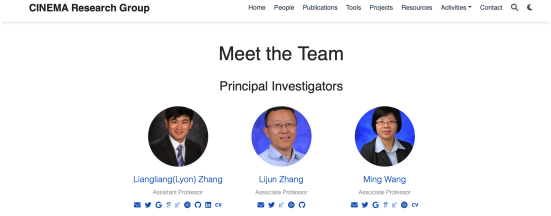
Decoding Microbiome dual-Mediation: A Tool for Advanced Zero-Inflated Data Analysis

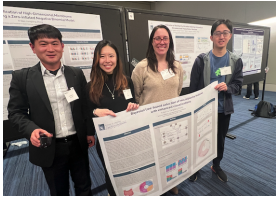
Xi Qiao, Liangliang Zhang

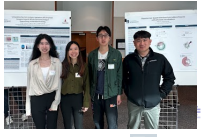
Presented by Liangliang Zhang
Department of Population and Quantitative Health Sciences

Our Group

CINEMA https://cwru-cinema.com/tour/



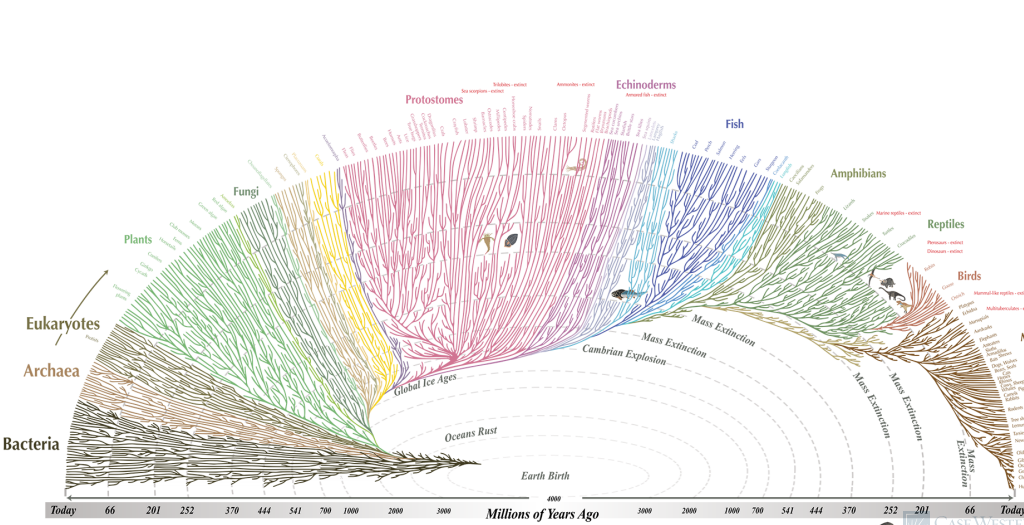




Outline

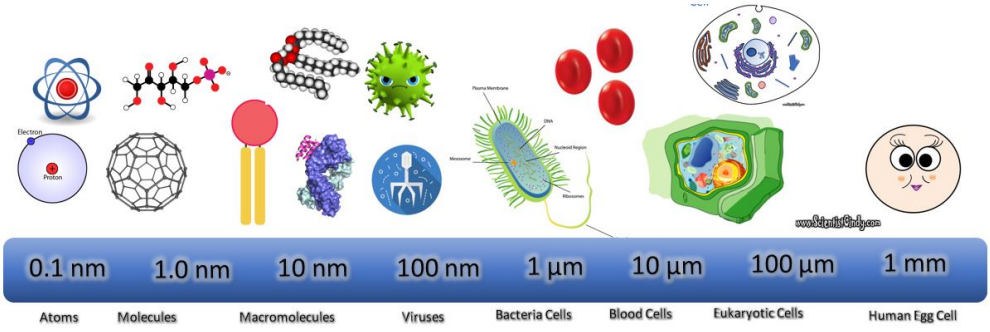
Microbiology and Human Microbiome Research

The world of bacteria holds far more genetic diversity



All the major and many of the minor living branches of life are shown on this diagram, but only a few of those that have gone extinct are shown. Example: Dinosaurs - extinct

Visual comparision of Microorganism Sizes



Microorganisms reside in every part of human body

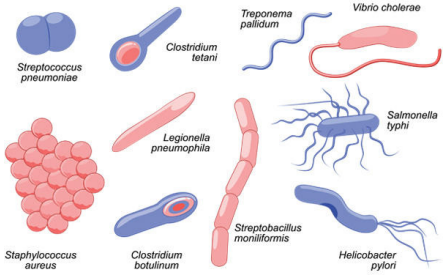


Figure: Various bacteria live on earth

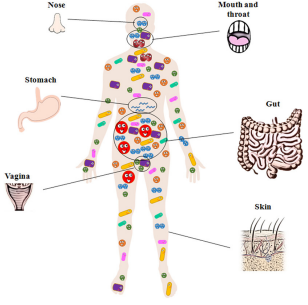


Figure: Distinct bacteria live in different body sites

Microbiota dysbiosis linked with health and diseases

Microbiome constitutes a human organ

- Microorganisms interact with body host's environment: diet, antibiotics, chemotherapy, etc.
- I have extensively worked on linking microbiome at different body sites to patient outcomes.

