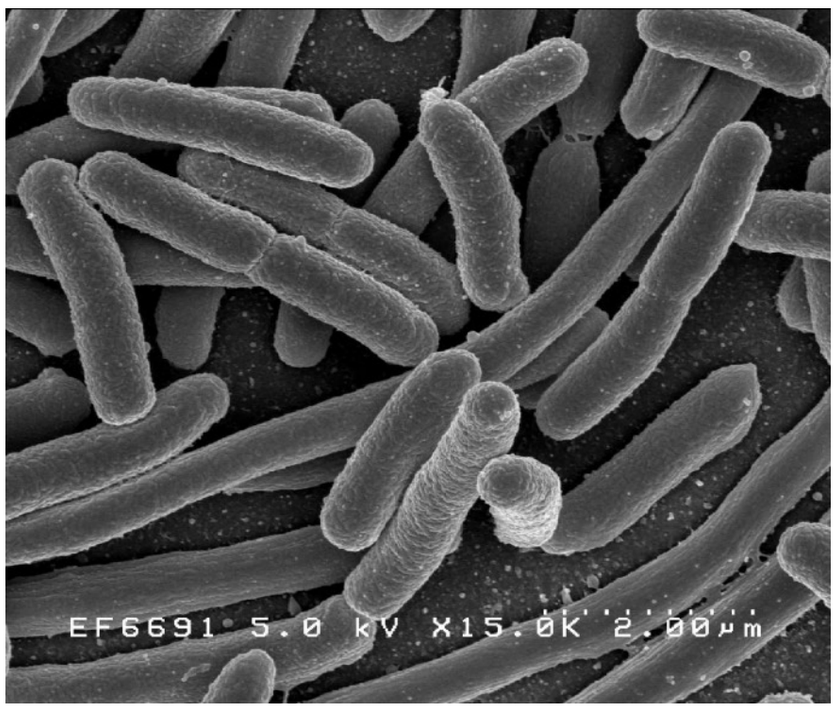
Biased* and Methodology-specific Measurement of Microbial Communities

STAMPS 2023

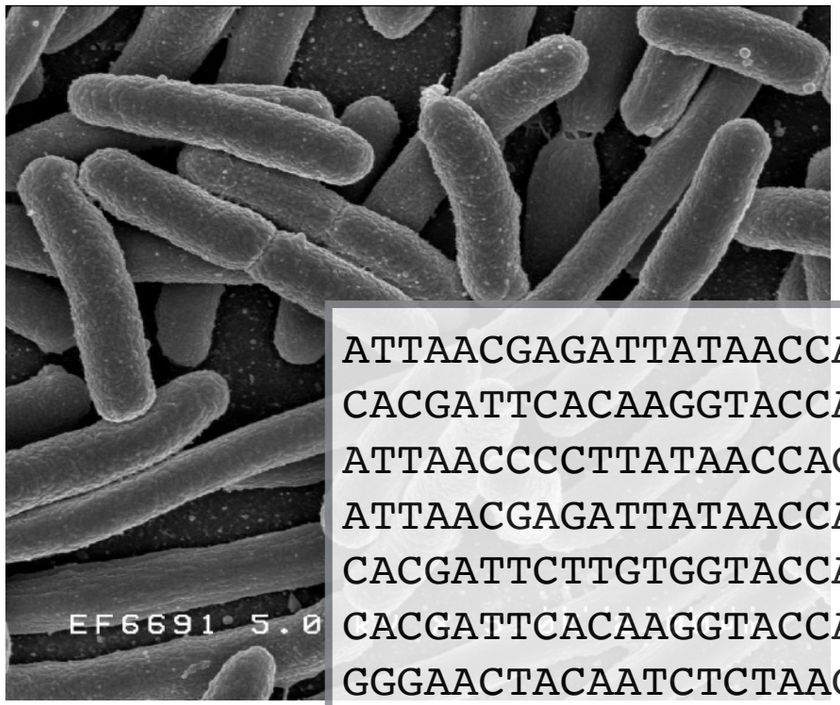
Microbial Census

Metagenomics or Marker-gene (MGS) Sequencing



Microbial Census

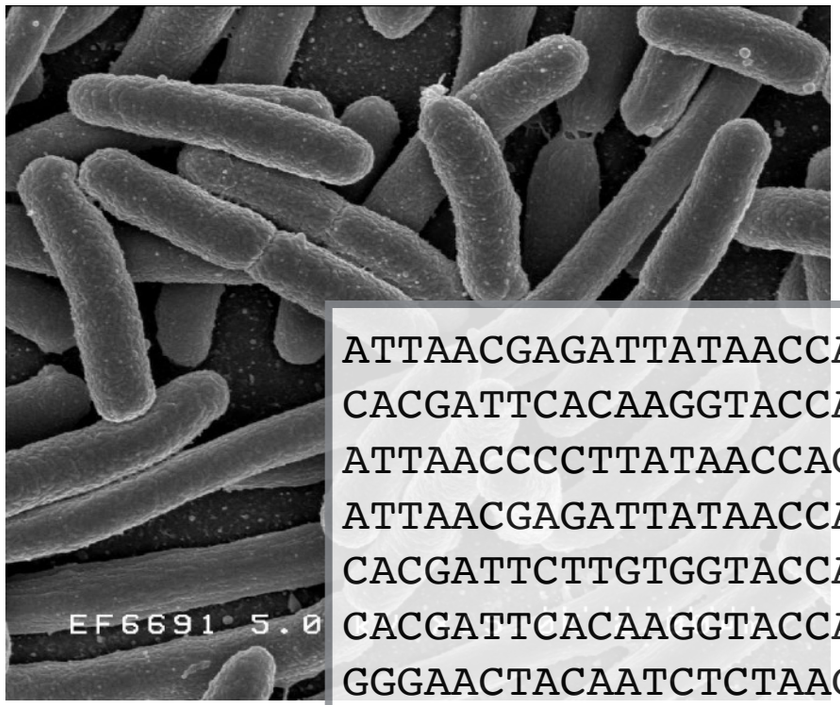
Metagenomics or Marker-gene (MGS) Sequencing




```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
```

Microbial Census

Metagenomics or Marker-gene (MGS) Sequencing

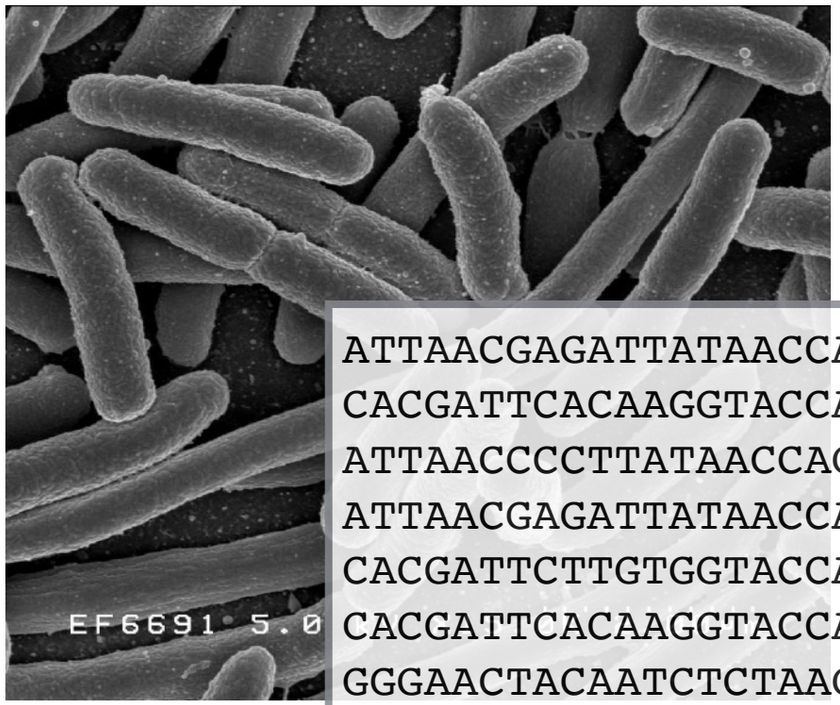



```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

Microbial Census

Metagenomics or Marker-gene (MGS) Sequencing




