A Bayesian Dirichlet-Multinomial Test for Cross-Group Differences

by

Yuhan Chen

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Supervisor

_____
Sayan Mukherjee

_____
Merlise Clyde

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2016

Abstract

A Bayesian Dirichlet-Multinomial Test for Cross-Group Differences

by

Yuhan Chen

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Li Ma, Supervisor

_____
Sayan Mukherjee

_____
Merlise Clyde

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2016

Abstract

Testing for differences within data sets is an important issue across various applications. Our work is primarily motivated by the analysis of microbiomial composition, which has been increasingly relevant and important with the rise of DNA sequencing. We first review classical frequentist tests that are commonly used in tackling such problems. We then propose a Bayesian Dirichlet-multinomial framework for modeling the metagenomic data and for testing underlying differences between the samples. A parametric Dirichlet-multinomial model uses an intuitive hierarchical structure that allows for flexibility in characterizing both the within-group variation and the cross-group difference and provides very interpretable parameters. A computational method for evaluating the marginal likelihoods under the null and alternative hypotheses is also given. Through simulations, we show that our Bayesian model performs competitively against frequentist counterparts. We illustrate the method through analyzing metagenomic applications using the Human Microbiome Project data.

To my family.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my supervisor, Li Ma, for his tremendous support and encouragement throughout the course of the thesis. Li's patience has been invaluable, and his insightful feedback and comments have been instrumental in my academic development at Duke.

Many thanks to my committee members, Sayan Mukherjee and Merlise Clyde, for their invaluable time and support. Since I first came to Duke, Sayan has been extremely friendly and has always provided sage advice whenever I feel troubled. Merlise was my linear models professor at Duke and her classes first showed me to the joys of statistics. Further thanks go to David Dunson for introducing me to exciting areas of statistical research as well as his amazing mentorship and patience.

I would like to thank my family and friends, whose enormous encouragement cannot be understated for all these years. I am extremely grateful to my parents, Yin and Ning, for their unconditional support through all the highs and lows of life. Without them, this thesis would not have been possible. I can also never thank all my classmates in the statistical science and biostatistics department enough. You guys are terrific!

Thank you!

1

Introduction

The comparison of two data sets to find the underlying differences has long been a key aspect of statistical inference. This thesis is primarily motivated by applications in microbial ecology, where the data can be represented using a count matrix with each row representing a sample and each entry within the row representing the number of times a microbial taxon is observed.

An average human is the host to a startling 10^{14} microbial cells, a quantity larger than the total number of stars in the Milky Way (Fujimura et al. 2010). Before the era of DNA sequencing, scientists interested in identifying and researching these microbes must directly use physical means through cultivation. As a result, these methods cannot fully capture the entire microbiome so the vast majority of microbial species remain unidentified. In recent years, with the introduction of direct DNA sequencing, the new advances allow microbial scientists to see the human microbiome with significantly higher clarity (Ng and Kirkness 2010).

Many factors can shape and form the microbial genetic diversity; various papers have linked diverse factors ranging from dietary patterns (Wu et al. 2011) to aging (Biagi et al. 2010) as being responsible for direct modulation of the composition.

Motivated by this potential, Human Microbiome Project (HMP) was launched in 2008 in order to further understand and test how different factors such as health or disease can change the human biocrobiome.

Understanding and characterizing the effects of these factors require the development of statistical tools. La Rosa et al. (2012) introduced the Dirichlet-multinomial multivariate hypothesis testing package based on earlier works on the subject (Wilson and Koehler 1984, Koehler and Wilson 1986) that models each microbial metagenomics data using a mixture of Dirichlet-multinomial distributions.

This thesis introduces a Bayesian Dirichlet-multinomial hypothesis testing framework that offers a fully Bayesian method of testing for cross-group differences. As with La Rosa et al. (2012), the method is multivariate parametric and hierarchical Dirichlet-multinomial structure similar to Holmes et al. (2012). This hierarchical structure can identify and capture both the within-group variation and between-group variation within each set of data and has fully interpretable parameters for the mean and the different sources of variation.

In Section 2, we review La Rosa et al. (2012)'s frequentist Dirichlet-multinomial test. In section 3, we consider the Bayesian interpretation and then introduce the Bayesian test for cross-group differences. A simple method for computing and estimating the marginal likelihood is provided. Section 4 contains a simulation study comparing our model with that of La Rosa et al. (2012)'s frequentist model. We also apply our method towards microbiome applications applications to analyze several HMP datasets. Section 5 provides a summary of the thesis as well as possible steps for future research.

2

Classical Dirichlet-multinomial testing

This thesis is primarily motivated by microbial metagenomics. In this section, we review the classical Dirichlet-Multinomial model as well as the testing strategy proposed by La Rosa et al. (2012). Assume we have S samples of data and each sample contains a count vector consisting of K components, e.g., one for each Bacteria taxon. Consider the matrix N, shown in Table 2.1, where each entry $n_j(h)$ represent the observed number of elements for component h and sample j, where $j = 1, 2, ..., S$ and $h = 1, ..., K$. We denote $\vec{n}_j = (n_j(1), \ldots, n_j(K))$ as the row vector and $n_j(+) = \sum_{h=1}^{K} n_j(h)$ as the total counts over all K taxon for a given sample j. For each j, it is common to assume that the sample $\vec{n}_j$ are observations from a multinomial distribution with parameter $\vec{p} = (p(1), \ldots, p(K))$:

$$\begin{aligned} \vec{n}_j &\sim \text{Multinom}(\vec{p}) \\ \text{L}(\text{N}|\vec{p}) &= \prod_{j=1}^{S} n_j(+)! \prod_{h=1}^{K} \frac{1}{n_j(h)!} p(h)^{n_j(h)}. \end{aligned}$$

The parameter $\vec{p}$ can be interpreted as the underlying probability an observation has of landing in each component. To account for overdispersion often caused by un-

accounted heterogeneity among the samples, one can adopt a Dirichlet-multinomial model that characterizes the overdispersion with parameter η, where we note that letting $\eta = 0$ implies the normal multinomial likelihood.

$$\begin{aligned}
\vec{n}_j &\sim \text{Dir-Multinom}(\vec{p}, \eta) \\
\text{L}(n_j|\vec{p}, \eta) &= n_j(+)! \frac{\prod_{h=1}^{K} \frac{1}{n_j(h)!} \prod_{r=1}^{n_j(h)} p(h)(1-\eta) + (r-1)\eta}{\prod_{r=1}^{n_j(+)} (1-\eta) + (r-1)\eta} \\
\text{L}(\text{N}|\vec{p}, \eta) &= \prod_{j=1}^{S} n_j(+)! \frac{\prod_{h=1}^{K} \frac{1}{n_j(h)!} \prod_{r=1}^{n_j(h)} p(h)(1-\eta) + (r-1)\eta}{\prod_{r=1}^{n_j(+)} (1-\eta) + (r-1)\eta}.
\end{aligned}$$

Now assume we are in a two group comparison case and let N_1 and N_2 be $S_1 \times K$ and $S_2 \times K$ matrices with K components and S_1 and S_2 samples respectively. The matrices have entries $n_{ij}(h)$ denoting the observed number of elements for group i, component h and sample j, where $i = 1, 2$, $j = 1, 2, ..., S_i$ and $h = 1, ..., K$. Let n_1 and n_2 be the total number of counts across all samples in N_1 and N_2 respectively. In many settings, scientists want to understand and characterize the differences between the two matrices and whether the samples come from the same distribution. An intuitive method would be to test whether $\vec{p}_1 = \vec{p}_2$, which can be formalized as

$$H_0 : \vec{p}_1 = \vec{p}_2 \quad \text{vs} \quad H_1 : \vec{p}_1 \neq \vec{p}_2.$$

We can calculate the generalized Wald-type test statistic (La Rosa et al. 2012)

$$\chi^2 = (\hat{\vec{p}}_1 - \hat{\vec{p}}_2)^T V^{-1} (\hat{\vec{p}}_1 - \hat{\vec{p}}_2),$$

where $\hat{\vec{p}}_1$ and $\hat{\vec{p}}_2$ are the MLE estimates of $\vec{p}_1$ and $\vec{p}_2$ respectively and the diagonal matrix V is given by

$$V = (\sum_{i=1}^{2} (n_i^2 C(\hat{\eta}, n_i)^{-1} (1 - \omega_i)^2))^{-1} D(\vec{p}_p),$$

Table 2.1: Data table with S samples and K components.

