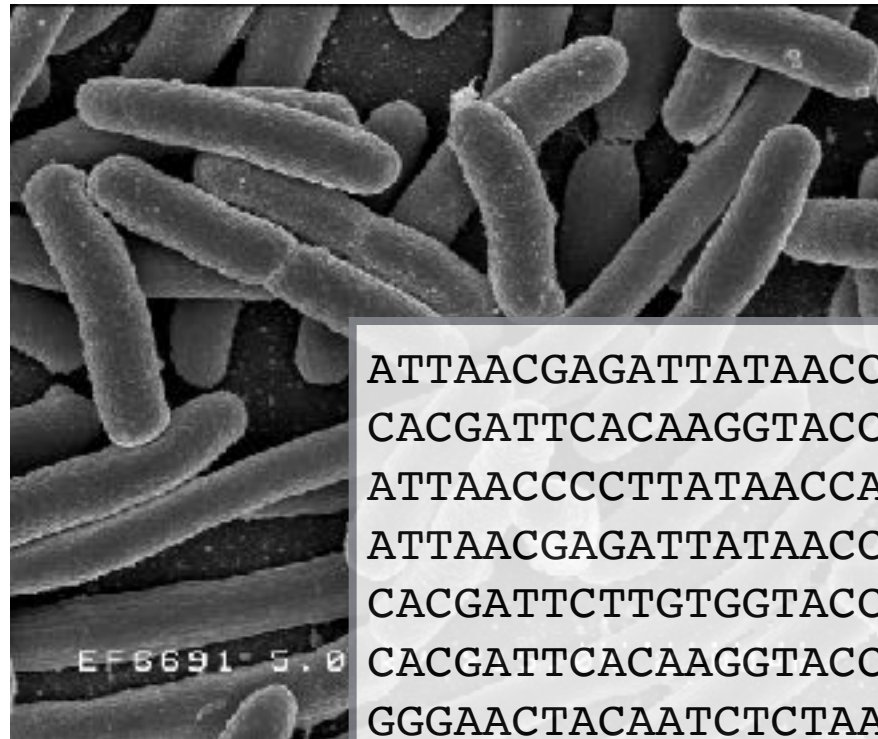


Biased* and Methodology-specific Measurement of Microbial Communities

Microbial Census

Metagenomics or Marker-gene (MGS) Sequencing



ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGA ACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC

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

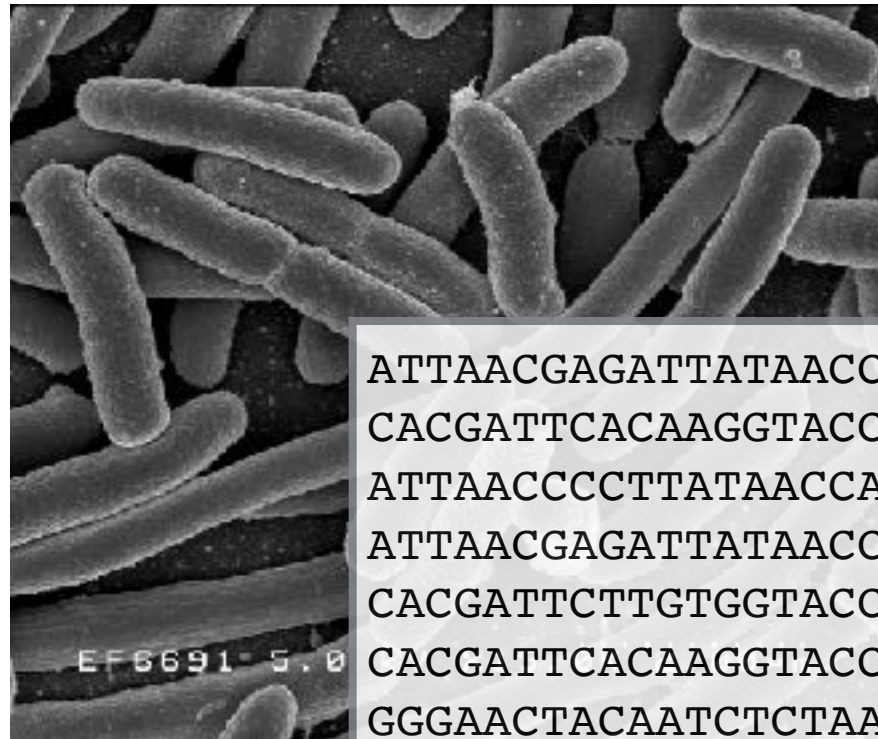
→ Inference

↓
Exploration

↘
Modeling

Microbial Census

Metagenomics or Marker-gene (MGS) Sequencing



ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGA ACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC

- **ASV table from DADA2**
- **Taxonomy table from MetaPhlAn**
- **Taxonomy table from Sourmash-gather**
- ...and many, many other methods...

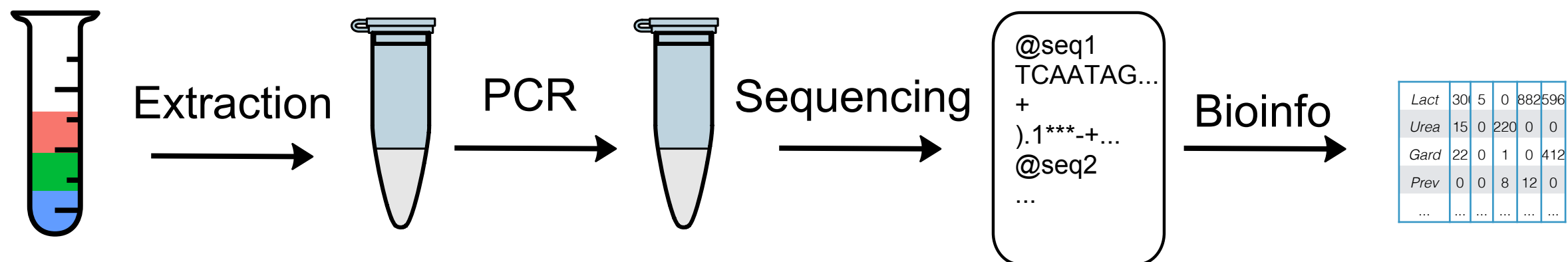
<i>Lactobacillus crispatus</i>	1300	5	0	882	596
<i>Ureaplasma urealytica</i>	15	0	220	0	0
<i>Gardnerella vaginalis</i>	22	0	1	0	412
<i>Prevotella intermedia</i>	0	0	8	12	0
...

Exploration

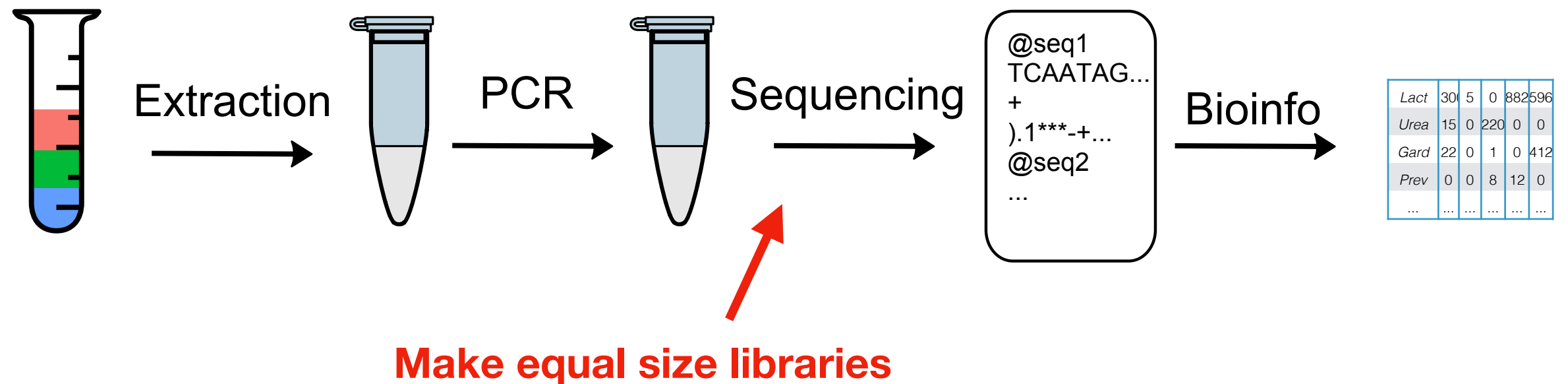
Modeling

Inference

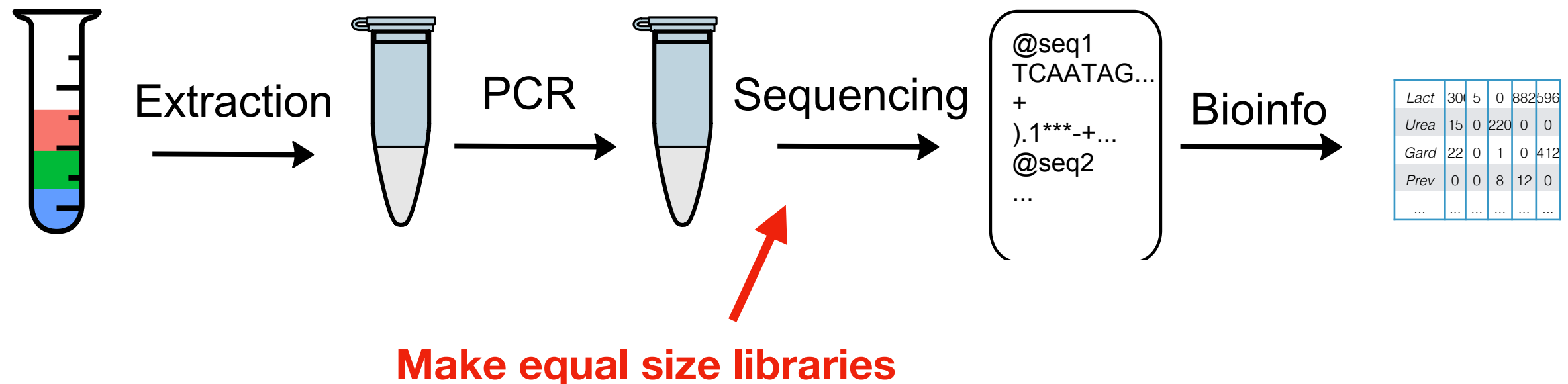
Metagenomic Compositionality



Metagenomic Compositionality



Metagenomic Compositionality

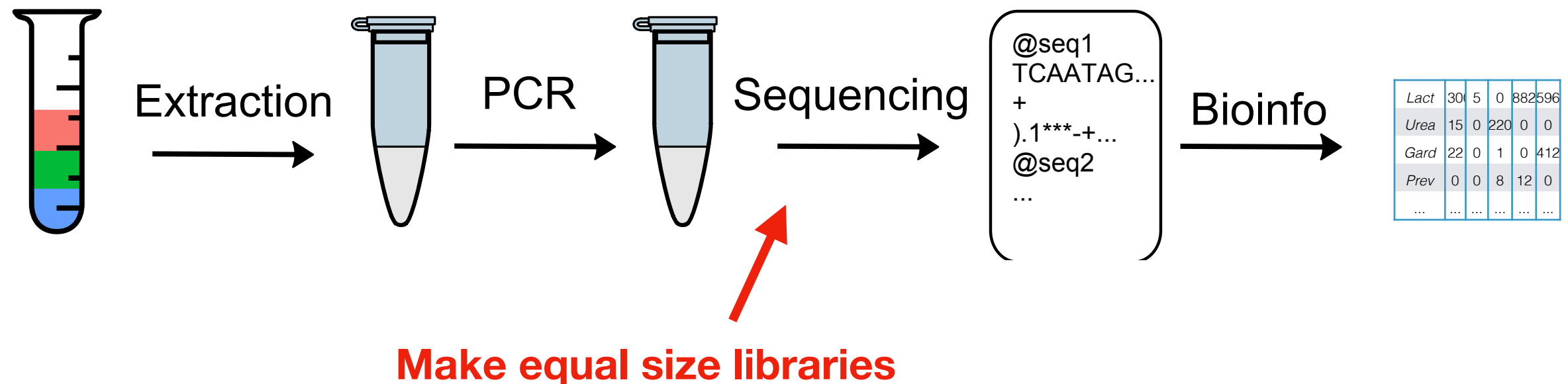


1000X microbial concentration



No change in **MGX** measurement

Metagenomic Compositionality

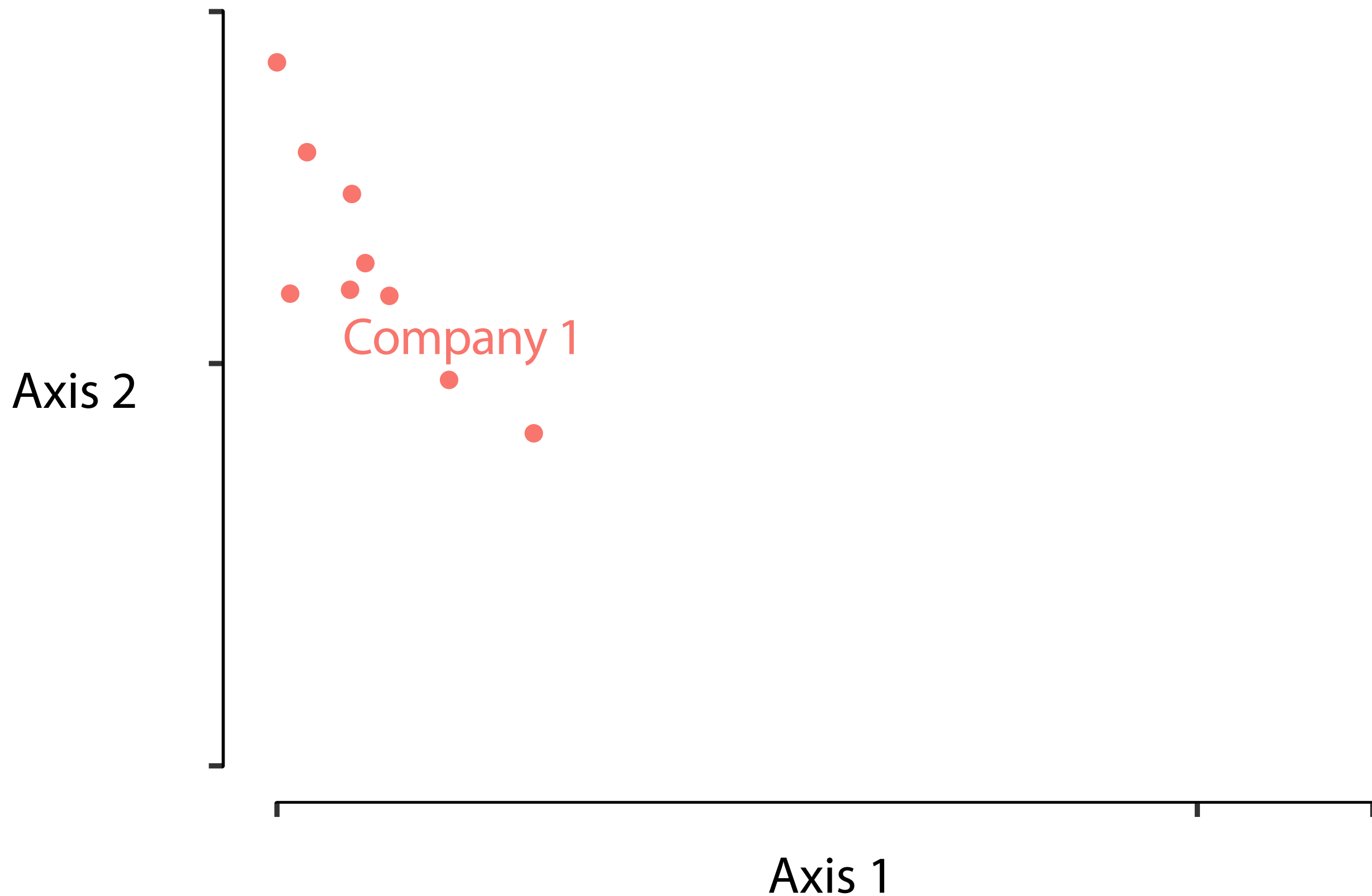


Every read in the data is the result of a

competition

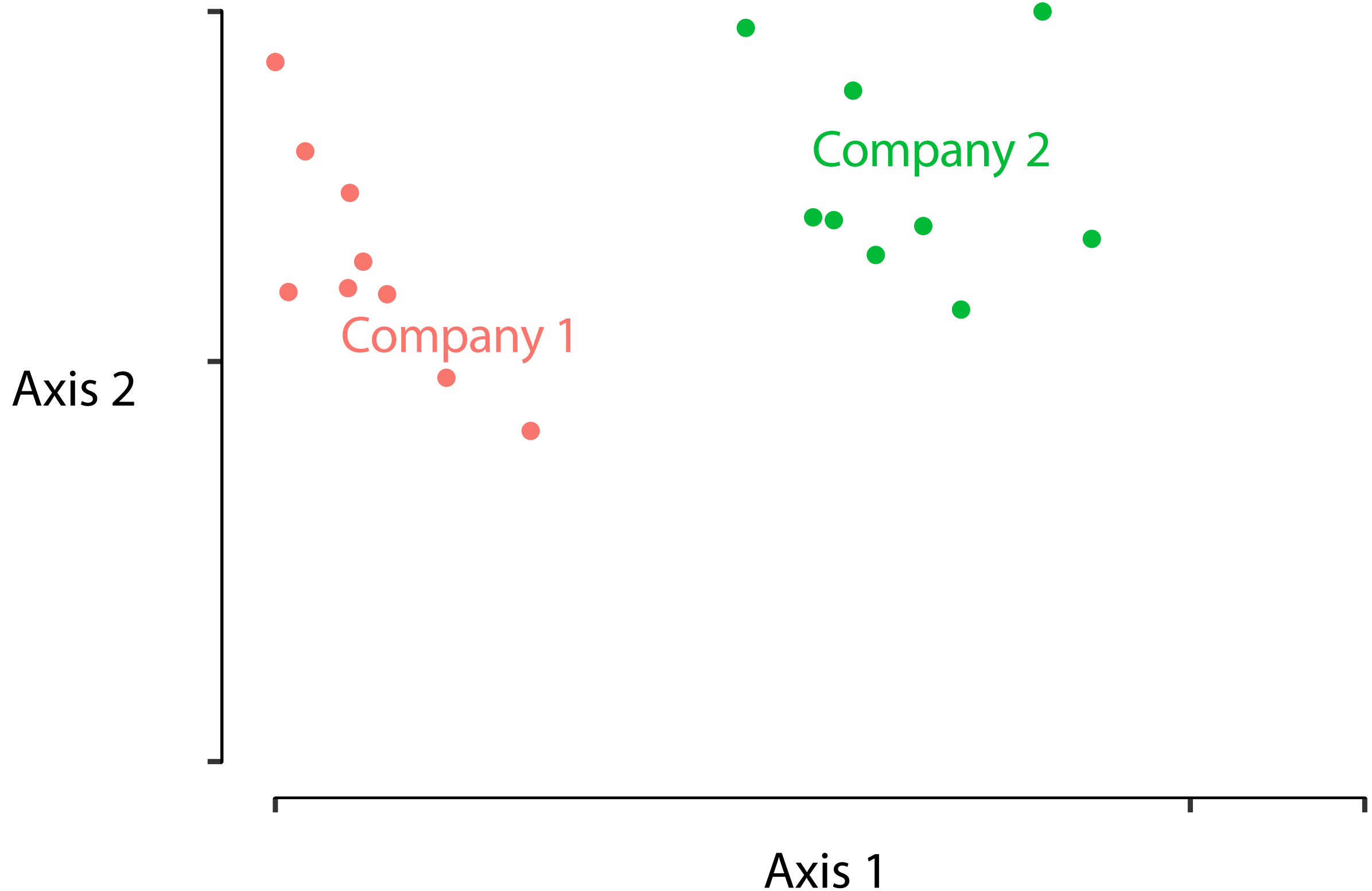
with all other sources of library-DNA for limited slots

Metagenomic Bias*



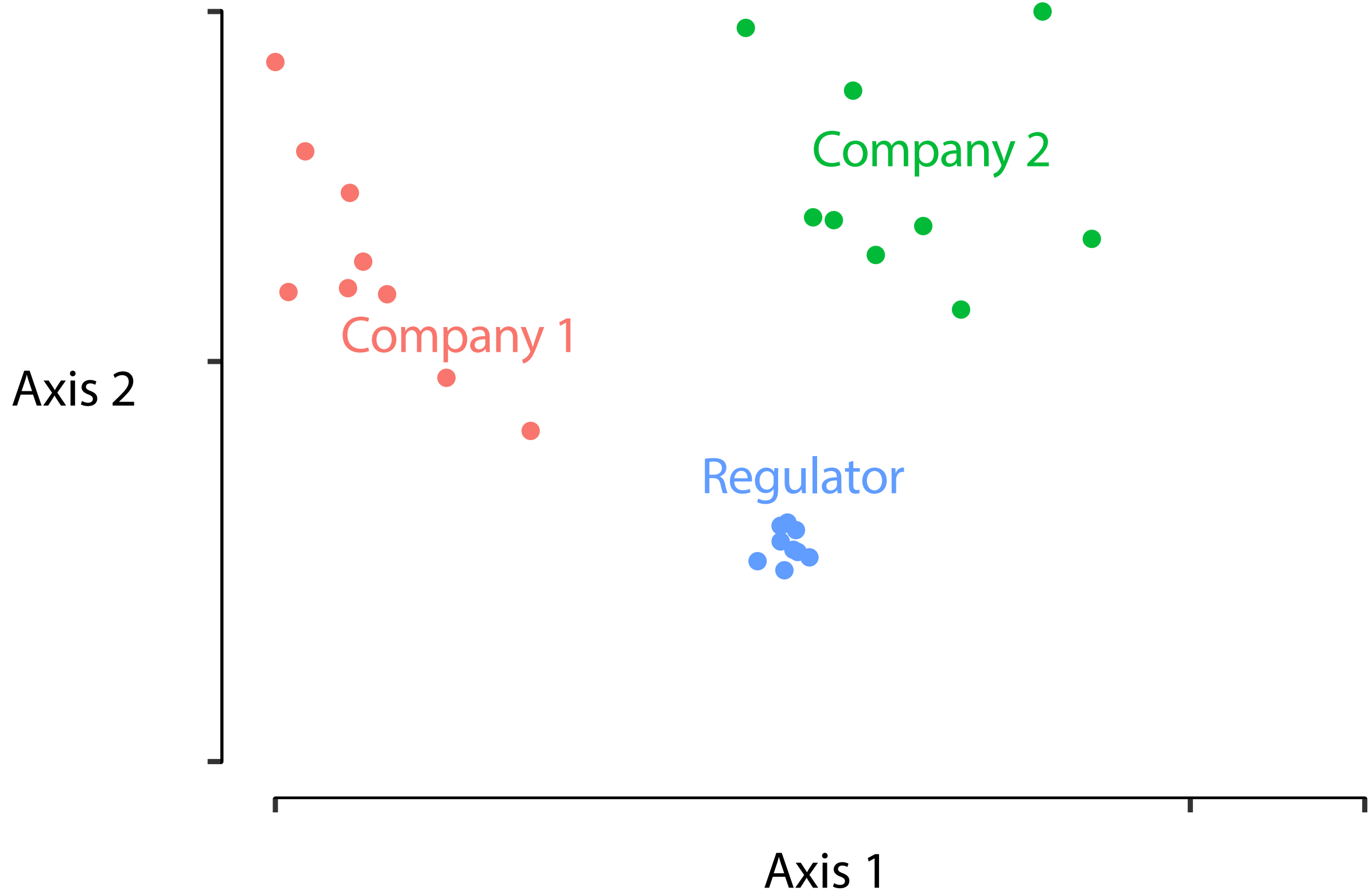
Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Bias*



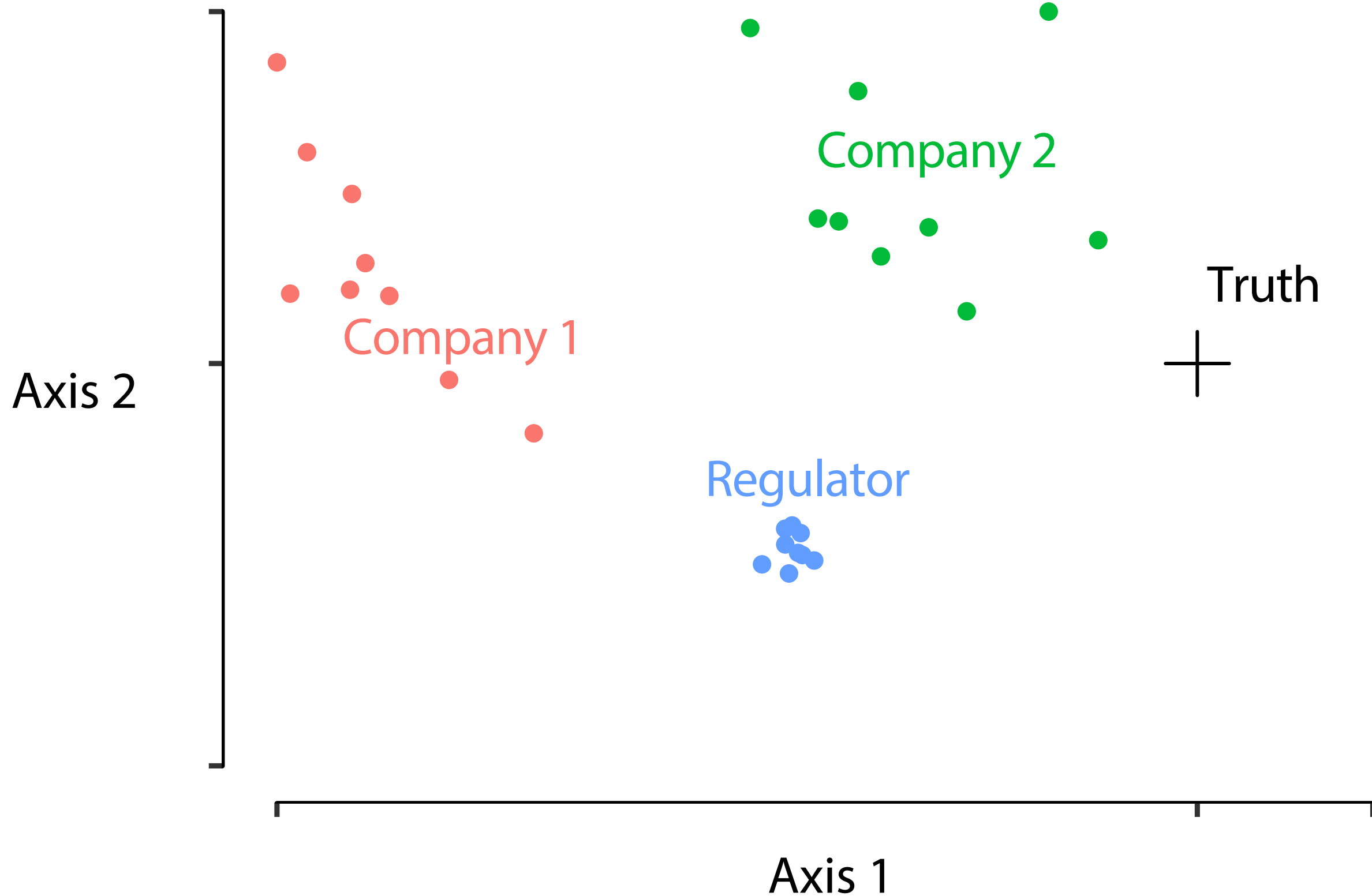
Data: Costea, et al. *Nature Biotechnology*, 2017.

