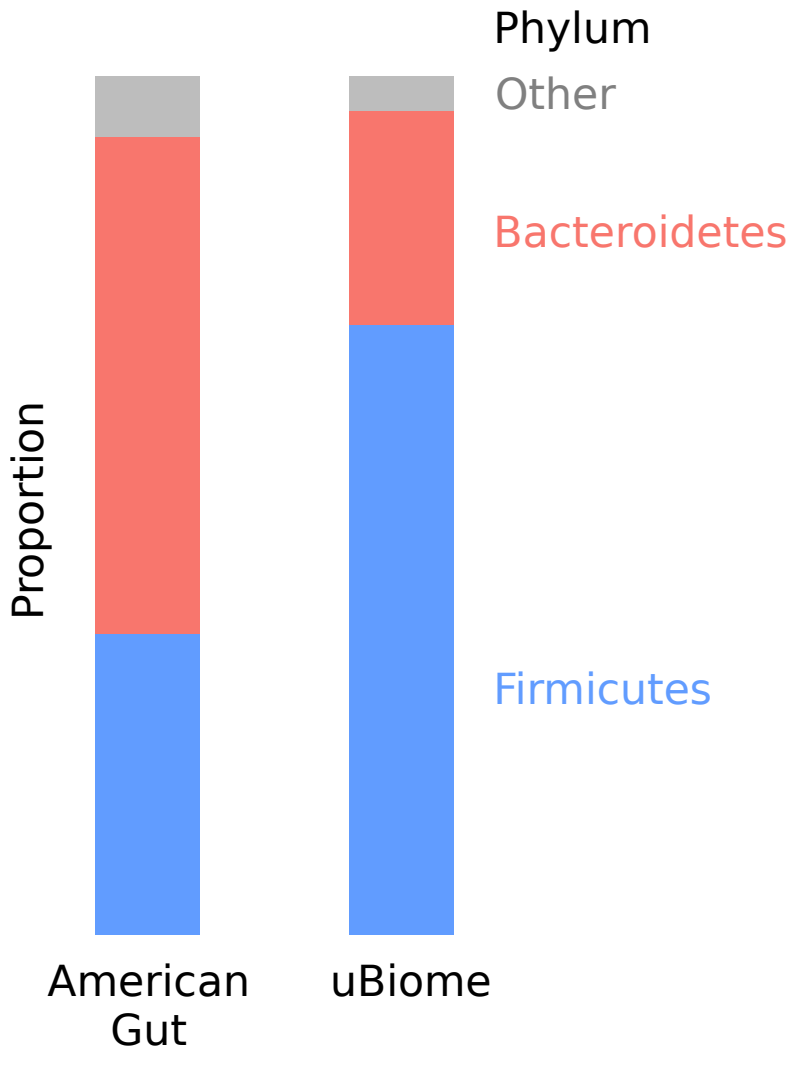
Consistent and comprehensible bias in metagenomic sequencing experiments

Michael McLaren

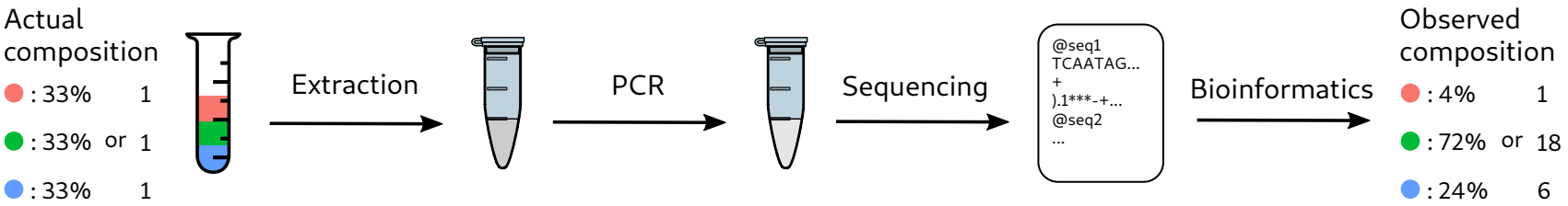
North Carolina State University

Different methods measure different taxonomic profiles

The same stool profiled by two testing services



Each step in the metagenomics workflow is *biased* towards detecting certain taxa over others



Measured abundances don't reflect actual abundances and are quantitatively incomparable between experiments

Key questions needing answers

How does bias affect a given microbiome analysis or diagnostics?
e.g., is an analysis ok if all samples were measured by the same experiment?

What is the best way to quantify bias?
are current metrics good indicators of a protocol's accuracy?

Can control samples be used to estimate bias and calibrate our measurements?
to make them accurate or at least quantitatively comparable between labs?

We seek statistical and experimental methods to accurately measure and correct bias

Co-authors:

Ben Callahan
(NC State)

Amy Willis
(U. of Washington)

Develop a model of bias
that reflects the characteristics
of real metagenomics experiments

Validate the model
with amplicon and shotgun experiments

Start using the model
to measure, account for, and correct bias

Our results suggest that. . .

Bias follows a simple pattern
observed relative abundances equal
actual abundances * taxon-specific factors

These factors are consistent
within an experiment
making bias possible to account for,
measure, and correct

Our model stems from two key features of metagenomics measurements

Many bias mechanisms are multiplicative

output = input * taxon-specific factors

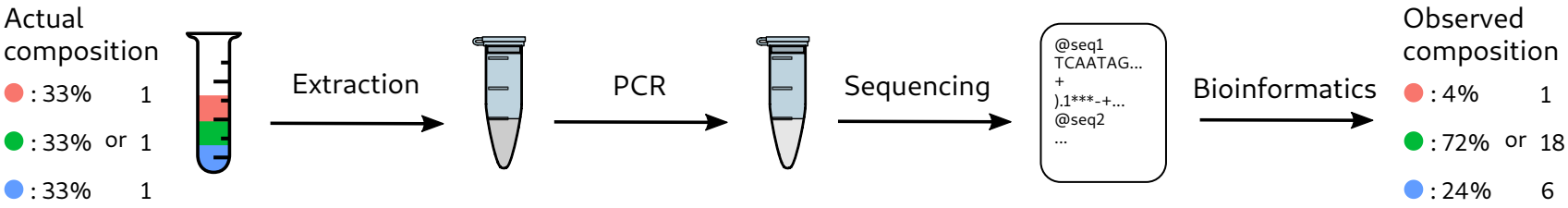
e.g., yield of DNA per cell of Species A