	Component					
Sample	1	2	3	...	K	Total
1	$n_1(1)$	$n_1(2)$	$n_1(3)$	...	$n_1(K)$	$n_1(+)$
2	$n_2(1)$	$n_2(2)$	$n_2(3)$	...	$n_2(K)$	$n_2(+)$
3	$n_3(1)$	$n_3(2)$	$n_3(3)$	...	$n_3(K)$	$n_3(+)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
S	$n_S(1)$	$n_S(2)$	$n_S(3)$	...	$n_S(K)$	$n_S(+)$
Total	$n_+(1)$	$n_+(2)$	$n_+(3)$	...	$n_+(K)$	$n_+(+)$

where $\hat{\eta}$ is the MoM estimates of the overdispersion parameter η and $D(\vec{p}_p)$ being a diagonal matrix with diagonal entries $\vec{p}_p = \sum_{i=1}^{2} \omega_i \hat{\vec{p}}_i$, where for all $i = 1, 2$,

$$\omega_i = n_i^2 C(\eta_i, n_i)^{-1} (\sum_{r=1}^{2} n_r^2 C(\eta_r, n_r)^{-1})^{-1}, \quad C(\eta_i, n_i) = \eta_i (\sum_{j=1}^{S_i} n_{ij}(+)^2 - n_i) + n_i,$$

where $n_{ij}(+)$ is the total number of observations for sample j in group i. Finally, the test statistic's null distribution is asymptotically a χ^2 distribution with $K - 1$ degrees of freedom.

3

Bayesian testing framework

3.1 Bayesian Dirichlet-multinomial model

After reviewing the frequentist version of the Dirichlet-multinomial model, we can introduce a Bayesian counterpart. Recall the multinomial model given in Section 2.3. We have S samples of a certain data and each sample contains K components. Consider the matrix N, shown in Table 2.1, where each entry $n_j(h)$ represent the observed number of elements for component h and sample j, where $j = 1, 2, ..., S$ and $h = 1, ..., K$. We denote $\vec{n}_j = (n_j(1), \ldots, n_j(K))$ as the row vector and $n_j(+) = \sum_{h=1}^{K} n_j(h)$ as the total number of observations for a given sample j. For each j, the sample $\vec{n}_j$ are observations from a multinomial likelihood with parameter $\vec{p}_j = (p_j(1), \ldots, p_j(K))$:

$$\begin{aligned} \vec{n}_j &\sim \text{Multinom}(\vec{p}_j) \\ \text{L}(n_j|\vec{p}_j) &= n_j(+)! \prod_{h=1}^{K} \frac{1}{n_j(h)!} p_j(h)^{n_j(h)} \\ \text{L}(\text{N}|\vec{p}_1, \ldots, \vec{p}_S) &= \prod_{j=1}^{S} n_j(+)! \prod_{h=1}^{K} \frac{1}{n_j(h)!} p_j(h)^{n_j(h)}. \end{aligned}$$

The parameter $\vec{p}_j$ can be interpreted as the underlying probability an observation has of landing in a certain component within sample j. Under the frequentist method, a Dirichlet-multinomial model with overdispersion was introduced, but taking a Bayesian perspective allows us to introduce a prior to $\vec{p}_j$ instead. Intuitively, both method allows us to introduce more variability. Because $\vec{p}_j$ lies on the probability simplex as $p_j(h) \geqslant 0, p_j(1) + p_j(2) + ... + p_j(K) = 1$, a natural choice would be a Dirichlet prior

$$\vec{p}_j \sim \text{Dir}(\eta\vec{p}_0),$$

where $\vec{p}_0$ represents the underlying probability vector beneath all the samples and η represents the precision that controls the degree of variation each sample probability $\vec{p}_j$ has around $\vec{p}_0$.

3.2 Bayesian Dirichlet-multinomial test

Using the Dirichlet-multinomial model from above, we can introduce our main framework for testing two samples. Extending the notation given in Table 2.1, let N_1 and N_2 be $S_1 \times K$ and $S_2 \times K$ matrices with K components and S_1 and S_2 samples respectively. The matrices have entries $n_{ij}(h)$ denoting the observed number of elements for group i, component h and sample j, where $i = 1, 2$, $j = 1, 2, ..., S_i$ and $h = 1, ..., K$. Like before, for each matrix N_i, denote $\vec{n}_{ij} = (n_{ij}(1), \ldots, n_{ij}(K))$ as the row vector and $n_{ij}(+) = \sum_{h=1}^{K} n_{ij}(h)$ as the total number of observations for a given sample j in the group i. For each matrix N_i, we can fit a Dirichlet-multinomial model with a slight abuse of notation

$$\begin{aligned} \vec{n}_{ij} &\sim \text{Multinom}(\vec{p}_{ij}) \\ \vec{p}_{ij} &\sim \text{Dir}(\eta_i\vec{p}_i). \end{aligned}$$

Consider the problem of testing

$$H_0 : \vec{p}_1 = \vec{p}_2 \quad \text{vs} \quad H_1 : \vec{p}_1 \neq \vec{p}_2,$$

where $\vec{p_i} \sim \text{Dir}(\tau\vec{p_0})$, for precision parameter τ. In this setting, η_i characterizes the between-sample variation and τ characterizes the cross-sample variation. This model has the advantage of allowing us to take an ANOVA approach and decompose the variation. Given that our model can be highly sensitive to η_i, let $\eta_i \sim \text{Ga}(a, b)$ be a hierarchical prior where a and b are preset hyperparameters. We can do the same for τ, but in a two-sample setting, there isn't enough data to warrant such an approach. Instead, we adopt Jeffrey's prior and set $\tau = 1$, $\vec{p_0} = (\frac{1}{K}, \ldots, \frac{1}{K})$. Let $\Pr(H_0)$ be the prior probability of the null hypothesis H_0. By Bayes' Theorem, we can compute the posterior probability of H_0

$$\begin{aligned}\Pr(H_0|\text{N}) &= \frac{\Pr(H_0)\Pr(\text{N}|H_0)}{\Pr(H_0)\Pr(\text{N}|H_0) + \Pr(H_1)\Pr(\text{N}|H_1)} \\ &= \frac{1}{1 + \frac{1-\Pr(H_0)}{\Pr(H_0)}\frac{\Pr(\text{N}|H_1)}{\Pr(\text{N}|H_0)}},\end{aligned}$$

where $\text{N} = \{\text{N}_1, \text{N}_2\}$ through some abuse of notation. We call

$$BF = \frac{\Pr(\text{N}|H_1)}{\Pr(\text{N}|H_0)}$$

as the Bayes factor in favor of the alternative and

$$\frac{1 - \Pr(H_0)}{\Pr(H_0)} = \frac{\Pr(H_1)}{\Pr(H_0)}$$

the prior odds. Both the posterior probability and the Bayes factor can be used as evidence for the two sample test, with the Bayes factor holding the advantage of not containing the sensitivity of specifying a prior distribution for the null.

3.3 Computation of marginal likelihood

A major challenge in the Bayesian Dirichlet-multinomial testing model described above is the computation of the marginal likelihood $\Pr(\text{N}|H_a)$ for $a = 0, 1$. The full

conditional likelihood of N under H_a given every other parameter can be represented as

$$\Pr(\mathrm{N}|H_a, \{\vec{p}_i\}, \{\vec{p}_{ij}\}, \{\eta_i\}) = \prod_{i=1}^{2}\prod_{j=1}^{S_i}\left(\frac{(n_{ij}(1)+...+n_{ij}(K))!}{n_{ij}(1)!...n_{ij}(K)!}\prod_{h=1}^{K}p_{ij}(h)^{n_{ij}(h)}\right)$$

Calculating the marginal likelihood would require finding

$$\Pr(\mathrm{N}|H_a) = \int \Pr(\mathrm{N}|H_a, \{\vec{p}_i\}, \{\vec{p}_{ij}\}, \{\eta_i\})\pi(\{\vec{p}_i\}, \{\vec{p}_{ij}\}, \{\eta_i\})d(\{\vec{p}_i\}, \{\vec{p}_{ij}\}, \{\eta_i\}).$$

Because this likelihood does not have an analytic close form, we instead rely on computational methods to provide a reasonable approximation. We show one possible avenue for computation in this section.

