MODELING SHOTGUN SEQUENCING DATA

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Assistant Professor

David Clausen — @davidandacat — PhD Candidate

Sarah Teichman — @sarah_teichman — PhD Candidate

KEY IDEAS: MODELING SHOTGUN DATA

- Same principles as always
- Frame scientific question as statistical question
- Wrangle data into form accepted by software
- Fit model using software

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

THE PLAN

- Statistical methods for modeling metagenomic data
  - Abundance
  - Presence/absence
  - Comparing protocols
  - Trees - tomorrow!

ABUNDANCE

MODELING ABUNDANCE

- Many statistical methods for modeling microbial abundances can be applied to *both* amplicon and shotgun data

- Challenge: wrangling data into needed formats

MODELING ABUNDANCE VIA SHOTGUN DATA

- Examples
  - k-mer / groups-of-k-mers **counts** *can* be modeled with corncob
    - corncob scales nicely since it is easy to parallelise
    - Better: k-mer groups
      - k-mers that map to a common reference (sourmash gather)
      - k-mer that are adjacent in an assembly graph (spacegraphcats), ….
  - metaphlan2 **relative abundances** w/ radEmu
  - anvi-estimate-scg-taxonomy **coverages** w/ radEmu

MODELING ABUNDANCE

- Recall radEmu
  - Goal: identify taxa that are differentially abundant across samples — in "absolute" sense
    - "absolute sense" = DNA copies per uL (or cell counts per uL), not relative abundance
- Compromise: identify taxa that are changing the **most** compared to other taxa in absolute abundance

RADEMU

- We wish we could fit the model

$$\log\left(\text{mean total abundance}_{ij}\right) = \beta_0 + \beta_{1j}X_{i1} + \ldots + \beta_{pj}X_{ip}$$

  - $e^{\beta_{kj}}$ gives the ratio between the expected abundance of taxon j between samples that differ in $X_{\cdot k}$ by one unit but are the same wrt all other covariates

- We don't observe total abundance from HTS

  - We can't fit this model directly

RADEMU

- We *can* observe HTS data
- We fit the model

$$\log\left(\text{expected HTS observation}_{ij}\right) = \beta_{0j} + \beta_{1j}X_{i1} + \ldots + \beta_{pj}X_{ip} + s_i + e_j$$

- Critical fact: The $\beta_{1j}, \ldots, \beta_{pj}$'s across these two models are the same

RADEMU

- $\log\left(\text{expected HTS observation}_{ij}\right) = \beta_{0j} + \beta_{1j}X_{i1} + \ldots + \beta_{pj}X_{ip} + s_i + e_j$
- Challenge: can't pull apart β_{0j} and e_j
- Compromise: Don't try to estimate everything… instead…
  - Identify taxa j's such that β_{1j} is large relative to other β_{1j}'s
  - e.g., Constrain $\text{median}_j \beta_{1j} = 0$
    - Could also pick a reference taxon whose abundance isn't changing

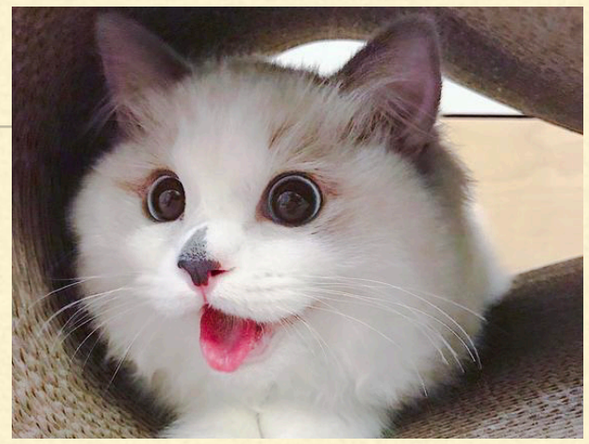
MODELING ABUNDANCE

- radEmu
  - Modeling absolute abundance from count data
  - Modeling absolute abundance from coverage (depth) data
  - Modeling absolute abundance from proportion data
- corncob
  - Modeling relative abundance from count data
- Many others
- Challenge: wrangling metagenome data into form software handles

MODELING ABUNDANCE LAB

- The `radEmu` lab looks at coverage (depth) data
  - So you've already worked through a differential abundance lab looking at shotgun data
- Feel free to return to this lab if you didn't finish it already

ACCESSING `RADEMU` LAB

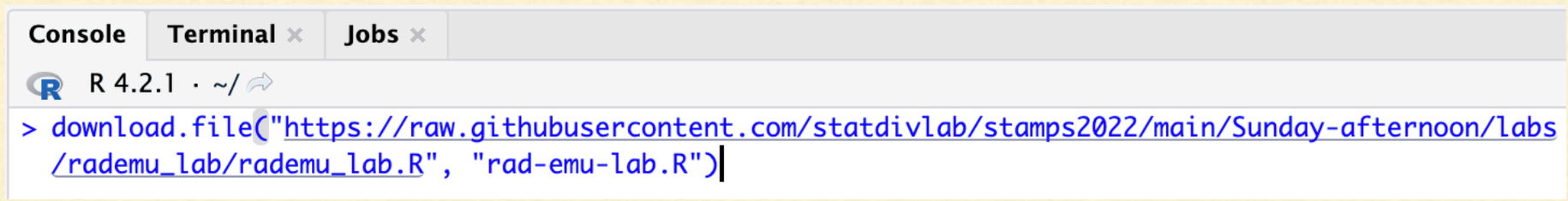


Get pumped!

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

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

3. Run this command in your RStudio Server console



PRESENCE/ABSENCE

PRESENCE/ABSENCE

- Sometimes, you care about abundance
- Sometimes, only *presence* is important
  - Presence of a gene
  - Presence of a pathway
  - Presence of a strain/genome/population/taxon
  - Presence of a pathogen 🏴‍☠️

PRESENCE/ABSENCE

- Challenge: Not everything that is *present* is *observed*
  - Non-detection ≠ Absence
    - Inexhaustive sequencing, incomplete use of reads, reads are short…
  - Detection ≠ Presence
    - Contamination (computational or physical)

PRESENCE/ABSENCE

- We know things about the quality of our data
  - Metagenomes: sequencing depth
  - Genomes: completeness, redundancy, length…

happi: a Hierarchical Approach to Pangenomics Inference

Pauline Trinh[1], David S. Clausen[2], and Amy D. Willis[2]

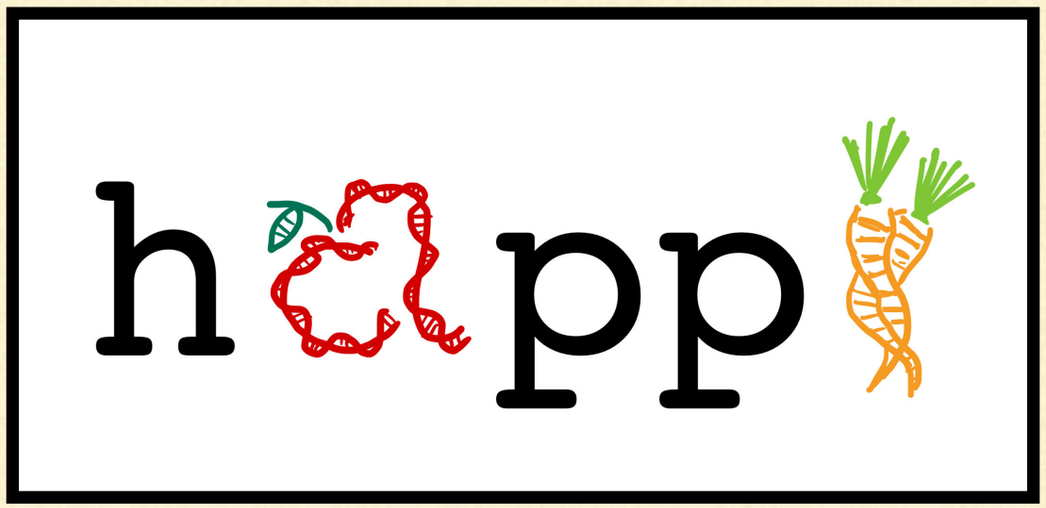
[1]Department of Environmental & Occupational Health Sciences, University of Washington
[2]Department of Biostatistics, University of Washington

April 26, 2022

Abstract

Recovering metagenome-assembled genomes (MAGs) from shotgun sequencing data is an increasingly common task in microbiome studies, as MAGs provide deeper insight into the functional potential of both culturable and non-culturable microorganisms. However, metagenome-assembled genomes vary in quality, and may contain omissions and contamination. These errors present challenges for detecting genes and comparing gene enrichment across sample types. To address this, we propose `happi`, an approach to testing hypotheses about gene enrichment that accounts for genome quality. We illustrate the advantages of `happi` over existing approaches using published Saccharibacteria MAGs and via simulation.

🍎🍎 github.com/statdivlab/happi🥕🥕



HAPPI

- `happi` addresses the mismatch between detection and presence

- Hierarchical model

1. odds gene j is present in sample $i = e^{\beta_{0j}+\beta_{1j}X_{i1}+\ldots+\beta_{pj}X_{ip}}$

2. probability gene is correctly detected = function of quality variables

3. probability gene is incorrectly detected = ε

HAPPI

- Allows test of null hypothesis that the odds that the gene will be present are equal when comparing groups of genomes that differ in $X_{.k}$ by one unit but are the same wrt all other covariates

- "Best competitor": pretend data doesn't have errors
  - happi produces larger p-values when differences in presence likely attributable to sequencing depth
  - happi produces smaller p-values when differences in presence likely attributable to biology

HAPPI

- Bonus
  - Can apply to presence/absence of anything
    - Marker genes (eg 16S)
    - k-mers
    - ….

ACCESSING `HAPPI` LAB

Get pumped!

1. Go to schedule on Wiki to Thursday afternoon, click on "Labs"

2. Choose your own adventure
   - Gene presence in MAGs recovered from shotgun sequencing
     - happi shotgun lab
   - 16S gene presence in 16S amplicon sequencing
     - happi 16S lab

3. Copy the command under the lab we're working on; run this command in your RStudio Server console; open file and start working through

USING CONTROL DATA

STEPS IN SEQUENCING MICROBIOMES

- Collecting, handling and storing samples
- Breaking open bacterial cells
- (Selecting for the gene being targeted)
- (Amplifying DNA)
- Sequencing DNA
- Processing sequences into quantitative data

SCIENTIFIC REPORTS

Open Access | Published: 17 November 2015

Sample storage conditions significantly influence faecal microbiome profiles

Jocelyn M Choo, Lex EX Leong & Geraint B Rogers

Scientific Reports **5**, Article number: 16350 (2015) | Cite this article

Storage conditions of intestinal microbiota matter in metagenomic analysis

Silvia Cardona, Anat Eck, Montserrat Cassellas, Milagros Gallart, Carmen Alastrue, Joel Dore, Fernando Azpiroz, Joaquim Roca, Francisco Guarner and Chaysavanh Manichanh

BMC Microbiology 2012 **12**:158

https://doi.org/10.1186/1471-2180-12-158

Received: 6 March 2012 | **Accepted:** 20 July 2012 | **Published:** 30 July 2012

Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome

Sanqing Yuan[1,4], Dora B. Cohen[1,4], Jacques Ravel[5], Zaid Abdo[2,3,4], Larry J. Forney[1,4]*

1 Department of Biological Sciences, University of Idaho, Moscow, Idaho, United States of America, **2** Department of Mathematics, University of Idaho, Moscow, Idaho, United States of America, **3** Department of Statistics, University of Idaho, Moscow, Idaho, United States of America, **4** Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, United States of America, **5** Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, United States of America

SC

Open A

Sample storag influence faec

Jocelyn M Choo, Lex EX Lec

Scientific Reports **5**, Article

4192 Accesses | **150** Cita

Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis

Agata Wesolowska-Andersen, Martin Iain Bahl, Vera Carvalho, Karsten Kristiansen, Thomas Sicheritz-Pontén, Ramneek Gupta and Tine Rask Licht

Microbiome 2014 **2**:19

https://doi.org/10.1186/2049-2618-2-19

Received: 3 February 2014 | **Accepted:** 25 April 2014 | **Published:** 5 June 2014

matter

Silvia Cardona, Anat Eck, Montserrat Cassellas, Milagros Gallart, Carmen Alastrue, Joel Dore, Fernando Azpiroz, Joaquim Roca, Francisco Guarner and Chaysavanh Manichanh

July 2012

Comparison of DNA extraction methods for human gut microbial community profiling

Mi Young Lim[a,*,1], Eun-Ji Song[a,b,1], Sang Ho Kim[c], Jangwon Lee[c], Young-Do Nam[a,b,*]

[a] *Research Group of Gut Microbiome, Division of Nutrition and Metabolism Research, Korea Food Research Institute, Jeollabuk-do 55365, Republic of Korea*
[b] *Department of Food Biotechnology, Korea University of Science and Technology, Daejeon 34113, Republic of Korea*
[c] *IVD Business Unit, SK Telecom, Seoul 04539, Republic of Korea*

Evaluation of Methods for the Extraction and Purification of DNA from the Hu

Sanqing Yuan[1,4], Dora B. Cohen[1,4], J

1 Department of Biological Sciences, University of Idaho, M
United States of America, 3 Department of Statistics, Univers
University of Idaho, Moscow, Idaho, United States of Americ
States of America

Library preparation methodology can influence genomic and functional predictions in human microbiome research

Marcus B. Jones, Sarah K. Highlander, Ericka L. Anderson, Weizhong Li, Mark Dayrit, Niels Klitgord, Martin M. Fabani, Victor Seguritan, Jessica Green, David T. Pride, Shibu Yooseph, William Biggs, Karen E. Nelson, and J. Craig Venter

PNAS November 10, 2015 112 (45) 14024-14029; published ahead of print October 28, 2015
https://doi.org/10.1073/pnas.1519288112

Check for updates

SC

Open A

Sample storag
influence faec

Jocelyn M Choo, Lex EX Lec

Choice
fecal m
evaluated by metagenomic analysis

n, Thomas Sicheritz-Pontén,

ee BioMed Central Ltd. 2014

matter

Carmen Alastrue, Joel Dore, Fernando Azpiroz,

see BioMed Central Ltd. 2012
July 2012

frontiers in Microbiology

Front Microbiol. 2016; 7: 459.
Published online 2016 Apr 20. doi: 10.3389/fmicb.2016.00459
PMCID: PMC4837688
PMID: 27148170

Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics

Juan Jovel,[1,*†] Jordan Patterson,[1,†] Weiwei Wang,[1] Naomi Hotte,[1] Sandra O'Keefe,[1] Troy Mitchel,[1] Troy Perry,[1] Dina Kao,[1] Andrew L. Mason,[1] Karen L. Madsen,[1] and Gane K.-S. Wong[1,2,3,*]

Mi Young Lim[a,*,1], Eun-Ji Song[a,b,1], Sang Ho Kim[c], Jangwon Lee[c], Young-Do Nam[a,b,*]

[a] *Research Group of Gut Microbiome, Division of Nutrition and Metabolism Research, Korea Food Research Institute, Jeollabuk-do 55365, Republic of Korea*
[b] *Department of Food Biotechnology, Korea University of Science and Technology, Daejeon 34113, Republic of Korea*
[c] *IVD Business Unit, SK Telecom, Seoul 04539, Republic of Korea*

Evaluation of Methods for the Extraction and Purification of DNA from the Hu

Sanqing Yuan[1,4], Dora B. Cohen[1,4], J

Library preparation methodology can influence genomic and functional han microbiome research

der, Ericka L. Anderson, Weizhong Li, Mark Dayrit,
tor Seguritan, Jessica Green, David T. Pride, Shibu Yooseph,
J. Craig Venter

4029; published ahead of print October 28, 2015

16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice

Alan W. Walker, Jennifer C. Martin, Paul Scott, Julian Parkhill, Harry J. Flint and Karen P. Scott

Microbiome 2015 **3**:26
https://doi.org/10.1186/s40168-015-0087-4
Received: 14 January 2015 **Accepted:** 4 June 2015 **Published:** 22 June 2015

ic analysis

n, Thomas Sicheritz-Pontén,

matter

ee BioMed Central Ltd. 2014

frontiers in Microbiology

Front Microbiol. 2016; 7: 459.
Published online 2016 Apr 20. doi: 10.3389/fmicb.2016.00459
PMCID: PMC4837688
PMID: 27148170

Character
Metagenc

Juan Jovel,[1,*†]
Dina Kao,[1] An

Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample

Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read, Konstantinos T. Konstantinidis

Published: February 10, 2012 • https://doi.org/10.1371/journal.pone.0030087

Mi Young

[a] *Research Group*
[b] *Department of F*
[c] *IVD Business Unit, SK Telecom, Seoul 04539, Republic of Korea*

CURRENT PARADIGM

- Everyone believes their choices are the best
- No one will ever agree on what's the best
- *There is no universal best*
- Not possible to confirm or refute scientific findings

SCOPE

- Use control data to better understand how we go from communities to data
  - via high-throughput sequencing
- Introduce a model to improve the accuracy of community composition estimates
  - Not *differential* abundance (neither relative nor absolute)

MICROBIOME DATA

- Common data structure
  - W_{ij} = number of times strain j observed in sample i
  - Alternatively, *coverage* of strain j observed in sample i

	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

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

MICROBIOME PARAMETERS

- p_{ij} = true relative abundance of taxon j in sample i

- Common working assumption:

observed relative abundance $\propto$ true relative abundance

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

- How can we investigate this assumption?