```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

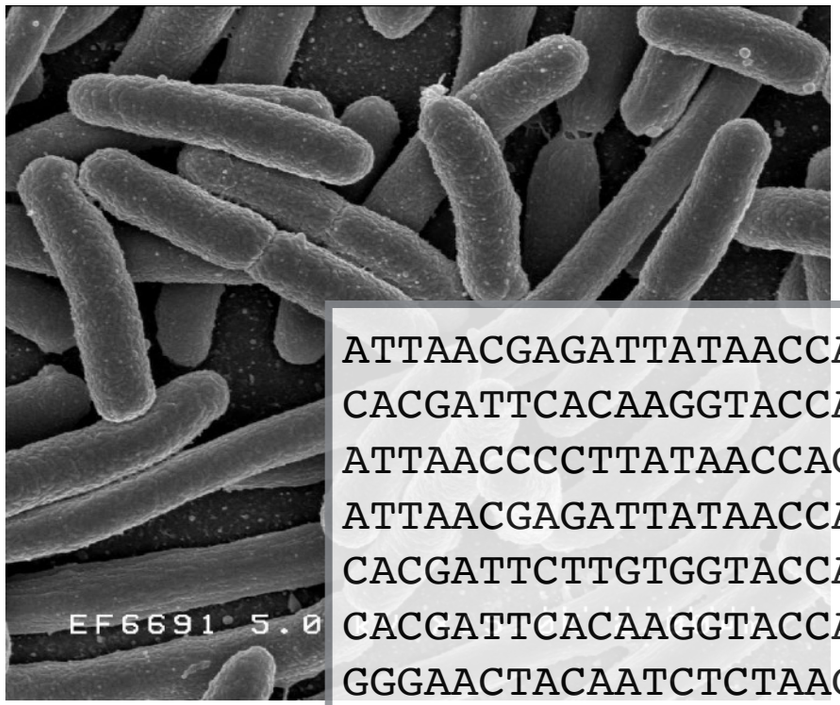
Inference

Exploration

Modeling

Microbial Census

Metagenomics or Marker-gene (MGS) Sequencing




```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

- ASV table from DADA2
- Taxonomy table from read recruitment
- Taxonomy table from Sourmash-gather

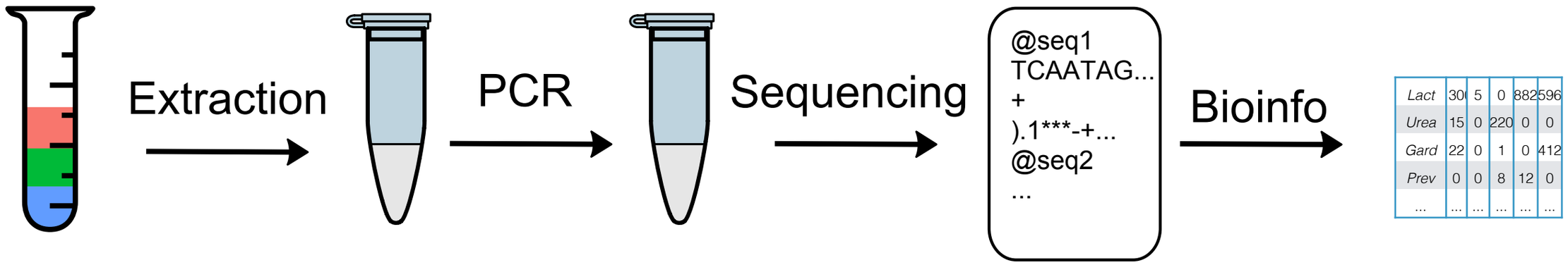
...and many, many other methods...

Inference

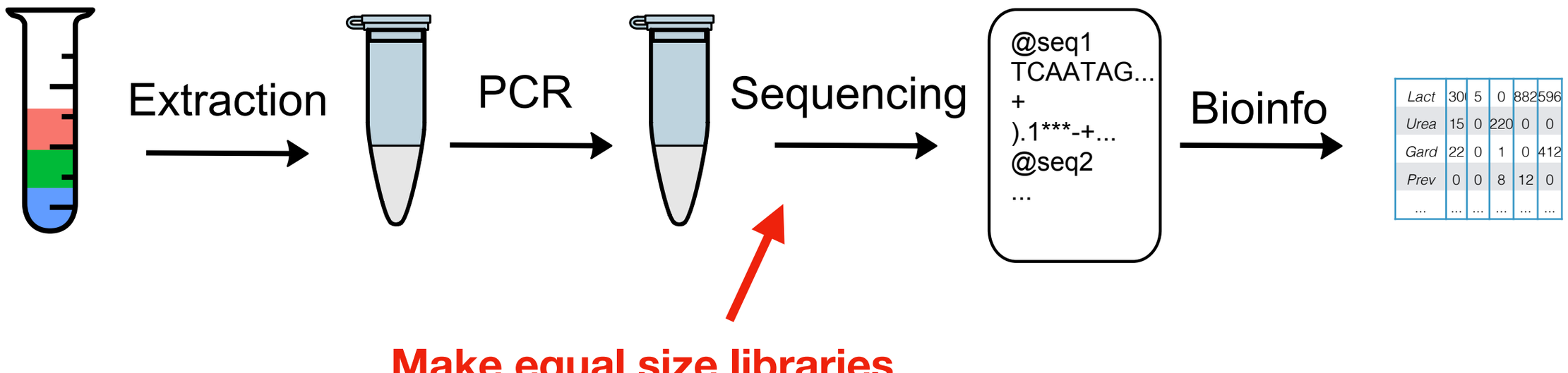
Exploration

Modeling

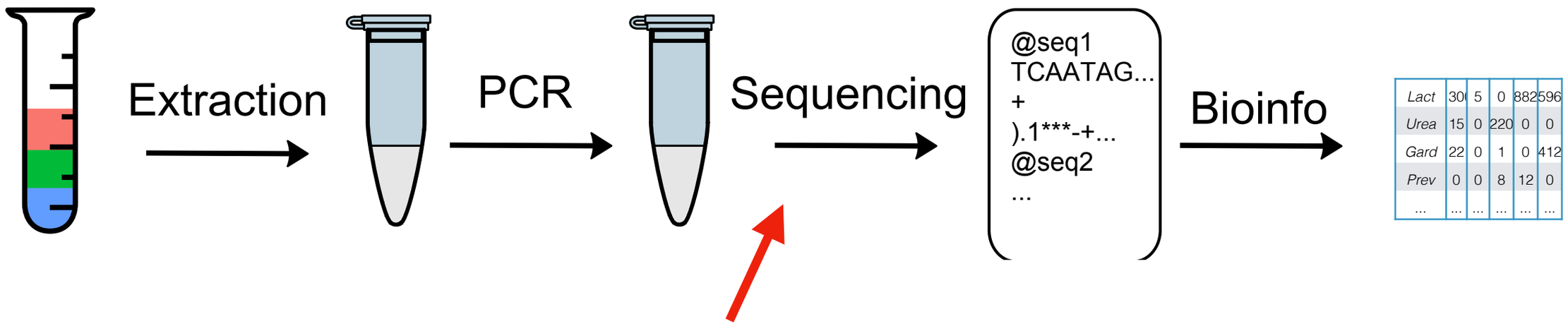
Metagenomic Compositionality



Metagenomic Compositionality



Metagenomic Compositionality



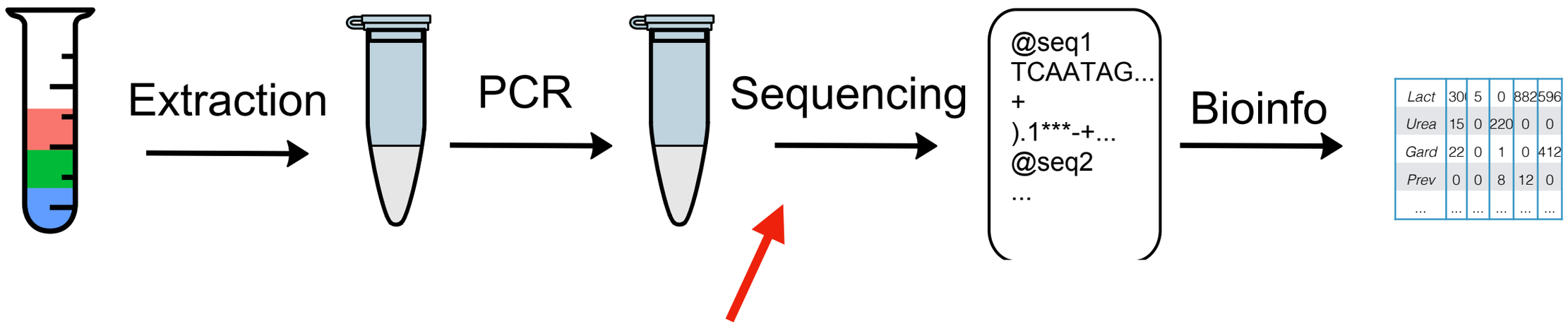
Lact	30	5	0	882	596	
Urea	15	0	220	0	0	
Gard	22	0	1	0	412	
Prev	0	0	8	12	0	
...	...	...	...	...	...	

1000X microbial concentration

↓

No change in MGX measurement

Metagenomic Compositionality



Lact	30	5	0	882	596	
Urea	15	0	220	0	0	
Gard	22	0	1	0	412	
Prev	0	0	8	12	0	
...	...	...	...	...	...	

Every read in the data is the result of a

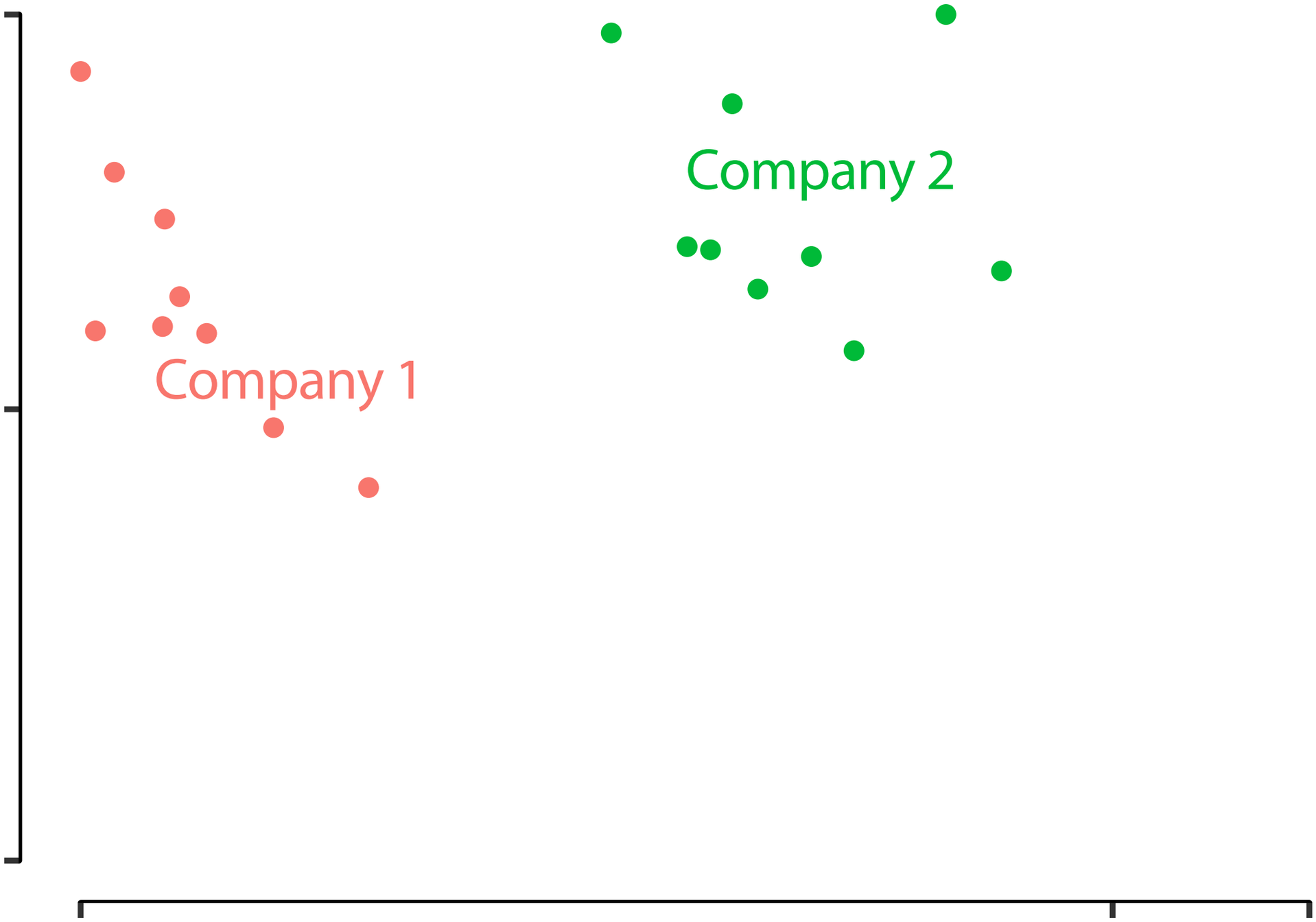
competition

with all other sources of library-DNA for limited slots

Metagenomic Bias*

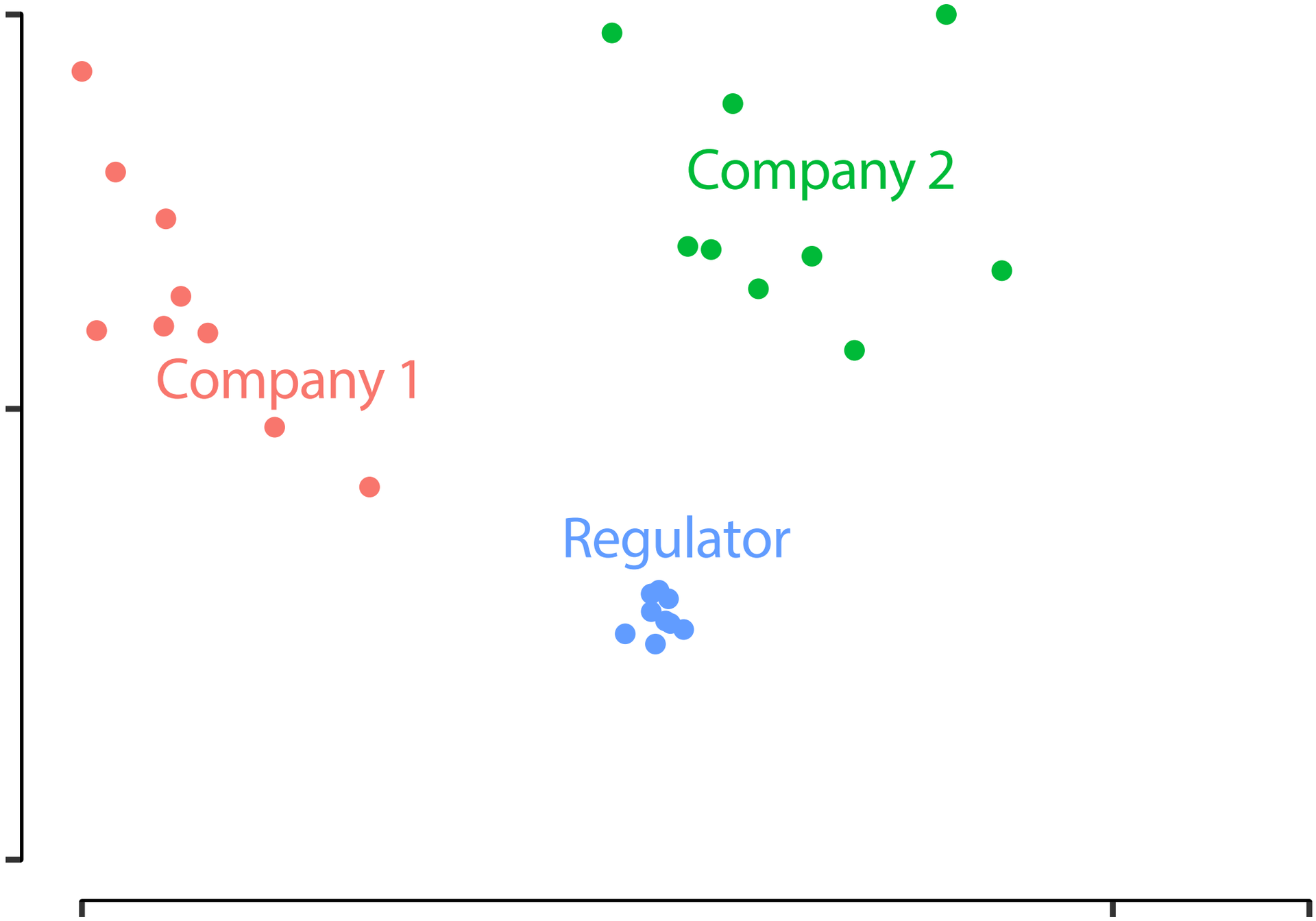
Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Bias*



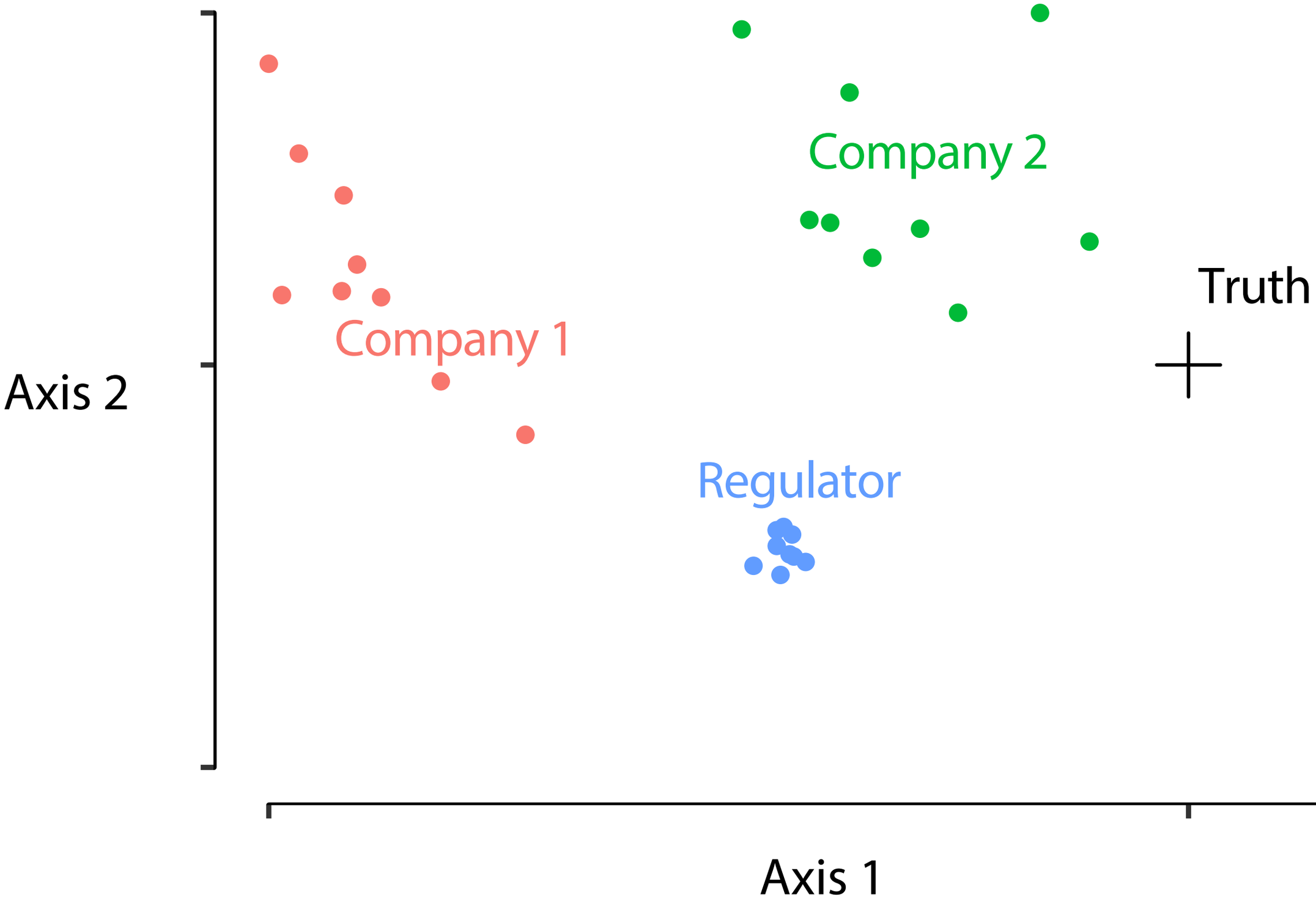
Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Bias*



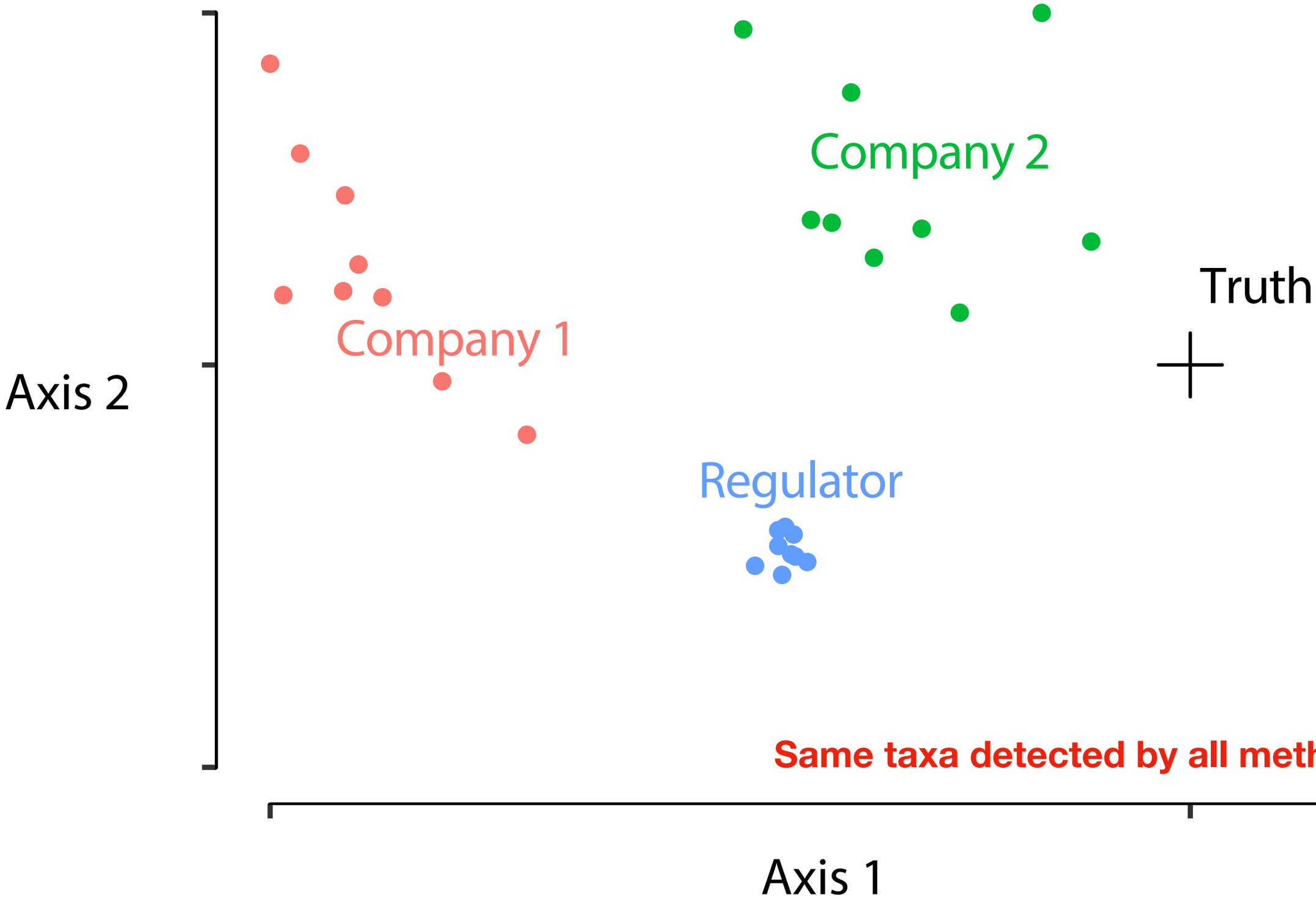
Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Bias*



Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Bias*



Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Bias*

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

Alan W

Microb

https:/

Receiv

CopyRighter: a rapid tool for improving the accuracy of microbial community profiles throu
y number correction

ka Vanwonterghem, Philip Hugenholtz and Gene

; licensee BioMed Central Ltd. 2014

014 | **Published:** 7 April 2014

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

in metagenomic analysis

Silvia Cardona, Anat Eck, Montserrat Cassellas
Joaquim Roca, Francisco Guarner and Chaysav

BMC Microbiology 2012 **12**:158
https://doi.org/10.1186/1471-2180-12-158 | ©
Received: 6 March 2012 | **Accepted:** 20 July 2

D
T

Sample

Chengwei Luo, Despina Tsementzi, Nikos K

Published: February 10, 2012 • https://doi.

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 | © Wesolowska-Andersen et al.; licensee BioMed Central Ltd. 2014
Received: 3 February 2014 | **Accepted:** 25 April 2014 | **Published:** 5 June 2014

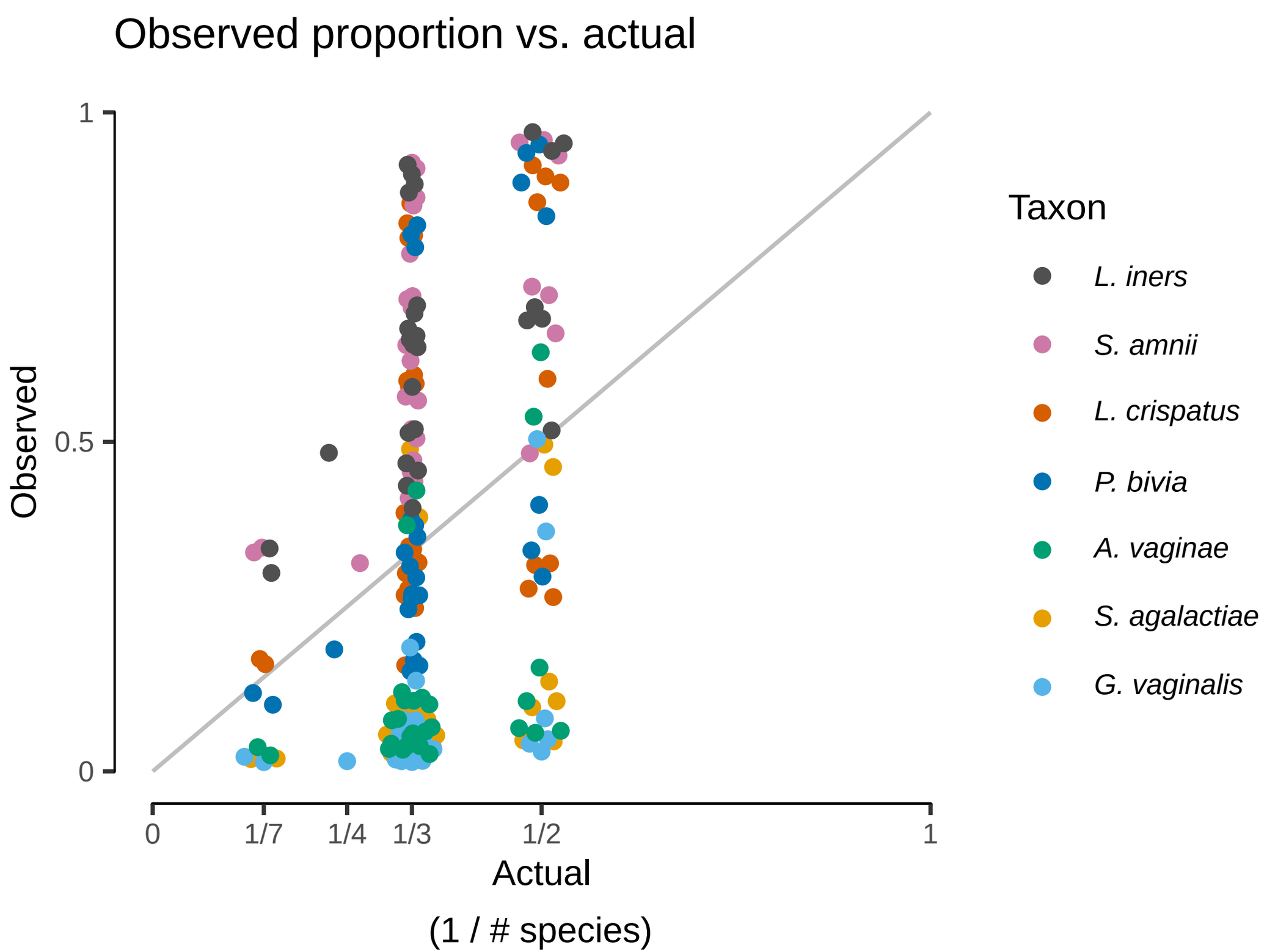
Every metagenomic measurement is **biased*** from the truth

Every metagenomic measurement is **biased*** from the truth

The relative abundances measured by metagenomics are systematically inaccurate

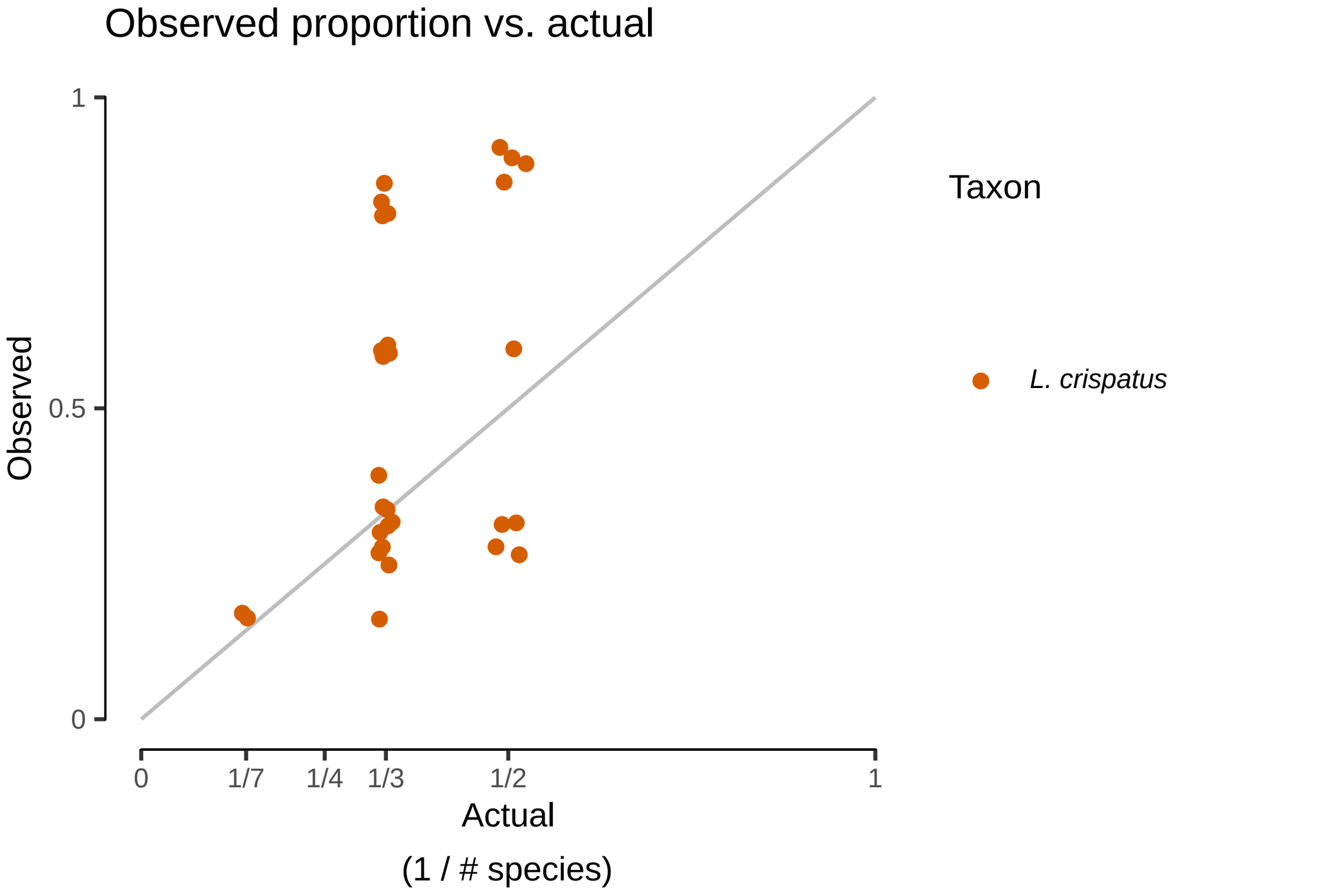
MGS measurements are not quantitatively reproducible across labs/methods

