurately estimate changes in ratios.

$$\frac{\frac{72}{4}}{\frac{32}{26}} = 15$$

$$\frac{\frac{33}{33}}{\frac{5}{75}} = 15$$

eLIFE elifesciences.org

RESEARCH ARTICLE

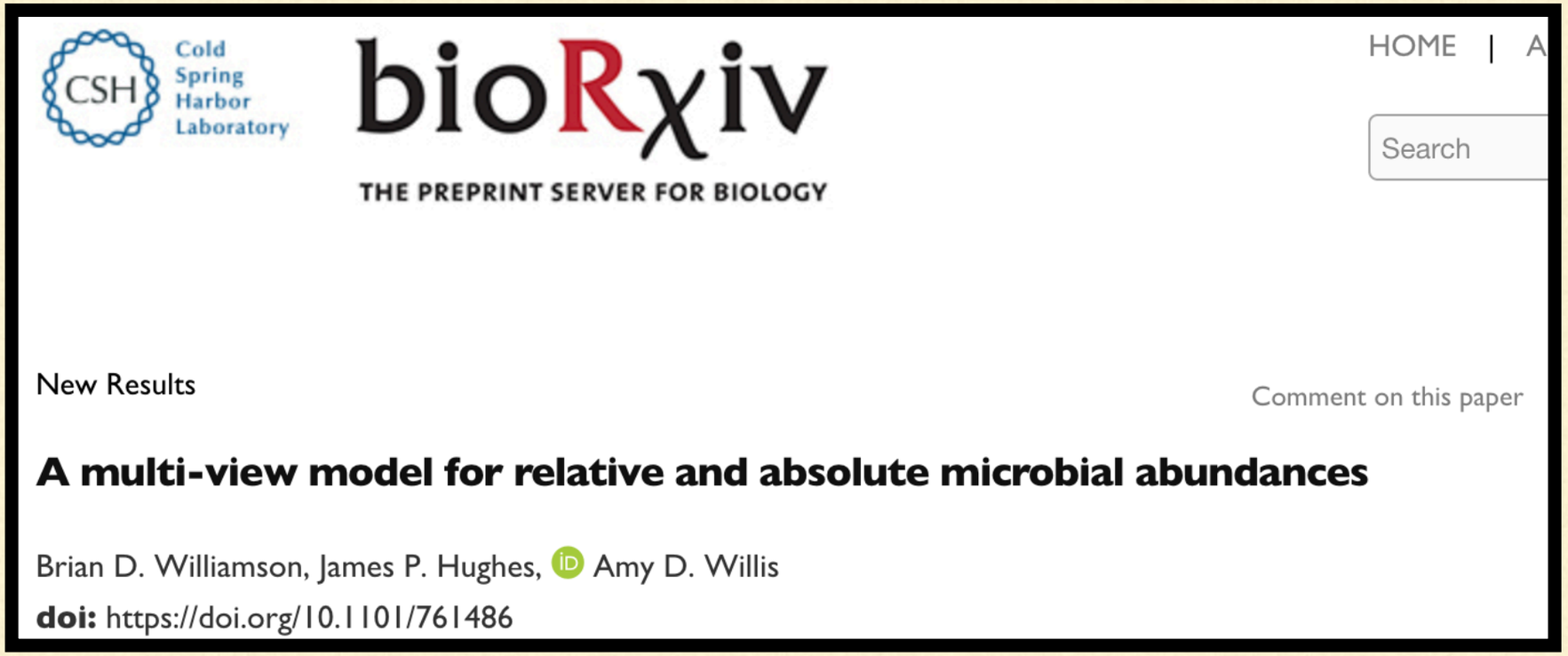
Consistent and correctable bias in metagenomic sequencing experiments

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

Michael McLaren (NCSU)

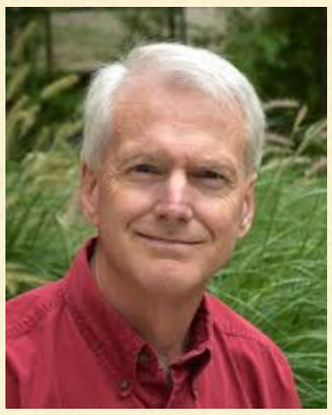
Ben Callahan (NCSU)

- Proposed model: taxon-specific efficiencies ("biases") confound estimation of relative abundance
- Model is validated in 16S & shotgun metagenomic data
  - Shotgun data is *also biased* — see *paper for details*



Brian Williamson (UW)

- Calibration with taxon-specific qPCR + 16S
  - Combing low-throughput + high-throughput
- *Key implication:* Do not multiply total 16S concentration and 16S relative abundance to estimate taxon-specific concentration



Jim Hughes (UW)

IMPLICATIONS

- Analyse ratios, not proportions
  - If you must analyse proportions, focus on largest signals
- Avoid aggregating abundances by taxonomy/phylogeny
- Avoid “quantitative” diversity indices

ONGOING & FUTURE WORK

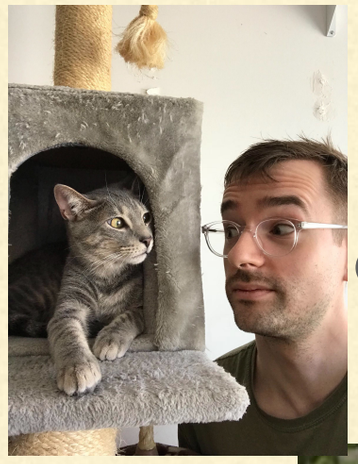
- Robust statistical methods
- Protocol selection
- Optimal experimental design (positive + negative controls)
- Bias prediction from phylogeny
- Calibration & meta-analysis (?)