We start by integrating out the parameters that are tractable. Note that

$$\begin{aligned}
\pi(\vec{p}_{ij}|\{\eta_i\}, \{\vec{p}_i\}) &= \frac{1}{\beta(\eta_i\vec{p}_i)}\prod_{h=1}^{K}p_{ij}(h)^{\eta_i p_i(h)-1} \\
\pi(\vec{p}_i) &= \frac{1}{\beta\left(\frac{1}{K},...,\frac{1}{K}\right)}\prod_{h=1}^{K}p(h)^{\frac{1}{K}-1} \\
\pi(\eta_i) &= \frac{b^a}{\Gamma(a)}\eta_i^{a-1}e^{-b\eta_i},
\end{aligned}$$

where β is the multivariate Beta function. The joint conditional distribution has the form

$$\begin{aligned}
\Pr(\mathrm{N}, \{\vec{p}_{ij}\}|H_a, \{\vec{p}_i\}, \{\eta_i\}) &= \prod_{i=1}^{2}\prod_{j=1}^{S_i}\left(\frac{(n_{ij}(1)+...+n_{ij}(K))!}{n_{ij}(1)!...n_{ij}(K)!}\prod_{h=1}^{K}p_{ij}(h)^{n_{ij}(h)}\right) \\
&\quad \cdot\prod_{i=1}^{2}\prod_{j=1}^{S_i}\left(\frac{1}{\beta(\eta_i\vec{p}_i)}\prod_{h=1}^{K}p_{ij}(h)^{\eta_i p_i(h)-1}\right).
\end{aligned}$$

Integrating out $\{\vec{p}_{ij}\}$ gives

$$\Pr(\mathrm{N}|H_a, \{\vec{p}_i\}, \{\eta_i\}) = \prod_{i=1}^{2}\prod_{j=1}^{S_i}\frac{(n_{ij}(1)+...+n_{ij}(K))!}{n_{ij}(1)!...n_{ij}(K)!}\frac{\beta(\vec{n}_{ij}+\eta_i\vec{p}_i)}{\beta(\eta_i\vec{p}_i)}.$$

Unfortunately, there is no closed formulas for integrating out either $\{\vec{p_i}\}$ or $\{\eta_i\}$. However, we can instead use Laplace approximation to calculate an estimate of the integral. Under H_0, $\vec{p}_1 = \vec{p}_2 = \vec{p}$, the joint distribution of $\mathrm{N}, \{\vec{p_i}\}, \{\eta_i\}$ under H_0 can be expressed as

$$\Pr(\mathrm{N}, \{\vec{p_i}\}, \{\eta_i\}|H_0) = A_0 \prod_{i=1}^{2} \prod_{j=1}^{S_i} \frac{\beta(\vec{n}_{ij} + \eta_i \vec{p})}{\beta(\eta_i \vec{p})} \cdot \prod_{h=1}^{K} p(h)^{\frac{1}{K}-1} \cdot \prod_{i=1}^{2} \eta_i^{a-1} e^{-b\eta_i}, \quad (3.1)$$

where

$$A_0 = \prod_{i=1}^{2} \prod_{j=1}^{S_i} \frac{(n_{ij}(1) + ... + n_{ij}(K))!}{n_{ij}(1)!...n_{ij}(K)!} \cdot \frac{1}{\beta\left(\frac{1}{K}, ..., \frac{1}{K}\right)} \cdot \left(\frac{b^a}{\Gamma(a)}\right)^2.$$

and the joint distribution of $\mathrm{N}, \{\vec{p_i}\}, \{\eta_i\}$ under H_1 is

$$\Pr(\mathrm{N}, \{\vec{p_i}\}, \{\eta_i\}|H_1) = A_1 \prod_{i=1}^{2} \prod_{j=1}^{S_i} \frac{\beta(\vec{n}_{ij} + \eta_i \vec{p_i})}{\beta(\eta_i \vec{p_i})} \cdot \prod_{i=1}^{2} \prod_{h=1}^{K} p_i(h)^{\frac{1}{K}-1} \cdot \prod_{i=1}^{2} \eta_i^{a-1} e^{-b\eta_i}, \quad (3.2)$$

and

$$A_1 = \prod_{i=1}^{2} \prod_{j=1}^{S_i} \frac{(n_{ij}(1) + ... + n_{ij}(K))!}{n_{ij}(1)!...n_{ij}(K)!} \cdot \frac{1}{\left(\beta\left(\frac{1}{K}, ..., \frac{1}{K}\right)\right)^2} \cdot \left(\frac{b^a}{\Gamma(a)}\right)^2.$$

Finding $\Pr(\mathrm{N}|H_0)$ would require integrating out $\{\vec{p}\}$ and $\{\eta_i\}$ from (3.1) using

$$\Pr(\mathrm{N}|H_0) = A_0 \int \prod_{i=1}^{2} \prod_{j=1}^{S_i} \frac{\beta(\vec{n}_{ij} + \eta_i \vec{p})}{\beta(\eta_i \vec{p})} \cdot \prod_{h=1}^{K} p(h)^{\frac{1}{K}-1} dp(h) \cdot \prod_{i=1}^{2} \eta_i^{a-1} e^{-b\eta_i} d\eta_i, \quad (3.3)$$

where the integral is over $(0, \infty)$ for η_1, η_2 and Ω for $p(1), ..., p(K-1))$, and Ω is the simplex

$$\Omega : p(1) \geqslant 0, ..., p(K-1) \geqslant 0, \quad \sum_{h=1}^{K-1} p(h) \leqslant 1.$$

To integrate out $\vec{p}$, we would like to use Laplace approximation, but since $\vec{p}$ is on the probability simplex, this make direct application of Laplace method difficult.

Instead, to simplify calculations, we can use a change-of-variable by using the transformation $T : \mathbb{R}^{K-1} \to \Omega$ to convert the region Ω in (3.3) to $\mathbb{R}^{K-1}$. Details about the transformation T as well as its Jacobian matrix are provided in Appendix A.

Set $\vec{u} = T^{-1}(\vec{p}) \in \mathbb{R}^{K-1}$. To extend the region of ν_i to $\mathbb{R}$, we set $\nu_i = \log(\eta_i) \in (-\infty, \infty)$. Under these change of variables, we can now express the integral in (3.3) as

$$\Pr(\mathrm{N}|H_0) = A_0 \int_{\mathbb{R}^{K+1}} e^{f_0(\vec{u},\vec{\nu})} d\nu_1 \, d\nu_2 \, du(1) \ldots du(K-1), \tag{3.4}$$

where

$$\begin{aligned} f_0(\vec{u},\vec{\nu}) &= \sum_{i=1}^{2} \sum_{j=1}^{S_i} (\log \beta(\vec{n}_{ij} + \eta_i \vec{p}) - \log \beta(\eta_i \vec{p})) + \log(|\det(J_0)|) \\ &\quad + \left(\frac{1}{K} - 1\right) \sum_{h=1}^{K} \log(p(h)) + \sum_{i=1}^{2} (a \log(\eta_i) - b\eta_i), \end{aligned}$$

and $\vec{p} = T(\vec{u})$ and $\eta_i = e^{\nu_i}$ and J_0 is the Jacobian matrix of T.

To use Laplace approximation of the integral in (3.3), we first need to find $\vec{u}_*, \vec{\nu}_*$, the maximizer of f_0. Then at the maximizer, we estimate the values of f_0 and the Hessian matrix Σ_0 of f_0 with respect to $\vec{u}, \vec{\nu}$ at $\vec{u}_*, \vec{\nu}_*$. Computation-wise, these values can be found using the BFGS algorithm in R on the function $-f_0$. Thus, the estimated value for $\log \Pr(\mathrm{N}|H_0)$ is

$$\log \Pr(\mathrm{N}|H_0) = \log(A_0) + f_0(\vec{u}_*, \vec{\nu}_*) + \frac{(K+1)\log(2\pi) - \log(|\det(\Sigma_0)|)}{2}.$$

Similarly,

$$\Pr(\mathrm{N}|H_1) = A_1 \int_{\mathbb{R}^{2K}} e^{f_1(\vec{u}_1,\vec{u}_2,\vec{\nu})} d\nu_1 \, d\nu_2 \, du_1(1) \, du_2(1) \ldots du_1(K-1) du_2(K-1), \tag{3.5}$$

where

$$f_1(\vec{u}_1, \vec{u}_2, \vec{\nu}) = \sum_{i=1}^{2}\sum_{j=1}^{S_i}(\log\beta(\vec{n}_{ij} + \eta_i\vec{p}_i) - \log\beta(\eta_i\vec{p}_i)) + \log(|\det(J_1)|) + \left(\frac{1}{K} - 1\right)\sum_{i=1}^{2}\sum_{h=1}^{K}\log(p_i(h)) + \sum_{i=1}^{2}(a\log(\eta_i) - b\eta_i),$$

and $\vec{p}_i = T(\vec{u}_i)$ and $\eta_i = e^{\nu_i}$ and $\det(J_1) = \det\left(\frac{\partial\vec{p}_1}{\partial\vec{u}_1}\right)\det\left(\frac{\partial\vec{p}_2}{\partial\vec{u}_2}\right)$ is the determinant of the Jacobian matrix. If $\vec{u}_{1*}, \vec{u}_{2*}, \vec{\nu}_*$ are the maximizer of f_1, then

$$\log\Pr(\mathrm{N}|H_1) = \log(A_1) + f_0(\vec{u}_{1*}, \vec{u}_{2*}\vec{\nu}_*) + \frac{2K\log(2\pi) - \log(|\det(\Sigma_1)|)}{2},$$

where Σ_1 is the hessian matrix of f_1 at $\vec{u}_{1*}, \vec{u}_{2*}, \vec{\nu}_*$ with respect to $\vec{u}_1, \vec{u}_2, \vec{\nu}$.

