Bioinformatics, 2024, **40(12)**, btae667
https://doi.org/10.1093/bioinformatics/btae667
Advance Access Publication Date: 19 November 2024
Original Paper

OXFORD

Genome analysis

Micro-DeMix: a mixture beta-multinomial model for investigating the heterogeneity of the stool microbiome compositions

Ruoqian Liu[1], **Yue Wang**[2,*], **Dan Cheng**[1]

[1]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85251, United States
[2]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO 80045, United States

*Corresponding author. Department of Biostatistics and Informatics, Colorado School of Public Health, 13001 E. 17th Place, Aurora, CO 80045, United States. E-mail: yue.2.wang@cuanschutz.edu

Associate Editor: Can Alkan

Abstract

Motivation: Extensive research has uncovered the critical role of the human gut microbiome in various aspects of health, including metabolism, nutrition, physiology, and immune function. Fecal microbiota is often used as a proxy for understanding the gut microbiome, but it represents an aggregate view, overlooking spatial variations across different gastrointestinal (GI) locations. Emerging studies with spatial microbiome data collected from specific GI regions offer a unique opportunity to better understand the spatial composition of the stool microbiome.

Results: We introduce Micro-DeMix, a mixture beta-multinomial model that deconvolutes the fecal microbiome at the compositional level by integrating stool samples with spatial microbiome data. Micro-DeMix facilitates the comparison of microbial compositions across different GI regions within the stool microbiome through a hypothesis-testing framework. We demonstrate the effectiveness and efficiency of Micro-DeMix using multiple simulated datasets and the inflammatory bowel disease data from the NIH Integrative Human Microbiome Project.

Availability and implementation: The R package is available at https://github.com/liuruoqian/MicroDemix.

1 Introduction

The human gut microbiome and its role in host health have been the subject of extensive research, establishing its involvement in human metabolism (Devaraj *et al.* 2013), nutrition (Zhang 2022), physiology (Andoh 2016), and immune function (Wu and Wu 2012, Bull and Plummer 2014). Recent studies have also demonstrated the association between the gut microbiota and the emergence of obesity (Aoun *et al.* 2020), metabolic syndrome and the onset of type 2 diabetes (Devaraj *et al.* 2013, Iatcu *et al.* 2021). Moreover, altered composition and function of the gut microbiota has been found associated with chronic diseases (Pellanda *et al.* 2021) ranging from gastrointestinal (GI) inflammatory (Shan *et al.* 2022) and metabolic conditions to neurological (Cryan *et al.* 2020), cardiovascular (Tang *et al.* 2017), and respiratory illnesses (Durack and Lynch 2019, Chunxi *et al.* 2020).

Most scientific findings regarding the gut microbiome have been derived from stool samples. However, it is increasingly recognized that the stool microbiome offers only an aggregate view, overlooking the spatial heterogeneity of the microbiome across different GI locations (Ahn *et al.* 2023, Levitan *et al.* 2023). Recent studies have revealed significant variations in microbial composition and function along distinct segments of the GI tract (Leite *et al.* 2020). For instance, the small intestine is dominated by Lactobacillaceae and Enterobacteriaceae, whereas the colon harbors species from families such as Bacteroidaceae, Prevotellaceae, Rikenellaceae, Lachnospiraceae, and Ruminococcaceae (Donaldson *et al.* 2016). Furthermore, spatial metagenomics applied to the mouse colonic microbiome has revealed a heterogeneous distribution of taxa throughout the gut (Sheth *et al.* 2019). Thus, understanding the spatial composition of the stool microbiome is essential for enhancing the biological relevance of stool-based microbiome analyses, allowing for more accurate interpretations of its role in health and disease. Although spatial microbiome data from specific GI locations remain relatively rare due to the invasive nature of sample collection, they offer a unique opportunity to deconvolute stool samples, thereby providing a deeper understanding of the gut microbiome's spatial heterogeneity.

We develop Micro-DeMix, a novel statistical model designed to deconvolute the fecal microbiome into specific GI locations by integrating spatial microbiome data. The development of Micro-DeMix was motivated by data from the inflammatory bowel disease (IBD) Multiomics Database (IBDMDB) project, part of the NIH Integrative Human Microbiome Project (iHMP). This project followed 132 individuals and collected 1785 stool samples and 651 intestinal biopsies over time (Integrative HMP *et al.* 2019). The integrated nature of this dataset, combining microbial profiles from both stool samples and distinct GI locations, highlighted the need for a tool like Micro-DeMix to effectively analyze and integrate these diverse sources of microbiome data.