MODEL VALIDATION

- Mock community: An artificially constructed community of known composition
  - A useful tool for understanding the data generating process

MODEL VALIDATION

- Mock community: An artificially constructed community of known composition
  - A useful tool for understanding the data generating process

	L.crispatus	L.iners	A.vaginae	S.agalactiae	G.vaginalis	S.amnii	P.bivia
1	0.00	0.00	0.0	1.00	0	0	0.00
2	0.00	0.00	0.5	0.00	0	0	0.50
3	0.33	0.33	0.0	0.00	0	0	0.33
4	0.33	0.33	0.0	0.33	0	0	0.00

MODEL VALIDATION

	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	0.00	0.00	0.0	1.00	0	0	0.00
2	0.00	0.00	0.5	0.00	0	0	0.50
3	0.33	0.33	0.0	0.00	0	0	0.33
4	0.33	0.33	0.0	0.33	0	0	0.00

MODEL VALIDATION

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

2. Despite being purportedly absent, taxa are observed

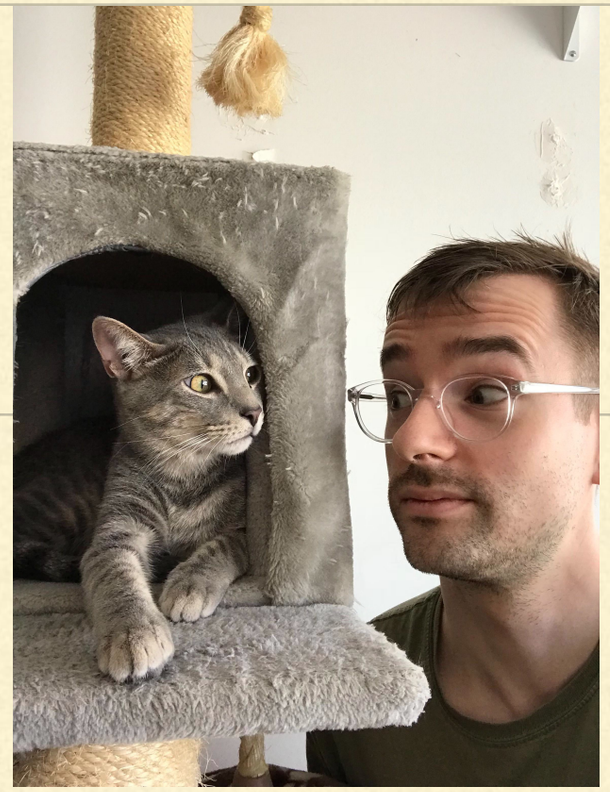
MODEL VALIDATION

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

2. Despite being purportedly absent, taxa are observed

Goal: Use control data to propose and validate a more sensible model

PROPOSED MODEL

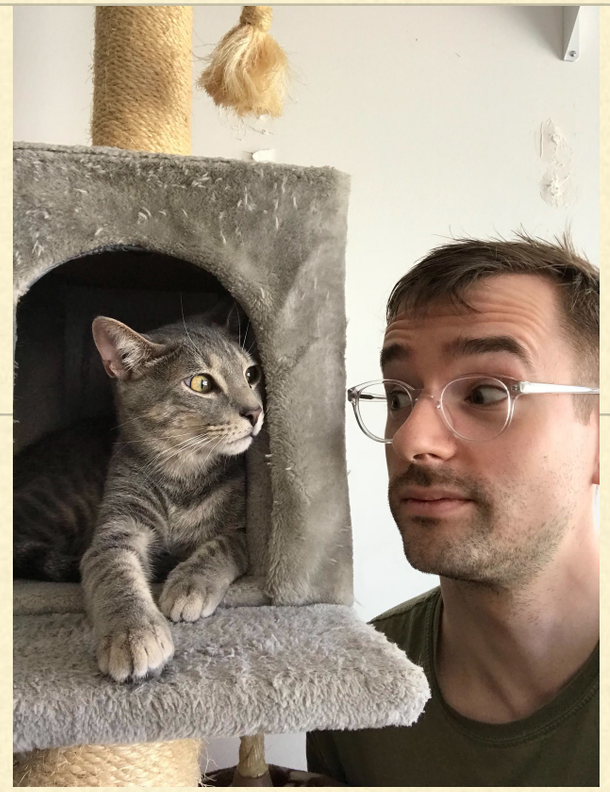


- We propose the following model:

Expected counts = Contribution from sample + Contributions from spurious sources

$$\mathbb{E}[W_{ij}] := \mu_{ij} = \mu_{ij}^{(s)} + \mu_{ij}^{(c)}$$

PROPOSED MODEL

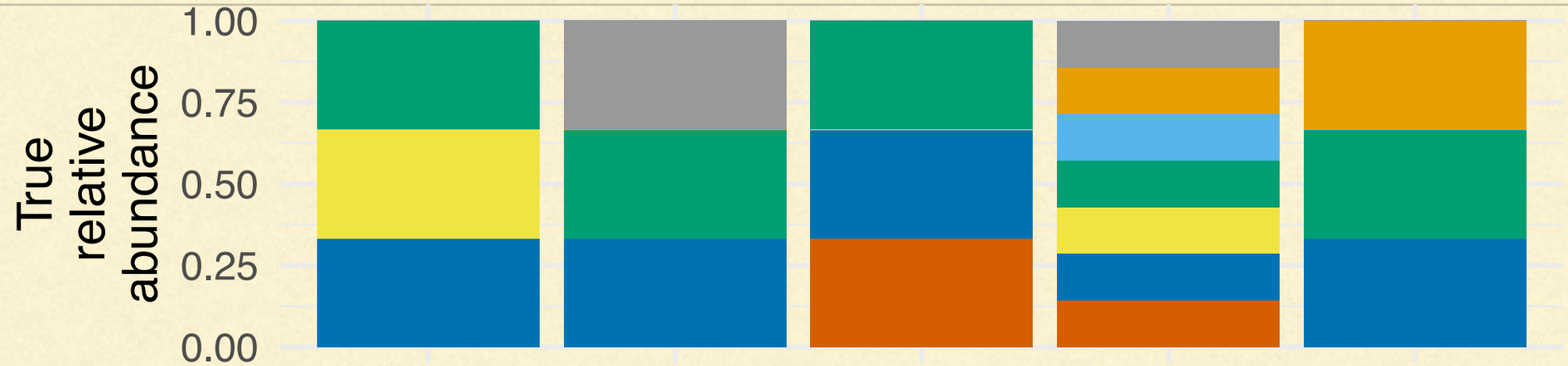


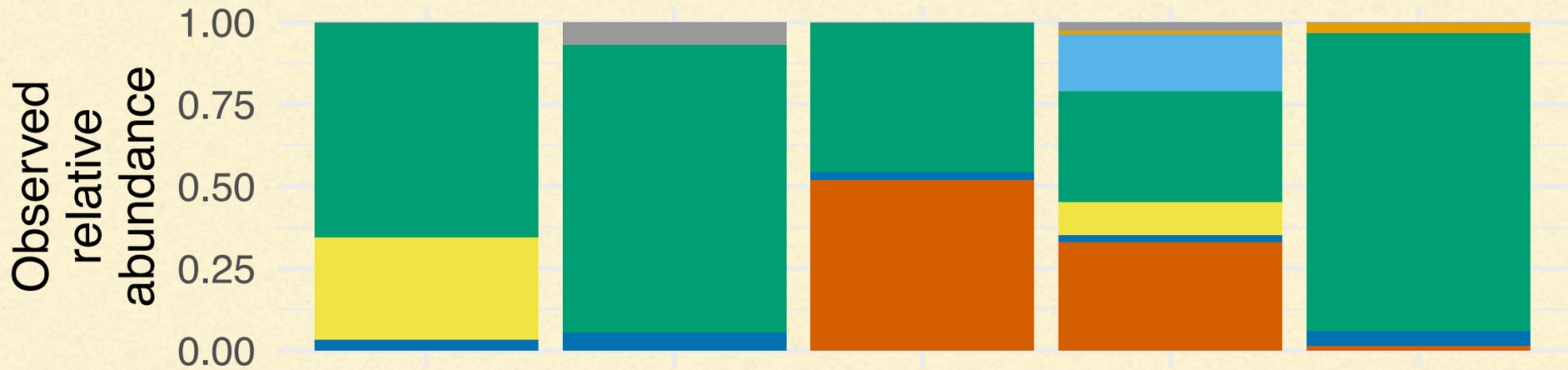
- We propose the following model:

Expected counts = **Contribution from sample** + Contributions from spurious sources

