Zero-Inflated Generalized Dirichlet Multinomial (ZIGDM) Regression Model for Microbiome Compositional Data

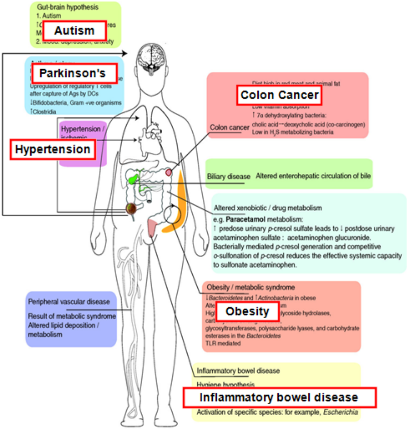
ZhengZheng Tang

Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

tang@biostat.wisc.edu

Human Microbiome Research

- Use high-throughput sequencing to quantify abundances of microbial taxa
- Link the abundance to human diseases and traits
- Proper modeling of microbial abundance is essential to the power of detecting this association



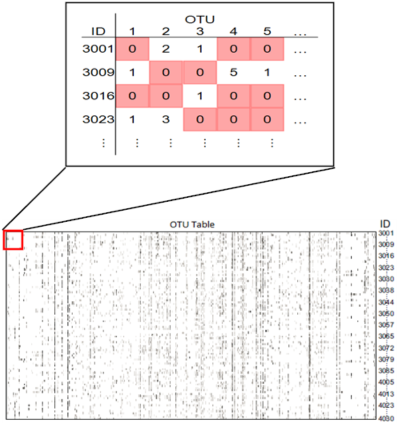
Gut-brain hypothesis
1. Autism
Autism
2. Mood: depression, anxiety
Parkinson's
Upregulation of regulatory T cells
after capture of Ags by DCs
↓Bifidobacteria, Gram +ve organisms
↑Clostridia
Hypertension / ischemic
Hypertension
Diet high in red meat and animal fat
Colon Cancer
Low vitamin absorption
↑ 7α dehydroxylating bacteria:
cholic acid→deoxycholic acid (co-carcinogen)
Low in H2S metabolizing bacteria
Colon cancer
Biliary disease
Altered enterohepatic circulation of bile
Altered xenobiotic / drug metabolism
e.g. Paracetamol metabolism:
↑ predose urinary p-cresol sulfate leads to ↓ postdose urinary acetaminophen sulfate : acetaminophen glucuronide.
Bacterially mediated p-cresol generation and competitive o-sulfonation of p-cresol reduces the effective systemic capacity to sulfonate acetaminophen.
Obesity / metabolic syndrome
↓Bacteroidetes and ↑Actinobacteria in obese
Obesity
glycoside hydrolases,
glycosyltransferases, polysaccharide lyases, and carbohydrate esterases in the Bacteroidetes
TLR mediated
Peripheral vascular disease
Result of metabolic syndrome
Altered lipid deposition / metabolism
Inflammatory bowel disease
Hygiene hypothesis
Inflammatory bowel disease
Activation of specific species: for example, Escherichia

Kinross et al., Genome Medicine 2011, 3:14

Microbiome Data

Data Characteristics

- Compositional
- Zero-inflated
- Over-dispersed
- Complex correlation structure

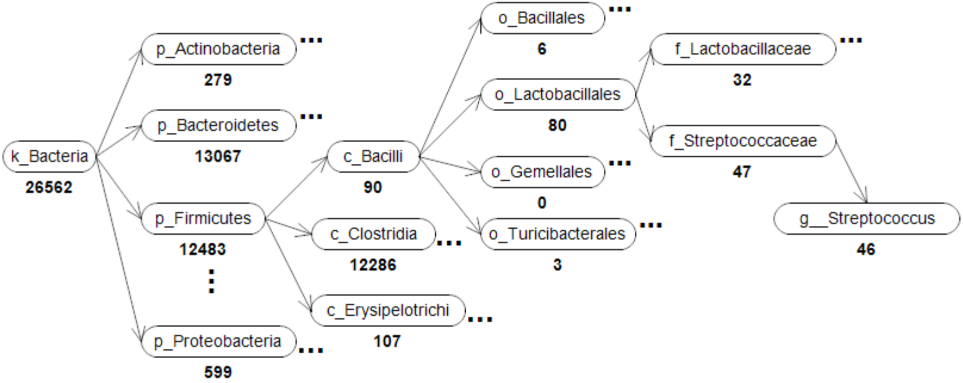


			OTU			
ID	1	2	3	4	5	...
3001	0	2	1	0	0	...
3009	1	0	0	5	1	...
3016	0	0	1	0	0	...
3023	1	3	0	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	

OTU Table
ID
3001
3009
3016
3023
3030
3038
3044
3050
3057
3065
3072
3079
3085
4005
4013
4023
4030

Count Data on a Taxonomic Tree

Counts for species can be summarized into higher taxonomic levels



Outline

- Probability distributions for microbial compositions
  - Dirichlet Multinomial (DM)
  - *Generalized* Dirichlet Multinomial (GDM)
  - *Zero-Inflated Generalized* Dirichlet Multinomial (ZIGDM)
- ZIGDM regression model
  - Differential mean and dispersion tests

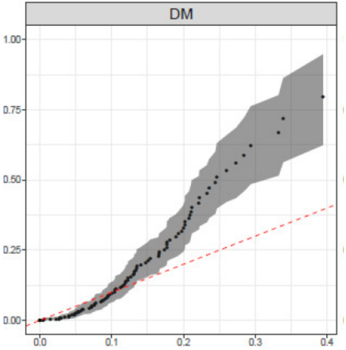
Dirichlet Multinomial (DM)
Dirichlet Prior for Multinomial