Metagenomic Bias



McLaren, Willis & Callahan. *eLife*, 2019.

Metagenomic Bias



McLaren, Willis & Callahan. *eLife*, 2019.

Modeling Metagenomics Bias

Truth

- 33%
- 33%
- 33%

A

Measurement

Measured

- 4%
- 72%
- 24%

O

Modeling Metagenomics Bias

Truth

- 33%
- 33%
- 33%

A

$$O = f(A)$$

Measured

- 4%
- 72%
- 24%

O

Modeling Metagenomics Bias

What is f( )?

Truth

- 33%
- 33%
- 33%

A

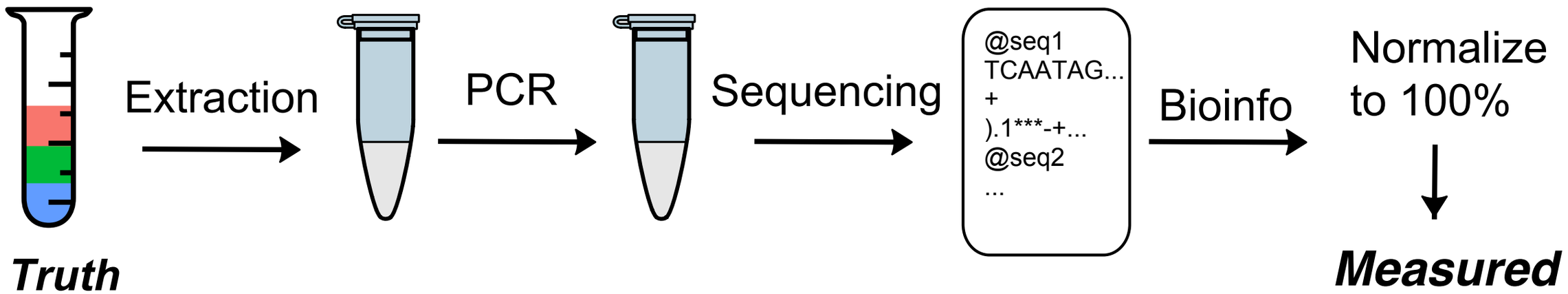
$$O = f(A)$$

Measured

- 4%
- 72%
- 24%

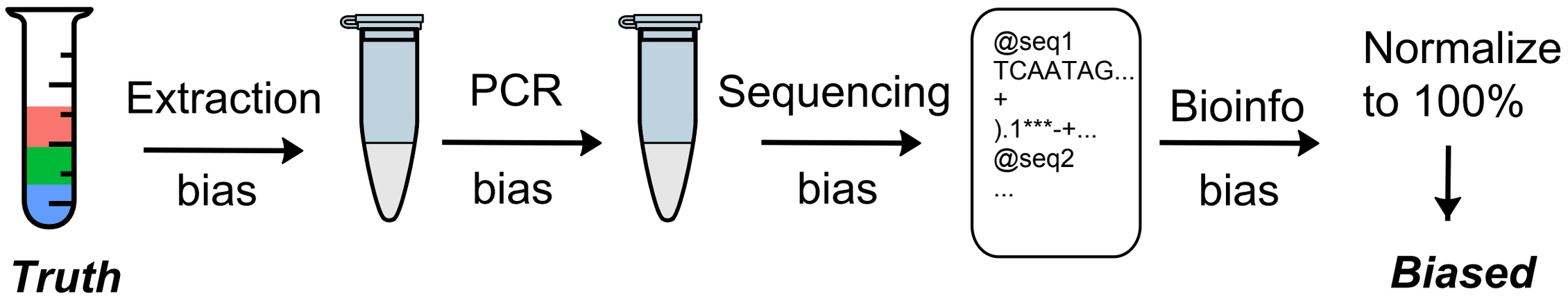
O

Modeling Metagenomics Bias



McLaren, Willis & Callahan. *eLife*, 2019.

Modeling Metagenomics Bias

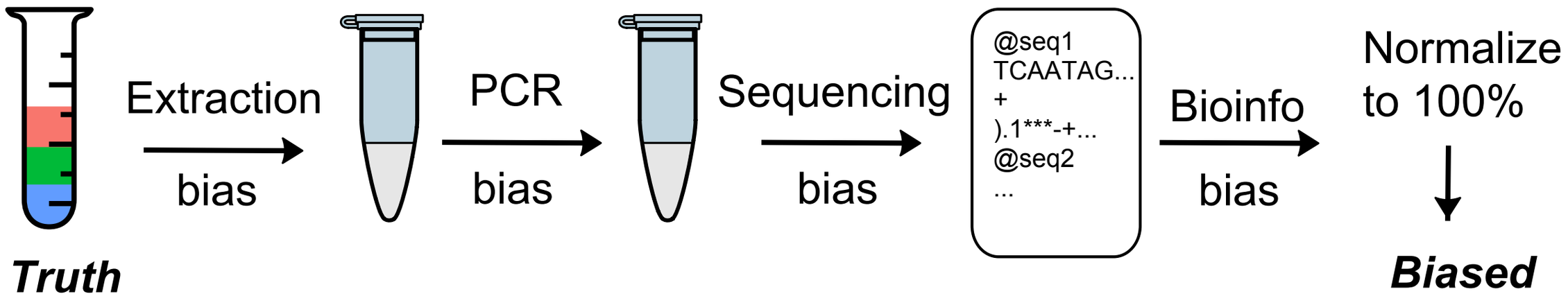


33%	x 1	x 1	x 1	x 1	4%
33%	x 4	x 6	x 0.5	x 1.5	72%
33%	x 15	x 2	x 0.25	x 0.8	24%

McLaren, Willis & Callahan. *eLife*, 2019.

Modeling Metagenomics Bias

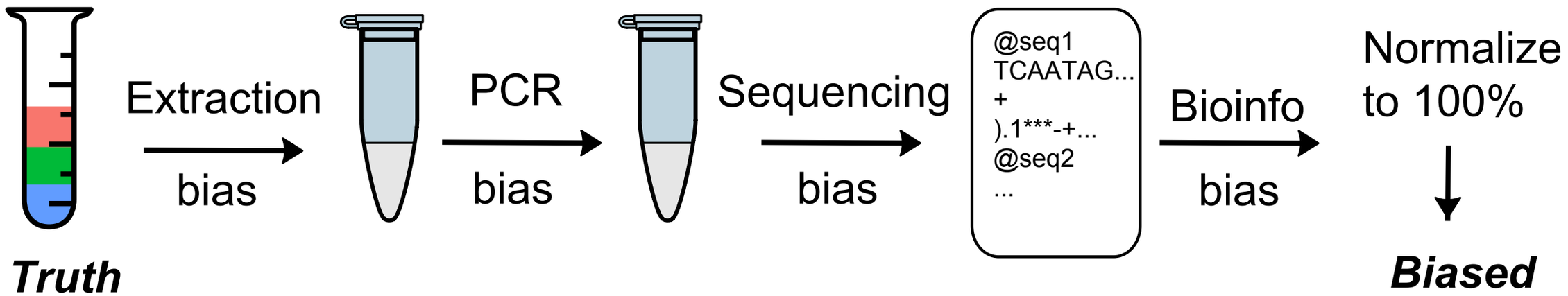
Strong evidence that some bias mechanisms act in this way.



33%	x 1	x 1	x 1	x 1	4%
33%	x 4	x 6	x 0.5	x 1.5	72%
33%	x 15	x 2	x 0.25	x 0.8	24%

McLaren, Willis & Callahan. *eLife*, 2019.

Modeling Metagenomics Bias

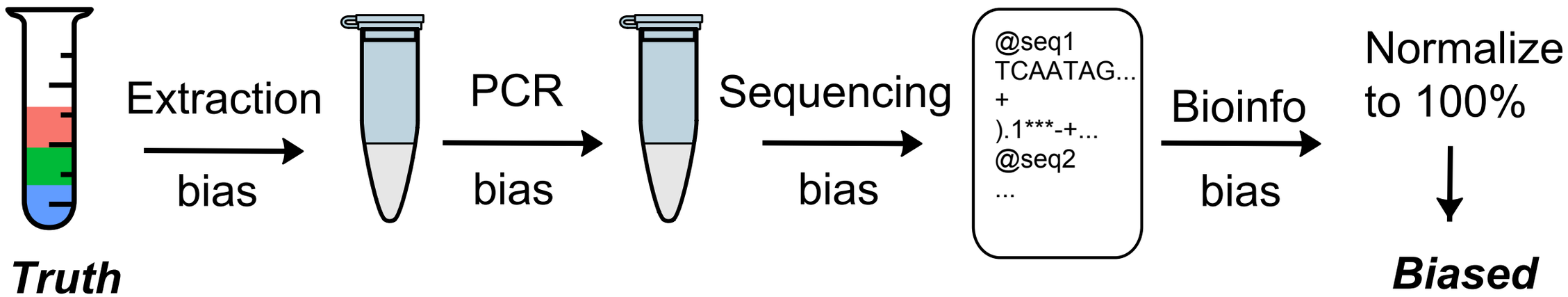


33%	x 1	x 1	x 1	x 1	4%
33%	x 4	x 6	x 0.5	x 1.5	72%
33%	x 15	x 2	x 0.25	x 0.8	24%

$$\mathbf{O} \sim \mathbf{A} \cdot \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2)} \cdot \ldots \cdot \mathbf{B}^{(P_L)}$$

McLaren, Willis & Callahan. *eLife*, 2019.

Modeling Metagenomics Bias



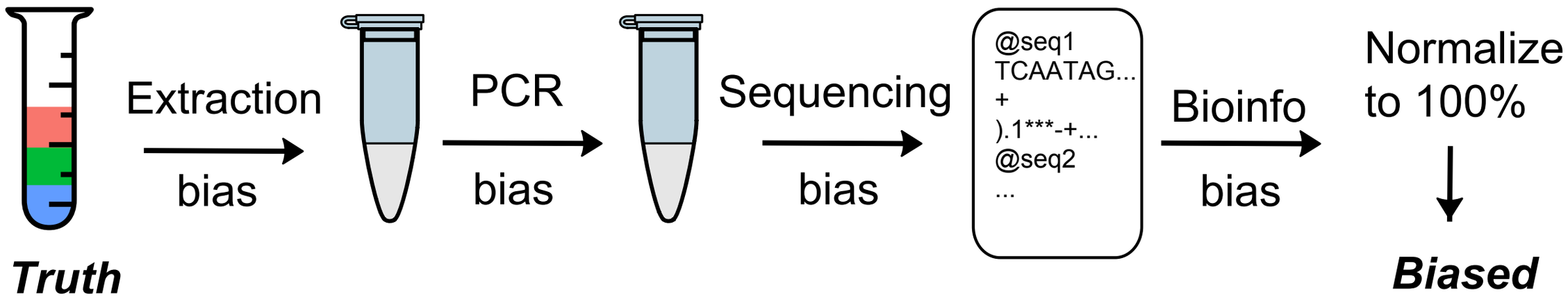
● 33%	x 1	x 1	x 1	x 1	● 4%
● 33%	x 4	x 6	x 0.5	x 1.5	● 72%
● 33%	x 15	x 2	x 0.25	x 0.8	● 24%
1 : 1 : 1					***1 : 18 : 6***

$$\mathbf{O} \sim \mathbf{A} \cdot \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2)} \cdot \ldots \cdot \mathbf{B}^{(P_L)}$$

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

McLaren, Willis & Callahan. *eLife*, 2019.

Modeling Metagenomics Bias



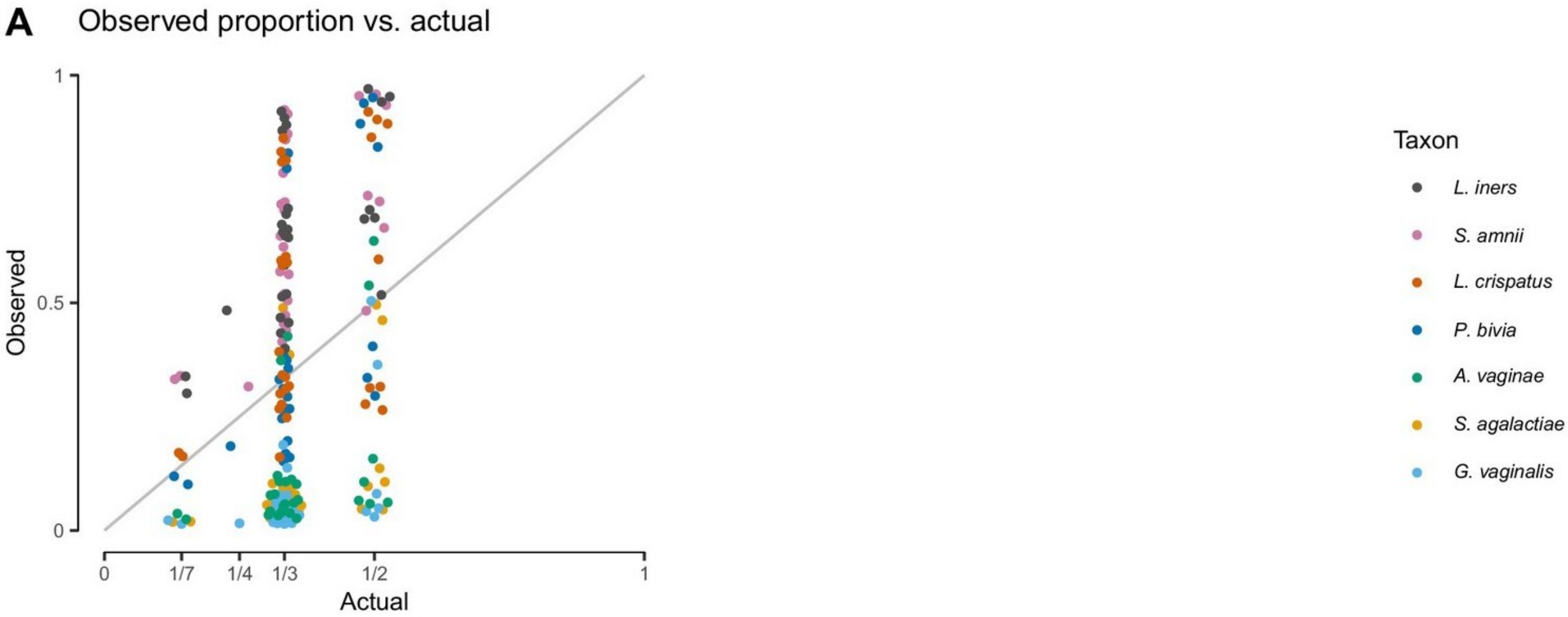
33%	x 1	x 1	x 1	x 1	4%
33%	x 4	x 6	x 0.5	x 1.5	72%
33%	x 15	x 2	x 0.25	x 0.8	24%
1 : 1 : 1					***1 : 18 : 6***

$$\mathbf{O} \sim \mathbf{A} \cdot \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2)} \cdot \ldots \cdot \mathbf{B}^{(P_L)}$$

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

McLaren, Willis & Callahan. *eLife*, 2019.

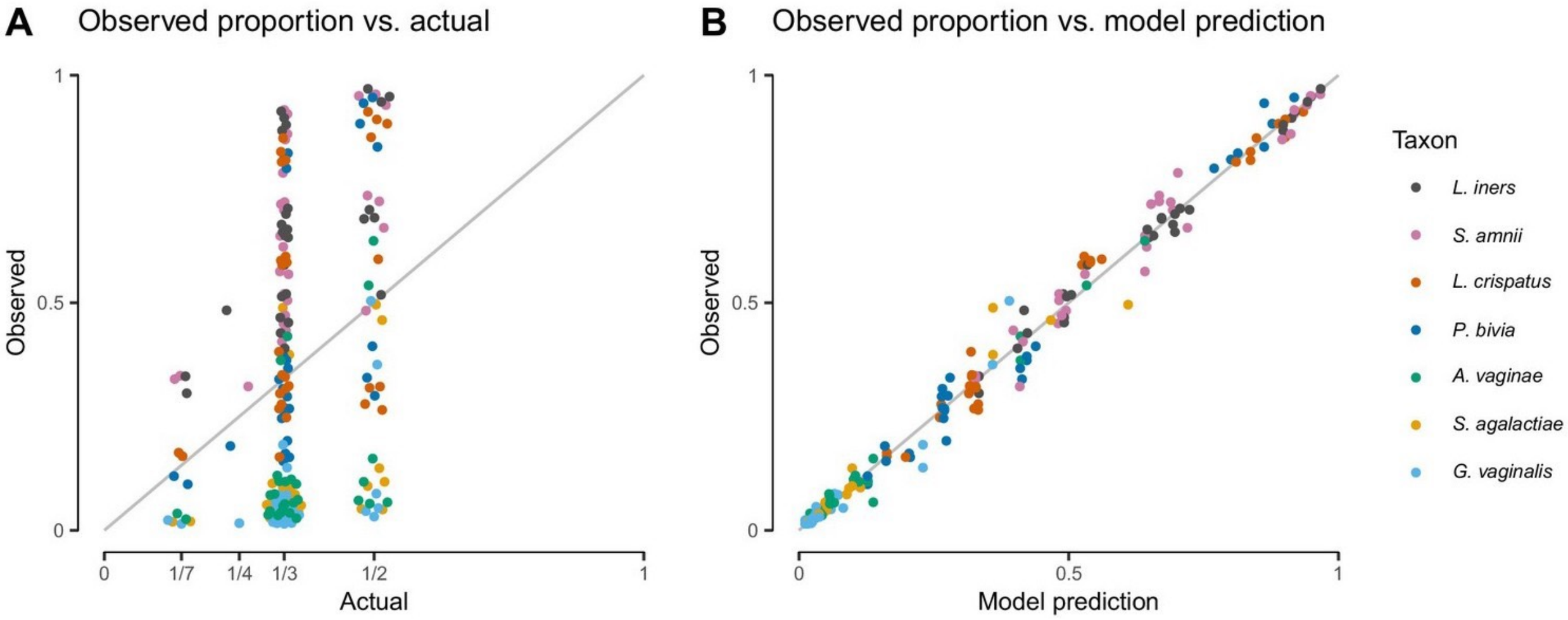
Testing the Model



McLaren, Willis & Callahan. *eLife*, 2019.

Testing the Model

It works!



This is f( )

$$Pr(\mathbf{O})_i = \frac{O_i}{\sum_{j=1}^{K} O_j} = \frac{Pr(\mathbf{A})_i B_i}{\sum_{j=1}^{K} Pr(\mathbf{A})_j B_j}$$

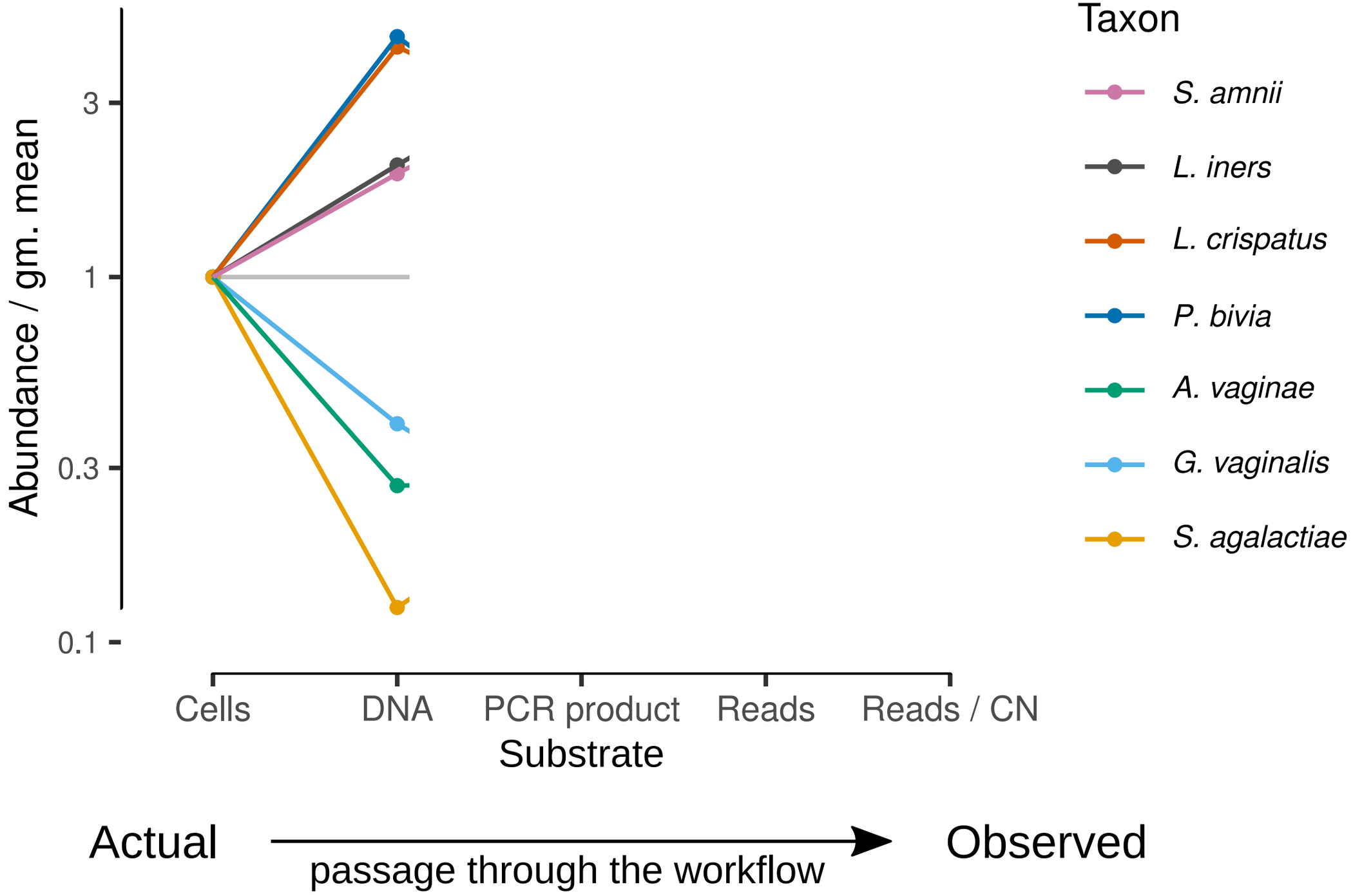
McLaren, Willis & Callahan. *eLife*, 2019.

A simple model of bias links measurements to the truth

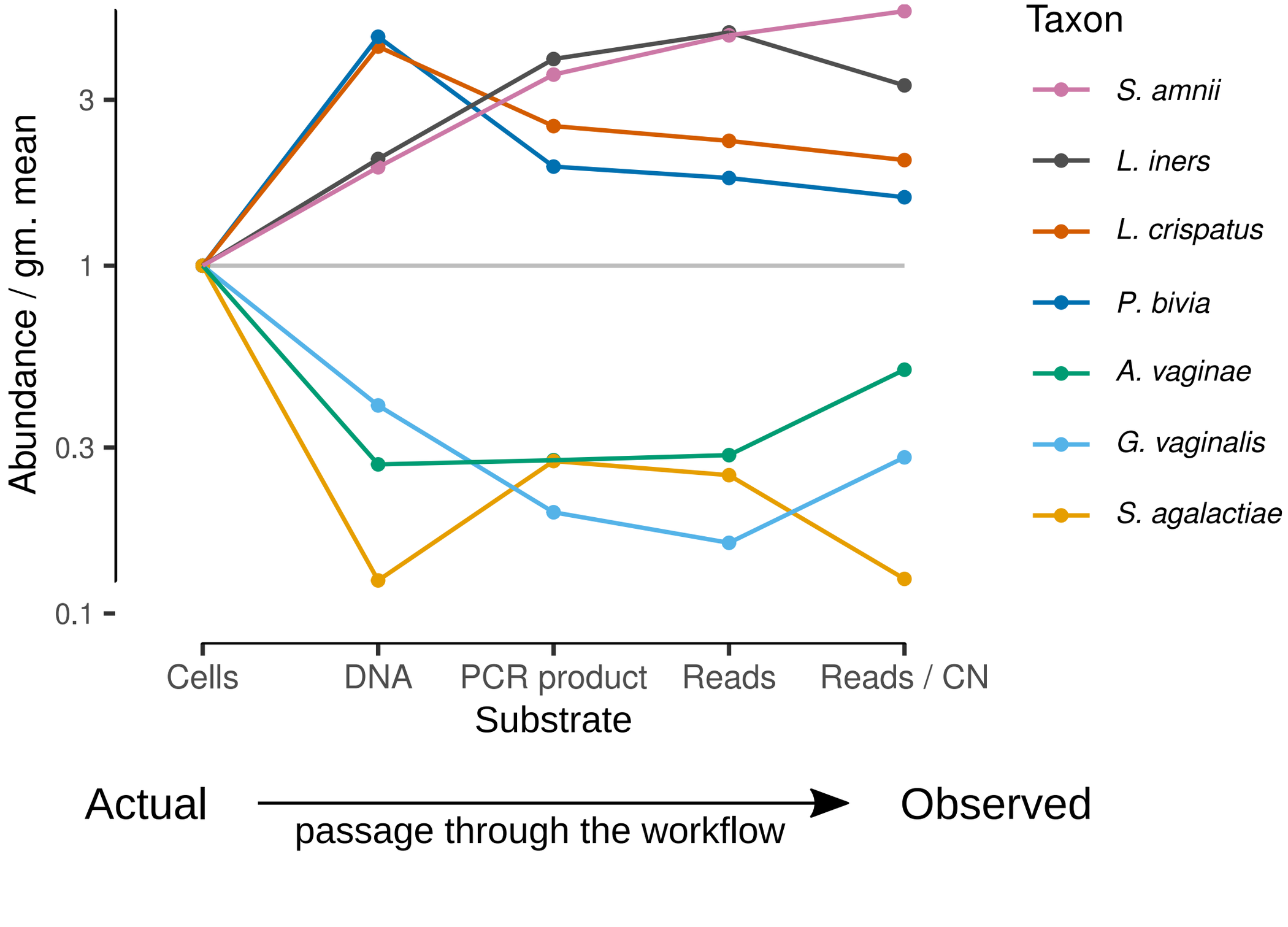
A simple model of bias links measurements to the truth

… at least in simple and well-controlled conditions.

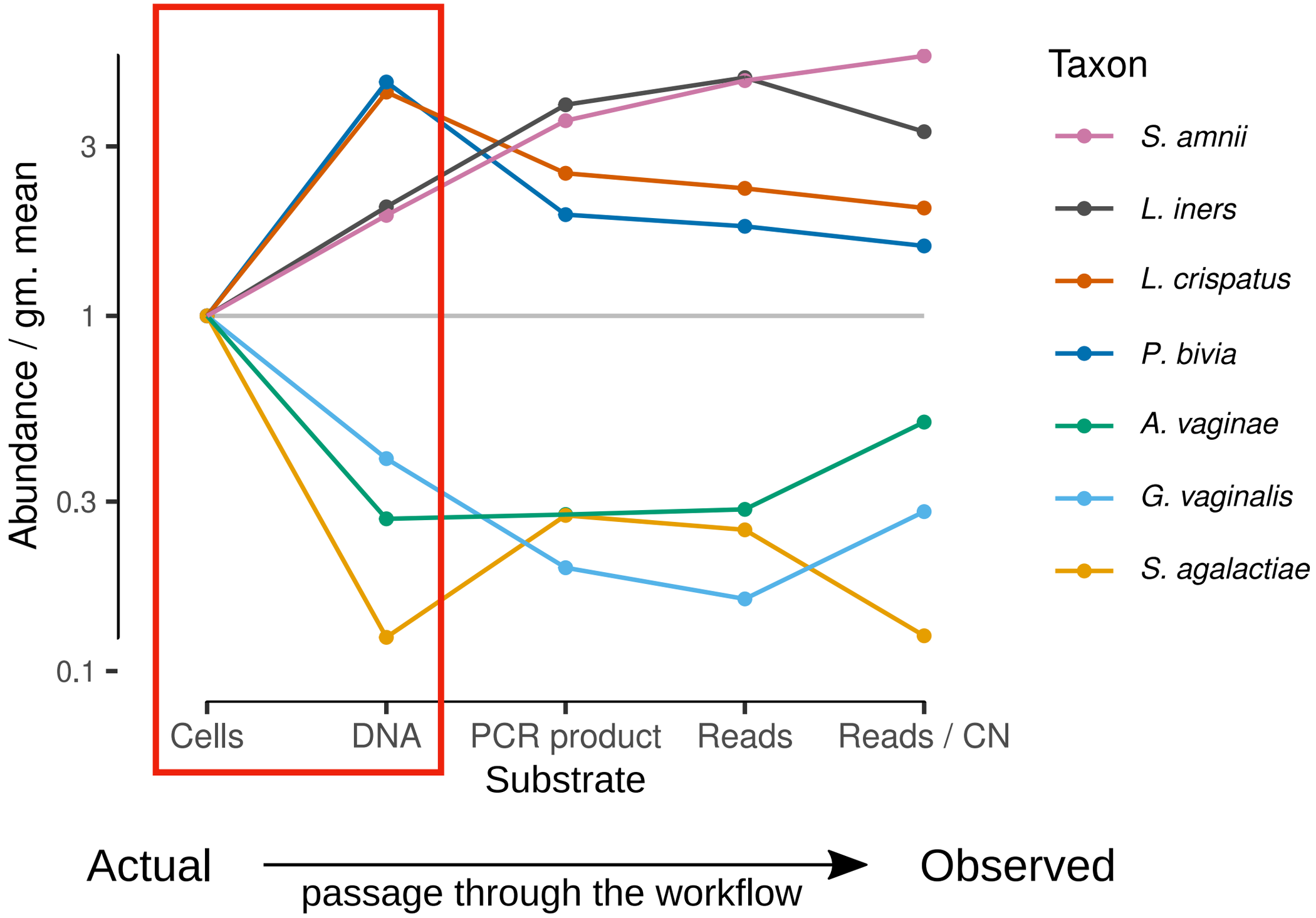
Protocol Optimization