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 influenced by DNA extraction, sequencing, and PCR primer choice

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>

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 2012

Sample

Chengwei Luo, Despina Tsementzi, Nikos K

Published: February 10, 2012 • <https://doi.org/10.1523/JNEUROSCI.4500-11.2012>

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

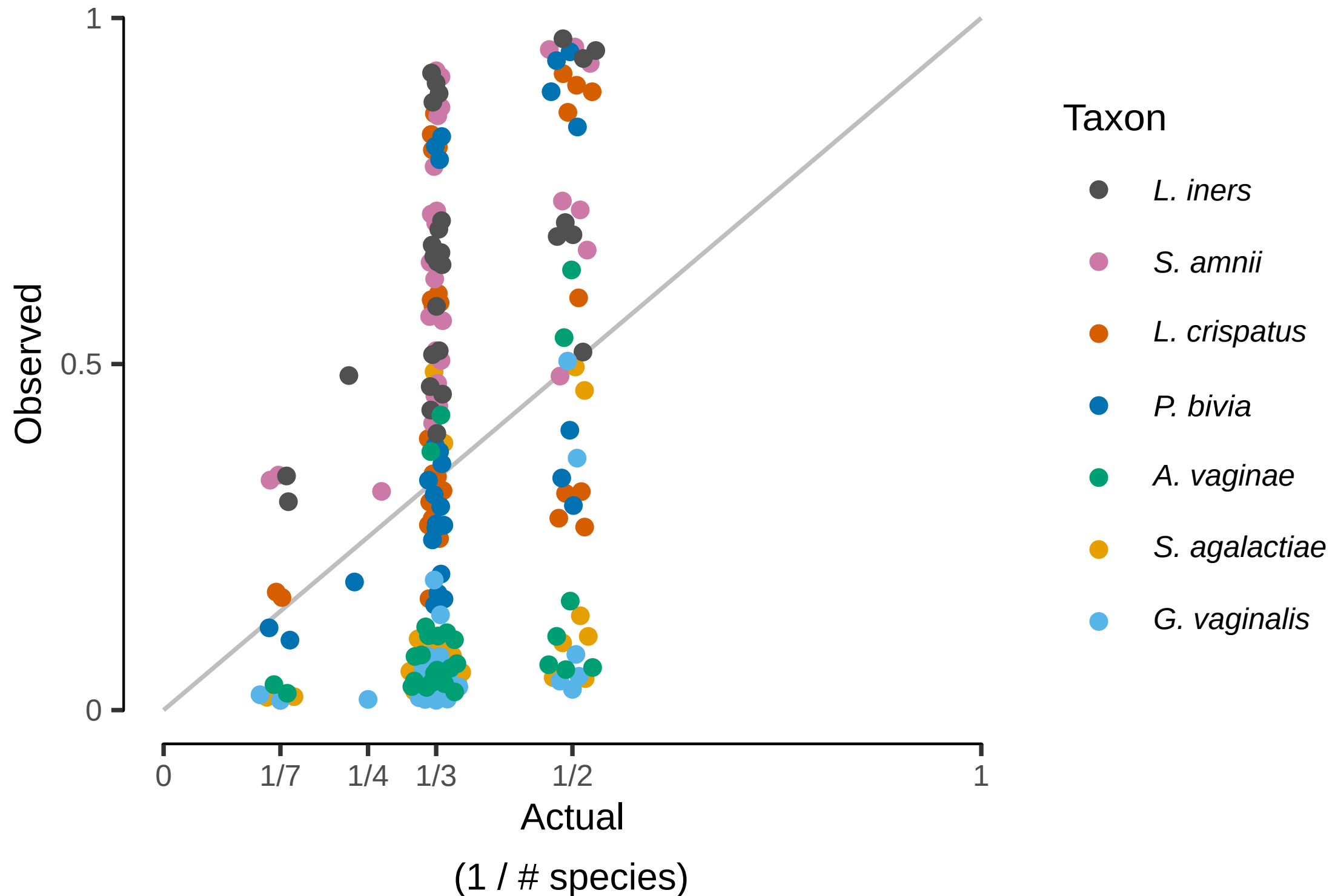
Every metagenomic measurement is **biased*** from the truth

The relative abundances measured by metagenomics
are **systematically inaccurate**

MGS measurements are **not quantitatively reproducible**
across labs/methods

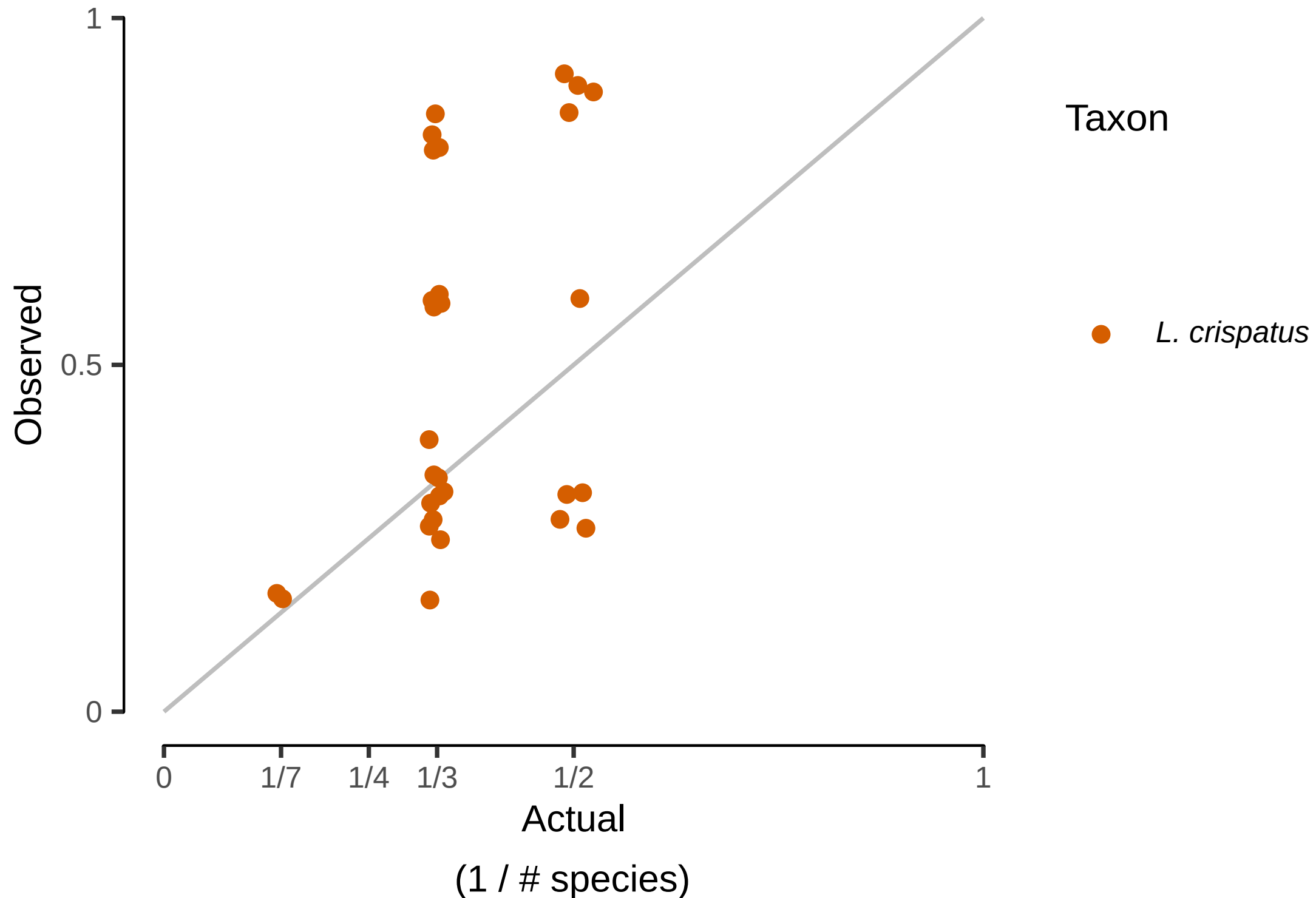
Metagenomic Bias

Observed proportion vs. actual



Metagenomic Bias

Observed proportion vs. actual



Modeling Metagenomics Bias

Truth
● 33%
● 33%
● 33%

A

Measured
● 4%
● 72%
● 24%

O

Measurement

Modeling Metagenomics Bias

Truth
● 33%
● 33%
● 33%

A

Measured
● 4%
● 72%
● 24%

O

$$O = f(A)$$

Modeling Metagenomics Bias

What is $f()$?

Truth

● 33%

● 33%

● 33%

A

Measured

● 4%

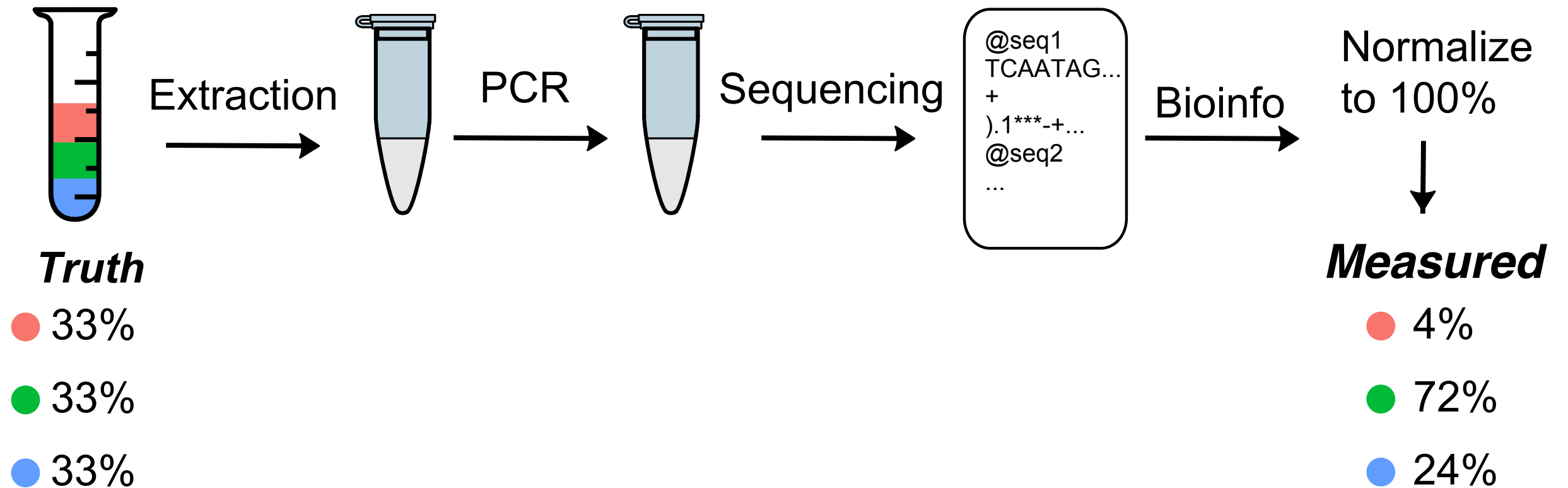
● 72%

● 24%

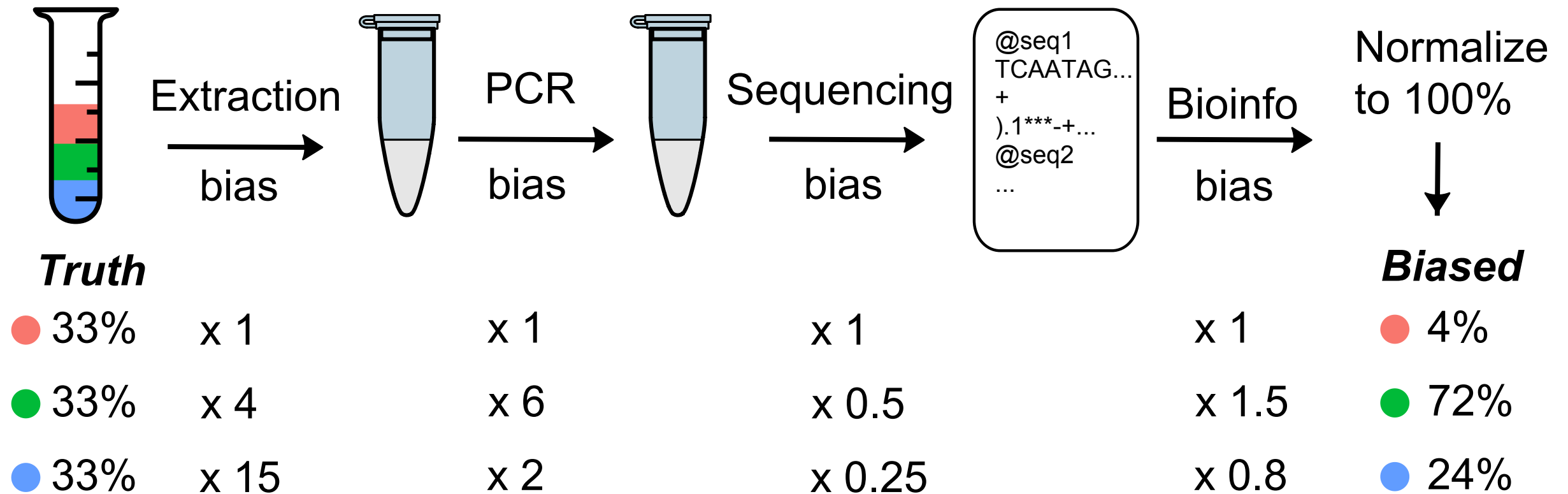
O

$$O = f(A)$$

Modeling Metagenomics Bias

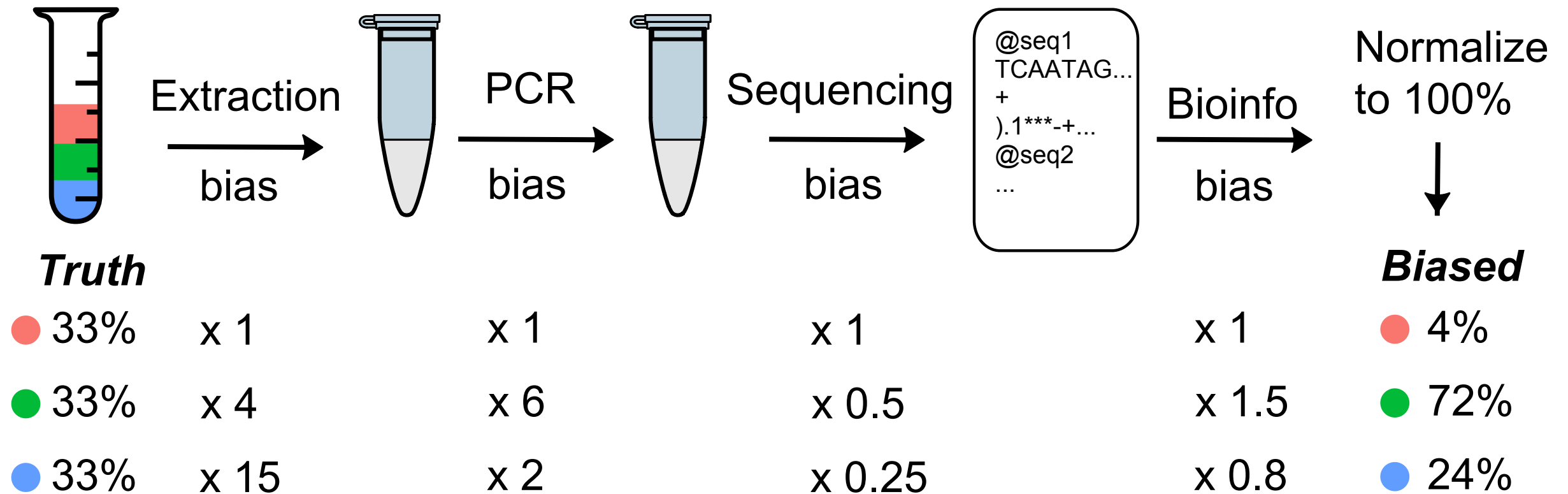


Modeling Metagenomics Bias

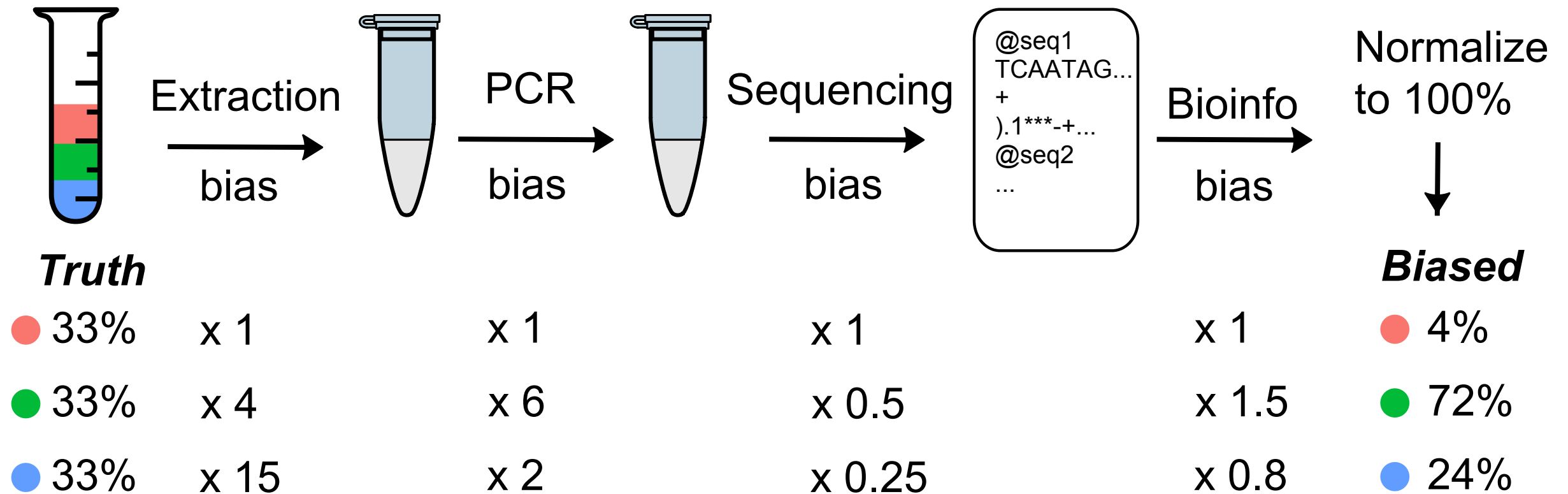


Modeling Metagenomics Bias

Strong evidence that some bias mechanisms act in this way.