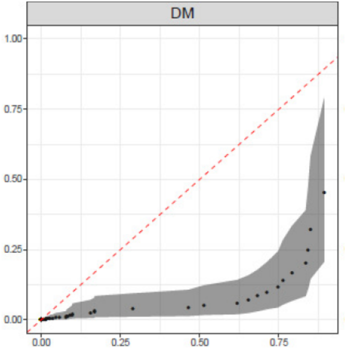
$K+1$: Number of taxa in the composition
$\mathbf{Y} = (Y_1, \ldots, Y_{(K+1)})$ with $N = \sum_{j=1}^{K+1} Y_j$
$\mathbf{P} = (P_1, \ldots, P_{(K+1)})$ with $\sum_{j=1}^{K+1} P_j = 1$

$$\mathbf{Y} \mid \mathbf{P} \sim \mathrm{Multinomial}(\mathbf{P}, N),$$
$$\mathbf{P} \sim \mathrm{Dirichlet}(\boldsymbol{\nu}, \theta)$$

Number of parameters: $K+1$

DM is Not Ideal





- Negative correlations
- Restrictive mean-variance relationships
- Limited ability to handle excessive zeros

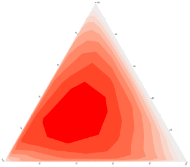
Generalized Dirichlet Multinomial (GDM)
Generalized Dirichlet (GD) Prior for Multinomial

Generalized Dirichlet *(Connor and Mosimann, JASA 1969)*

$$\mathbf{Y} \mid \mathbf{P} \sim \mathrm{Multinomial}(\mathbf{P}, N),$$
$$\mathbf{P} \sim \mathrm{GD}(\mathbf{a}, \mathbf{b}), \quad \mathbf{a} = (a_1, \ldots, a_K), \mathbf{b} = (b_1, \ldots, b_K)$$

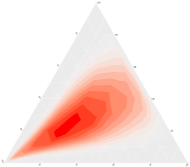
Number of parameters: $2K$

GD reduces to Dirichlet if $b_j = a_{j+1} + b_{j+1}$, $j = 1, \ldots, K-1$.



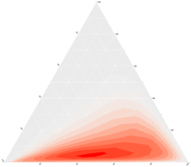
$\mathbf{a} = (0.5, 0.25)$
$\mathbf{b} = (0.5, 0.25)$

$$\begin{bmatrix} 1 & -0.59 & -0.55 \\ & 1 & -0.35 \\ & & 1 \end{bmatrix}$$



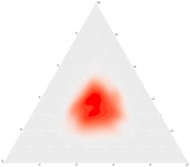
$\mathbf{a} = (0.5, 5)$
$\mathbf{b} = (0.5, 5)$

$$\begin{bmatrix} 1 & -0.89 & -0.88 \\ & 1 & 0.58 \\ & & 1 \end{bmatrix}$$



$\mathbf{a} = (0.5, 0.5)$
$\mathbf{b} = (0.6, 5)$

$$\begin{bmatrix} 1 & -0.39 & -0.97 \\ & 1 & 0.16 \\ & & 1 \end{bmatrix}$$



$\mathbf{a} = (5, 5)$
$\mathbf{b} = (10, 5)$

$$\begin{bmatrix} 1 & -0.54 & -0.50 \\ & 1 & -0.46 \\ & & 1 \end{bmatrix}$$

Advantages of GD Prior

- Comparing with Dirichlet
  Provide more general correlation structure

- Comparing with Logistic Normal
  Conjugate prior for multinomial *(Wong, Appl Math Comput. 1998)*

  $$\begin{aligned}
  &\mathbf{Y}|\mathbf{P} \sim \mathrm{Multinomial}(\mathbf{P}, N)\\
  &\mathbf{P} \sim \mathrm{GD}(\mathbf{a}, \mathbf{b})\\
  &\Longrightarrow \mathbf{P}|\mathbf{Y} \sim \mathrm{GD}(\mathbf{a}^*, \mathbf{b}^*),\\
  &\mathbf{a}^* = (a_1^*, \ldots, a_K^*),\ a_j^* = a_j + Y_j \text{ and}\\
  &\mathbf{b}^* = (b_1^*, \ldots, b_K^*),\ b_j^* = b_j + Y_{j+1} + \ldots + Y_{K+1},\\
  &j = 1, \ldots, K
  \end{aligned}$$

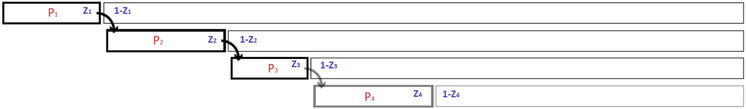
Can GD handle excessive zeros?

Construct GD from Independent Beta Variables

$$Z_j \sim \mathrm{Beta}(a_j,\ b_j),\ j = 1, \ldots, K$$

$$P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i) \ \Bigg\Updownarrow\ Z_j = P_j/(1 - \sum_{i=1}^{j-1} P_i)$$

$$\mathbf{P} = (P_1, \ldots, P_K) \sim \mathrm{GD}(\mathbf{a}, \mathbf{b})$$



“stick breaking process”

Doesn’t Permit Taxa Absence (Structural Zero)

Zero-Inflated Generalized **Dirichlet (ZIGD)**

$$Z_j \sim \begin{cases} 0 & \text{with probability } \pi_j, \\ \text{Beta}(a_j, b_j) & \text{with probability } 1 - \pi_j, \end{cases}$$

$$P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i) \Big\Updownarrow Z_j = P_j/(1 - \sum_{i=1}^{j-1} P_i)$$

$$\mathbf{P} = (P_1, \ldots, P_K) \sim \mathrm{ZIGD}(\boldsymbol{\pi}, \mathbf{a}, \mathbf{b}),\ \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$$

