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

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

Data Characteristics

- Compositional
- Zero-inflated
- Over-dispersed
- Complex correlation structure
Counts for species can be summarized into higher taxonomic levels
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  
\( \sum_{j=1}^{K+1} P_j = 1 \)

\[ \begin{align*} 
\mathbf{Y} \mid \mathbf{P} & \sim \text{Multinomial}(\mathbf{P}, N), \\
\mathbf{P} & \sim \text{Dirichlet}(\nu, \theta) 
\end{align*} \]

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*)

\[
\begin{align*}
Y \mid P & \sim \text{Multinomial}(P, N), \\
P & \sim \text{GD}(a, b), \quad a = (a_1, \ldots, a_K), \quad b = (b_1, \ldots, b_K)
\end{align*}
\]

Number of parameters: 2\(K\)

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

\[ Y|P \sim \text{Multinomial}(P, N) \]
\[ P \sim \text{GD}(a, b) \]
\[ \implies P|Y \sim \text{GD}(a^*, b^*), \]
\[ a^* = (a_1^*, \ldots, a_K^*), \quad a_j^* = a_j + Y_j \quad \text{and} \]
\[ b^* = (b_1^*, \ldots, b_K^*), \quad b_j^* = b_j + Y_{j+1} + \ldots + Y_{K+1}, \]
\[ j = 1, \ldots, K \]

Can GD handle excessive zeros?
Construct GD from Independent Beta Variables

\[ Z_j \sim \text{Beta}(a_j, b_j), \ j = 1, \ldots, K \]

\[ P_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i) \quad Z_j = P_j / (1 - \sum_{i=1}^{j-1} P_i) \]

\[ \mathbf{P} = (P_1, \ldots, P_K) \sim \text{GD}(a, b) \]

“stick breaking process”

Doesn’t Permit Taxa Absence (Structural Zero)
Zero-Inflated Generalized Dirichlet (ZIGD)

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

\[ P_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i) \quad \Leftrightarrow \quad Z_j = P_j / (1 - \sum_{i=1}^{j-1} P_i) \]

\[ \mathbf{P} = (P_1, \ldots, P_K) \sim \text{ZIGD}(\pi, \mathbf{a}, \mathbf{b}), \quad \pi = (\pi_1, \ldots, \pi_K) \]
**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 (\( \Delta_{\mathcal{U}} = 0 \), \( \Delta_{\overline{\mathcal{U}}} = 1 \))

\( \overline{\mathcal{U}} \): index set for the structural zeros

\( \mathcal{V} \): index set for taxa with zero counts (\( Y_{\mathcal{V}} = 0 \), \( Y_{\overline{\mathcal{V}}} > 0 \))

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).

\[
f(P \mid Y) = f(P \mid \Delta, Y)Pr(\Delta \mid Y)
\]

\[
\Rightarrow f(P \mid \Delta, Y) = I(P_{\overline{\mathcal{U}}} = 0)f(P_{\mathcal{U}} \mid \Delta_{\mathcal{U}} = 0, \Delta_{\overline{\mathcal{U}}} = 1, Y)
\]

\[
P_{\mathcal{U}} \mid (\Delta_{\mathcal{U}} = 0, \Delta_{\overline{\mathcal{U}}} = 1) \sim GD(a_{\mathcal{U}}, b_{\mathcal{U}})
\]

\[
P_{\mathcal{U}} \mid (\Delta_{\mathcal{U}} = 0, \Delta_{\overline{\mathcal{U}}} = 1, Y) \sim GD(a^*_{\mathcal{U}}, b^*_{\mathcal{U}})
\]
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

\[ Pr(\Delta \mid Y) = I(\Delta_{\mathcal{V}} = 0)Pr(\Delta_{\mathcal{V}} \mid Y_{\mathcal{V}} = 0, Y_{\mathcal{V}} > 0) \]

\[ Pr(\Delta_{\mathcal{V}} \mid Y_{\mathcal{V}} = 0, Y_{\mathcal{V}} > 0) \propto \prod_{j \in \mathcal{V}} \left\{ \pi_j^{\Delta_j} \left[ (1 - \pi_j) \frac{B(a_j^*, b_j^*)}{B(a_j, b_j)} \right]^{(1-\Delta_j)} \right\} , \]

For the taxon \( j \) with zero count, the probability of this observed zero being structural zero is

\[ \frac{\pi_j}{\pi_j + (1 - \pi_j) \frac{B(a_j^*, b_j^*)}{B(a_j, b_j)}} . \]
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
$X_i$: $d$-dimensional vector including intercept and covariates

$Y_i = (Y_{i1}, \ldots, Y_{iK}) \sim \text{ZIGDM}(\pi_i, a_i, b_i)$, where
$\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$, $a_i = (a_{i1}, \ldots, a_{iK})$, and $b_i = (b_{i1}, \ldots, b_{iK})$.