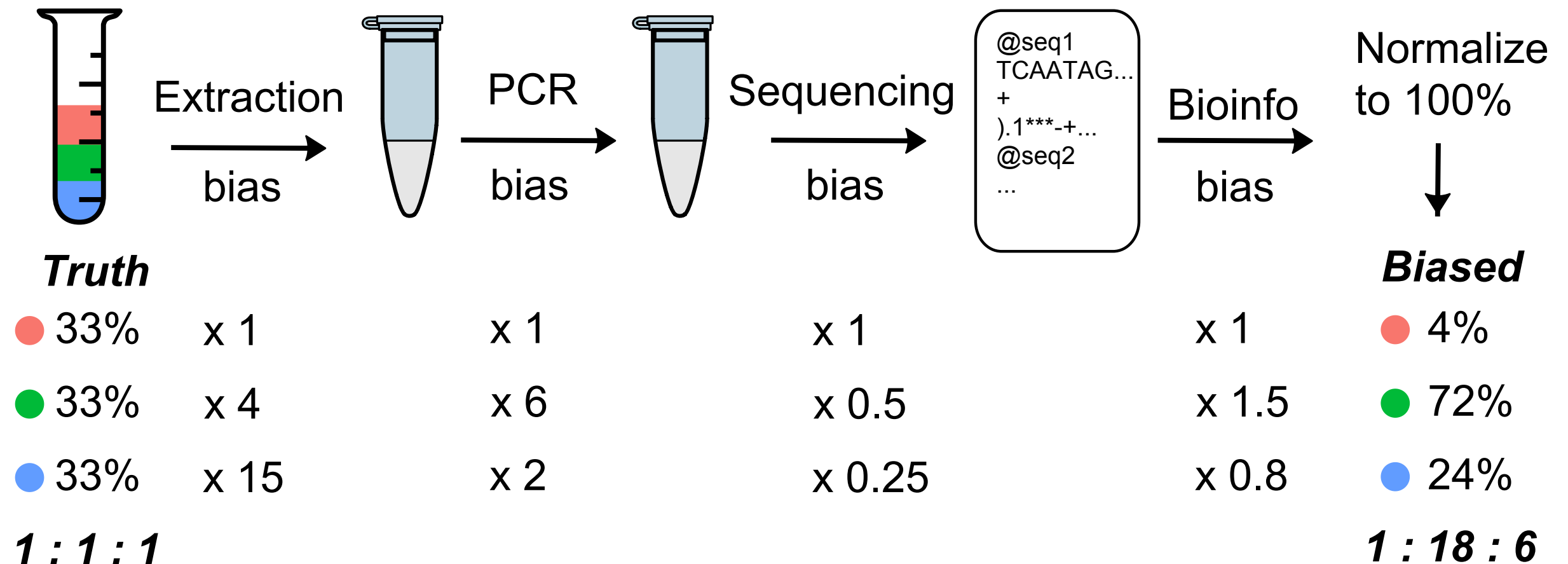


Modeling Metagenomics Bias



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

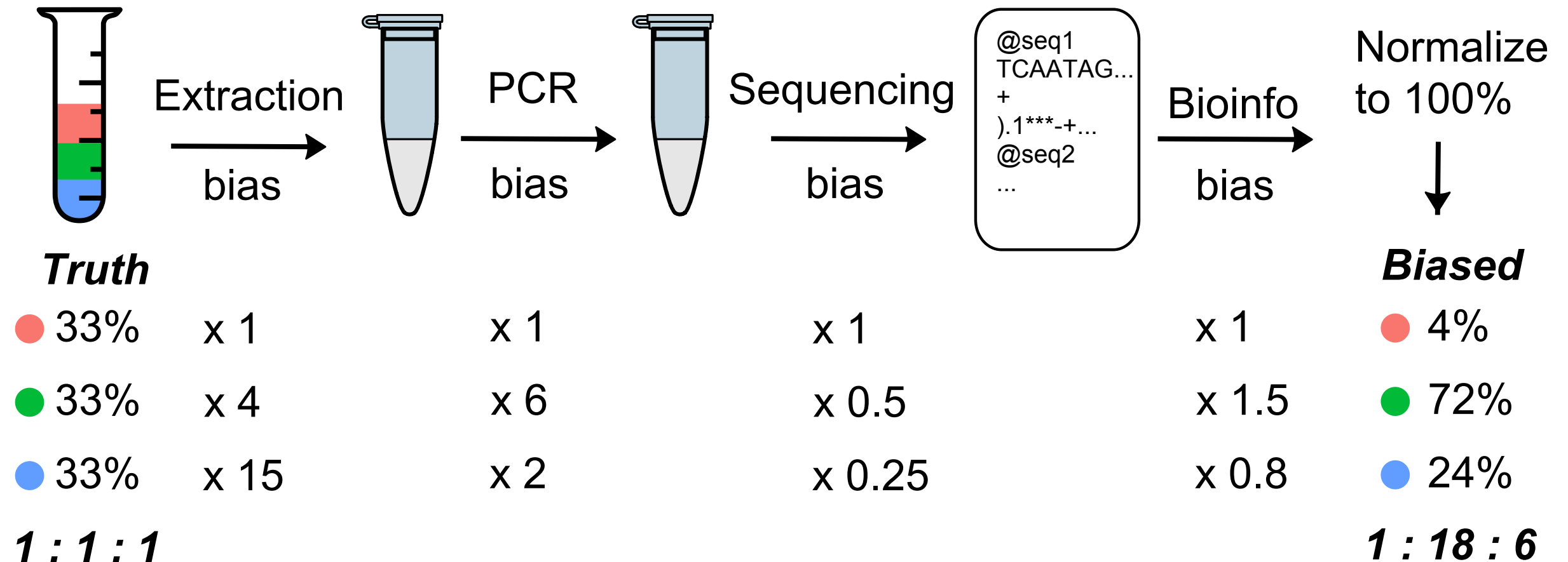
Modeling Metagenomics Bias



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

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

Modeling Metagenomics Bias

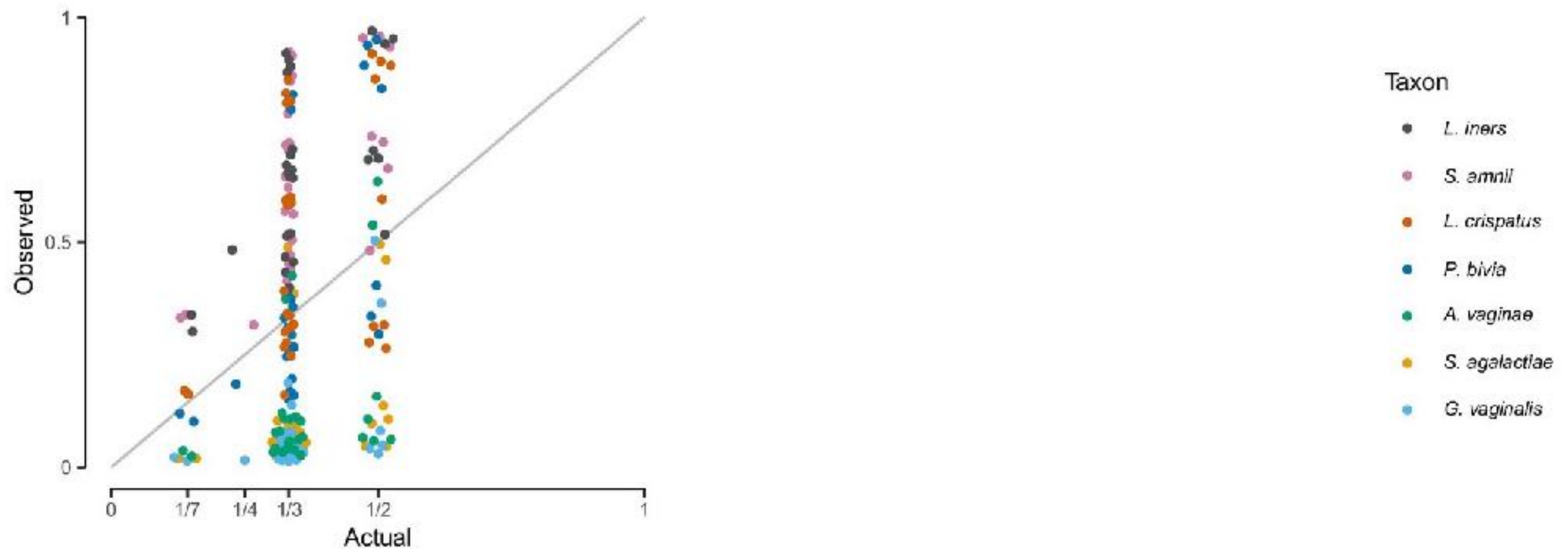


$$O \sim A \cdot B^{(P_1)} \cdot B^{(P_2)} \cdot \dots \cdot B^{(P_L)}$$

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

Testing the Model

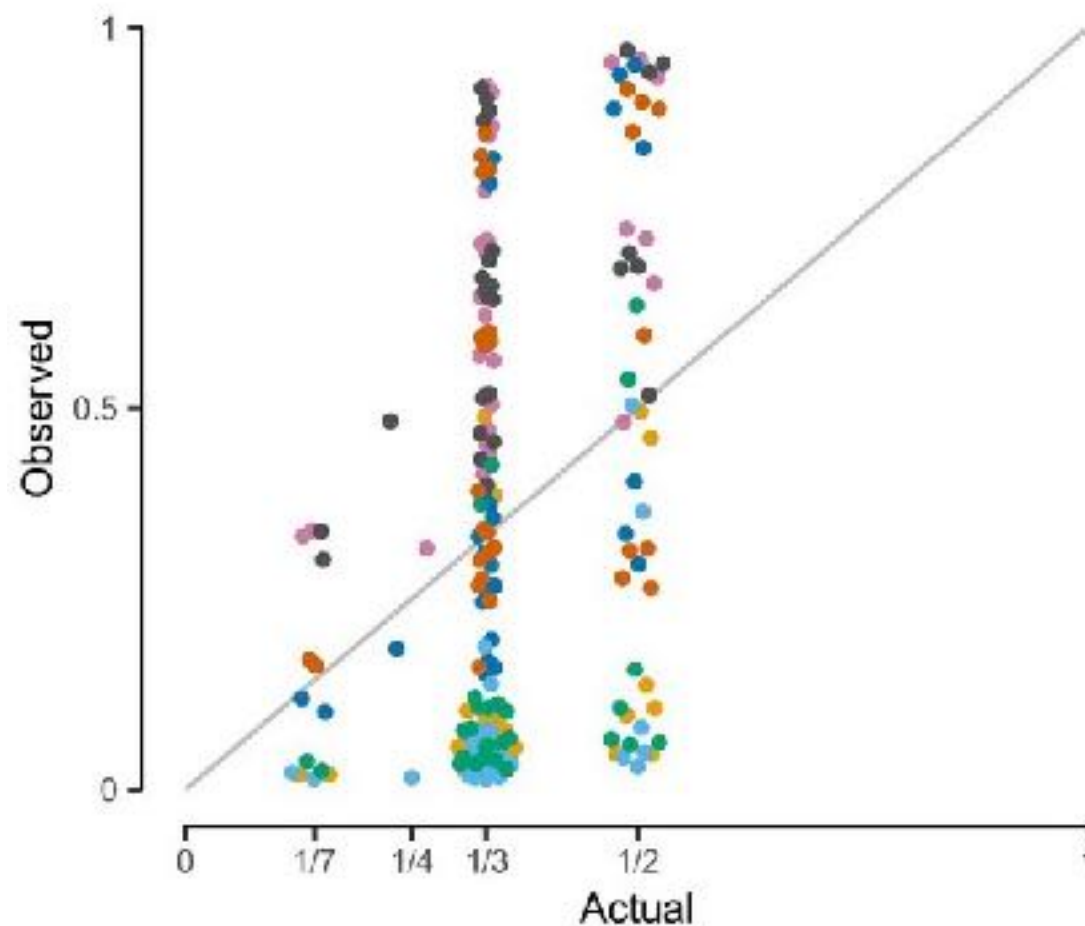
A Observed proportion vs. actual



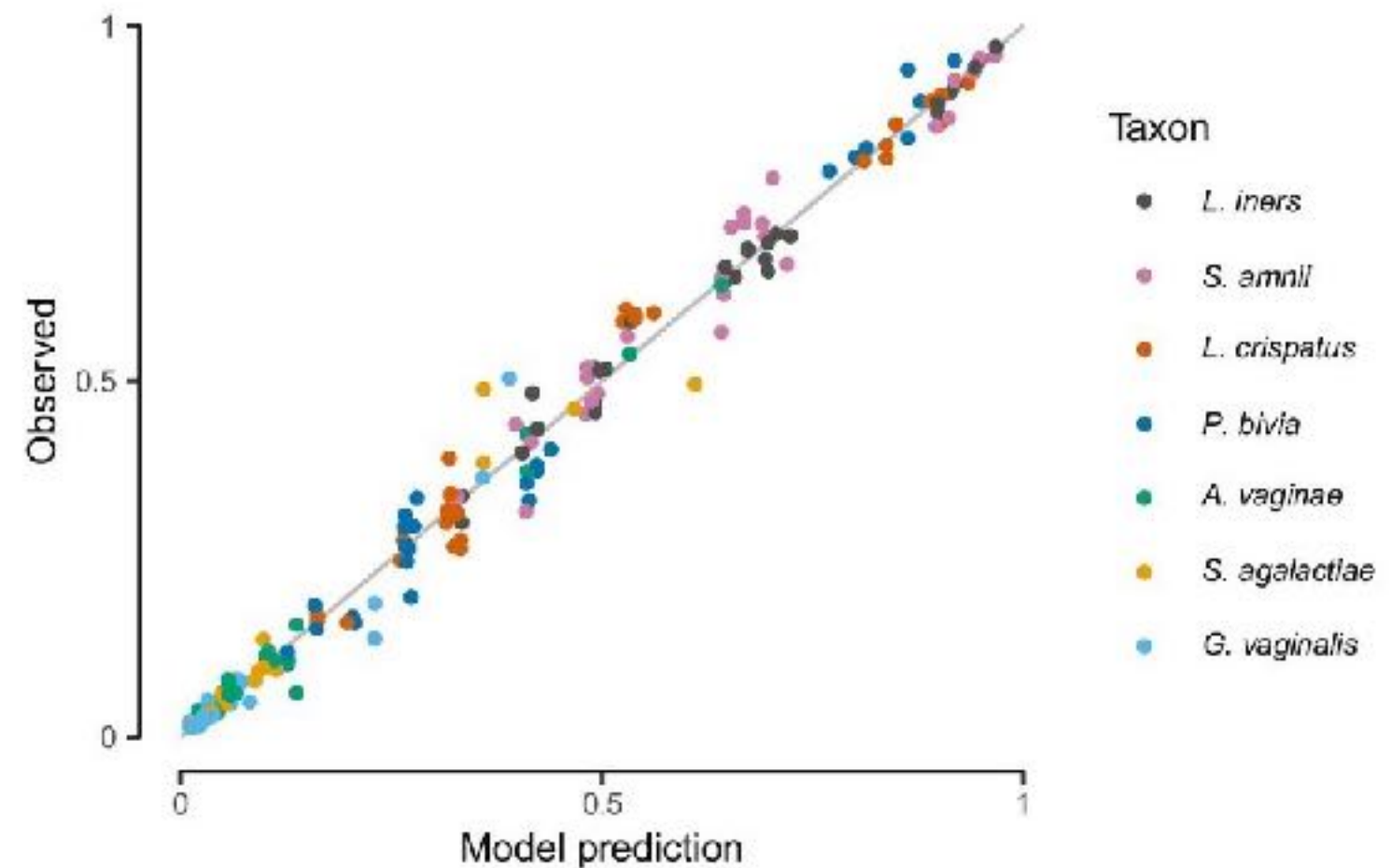
Testing the Model

It works!

A Observed proportion vs. actual



B Observed proportion vs. model prediction



This is $f()$



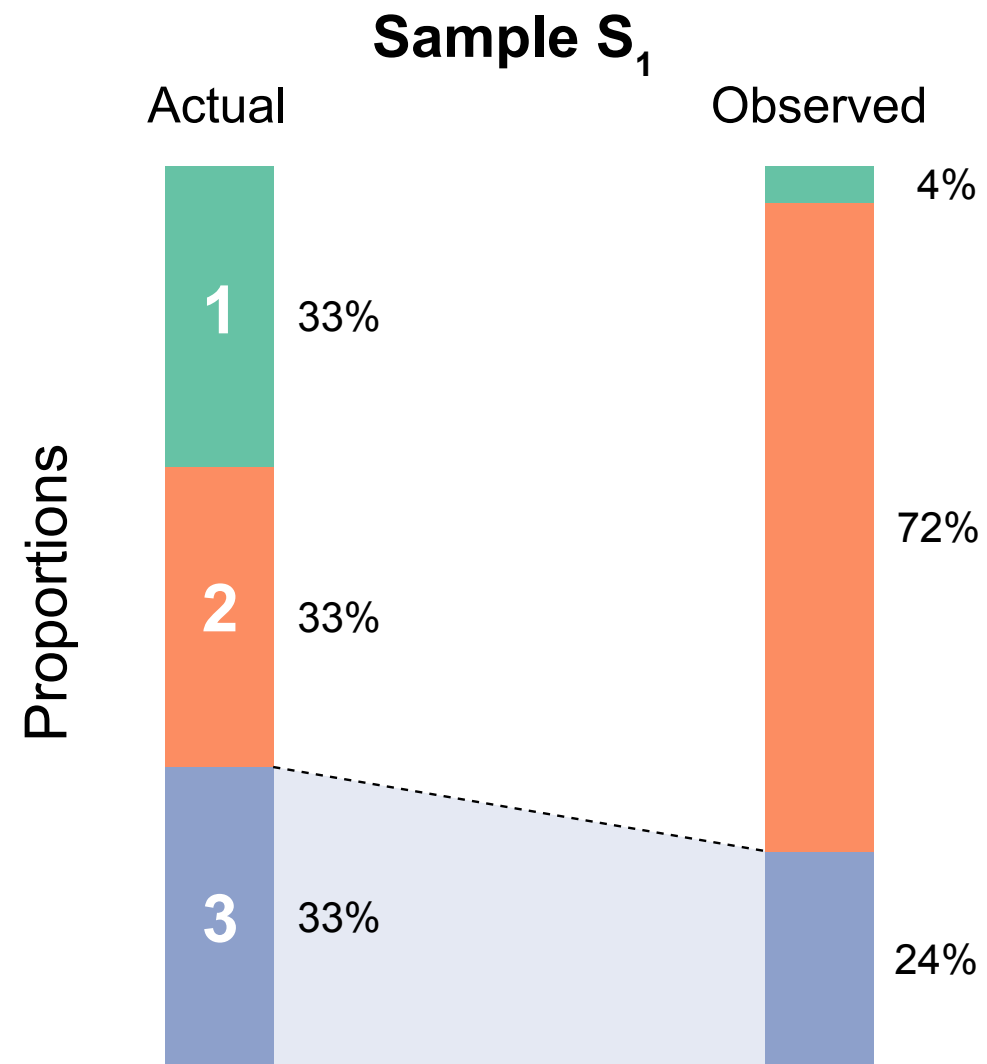
$$Pr(\mathbf{O})_i = \frac{O_i}{\sum_{j=1}^K O_j} = \frac{Pr(\mathbf{A})_i B_i}{\sum_{j=1}^K Pr(\mathbf{A})_j B_j}$$

A simple model of bias links
measurements to the truth

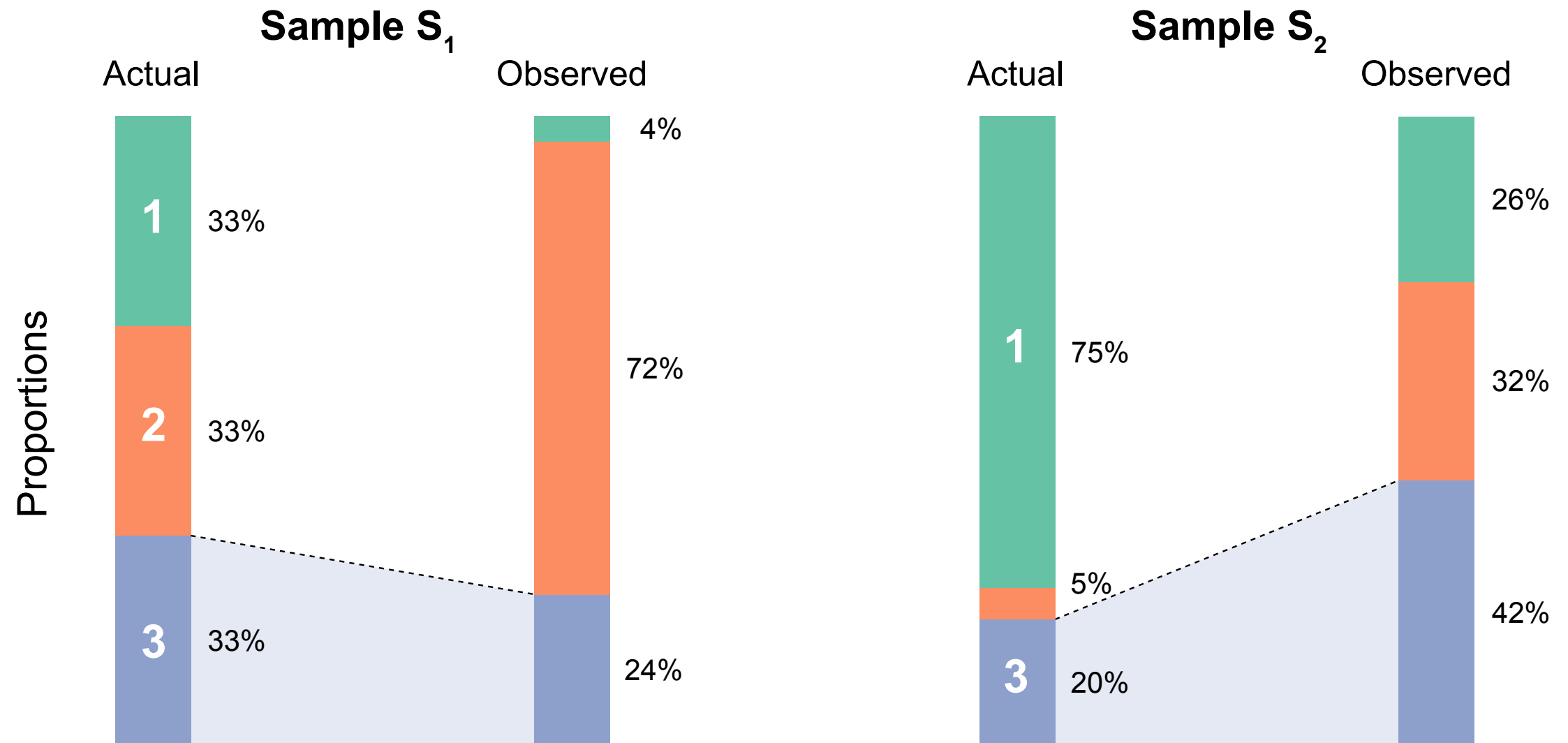
A simple model of bias links measurements to the truth

... at least in simple and well-controlled conditions.

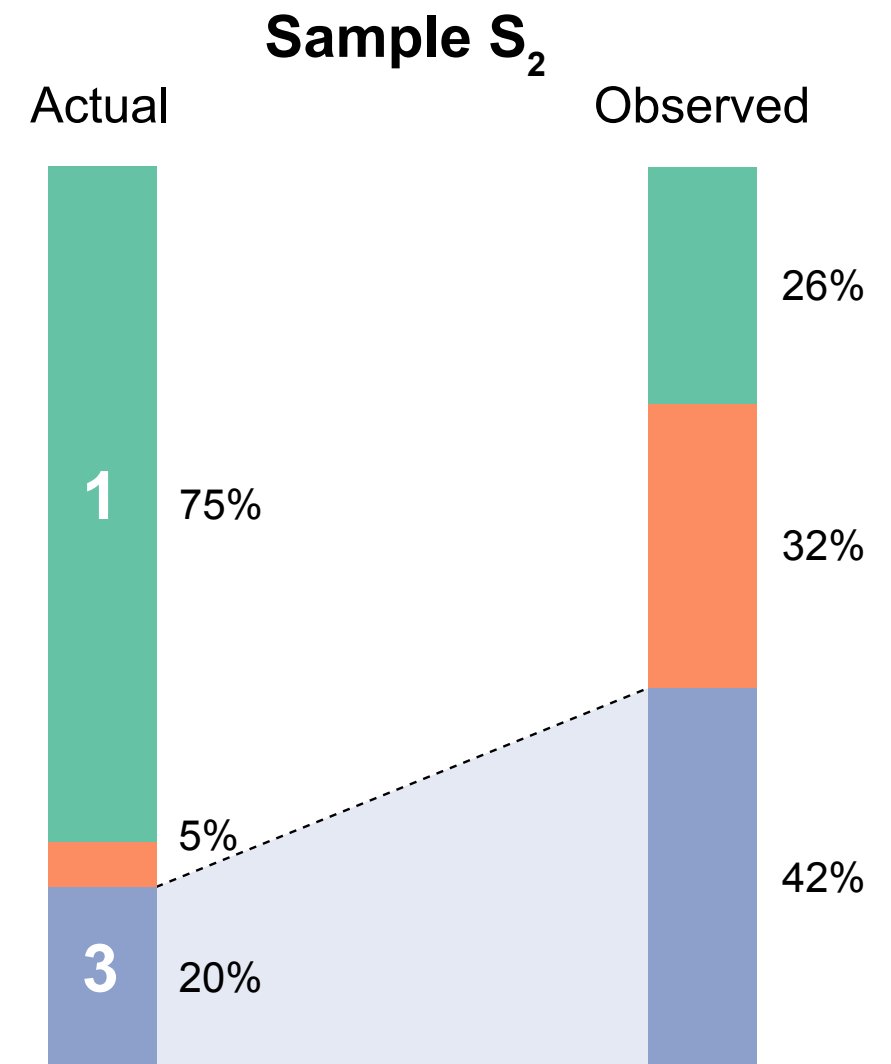
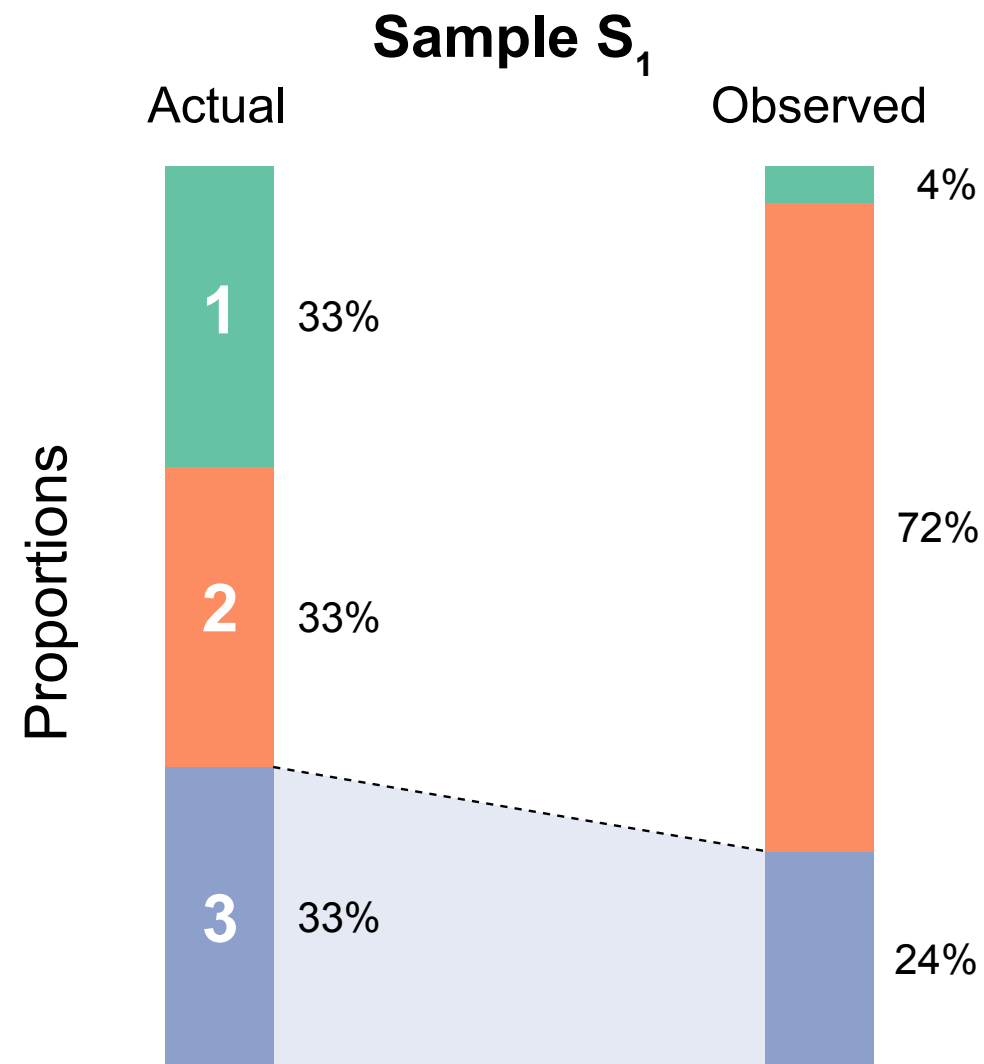
Better estimators for diff-abund



Better estimators for diff-abund



Better estimators for diff-abund

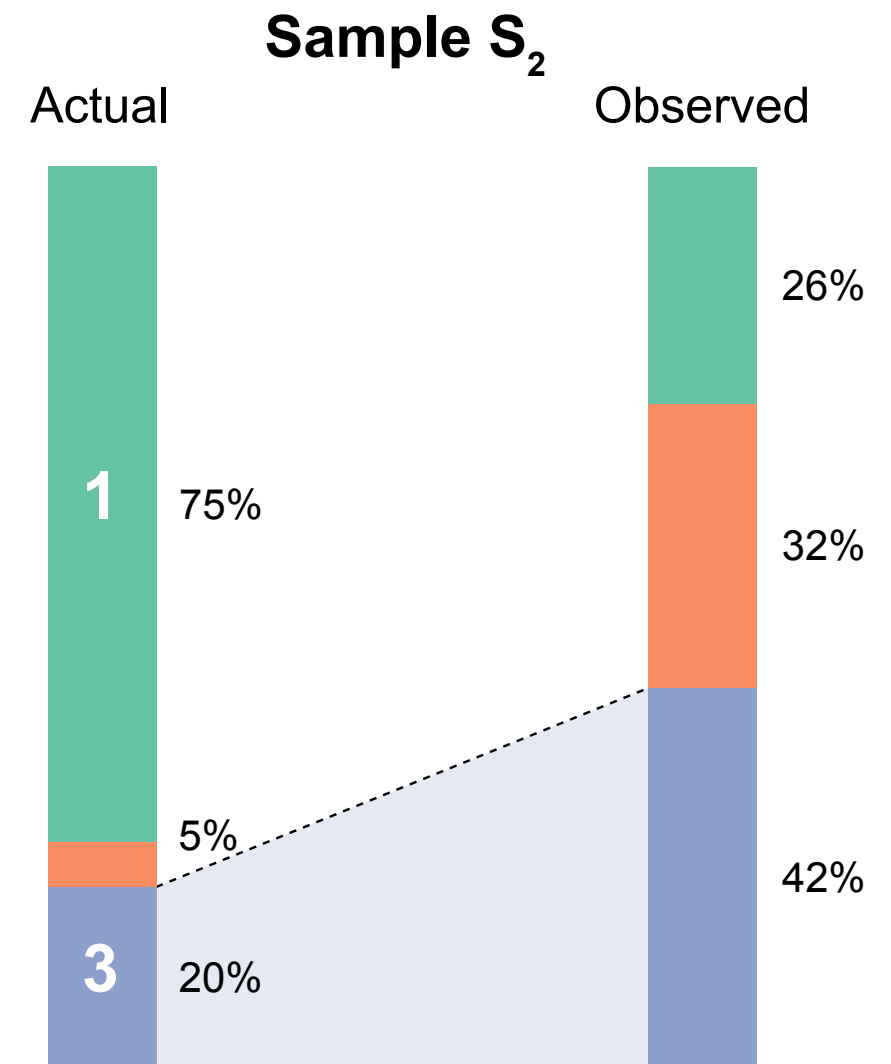
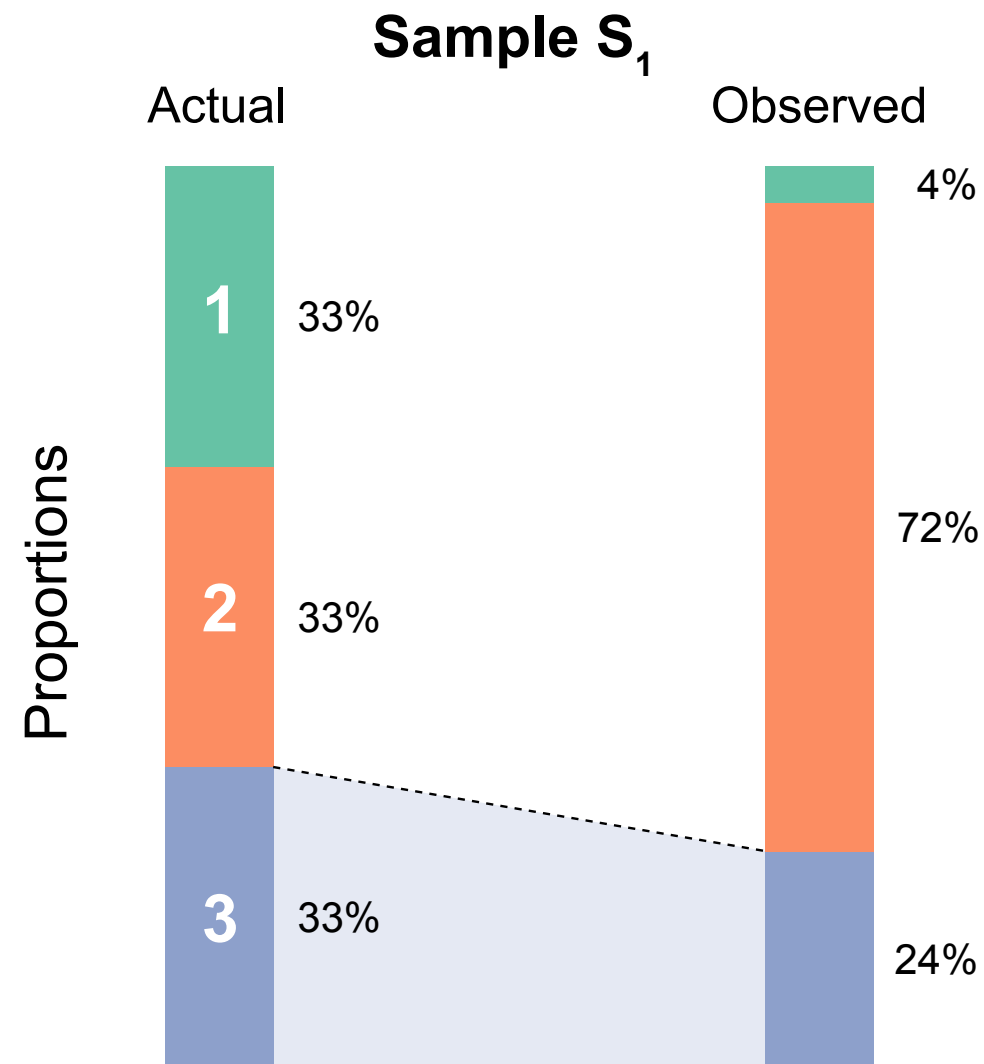


Ratios to
Taxon 1

$$\begin{array}{ccc}
 \text{Actual} & \text{Bias} & \text{Observed} \\
 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} & \cdot \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} & = \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \sim \begin{pmatrix} 0.04 \\ 0.72 \\ 0.24 \end{pmatrix} \\
 \mathbf{A}(S_1) & \mathbf{B} & \mathbf{O}(S_1)
 \end{array}$$

$$\begin{array}{ccc}
 \text{Actual} & \text{Bias} & \text{Observed} \\
 \begin{pmatrix} 1 \\ 1/15 \\ 4/15 \end{pmatrix} & \cdot \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} & = \begin{pmatrix} 1 \\ 18/15 \\ 24/15 \end{pmatrix} \sim \begin{pmatrix} 0.26 \\ 0.32 \\ 0.42 \end{pmatrix} \\
 \mathbf{A}(S_2) & \mathbf{B} & \mathbf{O}(S_2)
 \end{array}$$