ZIGD is a Conjugate Prior to Multinomial

Notation:
$\Delta_j = I(P_j = 0) = I(Z_j = 0)$
$\mathcal{U}$: index set for taxa present in the sample ($\mathbf{\Delta}_{\mathcal{U}} = \mathbf{0}$, $\mathbf{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1}$)
$\overline{\mathcal{U}}$: index set for the structural zeros
$\mathcal{V}$: index set for taxa with zero counts ($\mathbf{Y}_{\mathcal{V}} = \mathbf{0}$, $\mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0}$)
Sets $\mathcal{U}$ and $\mathcal{V}$ are not exclusive: their intersection $\mathcal{U} \cap \mathcal{V}$ indexed taxa that are present in the sample but have zero counts due to the undersampling in the sequencing experiment (i.e. sampling zeros).

$$f(\mathbf{P} \mid \mathbf{Y}) = f(\mathbf{P} \mid \mathbf{\Delta}, \mathbf{Y}) Pr(\mathbf{\Delta} \mid \mathbf{Y})$$

- $f(\mathbf{P} \mid \mathbf{\Delta}, \mathbf{Y}) = I(\mathbf{P}_{\overline{\mathcal{U}}} = \mathbf{0}) f(\mathbf{P}_{\mathcal{U}} \mid \mathbf{\Delta}_{\mathcal{U}} = \mathbf{0}, \mathbf{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1}, \mathbf{Y})$

$$\mathbf{P}_{\mathcal{U}} \mid (\mathbf{\Delta}_{\mathcal{U}} = \mathbf{0}, \mathbf{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1}) \sim GD(\mathbf{a}_{\mathcal{U}}, \mathbf{b}_{\mathcal{U}})$$
$$\mathbf{P}_{\mathcal{U}} \mid (\mathbf{\Delta}_{\mathcal{U}} = \mathbf{0}, \mathbf{\Delta}_{\overline{\mathcal{U}}} = \mathbf{1}, \mathbf{Y}) \sim GD(\mathbf{a}^*_{\mathcal{U}}, \mathbf{b}^*_{\mathcal{U}})$$

ZIGD is a Conjugate Prior to Multinomial

$\mathcal{U}$: index set for the taxa present in the sample
$\mathcal{V}$: index set for the taxa with zero counts

- $Pr(\mathbf{\Delta} \mid \mathbf{Y}) = I(\mathbf{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) Pr(\mathbf{\Delta}_{\mathcal{V}} \mid \mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0})$

$$\begin{aligned} & Pr(\mathbf{\Delta}_{\mathcal{V}} \mid \mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0}) \\ \propto \quad & \prod_{j \in \mathcal{V}} \left\{ \pi_j^{\Delta_j} \left[(1 - \pi_j) \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)} \right]^{(1-\Delta_j)} \right\}, \end{aligned}$$

For the taxon j with zero count, the probability of this observed zero being structural zero is $\frac{\pi_j}{\pi_j + (1-\pi_j) \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)}}$.

ZIGDM Regression Model

Use ZIGD as a prior for multinomial $\rightarrow$ ZIGDM

- ZIGDM regression model can link mean, dispersion, presence-absence frequency of the microbial abundance to the covariates of interest
- An efficient EM for fitting the model and estimating parameters

ZIGDM Regression Model

n subjects measured on $K+1$ taxa
$i = 1, \ldots, n; j = 1, \ldots, K+1$
Y_{ij}: observed taxon count
P_{ij}: underlying true proportion
$\mathbf{X}_i$: d-dimensional vector including intercept and covariates

$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK}) \sim \mathrm{ZIGDM}(\boldsymbol{\pi}_i, \mathbf{a}_i, \mathbf{b}_i)$, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, $\mathbf{a}_i = (a_{i1}, \ldots, a_{iK})$, and $\mathbf{b}_i = (b_{i1}, \ldots, b_{iK})$. We model $\mu_{ij} = a_{ij}/(a_{ij} + b_{ij})$ and $\sigma_{ij} = 1/(1 + a_{ij} + b_{ij})$ as they are Beta mean and dispersion parameters.

$$\mu_{ij} = \frac{e^{\boldsymbol{\alpha}_j^{\mathrm{T}} \mathbf{X}_i}}{1 + e^{\boldsymbol{\alpha}_j^{\mathrm{T}} \mathbf{X}_i}}, \quad \sigma_{ij} = \frac{e^{\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{X}_i}}{1 + e^{\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{X}_i}}, \quad \text{and} \quad \pi_{ij} = \frac{e^{\boldsymbol{\gamma}_j^{\mathrm{T}} \mathbf{X}_i}}{1 + e^{\boldsymbol{\gamma}_j^{\mathrm{T}} \mathbf{X}_i}},$$

where $\boldsymbol{\alpha}_j = (\alpha_{1j}, \ldots, \alpha_{dj})$, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{dj})$, and $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{dj})$ are regression coefficients for taxon j.

Equivalent Hierarchical Model

$$\Delta_{ij} \sim \mathrm{Bernoulli}(\pi_{ij}), \quad j = 1, \ldots, K,$$
$$Z_{ij} = 0 \text{ if } \Delta_{ij} = 1, \quad Z_{ij} \mid \Delta_{ij} = 0 \sim \mathrm{Beta}(a_{ij}, b_{ij}), \quad j = 1, \ldots, K,$$
$$P_{i1} = Z_{i1}, \quad P_{ij} = Z_{ij} \textstyle\prod_{k=1}^{j-1} (1 - Z_{ik}), \quad j = 2, \ldots, K,$$
$$\mathbf{Y}_i \mid \mathbf{P}_i \sim \mathrm{Multinomial}(\mathbf{P}_i, N_i),$$
$$\text{where } \mathbf{P}_i = (P_{i1}, \ldots, P_{iK}) \text{ and } N_i = \textstyle\sum_{j=1}^{K+1} Y_{ij}.$$

EM algorithm

Complete set of parameters: $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$
Complete data log-likelihood expressed in terms of Z's:

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \log \left[\prod_{i=1}^{n} \left\{ Pr(\mathbf{Y}_i \mid \mathbf{Z}_i) \prod_{j=1}^{K} f(Z_{ij}) \right\} \right] \\
&= \sum_{i=1}^{n} \log \left\{ Pr(\mathbf{Y}_i \mid \mathbf{Z}_i) \right\} \\
&\quad + \sum_{j=1}^{K} \sum_{i=1}^{n} \Big\{ \Delta_{ij} \log \pi_{ij} + (1 - \Delta_{ij}) \log(1 - \pi_{ij}) + \\
&\quad (1 - \Delta_{ij}) \left[- \log(\mathcal{B}(a_{ij}, b_{ij})) + (a_{ij} - 1) \log(Z_{ij}) + (b_{ij} - 1) \log(1 - Z_{ij}) \right] \Big\},
\end{aligned}
$$

where $a_{ij} = \mu_{ij}(1/\sigma_{ij} - 1)$ and $b_{ij} = (1 - \mu_{ij})(1/\sigma_{ij} - 1)$.