Received: 15 December 2023; **Revised:** 29 October 2024; **Editorial Decision:** 31 October 2024; **Accepted:** 15 November 2024

Micro-DeMix is a mixture beta-multinomial model developed to deconvolute the fecal microbiome into specific GI locations. The multinomial component effectively handles the compositional nature of microbiome data by jointly modeling microbial proportions, while the beta component adjusts for sample heterogeneity, incorporating key demographic and clinical variables that may influence the composition of the fecal microbiome. We introduce two estimation procedures for the parameters in the Micro-DeMix model. The first procedure directly maximizes the likelihood of the observed microbial counts, offering computational efficiency. This approach is also equipped with an efficient hypothesis-testing framework for comparing microbial compositions from different GI regions within the mixed stool sample based on the asymptotic null distribution. The second procedure is based on the Expectation-Maximization (EM) algorithm, which provides greater numerical robustness than the first method. It is particularly useful for scenarios with extremely rare microbes, where the first procedure may encounter limitations, although it is computationally more intensive. This EM approach includes a permutation test to identify differentially abundant microbial groups across GI locations. Unlike conventional differential abundance (DA) tools, which compare microbial abundances between predefined groups (e.g. disease versus healthy or across body sites), Micro-DeMix is specifically designed to detect compositional changes within a mixed population (e.g. the stool sample), where the groups are not prespecified and must be statistically estimated. This enables Micro-DeMix to uncover group differences and spatial heterogeneity within the stool microbiome by integrating spatial microbiome data. Traditional DA tools, which rely on predefined group distinctions, are less suited to detect these types of within-sample variations. We demonstrate the effectiveness of Micro-DeMix through extensive simulation studies and an application to IBD cohorts, where we integrate stool microbiome data with rectum microbiome data.

The rest of the paper is organized as follows. In Section 2, we describe the Micro-DeMix model, develop two maximum-likelihood-based estimation procedures, and propose hypothesis-testing frameworks for detecting differentially abundant microbes between GI locations. In Section 3, we conduct simulation studies to demonstrate the effectiveness of Micro-DeMix in terms of the estimation accuracy and the power of the hypothesis testing method. In Section 4, we apply our method to the IBD microbiome data from iHMP to elucidate the composition of fecal microbiome in IBD patients. We close with a discussion of our method in Section 5.

2 Materials and methods

2.1 The Micro-DeMix model

In this section, we present a beta-multinomial framework for modeling microbial abundance in stool samples. Our model considers a multivariate extension of the beta-binomial model proposed in a recent research (Martin *et al.* 2020) which models individual taxa separately. As shown later, while this multivariate extension allows for joint modeling of multiple microbes, it also brings several new statistical and computational challenges.

For ease of presentation and biological relevance, we assume the reference spatial microbiome data are collected at the rectum. However, the methodology proposed below can seamlessly extend to other segments of the GI tract, including the ileum, jejunum, colon, and beyond, given the data availability. Let $y_{jg}^{(r)}$ denotes the observed absolute abundance for taxon g in the jth subject of the reference rectum microbiome data for $j = 1, \ldots, S^{(r)}$. Let y_{ig} denotes the observed absolute abundance for taxon g in the ith stool samples for $g = 1, \ldots, G$ and $i = 1, \ldots, S$. For all subjects, we also collect covariate information $\mathbf{x}_j^{(r)}$ and $\mathbf{x}_i$ for $j = 1, \ldots, S^{(r)}$ and $i = 1, \ldots, S$. We model the stool microbial abundance $\mathbf{y}_i = (y_{i1}, \ldots, y_{iG})^\top$ using a multinomial model $\mathbf{y}_i | \boldsymbol{p}_i \sim \text{Multinomial}(N_i, \boldsymbol{p}_i)$, where $N_i = \sum_{g=1}^{G} y_{ig}$ and $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iG})^\top$ is the underlying true microbial proportions. We are particularly interested in the spatial composition of the stool microbiome. Acknowledging that the stool microbiome consists of microbes from rectum and other GI locations (mostly, the small intestine), we consider

$$p_{ig} = \pi_i p_g^{(r)} + (1 - \pi_i) p_g^{(o)}; \tag{1}$$

here, $\pi_i \in (0, 1)$ denotes the latent proportion of the rectal microbiome in the ith stool sample; $p_g^{(r)}$ and $p_g^{(o)}$ denotes the proportion of taxon g in the rectum and other GI locations, respectively.