$$\mathbb{E}[W_{ij}] := \mu_{ij} = \mu_{ij}^{(s)} + \mu_{ij}^{(c)}$$

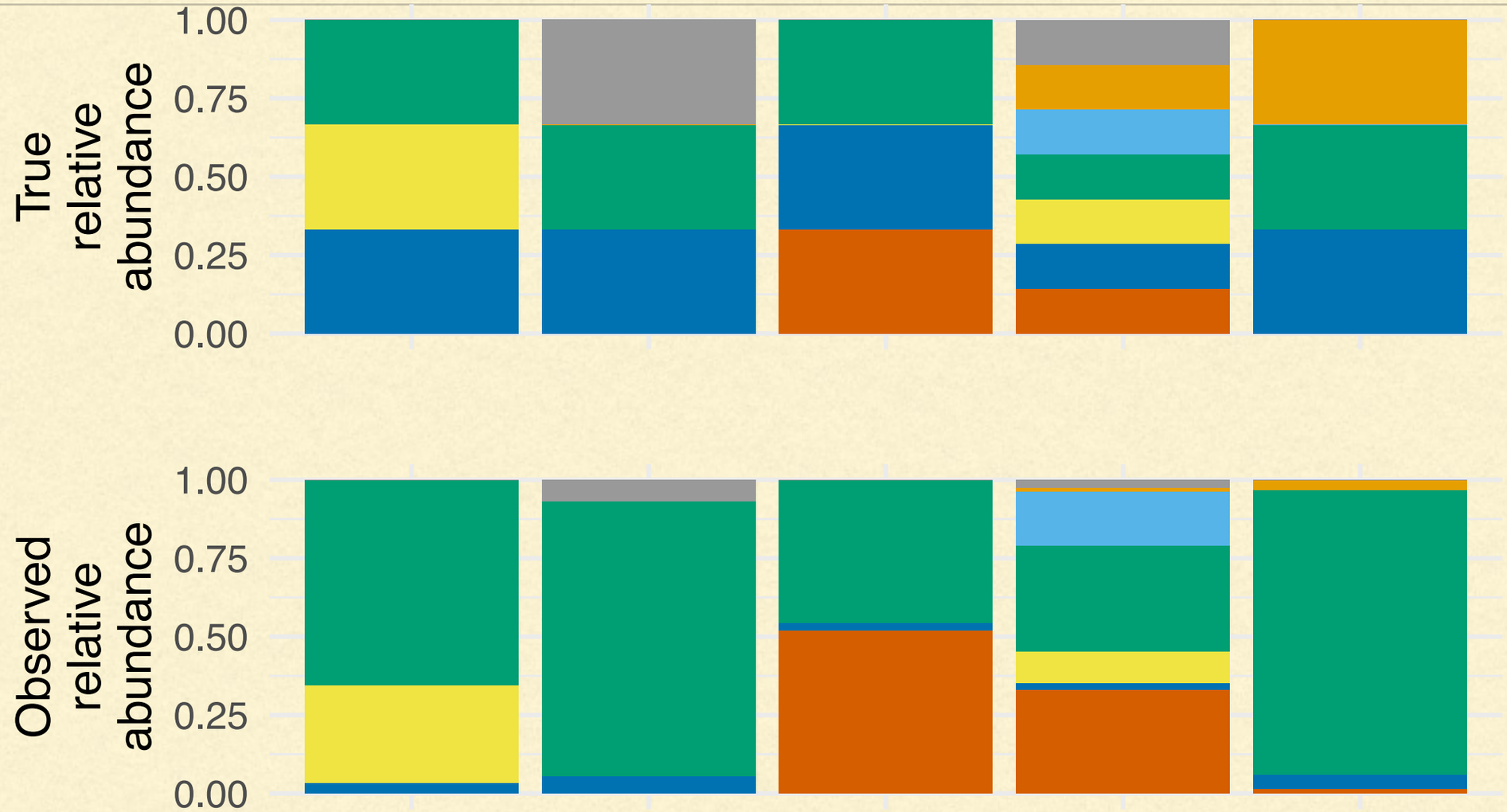
MODEL BUILDING





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

MODEL BUILDING



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

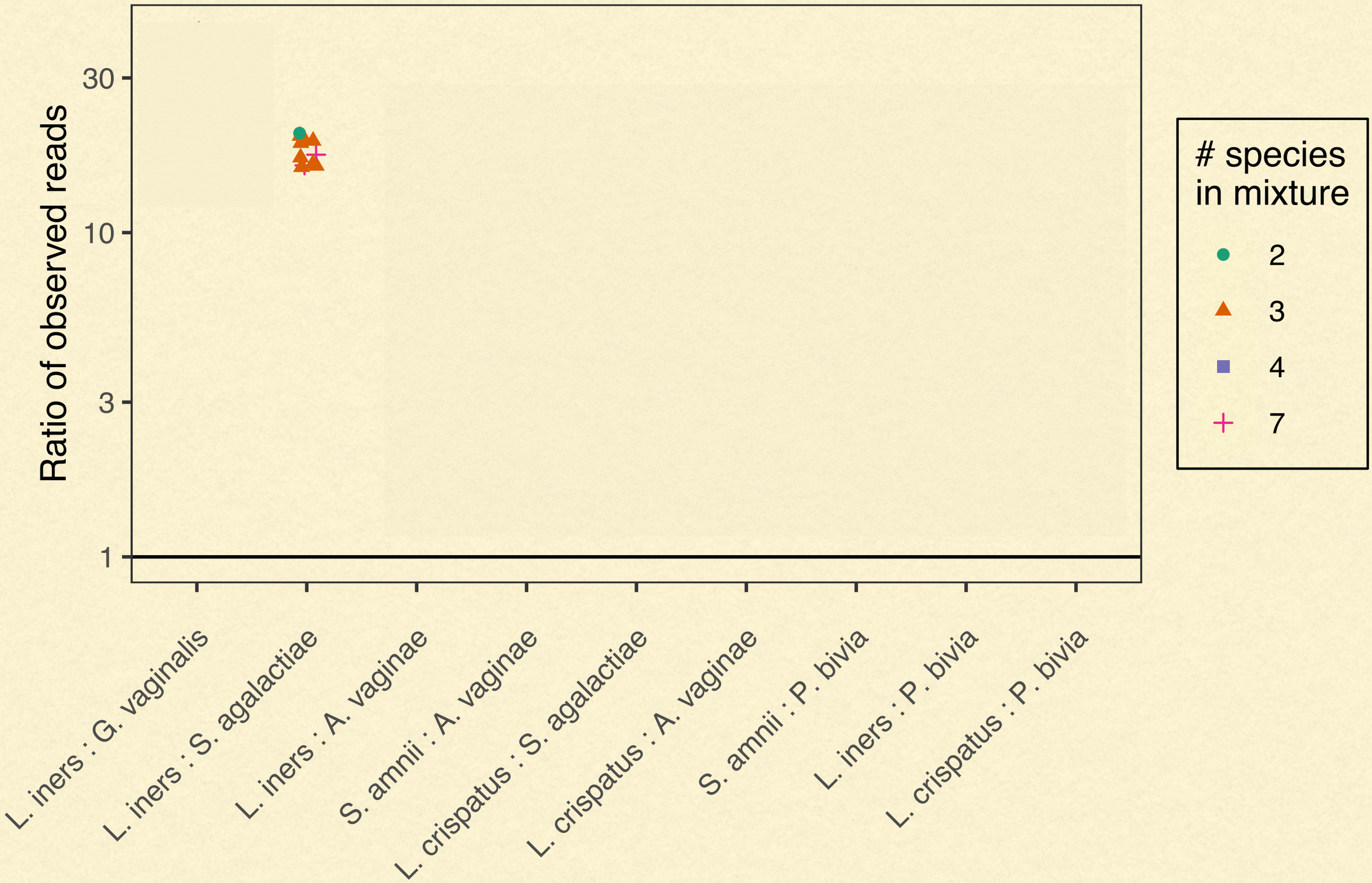


Figure adapted from McLaren, Willis & Callahan, 2019, *eLife*

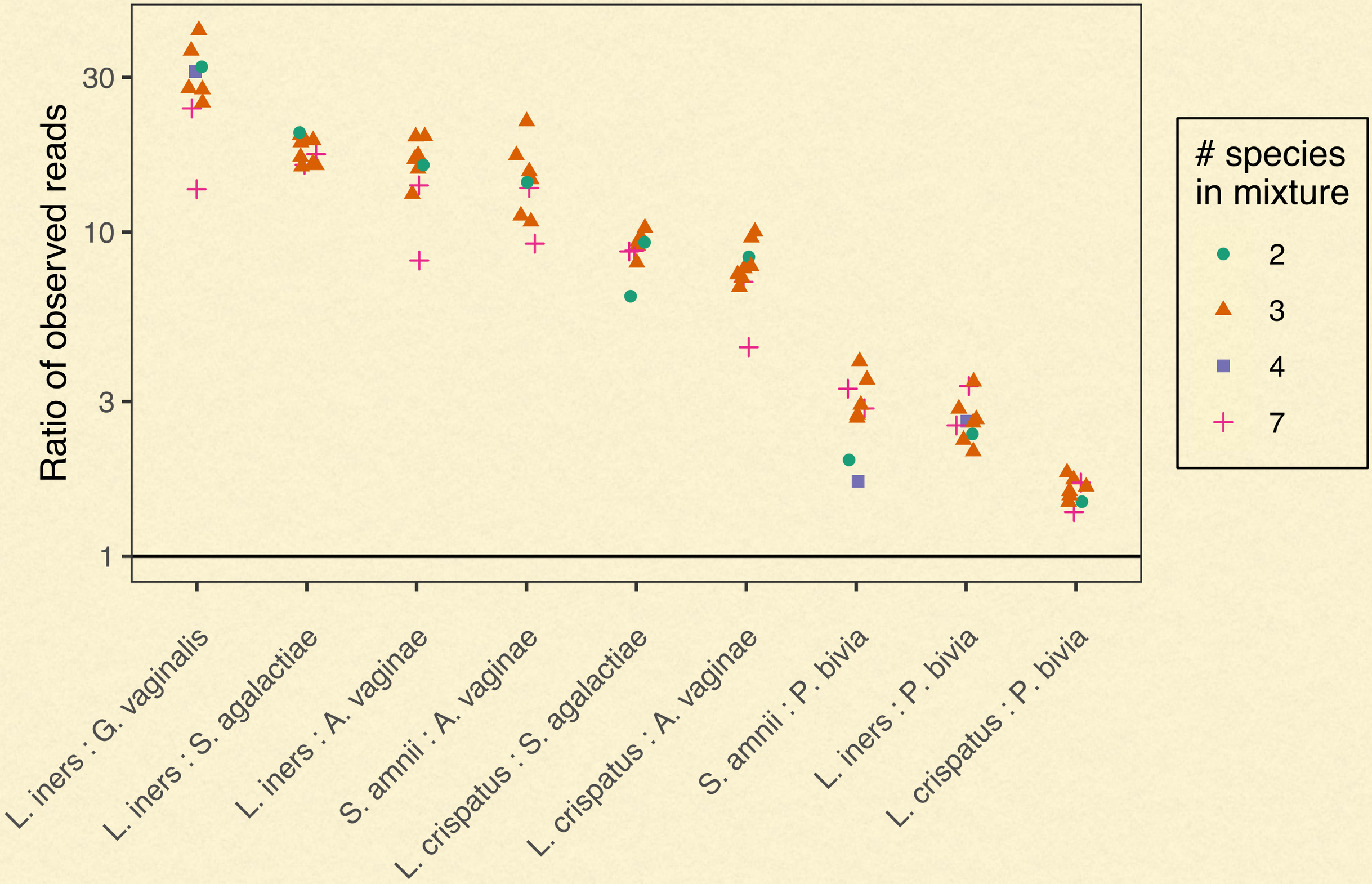


Figure adapted from McLaren, Willis & Callahan, 2019, *eLife*

MODEL SPECIFICATION

- Strong support *against*

$$\text{expected count}_{ij} = \text{true proportion}_{ij} \times \text{sampling intensity}_{i}$$

MODEL SPECIFICATION

- Strong support *against*

$$\text{expected count}_{ij} = \text{true proportion}_{ij} \times \text{sampling intensity}_{i}$$

- Better support for

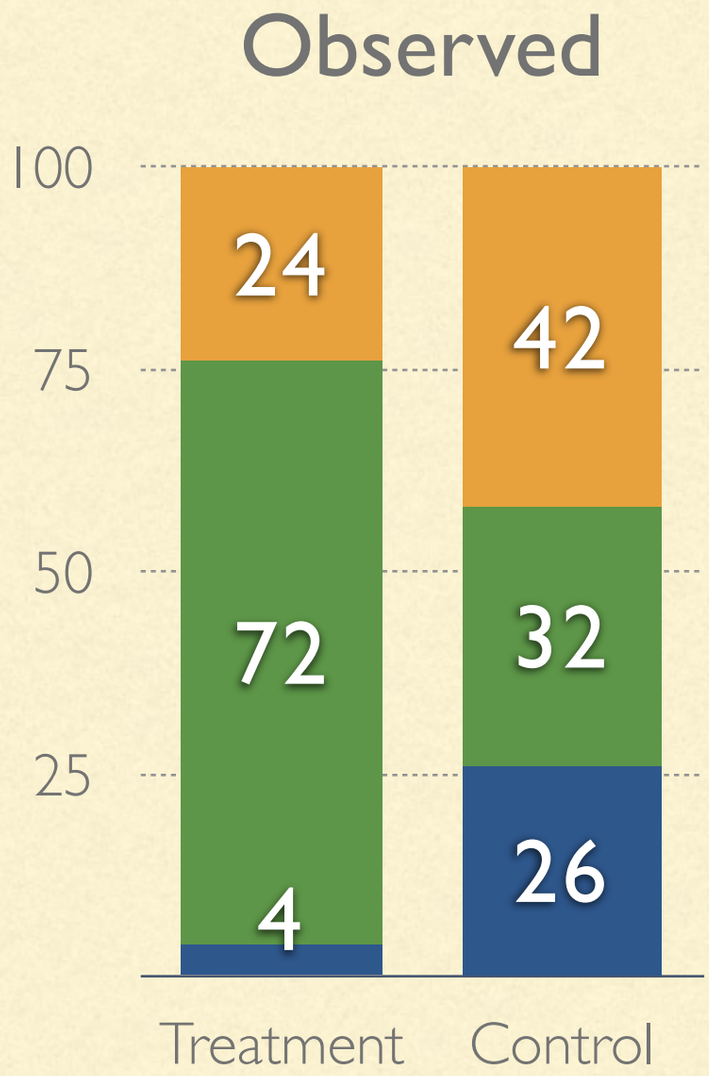
$$\text{expected count}_{ij} = \text{true proportion}_{ij} \times \text{sampling intensity}_{i} \times \text{detection efficiency}_{j}$$

MODEL SPECIFICATION

- Stated differently,

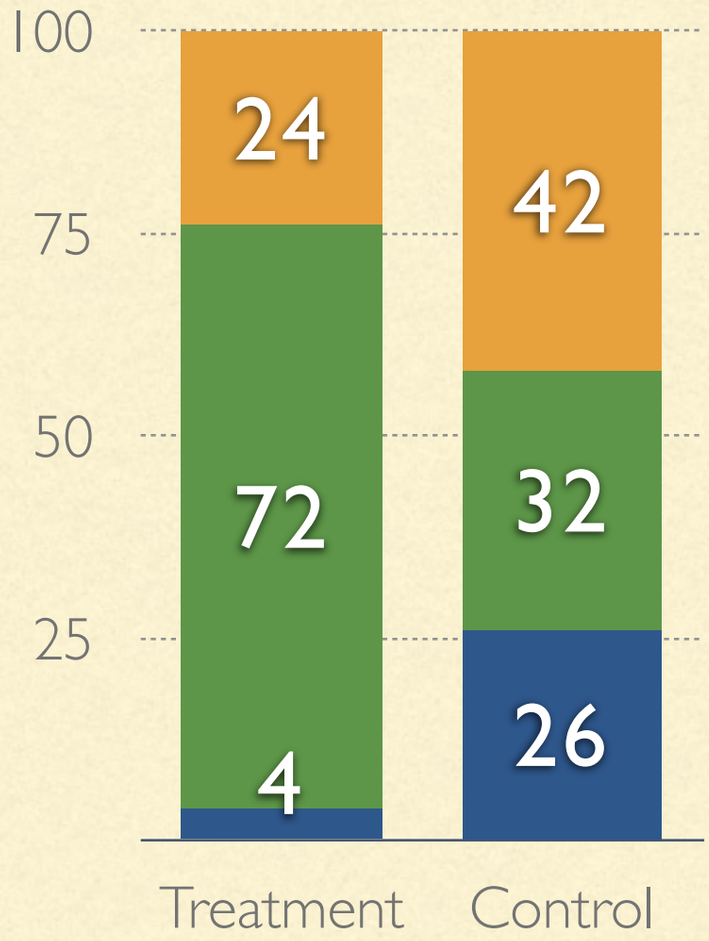
Observed relative abundance $\propto$ True relative abundance $\times$ Taxon-specific efficiencies

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

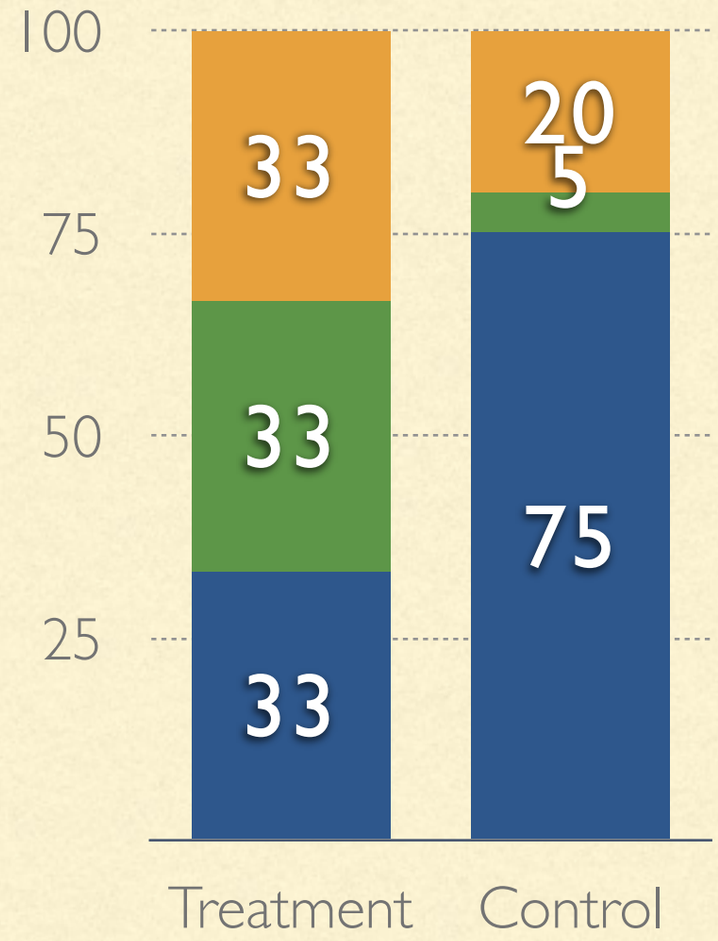


- A tempting conclusion:
  - The relative abundance of **taxon orange** *decreased* in the Treatment sample (left) compared to the Control sample (right)