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.

$$
\mu_{ij} = \frac{e^{\alpha_j^T X_i}}{1 + e^{\alpha_j^T X_i}}, \quad \sigma_{ij} = \frac{e^{\beta_j^T X_i}}{1 + e^{\beta_j^T X_i}}, \quad \text{and} \quad \pi_{ij} = \frac{e^{\gamma_j^T X_i}}{1 + e^{\gamma_j^T X_i}},
$$

where $\alpha_j = (\alpha_{1j}, \ldots, \alpha_{dj})$, $\beta_j = (\beta_{1j}, \ldots, \beta_{dj})$, and $\gamma_j = (\gamma_{1j}, \ldots, \gamma_{dj})$ are regression coefficients for taxon $j$. 
Equivalent Hierarchical Model

\[ \Delta_{ij} \sim \text{Bernoulli}(\pi_{ij}), \quad j = 1, \ldots, K, \]
\[ Z_{ij} = 0 \text{ if } \Delta_{ij} = 1, \quad Z_{ij} \mid \Delta_{ij} = 0 \sim \text{Beta}(a_{ij}, b_{ij}), \quad j = 1, \ldots, K, \]
\[ P_{i1} = Z_{i1}, \quad P_{ij} = Z_{ij} \prod_{k=1}^{j-1}(1 - Z_{ik}), \quad j = 2, \ldots, K, \]
\[ Y_i \mid P_i \sim \text{Multinomial}(P_i, N_i), \]
where \( P_i = (P_{i1}, \ldots, P_{iK}) \) and \( N_i = \sum_{j=1}^{K+1} Y_{ij} \).
EM algorithm

Complete set of parameters: $\theta = (\gamma_1, \ldots, \gamma_K, \alpha_1, \ldots, \alpha_K, \beta_1, \ldots, \beta_K)$

Complete data log-likelihood expressed in terms of $Z$'s:

$$l(\theta) = \log \left[ \prod_{i=1}^{n} \left\{ \Pr(Y_i | Z_i) \prod_{j=1}^{K} f(Z_{ij}) \right\} \right]$$

$$= \sum_{i=1}^{n} \log \left\{ \Pr(Y_i | Z_i) \right\}$$

$$+ \sum_{j=1}^{K} \sum_{i=1}^{n} \left\{ \Delta_{ij} \log \pi_{ij} + (1 - \Delta_{ij}) \log(1 - \pi_{ij}) + (1 - \Delta_{ij}) \left[ -\log(B(a_{ij}, b_{ij})) + (a_{ij} - 1) \log(Z_{ij}) + (b_{ij} - 1) \log(1 - Z_{ij}) \right] \right\},$$

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,

$$Q^*_\theta = \sum_{j=1}^{K} \sum_{i=1}^{n} \mathbb{E}\left\{ \Delta_{ij} \log \pi_{ij} + (1 - \Delta_{ij}) \log(1 - \pi_{ij}) + (1 - \Delta_{ij}) \left[ -\log(B(a_{ij}, b_{ij})) + (a_{ij} - 1) \log Z_{ij} + (b_{ij} - 1) \log(1 - Z_{ij}) \right] \right\},$$

where the expectation is with respect to the posterior distributions of $(\Delta_i \mid Y_i; \theta^{(t-1)})$ and $(Z_i \mid \Delta_i, Y_i; \theta^{(t-1)})$ with $\theta^{(t-1)}$ being the parameter estimates in the $(t - 1)$-th M-step.
Based on the results for ZIGD posterior distribution, we can derive the explicit form for the posterior means:

\[
\Delta_{ij}^* = \mathbb{E}(\Delta_{ij} \mid Y_i) = \begin{cases} 
0 & \text{if } Y_{ij} > 0 \\
\frac{\pi_{ij}}{\pi_{ij} + (1-\pi_{ij})} \frac{\mathcal{B}(a_{ij}^*, b_{ij}^*)}{\mathcal{B}(a_{ij}, b_{ij})} & \text{if } Y_{ij} = 0
\end{cases},
\]

\[
A_{ij}^* = \mathbb{E}(\log Z_{ij} \mid Y_i, \Delta_{ij} = 0) = \psi(a_{ij}^*) - \psi(a_{ij}^* + b_{ij}^*),
\]

\[
B_{ij}^* = \mathbb{E}(\log(1 - Z_{ij}) \mid Y_i, \Delta_{ij} = 0) = \psi(b_{ij}^*) - \psi(a_{ij}^* + b_{ij}^*),
\]

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, $Q_\theta^*$ can be rewritten as

$$Q_\theta^* = \sum_{j=1}^{K} Q_{\gamma j}^* + \sum_{j=1}^{K} Q_{\alpha_j,\beta_j}^*,$$

(3)

where $Q_{\gamma j}^* = \sum_{i=1}^{n} \{ \Delta_{ij}^* \log \pi_{ij} + (1 - \Delta_{ij}^*) \log (1 - \pi_{ij}) \}$ and $Q_{\alpha_j,\beta_j}^* = \sum_{i=1}^{n} (1 - \Delta_{ij}^*) \{ - \log (B(a_{ij}, b_{ij})) + (a_{ij} - 1) A_{ij}^* + (b_{ij} - 1) B_{ij}^* \}$.