SUMMARY

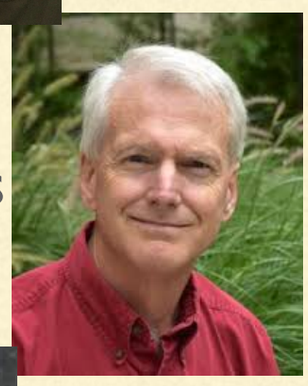
- Proposed a mathematical description of bias in sequencing
  - Validated on 16S and shotgun metagenomics data
- Model has serious implications for analysing microbiome data
  - *Proportions can lead to conclusions in the wrong direction*
  - Certain analysis methods are robust to bias; others are not

Michael McLaren (NCSU)

Ben Callahan (NCSU)



David Clausen (UW)



Jim Hughes (UW)

Brian Williamson (UW[2,3])

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]*

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | A

Search

New Results

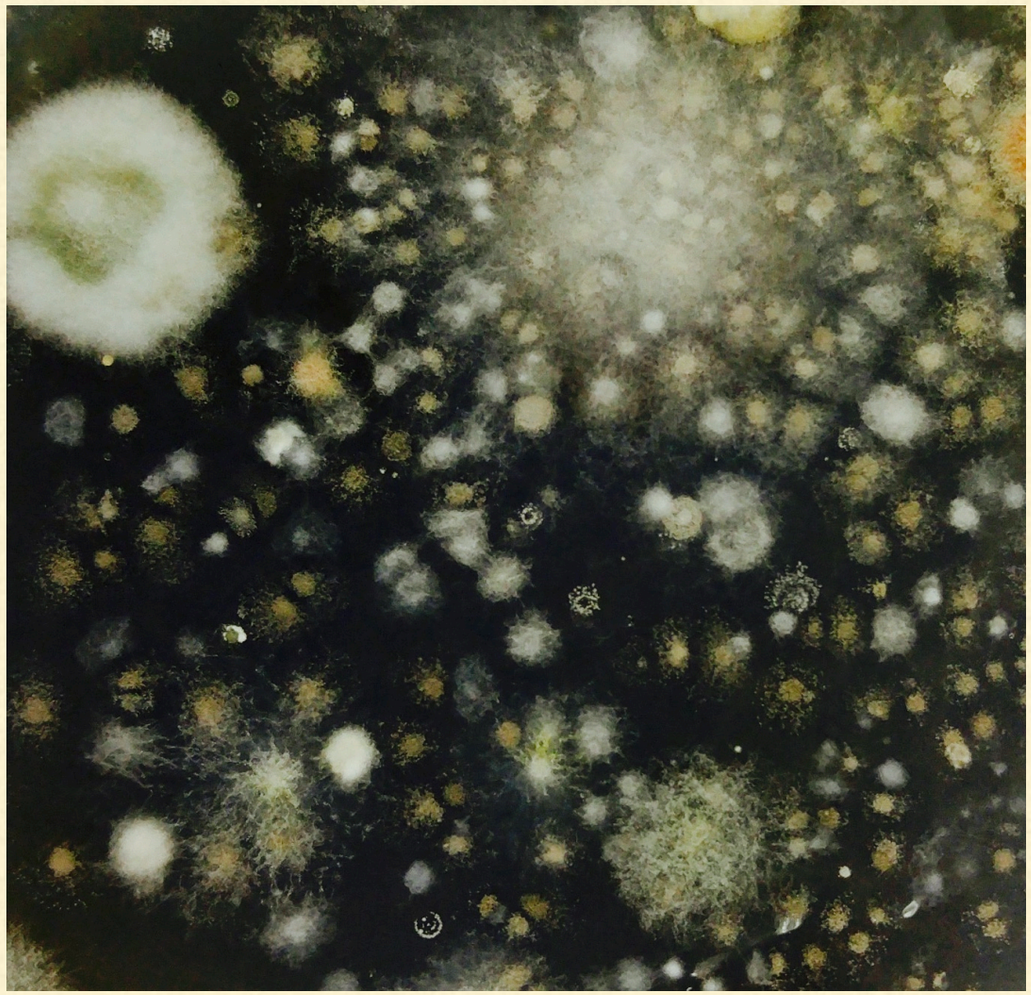
Comment on this paper

A multi-view model for relative and absolute microbial abundances

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