4

Numerical Examples

4.1 Simulations

In this section, we run simulations to evaluate the Bayesian Dirichlet-multinomial test's performance relative to the La Rosa et al. (2012)'s frequentist counterpart. Under each of the following scenarios, we simulate 500 datasets. Each dataset has two groups, with each group containing K components and S samples. Let η_1 and η_2 represent the between-group variations and $\vec{p}_1$ and $\vec{p}_2$ represent the underlying probability vector for each of the two groups of samples. For sample $1 \leqslant j \leqslant S$, we simulate under the alternate

$$n_{1j} \sim \text{Multinom}(n, \vec{p}_{1j}), \quad \vec{p}_{1j} \sim \text{Dir}(\vec{p}_1\eta_1),$$
$$n_{2j} \sim \text{Multinom}(n, \vec{p}_{2j}), \quad \vec{q}_{2j} \sim \text{Dir}(\vec{p}_2\eta_2),$$

where n represents the total number of observations for each sample. For each dataset, we construct a corresponding null dataset by setting $\vec{p}_1 = \vec{p}_2$ while still maintaining the different η_1 and η_2. Results can be sensitive to the selection of $\vec{p}_1$ and $\vec{p}_2$, so we test how well they perform under a variety of settings. We begin by first fixing a base $\vec{p}_1$ and then randomly generate $\vec{p}_2$ by fluctuating it around $\vec{p}_1$.

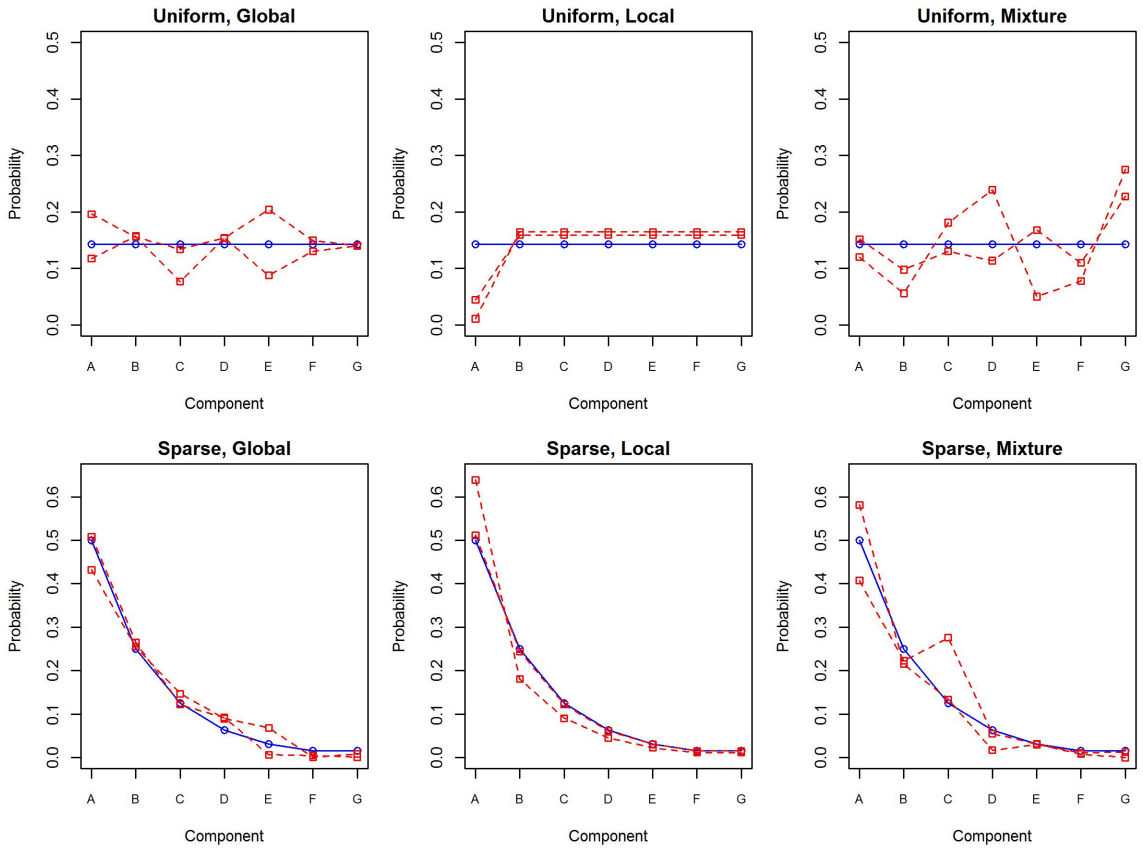

FIGURE 4.1: Intuitive description of the six settings we consider. The six graphs show the component probability for the first group $\vec{p}_1$ in blue under uniform or sparse settings and potential values of that of the second group $\vec{p}_2$ in red under global, local and mixture changes.

Consider two such ways of setting a base $\vec{p}_1$:

1. Uniform $\vec{p}_1$: we set $\vec{p}_1 = \frac{1}{K}, \ldots, \frac{1}{K}$.

2. Sparse $\vec{p}_1$: in applied settings, assuming a uniform $\vec{p}_1$ can be extremely inaccurate, as real bacterial data can be highly sparse. Hence in this setting, we set $\vec{p}_1 = \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^{K-1}}, \frac{1}{2^{K-1}}$.

After having a $\vec{p}_1$, we consider three ways of generating $\vec{p}_2$, the base underlying probability for the second group of samples:

1. Global change: we set $\vec{p}_2 \sim \text{Dir}(\phi\vec{p}_1)$ with precision parameter ϕ.

2. Local change: we simulate $\vec{p}_2$ from $\vec{p}_1$ under local variation where only the first component experiences a major change and the other components are scaled to keep $\vec{p}_2$ on the probability simplex. Hence we first generate $p_2(1) \sim \text{Beta}(\mu = p_1(1), \sigma^2 = \frac{1}{\phi})$, a reparametrized beta distribution centered on $p_1(1)$ with precision parameter ϕ, and then $p_2(2), \ldots p_2(K)$ take the rescaled values of $p_1(2), \ldots p_1(K)$.

3. Mixture change: this setting considers $\vec{p}_2$ being generated using a mixture of Dirichlet distributions. Hence $\vec{p}_2 \sim \frac{1}{2}\text{Dir}(\phi\vec{q}_1) + \frac{1}{2}\text{Dir}(\phi\vec{q}_2)$, where $\vec{q}_1, \vec{q}_2 \sim \text{Dir}(\phi\vec{p}_2)$.

An intuitive plot of the different settings is given in Figure 4.1. We run our model against La Rosa et al. (2012) using the Xmcupo function from the R package `HMP`. We generate ROC curves that compare the true positive rate against the false positive rate by using the Bayes Factor and Xmcupo's χ^2 test statistic. For each of the six cases, we alter S and n, the total number of samples and observations as well as K, the number of components and ϕ the precision. We set the prior with hyperparameters $\eta_i \sim \text{Ga}(10, 0.1)$.

First, consider a small sample case where $S = 2$, $n = 50$, $\phi = 20$ for global, local and mixture changes. We test two situations where the within-group variations are the same for both groups at $\eta_1 = \eta_2 = 100$ and when they are different at $\eta_1 = 20$ and $\eta_1 = 100$. Figure 4.2 and Figure 4.3 suggest that the Baysian approach performs better than the frequentist approach in almost every setting. This is to be expected since Bayesian methods incur smoothing for smaller data sets and La Rosa et al. (2012)'s estimated MLE can be 0 for some parameters. Figure 4.4 and Figure 4.5 show with larger sample sizes, where $S = 10$, $n = 500$, $\phi = 100$. The

Bayesian Dirichlet-multinomial hypothesis testing framework especially exhibits very close performance to La Rosa et al. (2012)'s model. This is perhaps unsurprising because both models operate under a similar Dirichlet-multinomial model. Xmcupo behaves oddly for larger values of K with sparsity where the issue of having MLEs of 0 can occur.