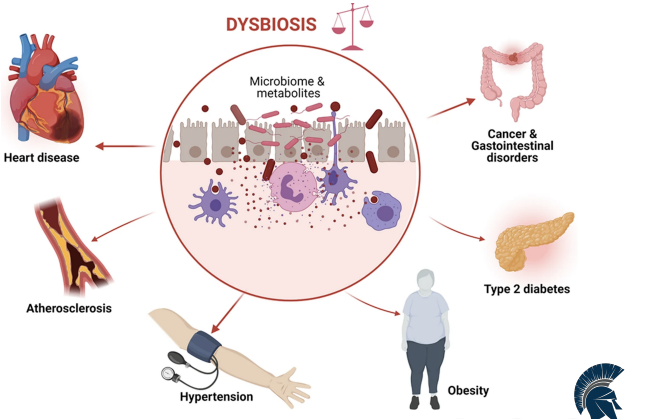
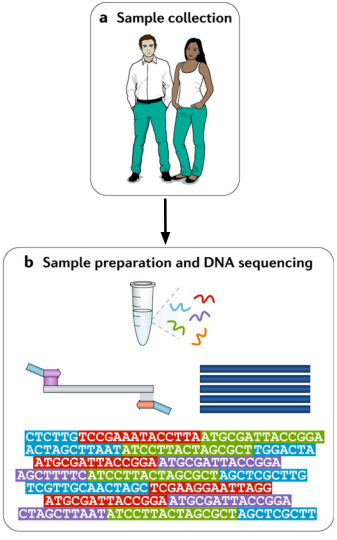
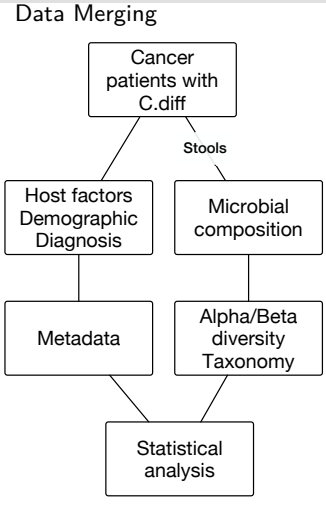


Figure from Masenga SK, et al. (2022)

Microbiome Data and Host-Microbiome Association

Steps of quantifying bacteria composition





Typical formats of microbiome data

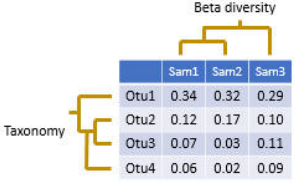
- **OTU/ASV table**

	Sam1	Sam2	Sam3
Otu1	660	605	560
Otu2	362	440	180
Otu3	153	60	170
Otu4	86	20	120

-
  - Operational taxonomic unit (OTU) are used to categorize bacteria based on sequence similarity.
  - An amplicon sequence variant (ASV) is referred to as exact sequence variants, zero-radius OTUs or sub-OTUs.

Typical formats of microbiome data

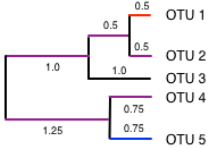
- **Proportion table**



	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

Typical formats of microbiome data

- Phylogenetic tree



Host-Microbiome Association Study

Tumor Microbiome and Pancreatic Cancer

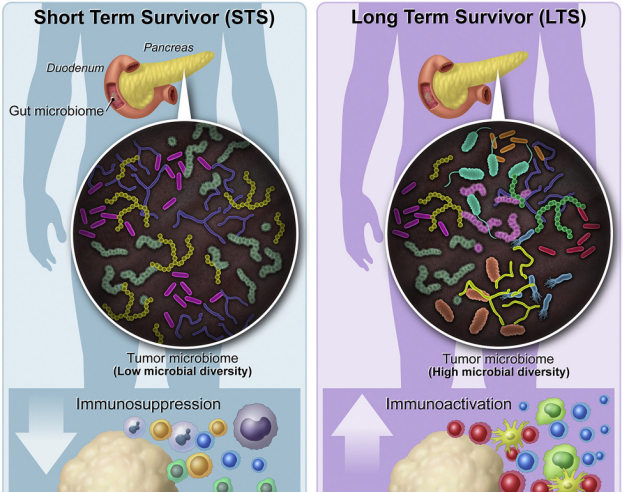
Cell
Volume 178, Issue 4, 8 August 2019, Pages 795-806.e12
CellPress

Article

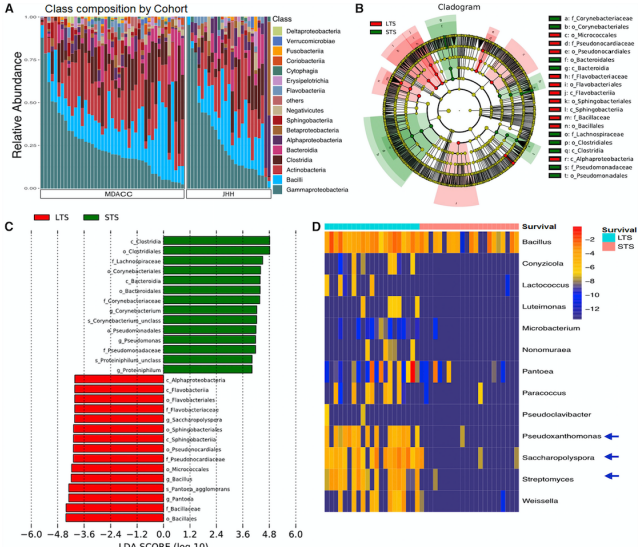
Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes

Erick Riquelme [1, 2, 18], Yu Zhang [1, 18], Liangliang Zhang [3, 4], Maria Montiel [1], Michelle Zoltan [1], Wenli Dong [3], Pompeyo Quesada [1], Ismet Sahin [5], Vidhi Chandra [1], Anthony San Lucas [6], Paul Scheet [6], Hanwen Xu [1], Samir M. Hanash [1, 7], Lei Feng [3], Jared K. Burks [8], Kim-Anh Do [3], Christine B. Peterson [3], Deborah Nejman [9] ... Florencia McAllister [1, 16, 17, 19]

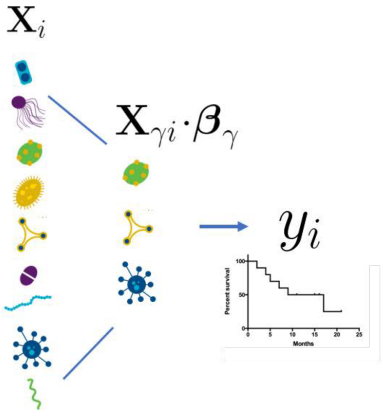
Binary outcome in the Pancreatic cancer project



Identify differential features between two groups

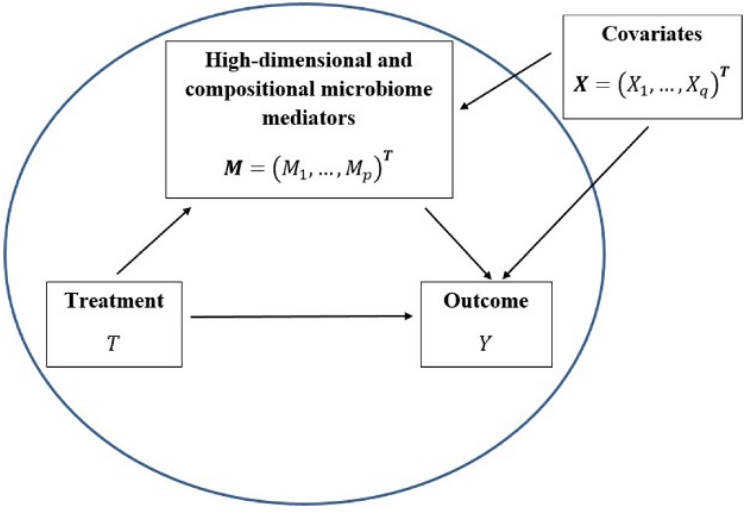


Linear Model and Variable Selection



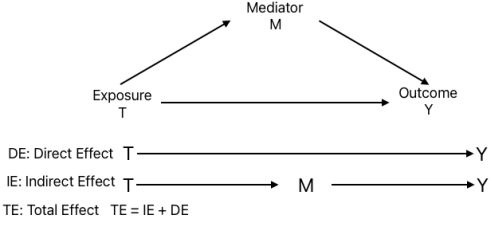
Mediation model

General Structure of Mediation Model



Causal Mediation Analysis

- In clinical trials and epidemiological studies, causal mediation analysis is to explain the underlying mechanism by which the effect of an exposure on the outcome is mediated through a casual intermediate variable or mediator.
- General Approaches
  - Structural equation modeling (SEM)
    [Baron and Kenny, 1986, MacKinnon and Dwyer, 1993, MacKinnon et al., 2002].
  - Counterfactual framework with potential outcomes [Albert, 2008, Robins and Greenland, 1992].



Zero-Inflated Microbiome Mediators

How to characterize the microbiome mediators?

- Count Data
- Zero-inflated Data
- High-dimensional Data

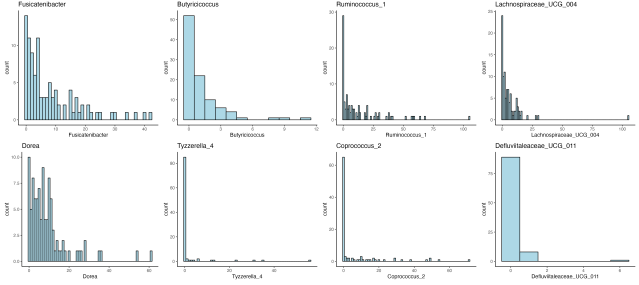


Figure: Histograms of genus level microbiome features from real human gut microbiome data [Wu et al., 2011]