Using Z's instead of P's allows us to derive the explicit form of posterior expectations in the E-step and estimate parameters for each taxon independently in the M-step.

EM algorithm – E Step

In the t-th E-step, we need to compute the expected complete data log-likelihood,

$$Q_\theta^* = \sum_{j=1}^{K} \sum_{i=1}^{n} \mathsf{E}\Big\{ \Delta_{ij} \log \pi_{ij} + (1 - \Delta_{ij}) \log(1 - \pi_{ij}) +$$

$$(1 - \Delta_{ij}) \left[-\log(\mathcal{B}(a_{ij}, b_{ij})) + (a_{ij} - 1) \log Z_{ij} + (b_{ij} - 1) \log(1 - Z_{ij}) \right]$$

where the expectation is with respect to the posterior distributions of $(\mathbf{\Delta}_i \mid \mathbf{Y}_i; \boldsymbol{\theta}^{(t-1)})$ and $(\mathbf{Z}_i \mid \mathbf{\Delta}_i, \mathbf{Y}_i; \boldsymbol{\theta}^{(t-1)})$ with $\boldsymbol{\theta}^{(t-1)}$ being the parameter estimates in the $(t-1)$-th M-step.

EM algorithm – E Step

Based on the results for ZIGD posterior distribution, we can derive the explicit form for the posterior means:

$$\Delta_{ij}^{*} = \mathrm{E}\left(\Delta_{ij} \mid \mathbf{Y}_i\right) = \begin{cases} 0 & \text{if } Y_{ij} > 0 \\ \frac{\pi_{ij}}{\pi_{ij} + \left(1 - \pi_{ij}\right)\frac{\mathcal{B}\left(a_{ij}^{*}, b_{ij}^{*}\right)}{\mathcal{B}\left(a_{ij}, b_{ij}\right)}} & \text{if } Y_{ij} = 0 \end{cases},$$

$$A_{ij}^{*} = \mathrm{E}\left(\log Z_{ij} \mid \mathbf{Y}_i, \Delta_{ij} = 0\right) = \psi\left(a_{ij}^{*}\right) - \psi\left(a_{ij}^{*} + b_{ij}^{*}\right),$$

$$B_{ij}^{*} = \mathrm{E}\left(\log(1 - Z_{ij}) \mid \mathbf{Y}_i, \Delta_{ij} = 0\right) = \psi\left(b_{ij}^{*}\right) - \psi\left(a_{ij}^{*} + b_{ij}^{*}\right),$$

where $a_{ij}^{*} = a_{ij} + Y_{ij}$, $b_{ij}^{*} = b_{ij} + Y_{i(j+1)} + \ldots + Y_{i(K+1)}$, and $\psi(\cdot)$ is the digamma function.

EM algorithm – M Step

Thus, Q^*_θ can be rewritten as

$$Q^*_\theta = \sum_{j=1}^{K} Q^*_{\gamma_j} + \sum_{j=1}^{K} Q^*_{\alpha_j,\beta_j}, \tag{3}$$

where $Q^*_{\gamma_j} = \sum_{i=1}^{n}\{\Delta^*_{ij}\log\pi_{ij} + (1-\Delta^*_{ij})\log(1-\pi_{ij})\}$ and $Q^*_{\alpha_j,\beta_j} = \sum_{i=1}^{n}(1-\Delta^*_{ij})\{-\log(\mathcal{B}(a_{ij},b_{ij})) + (a_{ij}-1)A^*_{ij} + (b_{ij}-1)B^*_{ij}\}$.

In the t-th M-step, for each taxon j, we obtain $\boldsymbol{\gamma}_j^{(t)}$ from maximizing the function $Q^*_{\gamma_j}$ and obtain $\boldsymbol{\alpha}_j^{(t)}$ and $\boldsymbol{\beta}_j^{(t)}$ from maximizing the function $Q^*_{\alpha_j,\beta_j}$.

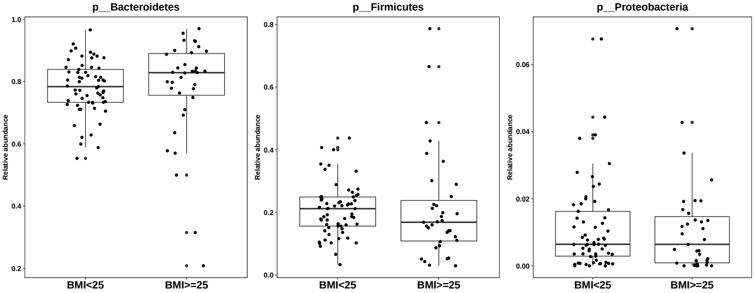
Association Tests

Test for the mean: $H_0 : \boldsymbol{\alpha}_{*1} = \ldots = \boldsymbol{\alpha}_{*K} = 0$
Test for the dispersion: $H_0 : \boldsymbol{\beta}_{*1} = \ldots = \boldsymbol{\beta}_{*K} = 0$
Test for the presence-absence frequency: $H_0 : \boldsymbol{\gamma}_{*1} = \ldots = \boldsymbol{\gamma}_{*K} = 0$

- When performing a test on a particular set of parameters (e.g. α's), we include only the intercept coefficient in the model for the other sets of parameters (e.g. β's and γ's)
- We adopted score statistics, which are computationally faster and more stable than Wald and LR statistics *(Lin and Tang, AJHG 2011)*
- The asymptotic approximation of the test statistics may not be accurate when most of the observations are zero, especially when the sample size is small. Therefore, we need to use permutation techniques to obtain p-values.