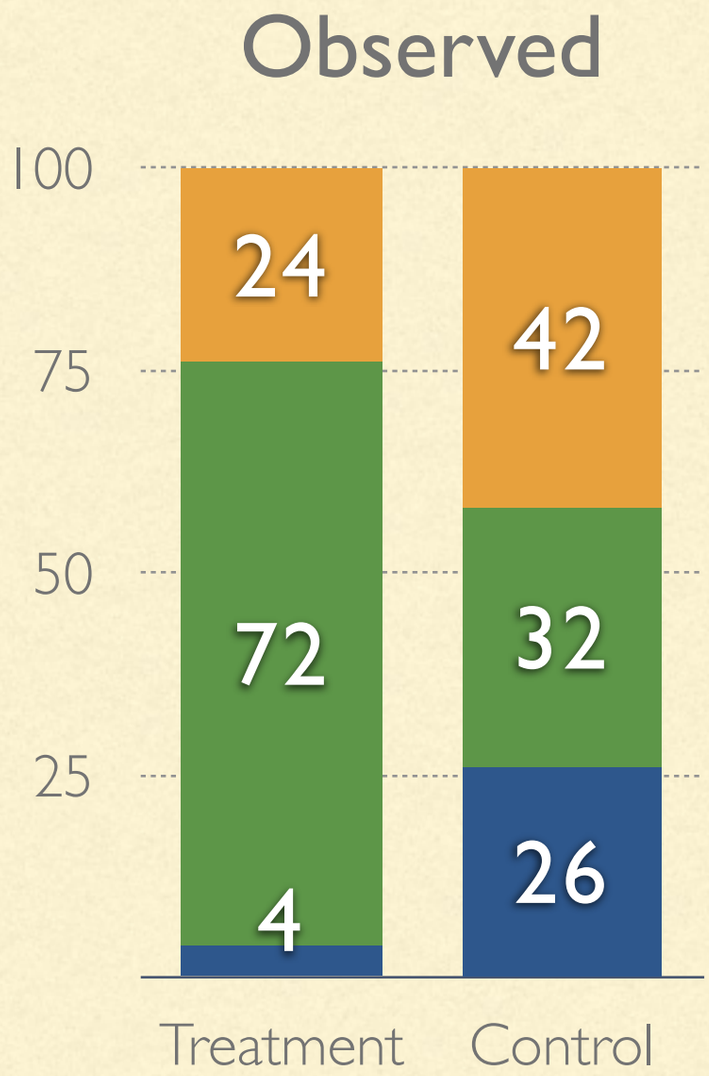
Observed

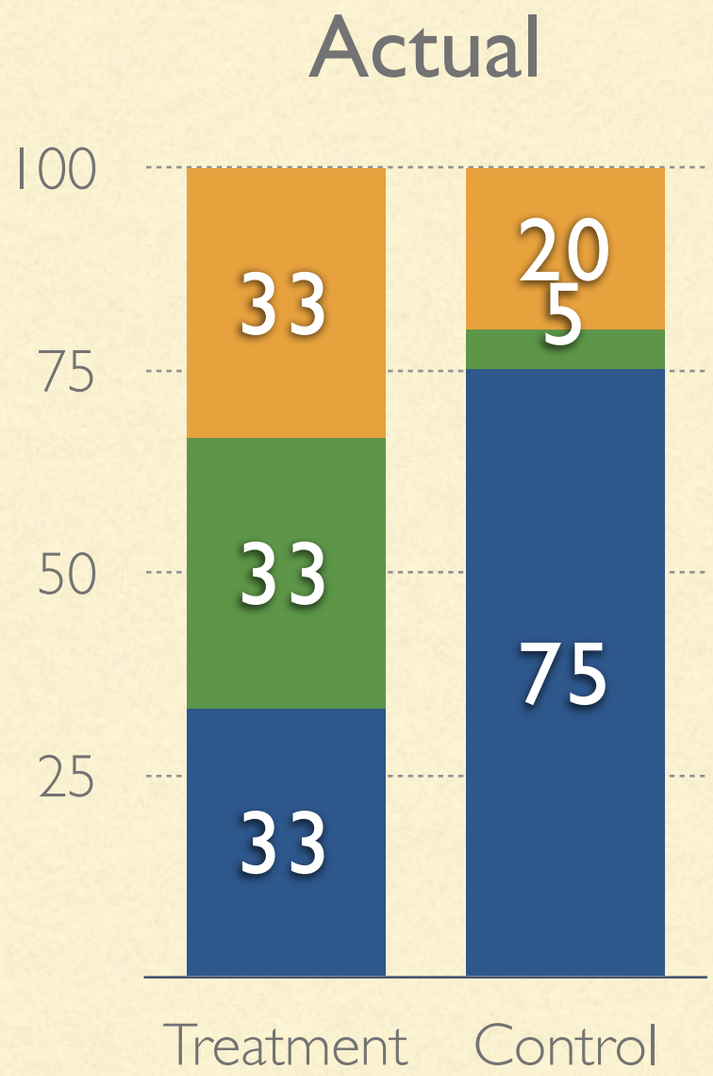


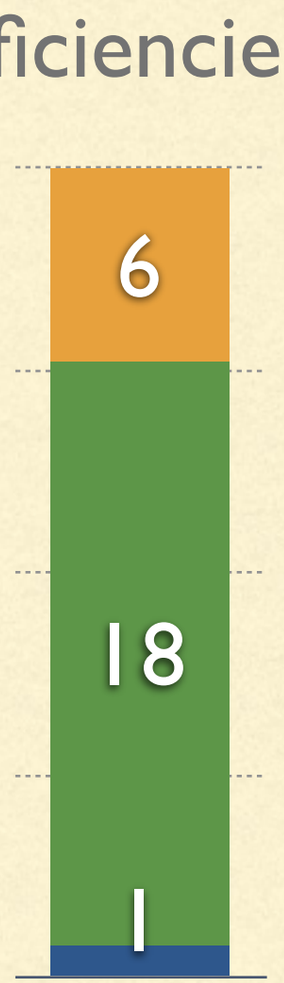
Actual



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

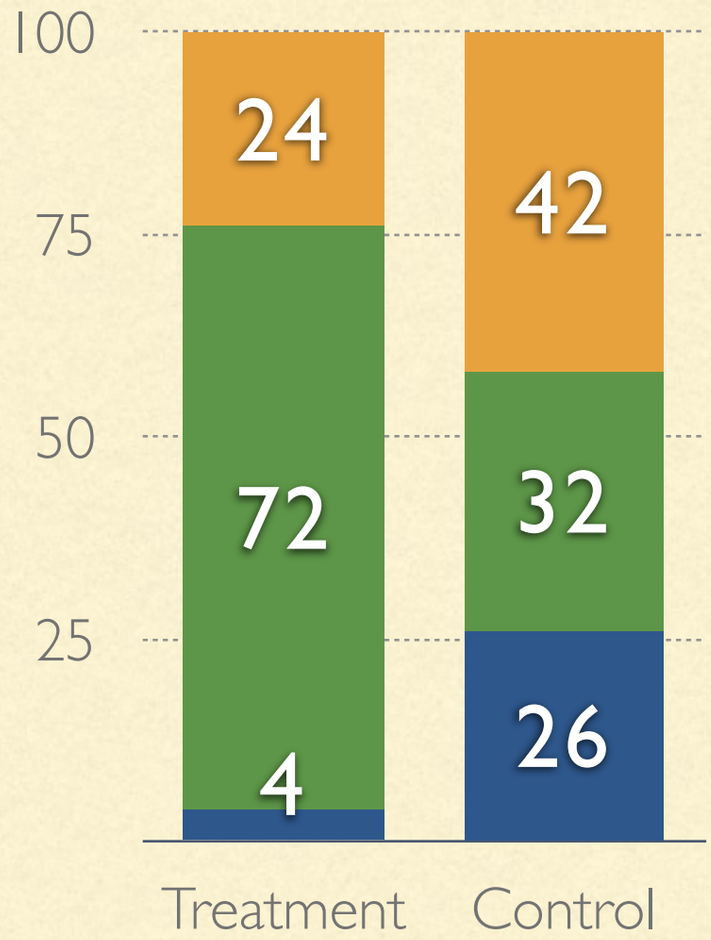




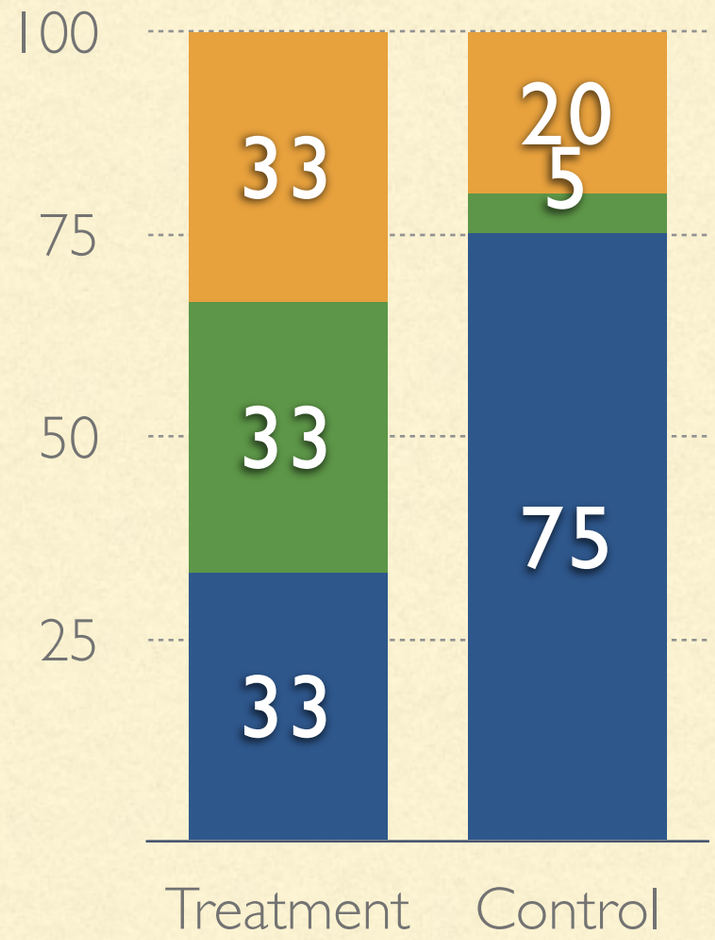


- In fact, the relative abundance of **taxon orange** *increased* in the Treatment sample compared to the Control sample

Observed



Actual



Efficiencies

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

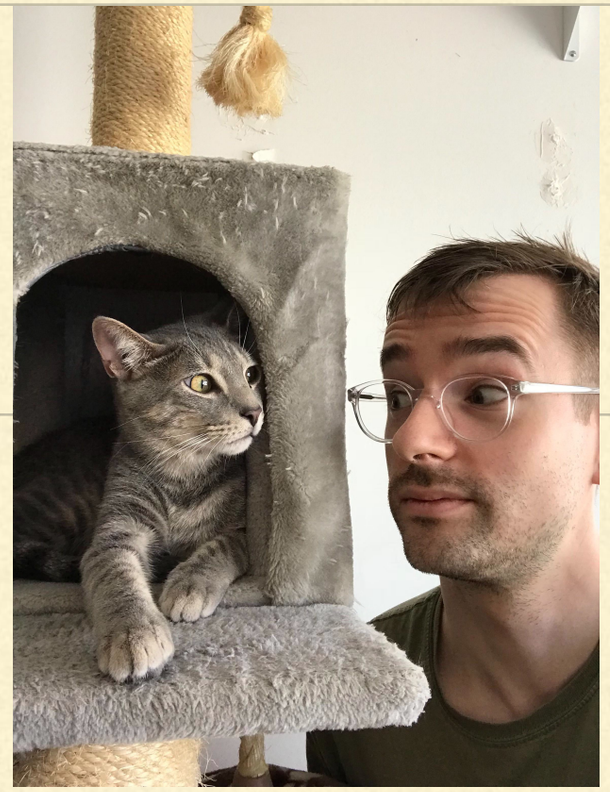
- **Taxon orange**'s abundance depends on the abundance of **the other taxa**

PROPOSED MODEL

$$\text{Contribution from sample} = \text{True relative abundance} \times \text{Sample-specific "intensity"} \times \text{Taxon-specific "efficiency"}$$

	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

PROPOSED MODEL

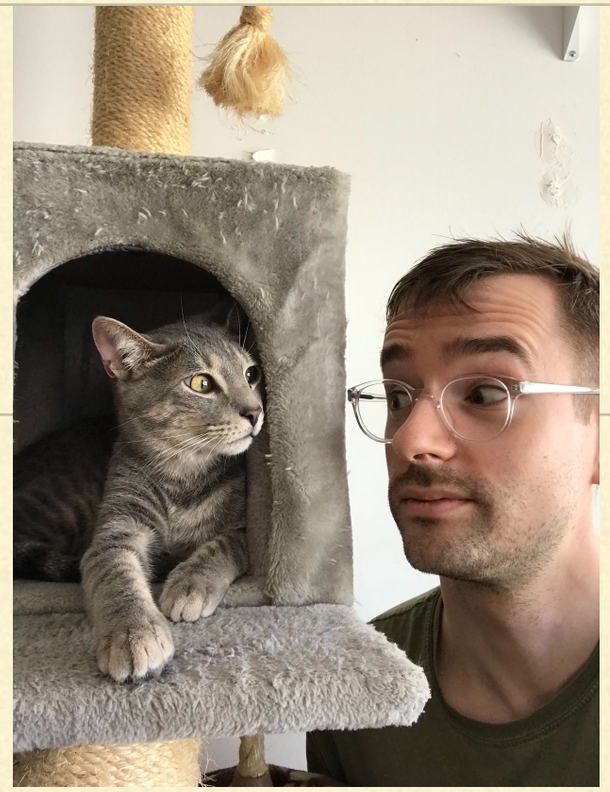


- We propose the following model:

Expected counts = **Contribution from sample** + Contributions from spurious sources

$$\mathbb{E}[W_{ij}] := \mu_{ij} = \mu_{ij}^{(s)} + \mu_{ij}^{(c)}$$

PROPOSED MODEL



- We propose the following model:

Expected counts = Contribution from sample + **Contributions from spurious sources**

$$\mathbb{E}[W_{ij}] := \mu_{ij} = \mu_{ij}^{(s)} + \mu_{ij}^{(c)}$$

PROPOSED MODEL

Contribution from spurious source $\tilde{k}$ to sample i = Sample i intensity × Source $\tilde{k}$ relative abundance × Source $\tilde{k}$ "intensity"

$$\mu_{ij}^{(c)} = \exp(\gamma_i) \times \sum_{\tilde{k}=1}^{\tilde{K}} \tilde{p}_{\tilde{k}j} \exp(\tilde{\gamma}_{\tilde{k}})$$

PROPOSED MODEL

$$\mathbb{E}[W_{ij}] = p_{ij} \times \exp(\gamma_i + \beta_j) \quad + \quad \exp(\gamma_i) \times \sum_{\tilde{k}=1}^{\tilde{K}} \tilde{p}_{\tilde{k}j} \exp(\tilde{\gamma}_{\tilde{k}})$$

PROPOSED MODEL

$$\mathbb{E}[W_{ij}] = p_{ij} \times \exp(\gamma_i + \beta_j) \quad + \quad \exp(\gamma_i) \times \sum_{\tilde{k}=1}^{\tilde{K}} \tilde{p}_{\tilde{k}j} \exp(\tilde{\gamma}_{\tilde{k}})$$

$$= \Big[\underbrace{\mathbf{D}_\gamma \left(\mathbf{Zp}\right) \circ \exp\left(\mathbf{X}\boldsymbol{\beta}\right)}_{\text{contribution of sample}} + \underbrace{\mathbf{D}_\gamma \tilde{\mathbf{Z}} \left[\tilde{\mathbf{p}} \circ \exp\left(\tilde{\boldsymbol{\gamma}} \mathbf{1}_J^T\right)\right]}_{\text{contribution of spurious sources}} \Big]_{ij}$$

- $\mathbf{Z} \in \mathbb{R}^{n \times K}$ links sample i to specimen k
- $\tilde{\mathbf{Z}} \in \mathbb{R}^{n \times \tilde{K}}$ links sample i to spurious source $\tilde{k}$
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ links sample i to batch p

arXiv > stat > arXiv:2204.12733

Statistics > Methodology

Modeling complex measurement error in microbiome experiments

David S Clausen, Amy D Willis

David Clausen (UW)

Ben Callahan (NCSU)

DOI: 10.1111/biom.13503

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

A multiview model for relative and absolute microbial abundances

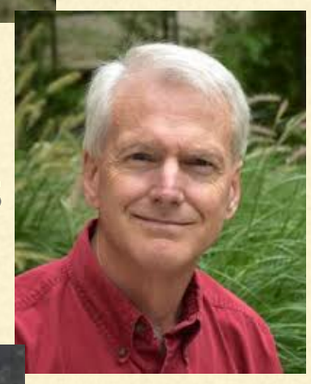
Brian D. Williamson | James P. Hughes | Amy D. Willis

Michael McLaren (MIT)

eLIFE elifesciences.org

Consistent and correctable bias in metagenomic sequencing experiments

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



Jim Hughes (UW)

Evaluating replicability in microbiome data

David S Clausen, Amy D Willis

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

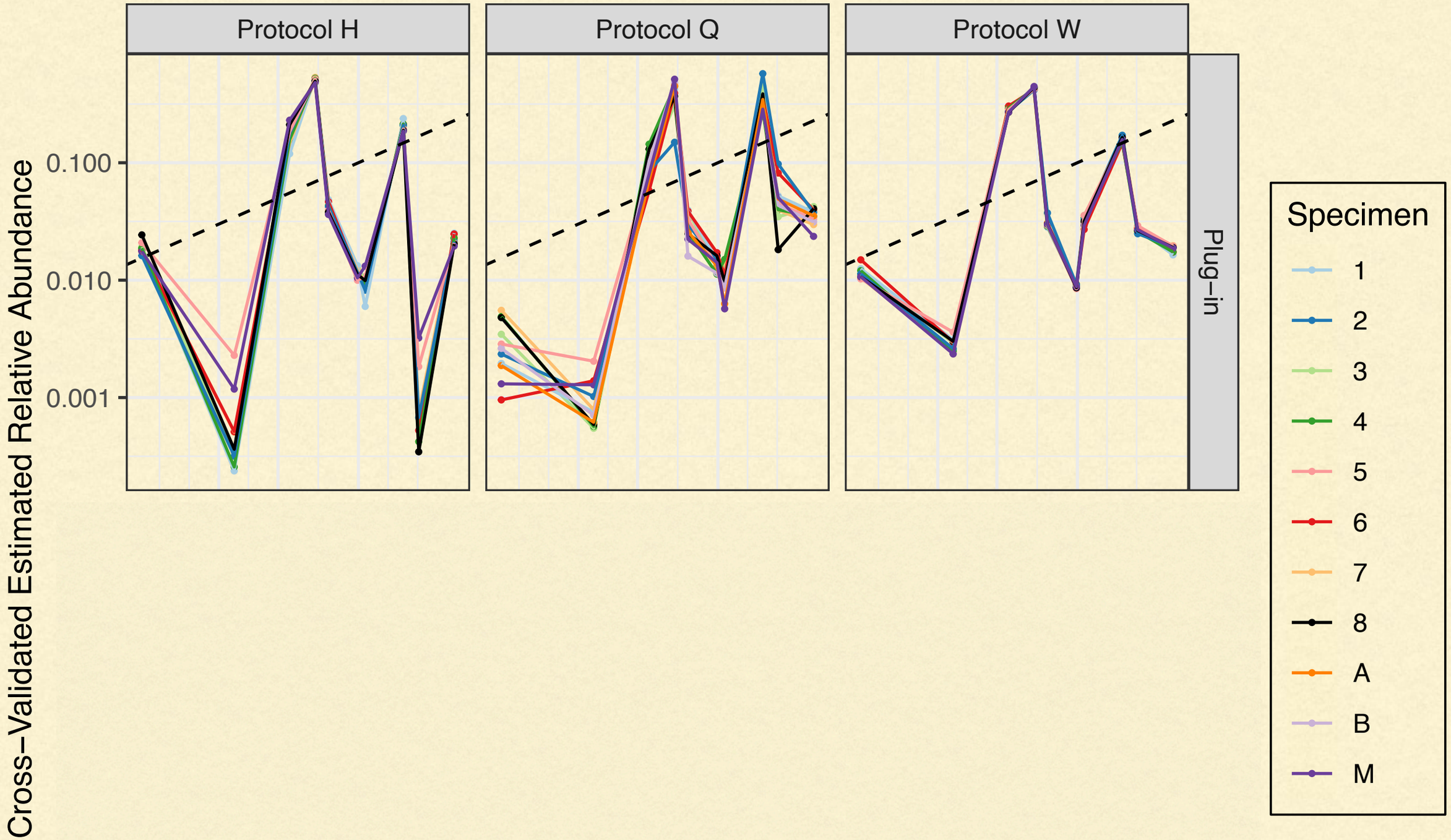
Brian Williamson (Kaiser)

- Details on estimation and inference
- Data examples
  - Estimating detection efficiencies with mock communities
  - **Comparing across experimental protocols (minimizing variance)**
  - **Removing contamination via dilution series**

