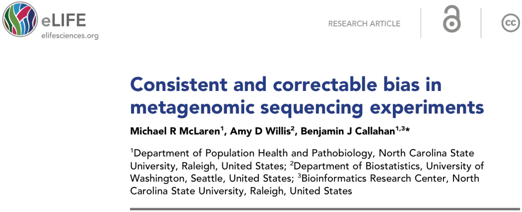
Biases and Ways to Address It



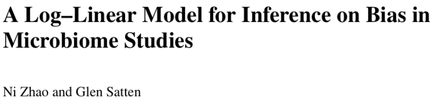
eLIFE
elifesciences.org

RESEARCH ARTICLE

Consistent and correctable bias in metagenomic sequencing experiments

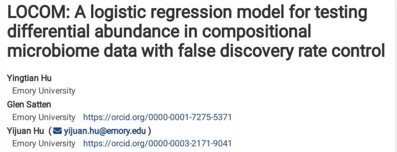
Michael R McLaren[1], Amy D Willis[2], Benjamin J Callahan[1,3]*

[1]Department of Population Health and Pathobiology, North Carolina State University, Raleigh, United States; [2]Department of Biostatistics, University of Washington, Seattle, United States; [3]Bioinformatics Research Center, North Carolina State University, Raleigh, United States



A Log–Linear Model for Inference on Bias in Microbiome Studies

Ni Zhao and Glen Satten



LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control

Yingtian Hu
Emory University
Glen Satten
Emory University https://orcid.org/0000-0001-7275-5371
Yijuan Hu (yijuan.hu@emory.edu)
Emory University https://orcid.org/0000-0003-2171-9041

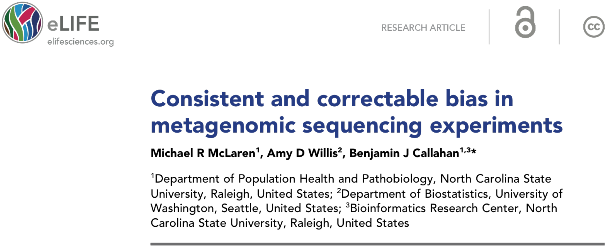
Bias is ubiquitous in microbiome sequencing

Bias: systematic distortion of measurement from the true values

- Sample collection and storage
- Fresh vs frozen samples
- DNA extraction – Gram positives and negatives (*Bacteroidetes-Firmicutes* ratios)
  - Mechanical and/or chemical/enzymatic lysis steps
- Sequencing strategy
  - Primer choices, PCR bias, GC content in genes, sequencing platforms

Bias is ubiquitous in microbiome sequencing

Bias: multiplicative bias generative model



eLIFE elifesciences.org

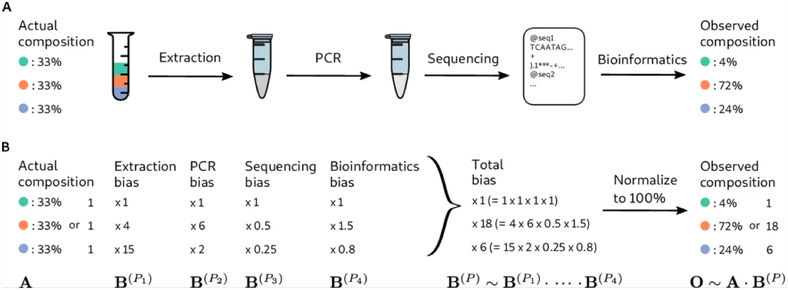
RESEARCH ARTICLE

Consistent and correctable bias in metagenomic sequencing experiments

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

[1]Department of Population Health and Pathobiology, North Carolina State University, Raleigh, United States; [2]Department of Biostatistics, University of Washington, Seattle, United States; [3]Bioinformatics Research Center, North Carolina State University, Raleigh, United States

MWC model



Bias arises throughout an MGS workflow, creating systematic error between the observed and actual compositions.

Panel **A** illustrates a hypothetical marker-gene measurement of an even mixture of three taxa. The observed composition differs fro the actual composition due to the bias at each step in the workflow. Panel **B** illustrates our mathematical model of bias, in which bi multiplies across steps to create the bias for the MGS protocol as a whole.