doi: https://doi.org/10.1101/761486

MEASUREMENT ERROR IN MICROBIOME DATA



Amy D Willis PhD

Assistant Professor

Department of Biostatistics

University of Washington, USA

@AmyDWillis

adwillis@uw.edu

📷 ✅

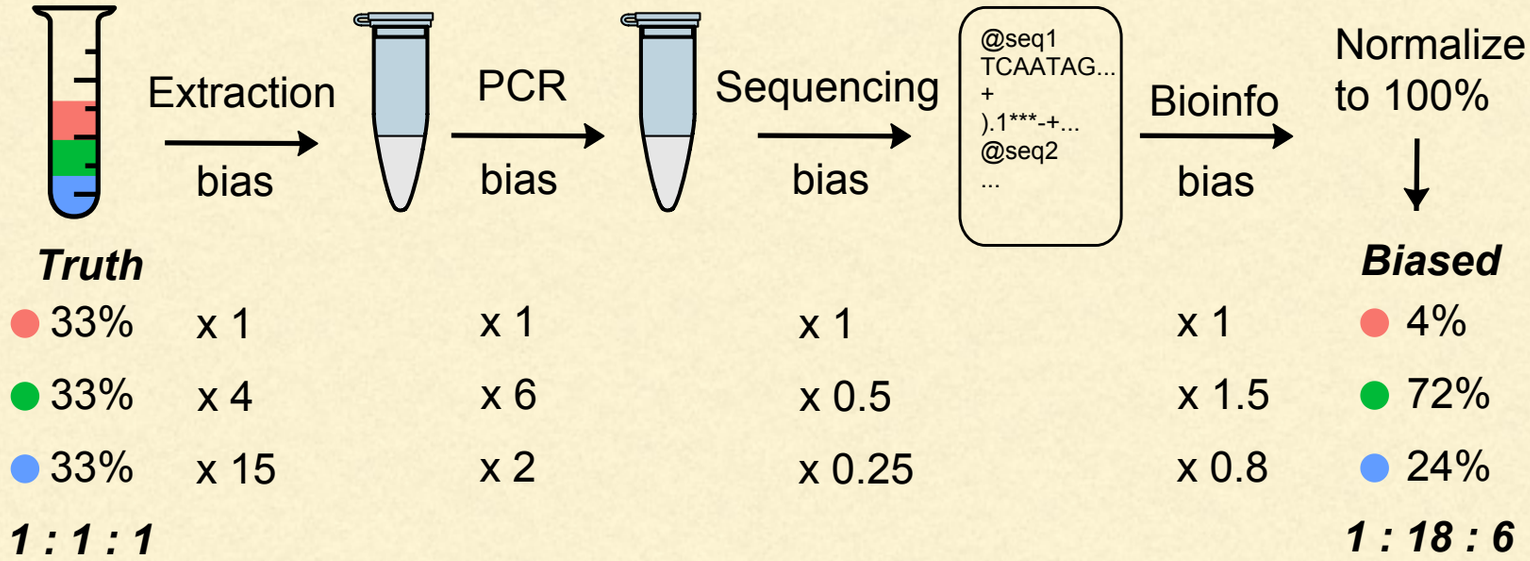
📬 ✅

🐦 ✅

Slides available at github.com/adw96/presentations

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

SUPPLEMENTARY SLIDES