Why we care about differential dispersion?

Microbiome compositions are very dynamic
Disease-microbe association can be moderated by many factors, resulting in heterogeneous dispersion levels between disease and healthy groups



Simulation Study

Methods: Differential-Mean, Differential-Dispersion

ZIGDM-based tests: ZIGDM_1 and ZIGDM_2

GDM-based tests: GDM_1 and GDM_2

DM-based tests: DM_1 and DM_2 *(La Rosa et al., PLOS ONE 2012)*

Non-parametric tests: QCAT_1 and QCAT_2 *(Tang et al., Bioinformatics 2017)*

Setup:

- Simulate 6 taxon counts for two groups with same sample sizes and tested differential abundance in the 6 taxa between the two groups.
- Sample sizes of 100 and 200 in all simulation studies
- In the power evaluation, we change either the mean abundance or the dispersion level in one group.
- 5000 simulated data sets to evaluate type I error and power of the tests at the 0.05 significance level.

Simulation Study

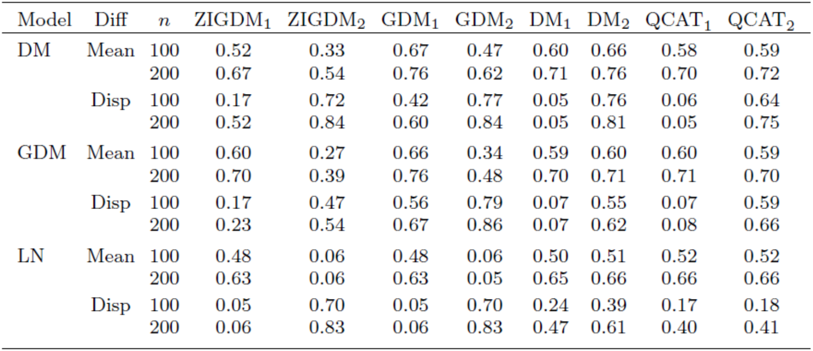
$$Y \sim \mathrm{Multinomial}(\mathbf{P}, N);\ N \sim \mathrm{Poisson}(1000)$$

model for proportion **P**	data generation	parameter specification
Dirichlet	$\mathbf{P} = \{P_j\}_{j=1}^{6}$ $\sim Dir(\boldsymbol{\mu}, \sigma)$ $\boldsymbol{\mu} = \{\mu_j\}_{j=1}^{6}$	mean of Dirichlet: $\mu_j = 1/6$ dispersion of Dirichlet: $\sigma = 0.3$
GD ZIGD	$Z_j \sim (ZI)B(\pi_j, \mu_j, \sigma_j)$ $P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i)$ $(j = 1, \ldots, 5)$ $P_6 = 1 - \sum_{i=1}^{5} P_i$	mean of Beta: $\mu_j = 0.2$ dispersion of Beta: $\sigma_j = 0.2$ zero-inflation: $\{\pi_j\}_{j=1}^{5} = \{0.1, 0.2, 0.4, 0.6, 0.8\}$
LN ZILN	$\{W_j\}_{j=1}^{5} \sim LN(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Omega})$ $P_j = W_j / (\sum_{i=1}^{5} W_i + 1)$ $(j = 1, \ldots, 5)$ replace counts with zero $P_6 = 1 - \sum_{i=1}^{5} P_i$	mean of Normal: $\mu_j = 0$ variance of Normal: $\sigma_j = 1$ correlation: $\boldsymbol{\Omega}_{jj'} = 0.5^{\lvert j-j' \rvert}$ zero-inflation: $\{\pi_j\}_{j=1}^{5} = \{0.1, 0.2, 0.4, 0.6, 0.8\}$