MWC model

- Model: $O/A \sim B$
- Taxa ratios are independent to sample compositions

$$\frac{O_i}{O_j} = \frac{A_i B_i}{A_j B_j}$$

- The observed proportion of taxon i

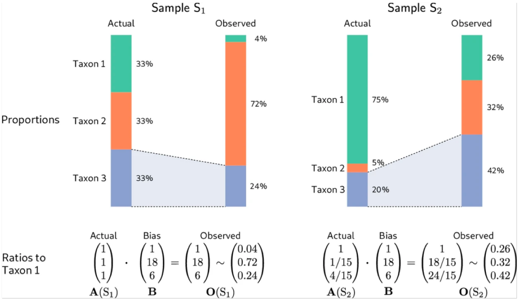
$$Pr(O)_i = \frac{O_i}{\sum_{j=1}^{K} O_j} = \frac{Pr(A)_i B_i}{\sum_{j=1}^{K} Pr(A)_j B_j}$$

- Ratios in taxa proportions depends on sample composition:

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

Analyses on taxon ratios are insensitive to bias

while analyses based on taxon proportions can give spurious results



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

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

- **Consistent multiplicative bias causes systematic error in taxon ratios, but not taxon proportions, that is independent of sample composition.**
- The even community from Figure 1 and a second community containing the same three taxa in different proportions are measured by a common MGS protocol. Measurements of both samples are subject to the same bias, but the magnitude and direction of error in the taxon proportions depends on the underlying composition (top row). In contrast, when the relative abundances and bias are both viewed as ratios to a fixed taxon (here, Taxon 1), the consistent action of bias across samples is apparent (bottom row).

Mock communities and Spike-in samples

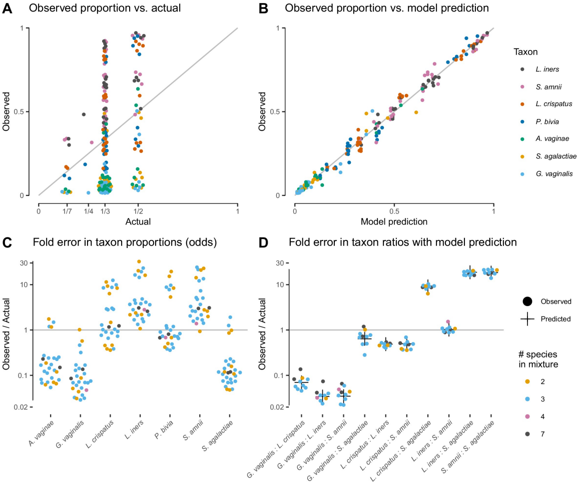
- Mock community
  - Control samples with predetermined abundances of bacterial mixture (DNA or whole cell mock community)
  - In-house mock community vs commercial mock community (American Type Culture Collection (ATCC) and Zymo Research)
- Spike-in
  - Bacteria with known identity and abundance added in real samples
  - Can also be used to calibrate measurements

The Brooks dataset

Mock community

- A mock community with seven bacteria
- 240 samples with 58 combinations of bacteria
- Processed in six plates processed in different ways:
  - cells, DNA, and PCR products

# of taxa present	1	2	3	4	7
	27	75	129	3	6



McLaren, Willis and Callahan, 2019. eLife

What does MWC model mean?

The MWC model:

$$\tilde{p}_j = \frac{p_j e^{\beta_j}}{\sum_{j'=1}^{J} p_{j'} e^{\beta_{j'}}}, \tag{1}$$

- Multiplicative and deterministic model of the bias factor
- the bias factor of one taxon is not impacted by other taxa in the sample ?
  - model bias one taxon instead of the whole community level.
- the bias depends on experimental (eg., extraction protocol and PCR parameters) and bacterial properities .
  - one taxon may have a low bias factor using one DNA extraction protocol, but may have a high bias factor using a different extraction protocol.
- The normalization factor depends on the true prevalence and bias factors of all taxa in the sample.

The Brooks dataset

- A mock community with seven bacteria
- 240 samples with 58 combinations of bacteria
- Processed in six plates processed in different ways:
  - cells, DNA, and PCR products

# of taxa present	1	2	3	4	7
	27	75	129	3	6

Research question:

- Do the plates have different bias between them?
- Do one taxon impact the bias of other taxa?