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

ZhengZheng Tang
Association Tests

Test for the mean: $H_0 : \alpha_{*1} = \ldots = \alpha_{*K} = 0$
Test for the dispersion: $H_0 : \beta_{*1} = \ldots = \beta_{*K} = 0$
Test for the presence-absence frequency: $H_0 : \gamma_{*1} = \ldots = \gamma_{*K} = 0$

- When performing a test on a particular set of parameters (e.g. $\alpha$’s), we include only the intercept coefficient in the model for the other sets of parameters (e.g. $\beta$’s and $\gamma$’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\textsubscript{1} and ZIGDM\textsubscript{2}
GDM-based tests: GDM\textsubscript{1} and GDM\textsubscript{2}
DM-based tests: DM\textsubscript{1} and DM\textsubscript{2} \textit{(La Rosa et al., PLOS ONE 2012)}
Non-parametric tests: QCAT\textsubscript{1} and QCAT\textsubscript{2} \textit{(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

\[ Y \sim \text{Multinomial}(\mathbf{P}, N); \ N \sim \text{Poisson}(1000) \]

<table>
<thead>
<tr>
<th>model for proportion ( \mathbf{P} )</th>
<th>data generation</th>
<th>parameter specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirichlet</td>
<td>[ \mathbf{P} = {P_j}_{j=1}^6 ]</td>
<td>mean of Dirichlet: ( \mu_j = 1/6 )</td>
</tr>
<tr>
<td></td>
<td>[ \sim \text{Dir}(\mu, \sigma) ]</td>
<td>dispersion of Dirichlet: ( \sigma = 0.3 )</td>
</tr>
<tr>
<td></td>
<td>[ \mu = {\mu_j}_{j=1}^6 ]</td>
<td></td>
</tr>
<tr>
<td>GD</td>
<td>[ Z_j \sim (Z</td>
<td>I)B(\pi_j, \mu_j, \sigma_j) ]</td>
</tr>
<tr>
<td>ZIGD</td>
<td>[ P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i) ]</td>
<td>dispersion of Beta: ( \sigma_j = 0.2 )</td>
</tr>
<tr>
<td></td>
<td>(( j = 1, \ldots, 5 ))</td>
<td>zero-inflation:</td>
</tr>
<tr>
<td></td>
<td>[ P_6 = 1 - \sum_{i=1}^{5} P_i ]</td>
<td>{( \pi_j }_{j=1}^5 = {0.1, 0.2, 0.4, 0.6, 0.8} }</td>
</tr>
<tr>
<td>LN</td>
<td>{W_j}_{j=1}^5 \sim LN(\mu, \sigma, \Omega)</td>
<td>mean of Normal: ( \mu_j = 0 )</td>
</tr>
<tr>
<td>ZILN</td>
<td>[ P_j = W_j/(\sum_{i=1}^{5} W_i + 1) ]</td>
<td>variance of Normal: ( \sigma_j = 1 )</td>
</tr>
<tr>
<td></td>
<td>(( j = 1, \ldots, 5 ))</td>
<td>correlation: ( \Omega_{jj'} = 0.5^{</td>
</tr>
<tr>
<td></td>
<td>replace counts with zero</td>
<td>zero-inflation:</td>
</tr>
<tr>
<td></td>
<td>[ P_6 = 1 - \sum_{i=1}^{5} P_i ]</td>
<td>{( \pi_j }_{j=1}^5 = {0.1, 0.2, 0.4, 0.6, 0.8} }</td>
</tr>
</tbody>
</table>
## Power Simulation Setup