Next, we investigate the behavior of the log Bayes factor as a function of the sample size, number of observations, within-group variation and the cross-group difference under both hypotheses. Let $\vec{p}_1 = (0.25, 0.25, 0.25, 0.25)$ in all situations and $\vec{p}_2 = (0.15, 0.15, 0.1, 0.6)$ under the alternate hypothesis. We simulate 300 datasets. Consider the following scenarios:

1. Within-group variation: $n = 100$ and $S = 15$. We vary $\eta_1 = \eta_2$ from 2 to 10.
2. Number of samples: $\eta_1 = \eta_2 = 100$, $n = 100$. We vary S from 2 to 11.
3. Number of observations: $\eta_1 = \eta_2 = 100$, $S = 15$. We vary S from 50 to 500.
4. Cross-group differences: $\eta_1 = \eta_2 = 100$, $S = 15$, $n = 100$. Under both local and global differences, we vary ϕ from 10 to 100.

The simulations from Figure 4.6 show that the absolute value of the log Bayes factor increases as n, S and η increase. This is to be expected as more data and less within-group variation provides stronger degree of certainty. The log Bayes factor also shrinks as the cross-group variation decreases until the alternate is no longer distinguishable from the null.

4.2 Applications

In this section, we use our Bayesian Dirichlet-multinomial test to analyze real metageonomic data from the Human Microbiome Project found in the R package `HMP`.

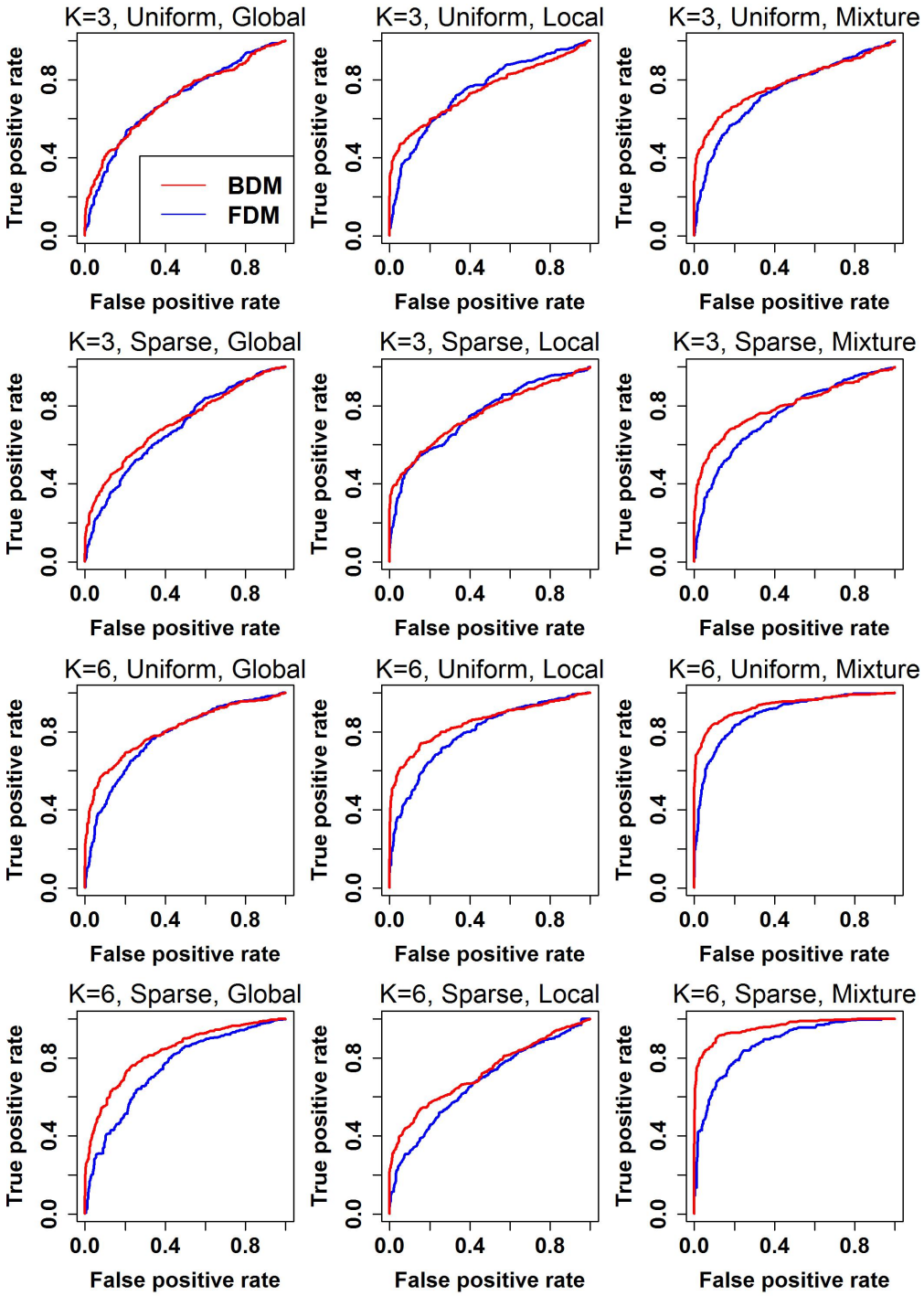

FIGURE 4.2: ROC curves for the six settings for the number of components $K = 3, 6$, the number of samples $S = 2$, within-group variations $\eta_1 = 100$, $\eta_2 = 100$, and number of observations per sample $n = 50$. BDM shows the Bayesian Dirichlet-multinomial test and FDM is the frequentist version.