Better estimators for diff-abund

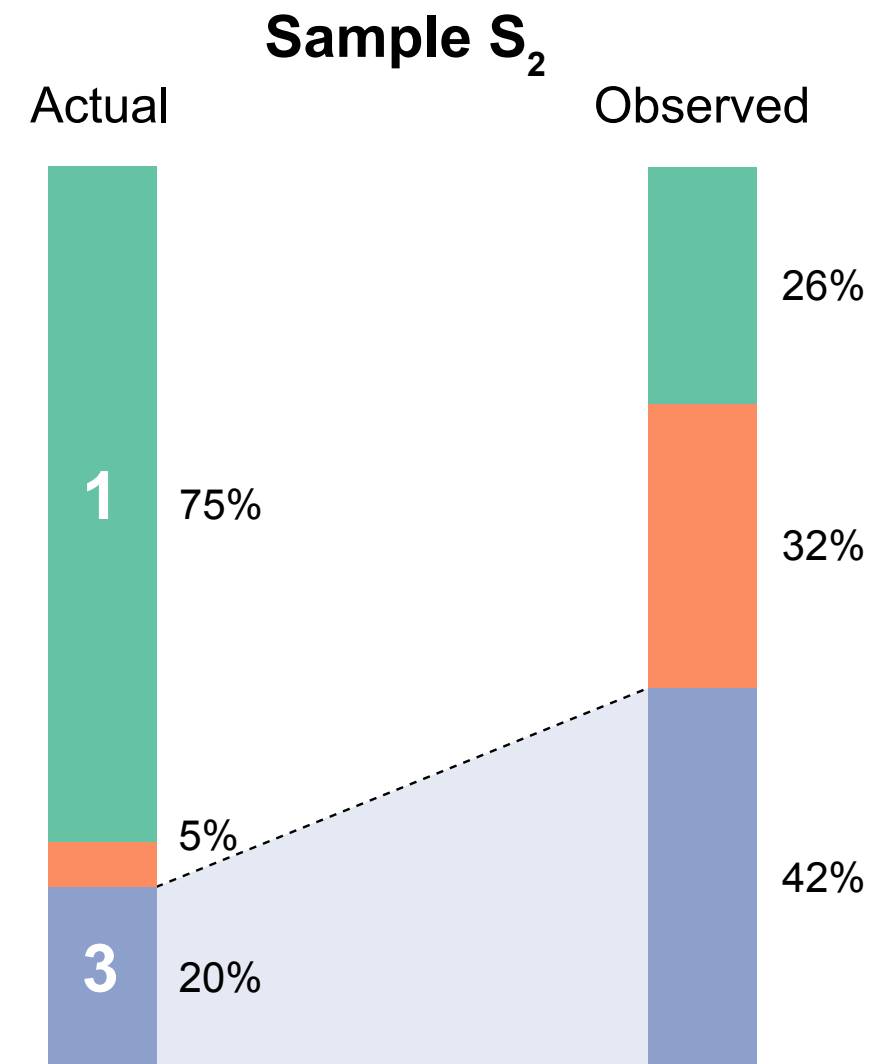
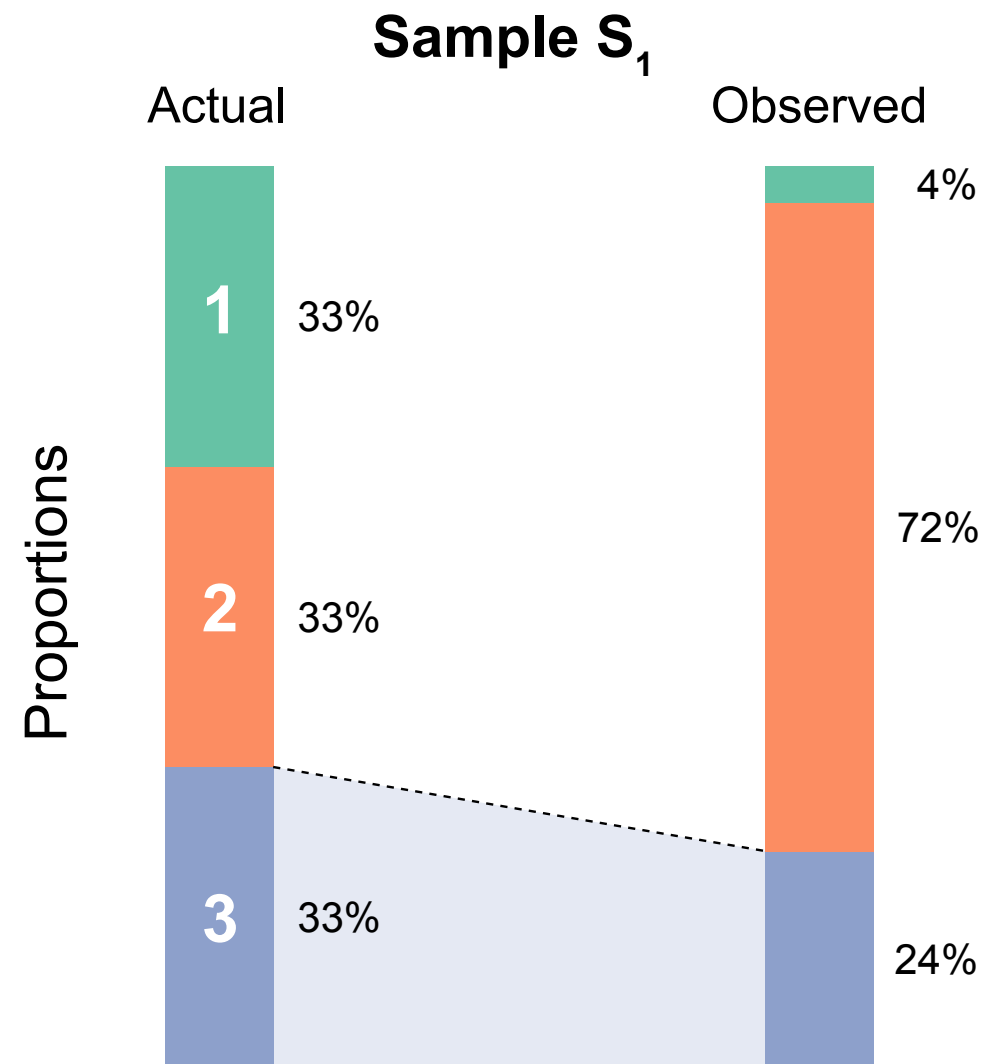


Ratios to
Taxon 1

$$\begin{array}{c} \text{Actual} \\ \blacktriangleright \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ \mathbf{A}(S_1) \end{array} \cdot \begin{array}{c} \text{Bias} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \mathbf{B} \end{array} = \begin{array}{c} \text{Observed} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \mathbf{O}(S_1) \end{array} \sim \begin{pmatrix} 0.04 \\ 0.72 \\ 0.24 \end{pmatrix}$$

$$\begin{array}{c} \text{Actual} \\ \begin{pmatrix} 1 \\ 1/15 \\ 4/15 \end{pmatrix} \\ \mathbf{A}(S_2) \end{array} \cdot \begin{array}{c} \text{Bias} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \mathbf{B} \end{array} = \begin{array}{c} \text{Observed} \\ \begin{pmatrix} 1 \\ 18/15 \\ 24/15 \end{pmatrix} \\ \mathbf{O}(S_2) \end{array} \sim \begin{pmatrix} 0.26 \\ 0.32 \\ 0.42 \end{pmatrix}$$

Better estimators for diff-abund



Ratios to
Taxon 1

$$\begin{array}{c} \text{Actual} \\ \rightarrow \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ \mathbf{A}(S_1) \end{array} \cdot \begin{array}{c} \text{Bias} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \mathbf{B} \end{array} = \begin{array}{c} \text{Observed} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \mathbf{O}(S_1) \end{array} \sim \begin{pmatrix} 0.04 \\ 0.72 \\ 0.24 \end{pmatrix}$$

$$\begin{array}{c} \text{Actual} \\ \rightarrow \begin{pmatrix} 1 \\ 1/15 \\ 4/15 \end{pmatrix} \\ \mathbf{A}(S_2) \end{array} \cdot \begin{array}{c} \text{Bias} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \mathbf{B} \end{array} = \begin{array}{c} \text{Observed} \\ \begin{pmatrix} 1 \\ 18/15 \\ 24/15 \end{pmatrix} \\ \mathbf{O}(S_2) \end{array} \sim \begin{pmatrix} 0.26 \\ 0.32 \\ 0.42 \end{pmatrix}$$

Better estimators for diff-abund

Sample S_1

Actual

Observed

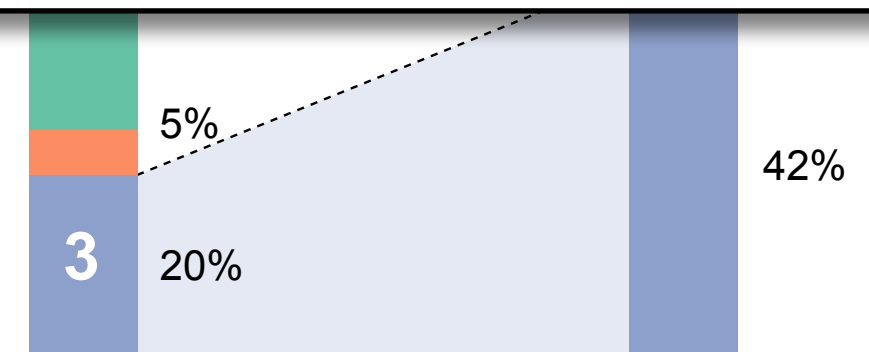
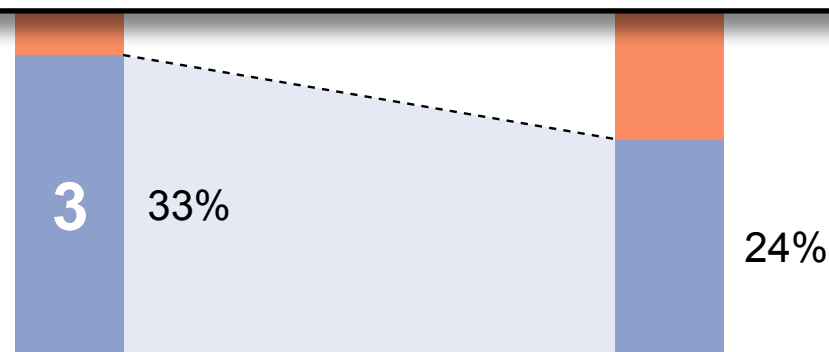
Sample S_2

Actual

Observed

Ratios are *consistently* affected by bias.

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



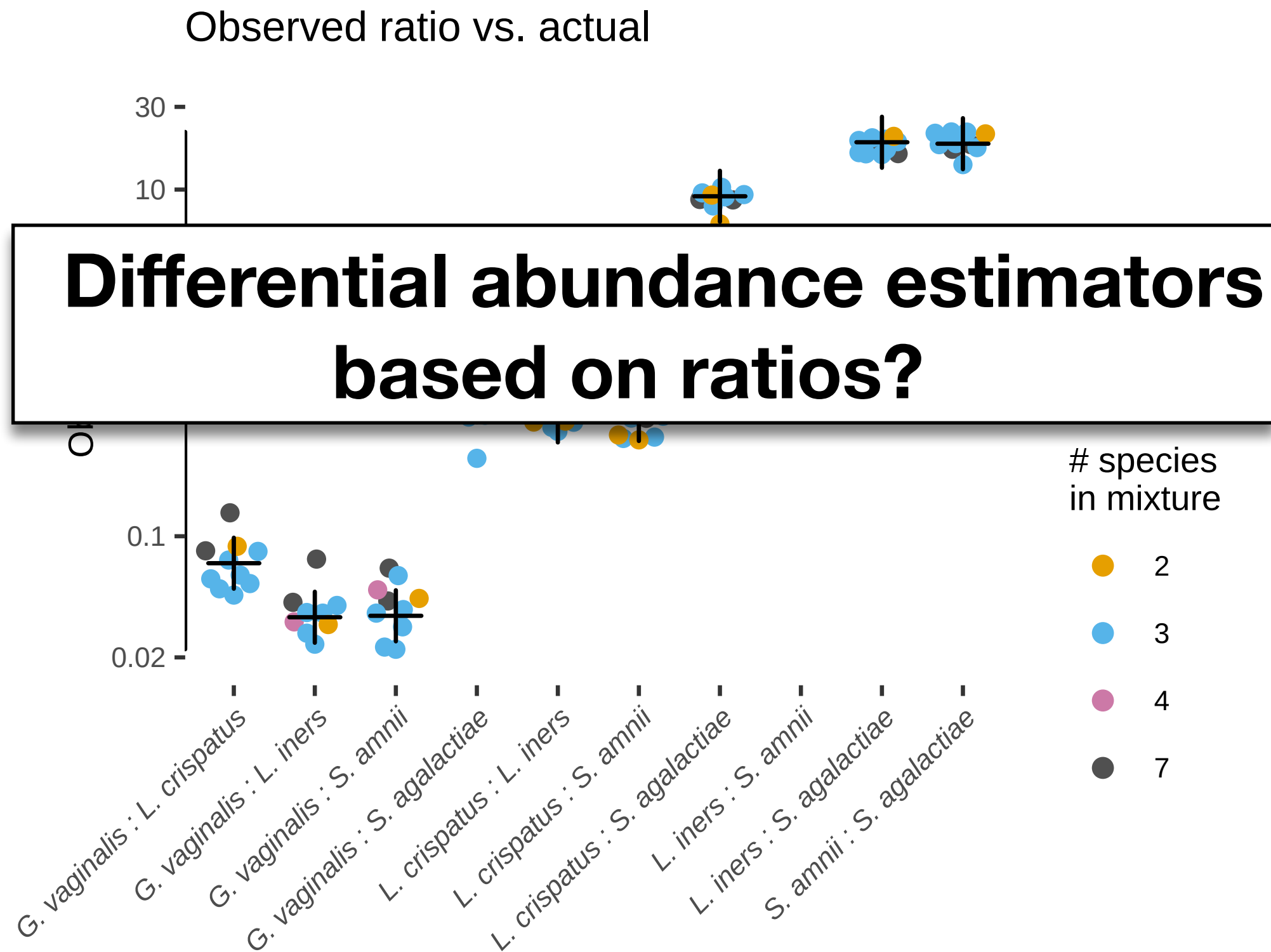
Ratios to
Taxon 1

$$\begin{array}{c} \text{Actual} \\ \rightarrow \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ \text{A}(S_1) \end{array} \cdot \begin{array}{c} \text{Bias} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \text{B} \end{array} = \begin{array}{c} \text{Observed} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \text{O}(S_1) \end{array} \sim \begin{pmatrix} 0.04 \\ 0.72 \\ 0.24 \end{pmatrix}$$

$$\begin{array}{c} \text{Actual} \\ \rightarrow \begin{pmatrix} 1 \\ 1/15 \\ 4/15 \end{pmatrix} \\ \text{A}(S_2) \end{array} \cdot \begin{array}{c} \text{Bias} \\ \begin{pmatrix} 1 \\ 18 \\ 6 \end{pmatrix} \\ \text{B} \end{array} = \begin{array}{c} \text{Observed} \\ \begin{pmatrix} 1 \\ 18/15 \\ 24/15 \end{pmatrix} \\ \text{O}(S_2) \end{array} \sim \begin{pmatrix} 0.26 \\ 0.32 \\ 0.42 \end{pmatrix}$$

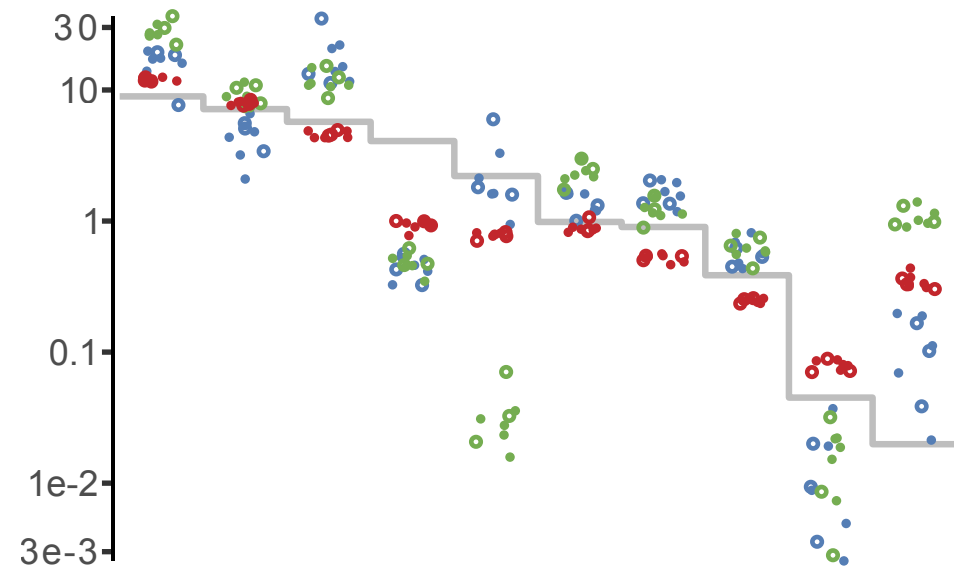


Better estimators for diff-abund



Metagenomic Calibration

Observed vs. True Abundances



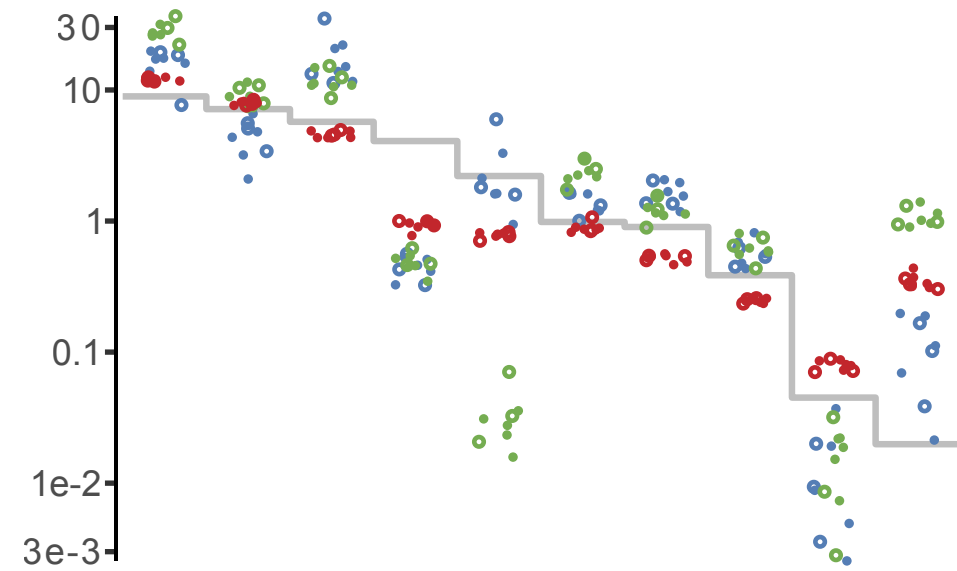
Protocol

H
Q
W

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

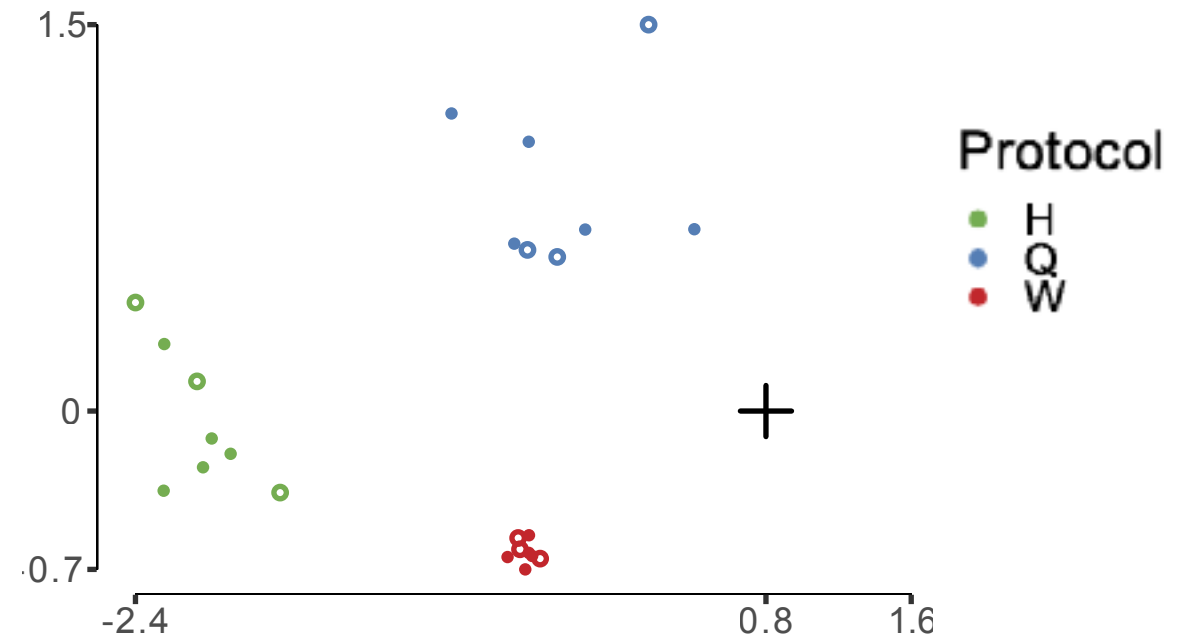
Metagenomic Calibration

Observed vs. True Abundances



Biased

Sample Ordination



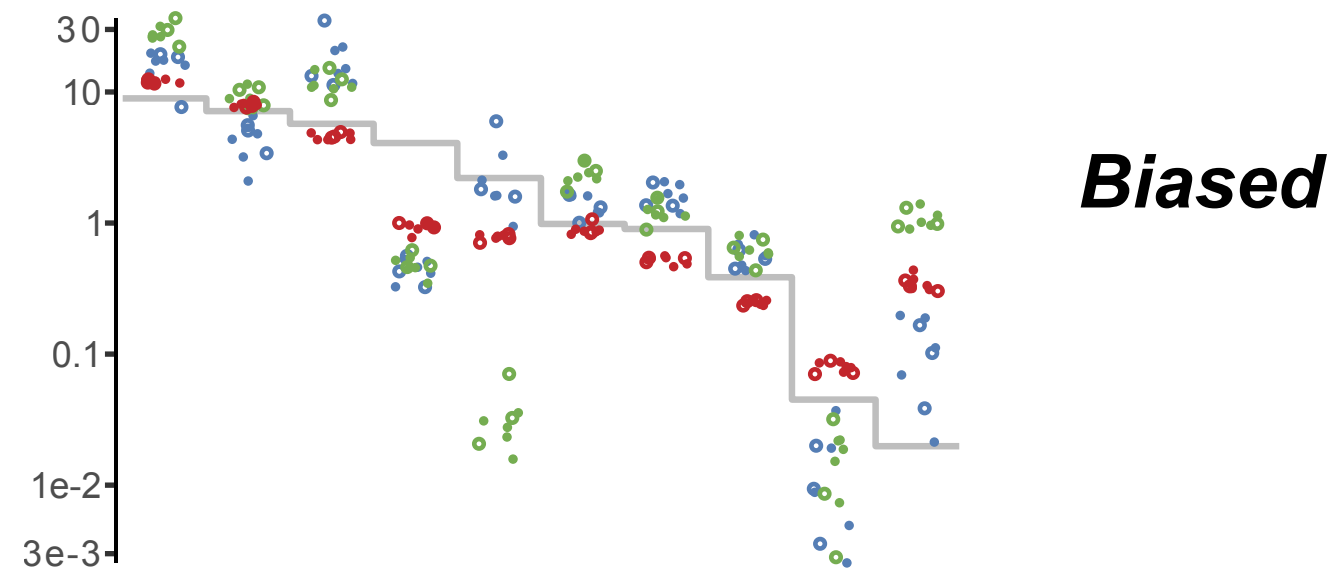
Protocol

H
Q
W

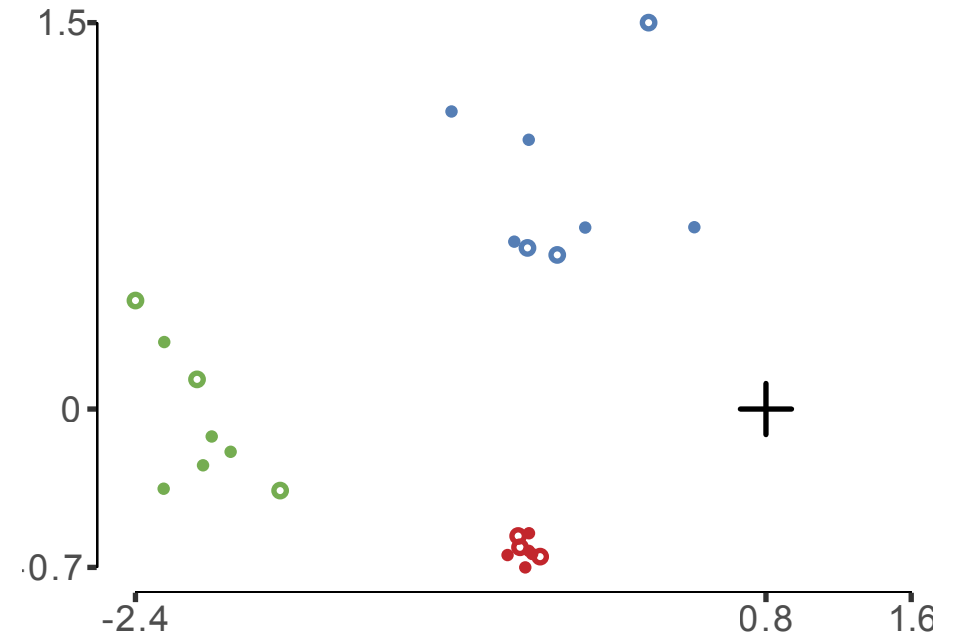
Data: Costea, et al. *Nature Biotechnology*, 2017.

