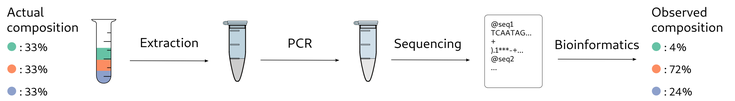
Consistent and correctable bias in metagenomic sequencing measurements

Michael R. McLaren[1], Amy D. Willis[2], and Benjamin J. Callahan[1]

[1]NC State University [2]University of Washington

A mathematical model of bias

Each step in a marker-gene or metagenomics workflow is **biased** towards detecting certain taxa over others



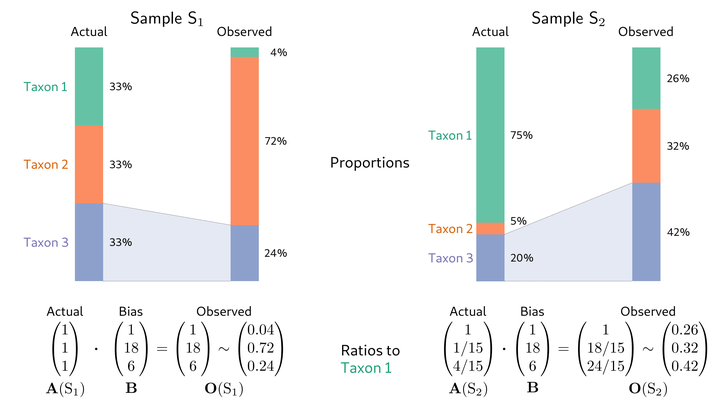
Bias makes measured abundances inaccurate and quantitatively incomparable between experiments. The input-output (actual-observed) relationship is currently unknown

Model of bias: Detection efficiencies multiply across steps

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)	Normalize to 100%	: 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
$\mathbf{A}$		$\mathbf{B}^{(P_1)}$	$\mathbf{B}^{(P_2)}$	$\mathbf{B}^{(P_3)}$	$\mathbf{B}^{(P_4)}$	$\mathbf{B}^{(P)} \sim \mathbf{B}^{(P_1)} \cdot \ldots \cdot \mathbf{B}^{(P_4)}$		$\mathbf{O} \sim \mathbf{A} \cdot \mathbf{B}^{(P)}$	

The set of **relative detection efficiencies** for all taxa for a given protocol determines the bias of that protocol for arbitrarily composed samples

Bias acts consistently on taxon ratios, not taxon proportions



Consistent bias can be estimated and corrected (**calibration**)

Estimation from a single control sample

$$\hat{\mathbf{B}} \sim \mathbf{O}(c)/\mathbf{A}(c)$$

Estimation from multiple controls containing different, partially-overlapping sets of taxa

$$\hat{\mathbf{B}} = \arg\min_B \textstyle\sum_c \|\mathbf{O}(c)/(\mathbf{A}(c) \cdot \mathbf{B})\|^2$$

Calibration to actual composition

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

Calibration to a reference protocol R

$$\hat{\mathbf{B}}^{(P/R)} \sim \mathbf{O}^{(P)}(c)/\mathbf{O}^{(R)}(c)$$
$$\hat{\mathbf{O}}^{(P)}(s) \sim \mathbf{O}^{(R)}(s)/\hat{\mathbf{B}}^{(P/R)}$$

Evaluation datasets

16S amplicon data from Brooks et al (2015)

7 bacterial species in 58 unique mixtures of cells, DNA, and PCR product

test that bias is independent of which species are in the sample

Mixtures of cells, of DNA, and of PCR product

test the model for extraction, PCR, and the total workflow

Shotgun data from Costea et al (2017)

10 bacterial species in fixed amounts measured alone and spiked into feces

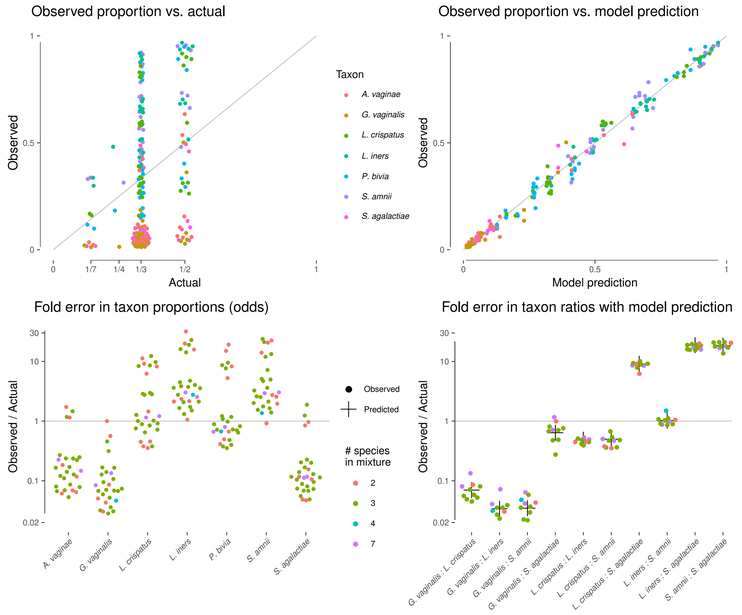
test that bias of spike-in is independent of the background composition