<table>
<thead>
<tr>
<th>Parameter Specification</th>
<th>Perturbation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of Dirichlet: $\mu_j = 1/6$</td>
<td>$\mu_k \sim Unif(0, 0.5)$ OR $\sigma \sim Unif(0, 0.5)$</td>
</tr>
<tr>
<td>Dispersion of Dirichlet: $\sigma = 0.3$</td>
<td></td>
</tr>
<tr>
<td>Mean of Beta: $\mu_j = 0.2$</td>
<td>$\mu_k \sim Unif(0, 0.5)$ OR $\sigma_k \sim Unif(0, 1)$</td>
</tr>
<tr>
<td>Dispersion of Beta: $\sigma_j = 0.2$</td>
<td></td>
</tr>
<tr>
<td>Zero-inflation: $\pi_j = {0.1, 0.2, 0.4, 0.6, 0.8}$</td>
<td></td>
</tr>
<tr>
<td>Mean of Normal: $\mu_j = 0$</td>
<td>For LN: $\mu_k \sim Unif(0, 1)$ OR $\sigma_k \sim Unif(1, 6)$</td>
</tr>
<tr>
<td>Variance of Normal: $\sigma_j = 1$</td>
<td></td>
</tr>
<tr>
<td>Correlation: $\Omega_{jj'} = 0.5</td>
<td>j-j'</td>
</tr>
<tr>
<td>Zero-inflation: $\pi_j = {0.1, 0.2, 0.4, 0.6, 0.8}$</td>
<td></td>
</tr>
</tbody>
</table>
Power under Non-zero-inflated Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Diff</th>
<th>$n$</th>
<th>ZIGDM$_1$</th>
<th>ZIGDM$_2$</th>
<th>GDM$_1$</th>
<th>GDM$_2$</th>
<th>DM$_1$</th>
<th>DM$_2$</th>
<th>QCAT$_1$</th>
<th>QCAT$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>Mean</td>
<td>100</td>
<td>0.52</td>
<td>0.33</td>
<td>0.67</td>
<td>0.47</td>
<td>0.60</td>
<td>0.66</td>
<td>0.58</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.67</td>
<td>0.54</td>
<td>0.76</td>
<td>0.62</td>
<td>0.71</td>
<td>0.76</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Disp</td>
<td>100</td>
<td>0.17</td>
<td>0.72</td>
<td>0.42</td>
<td>0.77</td>
<td>0.05</td>
<td>0.76</td>
<td>0.06</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.52</td>
<td>0.84</td>
<td>0.60</td>
<td>0.84</td>
<td>0.05</td>
<td>0.81</td>
<td>0.05</td>
<td>0.75</td>
</tr>
<tr>
<td>GDM</td>
<td>Mean</td>
<td>100</td>
<td>0.60</td>
<td>0.27</td>
<td>0.66</td>
<td>0.34</td>
<td>0.59</td>
<td>0.60</td>
<td>0.60</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.70</td>
<td>0.39</td>
<td>0.76</td>
<td>0.48</td>
<td>0.70</td>
<td>0.71</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Disp</td>
<td>100</td>
<td>0.17</td>
<td>0.47</td>
<td>0.56</td>
<td>0.79</td>
<td>0.07</td>
<td>0.55</td>
<td>0.07</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.23</td>
<td>0.54</td>
<td>0.67</td>
<td>0.86</td>
<td>0.07</td>
<td>0.62</td>
<td>0.08</td>
<td>0.66</td>
</tr>
<tr>
<td>LN</td>
<td>Mean</td>
<td>100</td>
<td>0.48</td>
<td>0.06</td>
<td>0.48</td>
<td>0.06</td>
<td>0.50</td>
<td>0.51</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.63</td>
<td>0.06</td>
<td>0.63</td>
<td>0.05</td>
<td>0.65</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Disp</td>
<td>100</td>
<td>0.05</td>
<td>0.70</td>
<td>0.05</td>
<td>0.70</td>
<td>0.24</td>
<td>0.39</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>0.06</td>
<td>0.83</td>
<td>0.06</td>
<td>0.83</td>
<td>0.47</td>
<td>0.61</td>
<td>0.40</td>
<td>0.41</td>
</tr>
</tbody>
</table>
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


- 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)
BMI-Associated Lineages

Use Benjamini-Hochberg procedure to control FDR at 0.05 level

Results from DM and QCAT tests:
Family *Ruminococcaceae*
\[(DM_1 \text{ p-value} = 0.0013 \text{ and } QCAT_1 \text{ p-value} = 0.00014)\]
Family *Veillonellaceae*
\[(DM_2 \text{ p-value} = 0.0012 \text{ and } QCAT_2 \text{ p-value} = 0.00080)\]

Results from GDM and ZIGDM tests:
Family *Ruminococcaceae*
Family *Prevotellaceae*
\[(ZIGDM_2 \text{ p-value} = 0.0014 \text{ and } GDM_2 \text{ p-value} = 0.0014)\]
Kingdom *Bacteria*
\[(ZIGDM_2 \text{ p-value} = 0.0016)\]
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.

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

Download miLineage v1.0  miLineage v2.0 at CRAN  Download miLineage v3.0
Acknowledgements

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

Thank you!