Metagenomic Calibration

Observed vs. True Abundances



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

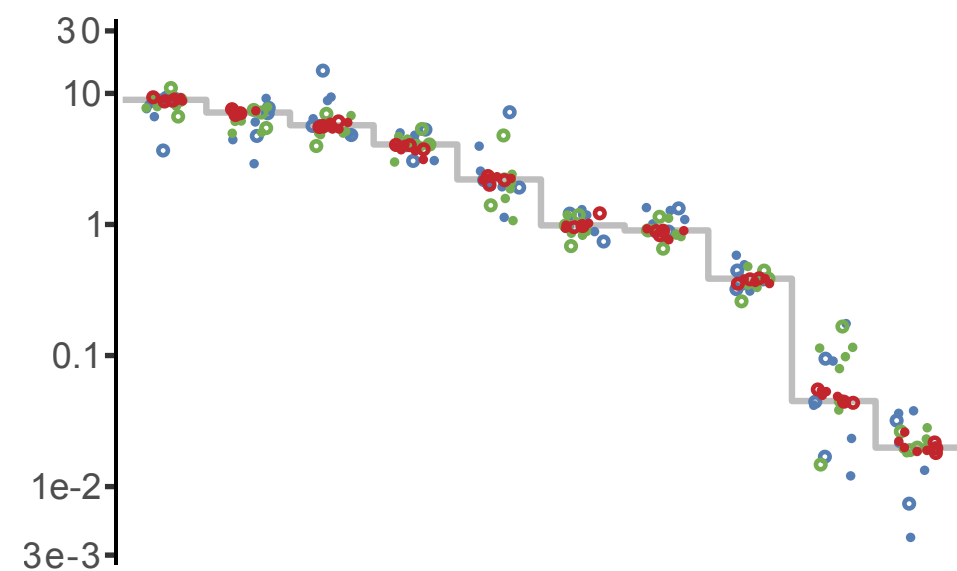
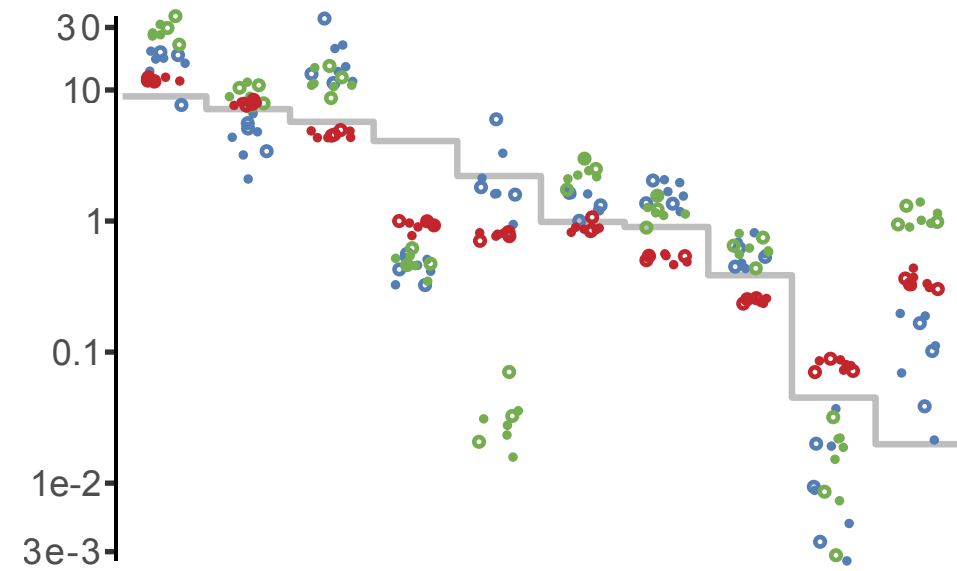
Protocol

● H
● Q
● W

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

Metagenomic Calibration

Observed vs. True Abundances

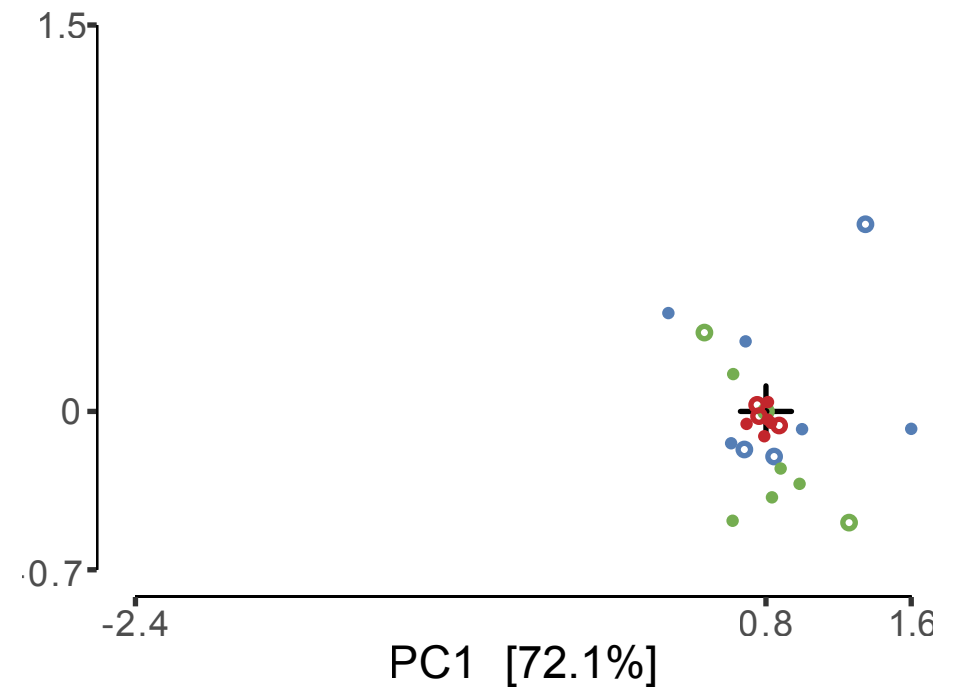
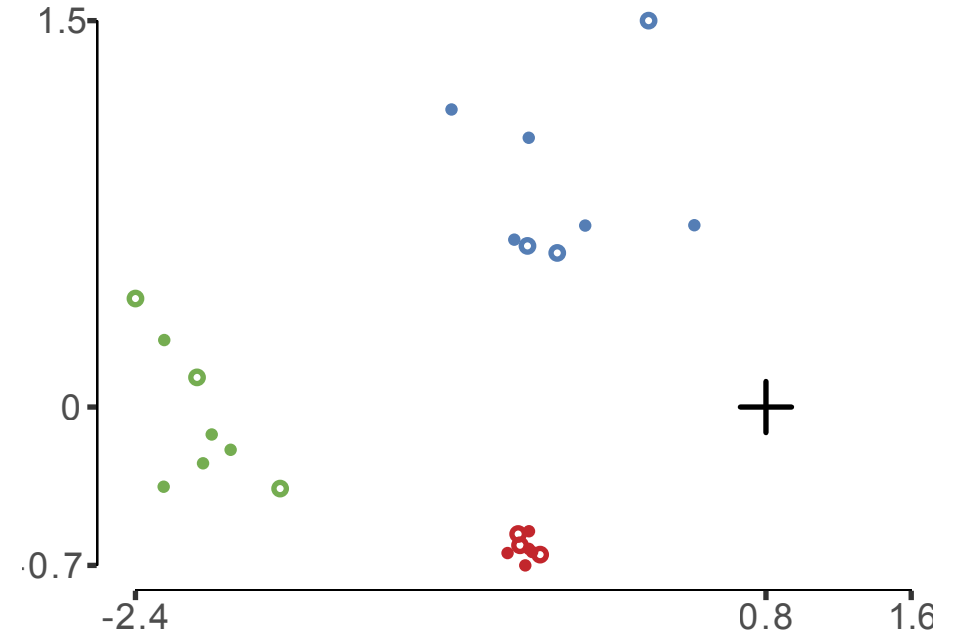


P. melaninogenica
C. perfringens
S. enterica
C. difficile
L. plantarum
V. cholerae
C. saccharolyticum
Y. pseudotuberculosis
B. hansenii
F. nucleatum

Protocol

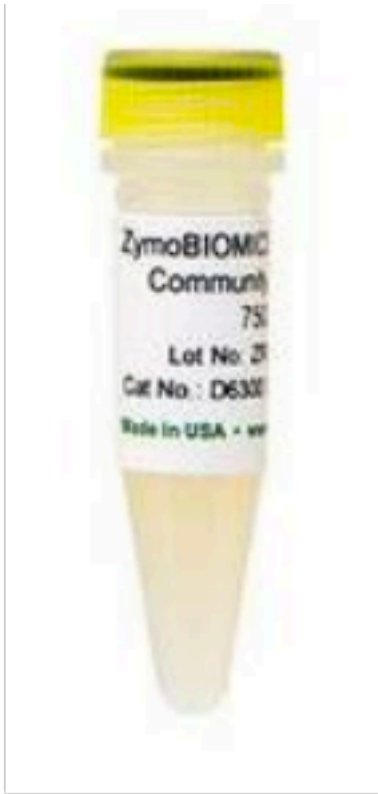
H
Q
W

Sample Ordination



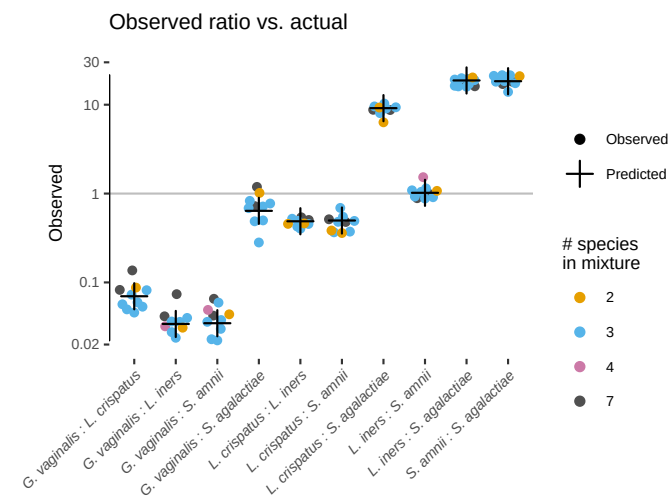
Data: Costea, et al. *Nature Biotechnology*, 2017.

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.

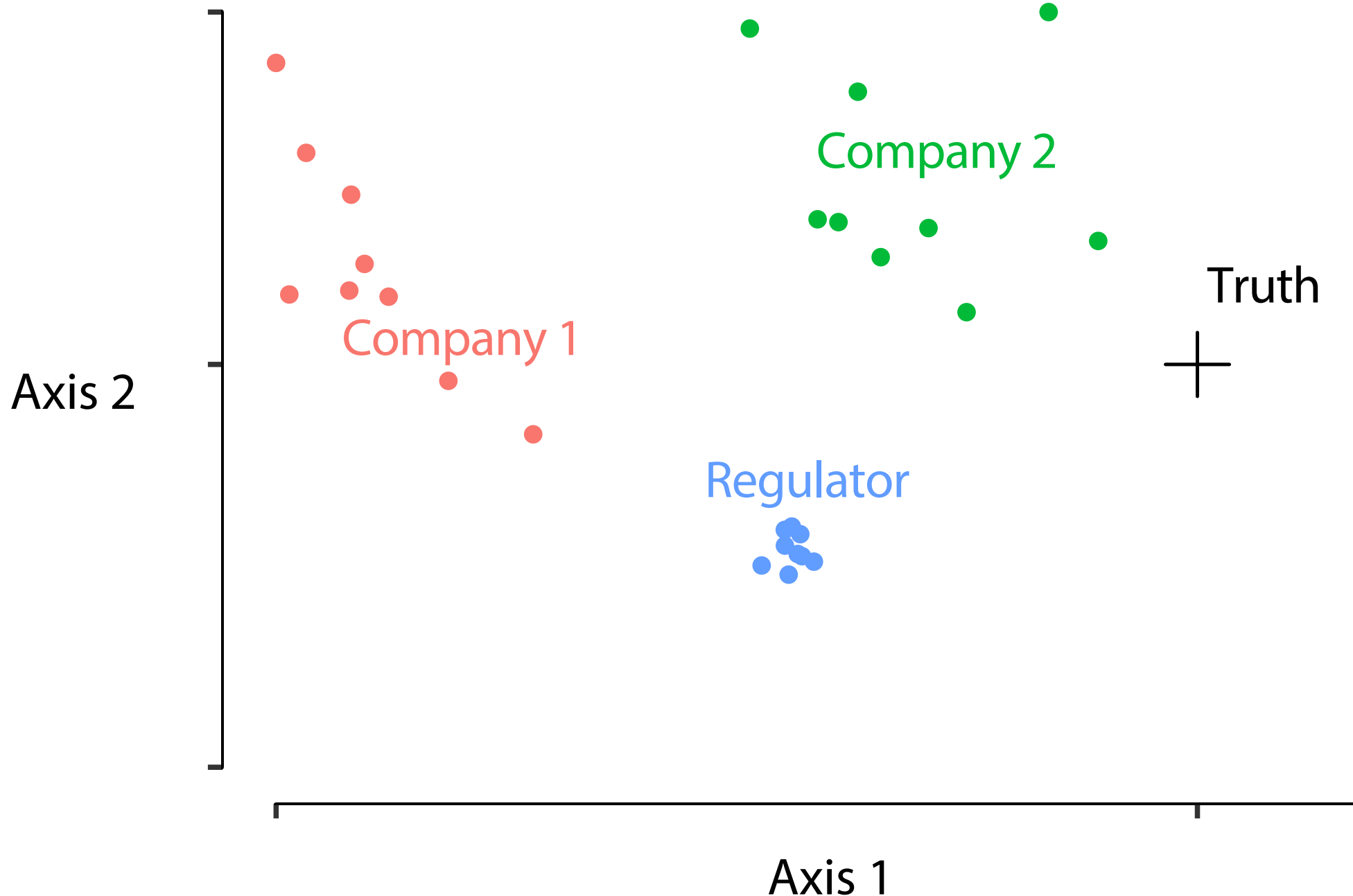
- Standard samples that are more than a process control?

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!



Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren¹, Amy D Willis², Benjamin J Callahan^{1,3*}

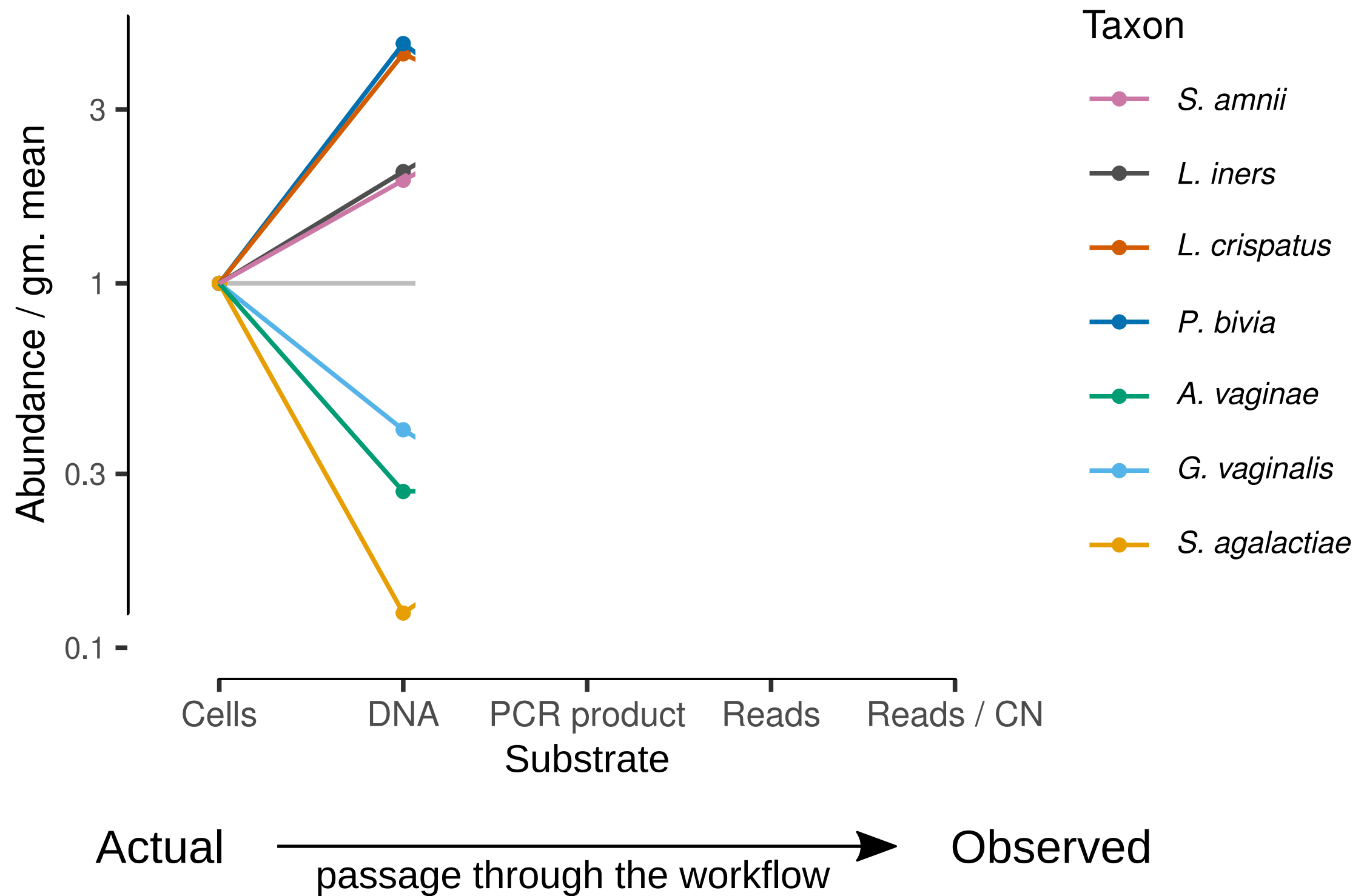


McLaren, Nearing, Willis, Lloyd, Callahan (2022). “Implications of taxonomic bias for microbial differential-abundance analysis”. *bioRxiv*. <https://doi.org/10.1101/2022.08.19.504330>

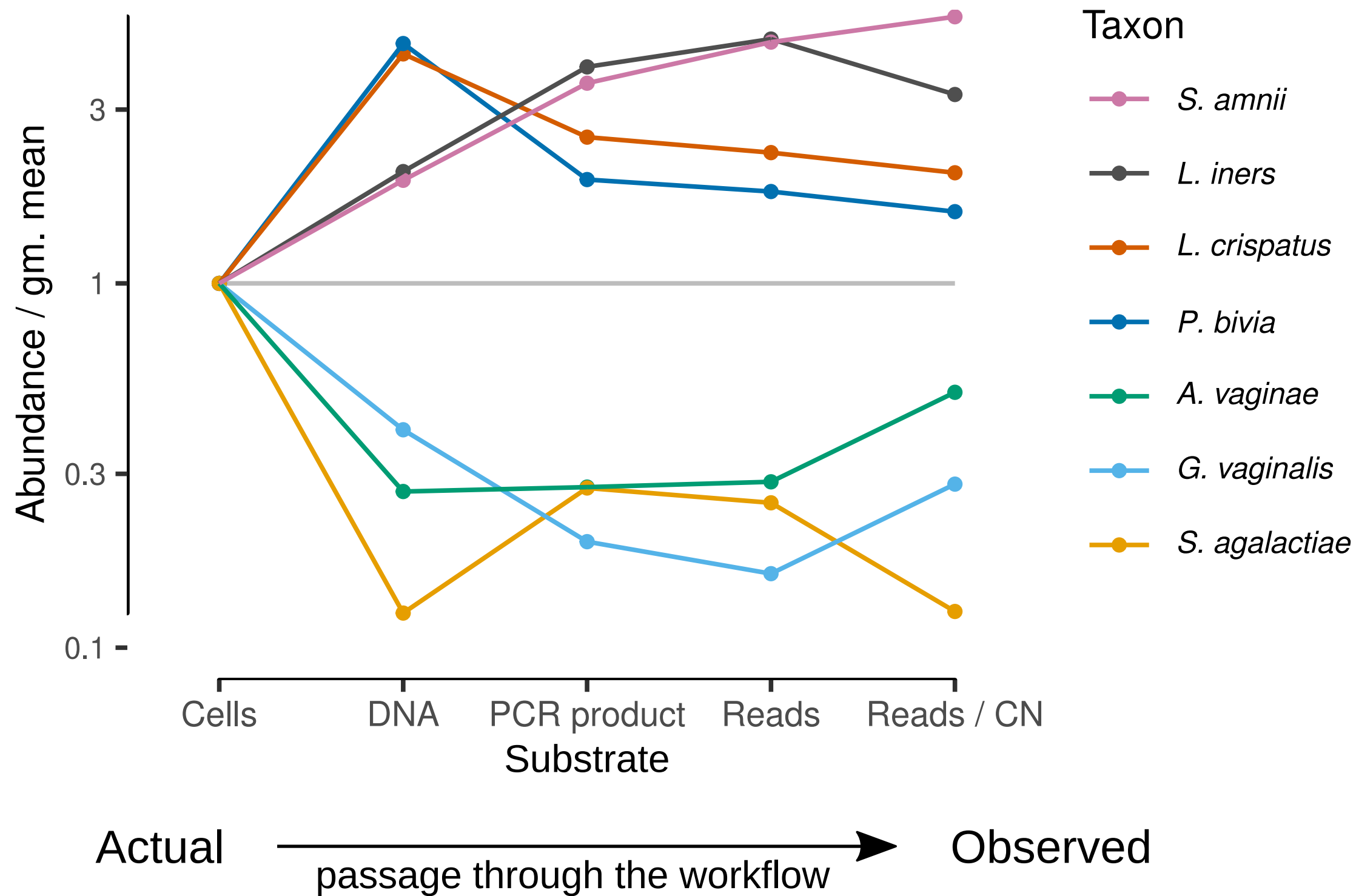
Williamson, Hughes, Willis (2021). “A multiview model for relative and absolute microbial abundances”. *Biometrics*, (2021).
<https://doi.org/10.1111/biom.13503>

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

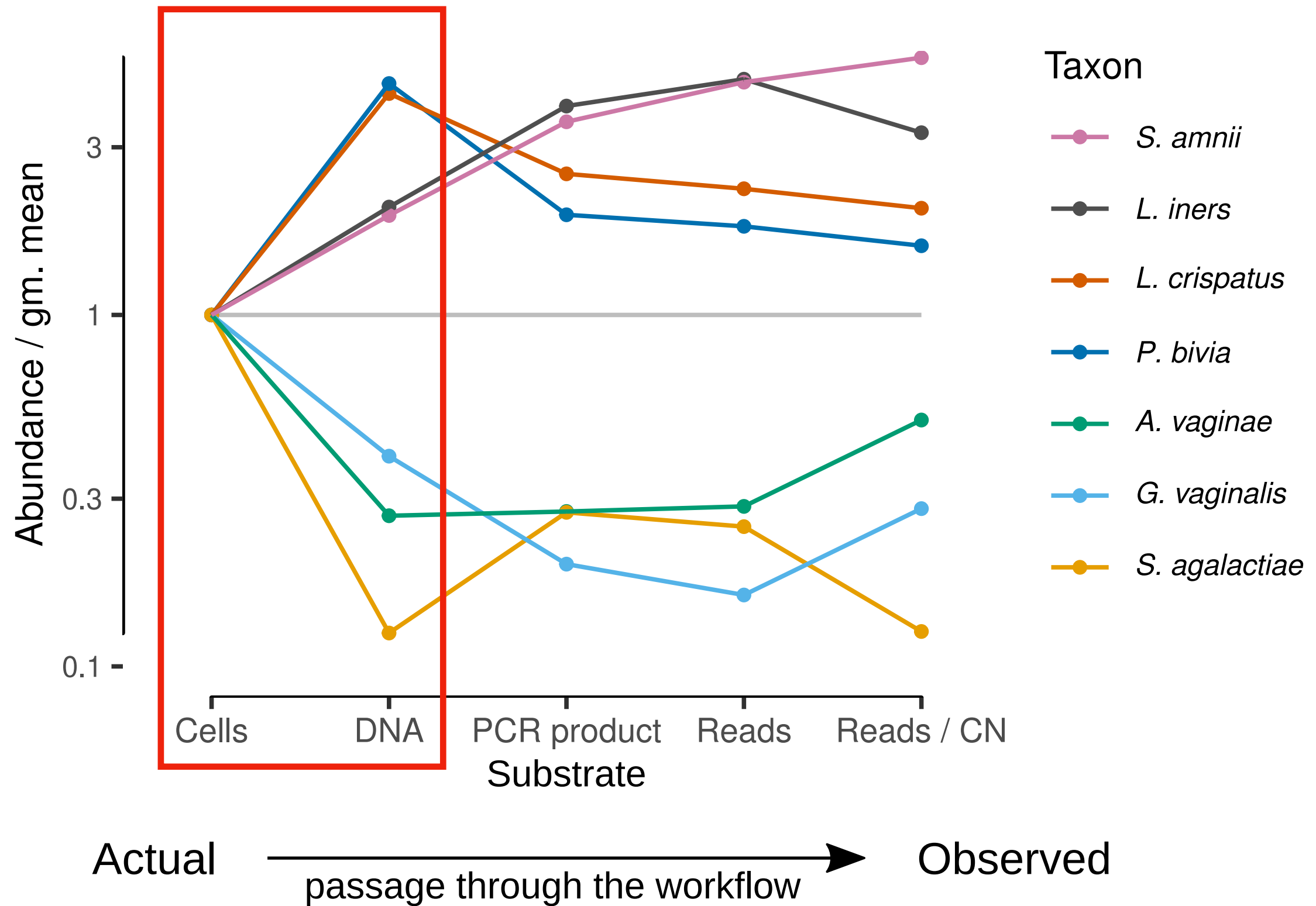
Protocol Optimization



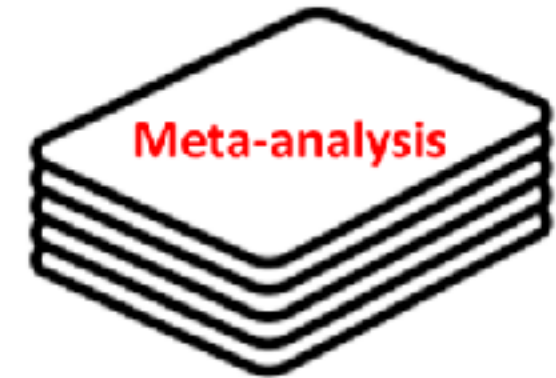
Protocol Optimization



Protocol Optimization



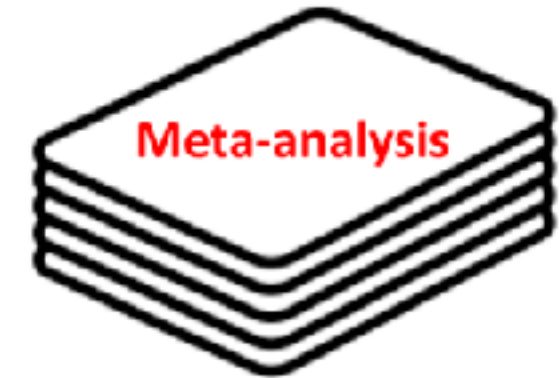
Synthesizing Across Studies



- 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.

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.

3

Different Methods = Different Biases

Image credit: Vaginal Microbiome Consortium.