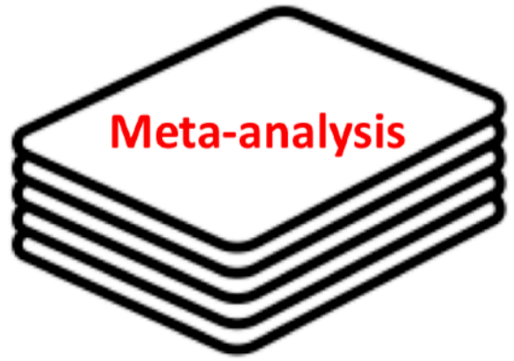
Protocol Optimization



Protocol Optimization



Synthesizing Across Studies

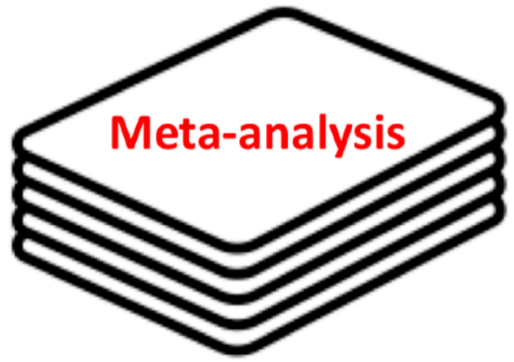


- The vaginal microbiome and PTB
- 16s rRNA gene sequencing
- Raw & metadata publicly available or reachable

Currently, we have 12 studies (6281 samples; 1926 subjects) and at least 2 studies are in progress.

Image credit: Vaginal Microbiome Consortium.

Synthesizing Across Studies



- The vaginal microbiome and PTB
- 16s rRNA gene sequencing
- Raw & metadata publicly available or reachable

Currently, we have 12 studies (6281 samples; 1926 subjects) and at least 2 studies are in progress.

Different Methods = Different Biases

Image credit: Vaginal Microbiome Consortium.

Synthesizing Across Studies

All-study differential prevalence analysis, Bayesian approach

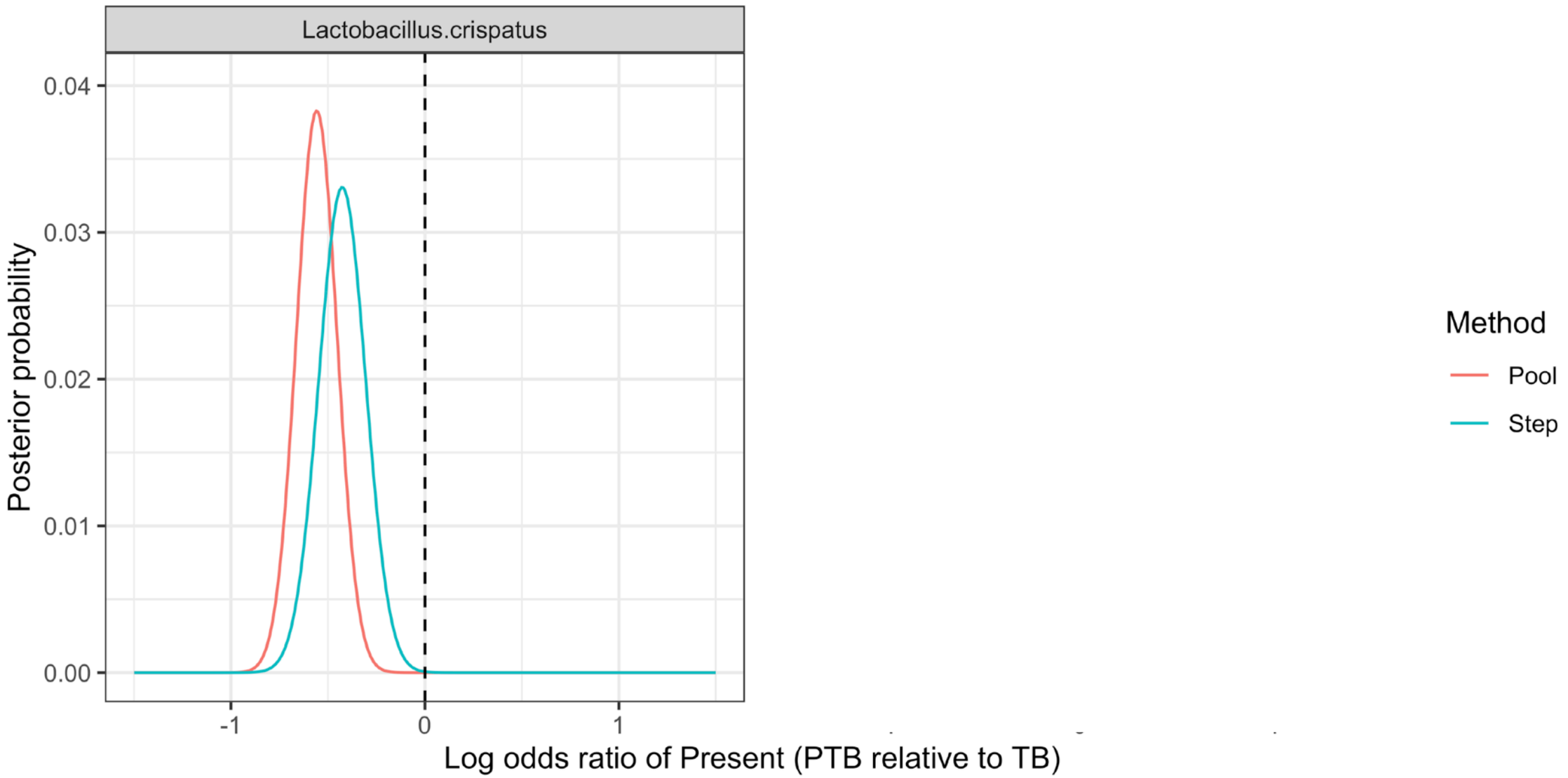
- **Pool**: Assume **same** detection rate in each study
- **Step**: Allow for **different** detection rates across studies

Huang et al., *bioRxiv*, 2022.

Synthesizing Across Studies

All-study differential prevalence analysis, Bayesian approach

- **Pool**: Assume **same** detection rate in each study
- **Step**: Allow for **different** detection rates across studies

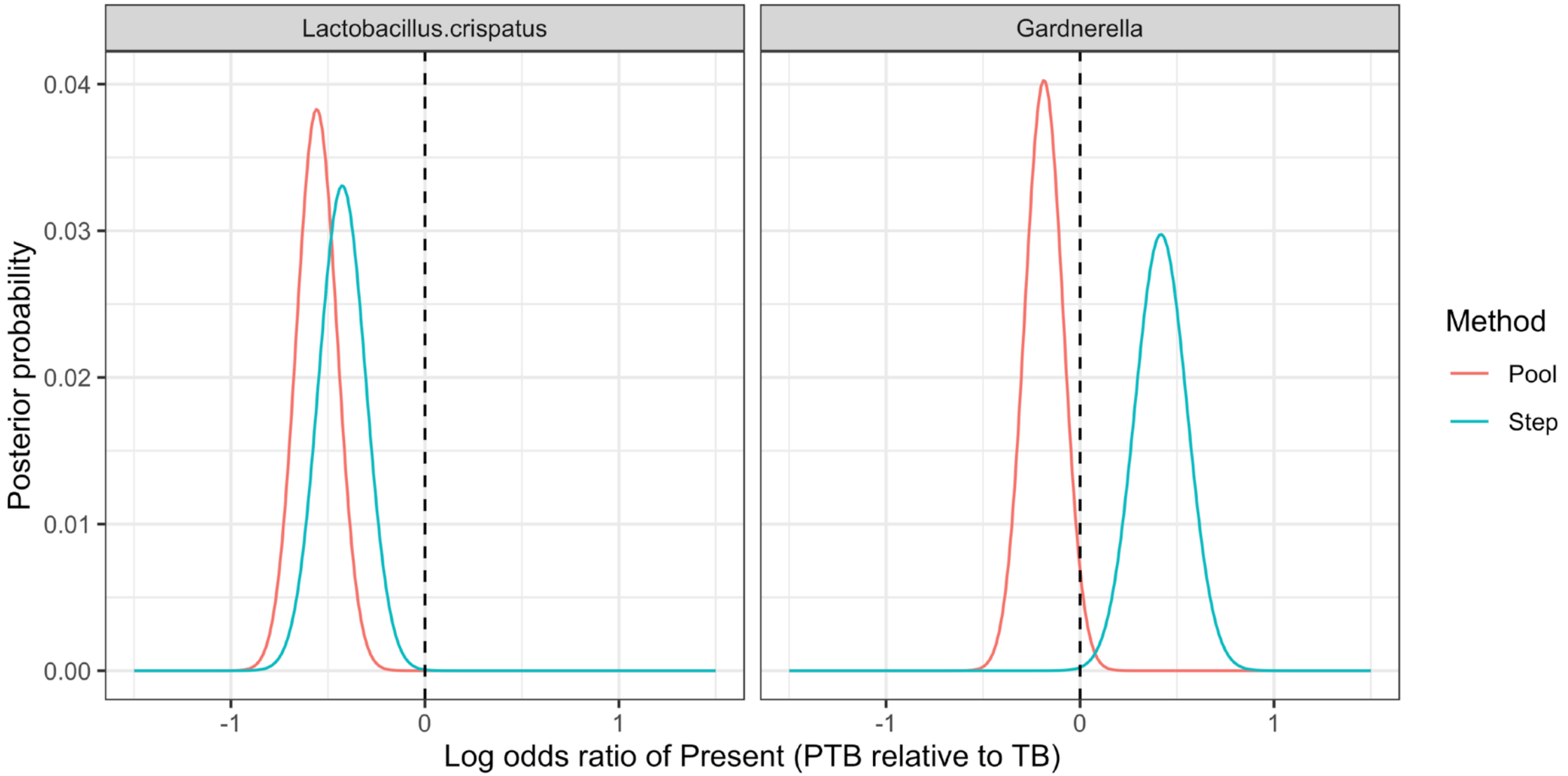


Huang et al., *bioRxiv*, 2022.

Synthesizing Across Studies

All-study differential prevalence analysis, Bayesian approach

- **Pool**: Assume **same** detection rate in each study
- **Step**: Allow for **different** detection rates across studies



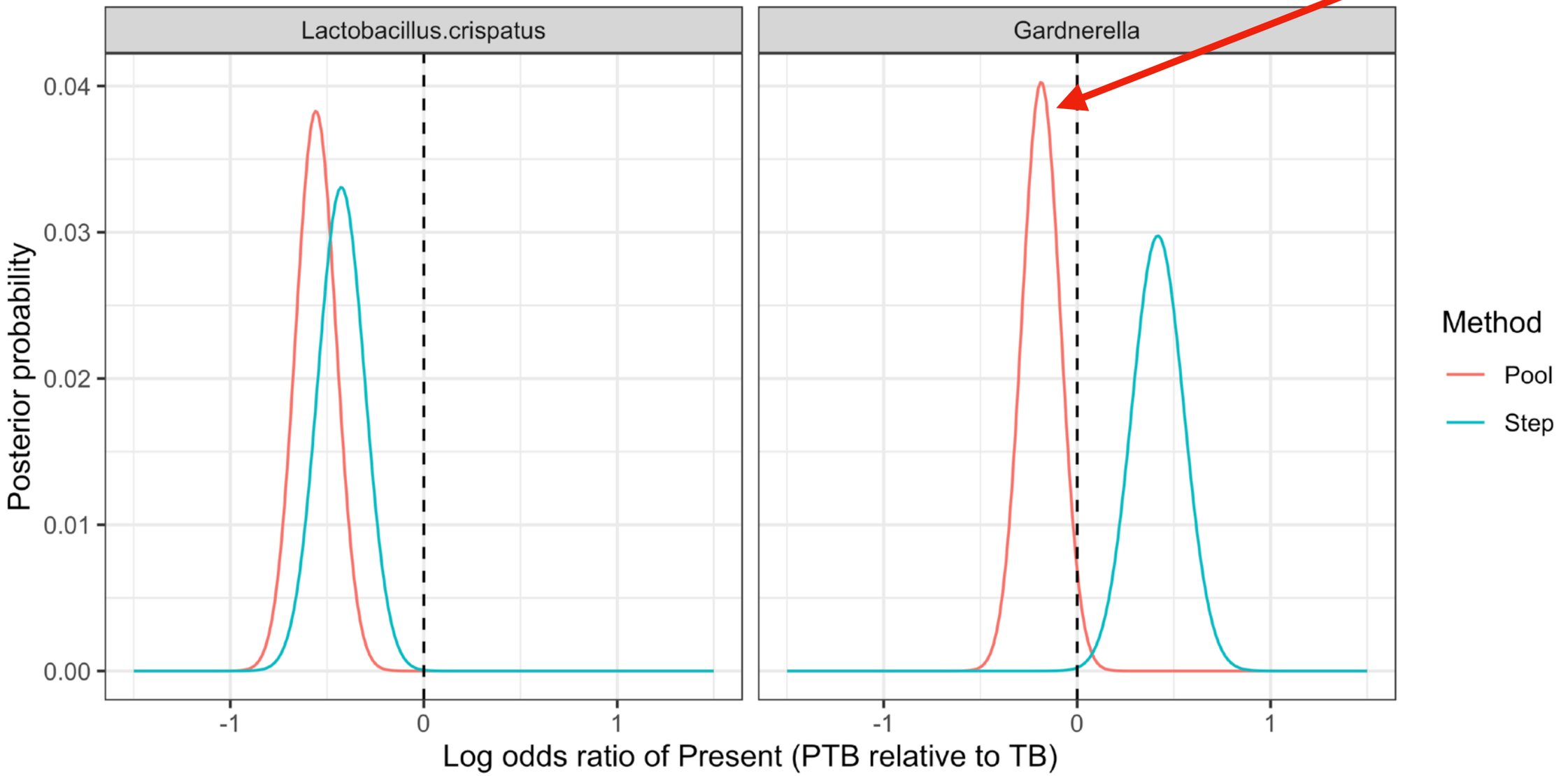
Huang et al., *bioRxiv*, 2022.

Synthesizing Across Studies

All-study differential prevalence analysis, Bayesian approach

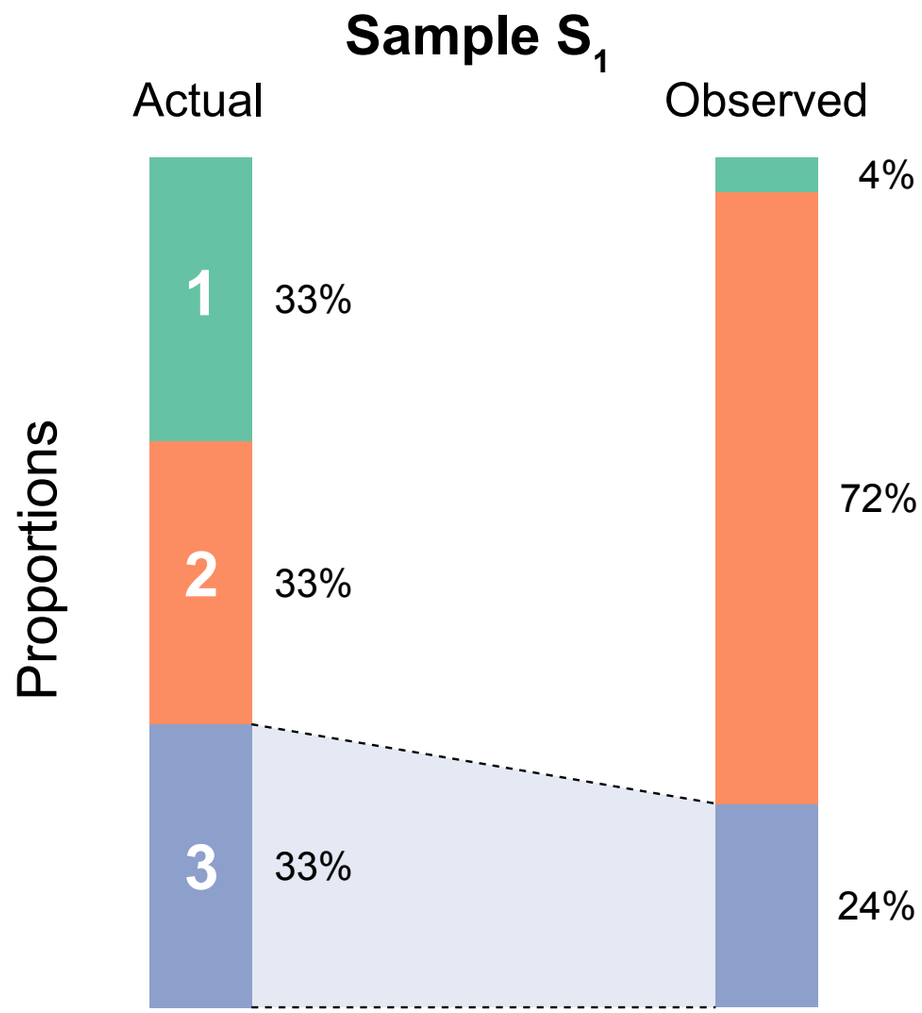
- **Pool**: Assume **same** detection rate in each study
- **Step**: Allow for **different** detection rates across studies

Sign Error!



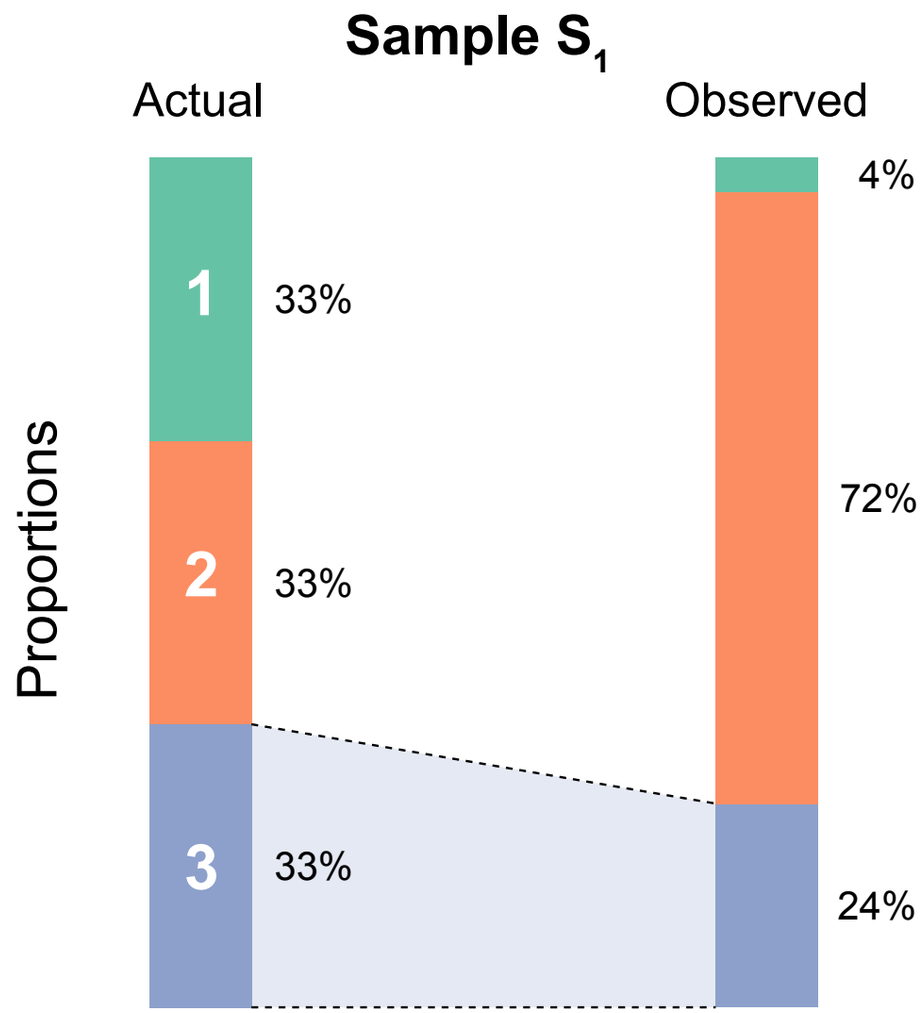
Huang et al., *bioRxiv*, 2022.

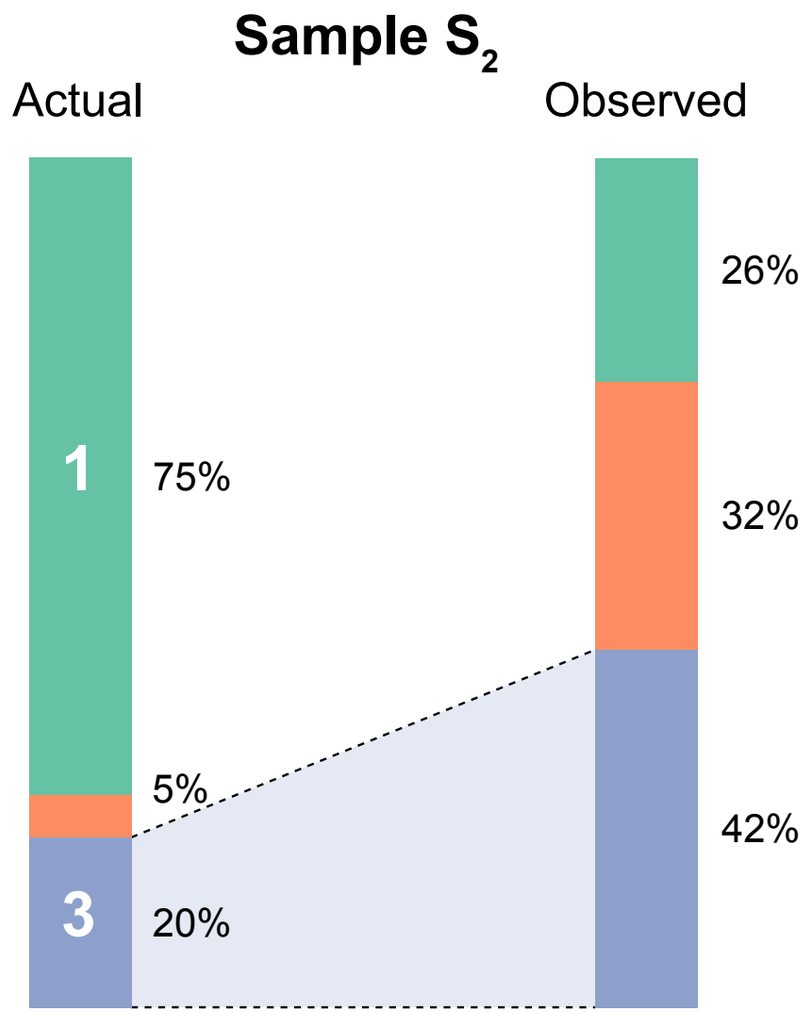
Better estimators for diff-abund



McLaren, Willis & Callahan. *eLife*, 2019.

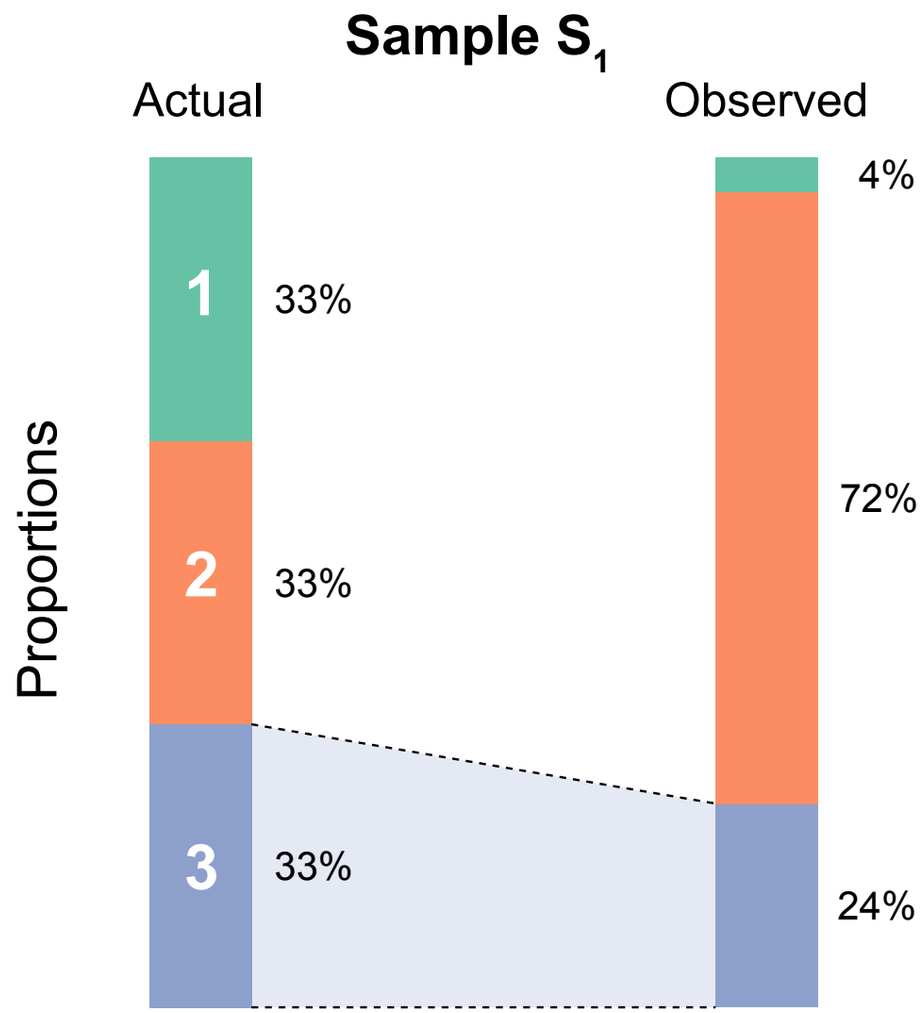
Better estimators for diff-abund

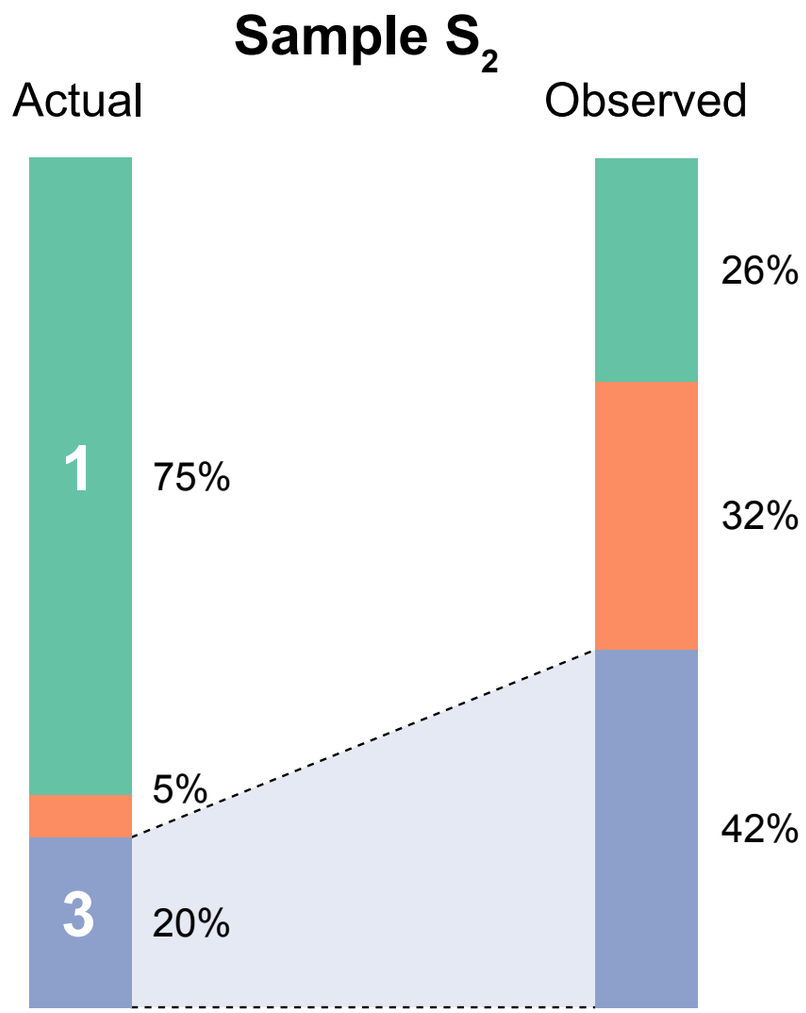




McLaren, Willis & Callahan. *eLife*, 2019.

Better estimators for diff-abund





Ratios to **Taxon 1**

$$\underset{\mathbf{A}(\mathrm{S}_1)}{\begin{pmatrix}1\\1\\1\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \begin{pmatrix}1\\18\\6\end{pmatrix} \sim \underset{\mathbf{O}(\mathrm{S}_1)}{\begin{pmatrix}0.04\\0.72\\0.24\end{pmatrix}}$$

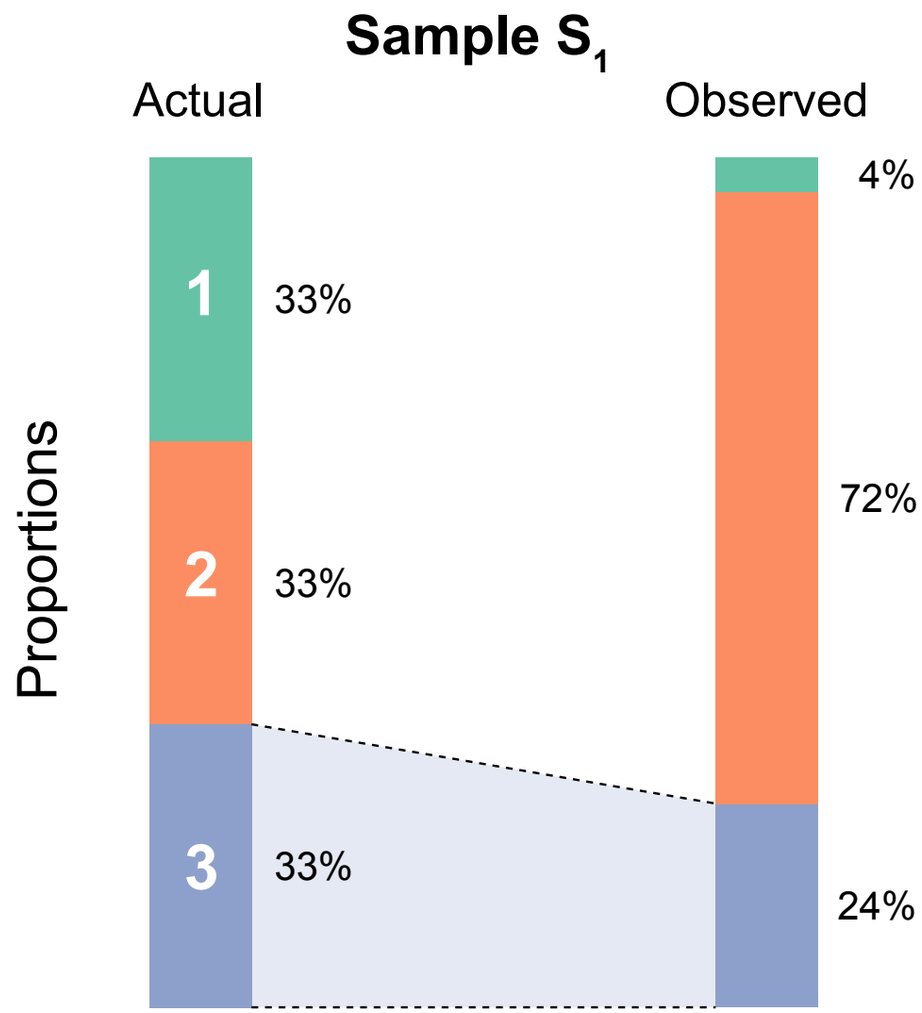
Actual · Bias = Observed

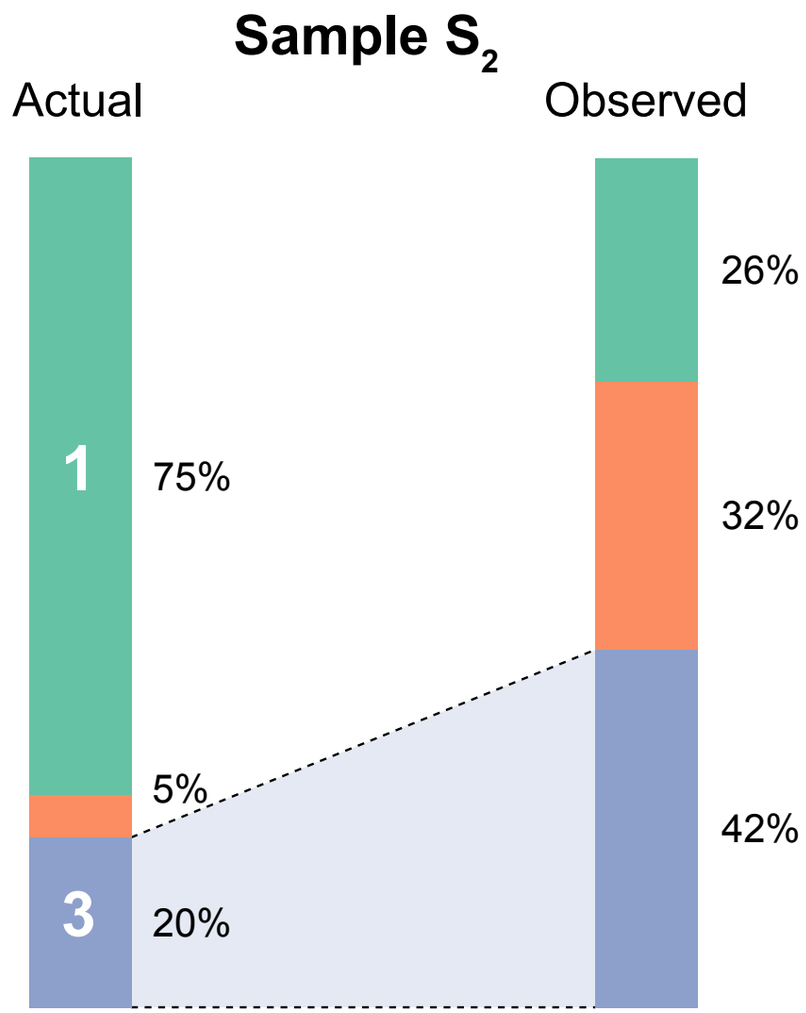
$$\underset{\mathbf{A}(\mathrm{S}_2)}{\begin{pmatrix}1\\1/15\\4/15\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \begin{pmatrix}1\\18/15\\24/15\end{pmatrix} \sim \underset{\mathbf{O}(\mathrm{S}_2)}{\begin{pmatrix}0.26\\0.32\\0.42\end{pmatrix}}$$