To allow π_i to vary across individuals, we assume $\pi_i \sim \text{Beta}(a_{1,i}, a_{2,i})$. Hence, we write the joint density function of (y_{ig}, π_i) as

$$f(y_{ig}, \pi_i | p_g^{(r)}, p_g^{(o)}, a_{1,i}, a_{2,i}) = \frac{N_i!}{y_{i1}! \ldots y_{iG}!} \prod_{g=1}^{G} \{\pi_i p_g^{(r)} + (1 - \pi_i) p_g^{(o)}\}^{y_{ig}} \frac{\pi_i^{a_{1,i}-1} (1 - \pi_i)^{a_{2,i}-1}}{B(a_{1,i}, a_{2,i})}, \tag{2}$$

where $B(a_{1,i}, a_{2,i}) = \Gamma(a_{1,i})\Gamma(a_{2,i})/\Gamma(a_{1,i} + a_{2,i})$

To capture the association between π_i and the subjects' characteristics, we further link each π_i with covariates $\mathbf{x}_i$ through a beta-regression model. With additional reparameterization

$$\mu_i = \frac{a_{1,i}}{a_{1,i} + a_{2,i}}, \tag{3}$$

$$\phi_i = a_{1,i} + a_{2,i}, \tag{4}$$

we consider

$$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{5}$$

$$\log(\phi_i) = \gamma_0 + \mathbf{x}_i^\top \boldsymbol{\gamma}. \tag{6}$$

Some calculations yield that μ_i is the expected proportion of the rectum microbiome in the ith stool sample, i.e. $\mu_i = \mathbb{E}(\pi_i)$, and $\text{Var}(\pi_i) = \mu_i(1 - \mu_i)/(1 + \phi_i)$, where ϕ_i is known as the precision parameter.

Thus, our first goal is to estimate each $p_g^{(r)}$ and $p_g^{(o)}$ to understand the spatial heterogeneity of the stool microbiome. Then, we aim to identify microbial groups that are differentially abundant between the rectum and other GI locations with the stool sample. Below we will present two procedures

for these two analyses, each with its own strengths for different types of applications.

Remark 1. *Under the proposed model, we derive the mean and variance of* y_{ig}:

$$\mathbb{E}(y_{ig}) = N_i\left[p_g^{(r)}\frac{a_{1,i}}{a_{1,i}+a_{2,i}} + p_g^{(o)}\left(1-\frac{a_{1,i}}{a_{1,i}+a_{2,i}}\right)\right]$$

$$\begin{aligned} Var(y_{ig}) = N_i\Bigg\{ & \frac{a_{1,i}}{a_{1,i}+a_{2,i}}\left(p_g^{(r)} - 2p_g^{(r)}p_g^{(o)} - p_g^{(o)} + 2\left(p_g^{(o)}\right)^2\right) \\ & + \left(\frac{a_{1,i}^2}{(a_{1,i}+a_{2,i})^2} + \frac{a_{1,i}a_{2,i}}{(a_{1,i}+a_{2,i})^2(a_{1,i}+a_{2,i}+1)}\right) \\ & \left(-\left(p_g^{(r)}\right)^2 + 2p_g^{(r)}p_g^{(o)} - \left(p_g^{(o)}\right)^2\right) + p_g^{(o)} - \left(p_g^{(o)}\right)^2\Bigg\} \\ & + N_i^2\left(p_g^{(r)} - p_g^{(o)}\right)^2\frac{a_{1,i}a_{2,i}}{(a_{1,i}+a_{2,i})^2(a_{1,i}+a_{2,i}+1)}. \end{aligned}$$

See derivations in the Supplementary Material.

2.2 Micro-DeMix-1

2.2.1 Estimation

The joint likelihood in Equation (2) involves π_i, which is unobserved. Thus, we calculate the log-likelihood of observing $\{y_{ig} : i = 1,\ldots,S; g = 1,\ldots,G\}$ in stool samples:

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{S}\log\int_0^1 f(y_{ig},\pi_i|\boldsymbol{\theta})d\pi_i, \tag{7}$$

where $\boldsymbol{\theta} = (\boldsymbol{p}^{(r)},\boldsymbol{p}^{(o)},\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma})^\top$. We estimate $\boldsymbol{\theta}$ in two steps. We first estimate $p_g^{(r)}$ using the reference rectum microbiome data for $g = 1,\ldots,G$. Recall that $y_{jg}^{(r)}$ denotes the observed absolute abundance for taxon g and sample j, $j = 1,\ldots,S^{(r)}$, collected from rectum. Let $N_j^{(r)}$ denotes the sequencing depth for the jth sample from rectum. We assume the absolute abundance to follow a multinomial distribution

$$\mathbf{y}_j^{(r)} \sim \text{Multinomial}(N_j^{(r)},\boldsymbol{p}^{(r)}). \tag{8}$$

Based on this model, we estimate $p_g^{(r)}$ with the sample proportions

$$\hat{p}_g^{(r)} = \frac{\sum_j y_{jg}^{(r)}}{\sum_j N_j^{(r)}}. \tag{9}$$

After replacing $p_g^{(r)}$ in Equation (7) by $\hat{p}_g^{(r)}$, we estimate the remaining parameters using an iterative procedure with the main steps given in Algorithm 1.Specifically, in the mth iteration, we obtain $\{p_g^{(o)}\}^m$ by maximizing the objective function (7) given $\{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^{m-1}$. Since the integral in Equation (7) does not have a closed form, we used the Gauss–Legendre quadrature (Golub and Welsch 1969) to approximate it. Solving for $\{p_g^{(o)}\}^m$ is a constrained optimization problem because $p_g^{(o)} \in [0,1]$ and $\sum_{g=1}^G p_g^{(o)} = 1$. To address this problem, we consider the parameterization

Algorithm 1. Estimating $\{p_g^{(o)},\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}$ for Micro-DeMix

Require: $y_{ig}, \mathbf{x}_i$, an integer M, a small number σ (e.g. $\sigma = 0.001$)
1: **Initialize:** $\{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\} = \{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^0$
2: **for** $m = 1,\ldots,M$ **do**
3: Update $\{p_g^{(o)}\}^m = \text{argmax}\log\boldsymbol{L}(\boldsymbol{\theta})$ as in (7) given $\{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\} = \{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^{m-1}$.
4: Update $\{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^m = \text{argmax}\log\boldsymbol{L}(\boldsymbol{\theta})$ as in (7) given $\{p_g^{(o)}\} = \{p_g^{(o)}\}^m$. Let L_m denote the value of (7) at this step.
5: **if** $|L_m - L_{m-1}| \le \sigma$ **then**
6: **return** $\hat{\boldsymbol{\theta}} = \{p_g^{(o)},\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^m$
7: **end if**
8: **end for**
9: **return** $\hat{\boldsymbol{\theta}} = \{p_g^{(o)},\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^M$

$$p_g^{(o)} = \frac{e^{u_g}}{\sum_{k=1}^G e^{u_k}}, \tag{10}$$

where $u_g \in \mathbb{R}$ for $g = 1,\ldots,G$. With this reparametrization, we first obtain $\{u_g\}^m$ using the Nelder–Mead optimization algorithm from `R` package `optimx` and then calculate $\{p_g^{(o)}\}^m$ from $\{u_g\}^m$ using Equation (10). Finally, given $\{p_g^{(o)}\}^m$, we obtain the optimal set $\{\beta_0,\boldsymbol{\beta},\gamma_0,\boldsymbol{\gamma}\}^m$ using the Nelder–Mead optimization algorithm by maximizing log-likelihood function (7). We repeat this iterative procedure until we achieve convergence or reach the maximum number of iterations.

2.2.2 Hypothesis testing

The proposed Micro-DeMix framework also facilitates the investigation of the biodiversity of the fecal microbiome by detecting microbial groups that are differentially abundant in rectum and other GI locations. Specifically, we consider a hypothesis testing problem with $H_0 : \boldsymbol{p}^{(o)} = \boldsymbol{p}^{(r)}$, where $\boldsymbol{p}^{(o)}$ and $\boldsymbol{p}^{(r)}$ are introduced in Section 2.1, representing the true proportions of the microbes of interest in the rectum and other GI locations, respectively.