Bayesian and related methods

Zero Inflated Mediation Analysis

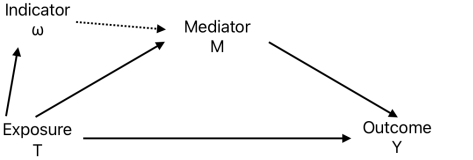
- Latent variable ω_{ij}: indicate the presence of structural zeros. For instance, patients undergoing antibiotic treatment are more likely to exhibit a zero count for a specific microbiome feature.
- In the context of the jth microbiome feature within the ith subject,

$$M_{ij} = \begin{cases} M_{ij}^* & \text{, if } \omega_{ij} = 0 \\ 0 & \text{, if } \omega_{ij} = 1 \end{cases}$$

$$\omega_{ij} \sim \text{Bernoulli}(\pi_{ij}).$$

Under counterfactuals,

$$\begin{aligned} \text{NDE} &= \text{E}(Y_{a^*, M_{a,\omega_a}} | C_i) - \text{E}(Y_{a, M_{a,\omega_a}} | C_i) \\ \text{NIE} &= \text{E}(Y_{a^*, M_{a^*,\omega_{a^*}}} | C_i) - \text{E}(Y_{a^*, M_{a^*,\omega_a}} | C_i) \\ &\quad + \text{E}(Y_{a^*, M_{a^*,\omega_a}} | C_i) - \text{E}(Y_{a^*, M_{a,\omega_a}} | C_i) \\ &= \text{NIE}_{prevelance} + \text{NIE}_{abundance} \\ \text{TE} &= \text{E}(Y_{a^*, M_{a^*,\omega_{a^*}}} | C_i) - \text{E}(Y_{a, M_{a,\omega_a}} | C_i) \\ &= \text{NIE} + \text{NDE} \end{aligned}$$



ZIMMA Framework

- Mediator Model

$$M_{ij} = \begin{cases} M_{ij}^* & \text{, if } \omega_{ij} = 0 \\ 0 & \text{, if } \omega_{ij} = 1 \end{cases}$$

$$\omega_{ij} \sim \text{Bernoulli}(\pi_{ij}),$$

  - Prevalence Model:

$$\text{logit}(\pi_{ij}) = \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \gamma_{0j} + \gamma_{Tj} T_i + \gamma_{Cj}^T \boldsymbol{C}_i. \quad (1)$$

  - Abundance Model:

$$M_{ij}^* \sim \text{NB}(\mu_{ij}, \tau_j)$$
$$\mu_{ij} = S_i A_{ij}$$

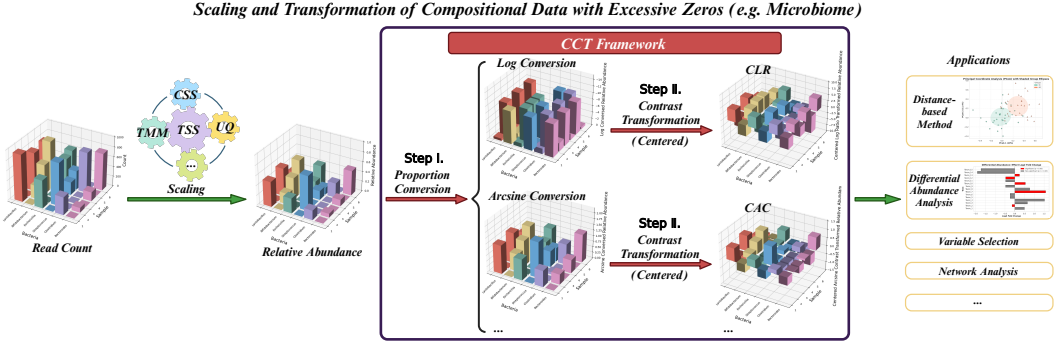
.

$$\log(A_{ij}) = \beta_{0j} + \beta_{Tj} T_i + \beta_{Cj}^T \boldsymbol{C}_i \quad (2)$$

- Outcome Model

$$\text{E}(Y_i) = \alpha_0 + \alpha_T T_i + \boldsymbol{\alpha}_M^T \boldsymbol{M}_i + \boldsymbol{\alpha}_C^T \boldsymbol{C}_i \quad (3)$$

Scaling and size factors



For more details, please refer to our paper
https://www.sciencedirect.com/science/article/pii/S200103702400374X

Expected abundance

Specifically, $M_{ij}^* \sim \text{NB}(\mu_{ij}, \tau_j)$ has the following probability mass function (PMF) [Pillow and Scott, 2012]:

$$p(M_{ij}^* = m^* | \mu_{ij}, \tau_j) = \frac{\Gamma(m^* + \tau_j)}{\Gamma(\tau_j) m^*!} \left(\frac{\tau_j}{\mu_{ij} + \tau_j} \right)^{\tau_j} \left(1 - \frac{\tau_j}{\mu_{ij} + \tau_j} \right)^{m^*} \tag{4}$$

where m^* is a non-negative integer, $\Gamma(\cdot)$ is the Gamma function, and the parameters μ_{ij} and τ_j control the mean and dispersion, respectively.

The expected value of the observed taxon counts, M_{ij}, given the treatment group T_i and pre-treatment confounding variables $\boldsymbol{C_i}$, is:

$$\begin{aligned} \text{E}(M_{ij} \mid T_i, \boldsymbol{C_i}) &= (1 - \pi_{ij}) \text{E}(M_{ij} \mid \omega_{ij} = 1, T_i, \boldsymbol{C_i}) + \pi_{ij} \text{E}(M_{ij} \mid \omega_{ij} = 0, T_i, \boldsymbol{C_i}) \\ &= (1 - \pi_{ij}) \text{E}(M_{ij}^* \mid T_i, \boldsymbol{C_i}) \\ &= (1 - \frac{1}{1 + \exp(\gamma_{0j} + \gamma_{Tj} T_i + \boldsymbol{\gamma_{Cj}^T C_i})}) S_i \exp(\beta_{0j} + \beta_{Tj} T_i + \boldsymbol{\beta_{Cj}^T C_i}) \end{aligned} \tag{5}$$