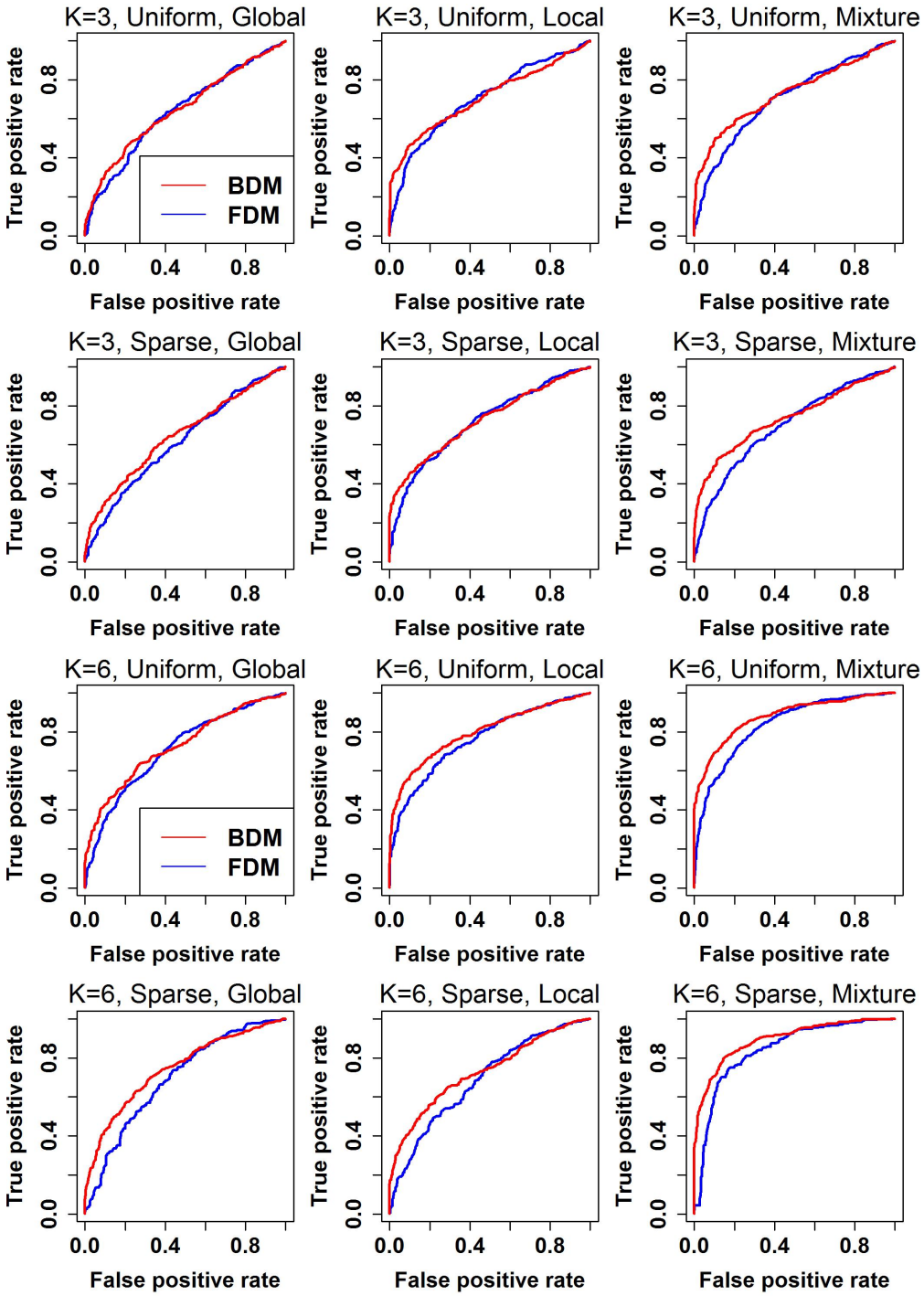

FIGURE 4.3: ROC curves for the six settings for the number of components $K = 3, 6$, the number of samples $S = 2$, within-group variations $\eta_1 = 20$, $\eta_2 = 100$, and number of observations per sample $n = 50$. BDM shows the Bayesian Dirichlet-multinomial test and FDM is the frequentist version.

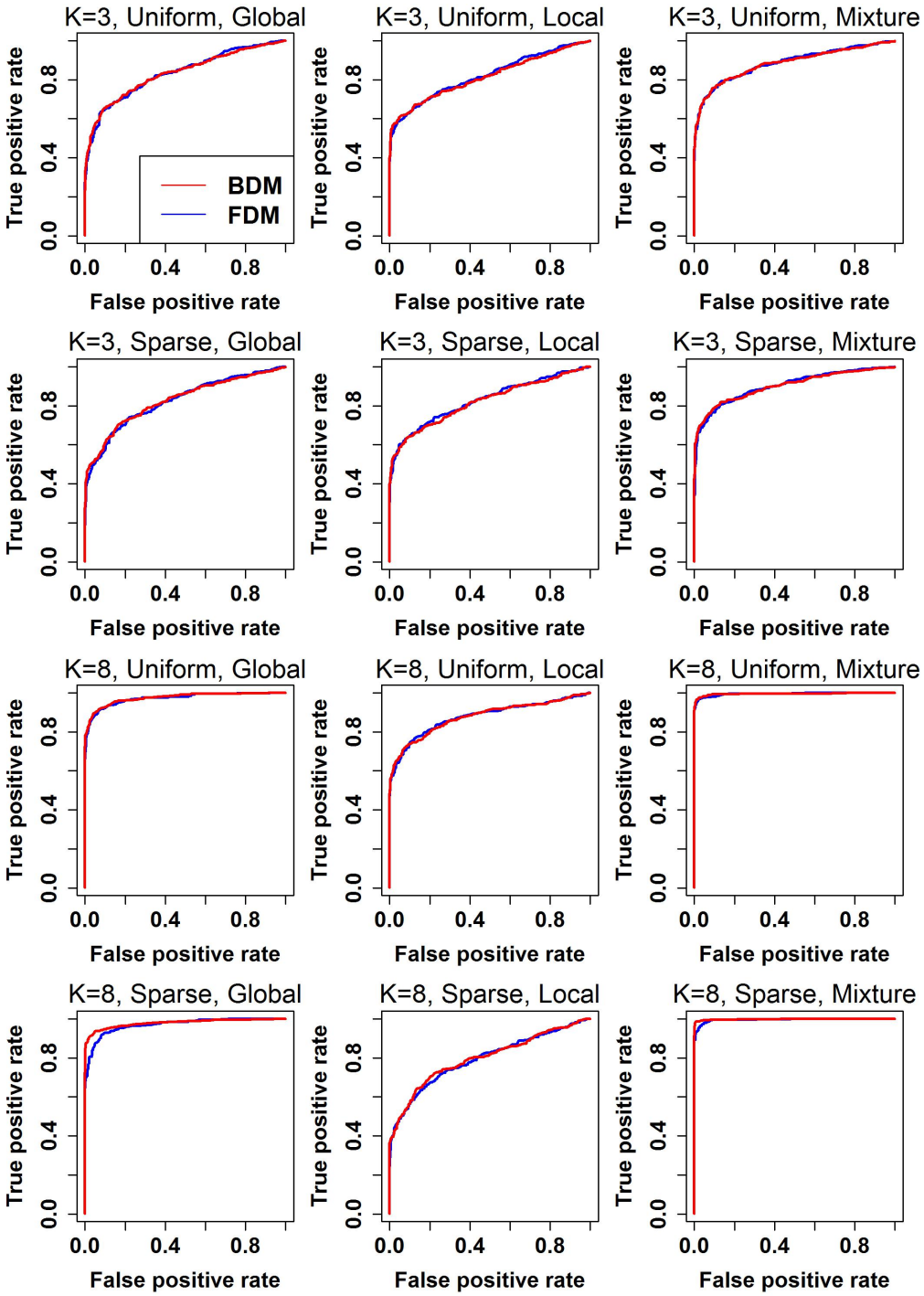

FIGURE 4.4: ROC curves for the six settings for the number of components $K = 3, 8$, the number of samples $S = 10$, within-group variations $\eta_1 = 100$, $\eta_2 = 100$, and number of observations per sample $n = 500$. BDM shows the Bayesian Dirichlet-multinomial test and FDM is the frequentist version.

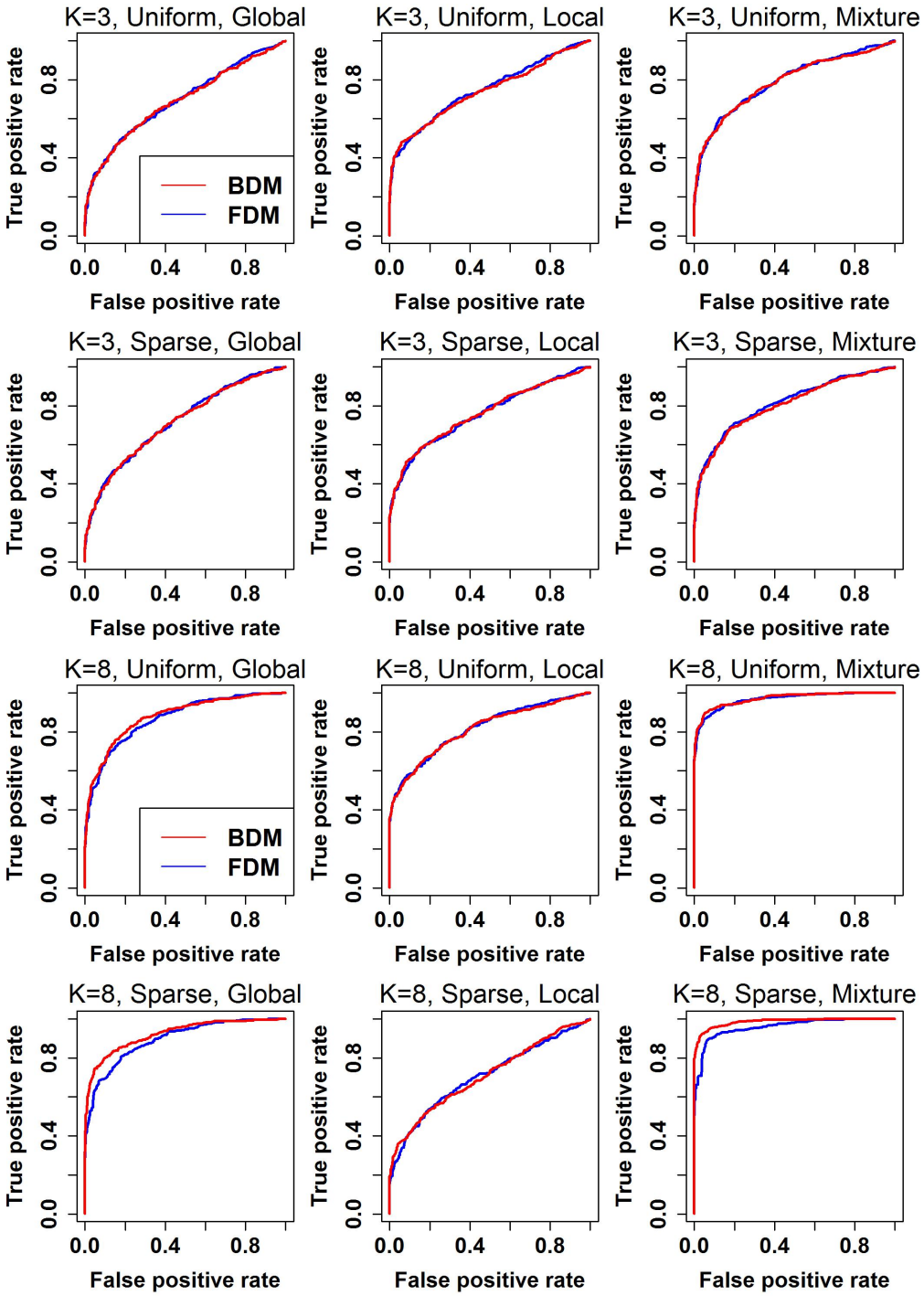

FIGURE 4.5: ROC curves for the six settings for the number of components $K = 3, 8$, the number of samples $S = 10$, within-group variations $\eta_1 = 20$, $\eta_2 = 100$, and number of observations per sample $n = 500$. BDM shows the Bayesian Dirichlet-multinomial test and FDM is the frequentist version.