Synthesizing Across Studies

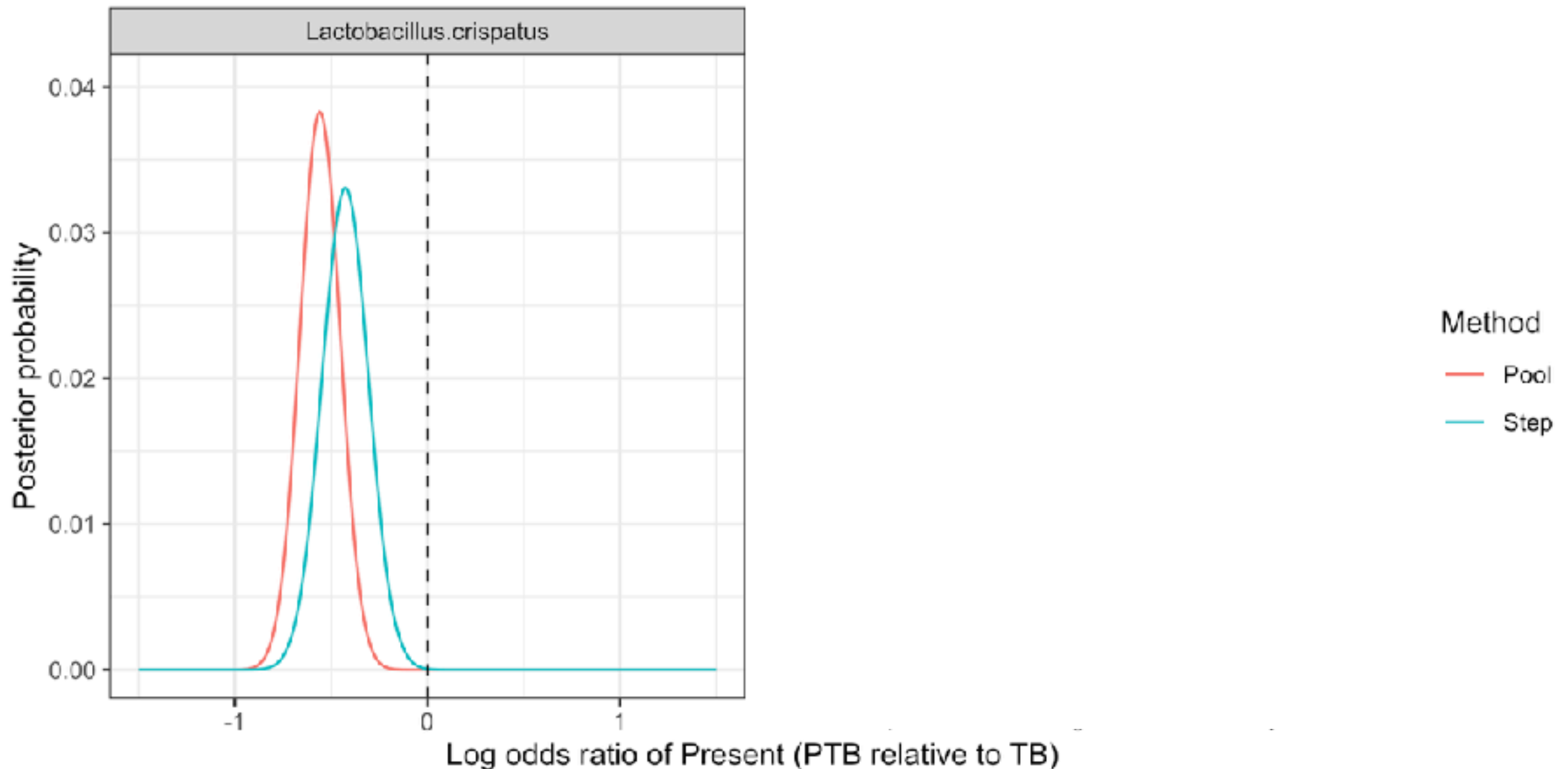
All-study differential prevalence analysis, Bayesian approach

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

Synthesizing Across Studies

All-study differential prevalence analysis, Bayesian approach

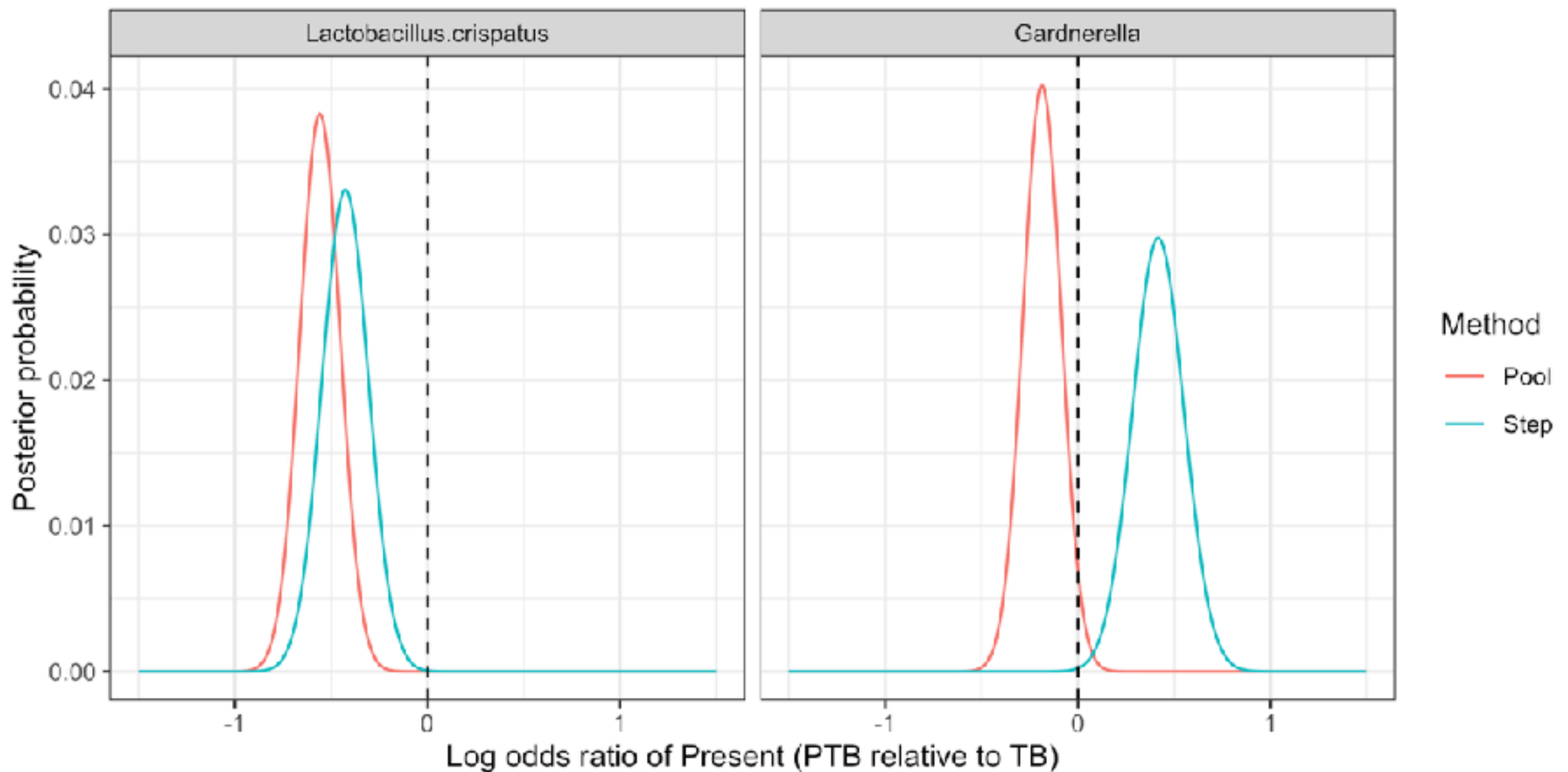
- **Pool**: Assume **same** detection rate in each study
- **Step**: Allow for **different** detection rates across studies



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

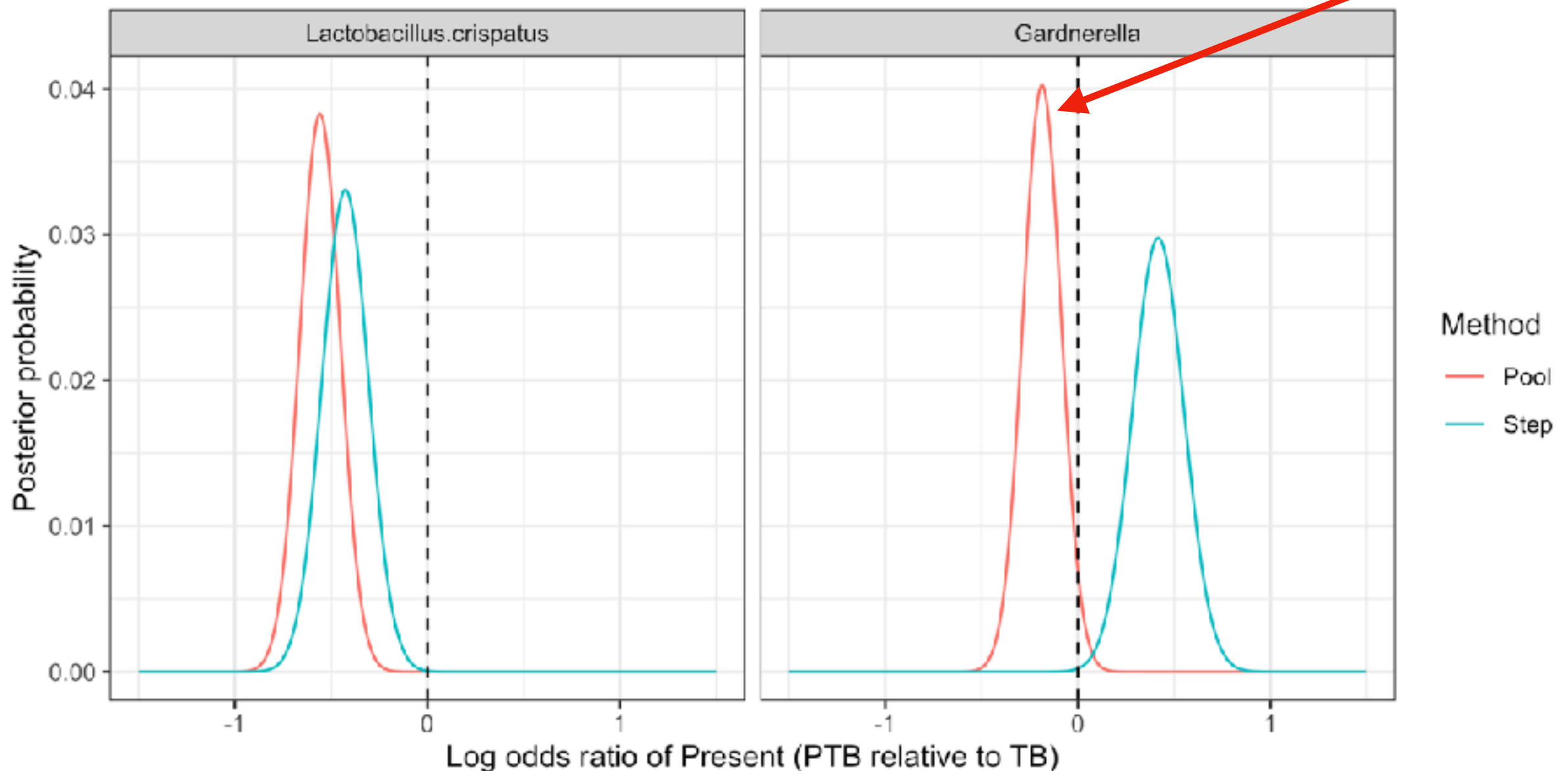


Synthesizing Across Studies

All-study differential prevalence analysis, Bayesian approach

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

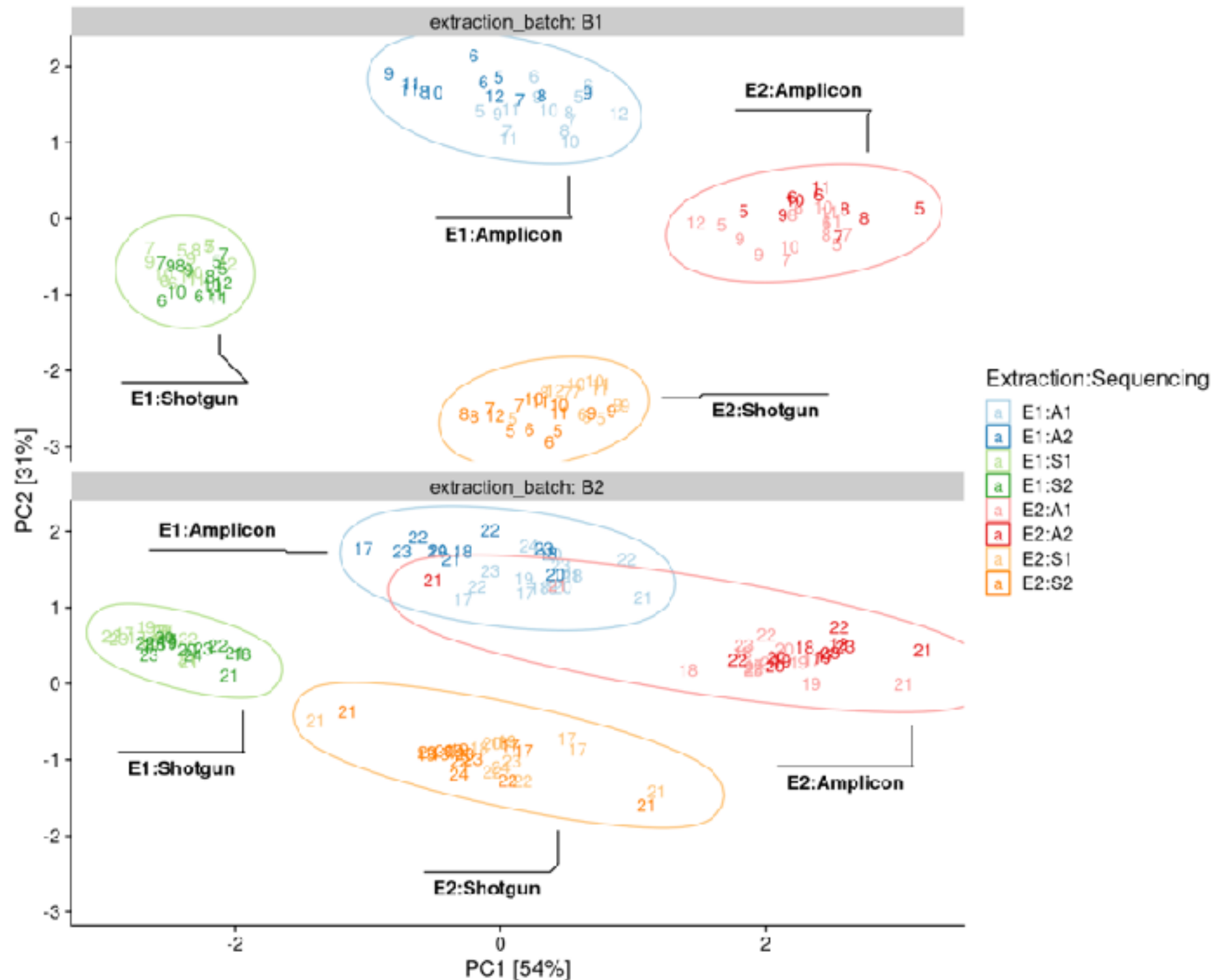
Sign Error!



Towards True Calibration?

Observed bias among extraction:sequencing combinations

PCA of CLR observations after subtracting specimen composition (for an arbitrary reference protocol)

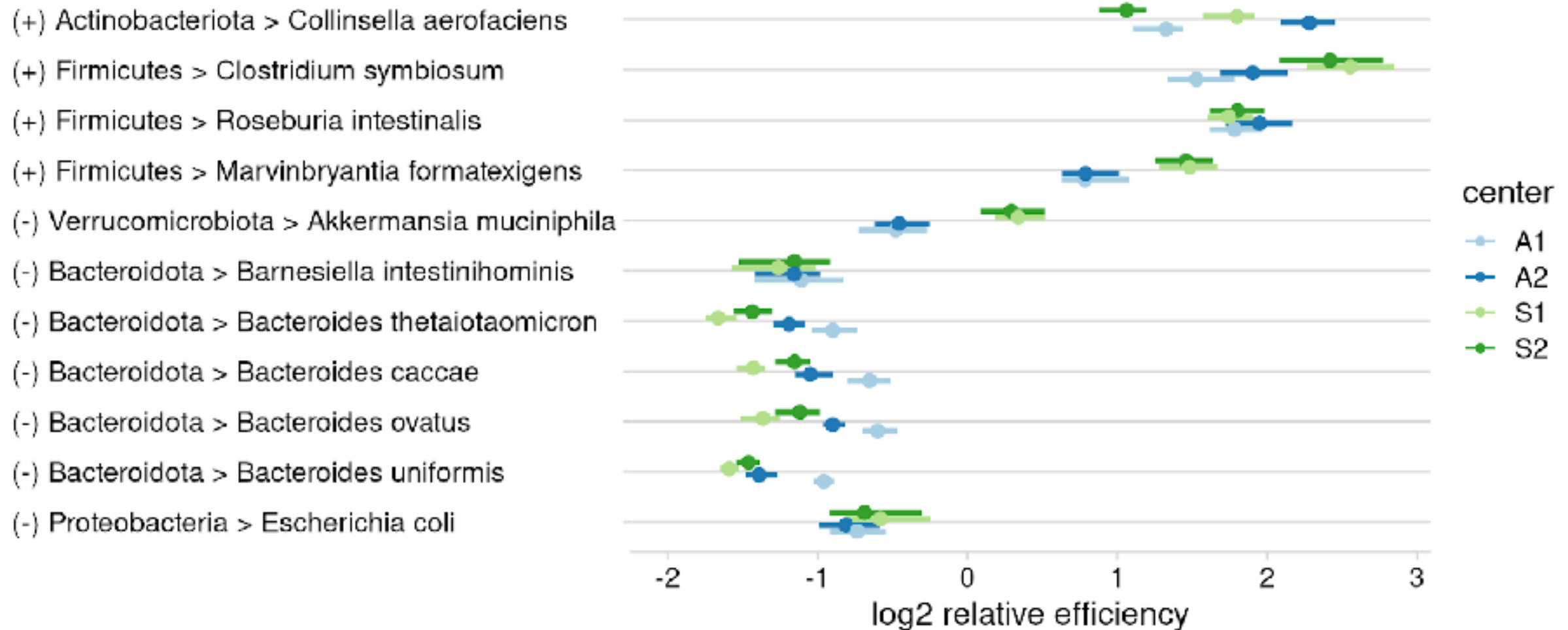


Acknowledgement: Michael McLaren, Angie Mordant, Manuel Kleiner.

Towards True Calibration?

Differential extraction bias conditional on sequencing center

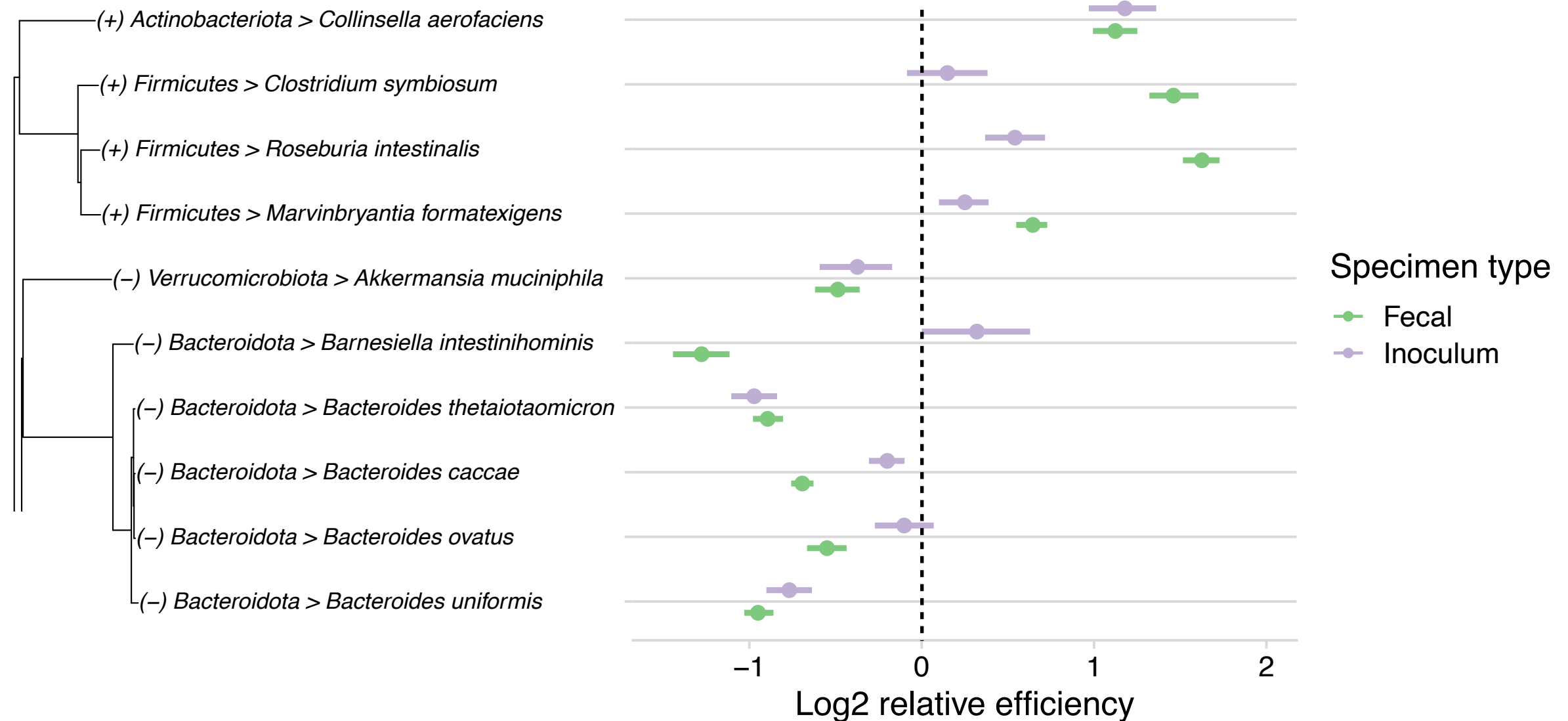
Efficiencies are relative to the average taxon (i.e., they are centered log ratios)



Acknowledgement: Michael McLaren, Angie Mordant, Manuel Kleiner.

A note of Caution

Differential extraction bias (E2/E1) in fecal and inoculum samples



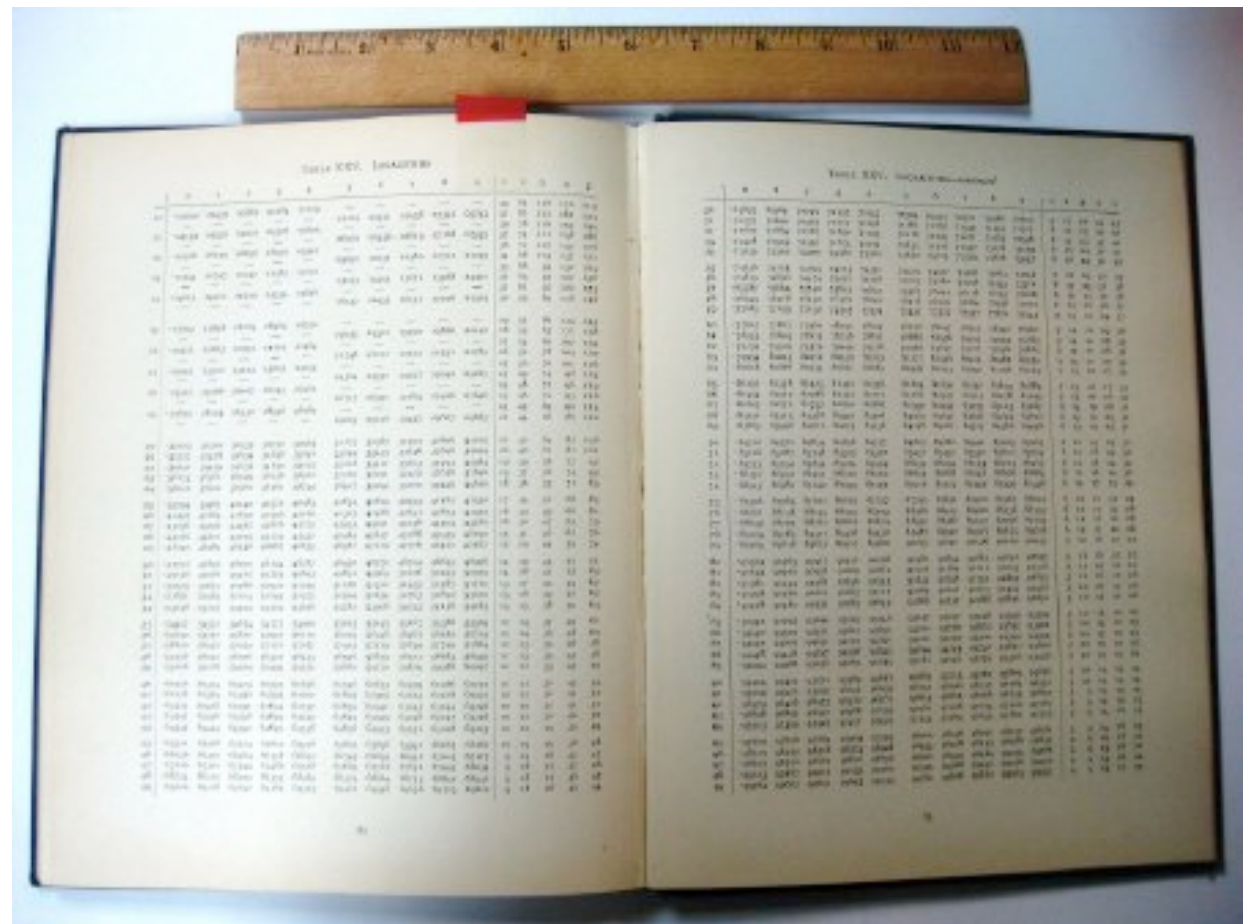
Acknowledgement: Michael McLaren, Angie Mordant, Manuel Kleiner.

Quantitative Qs about B(ias)

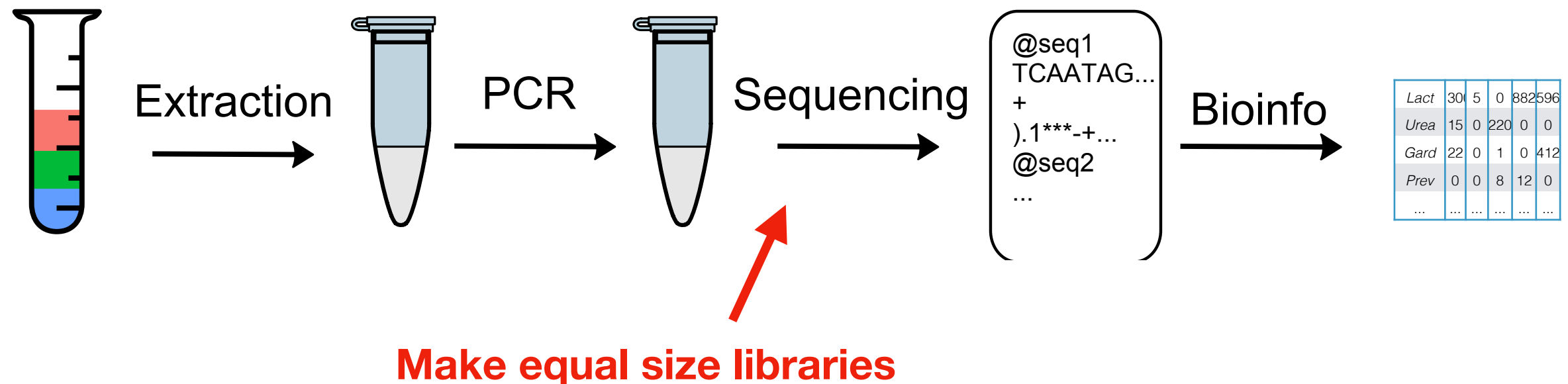
- **Scale? (within and between protocols)**
- **Phylogenetic coherence?**
- **Predictability?**

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. Macklaim¹, Vera Pawlowsky-Glahn² and Juan J. Egozcue³

¹Department of Biochemistry, University of Western Ontario, London, ON, Canada

²Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain

³Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

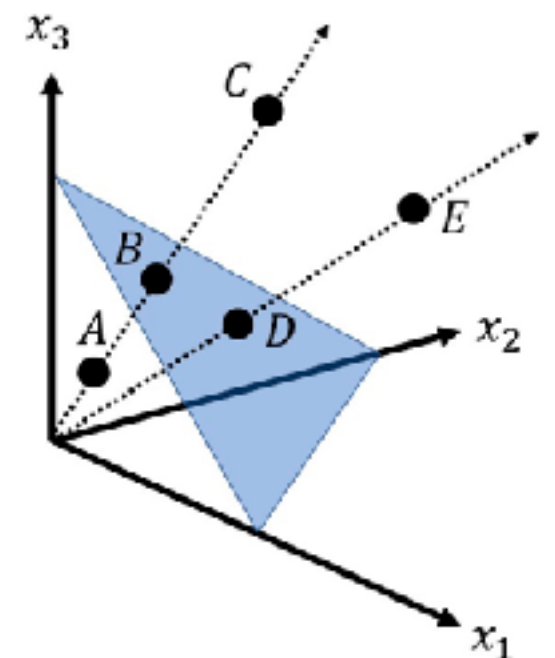


Image credit: Wikipedia.