Hypothesis on Indirect Effect

Under sequential ignorability assumption (no unmeasured confounding), for each of the mediator,

$$\text{Average NIE}_{prevelance} = \frac{1}{n}\sum_{i=1}^{n} \alpha_M^T(a^* - a) S_i e^{\beta_0+\beta_T a^*+\beta_C^T C_i}\left[\frac{e^{\gamma_0+\gamma_T a^*+\gamma_C^T C_i}}{e^{\gamma_0+\gamma_T a^*+\gamma_C^T C_i}+1} - \frac{e^{\gamma_0+\gamma_T a+\gamma_C^T C_i}}{e^{\gamma_0+\gamma_T a+\gamma_C^T C_i}+1}\right]$$

$$\text{Average NIE}_{abundance} = \frac{1}{n}\sum_{i=1}^{n} \alpha_M^T(a^* - a) \frac{e^{\gamma_0+\gamma_T a+\gamma_C^T C_i}}{e^{\gamma_0+\gamma_T a+\gamma_C^T C_i}+1} S_i\left[e^{\beta_0+\beta_T a^*+\beta_C^T C_i} - e^{\beta_0+\beta_T a+\beta_C^T C_i}\right]$$

$$\text{Average NIE}_{prevelance} = 0 \Leftrightarrow \alpha_M = 0 \text{ or } \gamma_T = 0,$$

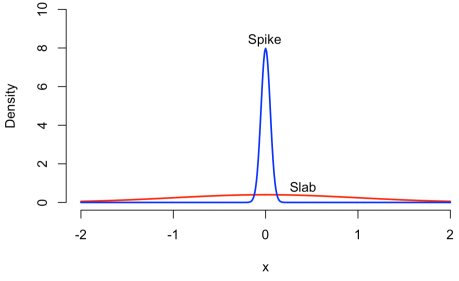
$$\text{Average NIE}_{abundance} = 0 \Leftrightarrow \alpha_M = 0 \text{ or } \beta_T = 0.$$

There is no indirect effect through $j$the microbiome mediator only if

$$\text{Average NIE}_{prevelance} = 0 \text{ and Average NIE}_{abundance} = 0$$

Mediator Selection through Spike and Slab prior

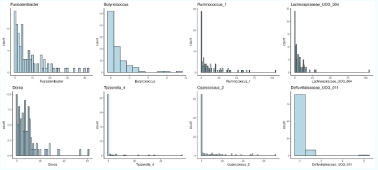
$$
\begin{aligned}
\alpha_M, \beta_T, \gamma_T &\sim \mathsf{N}(0, \delta\nu^2) \\
\delta &= (1 - \kappa_{\alpha,\beta,\gamma})\delta_0 + \kappa_{\alpha,\beta,\gamma}\delta_1 \\
\kappa_{\alpha,\beta,\gamma} &\sim \mathsf{Bernoulli}(\theta_{\alpha,\beta,\gamma}) \\
\nu^2 &\sim \mathsf{IG}(a, b) \\
\theta_{\alpha,\beta,\gamma} &\sim \mathsf{Beta}(\frac{1}{2}, \frac{1}{2})
\end{aligned}
$$



Unidentifiable source of zeros

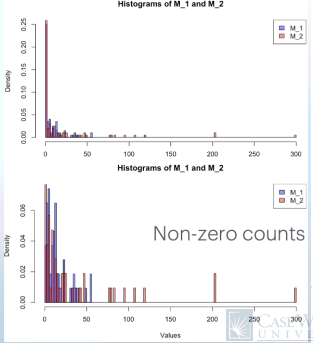
Unidentifiability

Current prior: $\gamma_0 \sim N(0,1)$, $\tau \sim gamma(0.01, 0.01)$



True (1/τ)	True (γ_0)	0%	True(τ)	Est.(τ)	Est.(γ_0)
0.5	***0 (p=0.5)***	67%	2	0.28 (0.08, 0.54)	-0.55 (-1.79, 0.52)
0.5	***-1 (p=0.27)***	46%	2	0.58(0.24, 0.98)	-0.87 (-1.87,-0.05)
0.5	***-2 (p=0.12)***	21%	2	2.02 (1.23, 2.93)	-1.63 (-2.37,-0.93)
0.5	***-3 (p=0.04)***	14%	2	2.18 (1.31, 3.06)	-1.95 (-2.70,-1.30)
1	***-3 (p=0.04)***	20%	1	0.84 (0.54, 1,18)	-2.08 (-3.06, -1.17)
2	***-3 (p=0.04)***	32%	0.5	0.47 (0.29, 0.68)	-1.86 (-2.88, -0.87)
5	***-3 (p=0.04)***	47%	0.2	0.22 (0.13, 0.33)	-1.53 (-2.71,-0.43)

Results: Different combination of over _dispersion and prevalence model intercept would result in similar zero%, but the non-zero counts would have different over dispersion (histograms on the right).



Empirical prior

1. $\tau_j \overset{\text{Ind}}{\sim} \text{Gamma}(m_{1j}, m_{2j})$
2. We fit an NB regression model using only the non-zero data and applying maximum likelihood estimation (MLE) to obtain an estimate of the dispersion, τ_j^+ [Venables and Ripley, 2002].
3. The mean of the gamma prior, m_{1j}/m_{2j}, is then set to τ_j^+, with a small variance, ν_τ^+ (e.g., 0.1) , specified as the prior variance $\frac{m_{1j}}{m_{2j}^2}$ to account for uncertainty, implying $m_{1j} = \frac{\tau_j^{+2}}{\nu_\tau^+}, m_{2j} = \frac{\tau_j}{\nu_\tau^+}$.