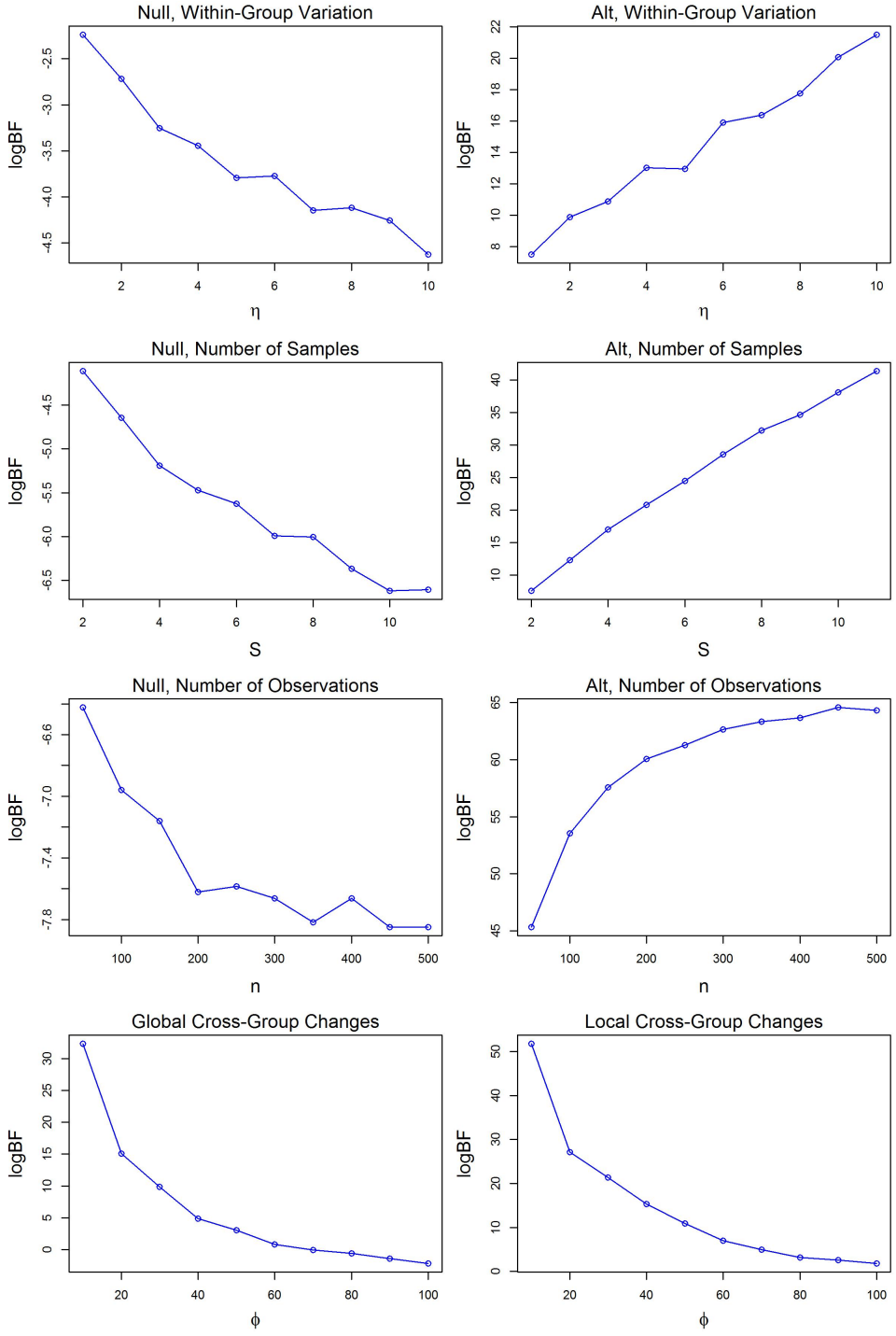

FIGURE 4.6: Behavior of the log Bayes factor under the null and alternative, after altering the within-sample variation, cross-group difference, total number of samples and total number of observations.

The HMP data analyzed comes from the tonsils, throat, and saliva sites of 24 subjects between the ages of 18 and 40 of both genders from Houston, Texas and St. Louis, Missouri and contains the 20 most abundant taxa found at these sites, with the remaining taxa grouped together into one component. These results are obtained using genetic sequencing on the 16S ribosomal RNA gene at the J. Craig Venter Institute, Broad Institute, Human Genome Sequencing Center at Baylor, and Genome Sequencing Center at Washington University in St. Louis. The data in `HMP` is presented using rank abundance distribution (RAD) format, which is the common ecological analysis done without taxa labels with the focus being on identifying the community structure (Whittaker 1965). The data can be visualized in Figure 4.7 and Figure 4.8.

The digestive tract has the most microbial diversity in the human body, so significant work has been done by biologists on its bacterial composition (Segata et al. 2012). We want to test whether underlying differences exist between the microbial composition found in the tonsils, throat and saliva sites, so analyses are run using both the Bayesian Dirichlet-multinomial test as well as the frequentist perspective using the Xmcupo function from the R package `HMP` which calculates the χ^2 test statistic and p-value.

First, we consider the tonsils and the throat samples. Figure 4.7 shows the relative proportion of the microbiome data found in the tonsils as as well as the throat. We run both the Bayesian Dirichlet-multinomial two group hypothesis test as well as La Rosa et al. (2012)'s Xmcupo. The Bayesian method gives a log Bayes factor of -76.49. Hence assuming a neutral hypothesis prior, this gives $\Pr(H_0) = 1$. On the other hand, Xmcupo calculates the χ^2 test statistic as 3.25 with a p-value of 0.99. This very high p-value can be a result of the paired nature of the data. Ultimately, both analyzes favor the null hypothesis that there isn't a fundamental cross-group difference between the tonsils samples and the throat samples. Using

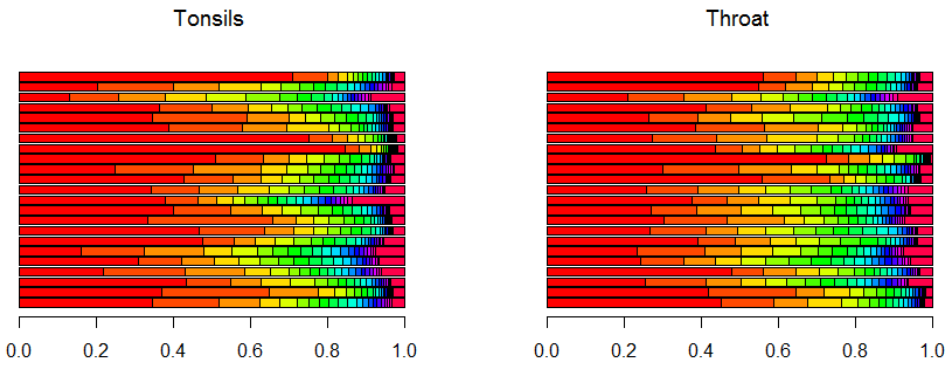

FIGURE 4.7: Comparison of the bacterial composition between tonsils and throat. Each colored bar represents the proportion one taxon makes up, with the taxa ranked by abundance from left to right.