Amy D Willis PhD

Assistant Professor

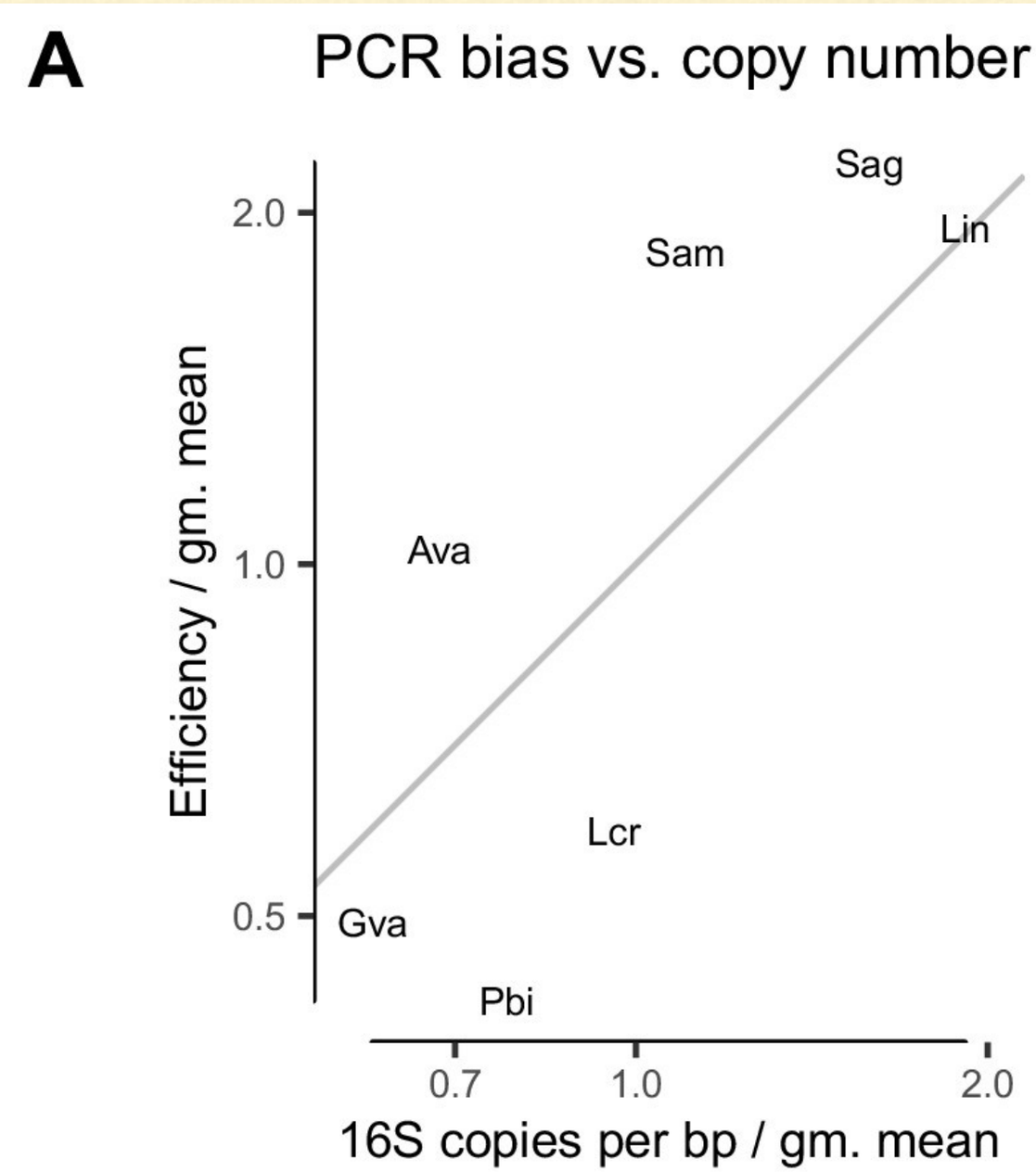
Department of Biostatistics

University of Washington, USA

@AmyDWillis

adwillis@uw.edu

Figure from McLaren et al., 2019, *eLife*



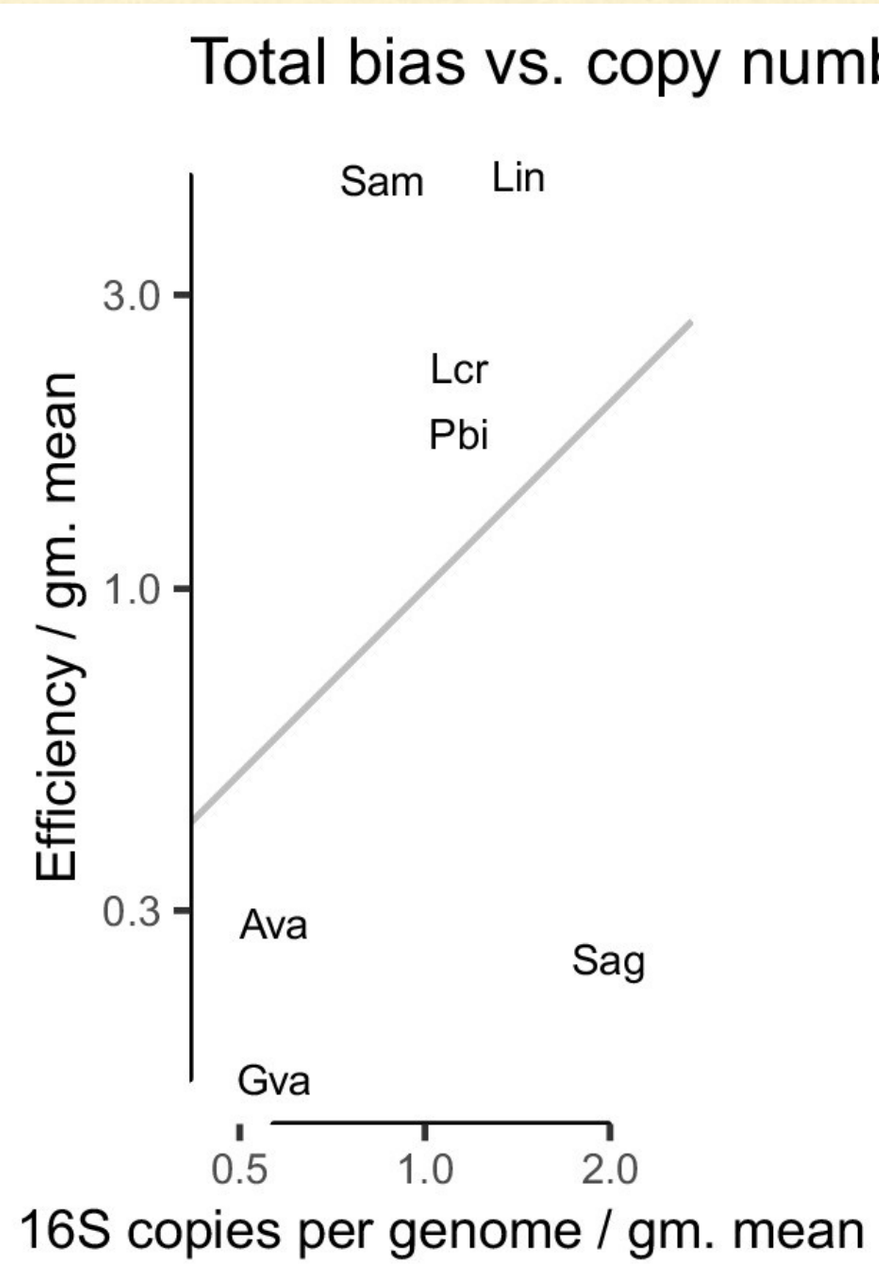


Figure from McLaren et al., 2019, *eLife*

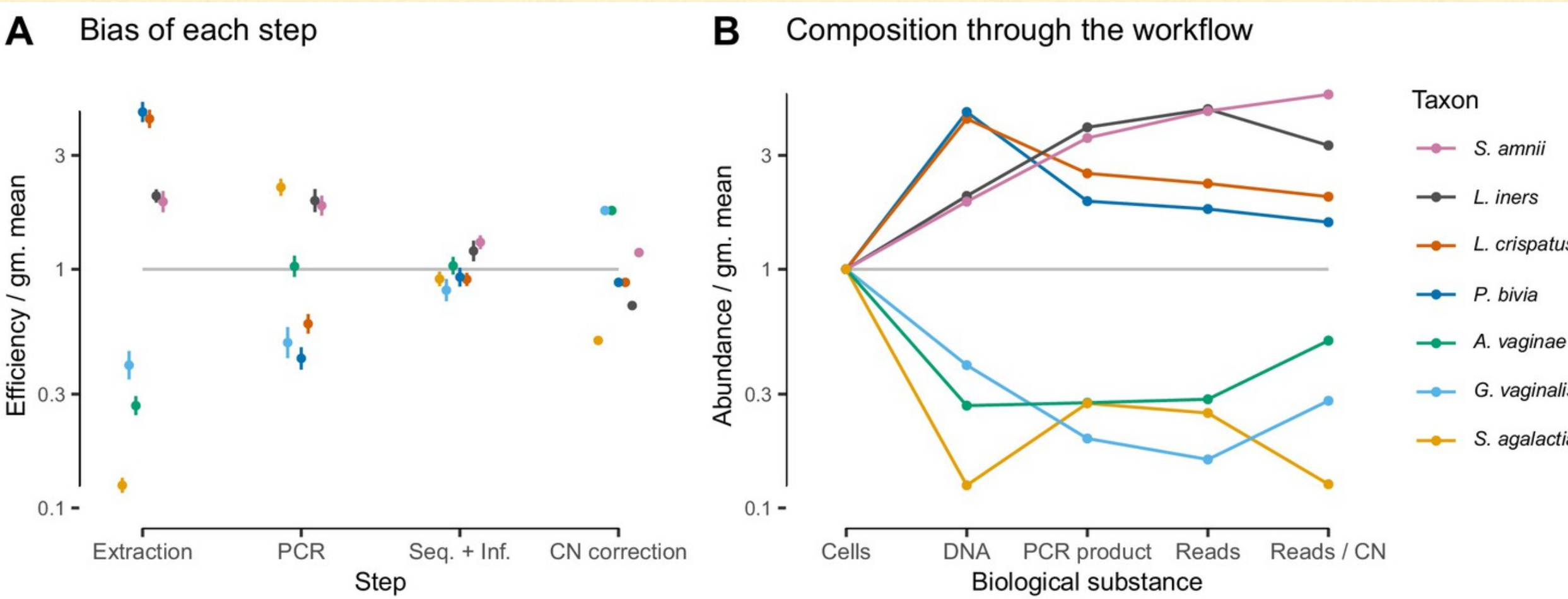
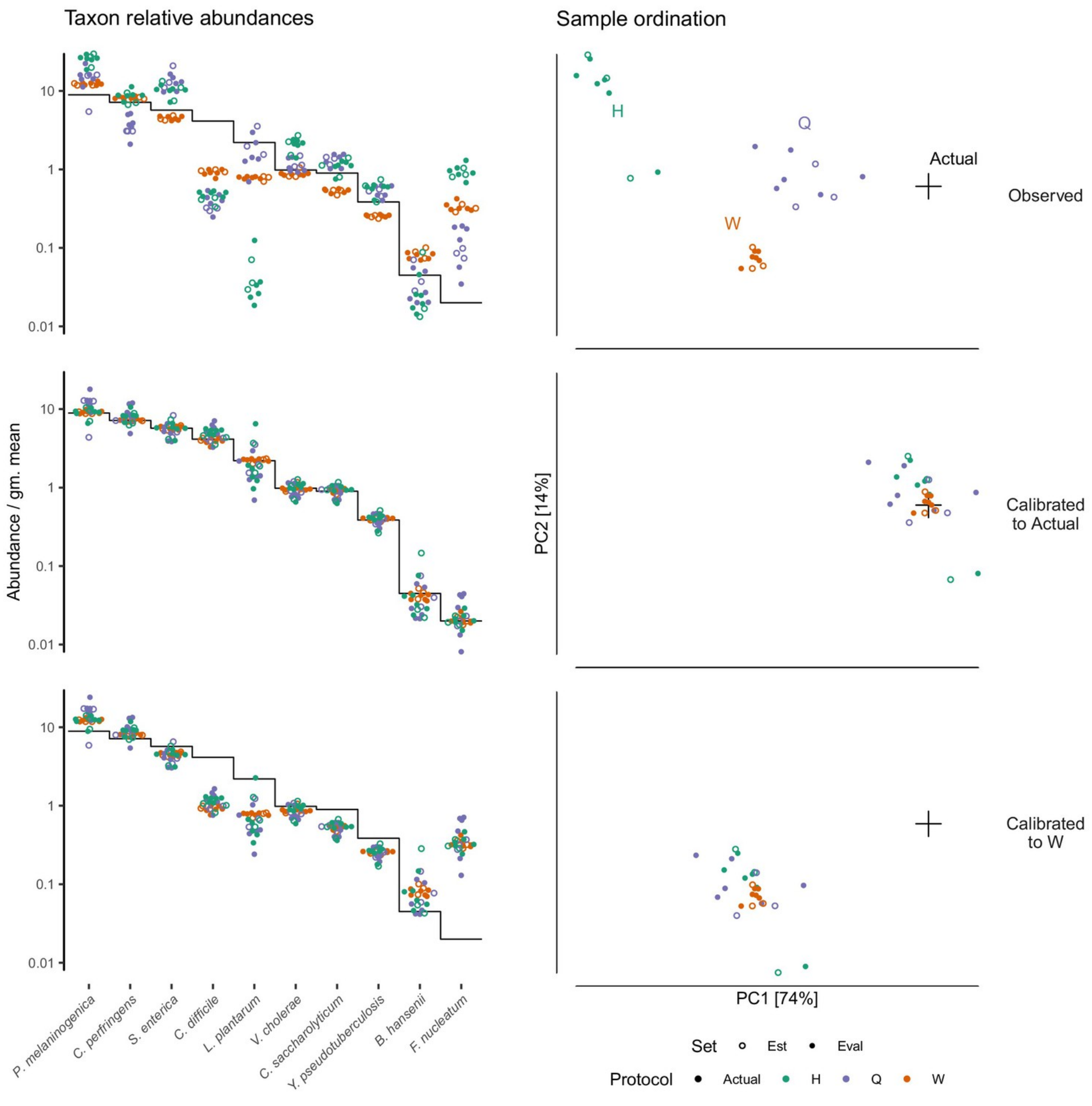