Algorithm

Algorithm 1 ZIMMA Posterior Sampling Algorithm

for each iteration from 1 to R **do**

Step 1: Update all parameter associated with j-th mediator, $j = 1, \ldots, P$.

for j from 1 to P **do**

draw latent structural zero indicator ω_{ij}:

$$p(\omega_{ij} = 1 | M_{ij}, \text{rest}) = \begin{cases} f(\omega_{ij} | M_{ij} = 0, \text{rest}), & \text{if } M_{ij} = 0 \\ 0, & \text{if } M_{ij} \neq 0 \end{cases}$$

draw Polya Gamma variable $\phi_{ij} | \omega_{ij}, \text{rest} \sim \text{PG}(1, \gamma_{0j} + \gamma_{Tj} T_i + \boldsymbol{\gamma_{Cj}} \boldsymbol{C_i})$.

draw $(\gamma_{0j}, \gamma_{Tj}, \boldsymbol{\gamma_{Cj}})^T | \kappa_{\gamma_j}, \text{rest} \sim \text{MVN}(\boldsymbol{\mu_\gamma}, \boldsymbol{\Sigma_\gamma})$.

draw $\beta_{0j}, \beta_{Tj}, \boldsymbol{\beta_{Cj}}, \tau_j$ using random walk Metropolis-Hastings sampling algorithm with a normal proposal distribution.

draw $\alpha_{Mj} | \kappa_{\alpha_j}, \text{rest} \sim \text{N}(\mu_{\alpha_j}, \sigma^2_{\alpha_j})$.

draw $\kappa_{\alpha_j}, \kappa_{\beta_j}, \kappa_{\gamma_j} | \alpha_{Mj}, \beta_{Tj}, \gamma_{Tj}, \text{rest} \sim \text{Bernoulli}\left(\frac{a_{\alpha_j,\beta_j,\gamma_j}}{a_{\alpha_j,\beta_j,\gamma_j} + b_{\alpha_j,\beta_j,\gamma_j}}\right)$.

draw $\nu^2_{\alpha_j}, \nu^2_{\beta_j}, \nu^2_{\gamma_j} | \kappa_{\alpha_j,\beta_j,\gamma_j}, \text{rest} \sim \text{IG}(l_1 + \frac{1}{2}, l_2 + \frac{(\alpha_{Mj}, \beta_{Tj}, \gamma_{Tj})^2}{2 b_{\alpha_j,\beta_j,\gamma_j}})$.

draw $\theta_{\alpha_j}, \theta_{\beta_j}, \theta_{\gamma_j} | \kappa_{\alpha_j,\beta_j,\gamma_j}, \text{rest} \sim \text{Beta}(a + \kappa_{\alpha_j,\beta_j,\gamma_j}, b + 1 - \kappa_{\alpha_j,\beta_j,\gamma_j})$.

end for

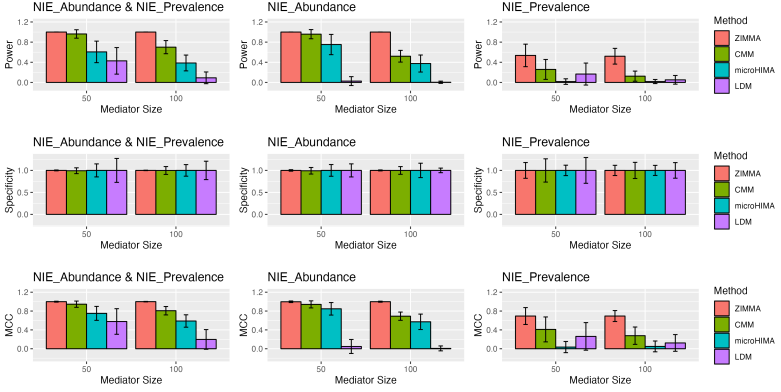
Step 2: Draw the rest coefficients in Equation (4): $\alpha_{0_j} | . \sim \text{N}(\mu_{\alpha_0}, \sigma^2_{\alpha_0})$, $\alpha_T \sim \text{N}(\mu_{\alpha_T}, \sigma^2_{\alpha_T})$, $\alpha_{C,\varepsilon} \sim \text{MVN}(\mu_{\alpha_{C,\varepsilon}}, \sigma_{\alpha^2_{C,\varepsilon}})$

Step 3: Draw the error term σ^2_Y Equation (4): $\sigma^2_Y | \alpha_0, \alpha_T, \boldsymbol{\alpha_M}, \boldsymbol{\alpha_C} \text{rest} \sim \text{IG}(\eta_{\sigma^2_Y}, \xi_{\sigma^2_Y})$.