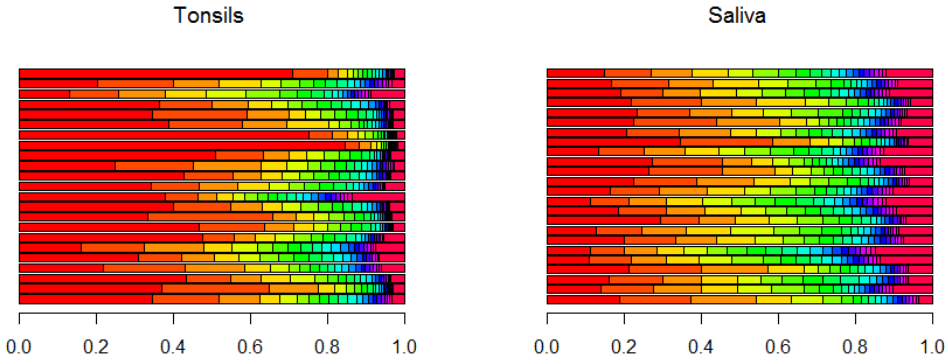

FIGURE 4.8: Comparison of the bacterial composition between tonsils and saliva. Each colored bar represents the proportion one taxon makes up, with the taxa ranked by abundance from left to right.

Table 4.1: Parameter estimates of HMP data for the tonsils and throat sites. The probability vector only shows the three most abundant taxa.

	tonsils η	throat η	tonsils $\vec{p}_1$	throat $\vec{p}_1$
Baye DM (MAP)	95.50	151.24	(0.36, 0.14, 0.09, ...)	(0.36, 0.14, 0.09, ...)
Freq DM (MLE)	-	-	(0.37, 0.14, 0.09, ...)	(0.36, 0.13, 0.09, ...)
Freq DM (MoM)	-	-	(0.39, 0.16, 0.09, ...)	(0.36, 0.13, 0.09, ...)

MAP estimation of our parameters, we present the estimates for the between-group variation as well as the underlying base probability for each group in Table 4.1. We compare these estimates with those frequentist estimates using the MLE and MoM found in the `HMP` package. Due to limited space, we only present the underlying probability vector from the three most frequent taxa.

Next, consider the question of whether the saliva samples and the tonsils samples differ. Figure 4.8 shows the relative proportion of the microbiome data in the tonsils as opposed to the saliva. Once again, we apply both methods. The Bayesian test finds a log Bayes factor of 32.33 with the posterior null probability $\Pr(H_0) = 10^{-15}$ under a neutral null prior. Meanwhile, Xmcupo calculates its own χ^2 test statistic as 88.42 with a p-value 10^{-20}. The two analyzes heavily favor the alternative hypothesis and believe that there exists significant a difference between the tonsils samples and the saliva samples. We present our parameter estimates in Table 4.2.

Both these tests are in line with what we would expect visually, given that from Figure 4.7, the tonsils and throat seem to have similar microbiome composition. This is opposed to Figure 4.8, in which there exists clear shifts in the probability vector, with the most abundant taxum having less weight in the saliva sitess. From our results in Table 4.1 and Table 4.2, we see that the probability vector estimated under the MAP and MLE are similar. There exists a slight variation between the MAP and MoM for tonsils, but overall the results show that the Dirichlet-multinomial methods exhibit alike inference.

Table 4.2: Parameter estimates of HMP data for the tonsils and saliva sites. The probability vector only show the three most abundant taxa.

	tonsils η	saliva η	tonsils $\vec{p}_1$	saliva $\vec{p}_1$
Baye DM (MAP)	96.46	236.2	(0.37, 0.14, 0.09, ...)	(0.19, 0.14, 0.12, ...)
Freq DM (MLE)	-	-	(0.37, 0.14, 0.09, ...)	(0.19, 0.14, 0.12, ...)
Freq DM (MoM)	-	-	(0.39, 0.16, 0.09, ...)	(0.20, 0.15, 0.12, ...)

5

Discussion

This paper reviews existing frequentist methods as well as introduces a new Bayesian framework for two group hypothesis testing for multivariate count data. This framework is motivated by recent advances in metagenomics and involves modeling multivariate samples from both groups as a hierarchical Dirichlet-multinomial model and then testing whether their underlying probabilities are the same. This model has the advantage of being able to decompose and differentiate the existing variation into between-group component and a cross-group component in an ANOVA-like fashion. This allows flexibility, as well as provides highly interpretable parameters. This new test also shows great computational performance. Using simulated data generating from a Dirichlet-multinomial distribution, we show that it performs closely to the frequentist counterpart. Future work will involve taking into account the microbial phylogeny to improve inference.

Appendix A

Transformation from $\mathbb{R}^{K-1}$ to Ω

Let

$$h(x) = \frac{x}{\sqrt{x^2+1}}, \quad x \in \mathbb{R},$$

where h is clearly a bijection from $\mathbb{R}$ to $(-1, 1)$. Let $\vec{p} = (p_1, p_2, \ldots, p_K) \in \Omega$ be a vector on the probability simplex. For any $\vec{u} = (u_1, u_2, \ldots, u_{K-1}) \in \mathbb{R}^{K-1}$, we can define the following transformation $T : \mathbb{R}^{K-1} \to \Omega$.

If $K = 2$, let T be the transformation

$$\begin{aligned} p(1) &= \frac{1}{2}(1 + h(u_1)) \\ p(2) &= \frac{1}{2}(1 - h(u_1)). \end{aligned}$$

If $K \geqslant 3$, let

$$\begin{aligned}
p(1) &= \frac{1}{2^2}(1+h(u_1))(1-h(u_2)) \\
p(2) &= \frac{1}{2^3}(1+h(u_1))(1+h(u_2))(1-h(u_3)) \\
p(3) &= \frac{1}{2^4}(1+h(u_1))(1+h(u_2))(1+h(u_3))(1-h(u_4)) \\
&\vdots \quad \vdots \quad \vdots \\
p(K-2) &= \frac{1}{2^{K-1}}(1+h(u_1))...(1+h(u_{K-2}))(1-h(u_{K-1})) \\
p(K-1) &= \frac{1}{2^{K-1}}(1+h(u_1))...(1+h(u_{K-2}))(1+h(u_{K-1})) \\
p(K) &= \frac{1-h(u_1)}{2}.
\end{aligned}$$

Under this transformation T, instead of integrating over Ω, we can instead integrate $\mathbb{R}^{K-1}$, which is significantly easier and much more manageable under Laplace transformation. In order to change the variables, we need to find the Jacobian J of the transformation T. For $i, j = 1, 2, \ldots, K-1$, define J by
if $j \leqslant i \leqslant K-1$,

$$J(i,j) = \frac{\partial p(i)}{\partial u_j} = p(i)\frac{h'(u_j)}{1+h(u_j)},$$

if $j = i+1$,

$$J(i,j) = \frac{\partial p(i)}{\partial u_j} = -p(i)\frac{h'(u_j)}{1-h(u_j)},$$

otherwise

$$J(i,j) = \frac{\partial p(i)}{\partial u_j} = 0.$$

Bibliography

Biagi, E., Nylund, L., Candela, M., Ostan, R., Bucci, L., Pini, E., Nikkla, J., Monti, D., Satokari, R., Franceschi, C., Brigidi, P., and De Vos, W. (2010), "Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians," *PloS one*, 5, e10667.

Fujimura, K. E., Slusher, N. A., Cabana, M. D., and Lynch, S. V. (2010), "Role of the gut microbiota in defining human health," *Expert Review of Anti-infective Therapy*, 8, 435–454.

Holmes, I., Harris, K., and Quince., C. (2012), "Dirichlet multinomial mixtures: generative models for microbial metagenomics," *PLoS one*, 7, e30126.

Koehler, K. and Wilson, J. (1986), "Chi-square tests for comparing vectors of proportions for several cluster samples," *Communications in Statistics: Theory and Methods*, 15, 2977–2990.

La Rosa, P., Brooks, J., Deych, E., Boone, E., Edwards, D., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W. (2012), "Hypothesis testing and power calculations for taxonomic-based human microbiome data," *PLoS one*, 7, e52078.

Ng, P. and Kirkness, E. (2010), "Whole genome sequencing," *Methods in Molecular Biology*, 628, 215–226.

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012), "Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples," *Genome Biology*, 13, R42.

Whittaker, R. (1965), "Dominance and diversity in land plant communities: numerical relations of species express the importance of competition in community function and evolution," *Science*, 147, 250–260.

Wilson, J. and Koehler, K. (1984), "Testing of equality of vectors of proportions for several cluster samples," *Proceedings of Joint Statistical Association Meetings Survey Research Methods.*

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., Nesse, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011), "Linking long-term dietary patterns with gut microbial enterotypes," *Science*, 334, 105–108.