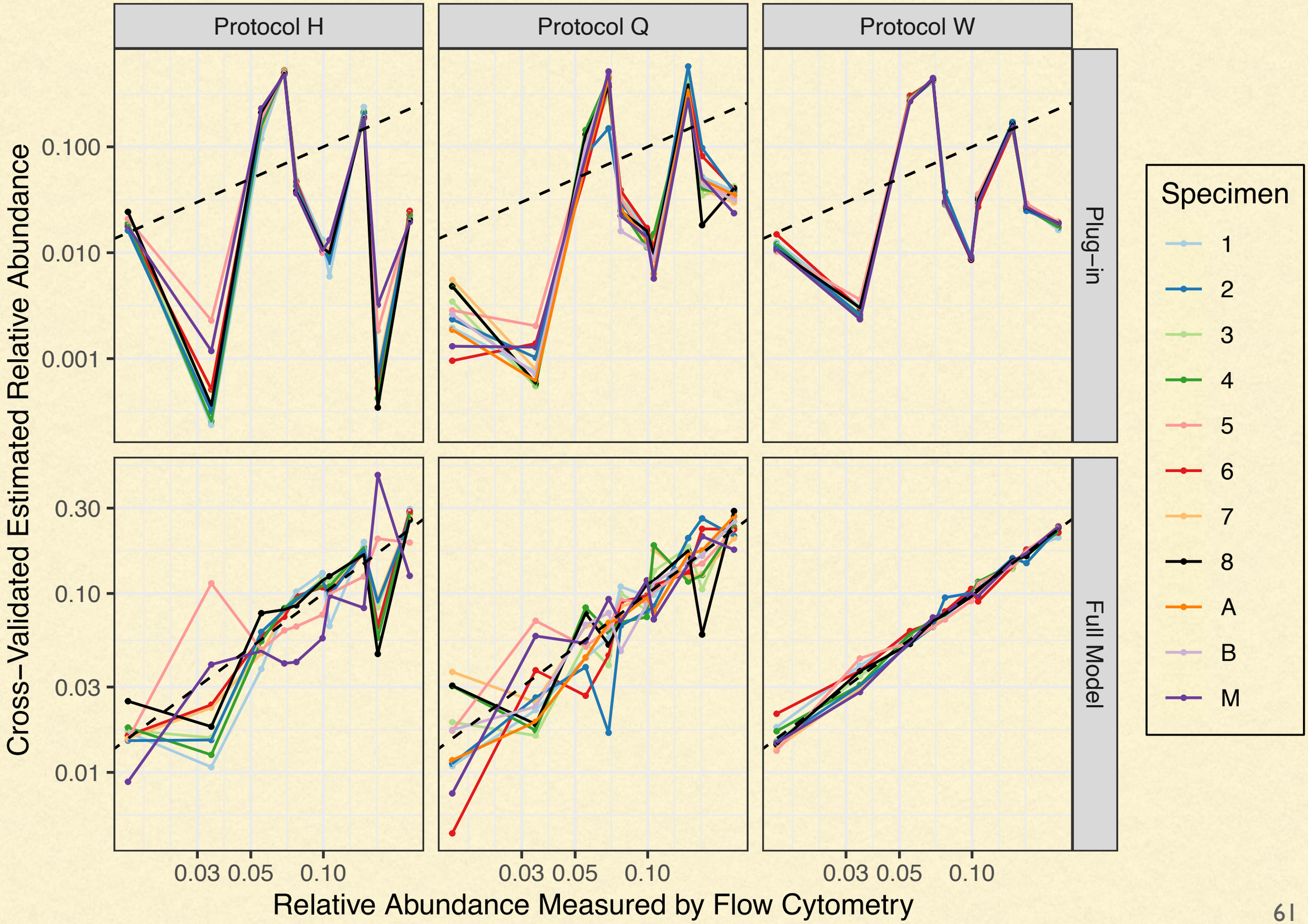
TINYVAMP IN ACTION: CHOOSING BETWEEN PROTOCOLS

- Goal: Choose between 3 different experimental protocols
  - 10 samples mixed with synthetic community
  - Sequencing + flow cytometry data
- Original comparison only allowed comparison of bias

Data from Costea et al, 2017, *Nature Biotech.*



Relative Abundance Measured by Flow Cytometry


Protocol H
Protocol Q
Protocol W
Plug–in
Full Model
Cross–Validated Estimated Relative Abundance
0.100
0.010
0.001
0.30
0.10
0.03
0.01
0.03 0.05 0.10
Relative Abundance Measured by Flow Cytometry
Specimen
1
2
3
4
5
6
7
8
A
B
M

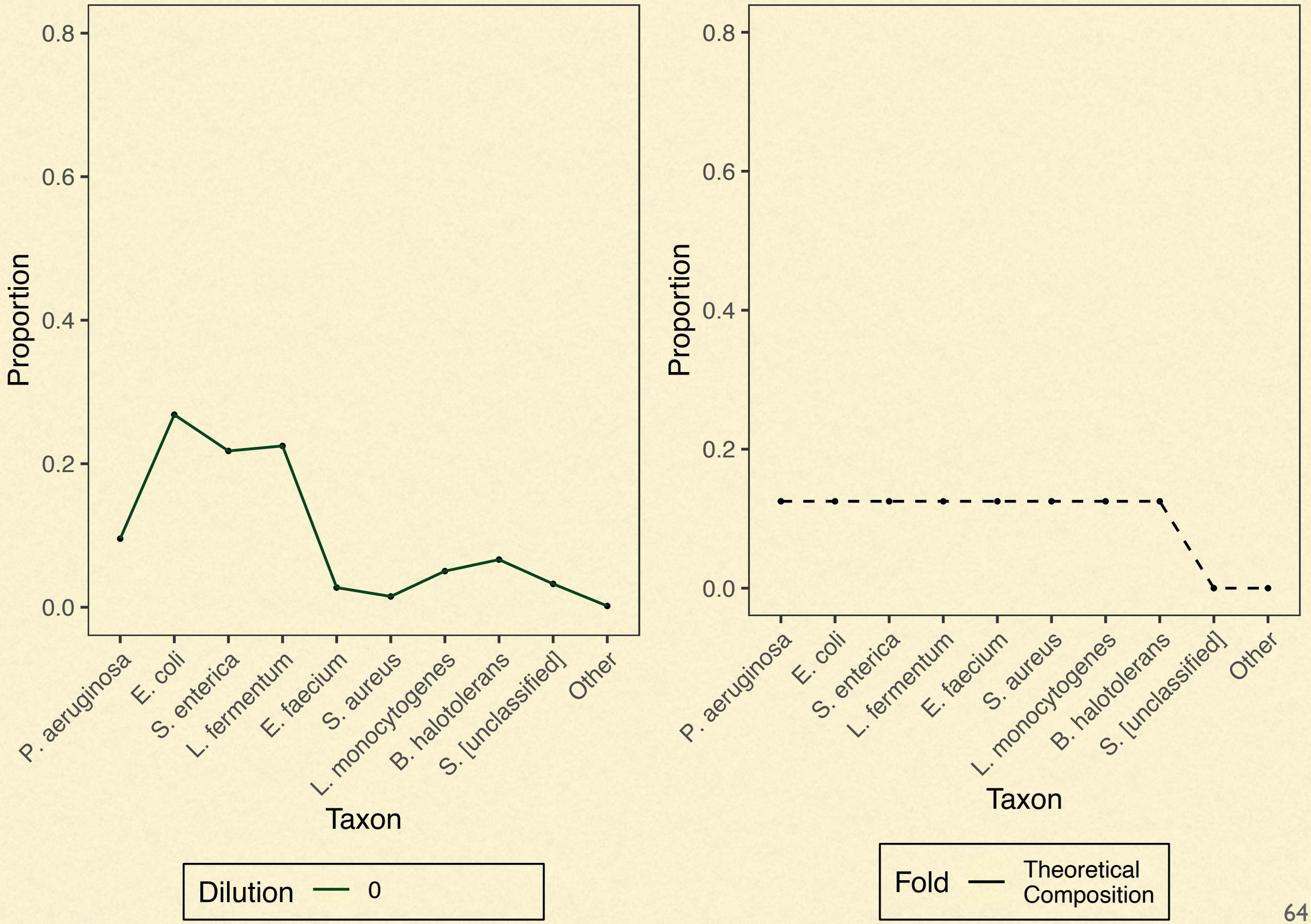
TINYVAMP IN ACTION: REMOVING CONTAMINATION

- Current tools for removing contamination are *coarse*
  - "Contaminant" taxa are removed from *all* samples
- Scalable solution: dilution series
  - Idea: More diluted samples have greater proportion of contamination
  - Proposal: Remove estimated contamination profile

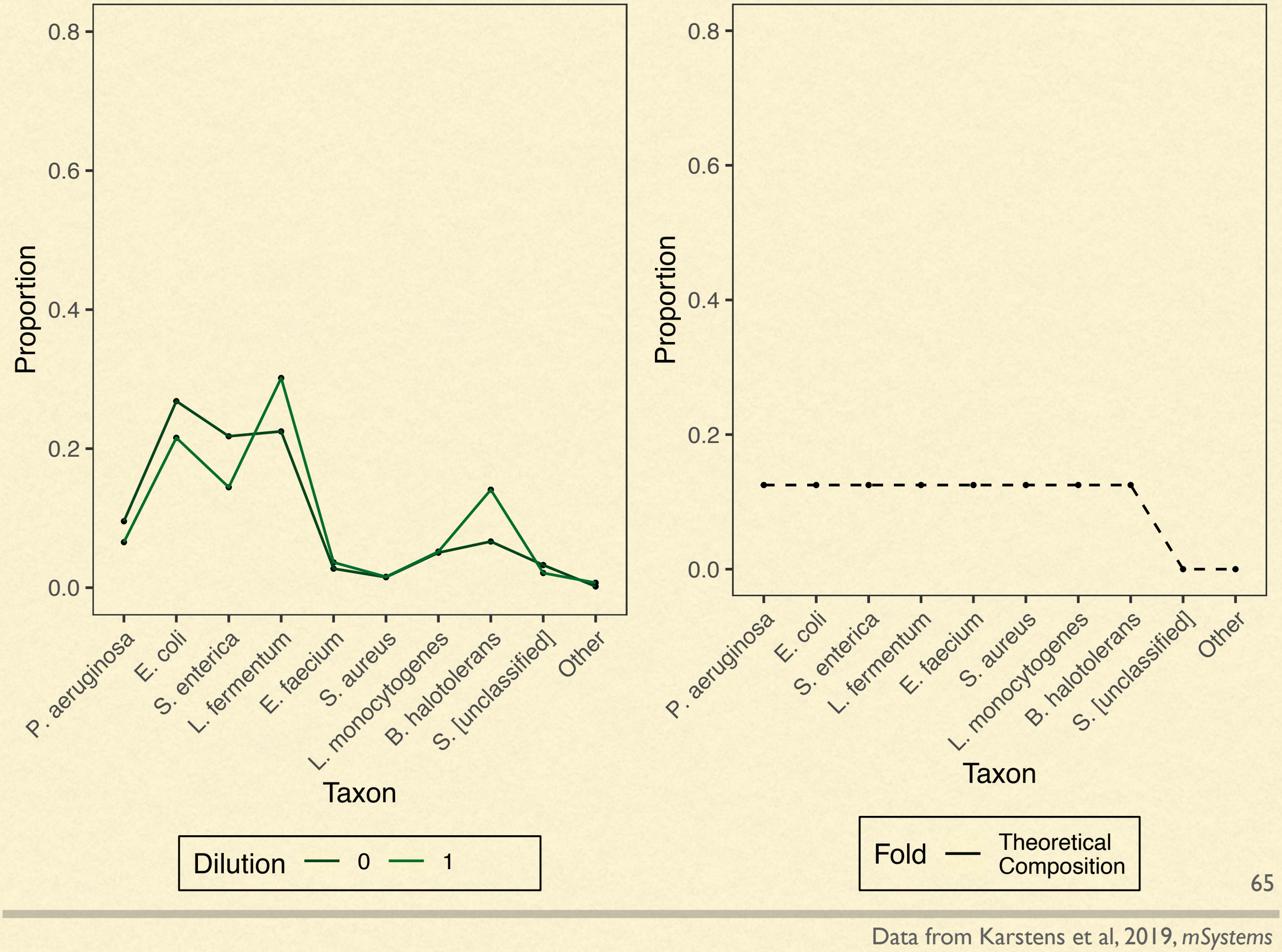
TINYVAMP IN ACTION: REMOVING CONTAMINATION

- Synthetic community: 8 bacterial strains, each 12.5%
  - Undiluted + sequence of 8 three-fold dilutions
- 248 total strains observed
- Model can account for both detection effects and contamination