Actual · Bias = Observed

McLaren, Willis & Callahan. *eLife*, 2019.

Better estimators for diff-abund





Ratios to **Taxon 1**

$$\underset{\mathbf{A}(S_1)}{\begin{pmatrix}1\\1\\1\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_1)}{\begin{pmatrix}1\\18\\6\end{pmatrix}} \sim \begin{pmatrix}0.04\\0.72\\0.24\end{pmatrix}$$

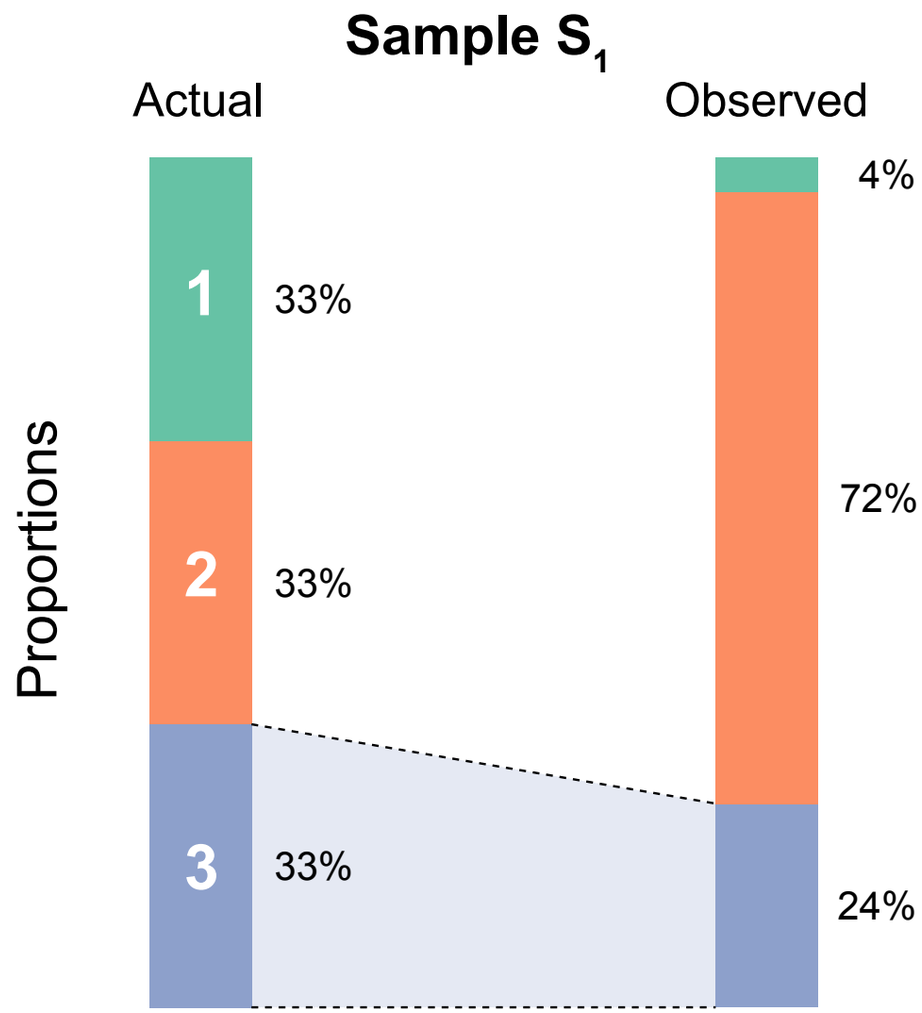
Actual · Bias · Observed

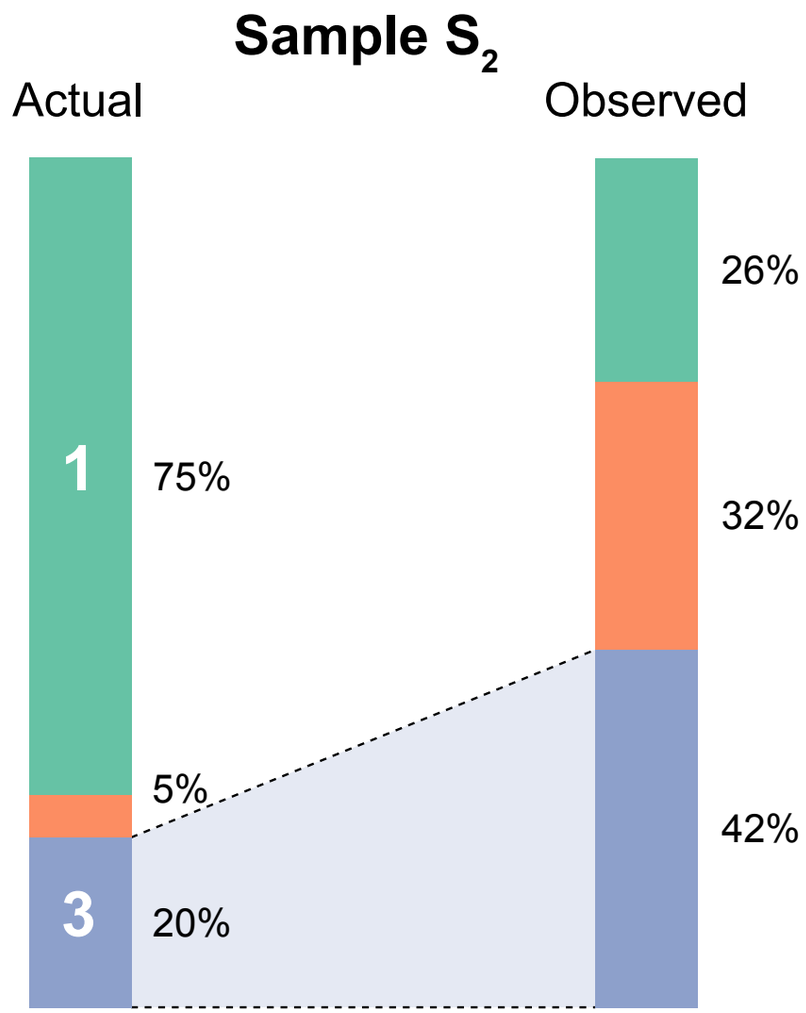
$$\underset{\mathbf{A}(S_2)}{\begin{pmatrix}1\\1/15\\4/15\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_2)}{\begin{pmatrix}1\\18/15\\24/15\end{pmatrix}} \sim \begin{pmatrix}0.26\\0.32\\0.42\end{pmatrix}$$

Actual · Bias · Observed

McLaren, Willis & Callahan. *eLife*, 2019.

Better estimators for diff-abund





Ratios to **Taxon 1**

$$\underset{\mathbf{A}(S_1)}{\begin{pmatrix}1\\1\\1\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_1)}{\begin{pmatrix}1\\18\\6\end{pmatrix}} \sim \begin{pmatrix}0.04\\0.72\\0.24\end{pmatrix}$$

Actual · Bias · Observed

$$\underset{\mathbf{A}(S_2)}{\begin{pmatrix}1\\1/15\\4/15\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_2)}{\begin{pmatrix}1\\18/15\\24/15\end{pmatrix}} \sim \begin{pmatrix}0.26\\0.32\\0.42\end{pmatrix}$$

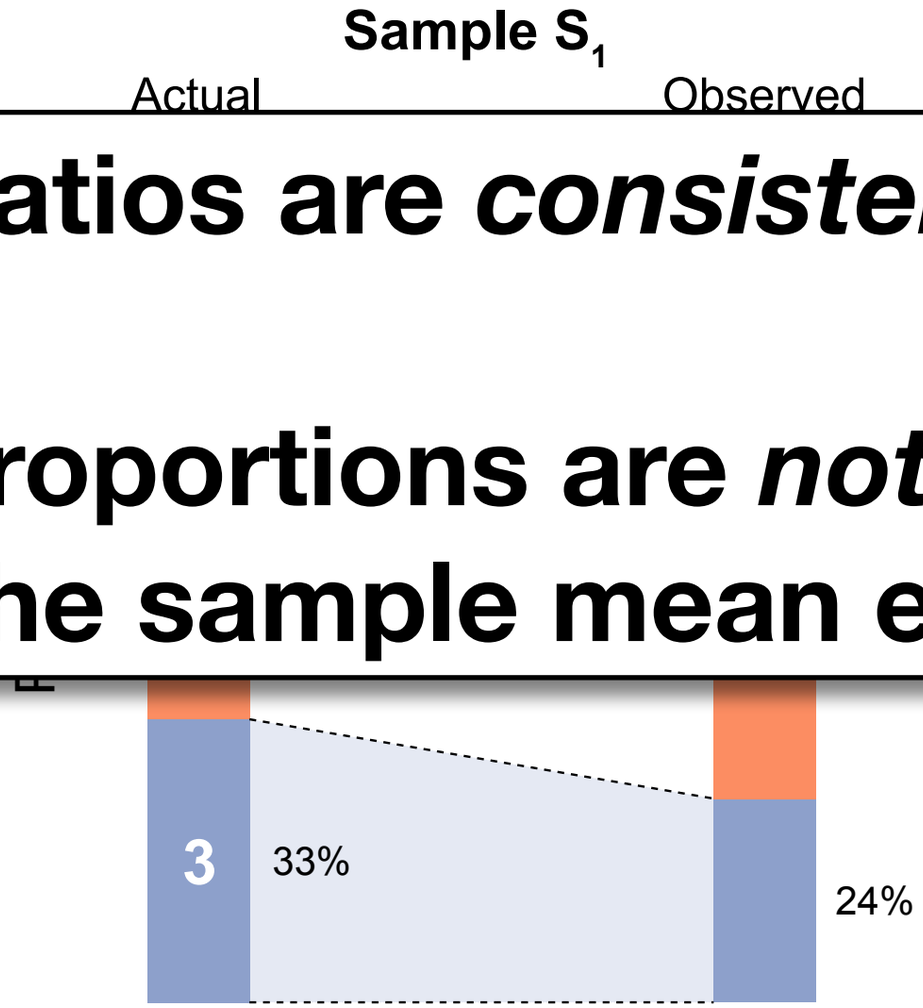
Actual · Bias · Observed

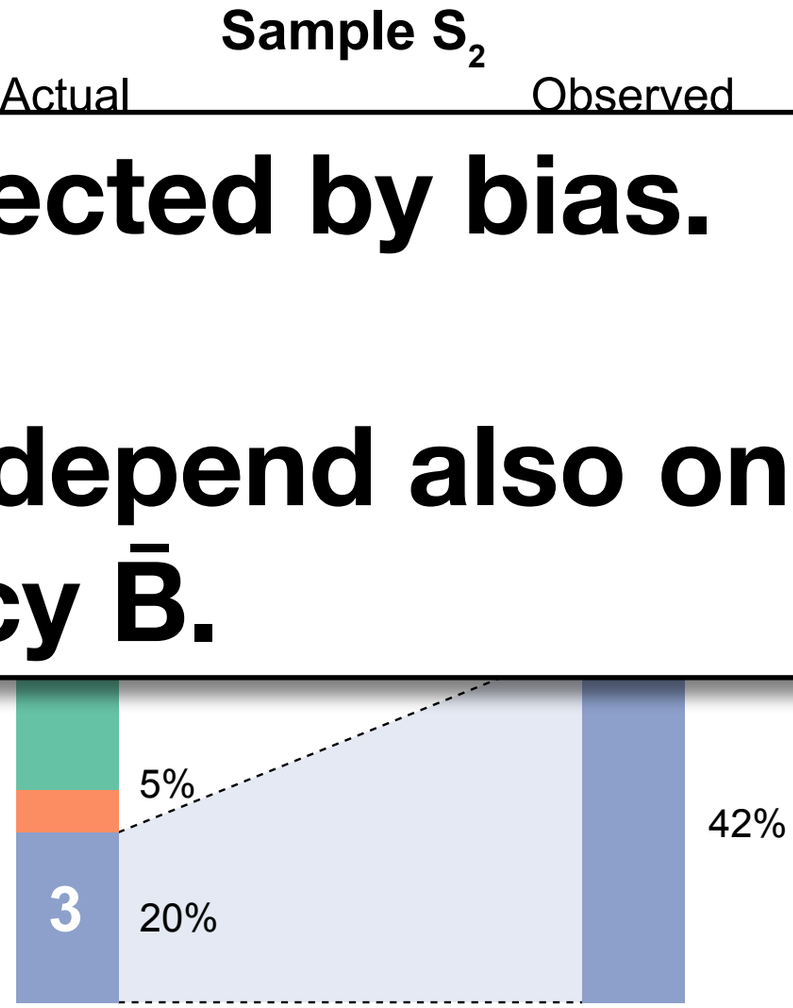
McLaren, Willis & Callahan. *eLife*, 2019.

Better estimators for diff-abund

Ratios are *consistently* affected by bias.

Proportions are *not*. They depend also on the sample mean efficiency $\bar{B}$.





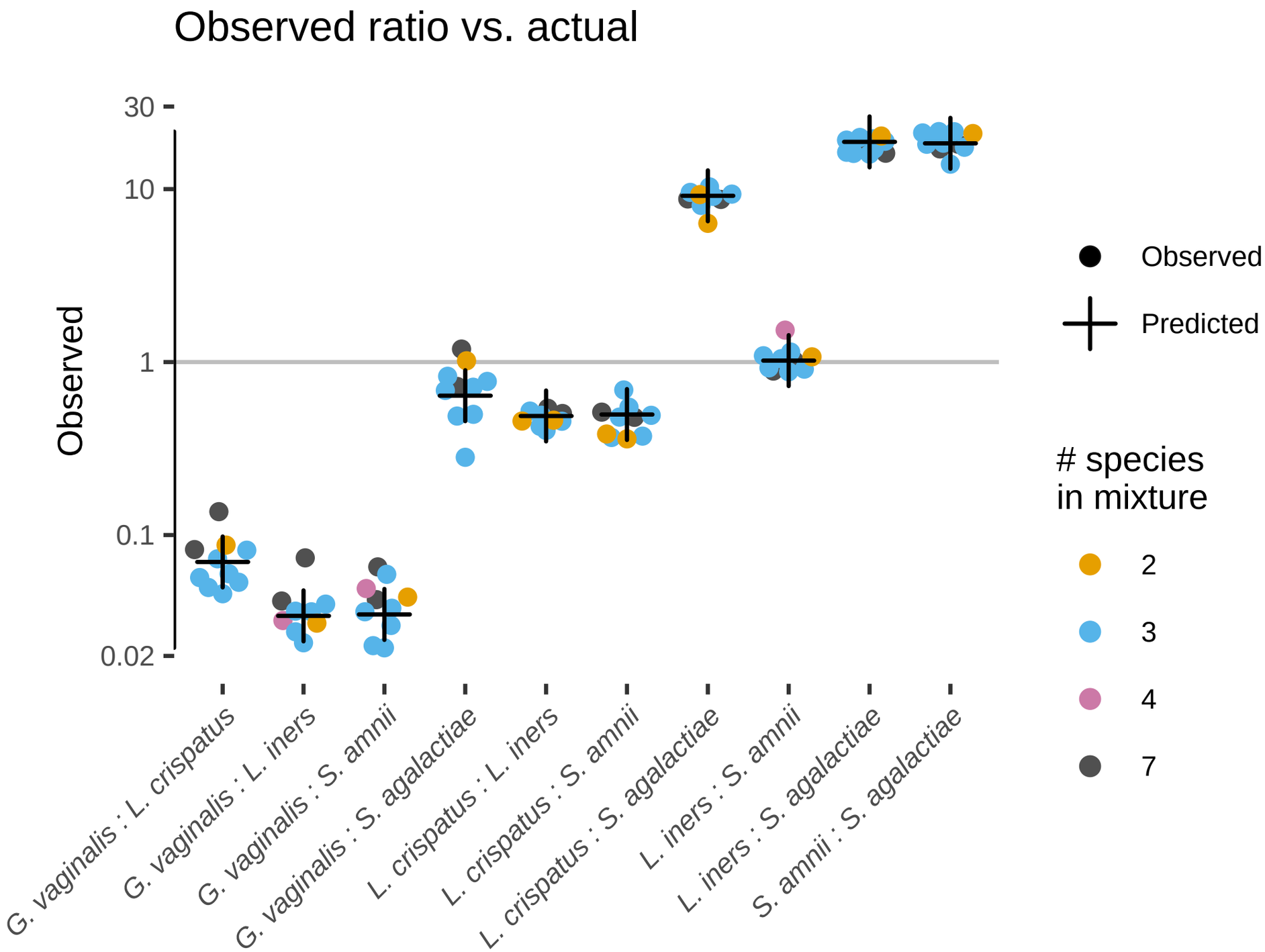
Ratios to **Taxon 1**

$$\underset{\mathbf{A}(S_1)}{\begin{pmatrix}1\\1\\1\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_1)}{\begin{pmatrix}1\\18\\6\end{pmatrix}} \sim \begin{pmatrix}0.04\\0.72\\0.24\end{pmatrix}$$

$$\underset{\mathbf{A}(S_2)}{\begin{pmatrix}1\\1/15\\4/15\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_2)}{\begin{pmatrix}1\\18/15\\24/15\end{pmatrix}} \sim \begin{pmatrix}0.26\\0.32\\0.42\end{pmatrix}$$

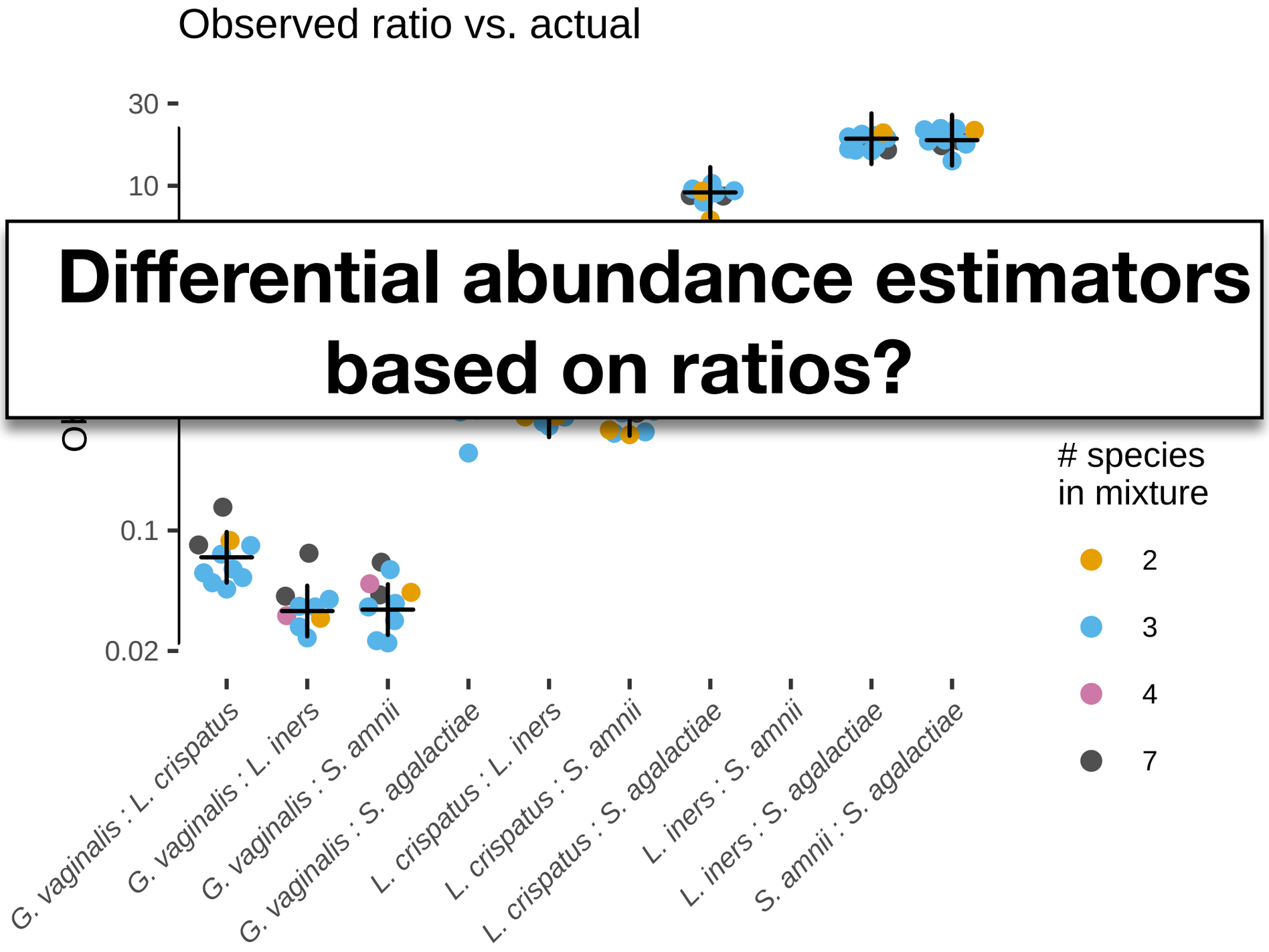
McLaren, Willis & Callahan. *eLife*, 2019.

Better estimators for diff-abund



McLaren, Willis & Callahan. *eLife*, 2019.

Better estimators for diff-abund

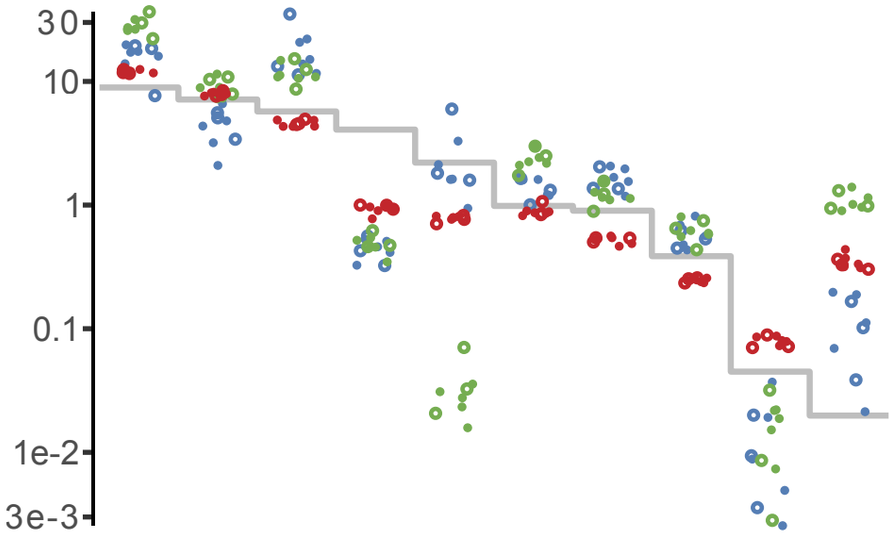


Differential abundance estimators based on ratios?

McLaren, Willis & Callahan. *eLife*, 2019.

Metagenomic Calibration

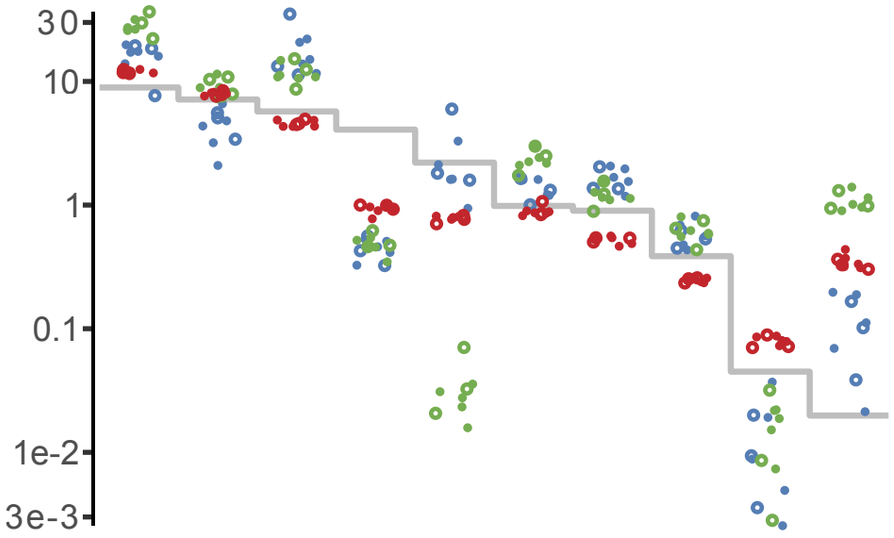
Observed vs. True Abundances



Data: Costea, et al. *Nature Biotechnology*, 2017.

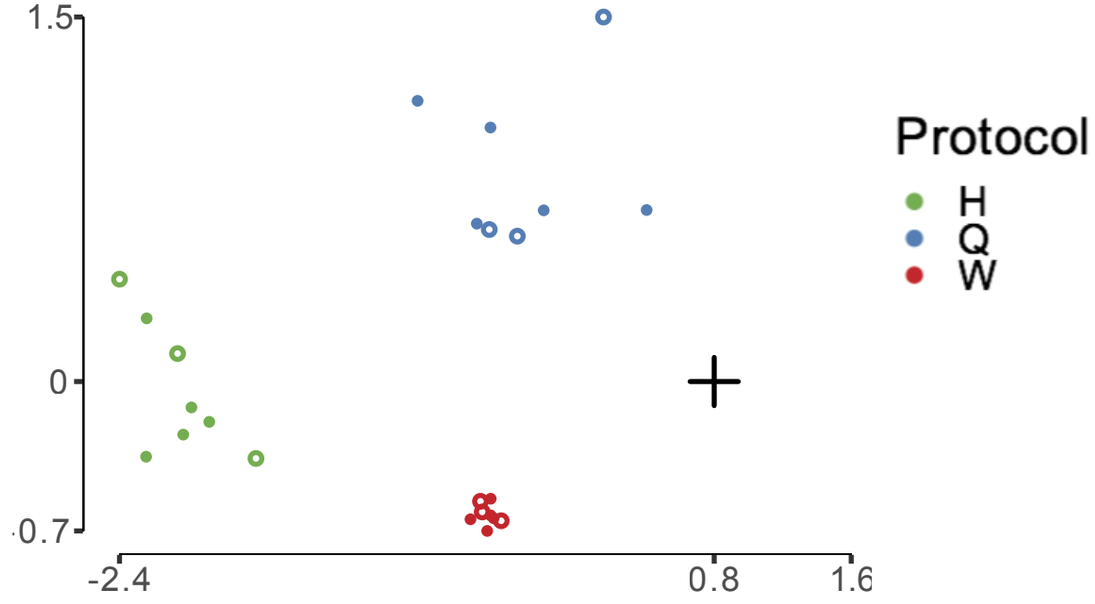
Metagenomic Calibration

Observed vs. True Abundances



Biased

Sample Ordination

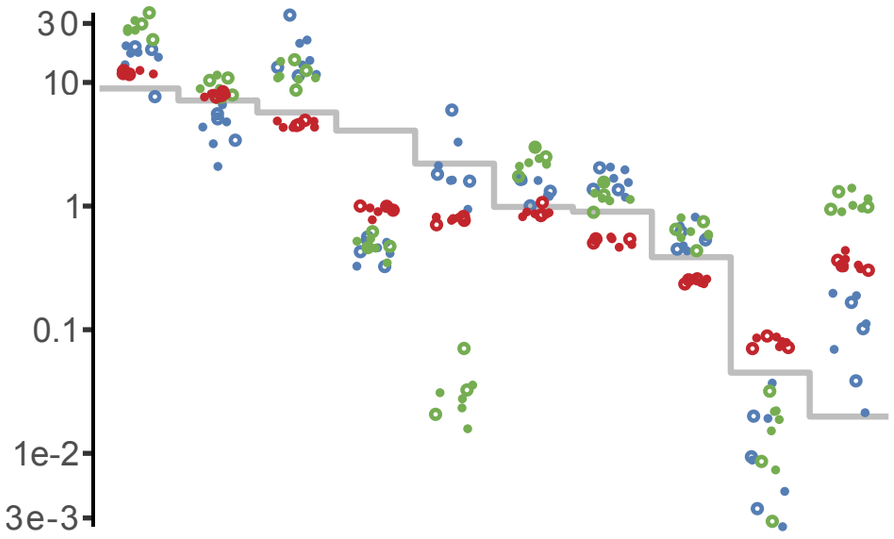


Protocol
H
Q
W

Data: Costea, et al. *Nature Biotechnology*, 2017.

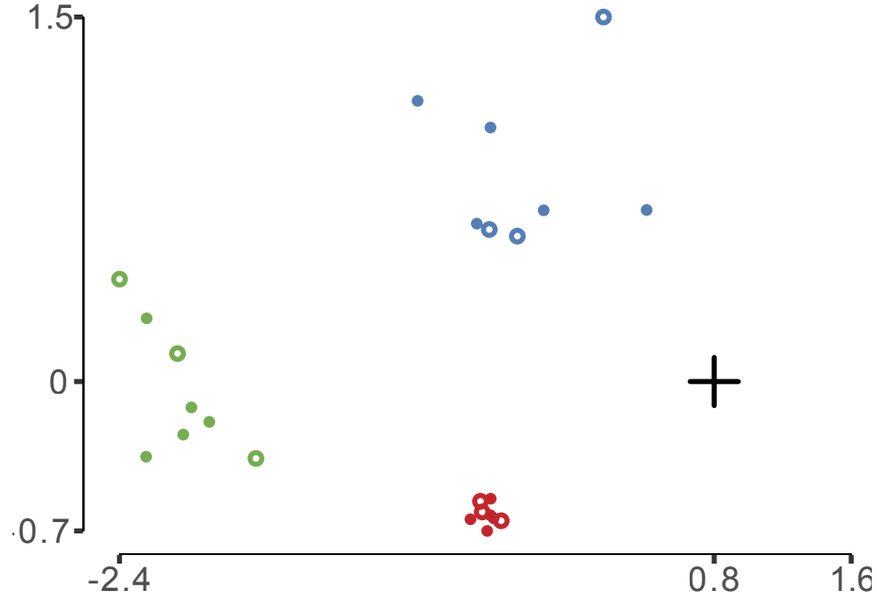
Metagenomic Calibration

Observed vs. True Abundances



Biased

Sample Ordination