3 DNA extraction protocols (H, Q, and W)

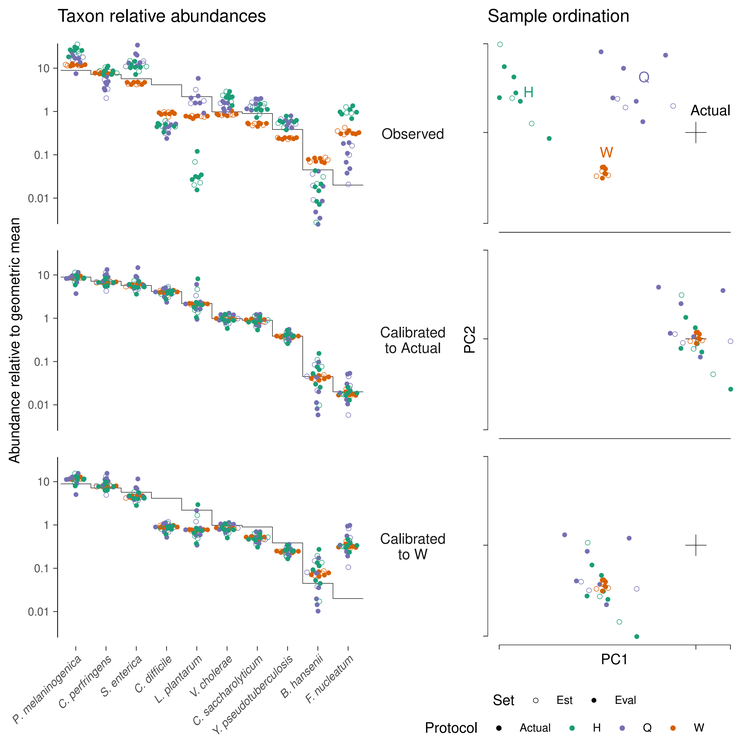
compare calibration to actual composition and calibration to a reference protocol

Results

Model performance in the 16S-sequenced cell mixtures



Bias and calibration in the shotgun-sequenced spike-ins

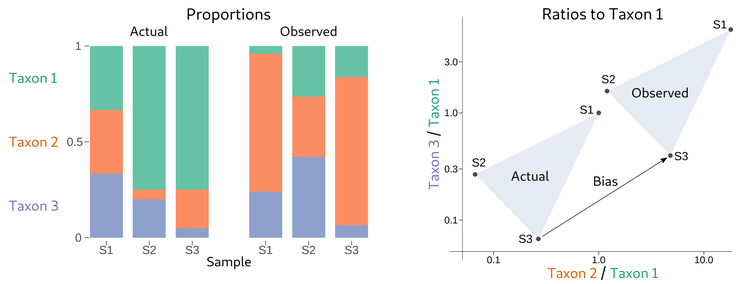


Next steps

Which (meta-)analyses are more robust to bias?

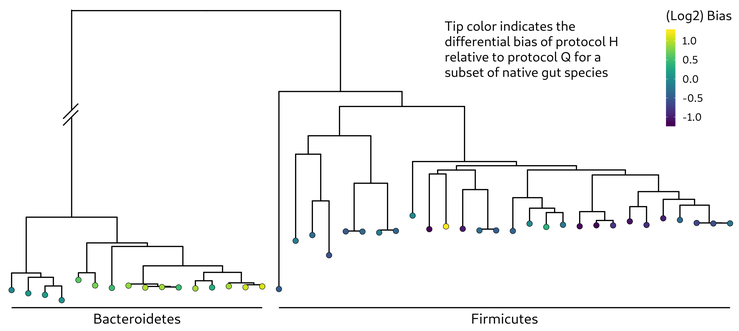
Analyses based on taxon ratios (e.g., CoDA methods) are theoretically bias invariant

$$\frac{O_i(S1)}{O_j(S1)} \Big/ \frac{O_i(S2)}{O_j(S2)} = \frac{A_i(S1)B_i}{A_j(S1)B_j} \Big/ \frac{A_i(S2)B_i}{A_j(S2)B_j} = \frac{A_i(S1)}{A_j(S1)} \Big/ \frac{A_i(S2)}{A_j(S2)}$$



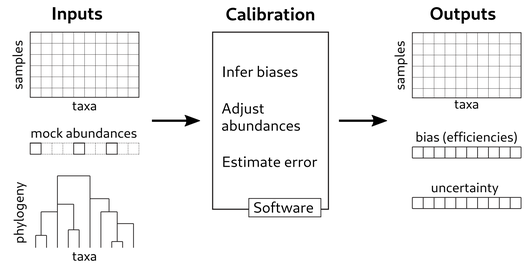
But factors such as biological zeros, cross-sample contamination, and taxonomic aggregation complicate this picture

How does bias evolve?



Natural communities often contain hundreds of species. If bias evolves slowly, we need only measure bias for a small fraction

Long-term goal: Software to estimate the bias of all taxa and calibrate the measurements of complex communities



References

McLaren MR, Willis AD, Callahan BJ. 2019. Consistent and correctable bias in metagenomic sequencing measurements. bioRxiv 559831.

Brooks JP, et al. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 15:66.

Costea PI, et al. 2017. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 35:1069-1076.