Model Setup

- Feature (OTU or ASV) table with N rows and J columns
- p_{ij}, $\tilde{p}_{ij}$: true RR and observed RR of taxon j in sample i
- $\Delta_{ij} = 1$ is the taxon j is present in sample i and is 0 otherwise.
- $\sum_{j=1}^{J} \tilde{p}_{ij} \Delta_{ij} = \sum_{j=1}^{J} p_{ij} \Delta_{ij} = 1$ for all samples.

Extend the MWC model:

$$E\left(\tilde{p}_{ij}\right) = \frac{p_{ij} e^{\beta_j}}{\sum_{j'=1}^{J} p_{ij'} e^{\beta_{j'}}}, \tag{2}$$

Full Model

$$ln\,\tilde{p}_{ij} = ln\,p_{ij} + X_{i\cdot} \cdot \beta_{\cdot j} + \alpha_i + \epsilon_{ij} \ , \tag{3}$$

Estimation

- Right multiplying $ln\,\tilde{p}_{i\cdot}$ given in equation (3) by the compositional projection operator $P_i = Diag(\Delta_{i\cdot}) - \frac{1}{n_i}\Delta_{i\cdot}^T\Delta_{i\cdot}$ (n_i is the number of taxa present in sample i) eliminates any term that is constant in i

Final Model:

$$Y_{ij} = X_{i\cdot}\beta P_i + e_{ij} \tag{4}$$

with $Y_{i\cdot} = (ln\,\tilde{p}_{i\cdot} - ln\,p_{i\cdot})\,P_i$ and $e_{i\cdot} = \epsilon_{i\cdot}P_i$.

Full Model

$$Y_{ij} = X_{i\cdot}\beta P_i + e_{ij}$$

that $Y_{ij} = 0$ if $\Delta_{ij} = 0$.

the vector trick for estimation

$$\text{vec}(Y_{i\cdot}) = P_i \otimes X_{i\cdot}\,\text{vec}(\beta) + \text{vec}(e_i)\ ,$$

$$\begin{pmatrix} \text{vec}(Y_{1\cdot}) \\ \text{vec}(Y_{2\cdot}) \\ \vdots \\ \text{vec}(Y_{N\cdot}) \end{pmatrix} = \begin{pmatrix} P_1 \otimes X_{1\cdot} \\ P_2 \otimes X_{2\cdot} \\ \vdots \\ P_N \otimes X_{N\cdot} \end{pmatrix} \text{vec}(\beta) + \text{vec}(e^T) =: \mathbb{X}\text{vec}(\beta) + \text{vec}(e^T). \quad (5)$$

We use least square for estimation

Example 1: Main effect

Design

- Z: single binary indicator on plates
- $X_{i\cdot} = (I[Z_i = 1], I[Z_i = 2])$
- $$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1J} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2J} \end{pmatrix}, \tag{6}$$

Adding a constant to the rows doesn't impact the bias.

Example 2: Interaction effects

Interactions *between taxa*.

- Having taxon j in the sample has no effect on the (relative) bias of taxon k.
- $X_{i\cdot} = (1, \Delta_{i1}, \Delta_{i2}, \cdots, \Delta_{iJ})$

$$\beta = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_J \\ 0 & \beta_{12} & \cdots & \beta_{1J} \\ \beta_{21} & 0 & \cdots & \beta_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{J1} & \beta_{J2} & \cdots & 0 \end{pmatrix}, \tag{7}$$

Example 2: Interaction effects

Alternative way to specify:

- $X_{i\cdot} = (\Delta_{i1}, \Delta_{i2}, \cdots, \Delta_{iJ})$

$$\beta = \begin{pmatrix} \beta_1 & \beta_{12} & \cdots & \beta_{1J} \\ \beta_{21} & \beta_2 & \cdots & \beta_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{J1} & \beta_{J2} & \cdots & \beta_J \end{pmatrix}, \tag{8}$$

- adding a constant to each row doesn't change the bias
- adding values to the diagonal elements doesn't change the interaction

Inference

Null hypothesis: H_0

$$\mathbb{C}\,\mathsf{vec}(\beta) = 0, \tag{9}$$

In a lot of context, $\mathbb{C}\,\mathsf{vec}(\beta) = C\beta Q$ that $\mathbb{C} = Q^T \otimes C$

- Q accounts for the nonidentifiability in β matrix.
- $Q = I - e_j 1^T$ subtracts β_j from every other column of β

Examples:

- Design: 3 taxa, no covariates, no missing taxa.
- H_0 : No bias across taxa $\mathsf{vec}(\beta) = \beta_{1\cdot}^T = (\beta_{11}, \beta_{12}, \beta_{13})$

$$\mathbb{C} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Example 1: Main effect