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Log-ratio Transforms:

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

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

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

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Log-ratio Transforms:

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

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

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

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

Compositional Data Analysis (CoDA)

- Scale invariance
- Perturbation invariance
- Sub-compositional coherence

Log-ratio Transforms:

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

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

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

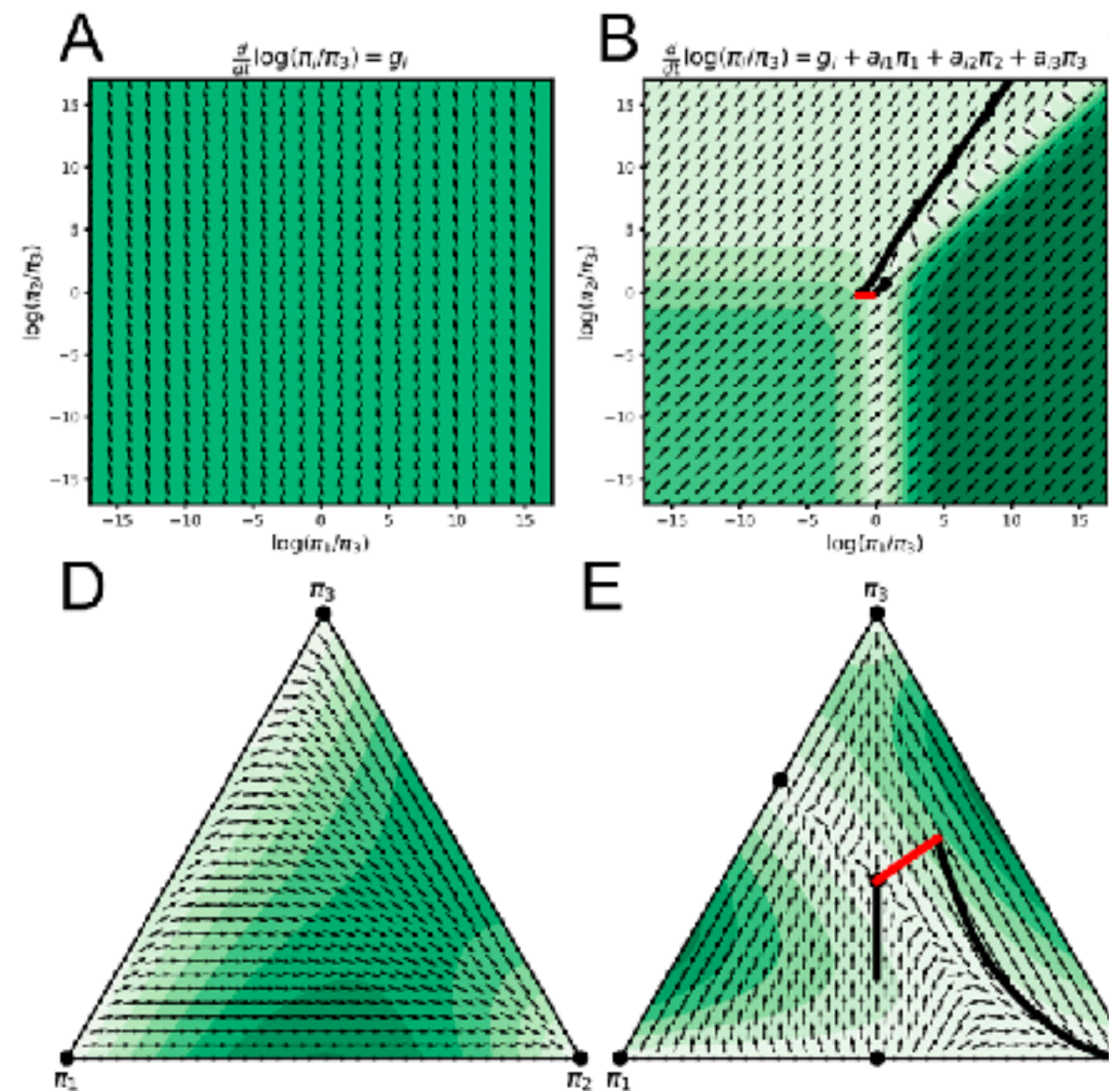
How to marry theory and compositional measurements?

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

Compositional Modeling

Compositional Lotka-Volterra describes microbial dynamics in the simplex

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



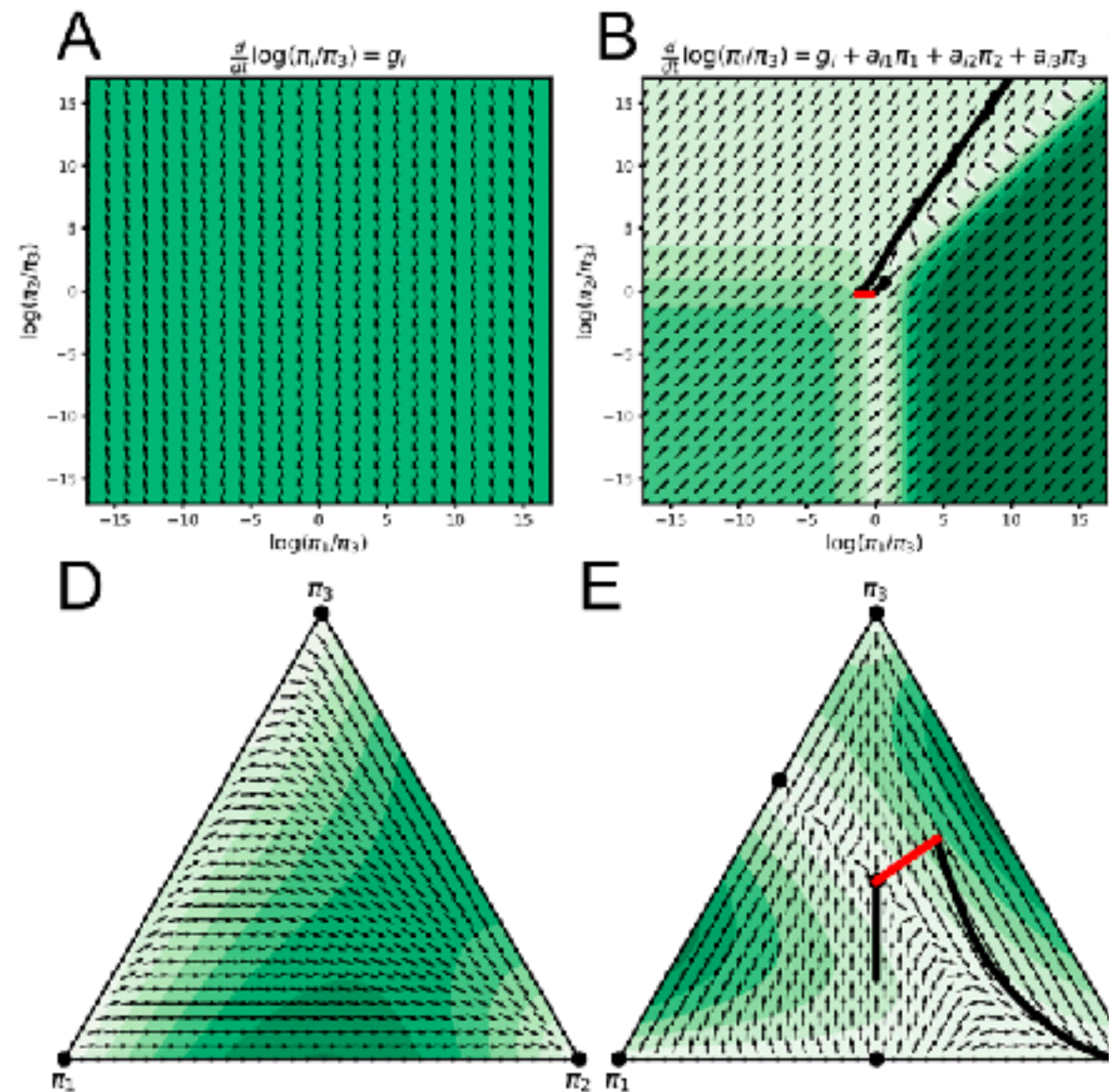
Credit: Joseph et al., *PLoS Comp Bio*, 2020

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, $\text{Var}(N(t)) = \mathbb{E}[(N(t) - 1)^2]$, is low. Then, the parameters of cLV approximately correspond to differences in parameters of gLV.”



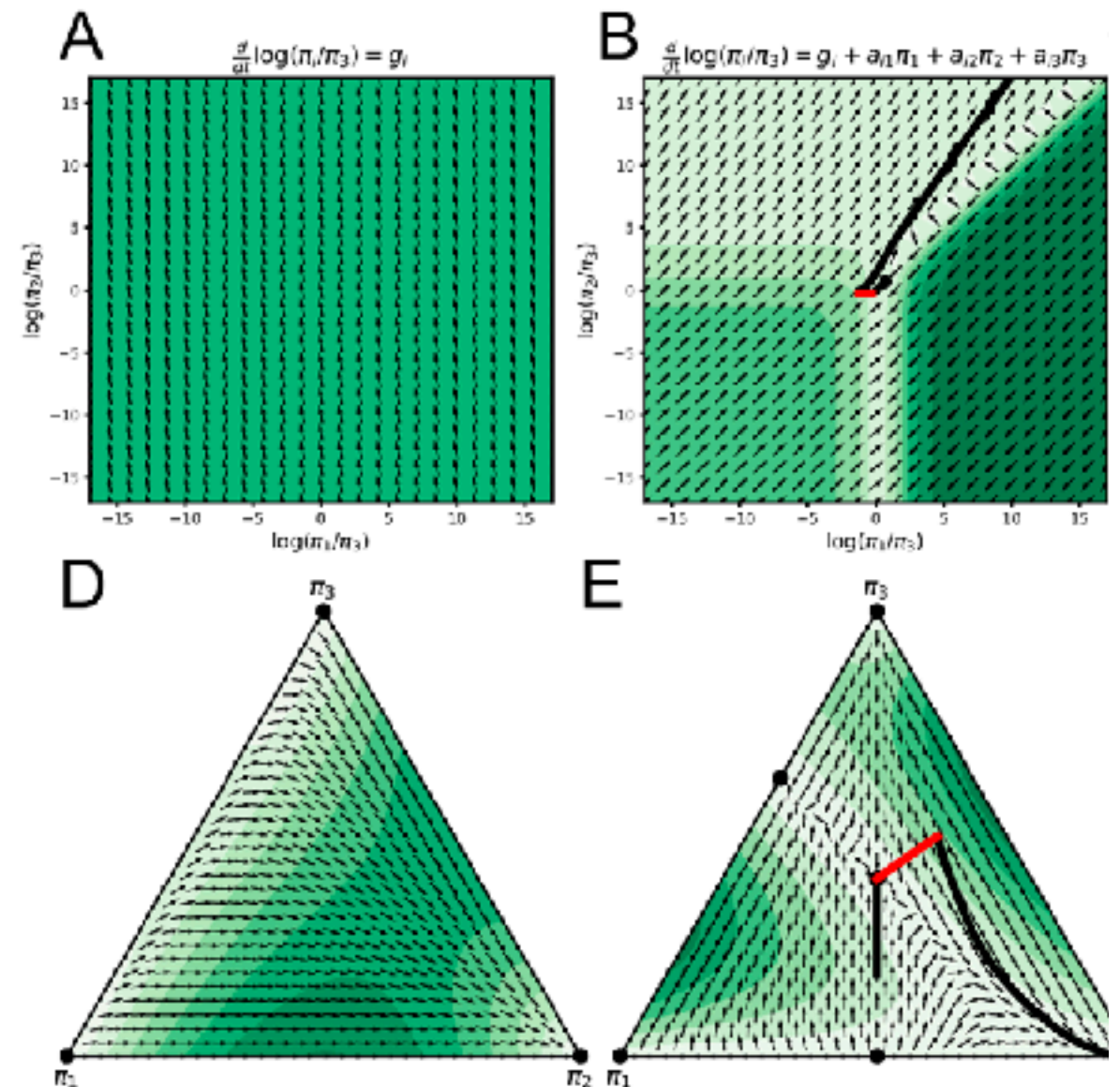
Compositional Modeling

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, $\text{Var}(N(t)) = \mathbb{E}[(N(t) - 1)^2]$, 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.



Trouble with zeros

$$\log(\mathbf{x}/0) = -\log(0/\mathbf{x}) = \text{bad}$$

Trouble with zeros

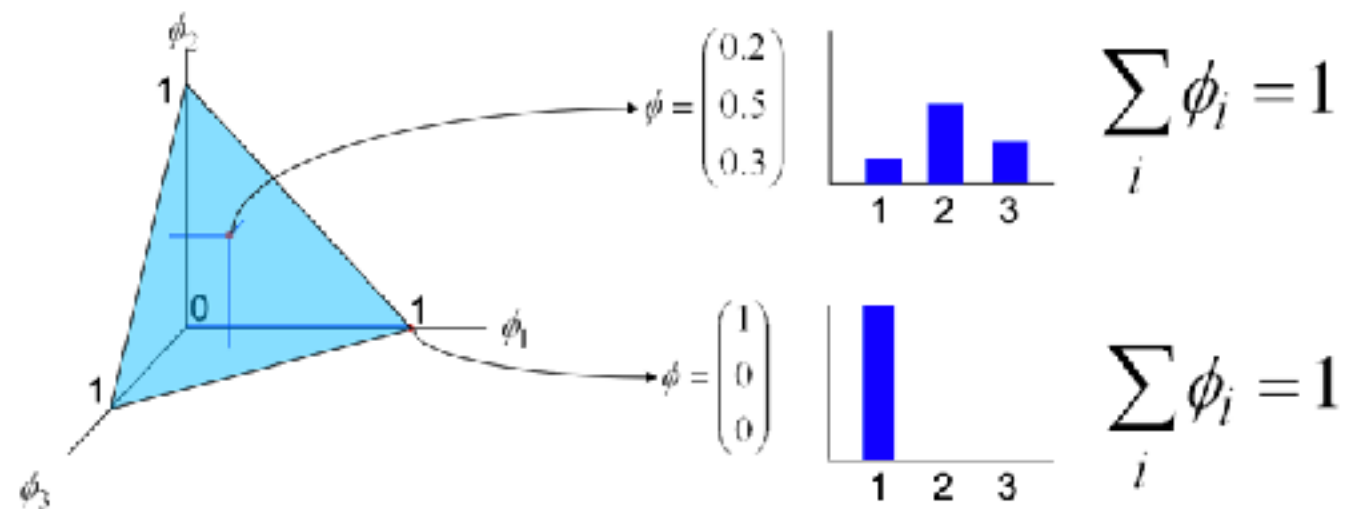
$$\log(\mathbf{x}/0) = -\log(0/\mathbf{x}) = \text{bad}$$

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:



Trouble with zeros

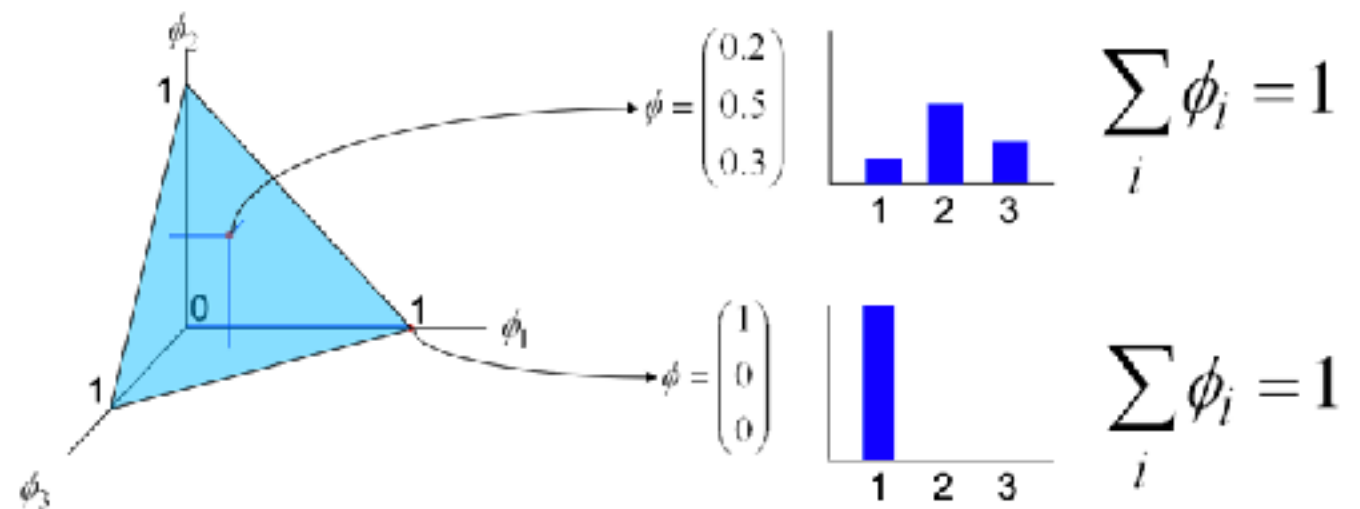
$$\log(\mathbf{x}/0) = -\log(0/\mathbf{x}) = \text{bad}$$

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

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

Bias and CoDA

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

$$\mathbf{O} \sim \mathbf{A} \cdot \underbrace{\mathbf{B}^{(P)}}_{\text{Compositional perturbation}}$$

Bias and CoDA

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

$$\mathbf{O} \sim \mathbf{A} \cdot \underbrace{\mathbf{B}^{(P)}}_{\text{Compositional perturbation}}$$

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

Bias and CoDA

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

$$\mathbf{O} \sim \mathbf{A} \cdot \underbrace{\mathbf{B}^{(P)}}_{\text{Compositional perturbation}}$$

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

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

Bias and CoDA

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

$$\mathbf{O} \sim \mathbf{A} \cdot \underbrace{\mathbf{B}^{(P)}}_{\text{Compositional perturbation}}$$

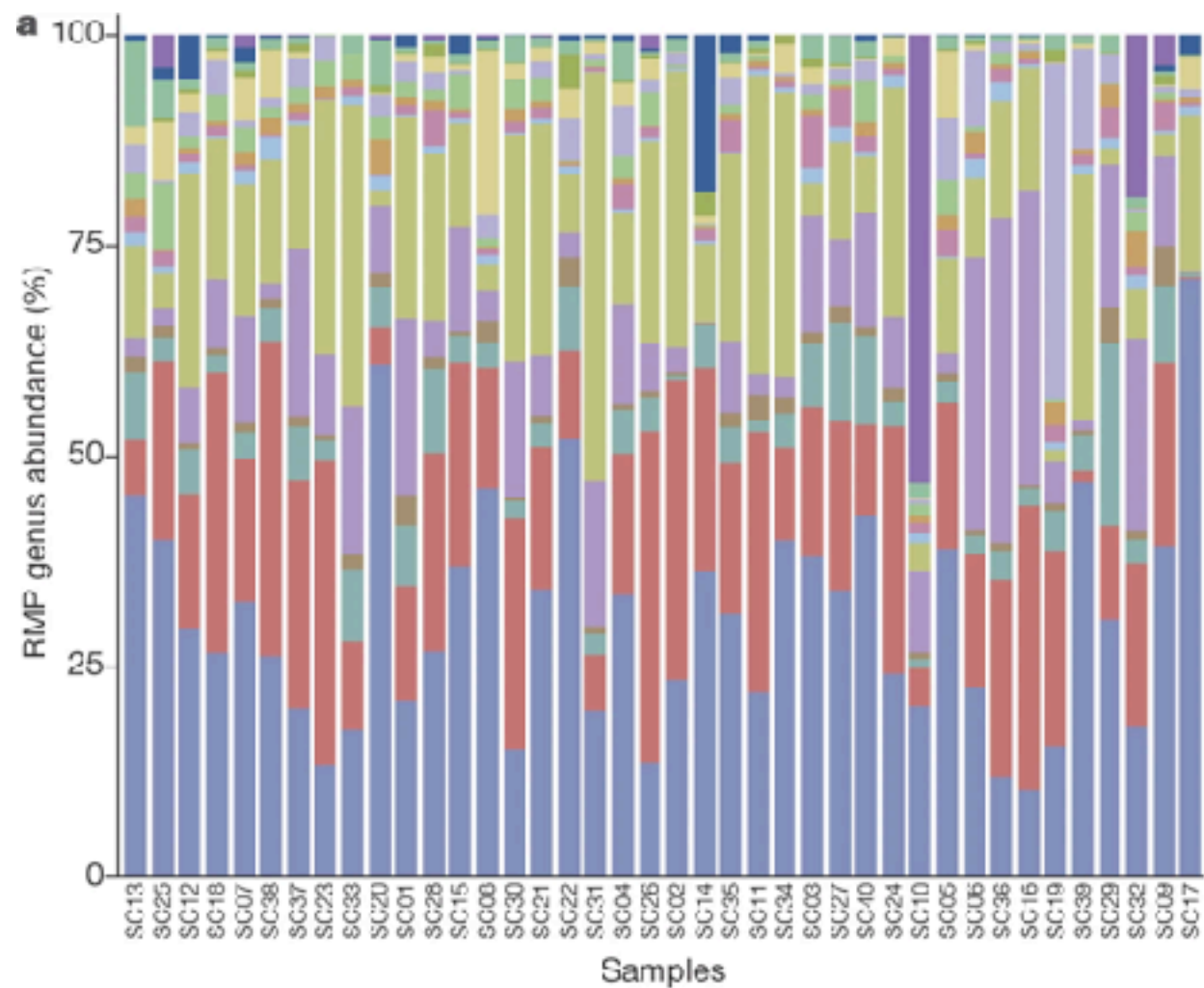
$$\mathbf{O}(s) / \mathbf{O}(t) \sim (\mathbf{A}(s) \cdot \mathbf{B}) / (\mathbf{A}(t) \cdot \mathbf{B}) \sim \mathbf{A}(s) / \mathbf{A}(t)$$

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

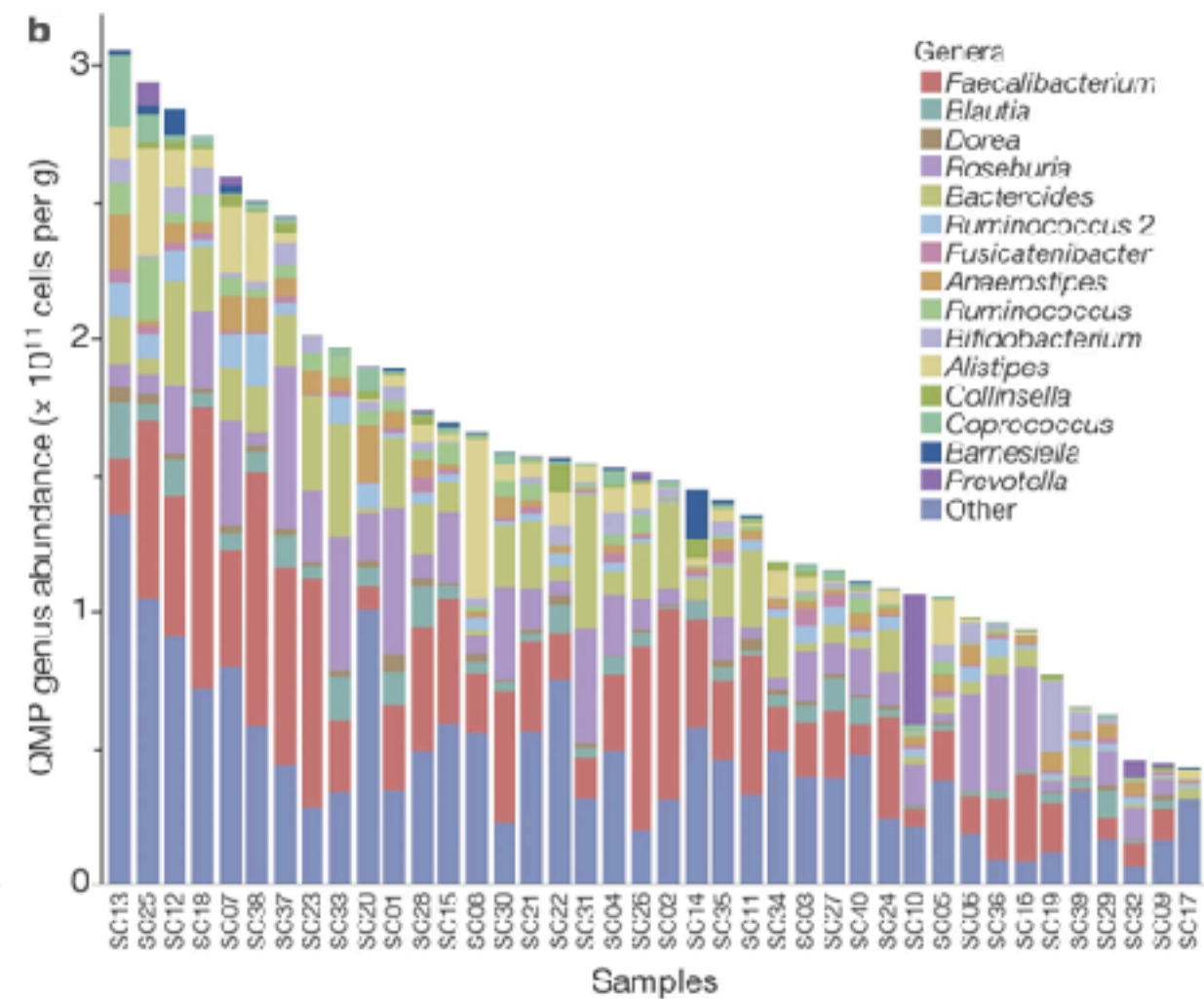
$$\mathbf{O}^{(P)} / \mathbf{O}^{(R)} \sim \mathbf{A} \cdot \mathbf{B}^{(P)} / (\mathbf{A} \cdot \mathbf{B}^{(R)}) = \mathbf{B}^{(P)} / \mathbf{B}^{(R)}$$