Calibration

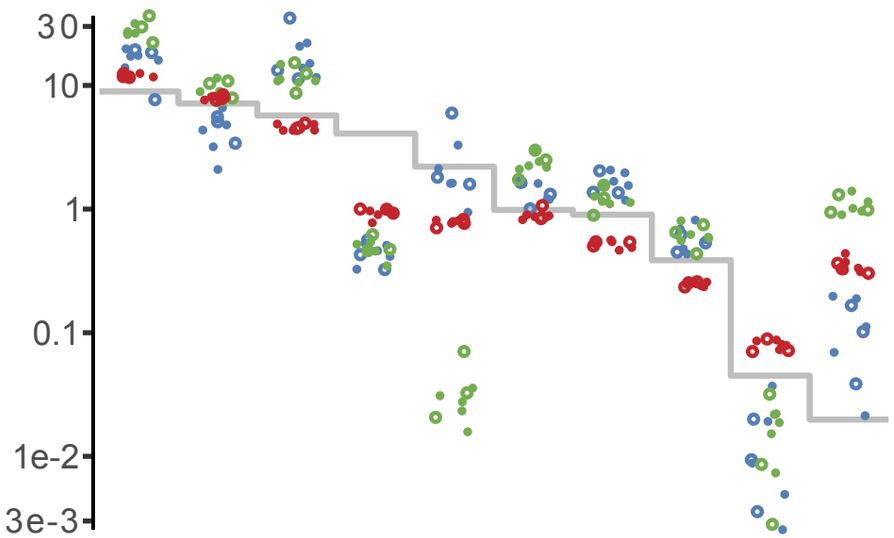
1. Measure control samples with known composition
2. Estimate bias from controls (B = O/A)
3. Use estimates bias to correct observations

Protocol
- H
- Q
- W

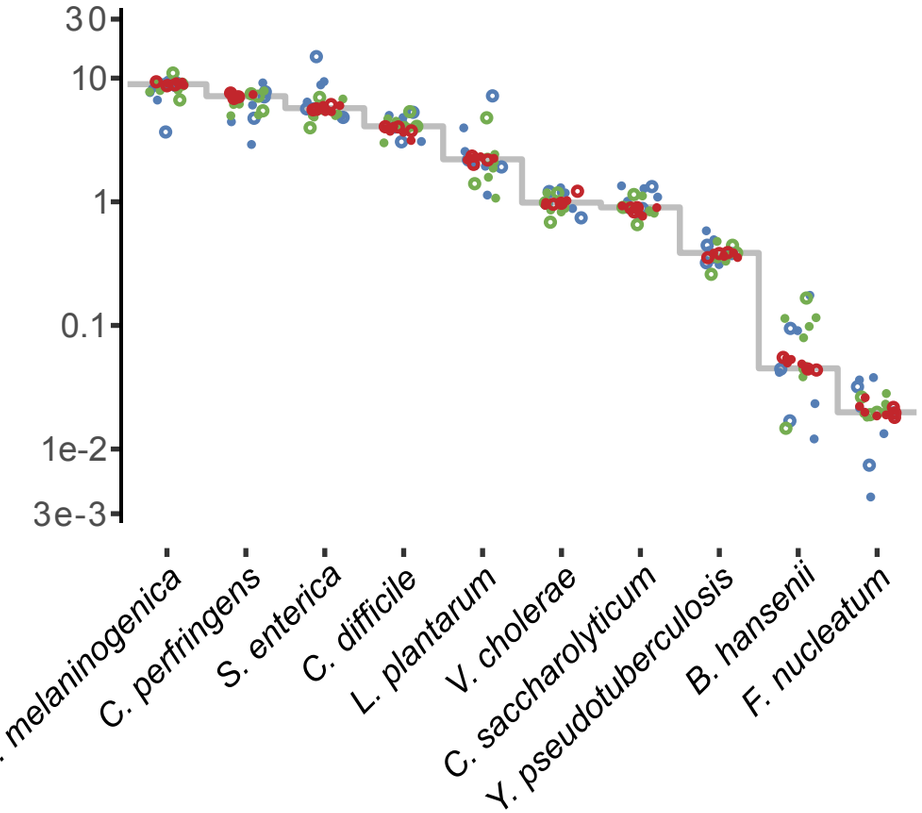
Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Calibration

Observed vs. True Abundances

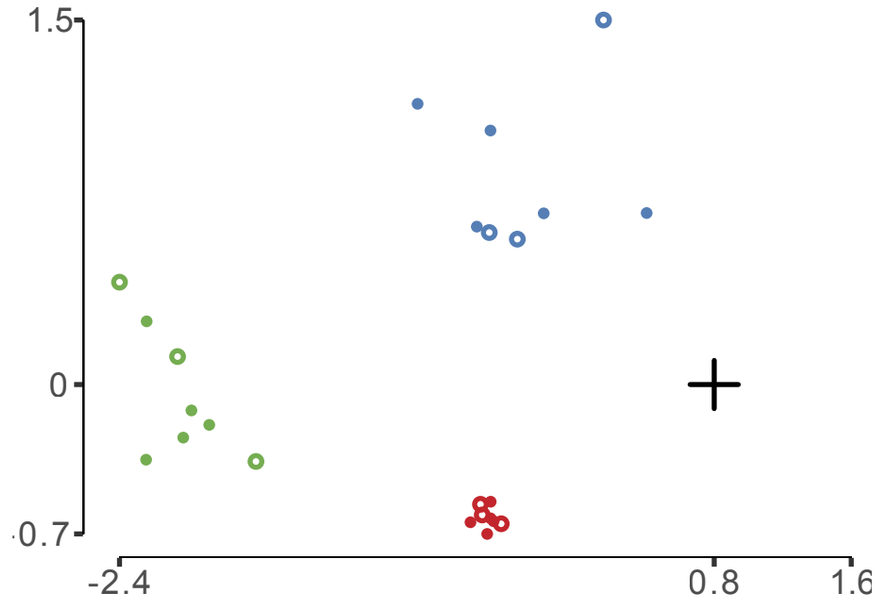


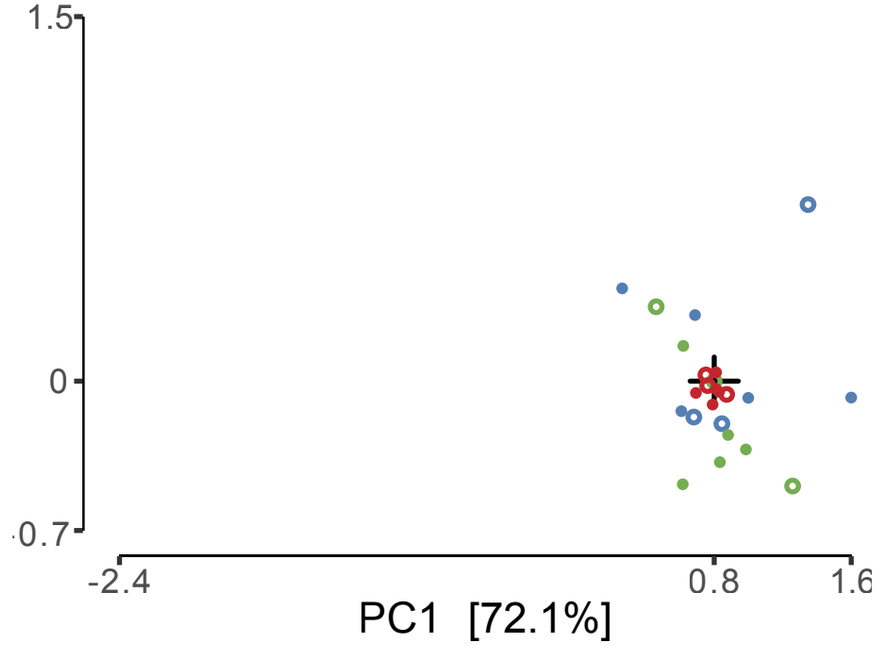
Biased



Calibrated

Sample Ordination

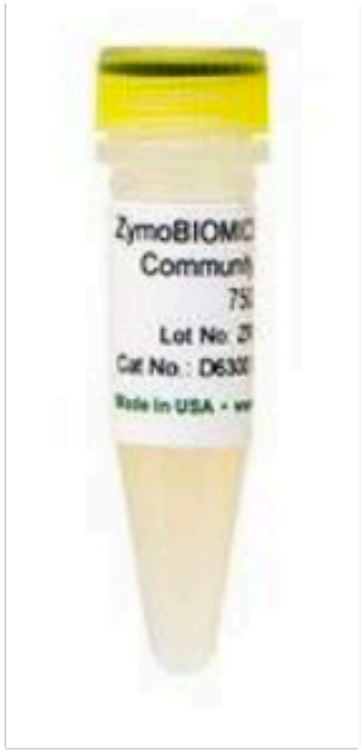




Protocol
- H
- Q
- W

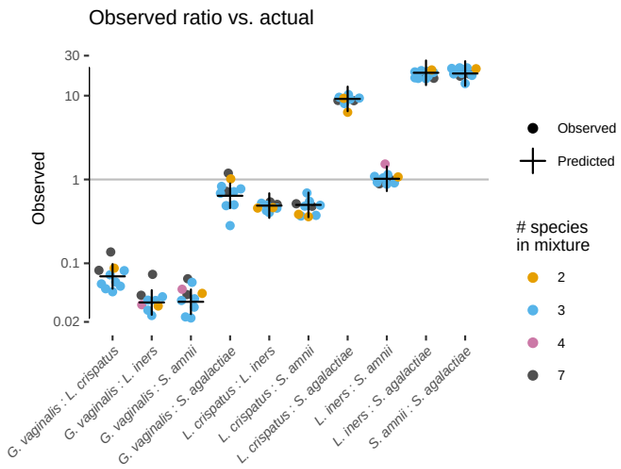
Data: Costea, et al. *Nature Biotechnology*, 2017.

Towards True Calibration?



Measure with multiple protocols

Estimate differential bias



Compare

Measure with multiple protocols

Estimate differential bias

Some thoughts

All measurements are wrong, but some are useful.

- (apologies to G.E.P. Box)

New opportunities from measurement models.

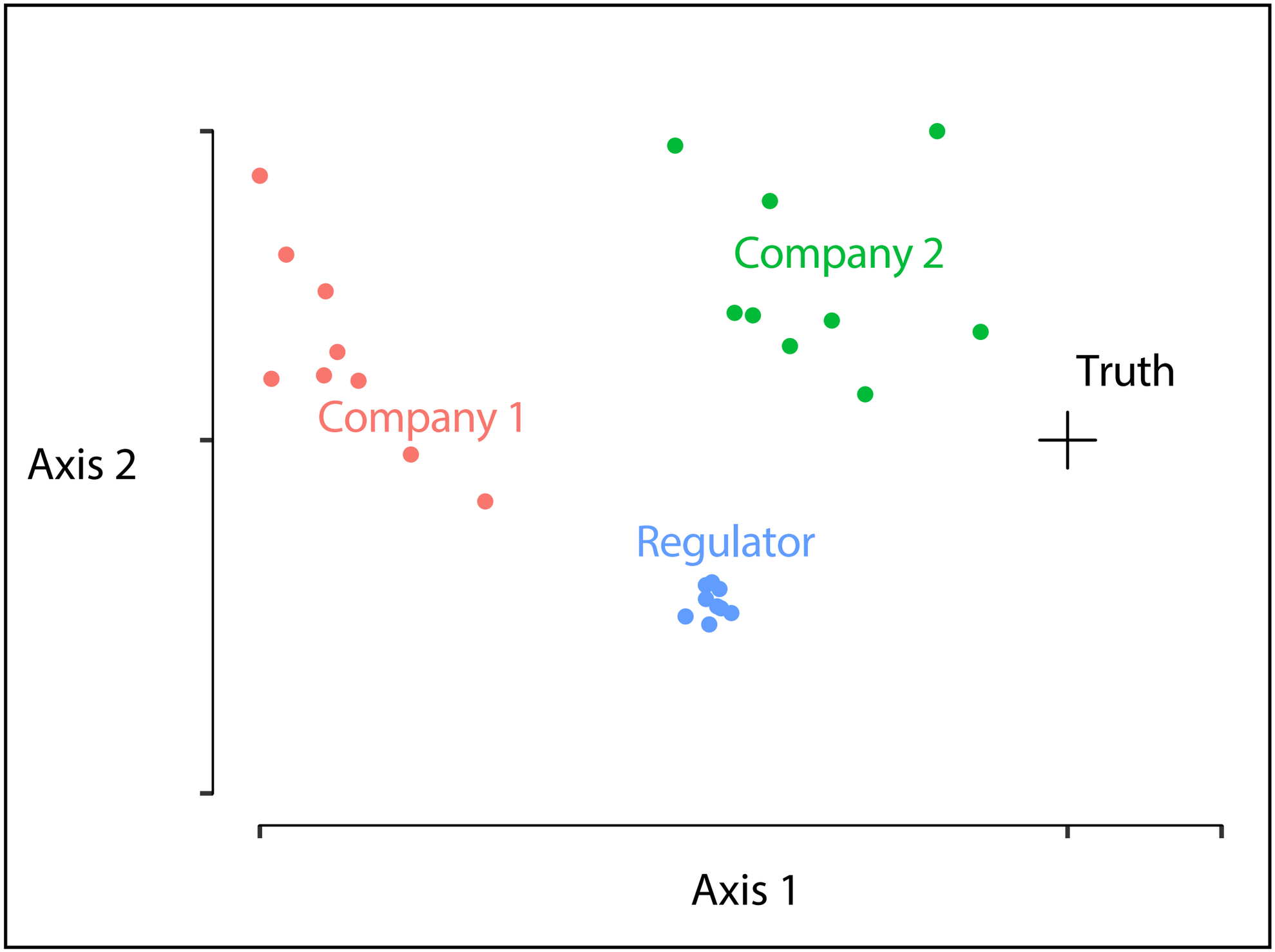
- Standard samples that are more than a process control?

What are the right units (e.g. genomes vs. cells vs. biomass…)?

- That matter? That can be consistently measured?

What estimates can we make? Should we make?

- That are robust to the realities of our measurements?



Michael McLaren

Amy Willis

Caizhi Huang

Jacob Nearing

Karen Lloyd

Manuel Kleiner

Thank You!

NIH NIGMS

USDA NIFA

Resources

Consistent and correctable bias in metagenomic sequencing experiments

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

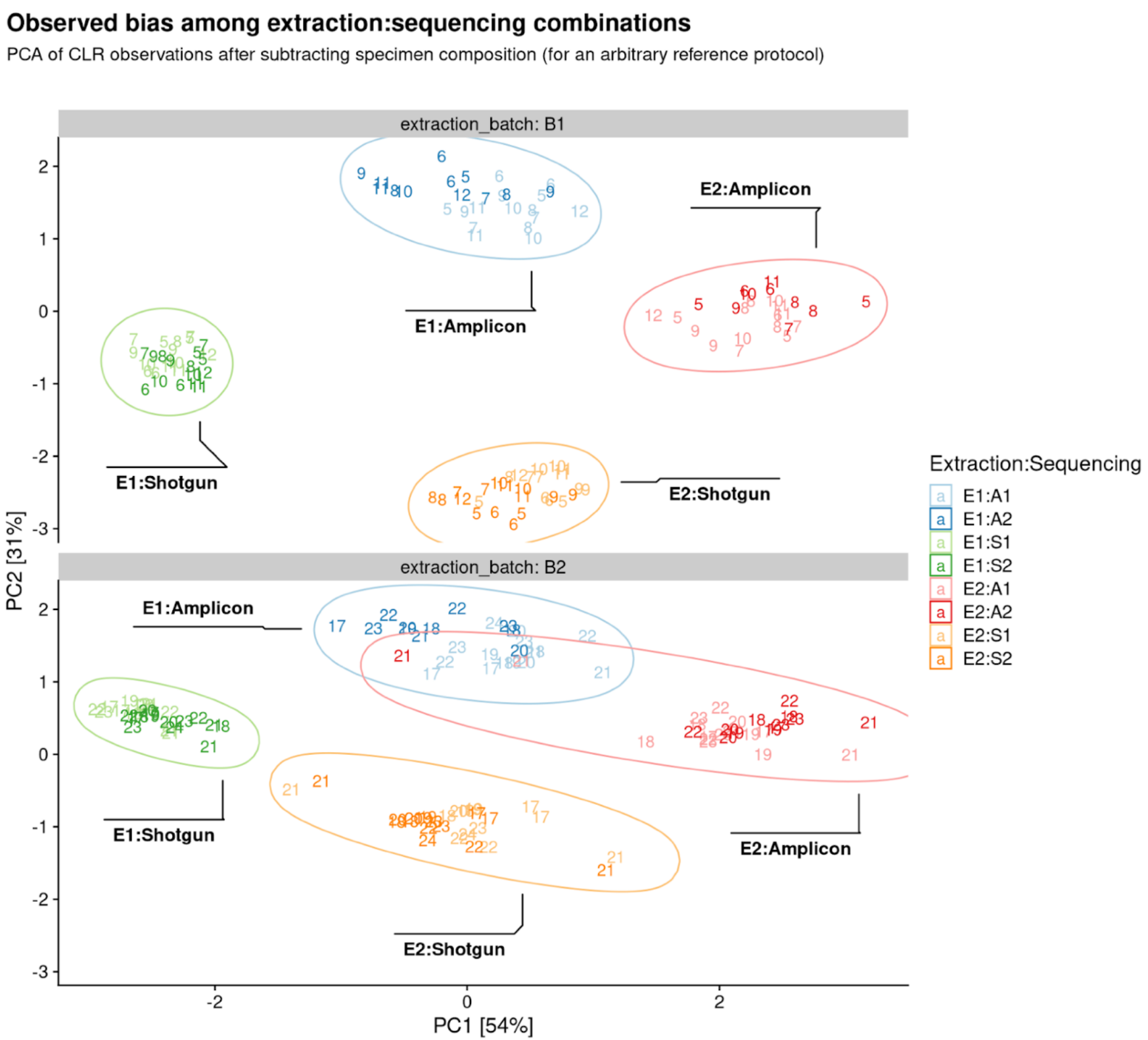
eLIFE

McLaren, Nearing, Willis, Lloyd, Callahan (2022). “Implications of taxonomic bias for microbial differential-abundance analysis”. *bioRxiv*. https://doi.org/10.1101/2022.08.19.504330

Williamson, Hughes, Willis (2021). “A multiview model for relative and absolute microbial abundances”. *Biometrics*, (2021). https://doi.org/10.1111/biom.13503

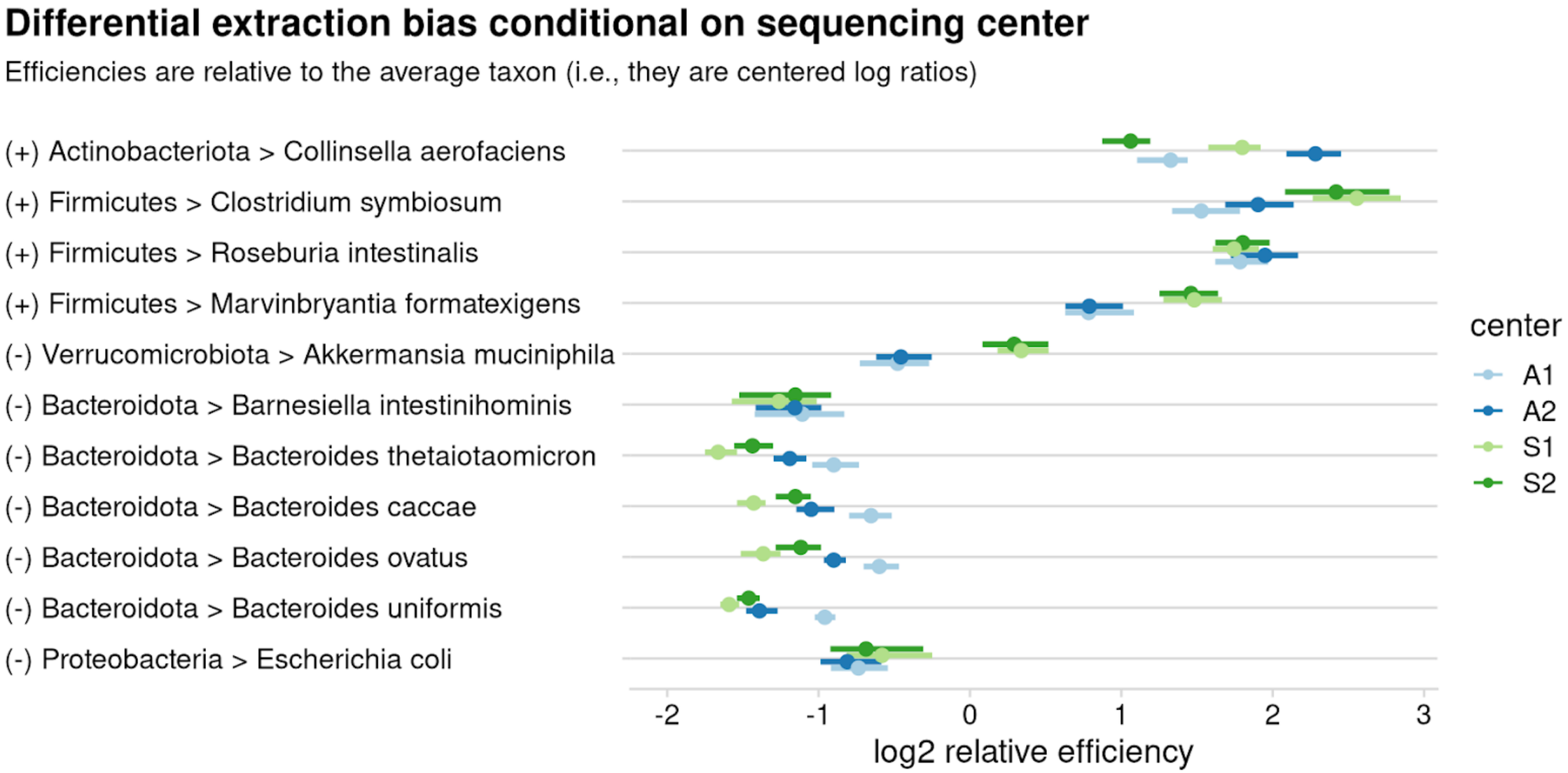
metacal: R package for Metagenomics calibration (2022). *Github*. https://doi.org/10.5281/zenodo.4380996

Towards True Calibration?



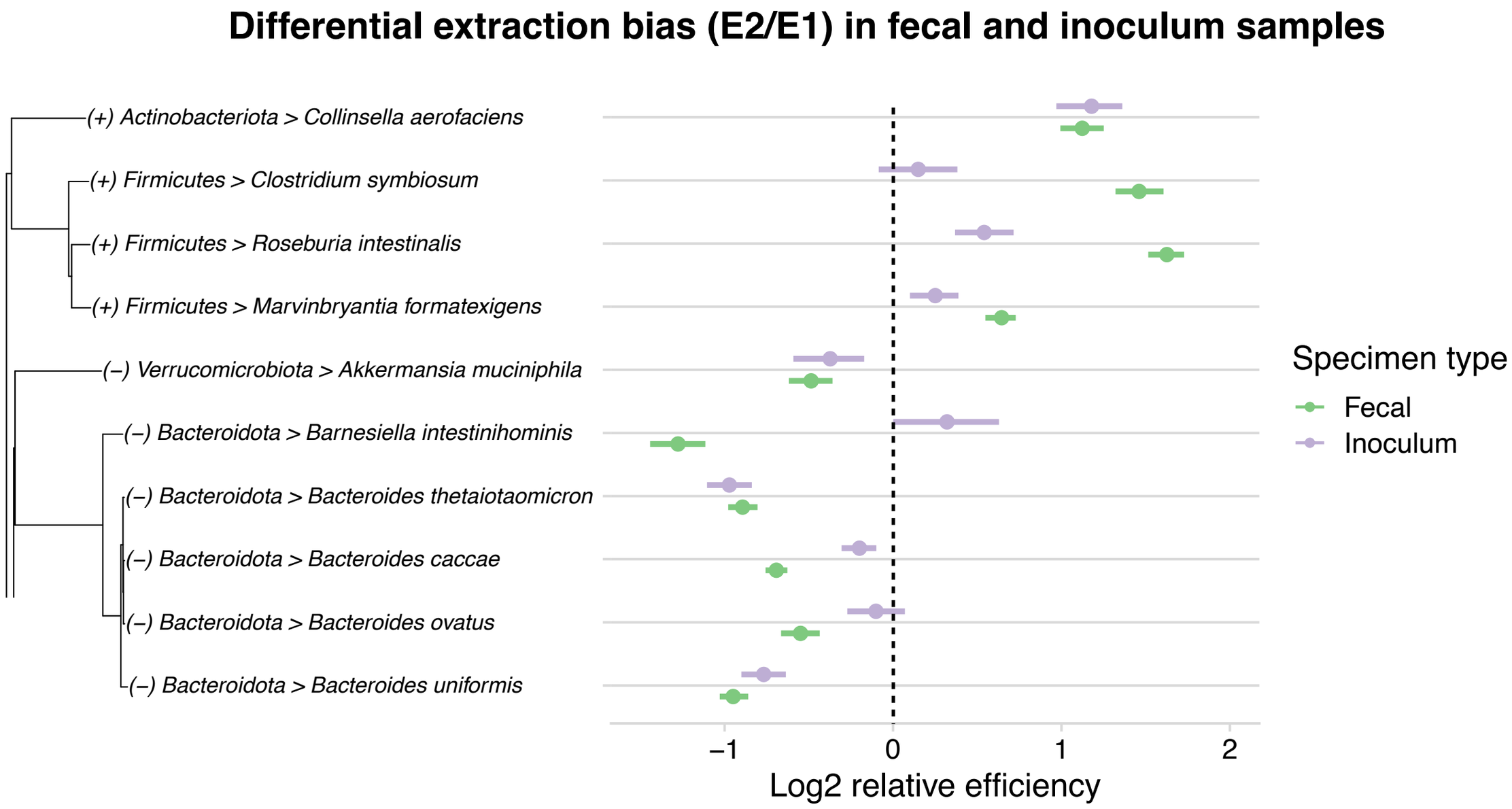
Acknowledgement: Michael McLaren, Angie Mordant, Manuel Kleiner.

Towards True Calibration?



Acknowledgement: Michael McLaren, Angie Mordant, Manuel Kleiner.

A note of Caution



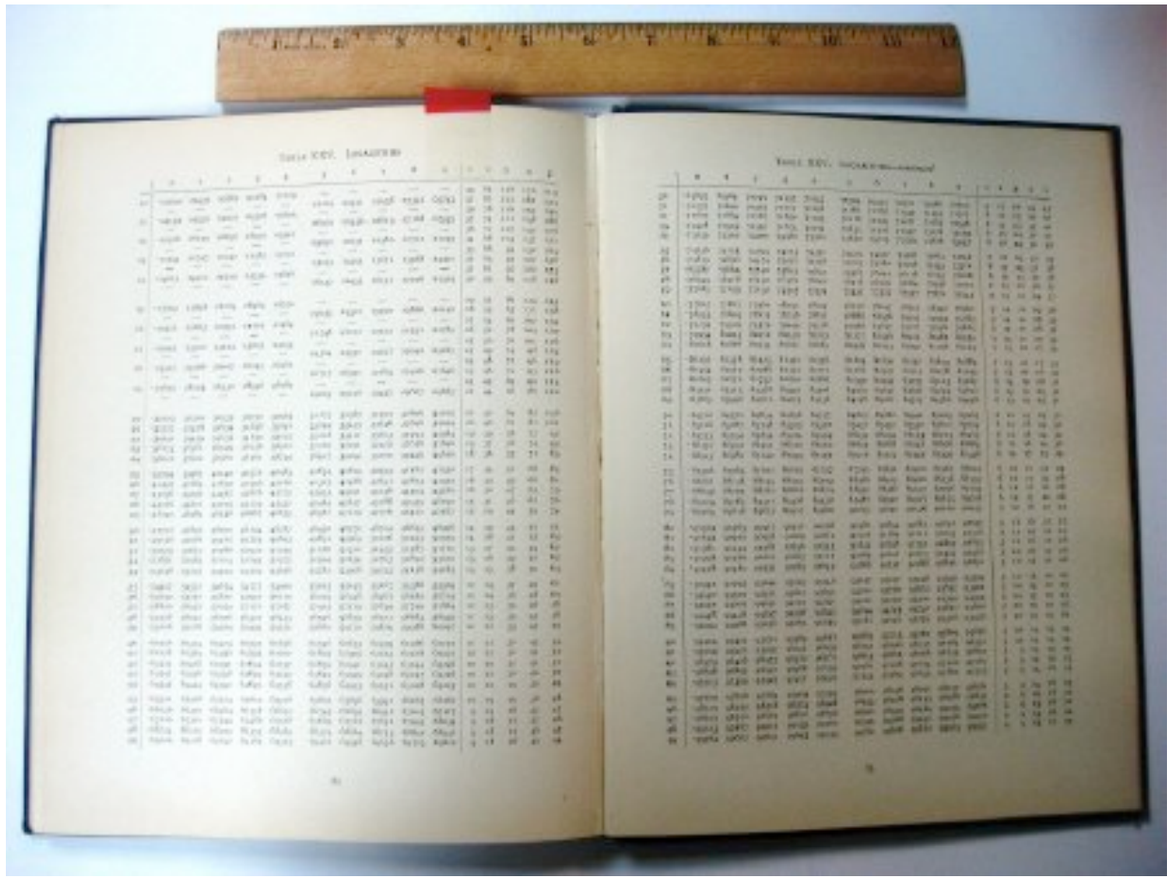
Acknowledgement: Michael McLaren, Angie Mordant, Manuel Kleiner.

Quantitative Qs about B(ias)

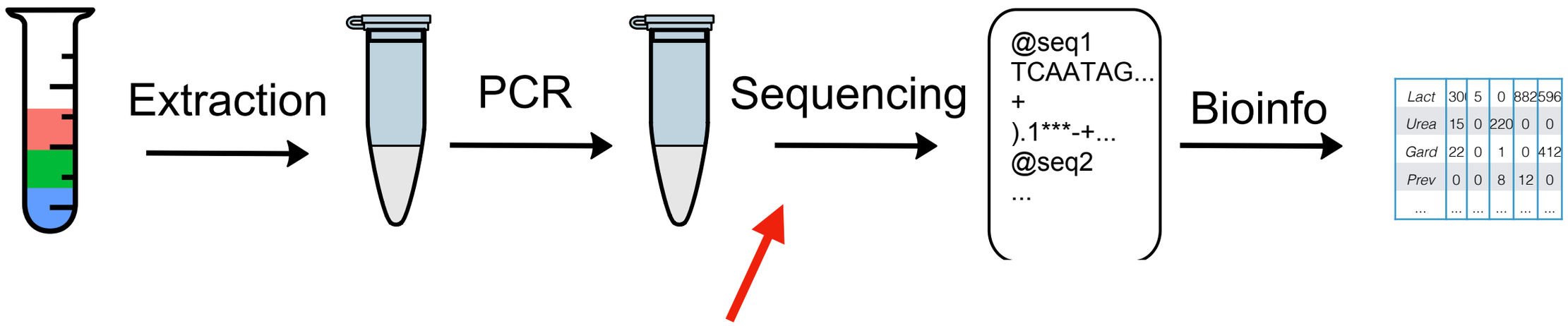
- Scale? (within and between protocols)
- Phylogenetic coherence?
- Predictability?

Quantitative Qs about B(ias)

- Scale? (within and between protocols)
- Phylogenetic coherence?
- Predictability?



Compositionality



Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*], Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]

[1]Department of Biochemistry, University of Western Ontario, London, ON, Canada