Design

- Z: single binary indicator on plates
- $X_{i\cdot} = (I[Z_i = 1], I[Z_i = 2])$
- Hypothesis: whether bias factors differ between plates.
- $$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1J} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2J} \end{pmatrix}, \tag{10}$$
- Contrast matrix: $\beta_{1j} - \beta_{1j'} = \beta_{2j} - \beta_{2j'}$ for $j \neq j'$

Example 2: Interaction effects

- $X_{i\cdot} = (\Delta_{i1}, \Delta_{i2}, \cdots, \Delta_{iJ})$

$$\beta = \begin{pmatrix} \beta_1 & \beta_{12} & \cdots & \beta_{1J} \\ \beta_{21} & \beta_2 & \cdots & \beta_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{J1} & \beta_{J2} & \cdots & \beta_J \end{pmatrix},$$

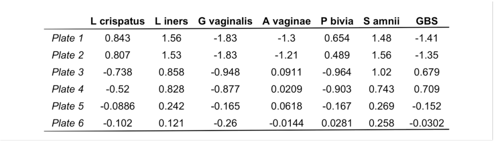
- taxon j does not affect the bias factors of the other taxa
  - $\beta_{jk} = \beta_{jk'}, k \neq k' \neq j$.
  - With 7 taxa in Brooks data, 5-DF test.
- any interaction effect between taxa
  - $\beta_{jk} = \beta_{jk'}, k \neq k' \neq j$ for all j, k, k'
  - With 7 taxa in Brooks data, 35-DF test.
- taxon j and j' have the same interaction effect on bias of other taxa

Inference

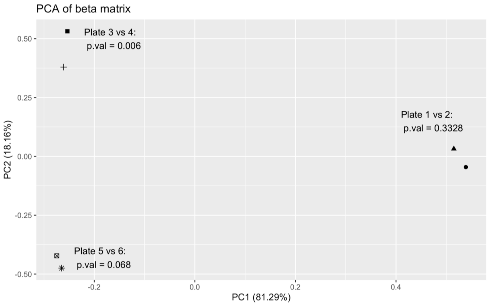
$$F = \frac{RSS_0 - RSS}{RSS} = \frac{\sum_i \sum_j (r_{0,ij}^2 - r_{ij}^2)}{\sum_i \sum_j r_{ij}^2} \ , \tag{11}$$

- The null distribution is hard to quantify.
- Permutation or bootstrapping: correlation between taxa, the compositionally and the structural missing.
- Permutation based on de-correlated residuals

Brooks data result



	L crispatus	L iners	G vaginalis	A vaginae	P bivia	S amnii	GBS
Plate 1	0.843	1.56	-1.83	-1.3	0.654	1.48	-1.41
Plate 2	0.807	1.53	-1.83	-1.21	0.489	1.56	-1.35
Plate 3	-0.738	0.858	-0.948	0.0911	-0.964	1.02	0.679
Plate 4	-0.52	0.828	-0.877	0.0209	-0.903	0.743	0.709
Plate 5	-0.0886	0.242	-0.165	0.0618	-0.167	0.269	-0.152
Plate 6	-0.102	0.121	-0.26	-0.0144	0.0281	0.258	-0.0302



Brooks data

Table: Interaction analysis results for the Brooks data

A: Interaction tests results using all data								
	L crispatus	*L iners*	*G vaginalis*	*A vaginae*	*P bivia*	*S amnii*	*GBS*	Overall
All plates	0.792	0.745	0.147	0.499	0.691	0.887	0.277	0.793
B: Interaction tests results in each stratum								
Plates 1 & 2	0.023	0.7098	0.612	0.016	0.645	0.263	0.003	0.009
Plates 3 & 4	0.052	0.0138	0.639	0.548	0.767	0.933	0.435	0.2198
Plates 5 & 6	0.442	0.537	0.482	0.947	0.862	0.656	0.679	0.616

Note: All the main effect are highly significant!

Discussion

- Currently only works for mock communities with known relative abundances
- Multivariate permutation with structural missing
- Extension to real communities and to samples with unknown true RR
- The fact that there is no (little) interaction opens the door to many future analyses.