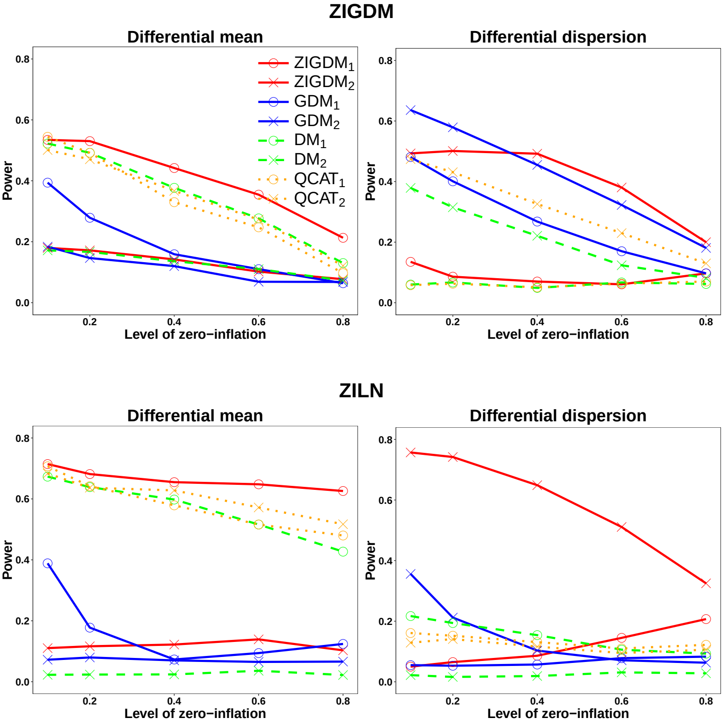
Power Simulation Setup

parameter specification	perturbation
mean of Dirichlet: $\mu_j = 1/6$	$\mu_k \sim \mathit{Unif}(0, 0.5)$
dispersion of Dirichlet: $\sigma = 0.3$	OR $\sigma \sim \mathit{Unif}(0, 0.5)$
mean of Beta: $\mu_j = 0.2$	$\mu_k \sim \mathit{Unif}(0, 0.5)$
dispersion of Beta: $\sigma_j = 0.2$	OR $\sigma_k \sim \mathit{Unif}(0, 1)$
zero-inflation:	
$\{\pi_j\}_{j=1}^5 = \{0.1, 0.2, 0.4, 0.6, 0.8\}$	
mean of Normal: $\mu_j = 0$	For LN: $\mu_k \sim \mathit{Unif}(0, 1)$
variance of Normal: $\sigma_j = 1$	OR $\sigma_k \sim \mathit{Unif}(1, 6)$
correlation: $\mathbf{\Omega}_{jj'} = 0.5^{\lvert j-j' \rvert}$	
zero-inflation:	For ZILN: $\mu_k \sim \mathit{Unif}(0, 2)$
$\{\pi_j\}_{j=1}^5 = \{0.1, 0.2, 0.4, 0.6, 0.8\}$	OR $\sigma_k \sim \mathit{Unif}(1, 8)$

Power under Non-zero-inflated Models



Model	Diff	n	ZIGDM_1	ZIGDM_2	GDM_1	GDM_2	DM_1	DM_2	QCAT_1	QCAT_2
DM	Mean	100	0.52	0.33	0.67	0.47	0.60	0.66	0.58	0.59
		200	0.67	0.54	0.76	0.62	0.71	0.76	0.70	0.72
	Disp	100	0.17	0.72	0.42	0.77	0.05	0.76	0.06	0.64
		200	0.52	0.84	0.60	0.84	0.05	0.81	0.05	0.75
GDM	Mean	100	0.60	0.27	0.66	0.34	0.59	0.60	0.60	0.59
		200	0.70	0.39	0.76	0.48	0.70	0.71	0.71	0.70
	Disp	100	0.17	0.47	0.56	0.79	0.07	0.55	0.07	0.59
		200	0.23	0.54	0.67	0.86	0.07	0.62	0.08	0.66
LN	Mean	100	0.48	0.06	0.48	0.06	0.50	0.51	0.52	0.52
		200	0.63	0.06	0.63	0.05	0.65	0.66	0.66	0.66
	Disp	100	0.05	0.70	0.05	0.70	0.24	0.39	0.17	0.18
		200	0.06	0.83	0.06	0.83	0.47	0.61	0.40	0.41


ZIGDM
Differential mean
Differential dispersion
ZIGDM$_1$
ZIGDM$_2$
GDM$_1$
GDM$_2$
DM$_1$
DM$_2$
QCAT$_1$
QCAT$_2$
Power
0.0
0.2
0.4
0.6
0.8
Level of zero−inflation
ZILN
Differential mean
Differential dispersion

Take Home Message

- The ZIGDM tests are more powerful to detect differential mean/dispersion and are more robust to the underlying distribution if the taxon counts are zero-inflated
- If the taxon counts are not zero-inflated, the GDM tests are more desirable
- The DM tests yield similar power to the GDM test even if data are DM distributed and the DM differential-dispersion test has substantial power loss if data are not DM distributed
- The QCAT tests have robust and decent power in detecting differential mean but cannot powerfully detect differential dispersion

Gut Microbiome and Body Mass Index

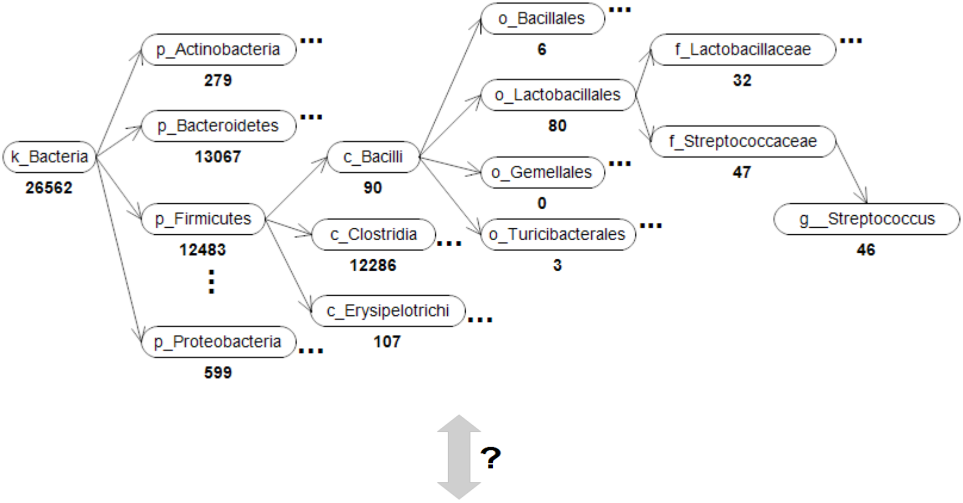
Wu, Gary D., et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 334.6052 (2011): 105-108.