Under H_0, we have $||\boldsymbol{p}^{(o)} - \boldsymbol{p}^{(r)}||_2 = 0$, motivating the following test statistic:

$$\hat{T} = (\hat{\boldsymbol{p}}^{(o)} - \hat{\boldsymbol{p}}^{(r)})^\top(\hat{\boldsymbol{p}}^{(o)} - \hat{\boldsymbol{p}}^{(r)}), \tag{11}$$

where $\hat{\boldsymbol{p}}^{(o)}$ are maximum likelihood estimators obtained from Algorithm 1 in Section 2.2, and $\hat{\boldsymbol{p}}^{(r)}$ is given in Equation (9). Under $H_0 : p_g^{(r)} = p_g^{(o)}$, we can rewrite Equation (11) as

$$\begin{aligned} \hat{T} = & \sum_{g=1}^G(\hat{p}_g^{(r)} - p_g^{(r)})^2 + \sum_{g=1}^G(\hat{p}_g^{(o)} - p_g^{(o)})^2 \\ & - 2\sum_{g=1}^G(\hat{p}_g^{(r)} - p_g^{(r)})(\hat{p}_g^{(o)} - p_g^{(o)}), \end{aligned} \tag{12}$$

indicating that $\hat{T}$ is a function of $\hat{\boldsymbol{p}}^{(r)} - \boldsymbol{p}^{(r)}$ and $\hat{\boldsymbol{p}}^{(o)} - \boldsymbol{p}^{(o)}$ under H_0. Thus, finding the null distribution of $\hat{T}$ reduces to finding the null distributions of $\hat{\boldsymbol{p}}^{(r)} - \boldsymbol{p}^{(r)}$ and $\hat{\boldsymbol{p}}^{(o)} - \boldsymbol{p}^{(o)}$. Specifically, applying the central limit theorem (CLT) based on model (8), we have

$$\sqrt{N^{(r)}}(\hat{\boldsymbol{p}}^{(r)} - \boldsymbol{p}^{(r)}) \xrightarrow{d} N_G(\mathbf{0}, V(\boldsymbol{p}^{(r)})), \tag{13}$$

where

$$V(\boldsymbol{p}^{(r)}) = \begin{pmatrix} p_1^{(r)}(1-p_1^{(r)}) & -p_1^{(r)}p_2^{(r)} & \dots & -p_1^{(r)}p_G^{(r)} \\ -p_1^{(r)}p_2^{(r)} & p_2^{(r)}(1-p_2^{(r)}) & \dots & -p_2^{(r)}p_G^{(r)} \\ \vdots & & \ddots & \vdots \\ -p_1^{(r)}p_G^{(r)} & -p_2^{(r)}p_G^{(r)} & \dots & p_G^{(r)}(1-p_G^{(r)}) \end{pmatrix}$$

and $N^{(r)} = \sum_{j=1}^{S^{(r)}} N_j^{(r)}$. Also, under H_0, the joint density function in Equation (2) reduces to

$$f(y_{ig}, \pi_i | p_g^{(o)}, a_{1,i}, a_{2,i}) = \frac{N_i!}{y_{i1}! \dots y_{iG}!} \prod_{g=1}^{G} \{p_g^{(o)}\}^{y_{ig}} \frac{\pi_1^{a_{1,i}-1}(1-\pi_i)^{a_{2,i}-1}}{\boldsymbol{B}(a_{1,i}, a_{2,i})}$$

and the density function (2) becomes

$$f(y_{ig}|\cdot) = \frac{N_i!}{y_{i1}! \dots y_{iG}!} \prod_{g=1}^{G} \{p_g^{(o)}\}^{y_{ig}} \int_0^1 \frac{\pi_1^{a_{1,i}-1}(1-\pi_i)^{a_{2,i}-1}}{\boldsymbol{B}(a_{1,i}, a_{2,i})} d\pi_i.$$

As the integral of the beta density is 1, we have

$$f(y_{ig} | p_g^{(o)}, a_{1,i}, a_{2,i}) = \frac{N_i!}{y_{i1}! \dots y_{iG}!} \prod_{g=1}^{G} \{p_g^{(o)}\}^{y_{ig}}.$$

This indicates that under H_0, the proposed beta-multinomial model in Section 2.1 reduces to a multinomial model. Indeed, when $p_g^{(r)} = p_g^{(o)}$, $\mathrm{Var}(y_{ig})$ in Remark 1 simplifies to $N_i p_g^{(o)}(1-p_g^{(o)})$, which equals the variance of y_{ig} under the multinomial model. Thus, under H_0, the maximum likelihood theory (MLE) derived from the mixture beta-multinomial model is equivalent to that derived from the multinomial model. Thus, we can apply the CLT again and obtain

$$\sqrt{N}(\hat{\boldsymbol{p}}^{(o)} - \boldsymbol{p}^{(o)}) \xrightarrow{d} N_G(\mathbf{0}, V(\boldsymbol{p}^{(r)})), \tag{14}$$

where $N = \sum_{i=1}^{S} N_i$ is the total counts of all samples collected from stool.