Absolute Abundance

Relative Abundances

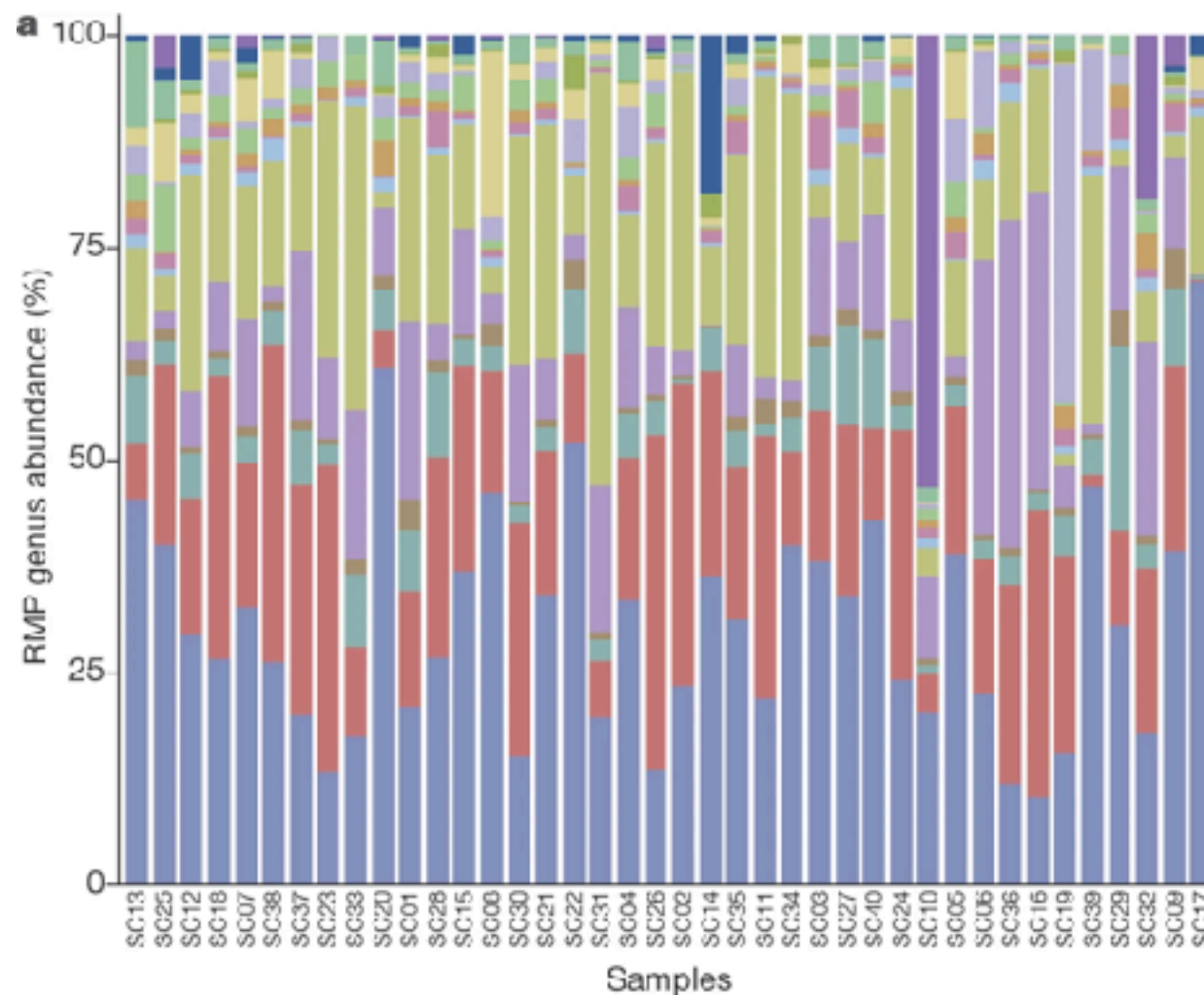


Absolute Abundances

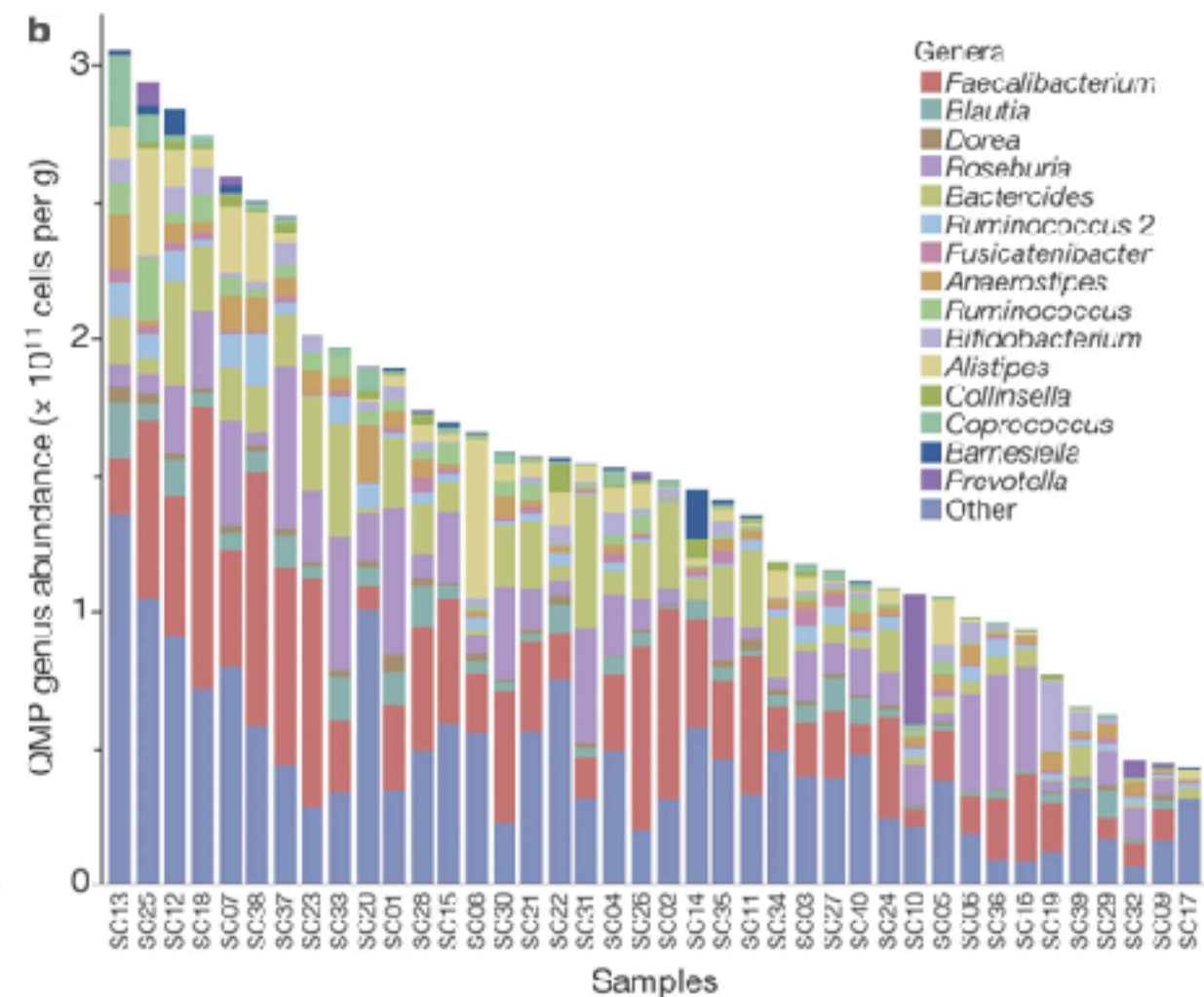


Absolute Abundance

Relative Abundances



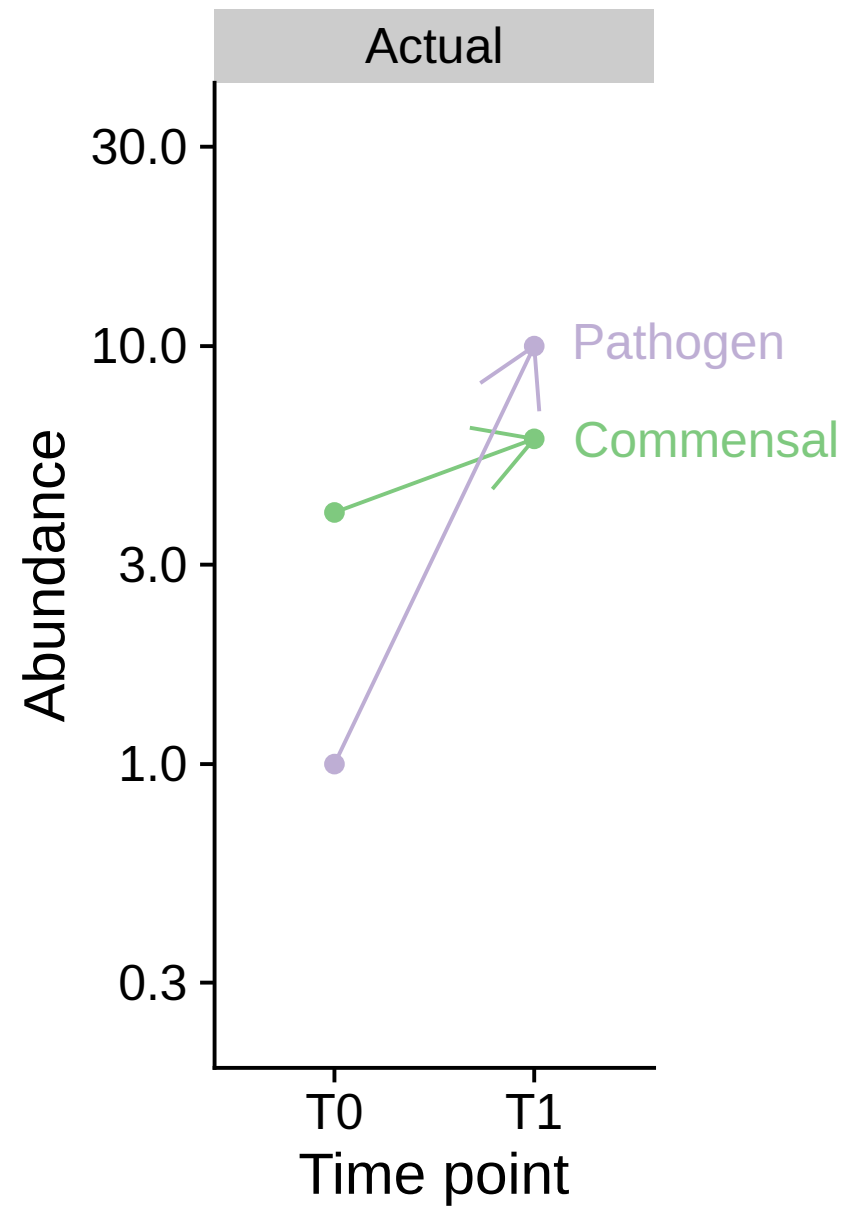
Absolute Abundances



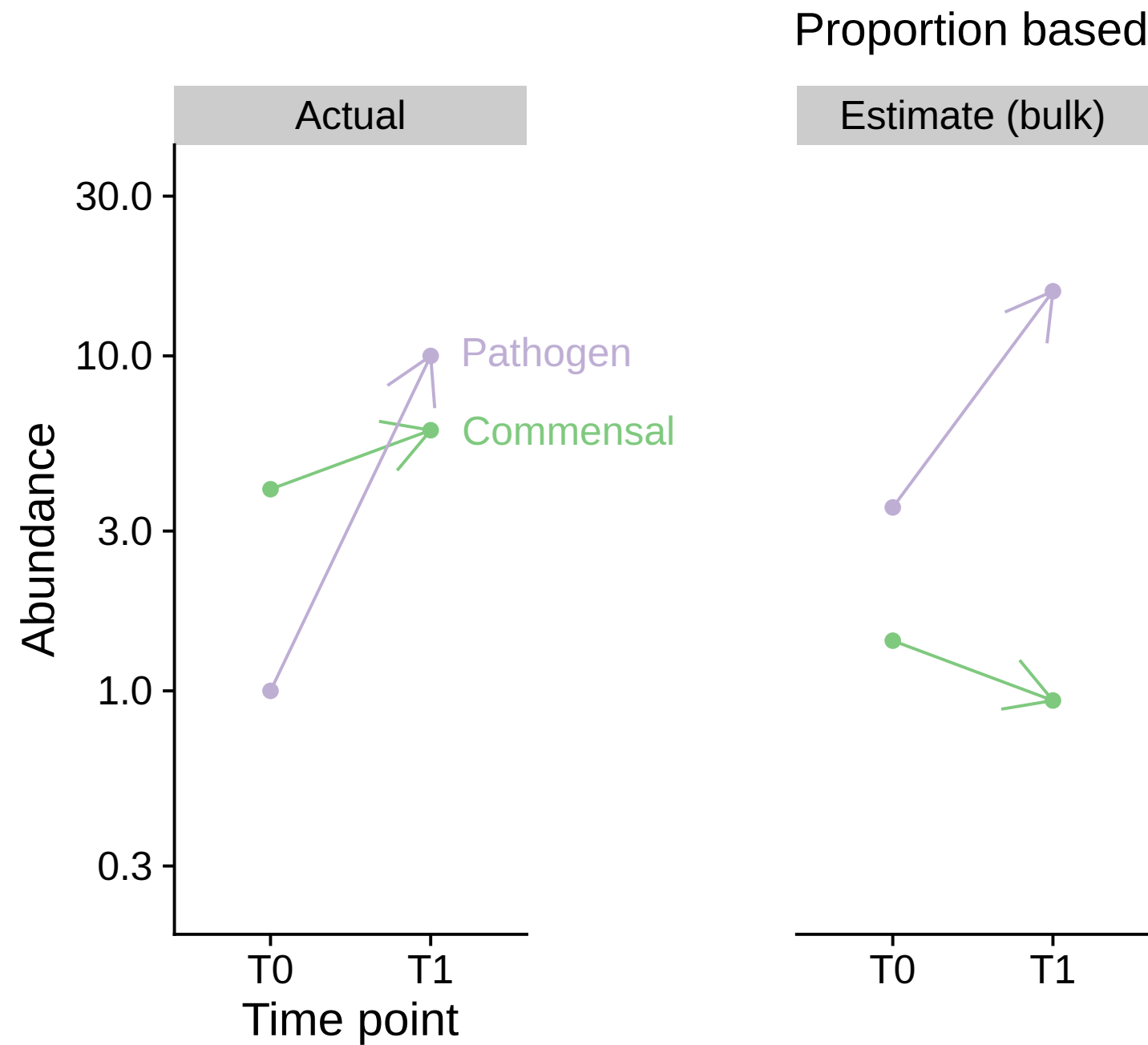
Two types of Absolute Abundance estimation methods

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

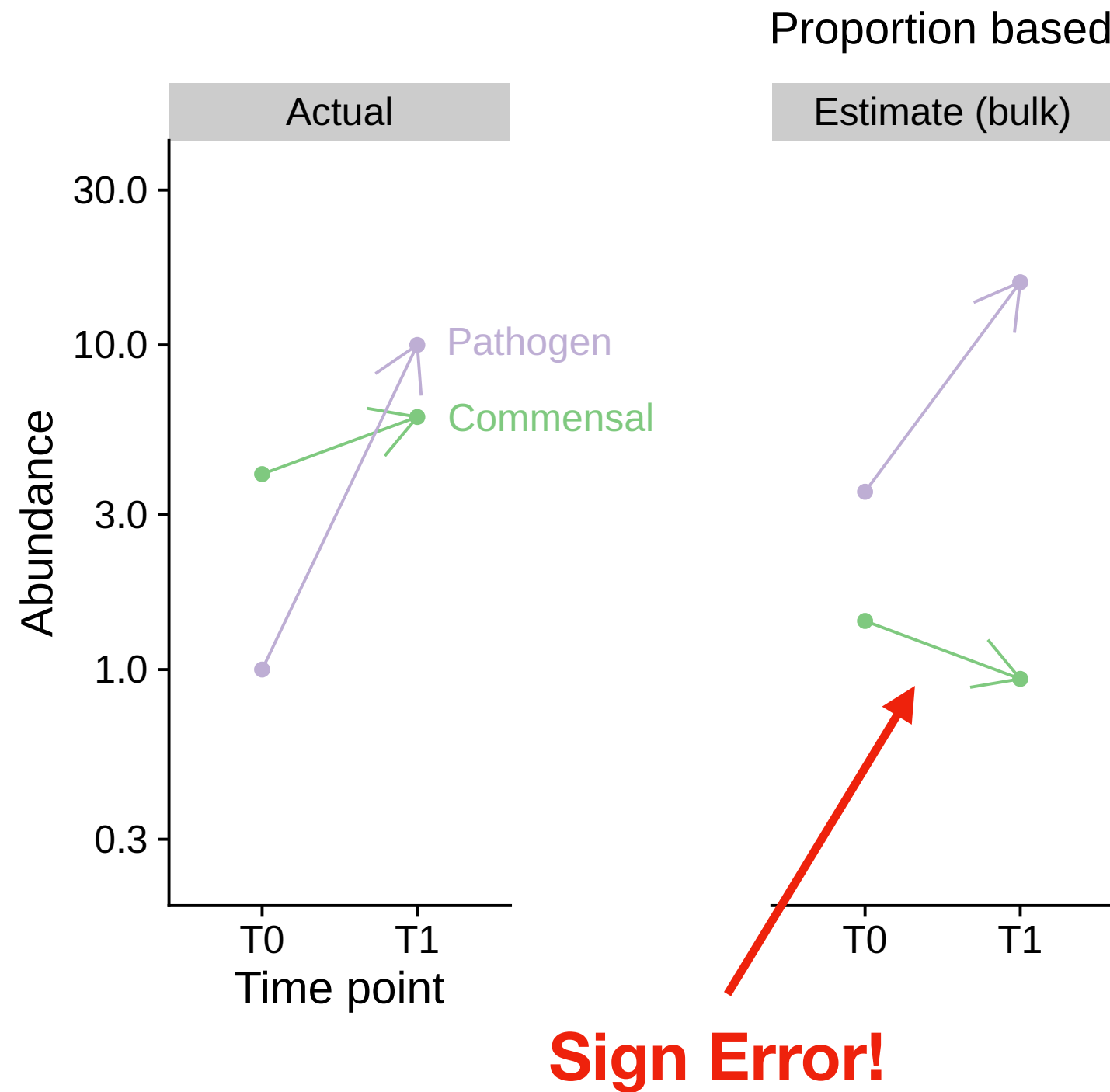
Absolute Abundance



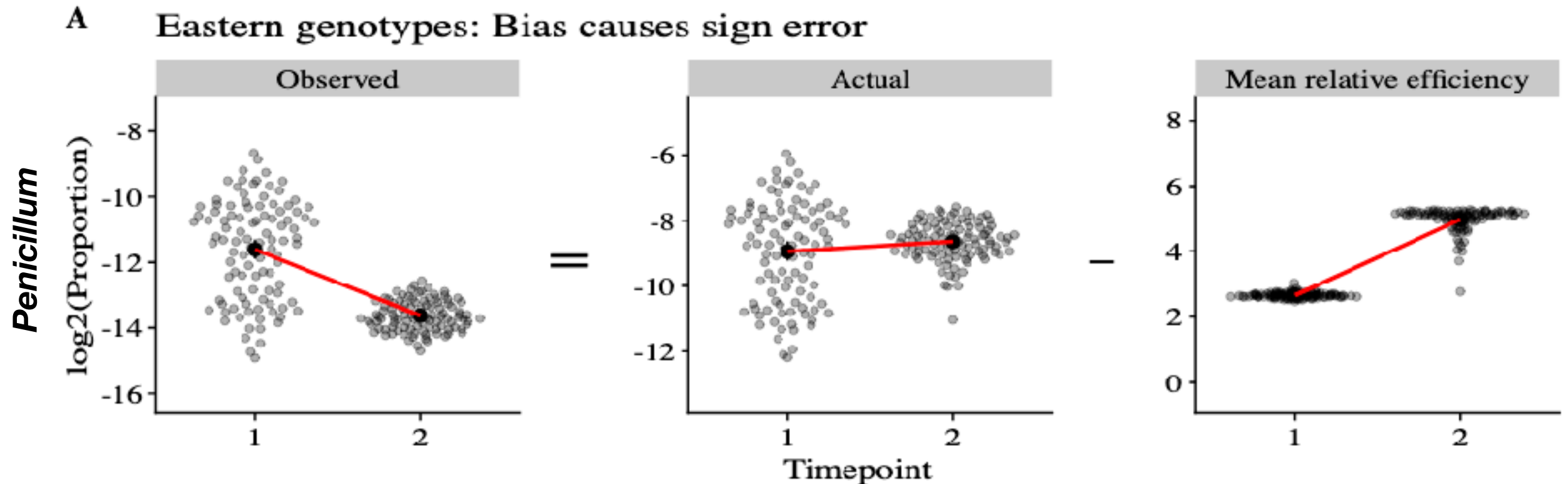
Absolute Abundance



Absolute Abundance

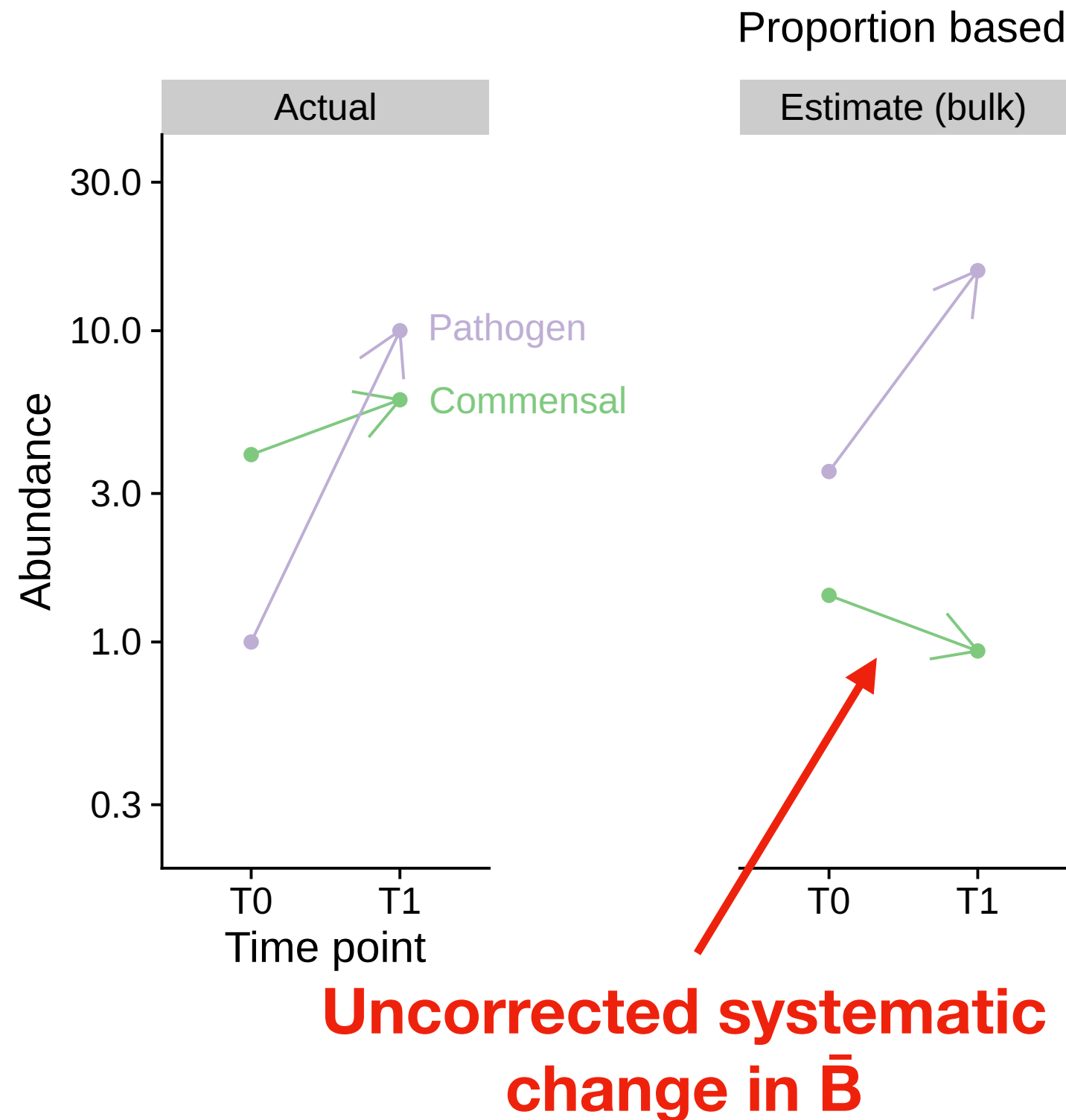


Absolute Abundance

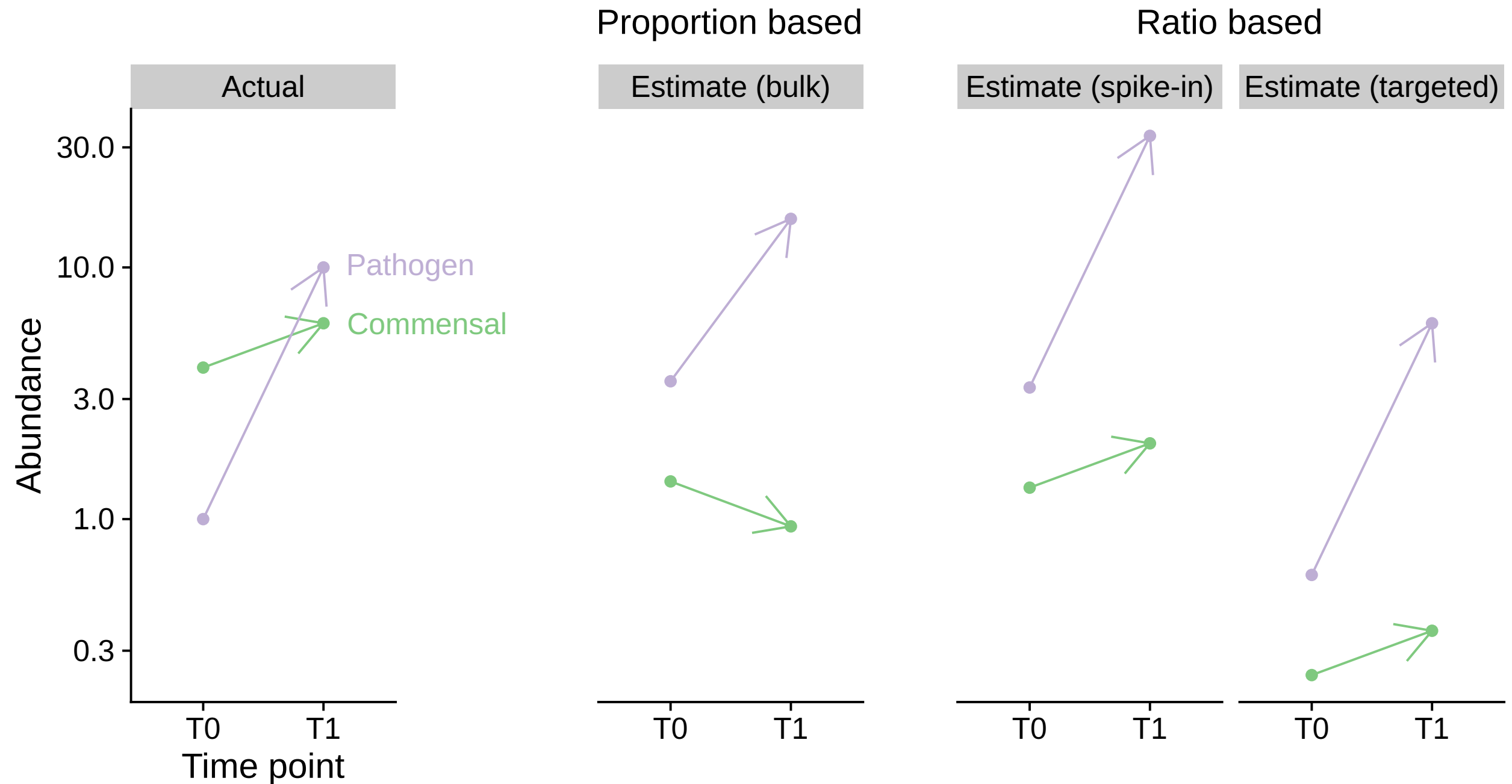


Experimental manipulation of plant microbiome assembly in cottonwood leaves
High-efficiency pathogen *Melampsora* × *columbiana* invades in T2.
Synthetic (defined) community.

Absolute Abundance



Absolute Abundance



Absolute Abundance