Bias resistant modeling of microbiome

ANCOM-BC:

Analysis of Compositions of Microbiomes with Bias Correction

- Observed read counts are a fraction of the total molecules in samples
- Assumes this fraction is the same across taxa (yet can be different between samples).
- Estimate the sampling fractions and then model the log of reads via a linear regression with an offset term (the estimated sampling fraction).

Lin & Peddada, *Nature Communication*, 2020

LOCOM: (LOgistic COMpositional)

- $$\log(p_{ij}) = \log(\pi_{ij}) + \gamma_j + \alpha_i$$
  - γ_j is taxon-specific *bias factor*
  - α_i is the sample-specific normalization factor
  - $\sum_{j=1}^{J} p_{ij} = 1$. $\sum_{j=1}^{J} \pi_{ij} = 1$
- Extend the model: describe π_{ij} by a baseline RA and covariate effect

$$\log(p_{ij}) = \log(\pi_j^0) + X_i^T \beta_j + \gamma_j + \alpha_i$$

LOCOM

- Extend the model: describe π_{ij} by a baseline RA and covariate effect

$$\log(p_{ij}) = \log(\pi_j^0) + X_i^T \beta_j + \gamma_j + \alpha_i$$

- β_j is not estimable due to γ_j and α_i.
- Odds ratios of observed relative abundances

$$\log\left(\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}\right) = (X_i - X_{i'})^T(\beta_j - \beta_{j'})$$

- $\beta_j - \beta_{j'} = 0$ corresponds to testing $\frac{p_{ij}}{p_{ij'}} = \frac{p_{i'j}}{p_{i'j'}}$

Equivalent individual logistic regression models

- The proposed model is a polychotomous logistic regression of the full $n \times J$ taxa count table.
- Transformed into individualized logistic regression of two taxa at a time
- Let $\mu_{ij} = p_{ij}/(p_{ij} + p_{iJ})$, take taxon J as the reference

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \theta_j + X_i^T(\beta_j - \beta_J), \quad 1 \leq j \leq J-1$$

- $\theta_j = [\log(\pi_j^0) - \log(\pi_J^0)] + (\gamma_j - \gamma_J)$ is a free nuisance parameter.
- Estimation is done via a Firth-corrected score function

$$U_j(\beta_j) = \sum_{i=1}^{n} [Y_{ij} - M_{ij}\mu_{ij} + h_i(0.5 - \mu_{ij})]X_i = 0$$

that $M_{ij} = Y_{ij} + Y_{iJ}$, and h_i is the i-th diagonal element of the weighted hat matrix.

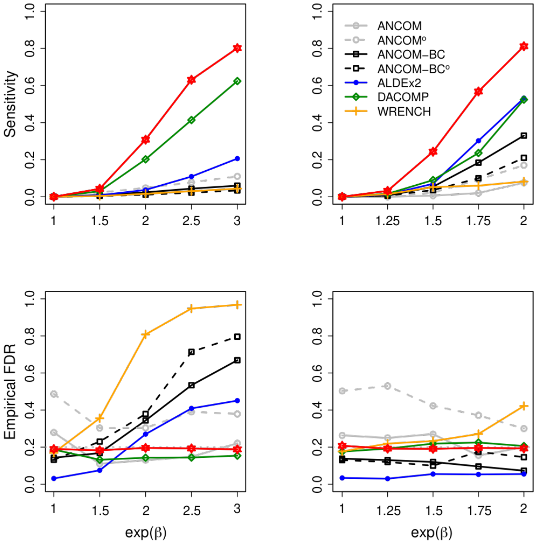
Testing individual taxa

- $\beta_{j1} = 0$ corresponds to a null taxon only when the reference is null.
- Assumes that at most half of all taxa can be alternative. Then $\text{median}_{j'=1,\cdots,J}\beta_{j',1}$ corresponds some estimate for a null taxon.

$$H_{j0} : \beta_{j1} - \text{median}_{j'=1,\cdots,J}\beta_{j',1} = 0$$

- A permutation framework was used to evaluate statistical significance.

Testing global hypothesis: p-value combination approach via the Harmonic means of p-values.


Sensitivity
Empirical FDR
exp(β)
exp(β)
0.0
0.2
0.4
0.6
0.8
1.0
1
1.5
2
2.5
3
1
1.25
1.5
1.75
2
ANCOM
ANCOM^{o}
ANCOM–BC
ANCOM–BC^{o}
ALDEx2
DACOMP
WRENCH