Sequencing only measures relative abundances

error only depends on the relative factors

e.g., yield from A = 2 * yield from B

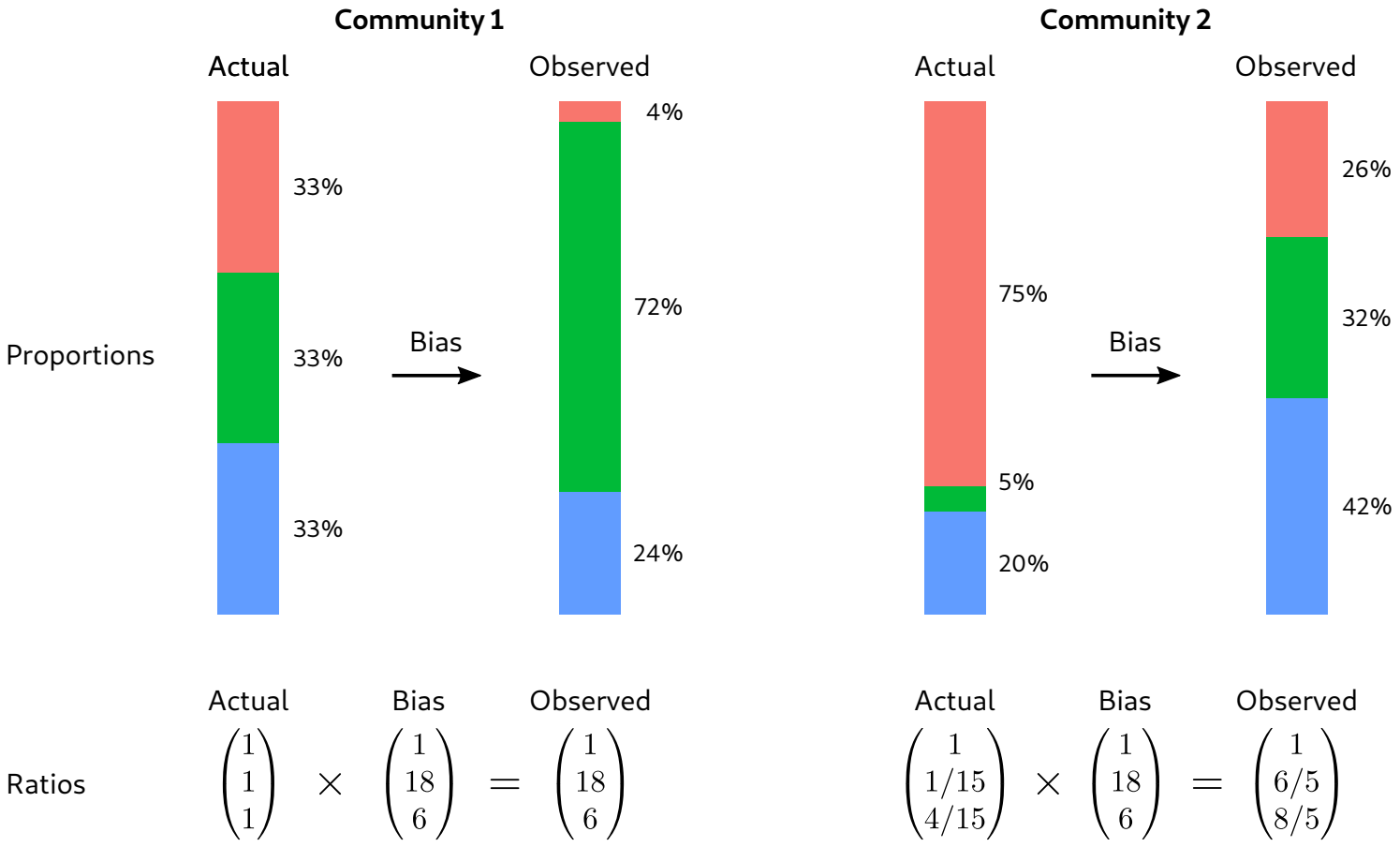
Relative bias factors multiply across steps to determine total error



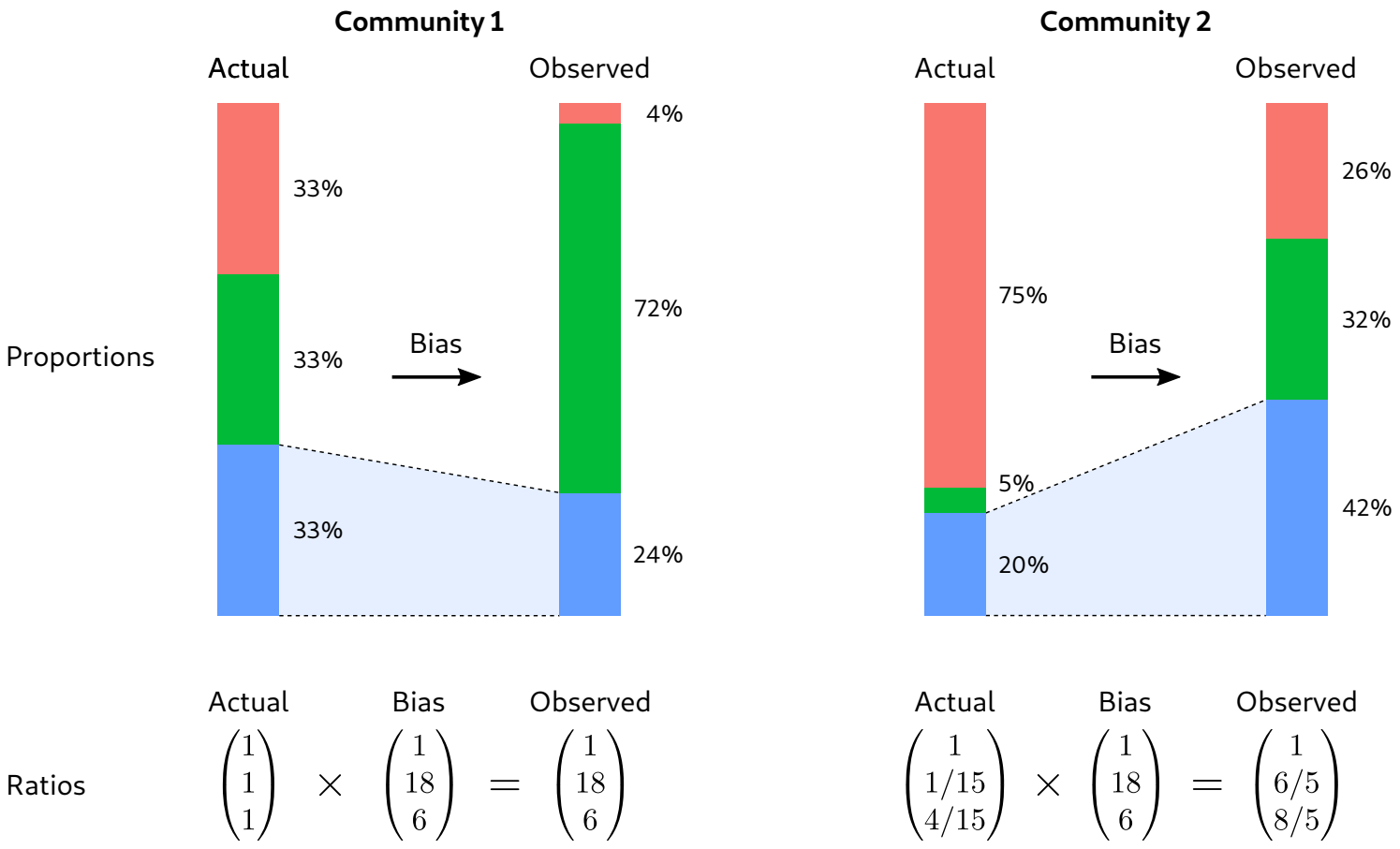
Actual composition		Extraction bias	PCR bias	Sequencing bias	Bioinformatics bias	Total bias	Observed composition	
● : 33%	1	x 1	x 1	x 1	x 1	x 1 (= 1 x 1 x 1 x 1)	● : 4%	1
● : 33%	or 1	x 4	x 6	x 0.5	x 1.5	x 18 (= 4 x 6 x 0.5 x 1.5)	● : 72%	or 18
● : 33%	1	x 15	x 2	x 0.25	x 0.8	x 6 (= 15 x 2 x 0.25 x 0.8)	● : 24%	6