[2]Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain
[3]Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

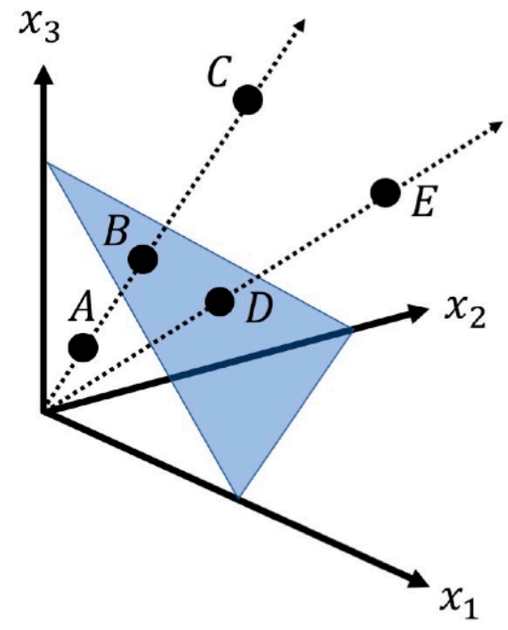


Image credit: Wikipedia.

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Credit: Aitchison, 1980s

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Log-ratio Transforms:

$$\mathrm{alr}(x) = \left[\log \frac{x_1}{x_D} \cdots \log \frac{x_{D-1}}{x_D}\right]$$

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)}\right]$$

$$\mathrm{ilr}(x) = \left[\langle x, e_1 \rangle, \ldots, \langle x, e_{D-1} \rangle\right]$$

Credit: Aitchison, 1980s

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Log-ratio Transforms:

$$\mathrm{alr}(x) = \left[\log \frac{x_1}{x_D} \cdots \log \frac{x_{D-1}}{x_D}\right]$$

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)}\right]$$

$$\mathrm{ilr}(x) = [\langle x, e_1 \rangle, \ldots, \langle x, e_{D-1} \rangle]$$

Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

Image Credit: Gloor et al., *Front. Microbiol.*, 2017

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Log-ratio Transforms:

$$\mathrm{alr}(x) = \left[\log \frac{x_1}{x_D} \cdots \log \frac{x_{D-1}}{x_D}\right]$$

$$\mathrm{clr}(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x)}\right]$$

$$\mathrm{ilr}(x) = [\langle x, e_1 \rangle, \ldots, \langle x, e_{D-1} \rangle]$$

How to marry theory and compositional measurements?

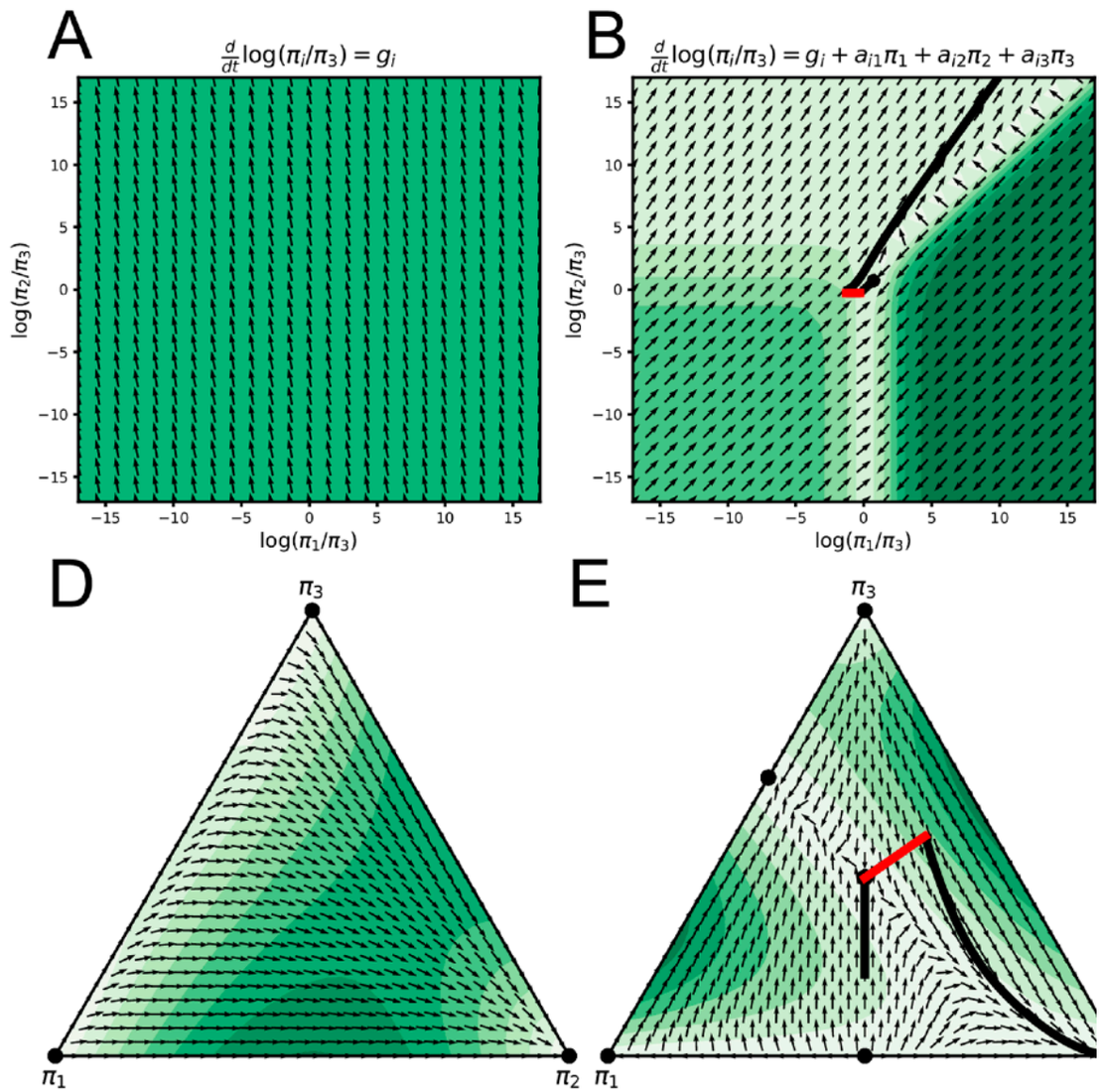
Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

Image Credit: Gloor et al., *Front. Microbiol.*, 2017

Compositional Modeling

Compositional Lotka-Volterra describes microbial dynamics in the simplex

Tyler A. Joseph, Liat Shenhav, Joao B. Xavier, Eran Halperin, Itsik Pe'er



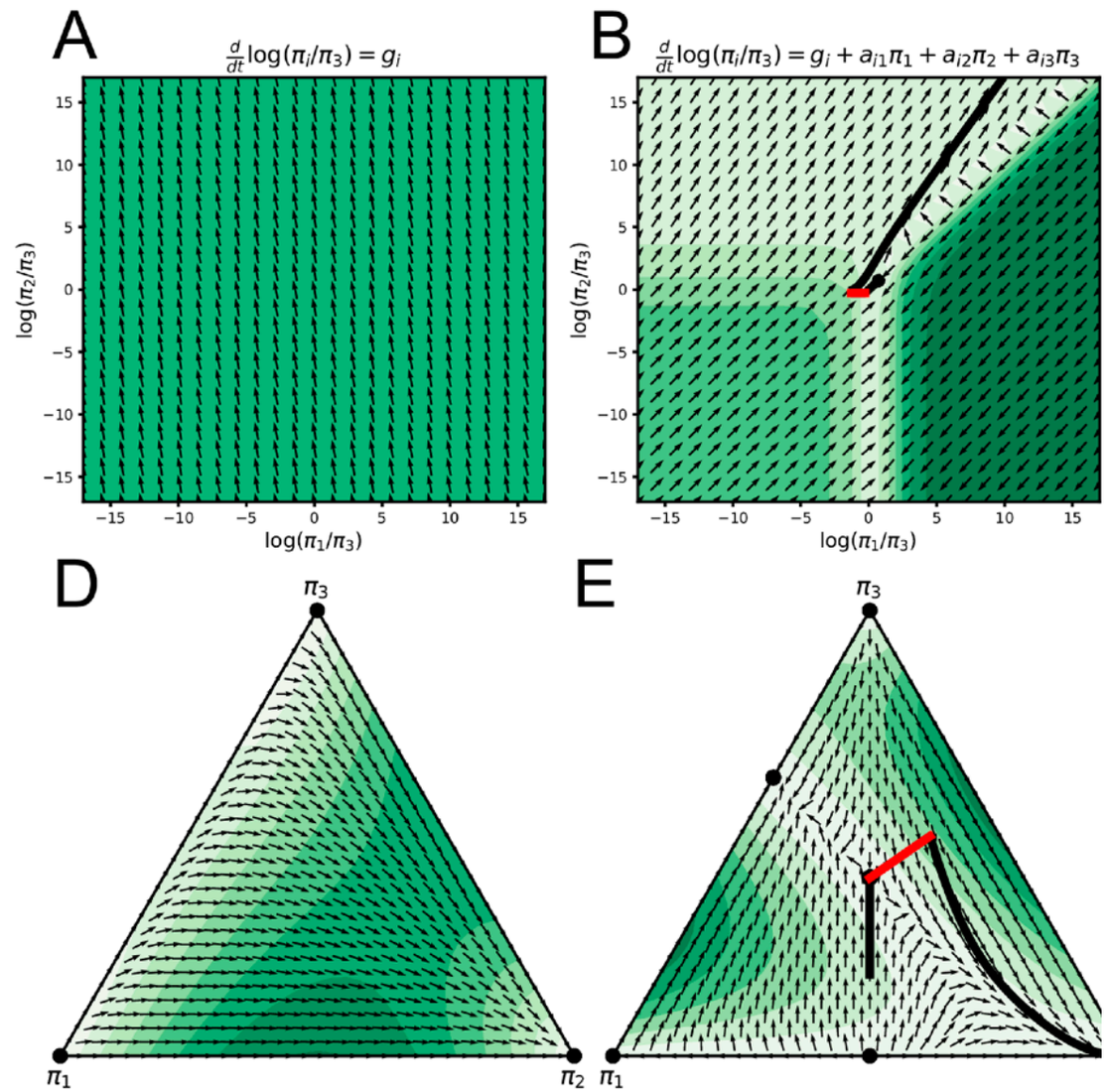
Credit: Joseph et al., *PLoS Comp Bio*, 2020

Compositional Modeling

Compositional Lotka-Volterra describes microbial dynamics in the simplex

Tyler A. Joseph, Liat Shenhav, Joao B. Xavier, Eran Halperin, Itsik Pe'er

"cLV is an approximation to gLV when the variance in community size, $\mathrm{Var}(N(t)) = \mathbb{E}[(N(t)-1)^2]$, is low. Then, the parameters of cLV approximately correspond to differences in parameters of gLV."



Credit: Joseph et al., *PLoS Comp Bio*, 2020

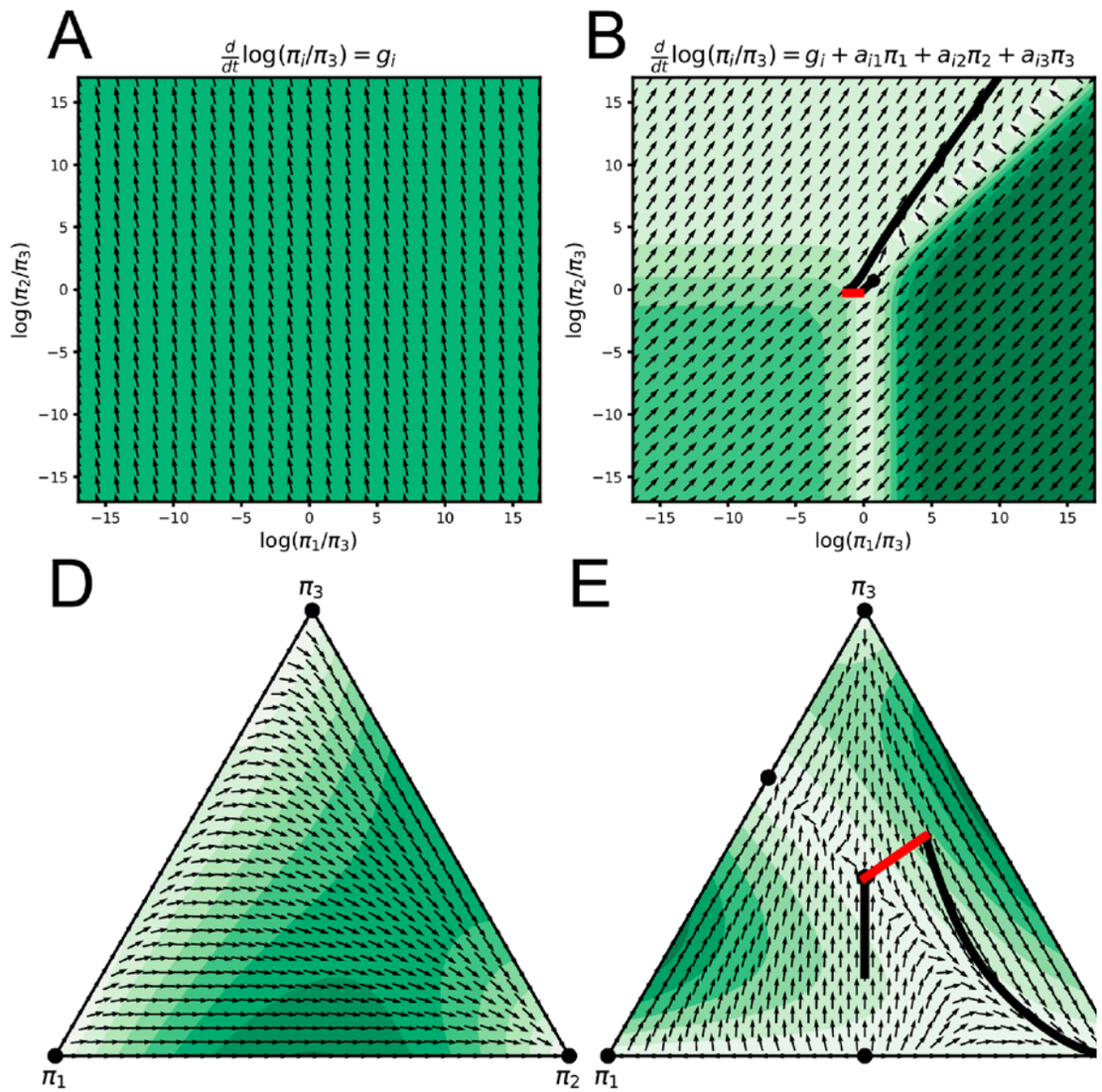
Compositional Modeling

Compositional Lotka-Volterra describes microbial dynamics in the simplex

Tyler A. Joseph, Liat Shenhav, Joao B. Xavier, Eran Halperin, Itsik Pe'er

"cLV is an approximation to gLV when the variance in community size, $\mathrm{Var}(N(t)) = \mathbb{E}[(N(t)-1)^2]$, is low. Then, the parameters of cLV approximately correspond to differences in parameters of gLV."

Not typical that differently scaled communities would follow same dynamics.



Credit: Joseph et al., *PLoS Comp Bio*, 2020

Trouble with zeros