Figure from McLaren et al., 2019, *eLife*


Taxon relative abundances
Sample ordination
Abundance / gm. mean
10
1
0.1
0.01
H
Q
W
Actual
Observed
Calibrated to Actual
Calibrated to W
PC2 [14%]
PC1 [74%]
P. melaninogenica
C. perfringens
S. enterica
C. difficile
L. plantarum
V. cholerae
C. saccharolyticum
Y. pseudotuberculosis
B. hansenii
F. nucleatum
Set Est Eval
Protocol Actual H Q W

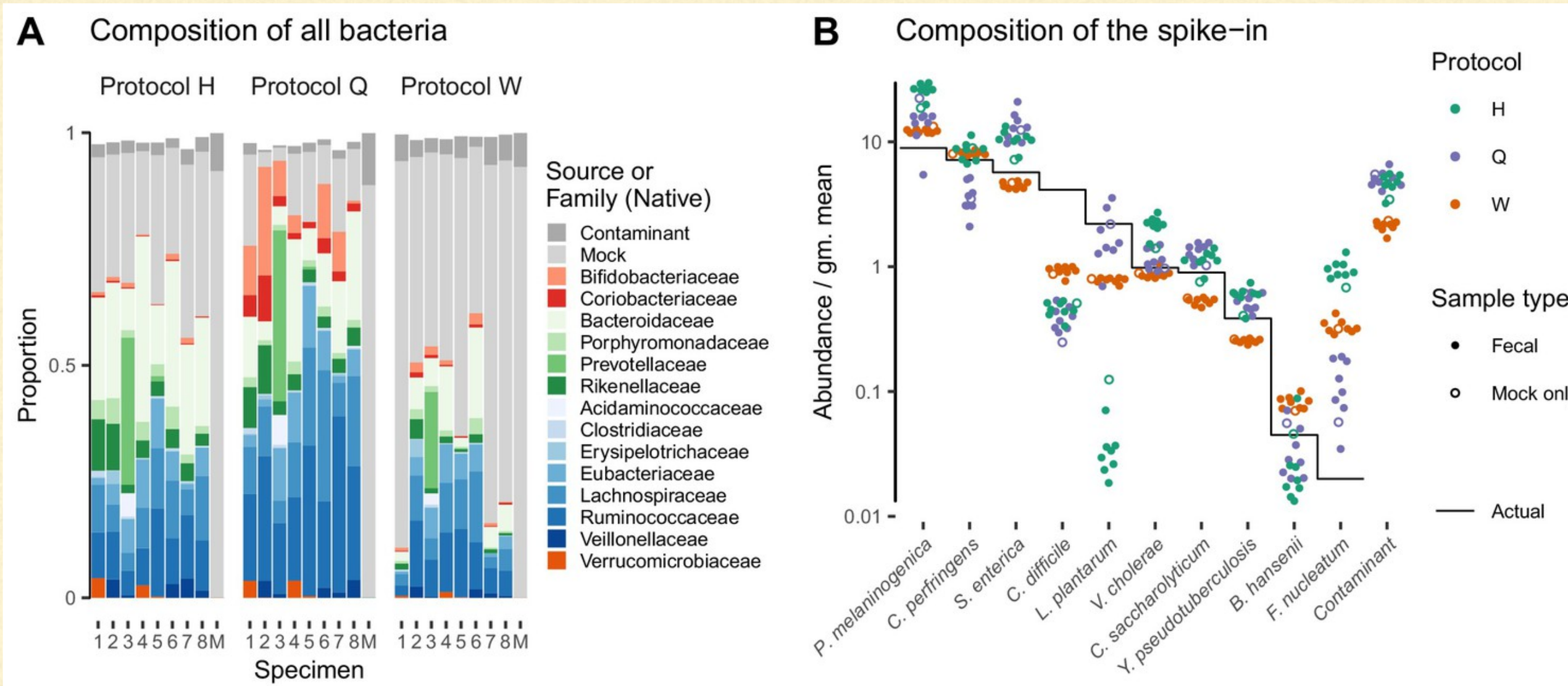


Figure from McLaren et al., 2019, *eLife*

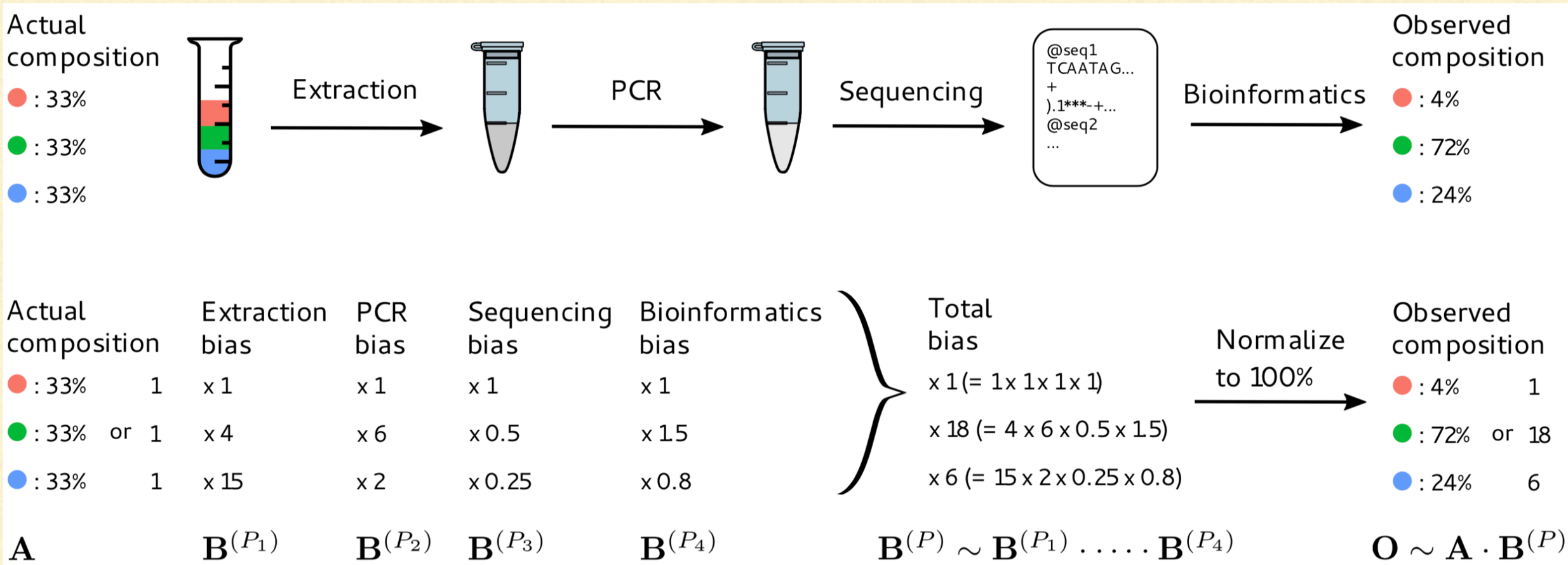


Figure from McLaren et al., 2019, *eLife*

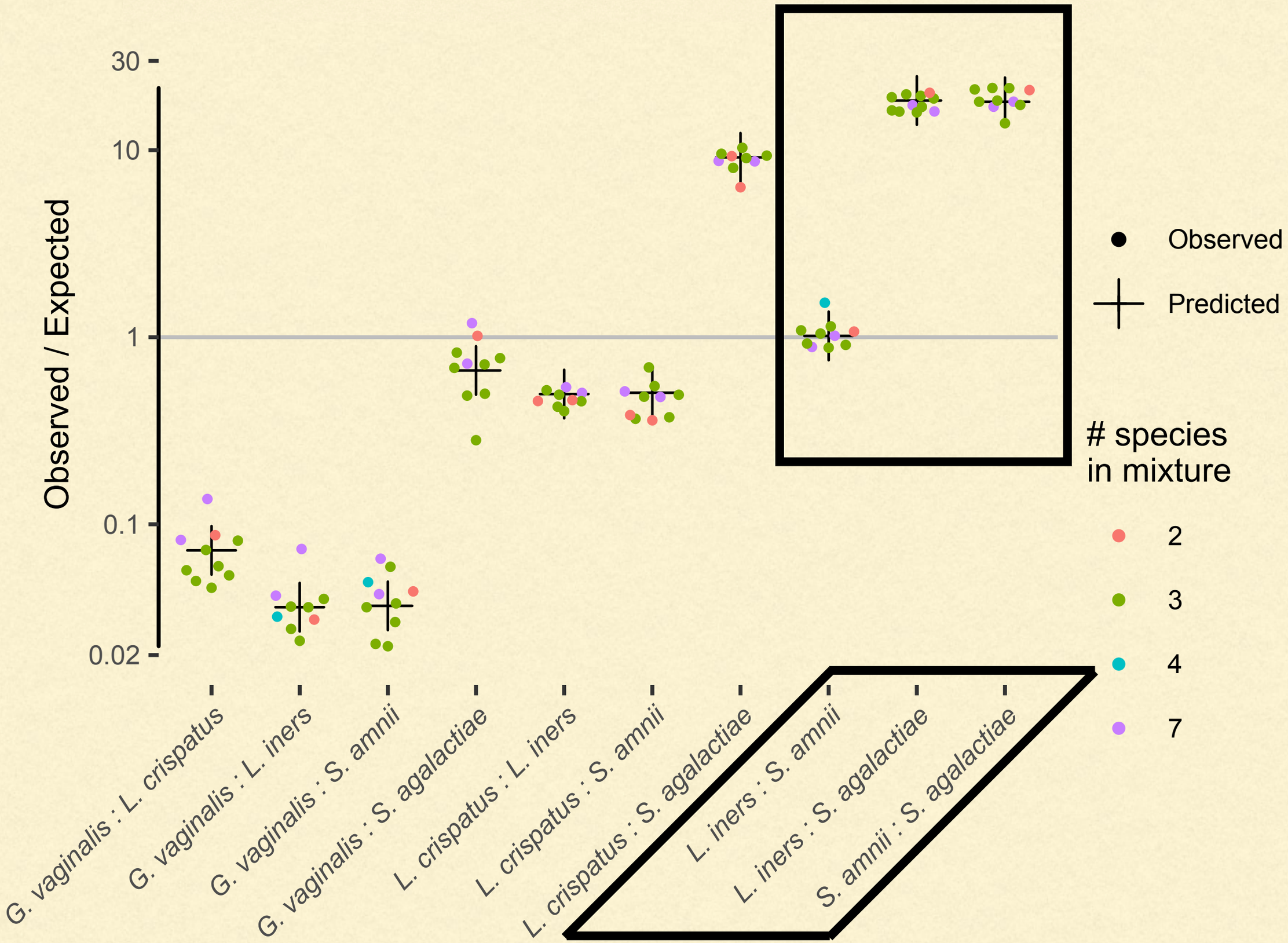


Figure from McLaren et al., 2019, *eLife*

WHY ARE RATIOS OF RATIOS UNAFFECTED BY BIAS?