$$24\% = \frac{6}{1 + 18 + 6}$$

One bias vector determines the error in arbitrarily composed communities



Focusing on proportions gives the incorrect impression that bias depends on sample composition



Validating the model with mock communities

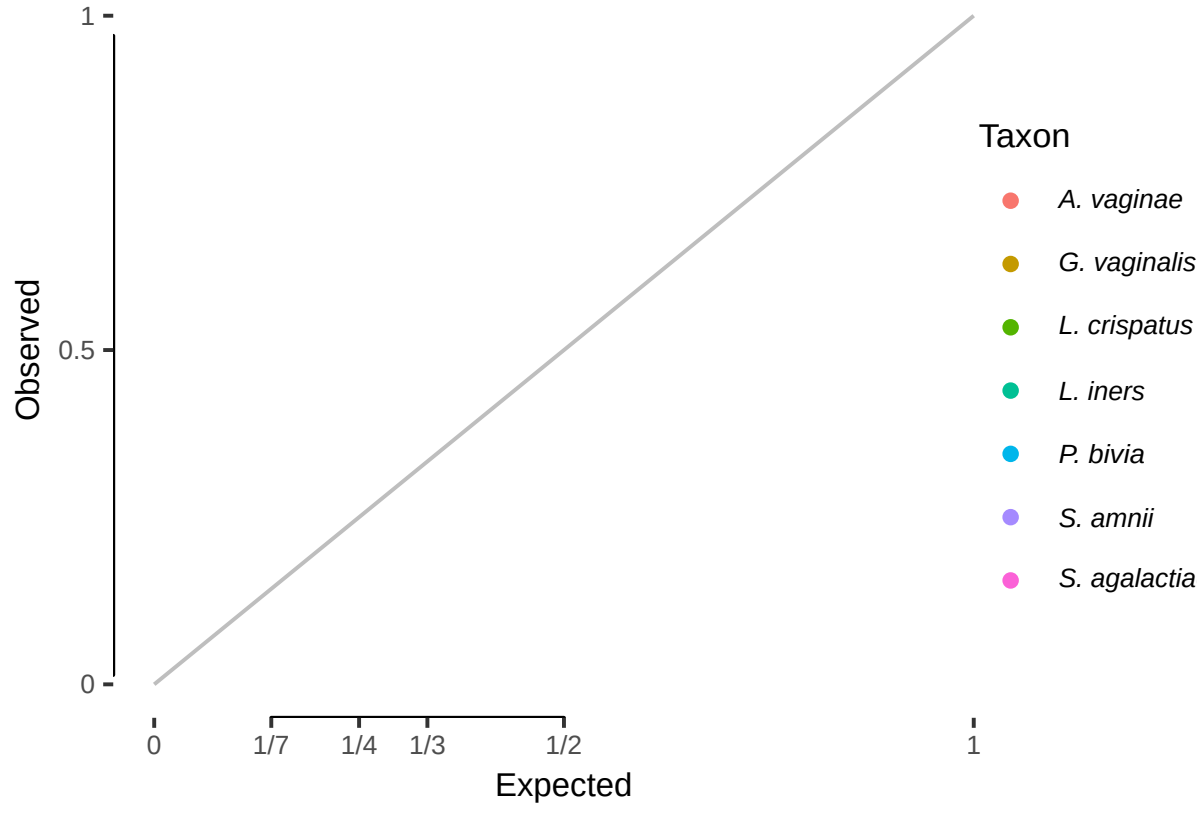
16S amplicon data from Brooks et al (2015)
The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies

Seven bacterial species in 58 unique mixtures
can test that bias is composition-independent

Mixtures of cells, DNA, and PCR product
allows testing the model for extraction, PCR, and the total workflow