- Gut microbiota play an important role in obesity
- Fecal samples were collected from 98 healthy volunteers, along with their demographic data and diet information
- Sample DNA was analyzed by sequencing the V1-V2 region of the 16S rRNA gene
- The sequencing reads were taxonomically classified to the 80 genera, and then mapped to a taxonomic tree with 74 lineages from family to kingdom

Identify the microbial lineages have differential mean or dispersion between high and normal BMI groups

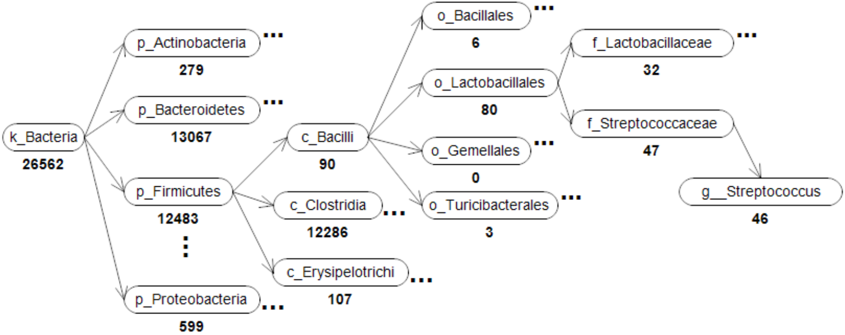
Count Data on a Taxonomic Tree

Counts for species can be summarized into higher taxonomic levels

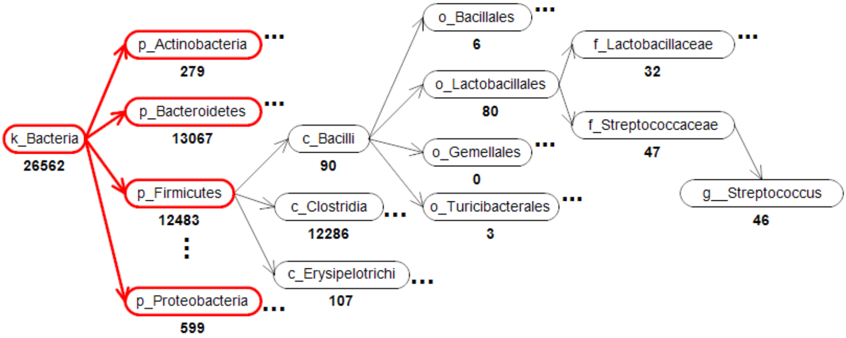


Covariates of interest (e.g. disease status)

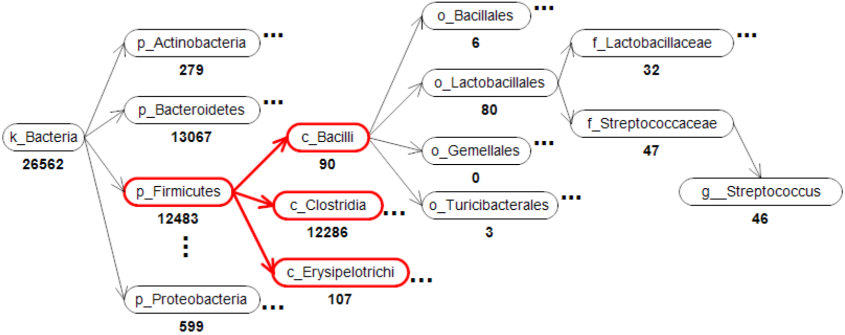
Apply Tests to Lineages (Subcompositions)



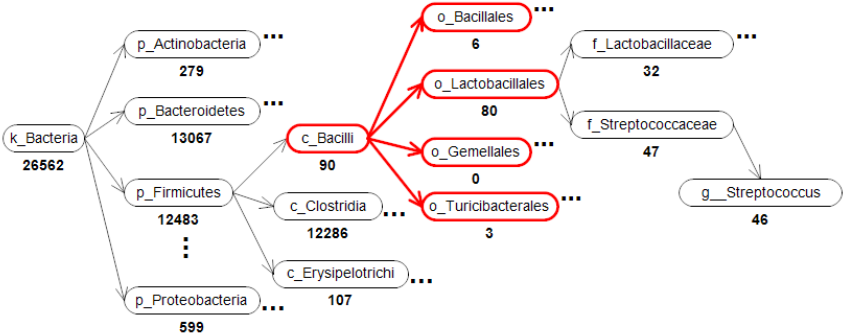
Apply Tests to Lineages (Subcompositions)



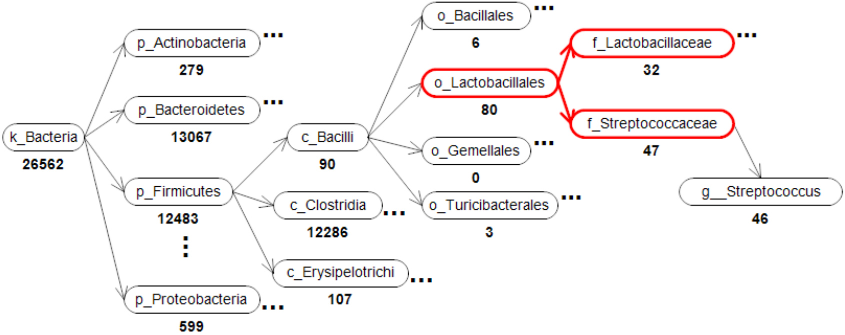
Apply Tests to Lineages (Subcompositions)



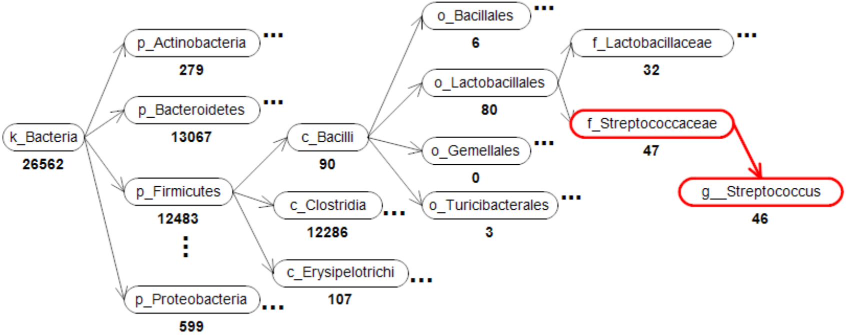
Apply Tests to Lineages (Subcompositions)



Apply Tests to Lineages (Subcompositions)



Apply Tests to Lineages (Subcompositions)