Data from Karstens et al, 2019, *mSystems*

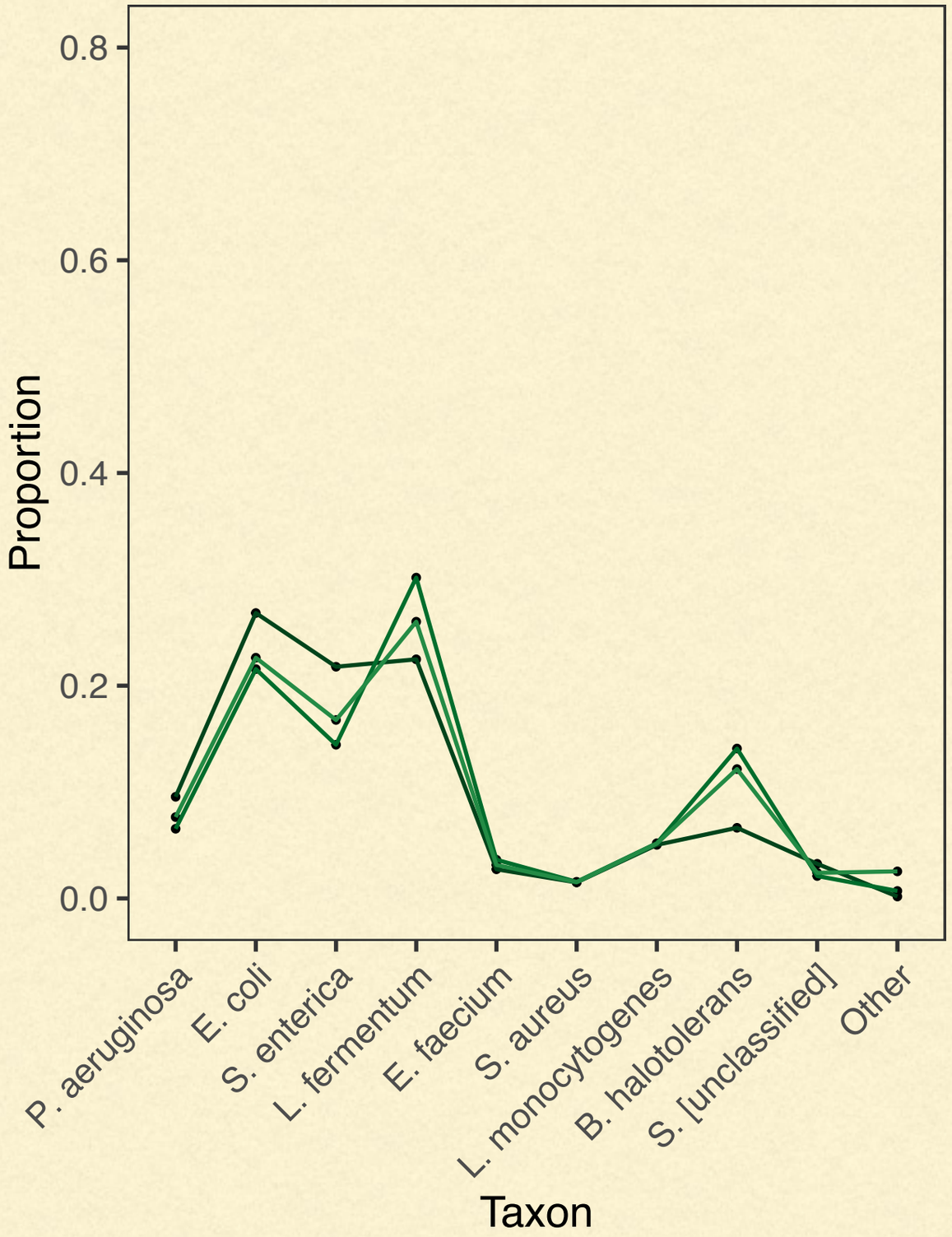


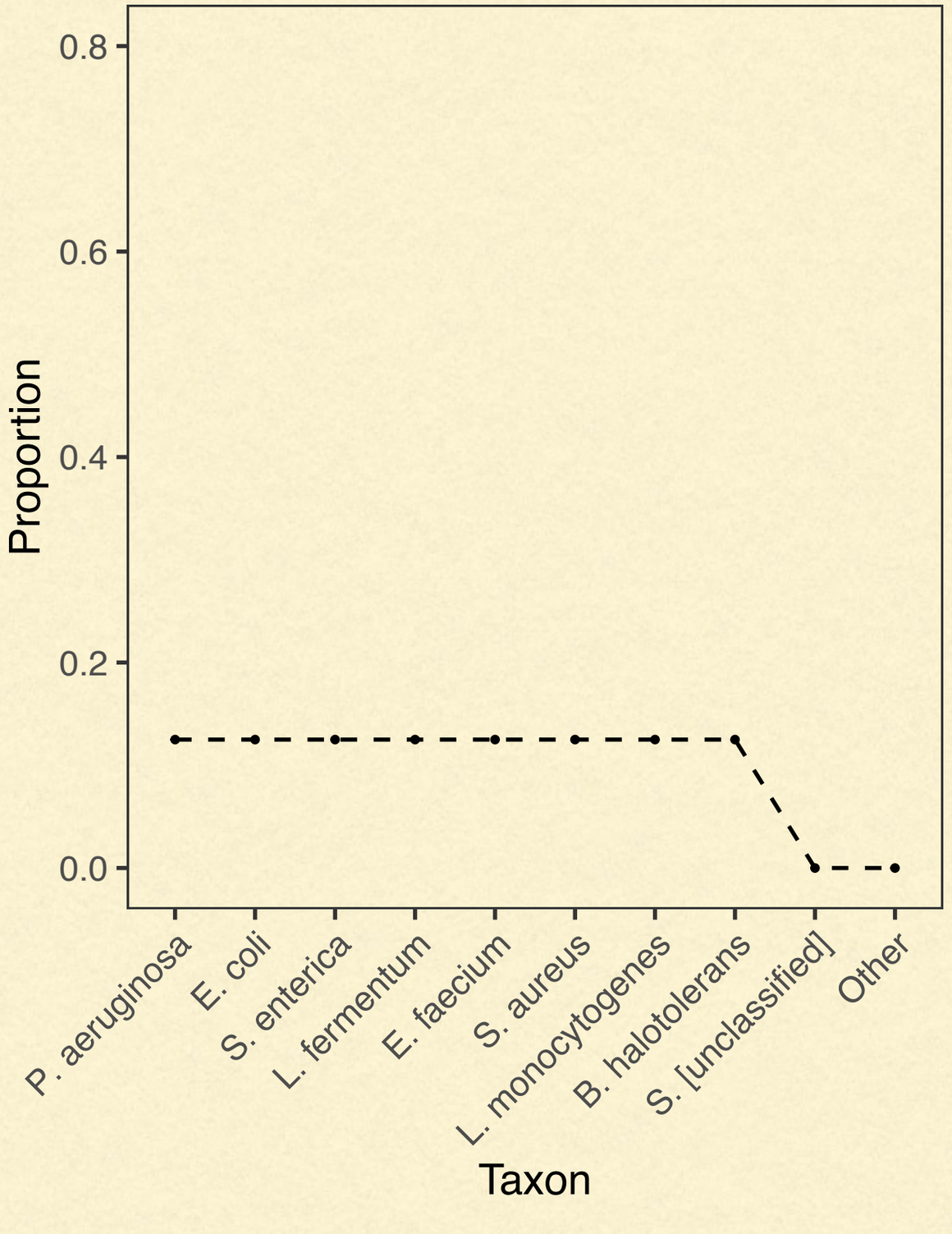
Data from Karstens et al, 2019, *mSystems*

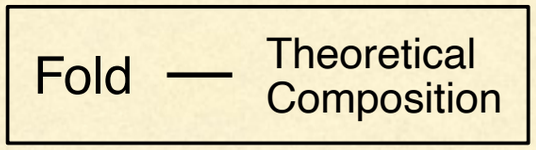


65

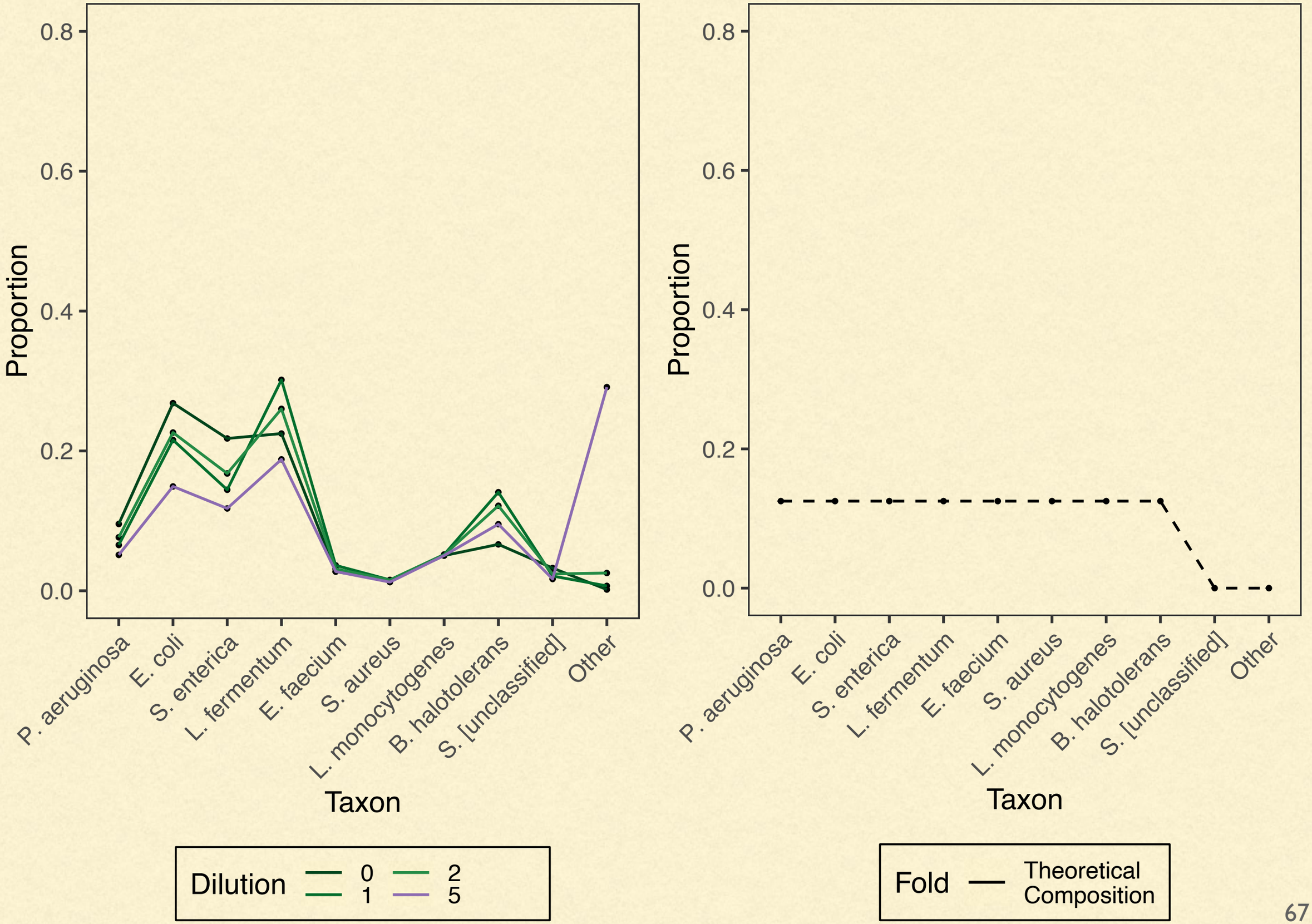
Data from Karstens et al, 2019, *mSystems*



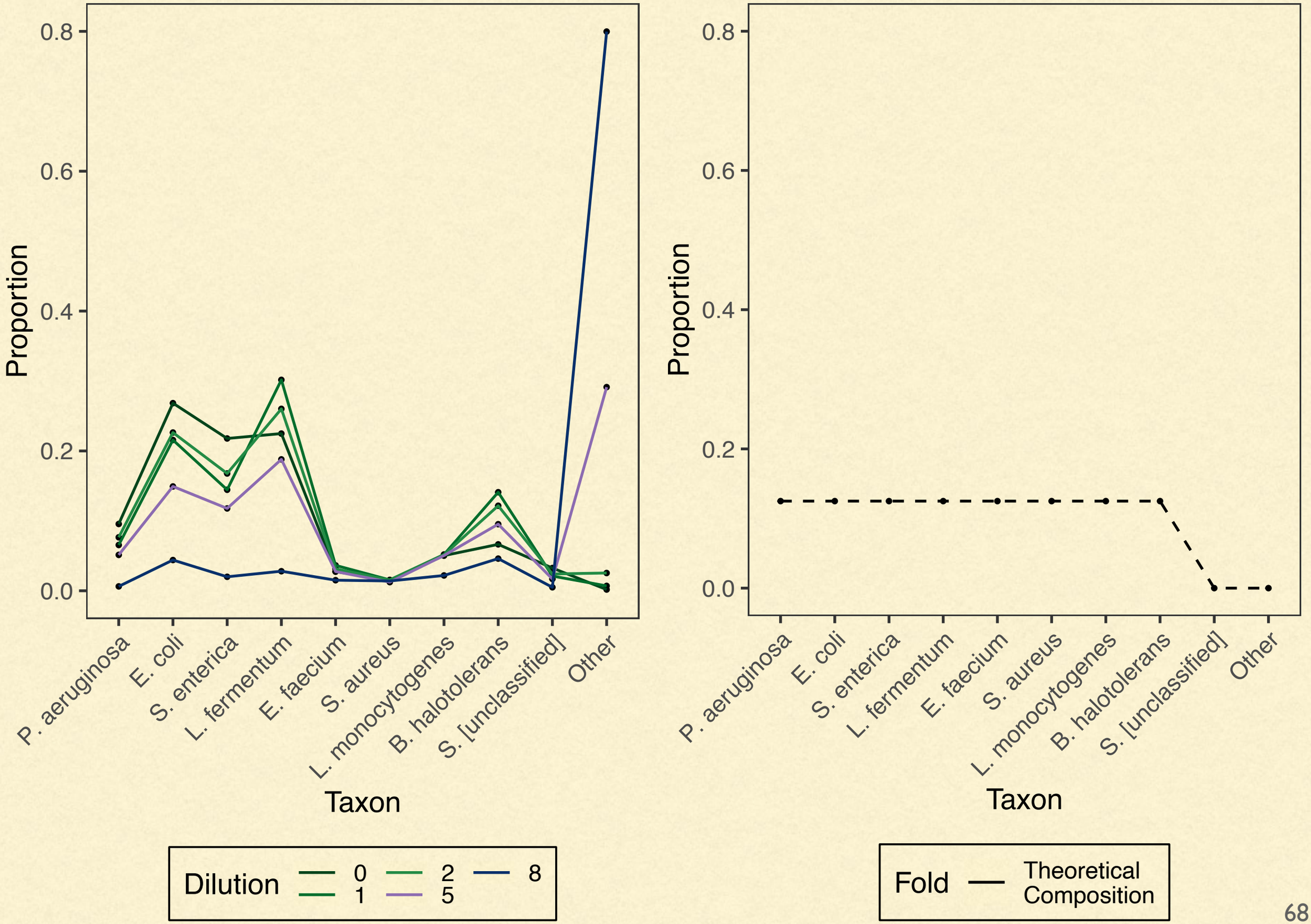




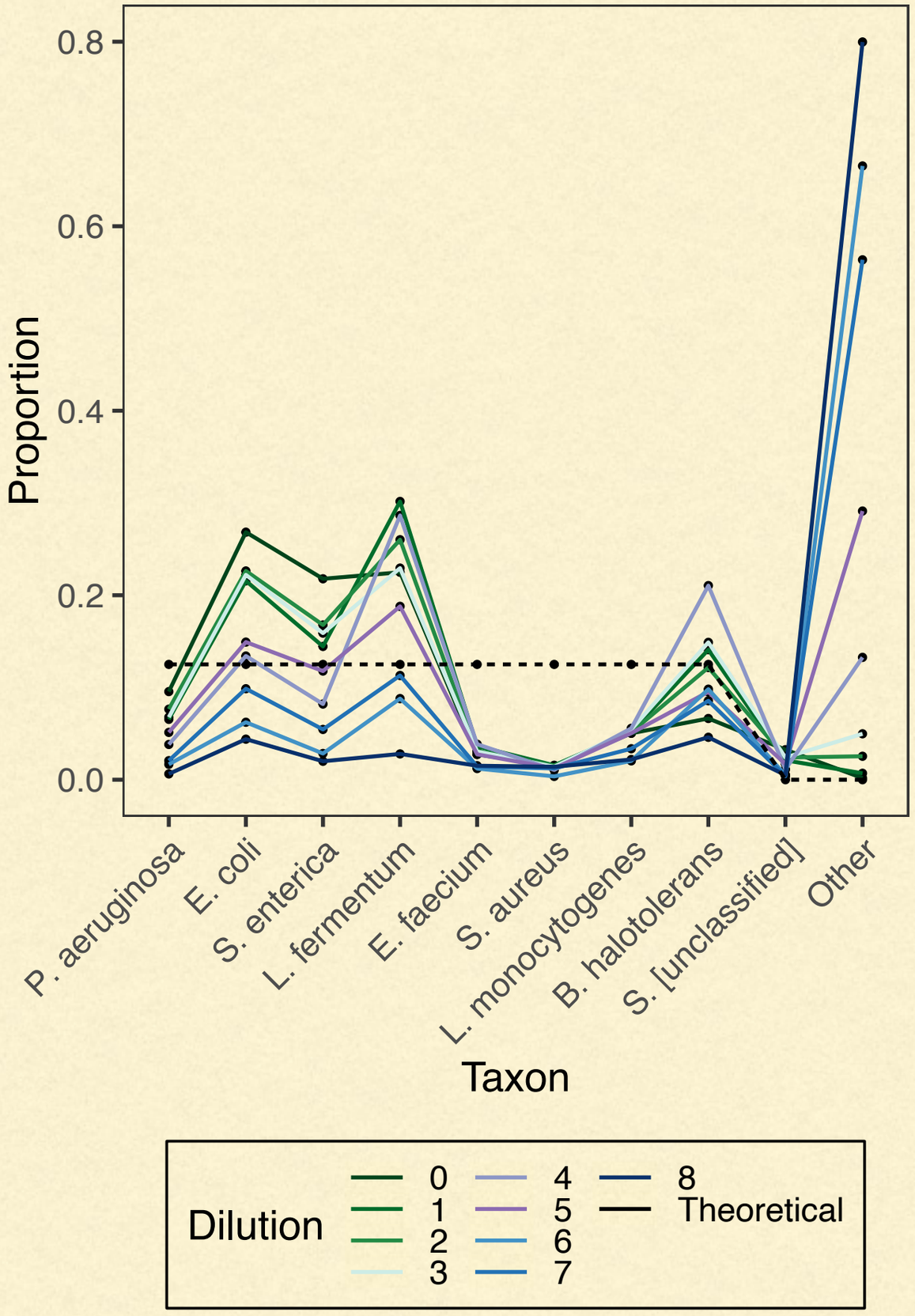
Data from Karstens et al, 2019, *mSystems*