The observed proportions deviate strongly from the even-mixture expectation

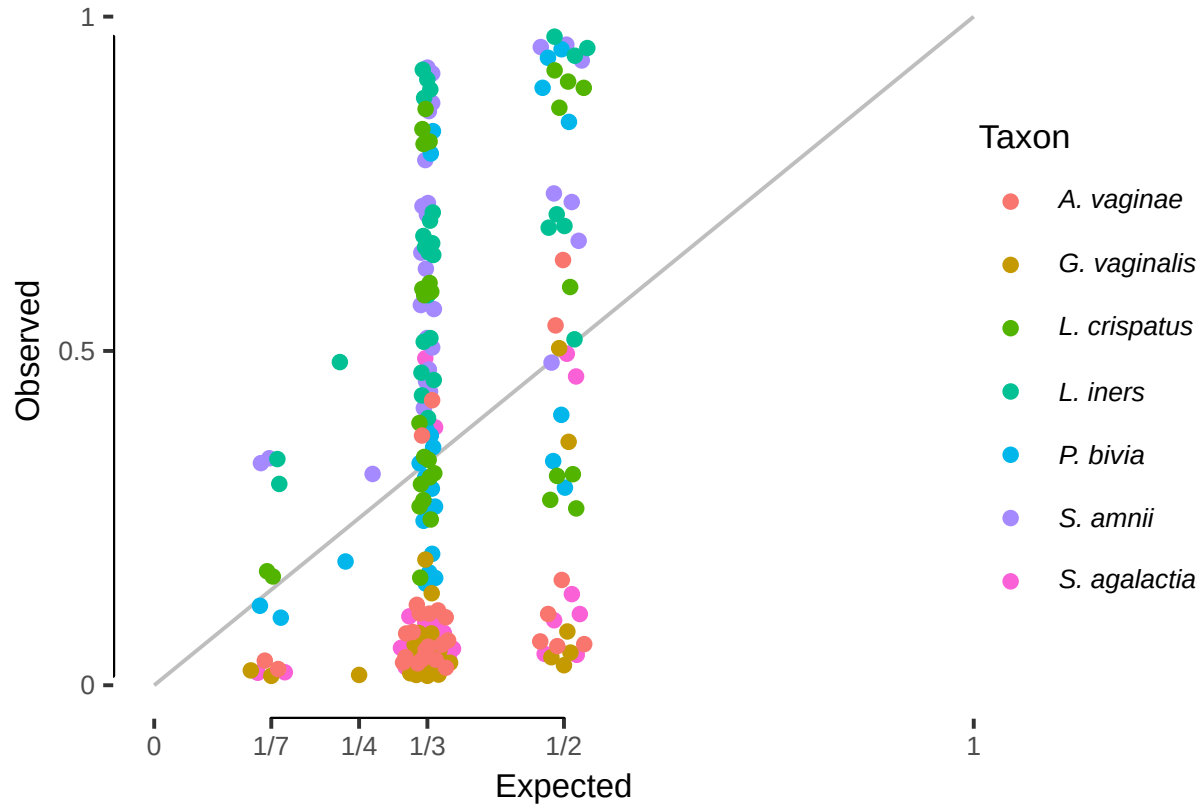
Observed proportion vs. expected



$$\text{Expected} = \frac{1}{\#\ \text{species}}$$

The observed proportions deviate strongly from the even-mixture expectation

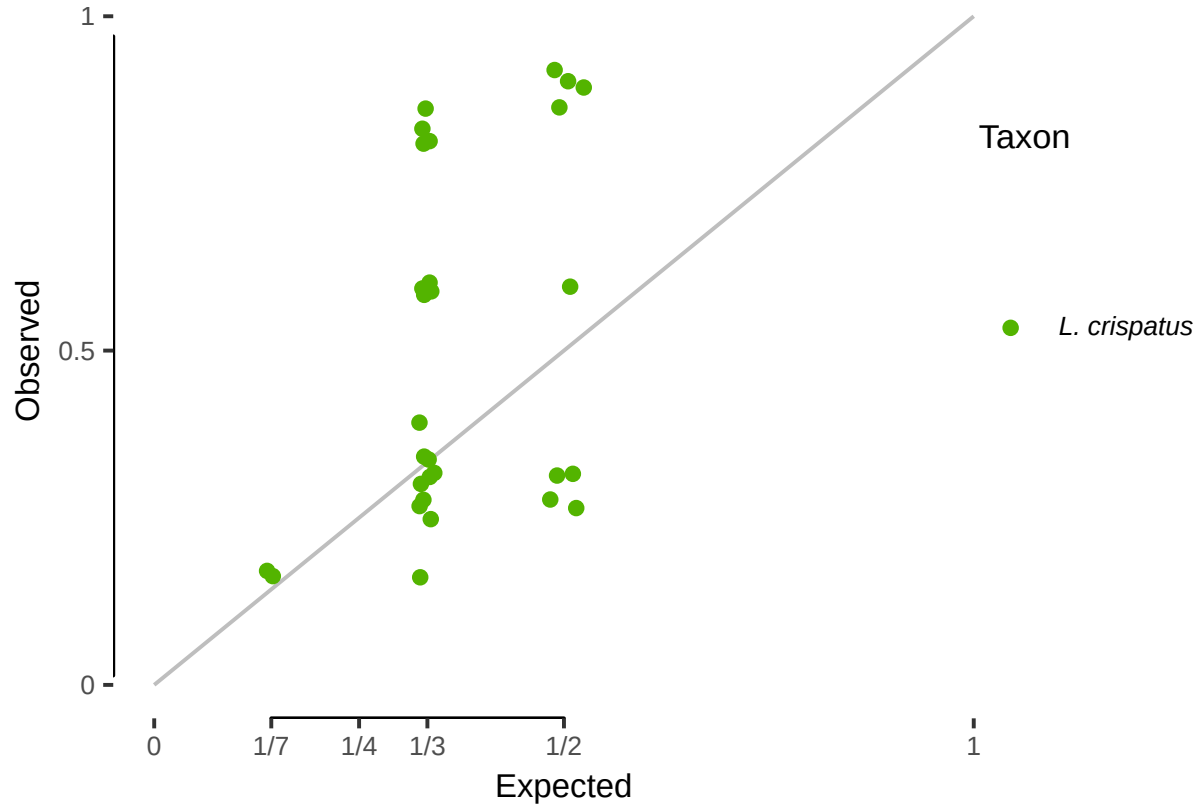
Observed proportion vs. expected



$$\text{Expected} = \frac{1}{\#\ \text{species}}$$

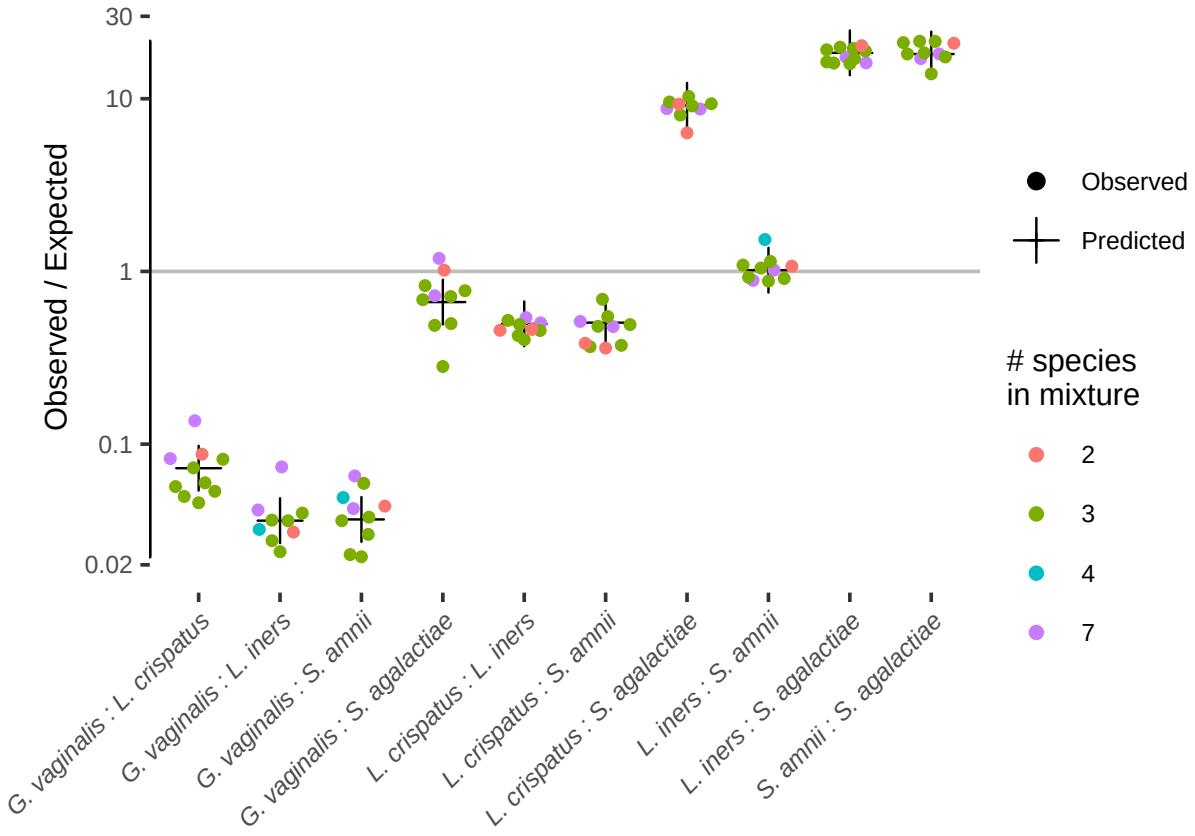
The error in proportions varies among samples

Observed proportion
vs. expected
for *L. crispatus*



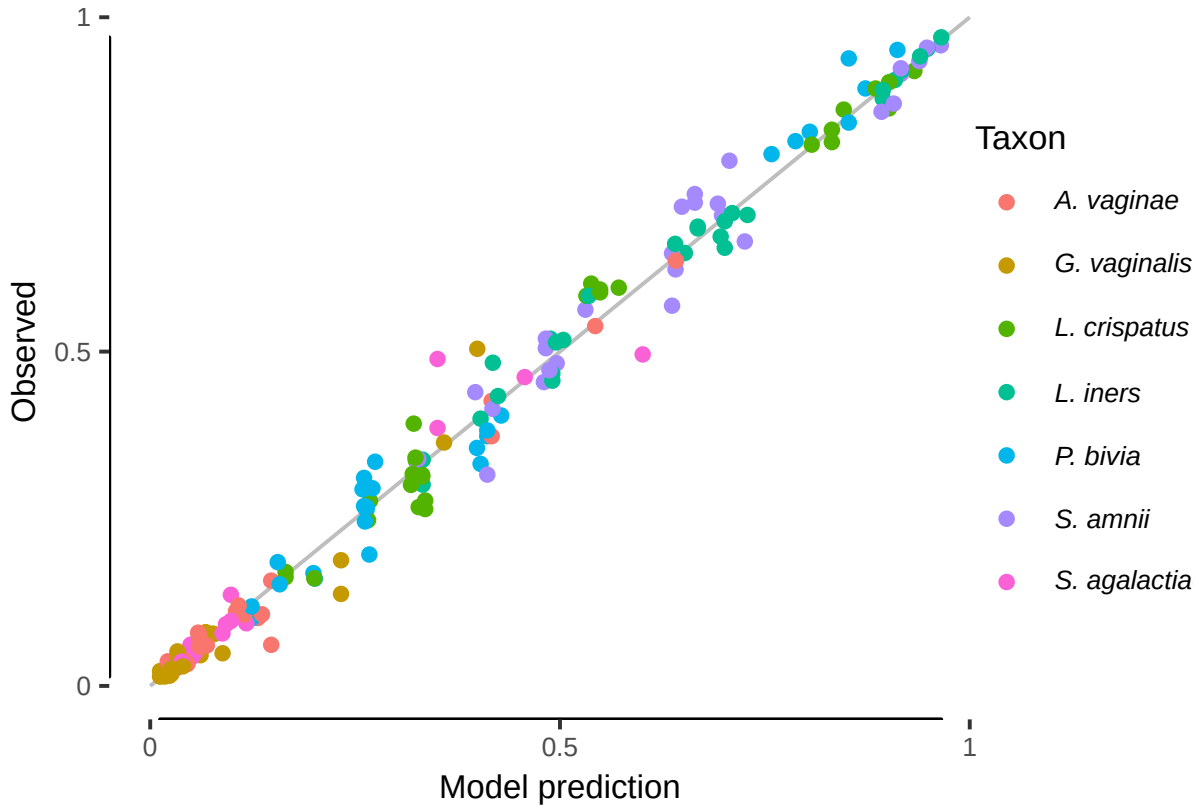
But the error in taxa ratios is consistent across samples

Fold error in taxa ratios



The estimated bias explains the observed error

Observed proportion
vs. model prediction



Having an accurate model of bias is highly useful

It lets us

Improve protocol optimization
by having a composition-independent, biologically-meaningful measure of bias

Evaluate the accuracy of downstream statistical analyses and diagnostics
by analyzing the effect of bias mathematically or with empirically-motivated simulations

Develop calibration methods
to correct bias or make measurements comparable between labs