Simulation and applications

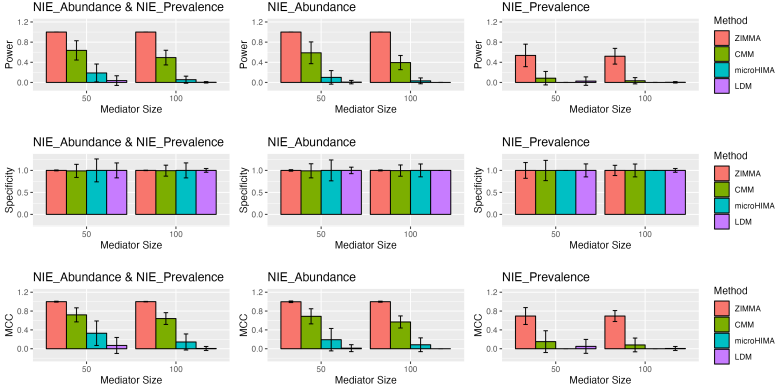
Simulation Results

- $N = 100$; $\tau = 0.5$; Effect size $= 1$
- Compared Methods: CMM [Sohn and Li, 2019]; microHIMA [Zhang et al., 2021]; LDM [Yue and Hu, 2022]



Simulation Results

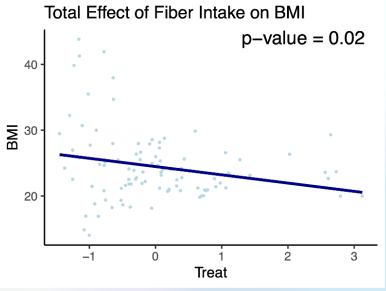
- $N = 100$; τ = real data median; Effect size = 1
- Compared Methods: CMM [Sohn and Li, 2019]; microHIMA [Zhang et al., 2021]; LDM [Yue and Hu, 2022]



Real Data Application on COMBO Study

Application1: COMBO

- N = 98
- Exposure/treatment: fiber intake
- Outcome: BMI
- P = 99 (Genus level, Prevalence > 10)

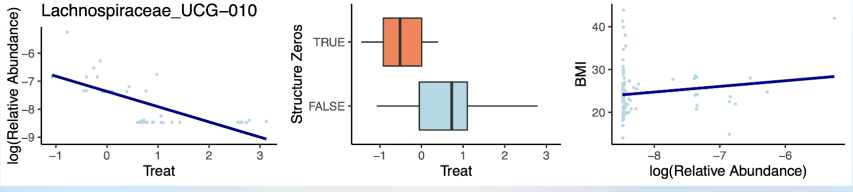


Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., & Lewis, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science (New York, N.Y.), 334(6052), 105–108. https://doi.org/10.1126/science.1208344

Real Data Application on COMBO Study

Indirect effect through prevalence and abundance (NIE_AP)

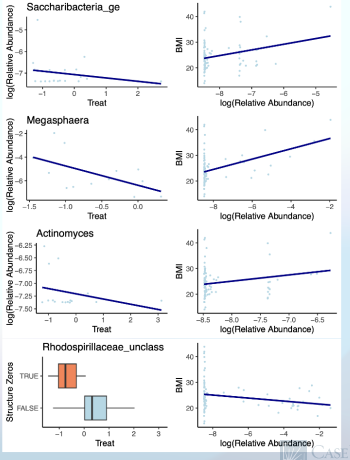
Feature Name	Abundance Estimate (PIP)	Prevalence Estimate (PIP)	Outcome Estimate (PIP)
Lachnispriraceae_UCG-010	-1.61 (0.90)	-1.50 (0.82)	0.96 (0.84)



Real Data Application on COMBO Study

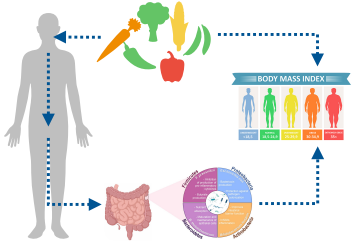
Indirect effect through abundance or prevalence (NIE_A, NIE_P)

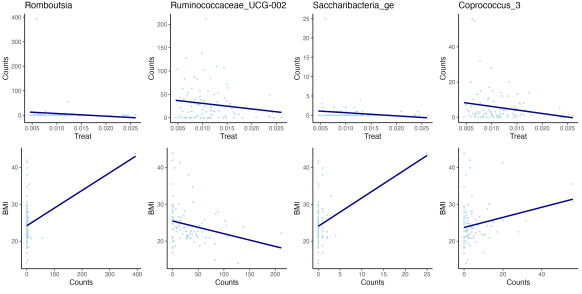
Feature Name	Abundance Estimate (PIP)	Prevalence Estimate (PIP)	Outcome Estimate (PIP)
Saccharibacteria	-1.02 (0.77)		0.50 (0.54)
Megasphaera	-2.68 (0.95)		0.64 (0.62)
Actinomyces	-3.60 (0.91)		0.44 (0.53)
Rhodospirillaceae_unclass		-0.64 (0.54)	-0.49 (0.61)



Real Data Application on COMBO Study

Active Genus Features	Phylum	NIE_Category	NIE Direction
Romboutsia	Firmicutes	$NIE_{abundance}$	- + →-
Ruminococcaceae_UCG-002	Firmicutes	$NIE_{abudance}$	- - →+
Saccharibacteria_ge	Saccharibacteria	$NIE_{abundance}$	- + →-
Coprococcus_3	Firmicutes	$NIE_{abundance}$	- + →-