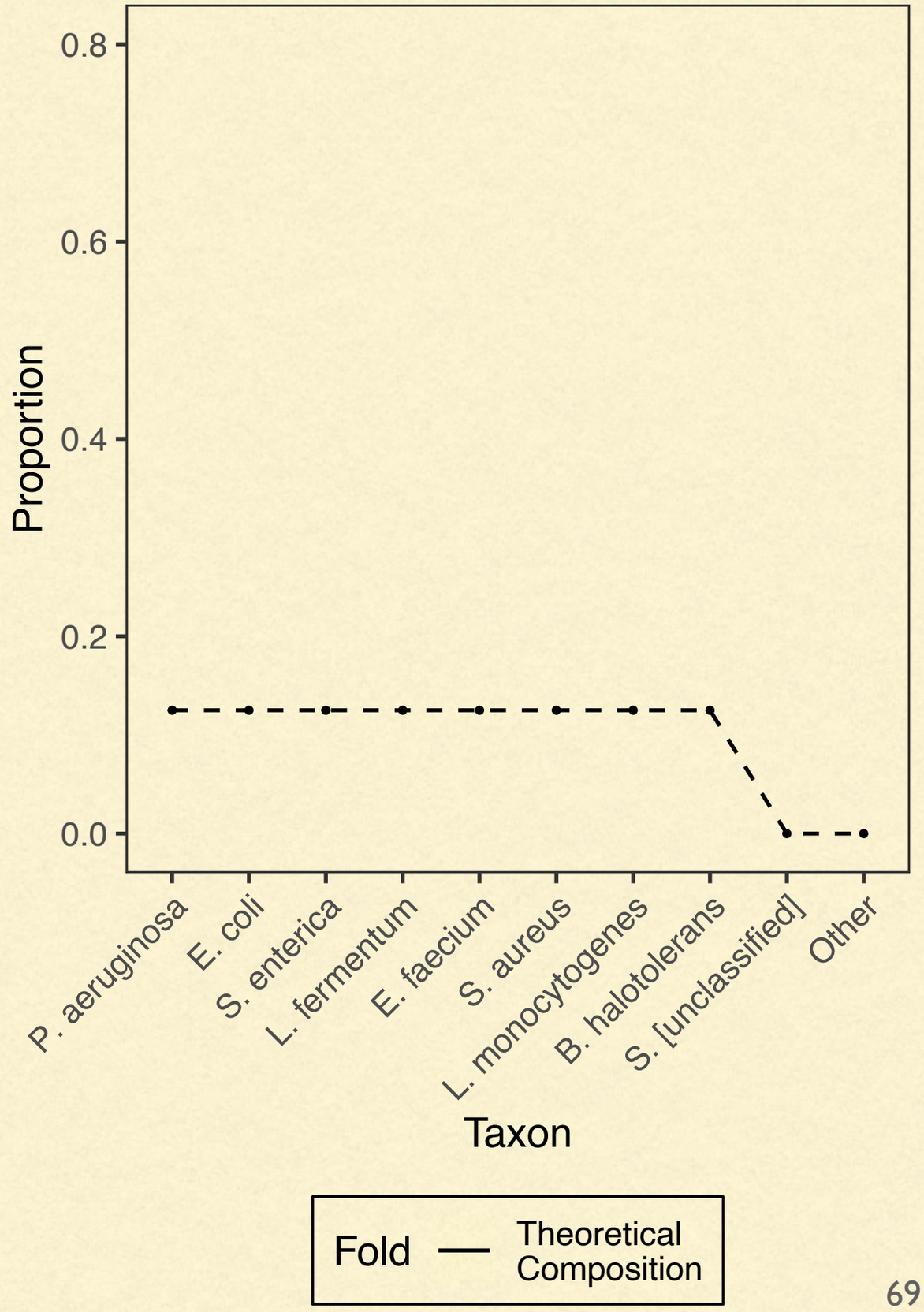


Data from Karstens et al, 2019, *mSystems*

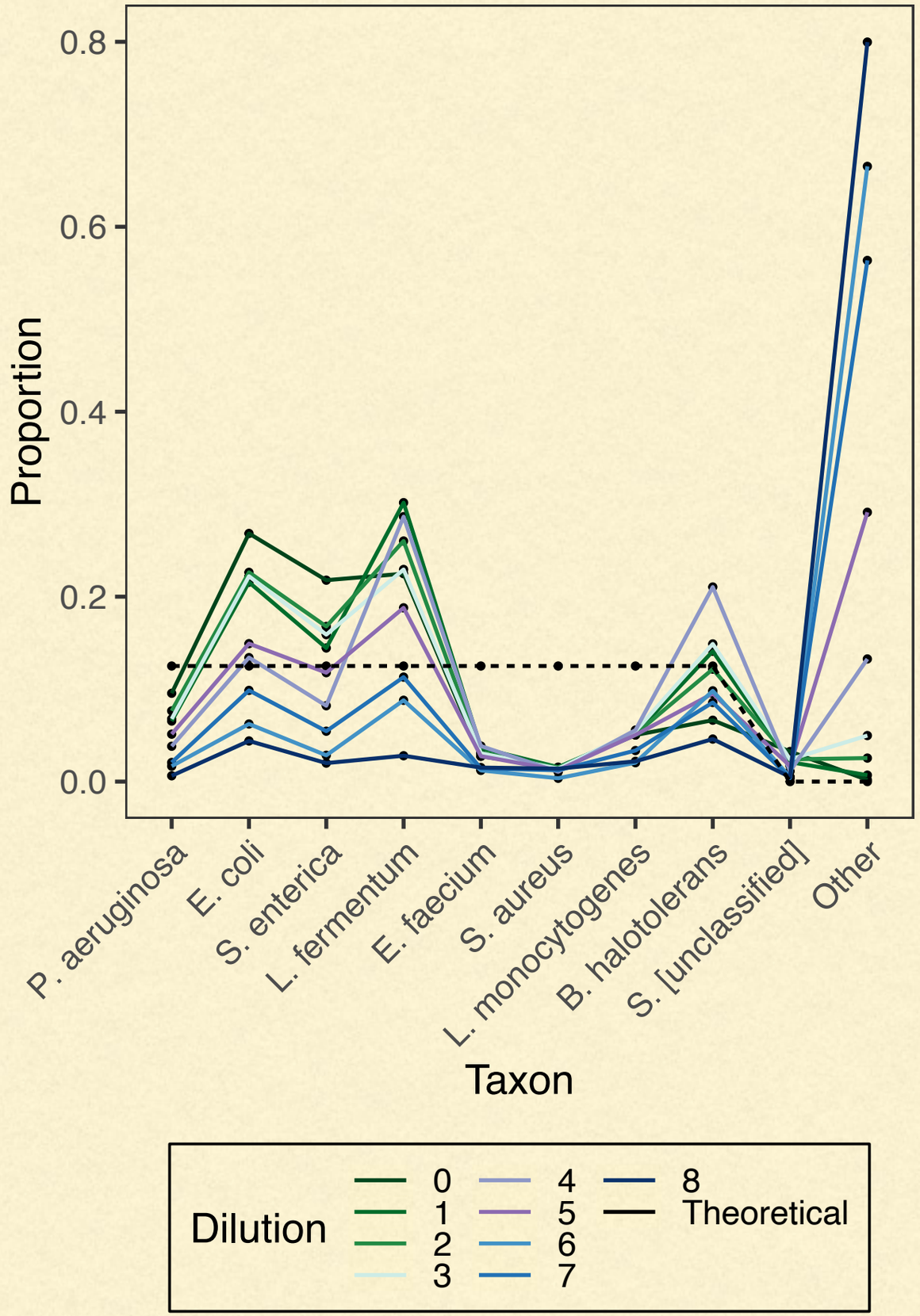


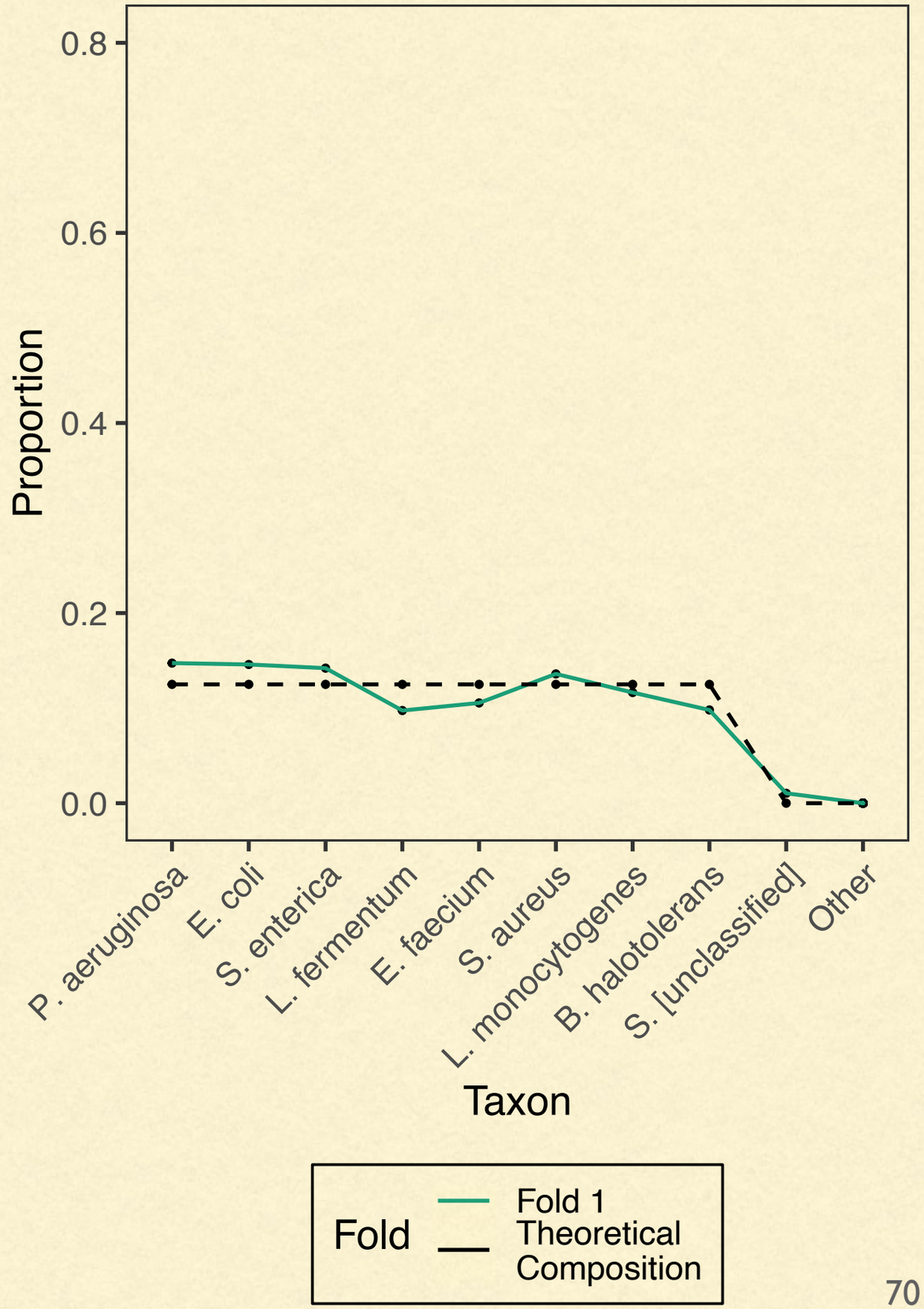
Data from Karstens et al, 2019, *mSystems*



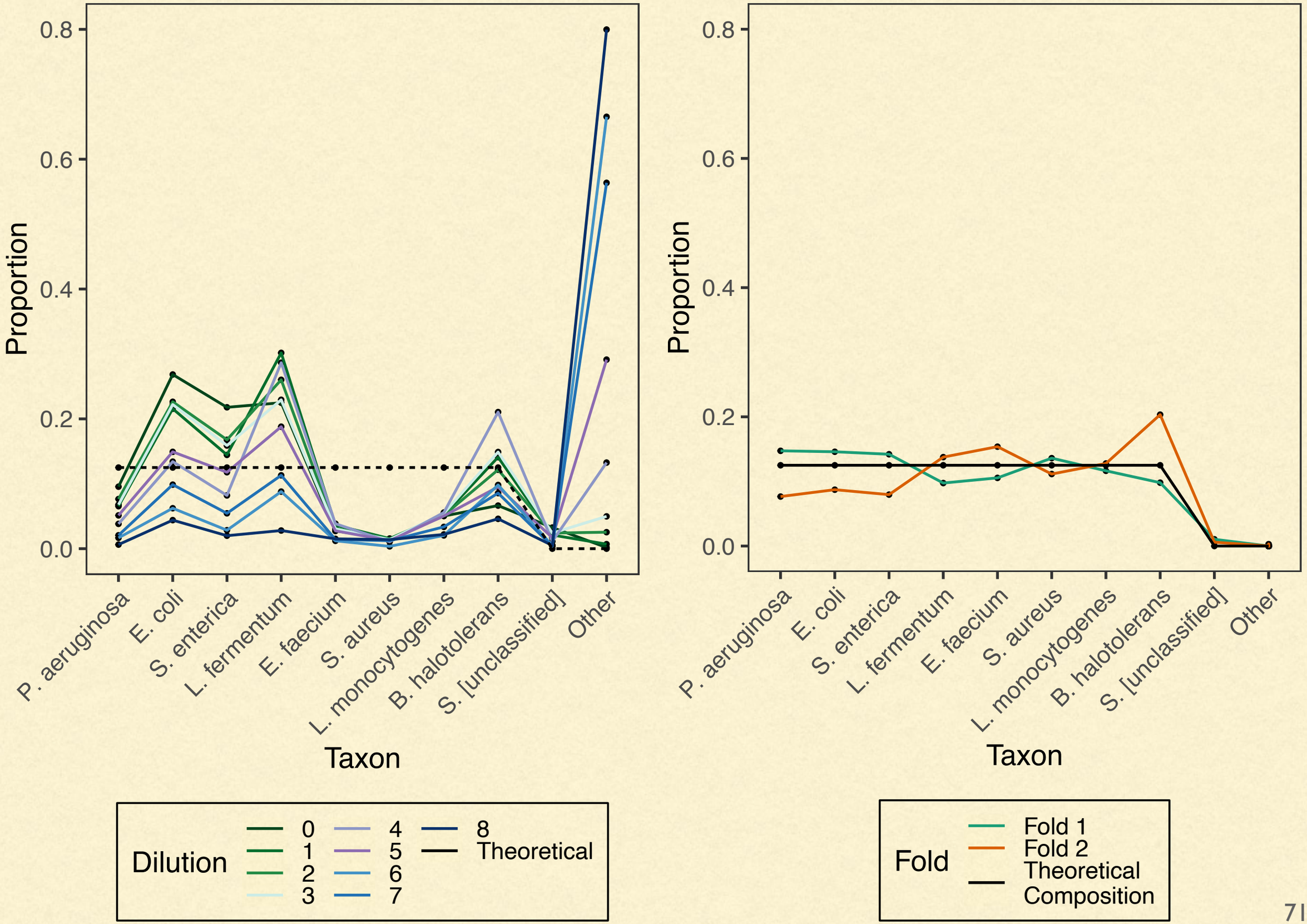


Data from Karstens et al, 2019, *mSystems*

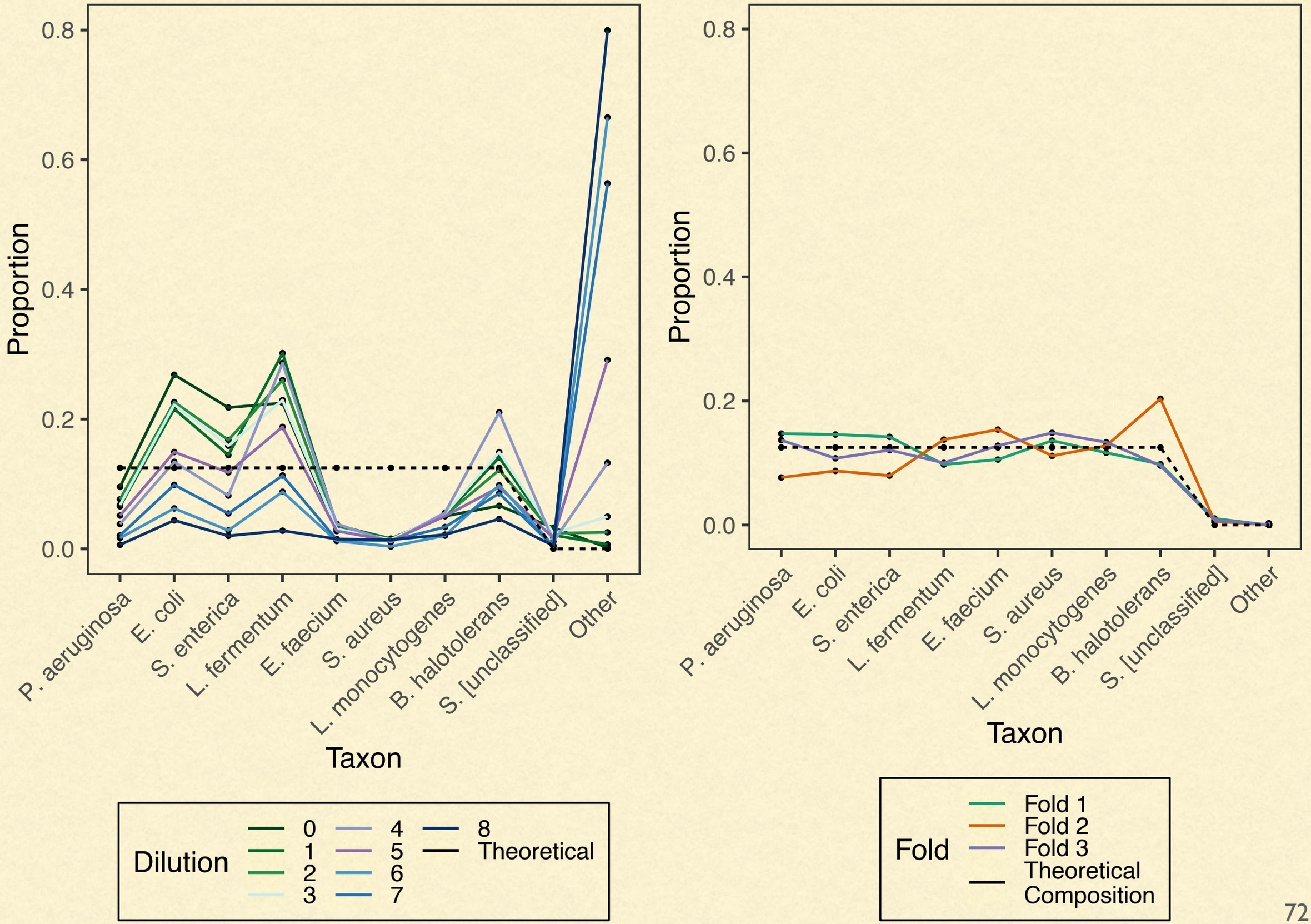




Data from Karstens et al, 2019, *mSystems*



Data from Karstens et al, 2019, *mSystems*



Data from Karstens et al, 2019, *mSystems*

DATA ANALYSIS: REMOVING CONTAMINATION

- 95% confidence intervals for elements of $\mathbf{p}$ included zero for 238 out of 240 off-target taxa