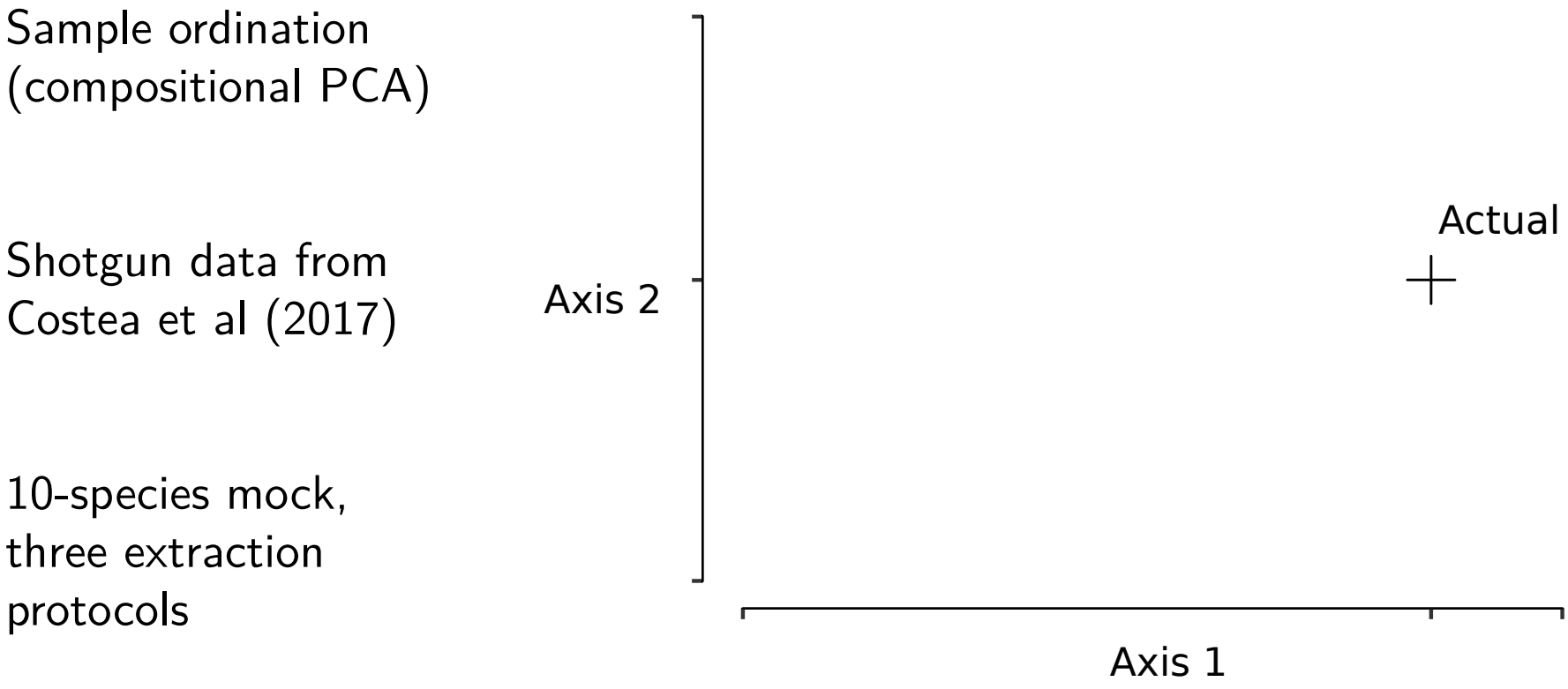
Using control samples to calibrate measurements

1. Measure controls along with target samples
mock communities or natural samples
measured by a reference protocol

2. Estimate bias from controls
Bias = Observed / Actual

3. Use estimate to calibrate targets
Calibrated = Observed / Bias

Uncalibrated measurements systematically differ between protocols and from the actual abundances

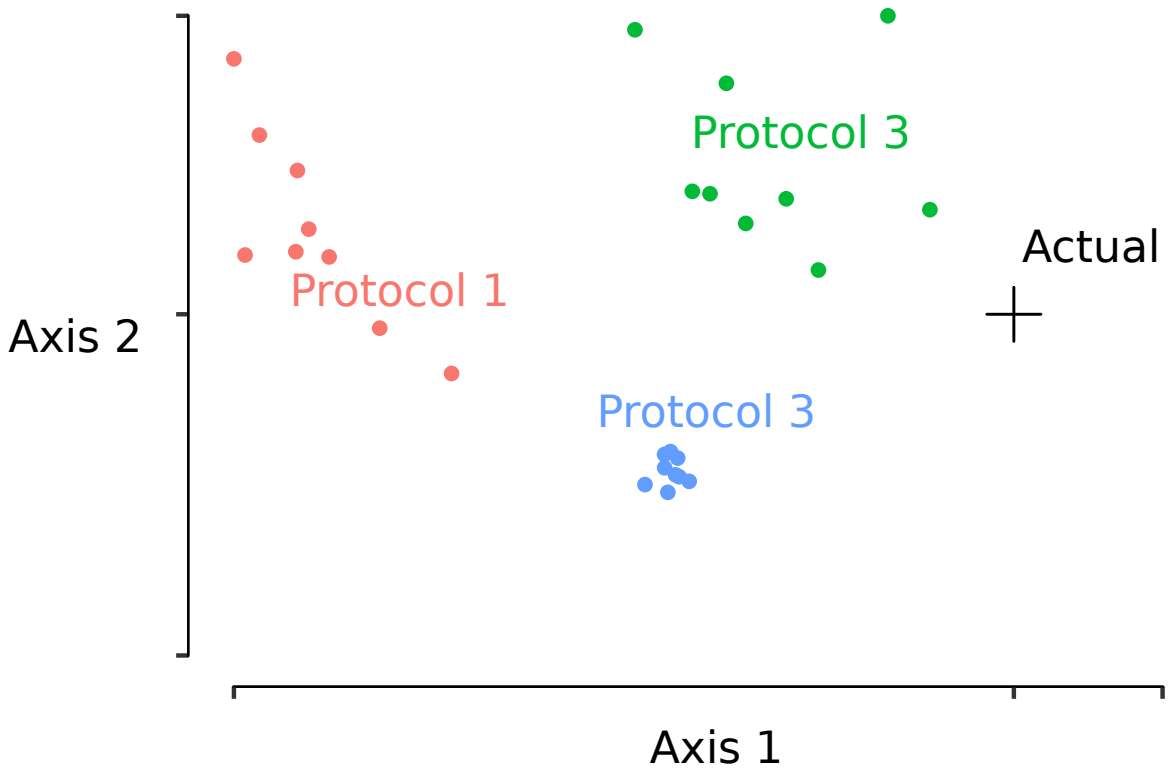


Uncalibrated measurements systematically differ between protocols and from the actual abundances

Sample ordination (compositional PCA)

Shotgun data from Costea et al (2017)