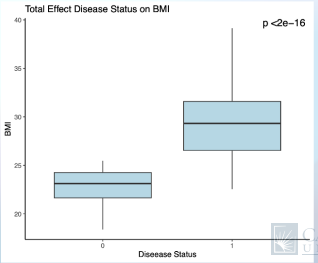




Real Data Application on Cardiovascular Study

Application2: Cardiometabolic Disease

- N = 220
- Male, No diabetes.
- Exposure/treatment: HC (Heathy Control) vs.MMC (individuals with features of the metabolic syndrome and, thus, at increased risk of ischemic heart disease (IHD)).
  - HC (Status = 0) = 104
  - MMC (Status = 1) = 116
- Outcome: BMI
- P = 106 (Genus level, Prevalence > 10%)

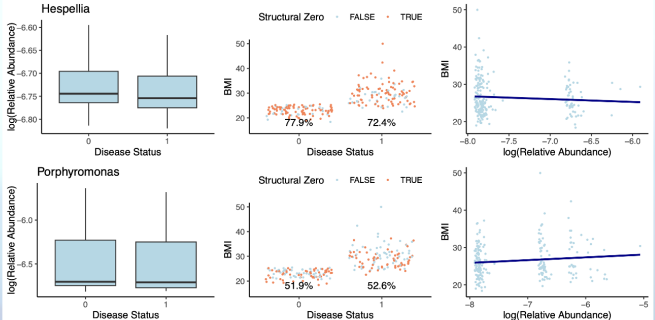


Fromentin, S., Forslund, S.K., Chechi, K. *et al.* Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat Med* **28**, 303–314 (2022). https://doi.org/10.1038/s41591-022-01688-4

Real Data Application on Cardiovascular Study

Indirect effect through prevalence and abundance (NIE_AP)

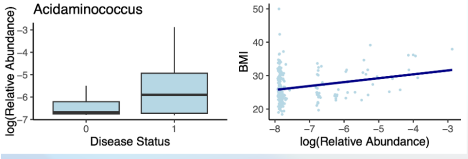
Feature Name	Abundance Estimate (PIP)	Prevalence Estimate (PIP)	Outcome Estimate (PIP)
Hespellia	-8.68 (0.99)	-0.51 (0.78)	-0.51 (0.78)
Porphyromonas	-4.28 (0.95)	0.41 (0.68)	0.30 (0.53)

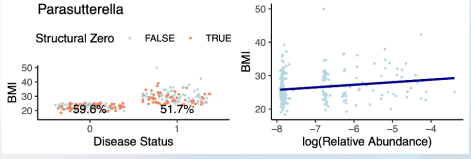


Real Data Application on Cardiovascular Study

Indirect effect through abundance or prevalence (NIE_A, NIE_P)

Feature Name	Abundance Estimate (PIP)	Prevalence Estimate (PIP)	Outcome Estimate (PIP)
Acidaminococcus	1.27 (0.83)		0.85 (0.97)
Parasutterella		-1.11 (0.62)	0.41 (0.68)





Conclusions and Future Work

Conclusions –

- **Addressing Zero-Inflation:** ZIMMA's ability to detect structural zeros avoids the bias introduced by pseudo counts, a common strategy in dealing with zero-inflated count data.
- **Precise Interpretation:** By decomposing the indirect effect into abundance and prevalence pathways, ZIMMA provides a more precise interpretation of active microbiome mediators.
- **High Power:** ZIMMA demonstrates superior statistical power compared to existing methods.

Future Works–

- **More Applications** To demonstrate the usage of ZIMMA.
- **Sensitivity Analysis** To what extent does violating assumptions affect the magnitude of bias?
- **Microbiome Correlation**

Thank You


Access to
Excellence
SCHOOL OF MEDICINE
1843
CASE WESTERN RESERVE UNIVERSITY
The School of Medicine
THANK YOU !

Albert, J. M. (2008).
Mediation analysis via potential outcomes models.
Statistics in medicine, 27(8):1282–1304.

Baron, R. and Kenny, D. (1986).
The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.
Journal of Personality and Social Psychology, 51:1173–1182.

MacKinnon, D. P. and Dwyer, J. H. (1993).
Estimating mediated effects in prevention studies.
Evaluation review, 17(2):144–158.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002).
A comparison of methods to test mediation and other intervening variable effects.
Psychological methods, 7(1):83.

Pillow, J. and Scott, J. (2012).
Fully bayesian inference for neural models with negative-binomial spiking.

In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Robins, J. M. and Greenland, S. (1992).
Identifiability and exchangeability for direct and indirect effects.
Epidemiology, pages 143–155.

Sohn, M. B. and Li, H. (2019).
Compositional mediation analysis for microbiome studies.
The Annals of Applied Statistics, 13(1):661 – 681.

Venables, W. N. and Ripley, B. D. (2002).
Modern Applied Statistics with S.
Springer, New York, fourth edition.
ISBN 0-387-95457-0.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011).
Linking long-term dietary patterns with gut microbial enterotypes.
Science, 334(6052):105–108.

Yue, Y. and Hu, Y.-J. (2022).
A new approach to testing mediation of the microbiome at both the community and individual taxon levels.
Bioinformatics, 38(12):3173–3180.

Zhang, H., Chen, J., Feng, Y., Wang, C., Li, H., and Liu, L. (2021).
Mediation effect selection in high-dimensional and compositional microbiome data.
Statistics in Medicine, 40(4):885–896.