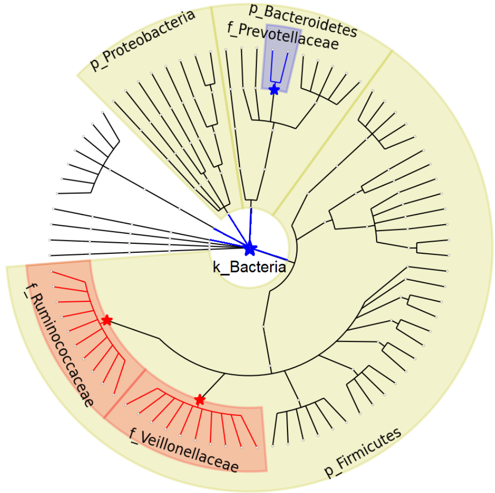
BMI-Associated Lineages

Use Benjamini-Hochberg procedure to control FDR at 0.05 level

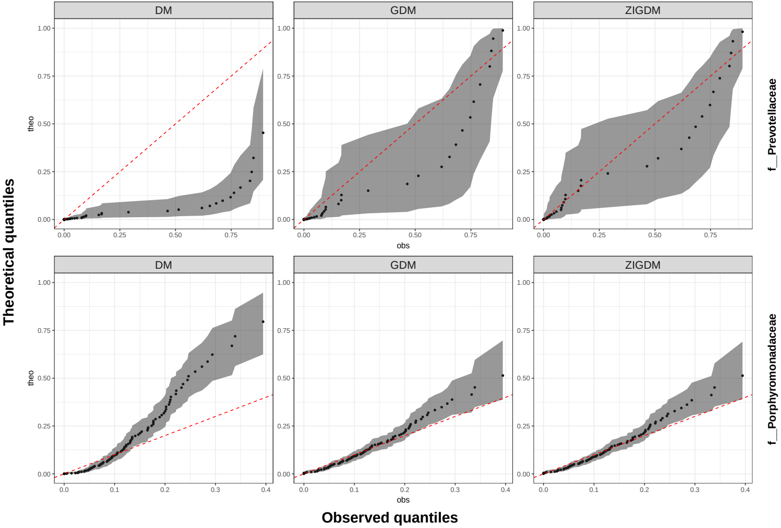
Results from DM and QCAT tests:
Family *Ruminococcaceae*
(DM_1 p-value $=$ 0.0013 and QCAT_1 p-value $=$ 0.00014)
Family *Veillonellaceae*
(DM_2 p-value $=$ 0.0012 and QCAT_2 p-value $=$ 0.00080)

Results from GDM and ZIGDM tests:
Family *Ruminococcaceae*
Family *Prevotellaceae*
(ZIGDM_2 p-value $=$ 0.0014 and GDM_2 p-value $=$ 0.0014)
Kingdom *Bacteria*
(ZIGDM_2 p-value $=$ 0.0016)

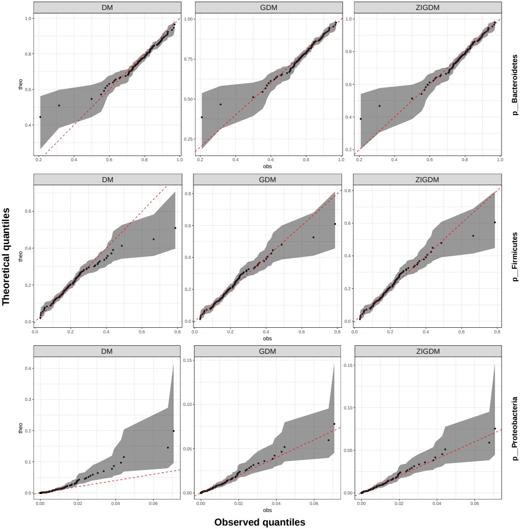
Differential Lineages



QQ plots for the two families under order Bacteroidales



QQ plots for the three most abundant phyla



Choose between ZIGDM and GDM based on AIC/BIC or LRT

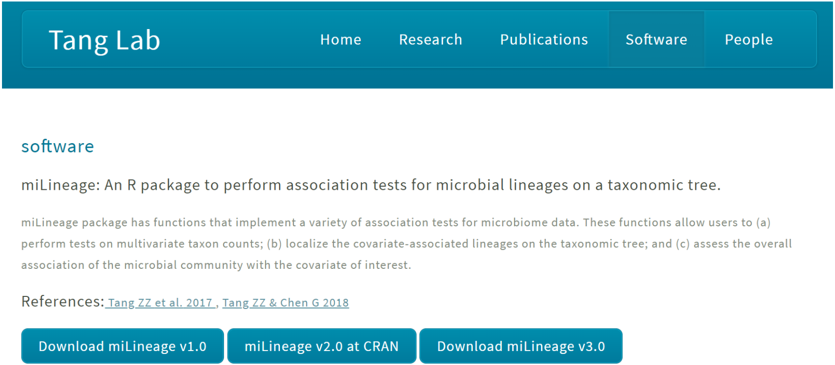
Summary

- The ZIGDM provides better fit to the microbiome data than DM
- The ZIGDM provides a more flexible way of accommodating excessive zeros and disentangle structural zeros and sampling zeros.
- Propose score tests based on the ZIGDM regression model to detect differential mean or dispersion level of microbial composition
- Develop an efficient EM algorithm to estimate parameters in the ZIGDM regression.

Tang, Zheng-Zheng, and Chen, Guanhua. "Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis." Biostatistics, kxy025, (2018).

Software

https://tangzheng1.github.io/tanglab/software.html



software

miLineage: An R package to perform association tests for microbial lineages on a taxonomic tree.

miLineage package has functions that implement a variety of association tests for microbiome data. These functions allow users to (a) perform tests on multivariate taxon counts; (b) localize the covariate-associated lineages on the taxonomic tree; and (c) assess the overall association of the microbial community with the covariate of interest.

References: Tang ZZ et al. 2017 , Tang ZZ & Chen G 2018

Acknowledgements

- Dr. Guanhua Chen, University of Wisconsin-Madison
- Dr. Hongzhe Li, University of Pennsylvania

Thank you!