10-species mock, three extraction protocols

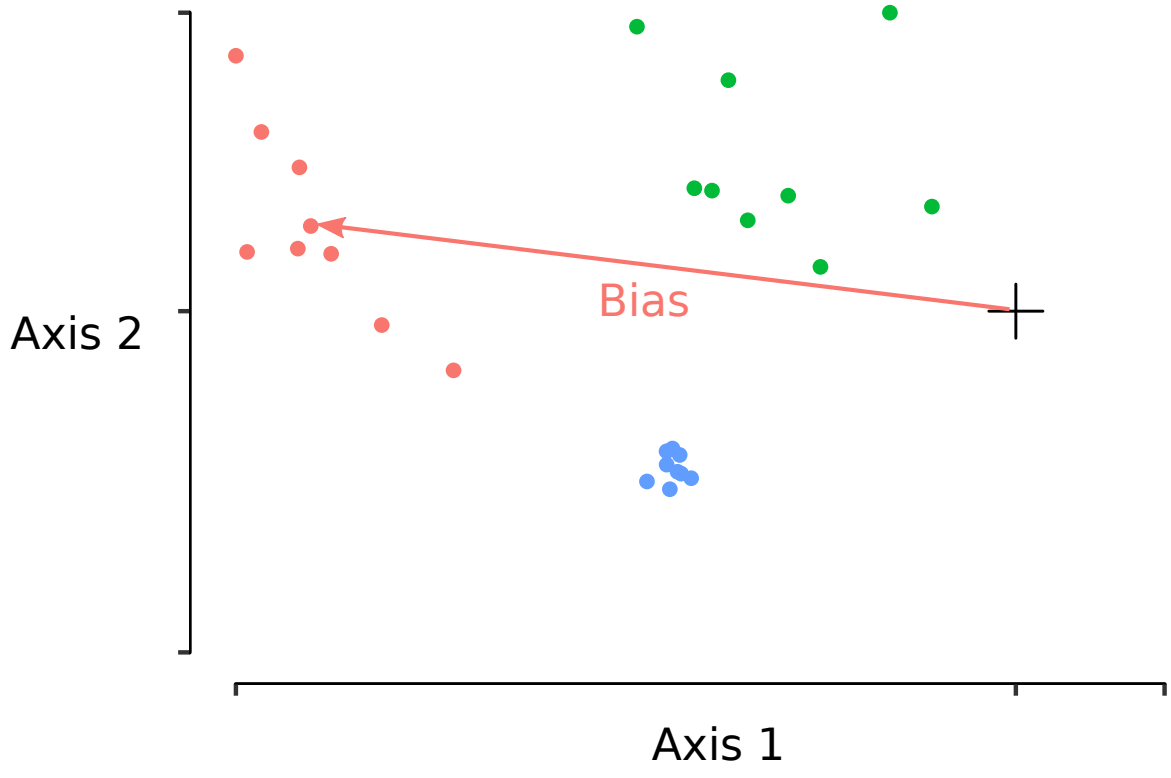


Uncalibrated measurements systematically differ between protocols and from the actual abundances

Sample ordination (compositional PCA)

Shotgun data from Costea et al (2017)

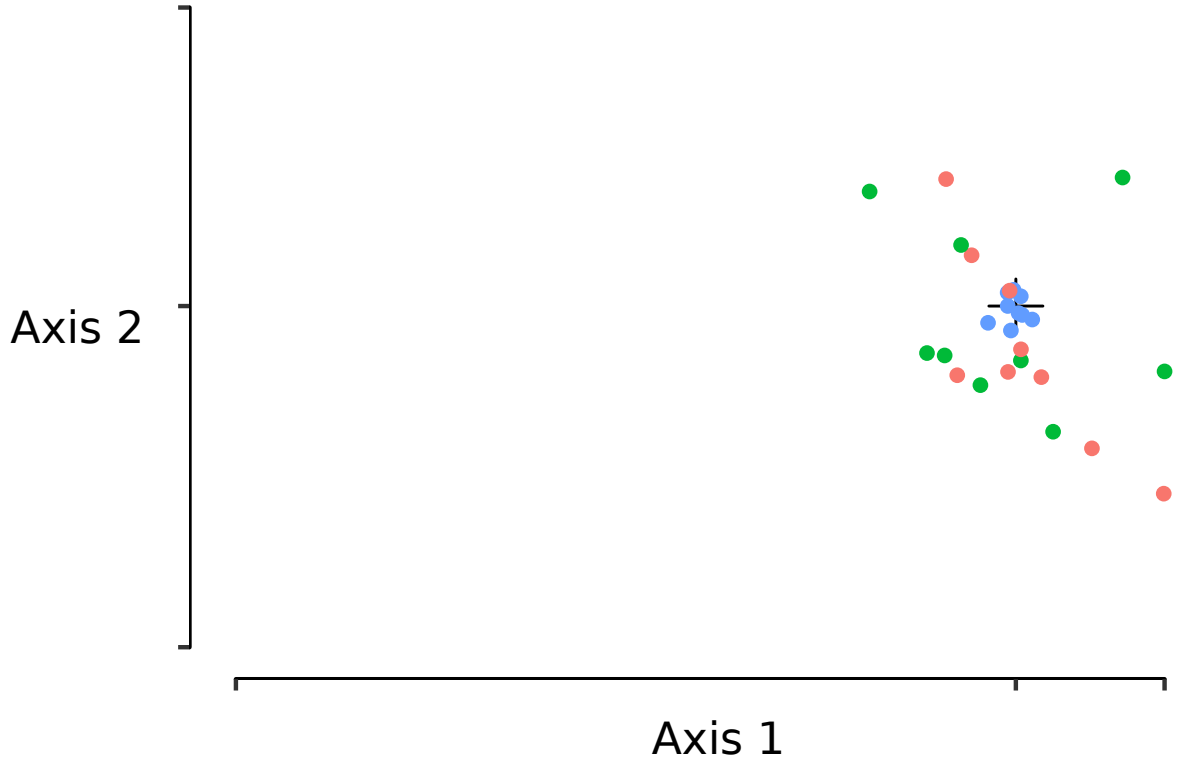
10-species mock, three extraction protocols