$$\frac{\frac{p_{11}e_1}{\sum p_{1j}e_j}}{\frac{p_{12}e_2}{\sum p_{1j}e_j}} \Big/ \frac{\frac{p_{21}e_1}{\sum p_{2j}e_j}}{\frac{p_{22}e_2}{\sum p_{2j}e_j}} = \frac{p_{11}e_1}{p_{12}e_2} \Big/ \frac{p_{21}e_1}{p_{22}e_2} = \frac{p_{11}}{p_{12}} \Big/ \frac{p_{21}}{p_{22}}$$

p_{ij} = proportion of taxon j in sample i
e_j = efficiency of taxon j

HOW CONSISTENT IS MICROBIOME DATA WITHIN AND ACROSS LABS?

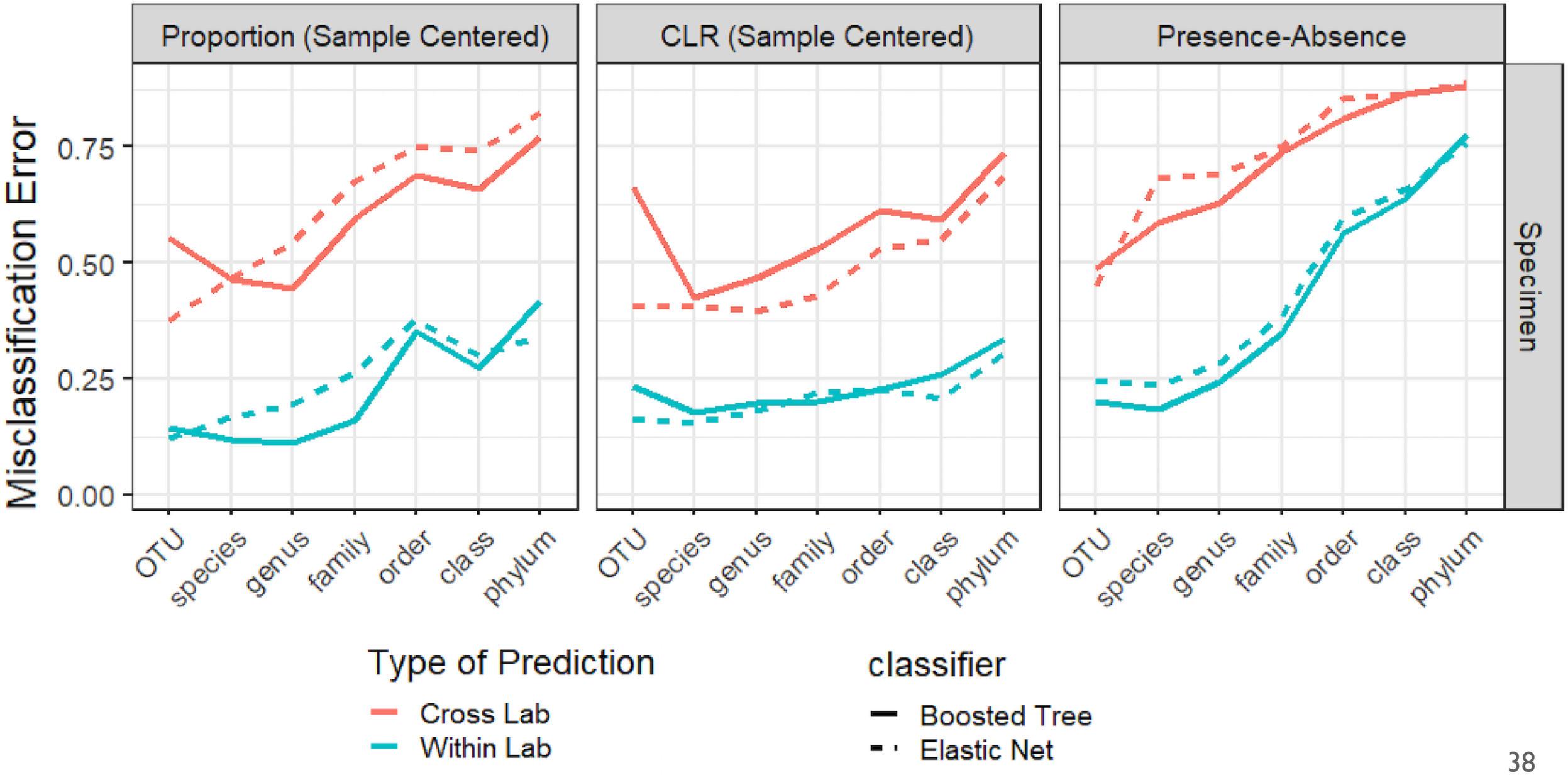


Figure from Clausen & Willis, 2019+