```
log(x/0) = -log(0/x) = bad
```

Trouble with zeros

```
log(x/0) = -log(0/x) = bad
```

Microbial survey data

Sampling zeros: Low abundance, no reads

- pseudo-counts
- imputation
- sampling layer

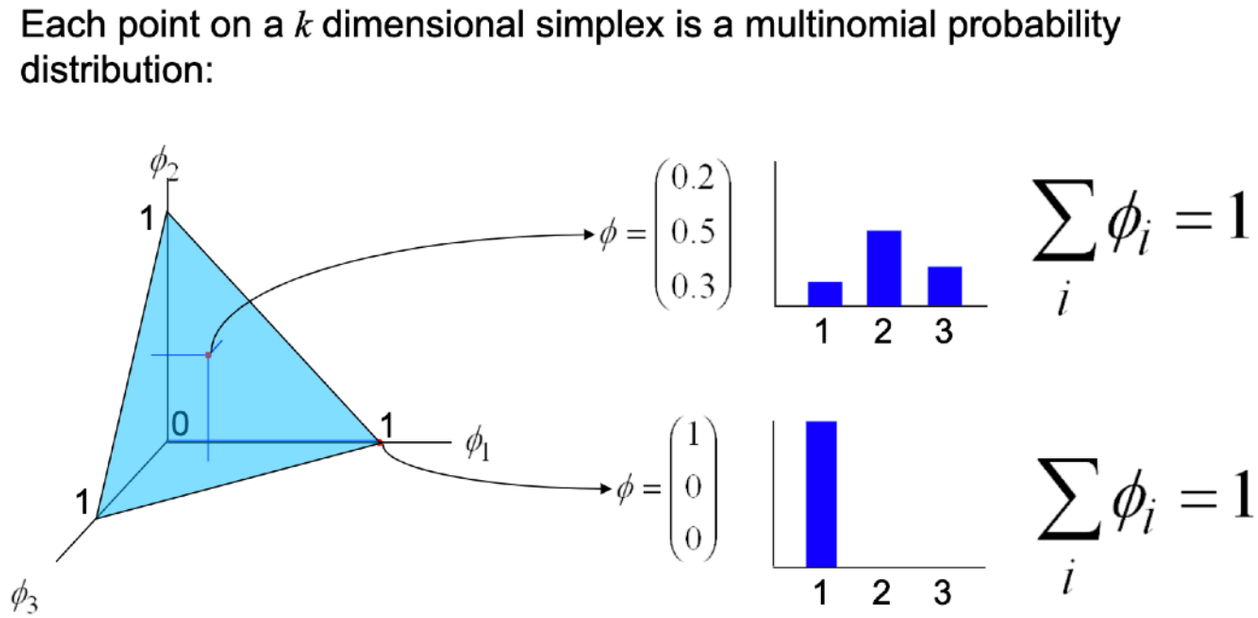


Image Credit: Luke Shelton

Trouble with zeros

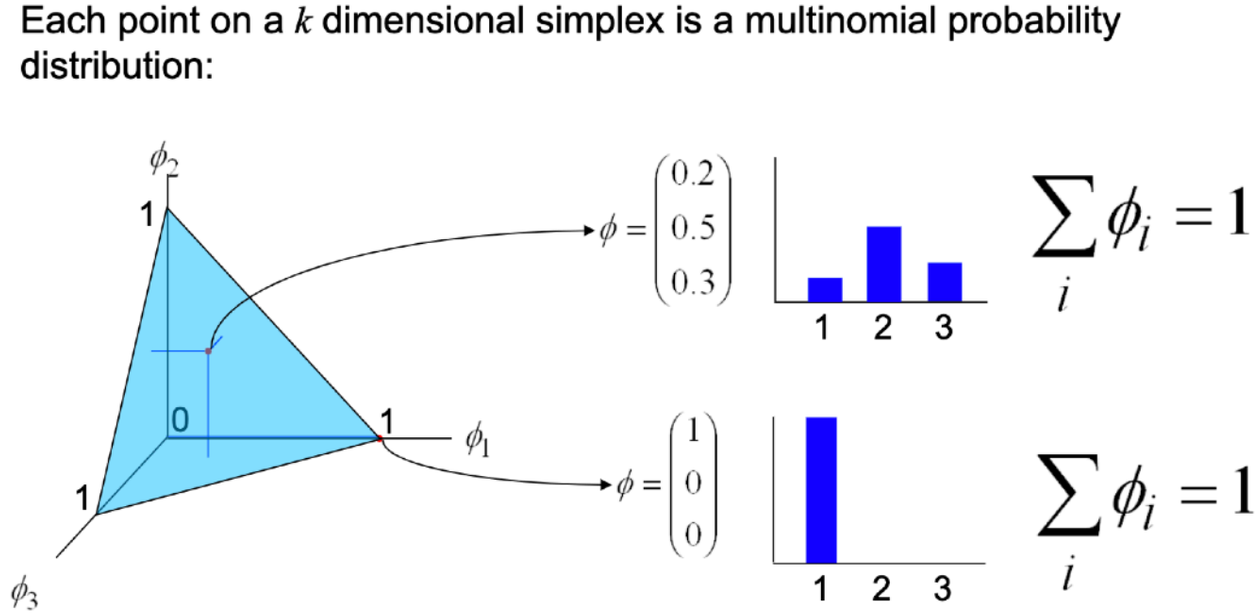
`log(x/0) = -log(0/x) = bad`

Microbial survey data

Sampling zeros: Low abundance, no reads

- pseudo-counts
- imputation
- sampling layer

Each point on a k dimensional simplex is a multinomial probability distribution:



Essential zeros: True zero abundance

- separate treatment

Image Credit: Luke Shelton

Compositionality creates substantial technical and interpretability challenges for metagenomic data in microbial ecology.

Bias and CoDA

- Scale invariance
- **Perturbation invariance**
- Sub-compositional coherence

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

Compositional perturbation

Credit: Aitchison, 1980s

Bias and CoDA

- Scale invariance
- **Perturbation invariance**
- Sub-compositional coherence

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

Compositional perturbation

$$\mathbf{O}(s) \,/\, \mathbf{O}(t) \sim (\mathbf{A}(s) \cdot \mathbf{B}) \,/\, (\mathbf{A}(t) \cdot \mathbf{B}) \sim \mathbf{A}(s) \,/\, \mathbf{A}(t)$$

Credit: Aitchison, 1980s

Bias and CoDA

- Scale invariance
- **Perturbation invariance**
- Sub-compositional coherence

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

Compositional perturbation

$$\mathbf{O}(s) \,/\, \mathbf{O}(t) \sim (\mathbf{A}(s) \cdot \mathbf{B}) \,/\, (\mathbf{A}(t) \cdot \mathbf{B}) \sim \mathbf{A}(s) \,/\, \mathbf{A}(t)$$

The differential bias between protocols is of the same mathematical form as bias

McLaren, Willis & Callahan. *eLife*, 2019.

Bias and CoDA

- Scale invariance
- **Perturbation invariance**
- Sub-compositional coherence

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

Compositional perturbation

$$\mathbf{O}(s) \,/\, \mathbf{O}(t) \sim (\mathbf{A}(s) \cdot \mathbf{B}) \,/\, (\mathbf{A}(t) \cdot \mathbf{B}) \sim \mathbf{A}(s) \,/\, \mathbf{A}(t)$$

The differential bias between protocols is of the same mathematical form as bias

$$\mathbf{O}^{(P)} / \mathbf{O}^{(R)} \sim \mathbf{A} \cdot \mathbf{B}^{(P)} \,/\, (\mathbf{A} \cdot \mathbf{B}^{(R)}) = \mathbf{B}^{(P)} \,/\, \mathbf{B}^{(R)}$$

McLaren, Willis & Callahan. *eLife*, 2019.

Absolute Abundance

Relative Abundances

Absolute Abundances

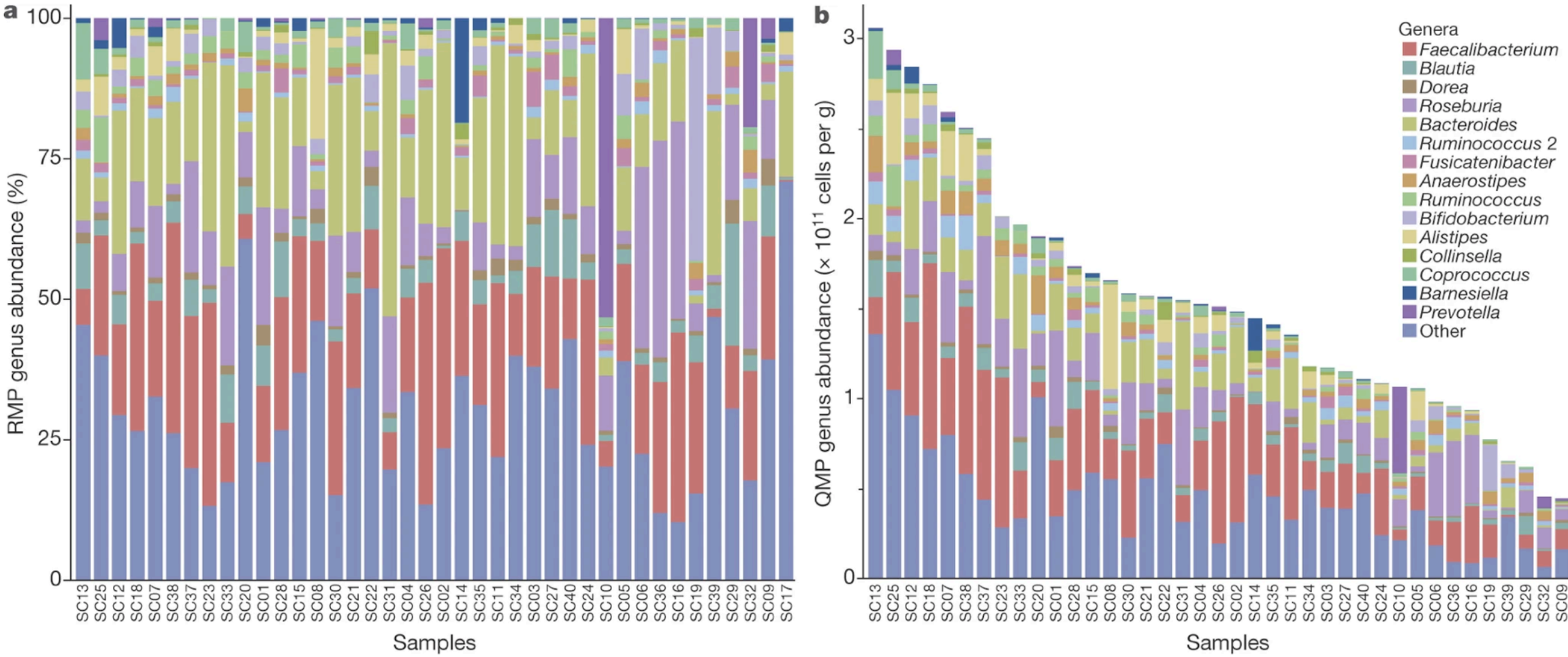
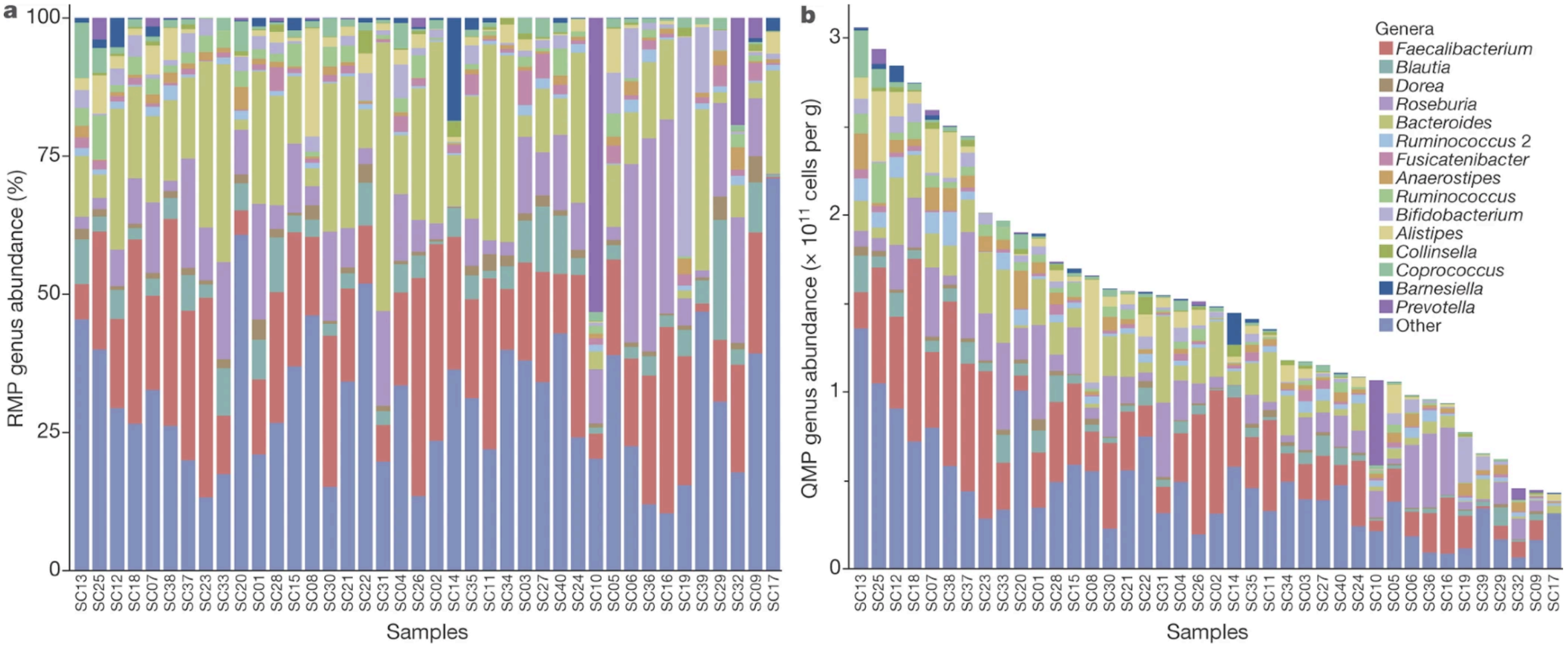


Image Credit: Vandeputte et al., Nature, 2017.

Absolute Abundance

Relative Abundances

Absolute Abundances

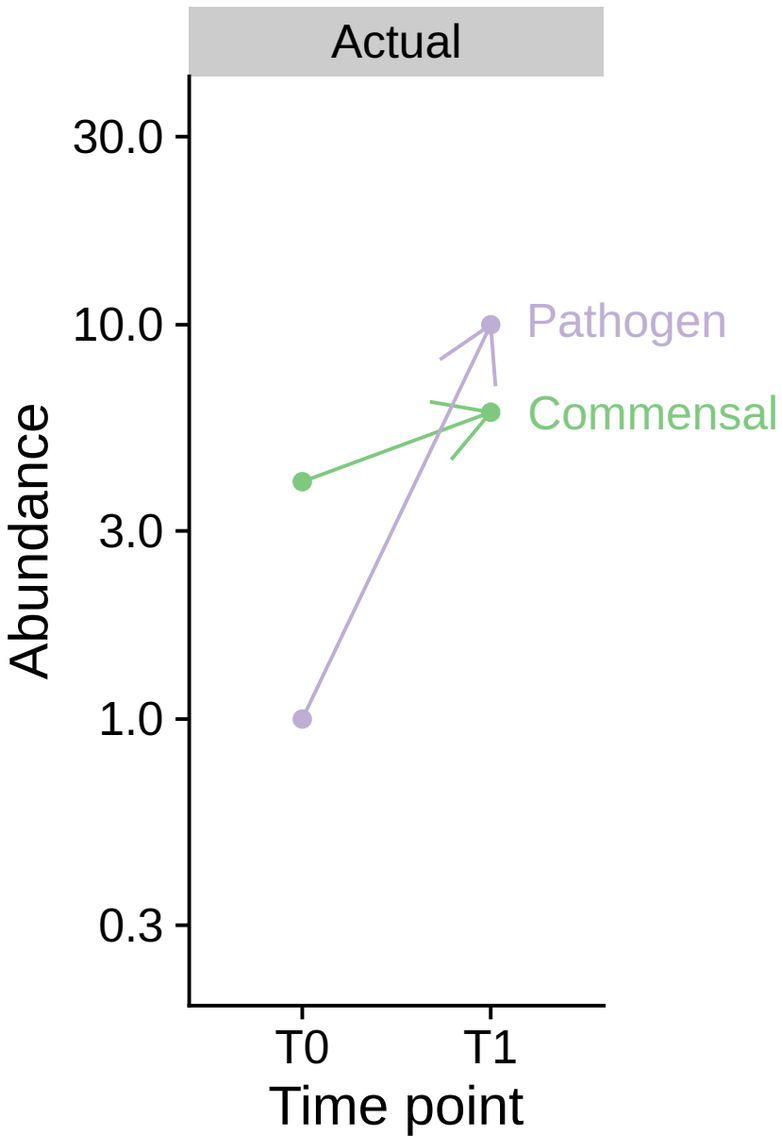


Two types of Absolute Abundance estimation methods

- Proportion-based “Bulk” estimation (e.g. normalize to cell counts)
- Ratio-based “Target” estimation (e.g. normalize to spike-in)

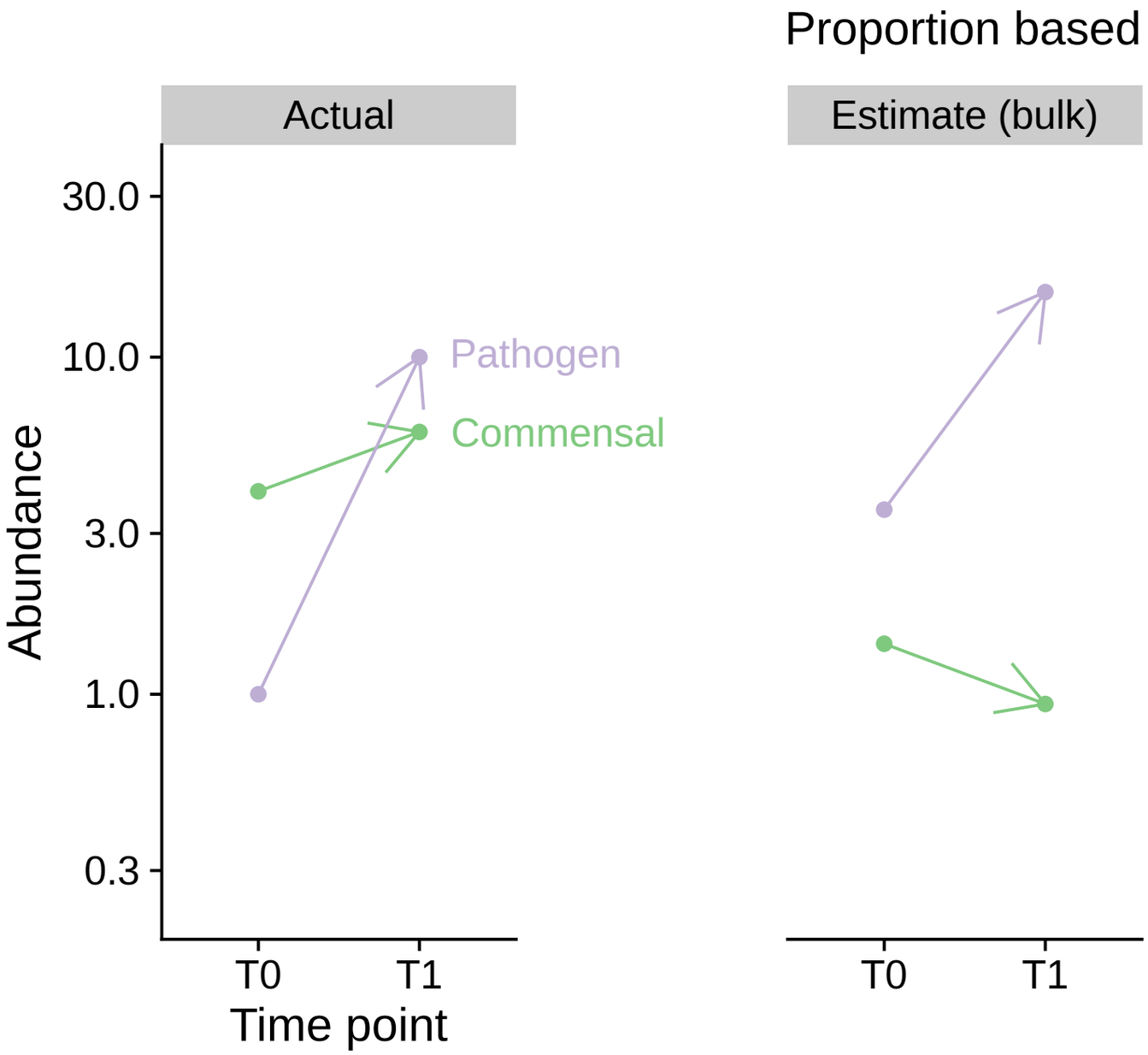
Image Credit: Vandeputte et al., Nature, 2017.

Absolute Abundance



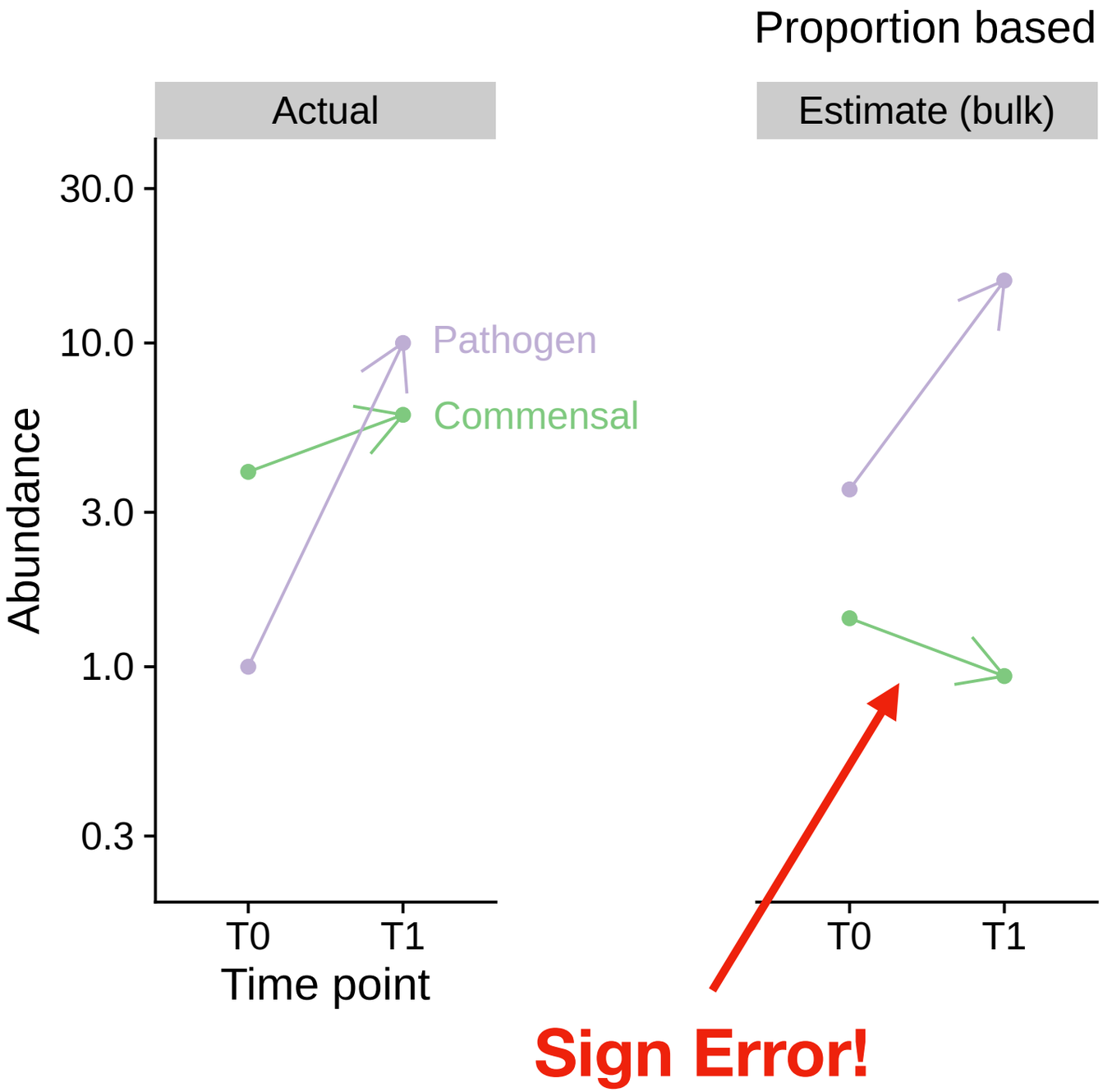
McLaren, Nearing, Willis, Lloyd, Callahan, *bioRxiv*, 2022.

Absolute Abundance



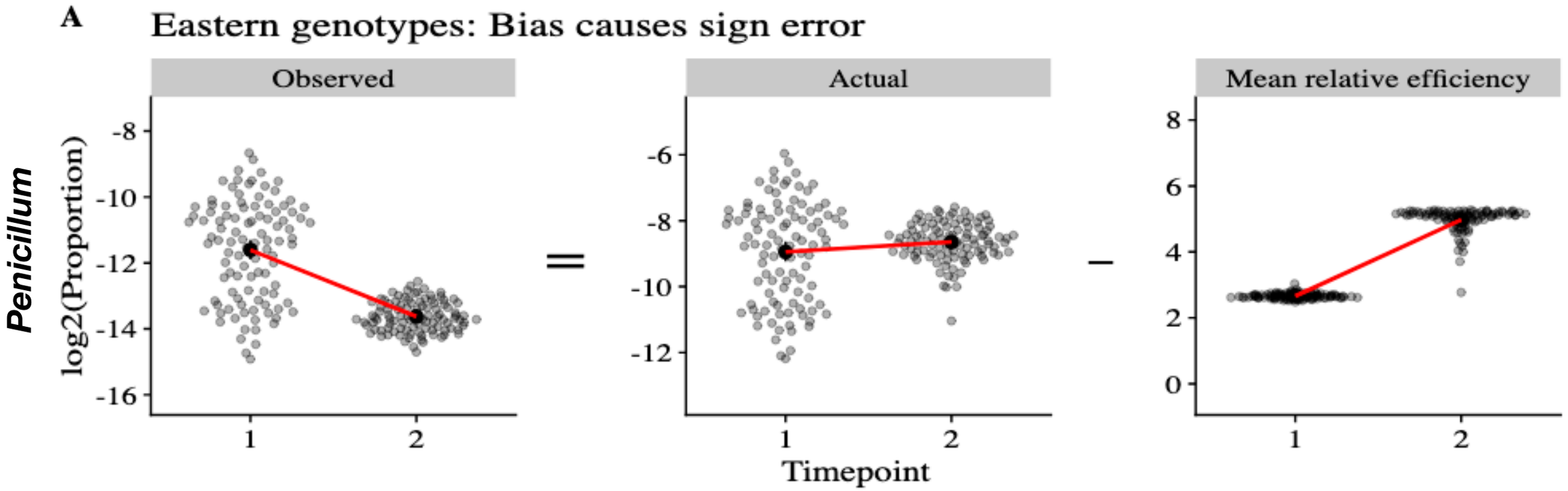
McLaren, Nearing, Willis, Lloyd, Callahan, *bioRxiv*, 2022.

Absolute Abundance



McLaren, Nearing, Willis, Lloyd, Callahan, *bioRxiv*, 2022.

Absolute Abundance



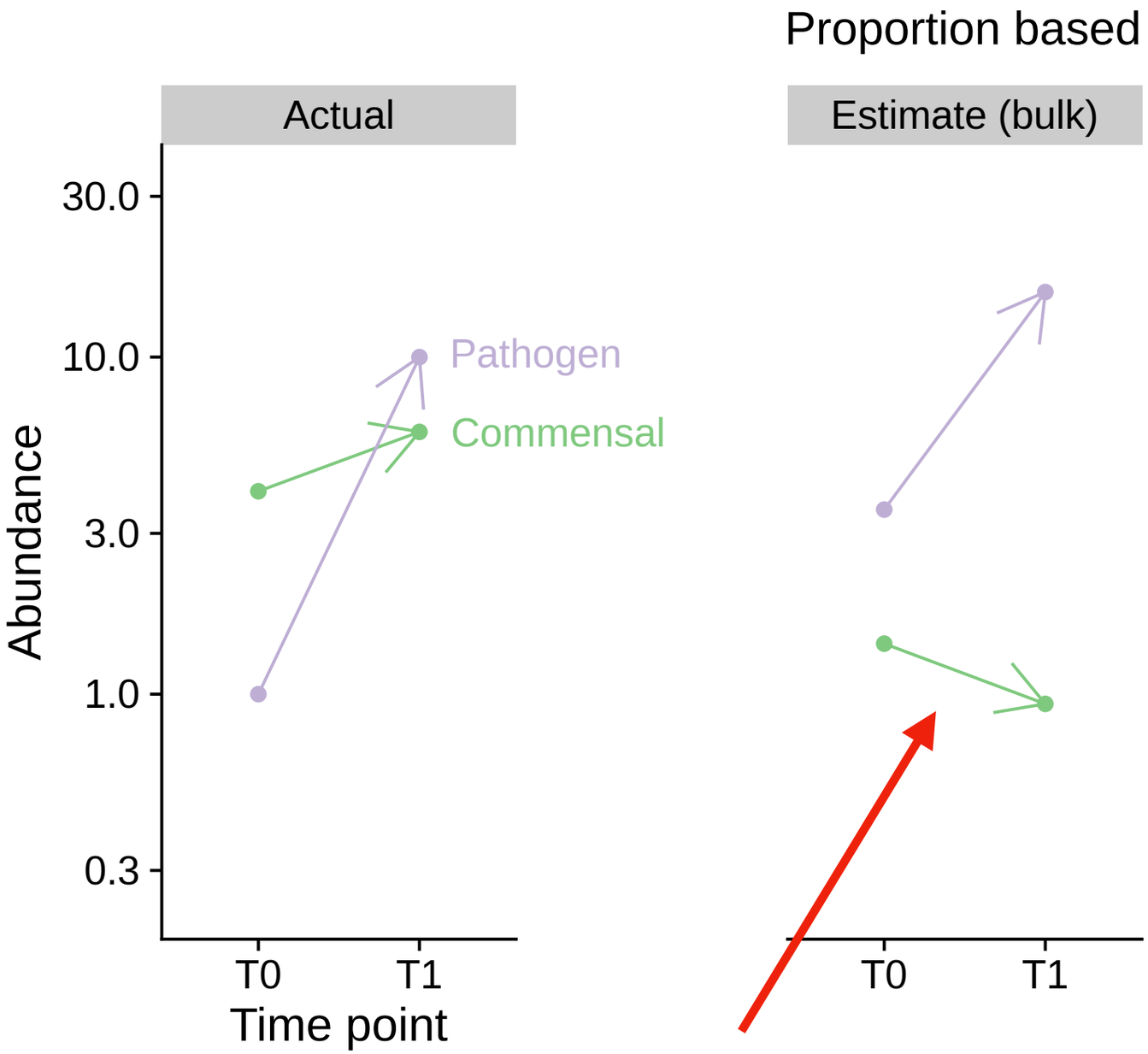
Experimental manipulation of plant microbiome assembly in cottonwood leaves

High-efficiency pathogen *Melampsora* × *columbiana* invades in T2.

Synthetic (defined) community.

Data Credit: Leopold and Busby, Current Biology, 2020.

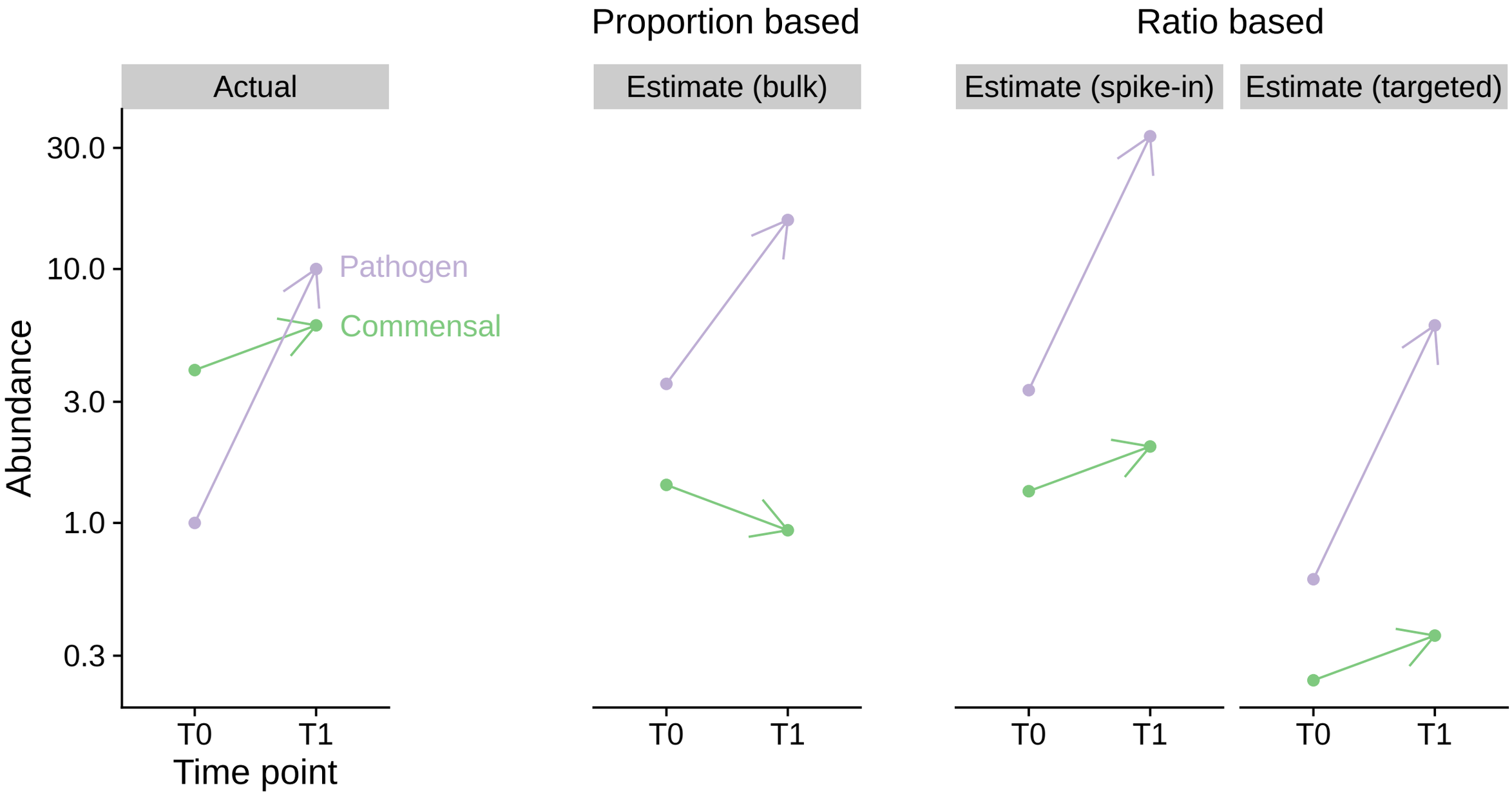
Absolute Abundance



Uncorrected systematic change in $\bar{B}$

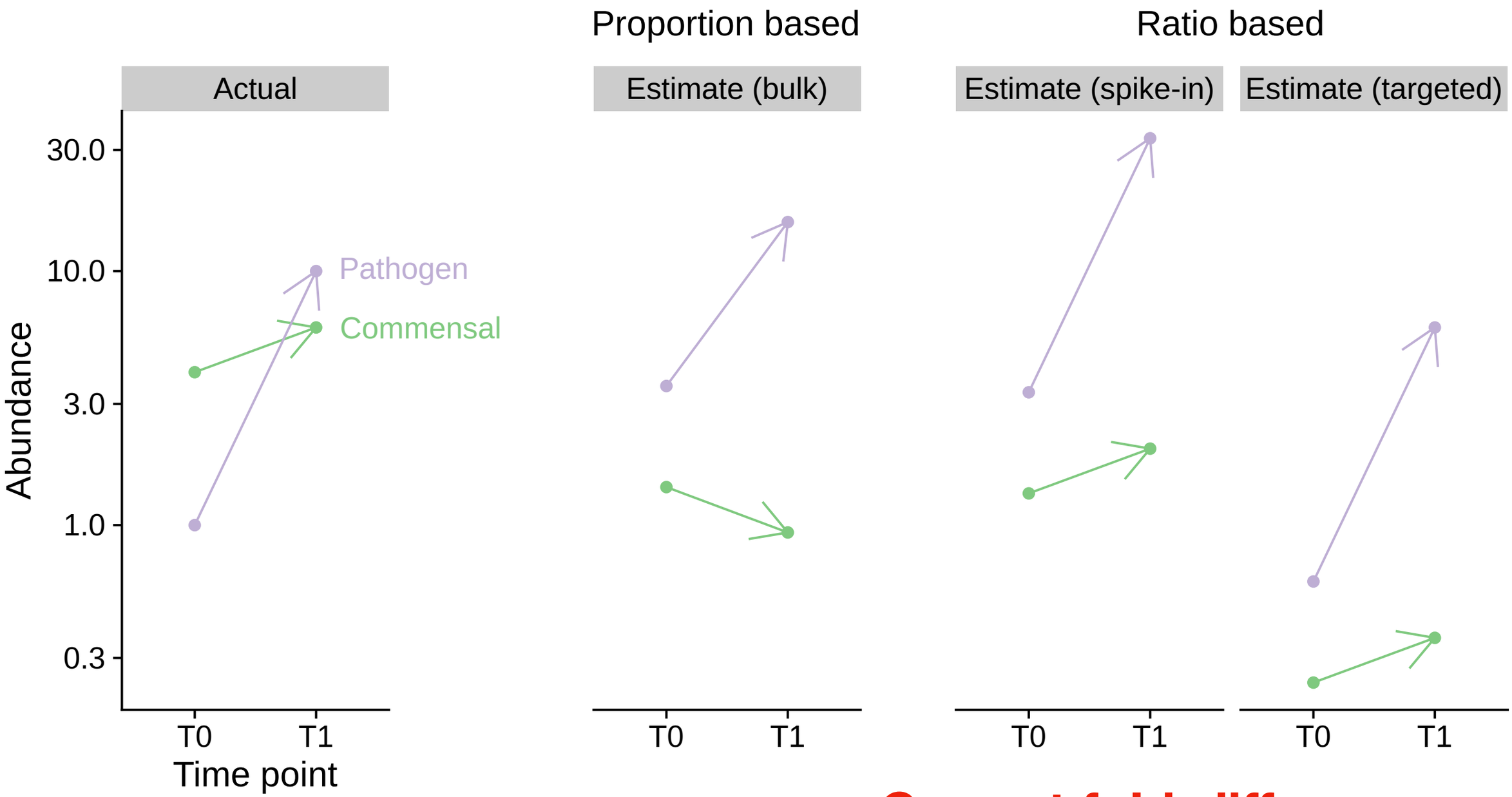
McLaren, Nearing, Willis, Lloyd, Callahan, *bioRxiv*, 2022.

Absolute Abundance



McLaren, Nearing, Willis, Lloyd, Callahan, *bioRxiv*, 2022.

Absolute Abundance



Correct fold-differences
Biased absolute abundances

McLaren, Nearing, Willis, Lloyd, Callahan, *bioRxiv*, 2022.