Calibration to known actual abundances can remove systematic error

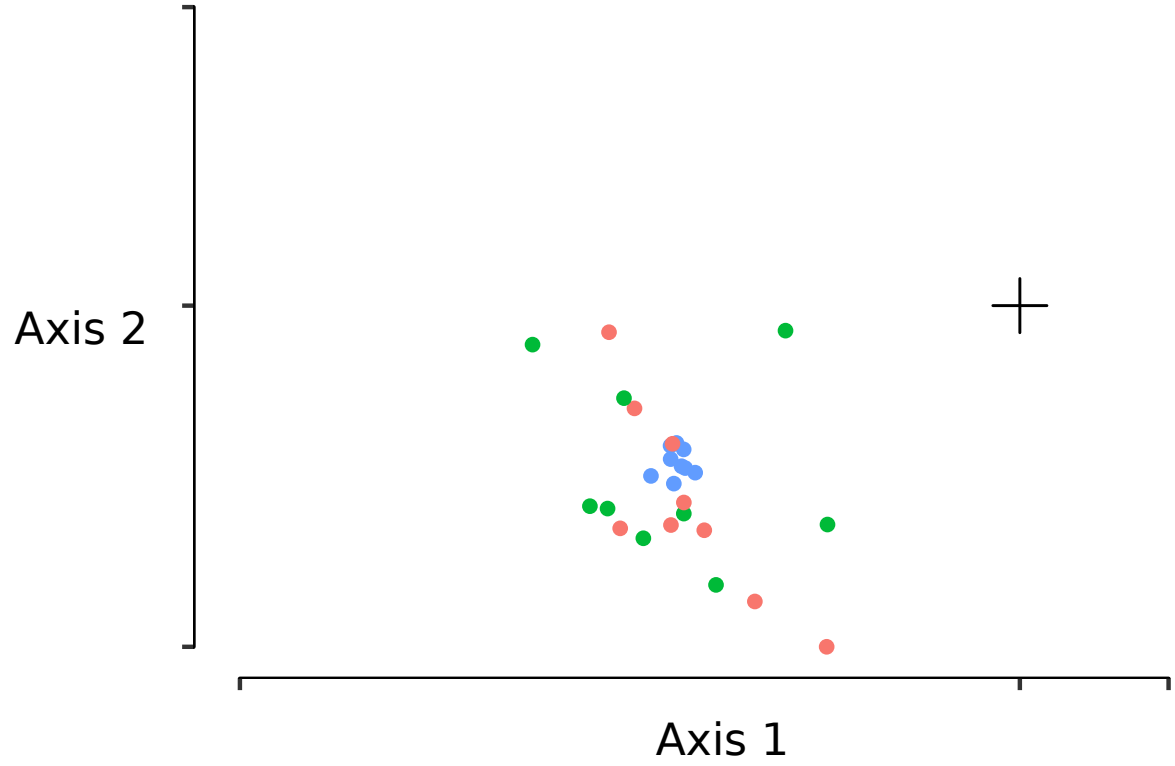
Calibrated to actual abundances

Bias estimated from 3 of 10 samples

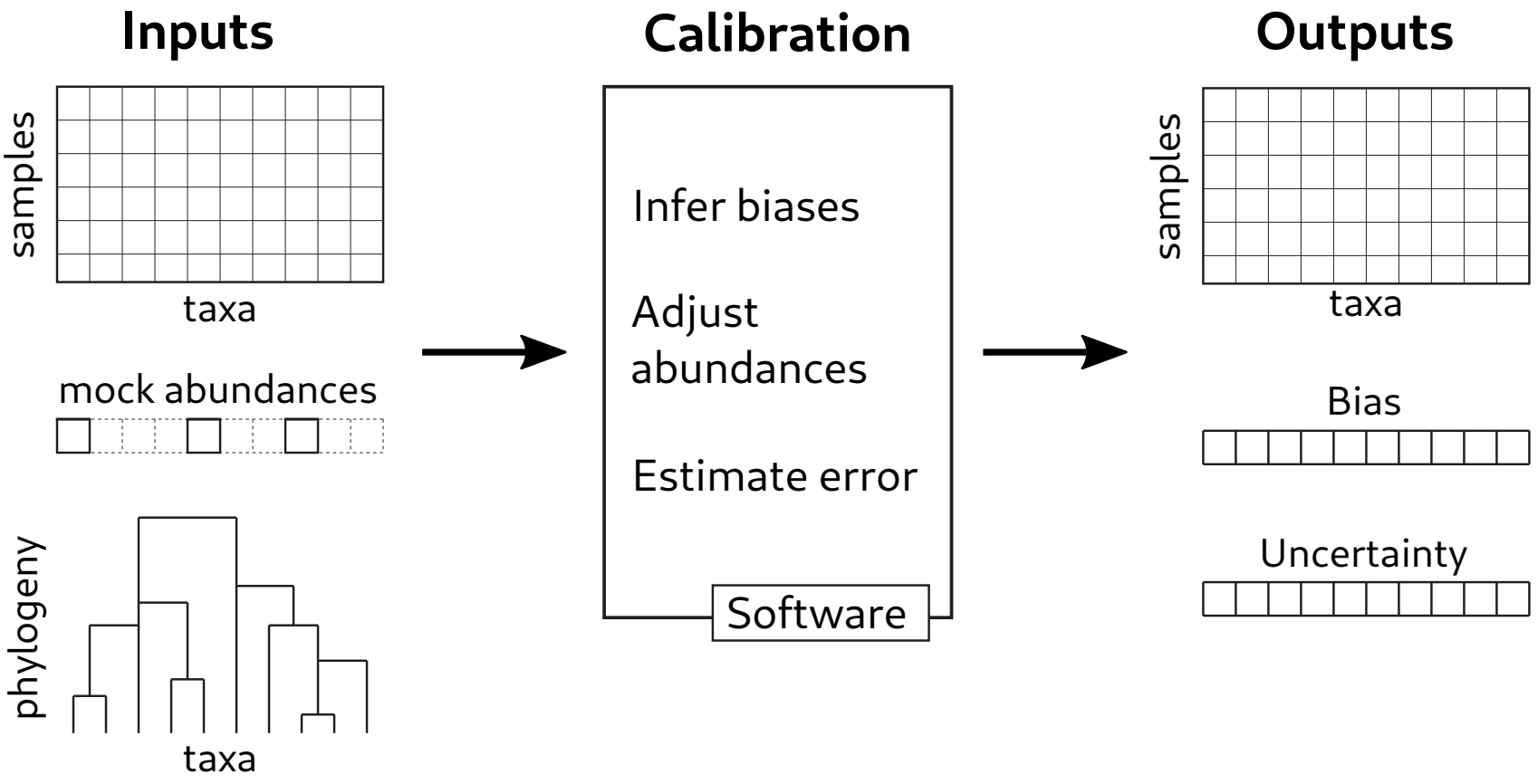


Without knowing actual abundances, we can still calibrate to a reference protocol

Calibrated to chosen reference (blue)



Full calibration may be possible by combining bias measurement with phylogenetic inference



In summary, our results suggest that. . .

Bias is consistent
within an experiment when properly measured

Bias is comprehensible
observed = actual * taxon-specific factors

Bias is tractable
it has predictable effects on analyses
and can be measured and corrected

A good model is just the beginning

Further work needed to

Test more protocols
fit may require protocol optimization

Test the model in natural communities
using differential bias between protocols

Develop and validate better control samples
taxonomically representative;
bias same in control and targets

Develop robust statistical methods
for estimation and calibration

A hopeful alternative to hopeless and wishful thinking

The hopeless view

metagenomic measurements at best can detect who's there

The wishful view

bias can be ignored within experiments

The hopeful view

experiments and analyses can be designed to account for bias

Thanks to

Co-authors
Amy Willis
Ben Callahan

Feedback and discussions
David Clausen
Callahan lab
NCSU CVM
4th floor

The M3 organizers!

How might these ideas be put to use in your metagenomics applications?

Let's discuss! Find me later today or email at m.mclaren42@gmail.com