Data from Karstens et al, 2019, *mSystems*

ONGOING & FUTURE WORK

- Experimental design
  - ~~Choosing between laboratory protocols~~
  - positive vs. negative controls
    - dilution series vs. replicates
- Leveraging phylogeny to predict β_j's
- Conditions for identifiability
- Uncertainty propagation to allow comparison across sample types
- Calibration & meta-analysis 💰🌈

SUMMARY

- Observed relative abundances are biased for actual relative abundance (16S & shotgun)
- Implications for analyzing microbiome data:
  - *Proportions can lead to conclusions in the wrong direction*
  - Efficiencies vary across taxa; avoid aggregating by taxonomy/phylogeny
  - Correction possible with mock communities or qPCR data
  - Remove contamination with dilution series

arXiv > stat > arXiv:2204.12733

Statistics > Methodology

Modeling complex measurement error in microbiome experiments

David S Clausen, Amy D Willis

David Clausen (UW)

Ben Callahan (NCSU)

DOI: 10.1111/biom.13503

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

A multiview model for relative and absolute microbial abundances

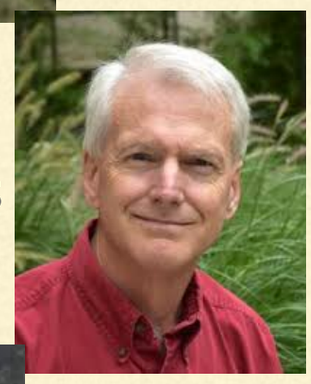
Brian D. Williamson | James P. Hughes | Amy D. Willis

Michael McLaren (MIT)

eLIFE elifesciences.org

Consistent and correctable bias in metagenomic sequencing experiments

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



Jim Hughes (UW)

Evaluating replicability in microbiome data

David S Clausen, Amy D Willis

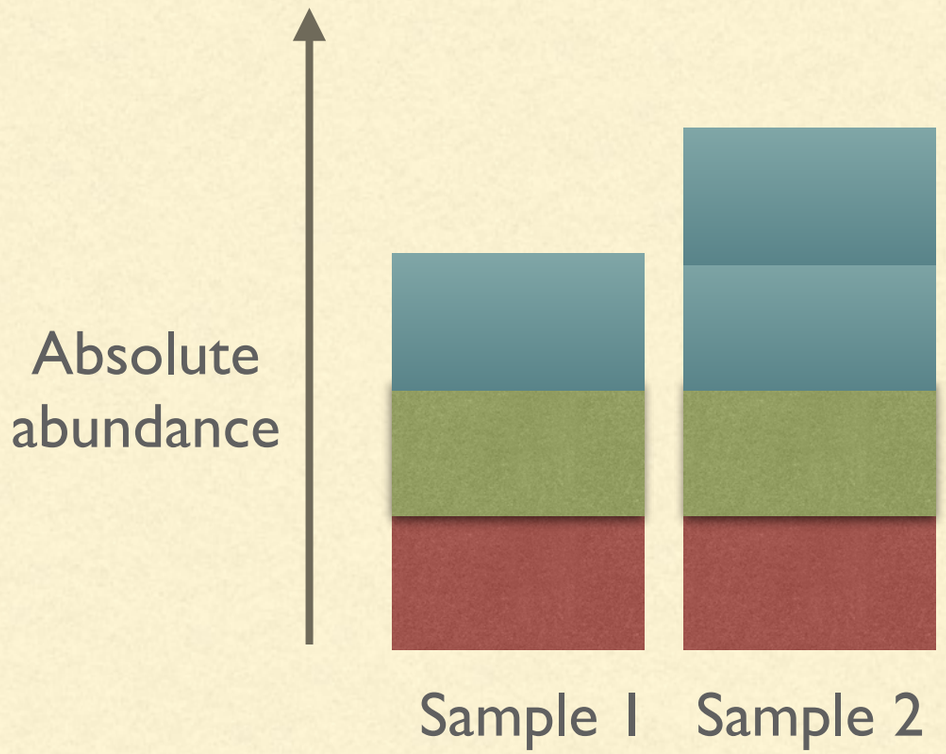
Biostatistics, kxab048, https://doi.org/10.1093/biostatistics/kxab048

Brian Williamson (Kaiser)

NIH NIGMS

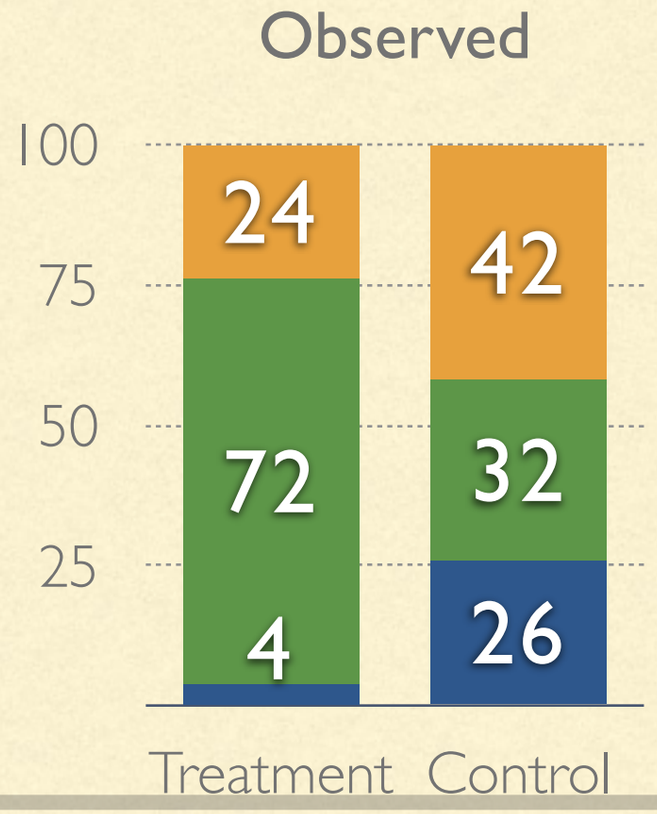
LIMITATIONS OF RELATIVE ABUNDANCE: # 1

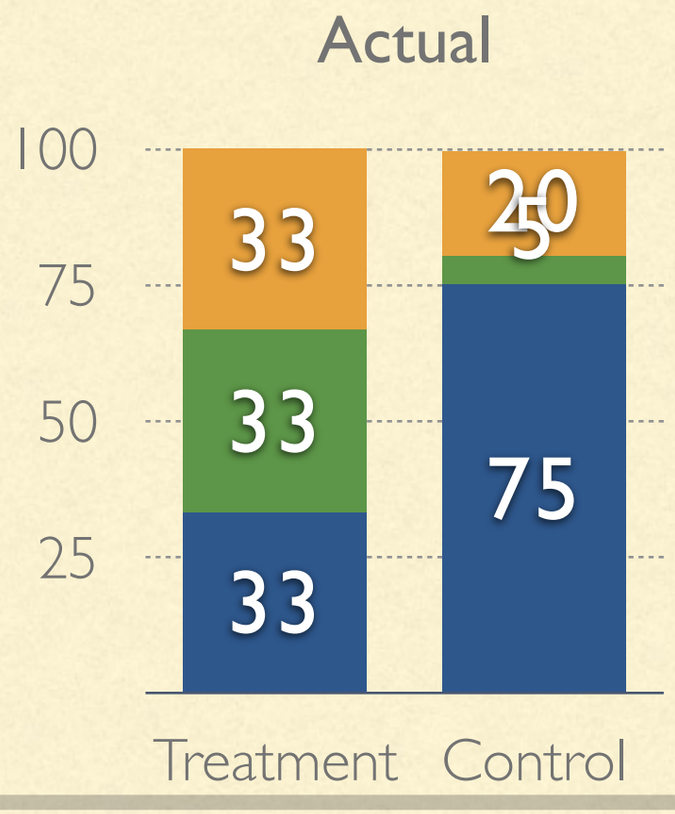
- Relative abundance of all taxa change even when only one relative abundance changes
  - Not "spurious" but misleading
  - **0.33** / **0.33** / **0.33**
  - **0.25** / **0.25** / **0.50**
- This is an inherent limitation of relative abundance analysis

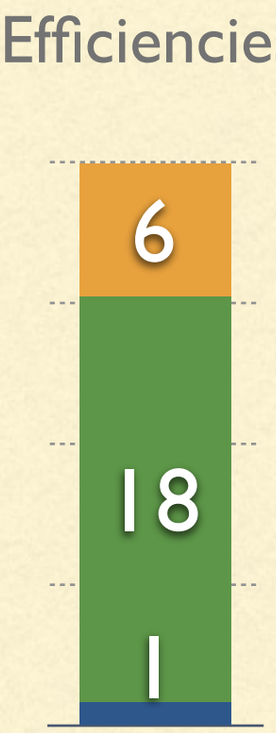


LIMITATIONS OF RELATIVE ABUNDANCE: #2

- Different detection efficiencies adversely impact relative abundance estimates
- Different detection efficiencies adversely impact estimates of changes in relative abundance (including direction)

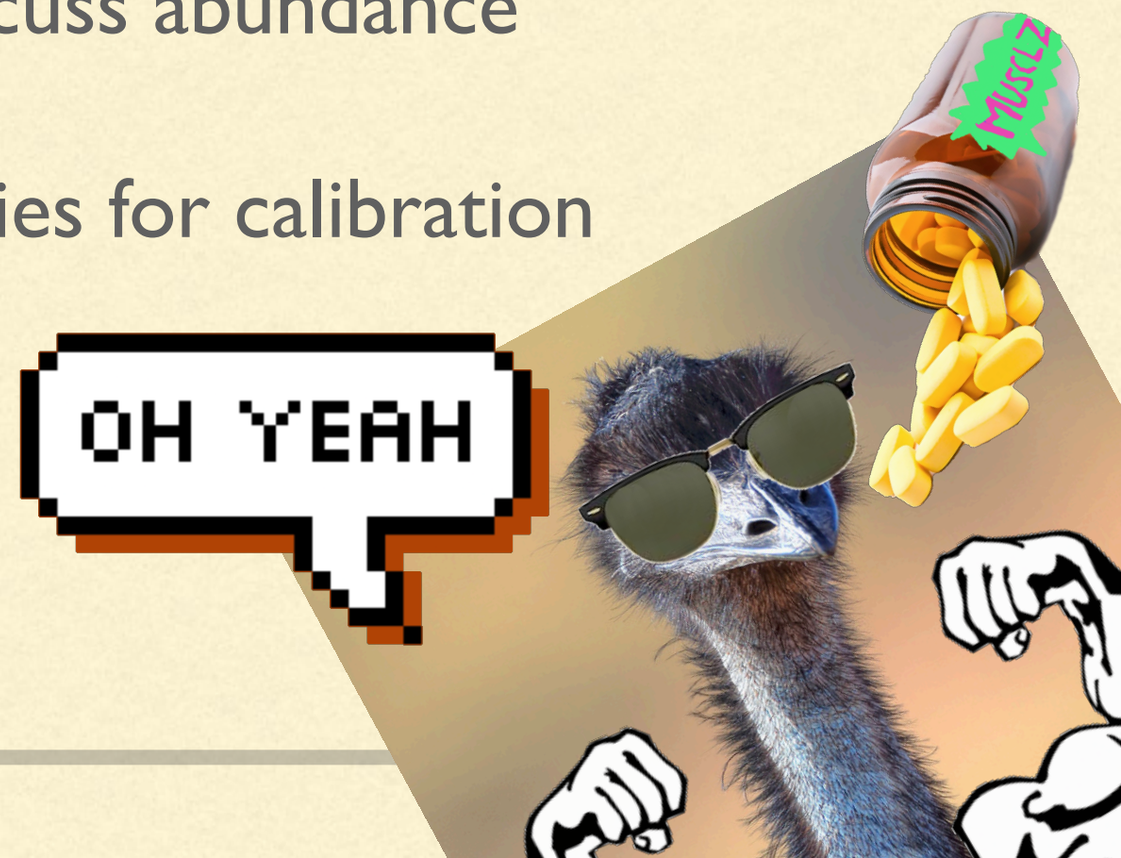






IMPLICATIONS

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



CLOSING THOUGHTS

- Methods for modeling microbiome data is a fast-moving field, and new methods are constantly emerging
- Talk to lots of people
  - "What's the biggest limitation of this?"
- Stay critical but open-minded

ACCESSING `TINYVAMP` LAB

1. Go to schedule on Wiki to Thursday afternoon, click on "Labs"
2. Copy the command under `tinyvamp` lab
3. Run this command in your RStudio Server console; open file and start working through

Get pumped!

MODELING SHOTGUN SEQUENCING DATA

Statistical Diversity Lab @ University of Washington

Amy Willis — @AmyDWillis — Assistant Professor

David Clausen — @davidandacat — PhD Candidate

Sarah Teichman — @sarah_teichman — PhD Candidate

CONSISTENCY OF EFFICIENCIES

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

Data from Brooks et al., 2015; Analysis from McLaren et al., 2019

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

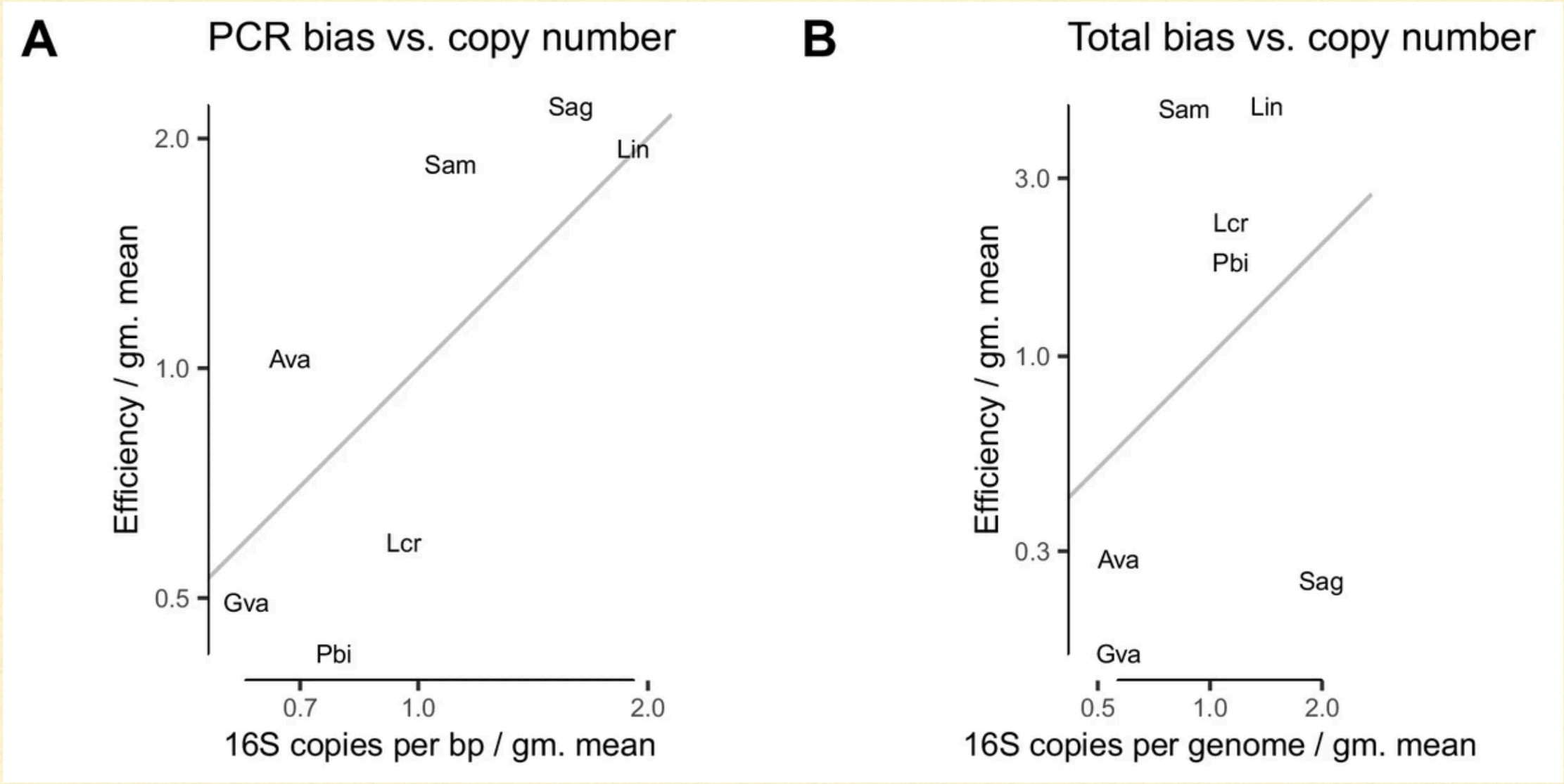


Image from McLaren et al., 2019, *eLife*