Assuming no overlapping between stool samples and rectum samples, we establish the asymptotic distribution of $\hat{T}$ in the following result.

Proposition 1. *Let random vector* $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ *follow the multivariate normal distribution*

$$N_{2G}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V(\boldsymbol{p}^{(r)}) & 0 \\ 0 & V(\boldsymbol{p}^{(r)}) \end{pmatrix} \right).$$

Suppose $\boldsymbol{p}^{(o)} = \boldsymbol{p}^{(r)}$ *and* $N^{(r)}/N \to K > 0$. *Then* $N\hat{T} \xrightarrow{d} g(\boldsymbol{y}_1, \boldsymbol{y}_2)$ *as* $N \to \infty$, *where* $g(\boldsymbol{y}_1, \boldsymbol{y}_2) = \boldsymbol{y}_1^\top \boldsymbol{y}_1 + (1/K)\boldsymbol{y}_2^\top \boldsymbol{y}_2 - (2/\sqrt{K})\boldsymbol{y}_1^\top \boldsymbol{y}_2$.

Algorithm 2. Simulation method of testing $H_0 : \boldsymbol{p}^{(o)} = \boldsymbol{p}^{(r)}$

Require: $y_{jg}^{(r)}, y_{ig}, \mathbf{x}_i$, a large integer B (e.g. $B = 10\,000$)
1: Estimate $\hat{\boldsymbol{p}}^{(o)}$ and $\hat{\boldsymbol{p}}^{(r)}$ as in Section 2.2.
2: Compute $\hat{T}$ as in (11) using $\hat{\boldsymbol{p}}^{(o)}$ and $\hat{\boldsymbol{p}}^{(r)}$.
3: **for** $b = 1, \dots, B$ **do**
4: Simulate vector $\boldsymbol{v}_1, \boldsymbol{v}_2$ independently from $N_G(0, V(\hat{\boldsymbol{p}}^{(r)}))$ in (13) and (14).
5: Approximate $\hat{\boldsymbol{p}}^{(r)} - \boldsymbol{p}^{(r)}$ and $\hat{\boldsymbol{p}}^{(o)} - \boldsymbol{p}^{(o)}$ with $\boldsymbol{v}_1/\sqrt{N^{(r)}}$ and $\boldsymbol{v}_2/\sqrt{N}$, respectively. Compute $\hat{T}^b$ as in (12).
6: **end for**
7: Calculate the p-value:

$$\hat{p} \leftarrow \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} 1\{\hat{T}^b \geq \hat{T}\}\right).$$

8: **return** $\hat{p}$

As a special case of Proposition 1, when $K = 1$ or $N^{(r)} = N$, we have $N\hat{T} \xrightarrow{d} g(\boldsymbol{y}_1, \boldsymbol{y}_2) = \boldsymbol{y}_1^\top \boldsymbol{y}_1 + \boldsymbol{y}_2^\top \boldsymbol{y}_2 - 2\boldsymbol{y}_1^\top \boldsymbol{y}_2$. Based on Proposition 1, we develop a simulation-based algorithm for calculating the p-value, as detailed in Algorithm 2.

2.3 Micro-DeMix-2: an EM approach

The proposed Micro-DeMix-1 procedure has strengths in easy and efficient computation: its estimation component can be implemented using standard optimization packages, and its hypothesis-testing component runs quickly due to the use of an asymptotic null distribution. However, it may have a limited scope of applicability due to the numerical evaluation of the log-likelihood in Equation (7). Specifically, based on Equation (2), we know that when $\{\pi_i p_g^{(r)} + (1-\pi_i)p_g^{(o)}\}^{y_{ig}}$ is close to 0, which can occur for rare microbes with extremely small proportions or more common microbes with very large counts, Equation (7) may be evaluated as negative infinity. This happens because, although $\{\pi_i p_g^{(r)} + (1-\pi_i)p_g^{(o)}\}^{y_{ig}}$ is technically not zero, it falls below the smallest representable number in standard software.

To address this issue, instead of working with the marginal density function of y_{ig} in Equation (7), we develop a new estimation procedure based on the joint density function of y_{ig} and π_i, as shown in Equation (2). Specifically, the joint log-likelihood is given by

$$\begin{aligned} l_{\text{joint}}(\boldsymbol{\theta}; \mathbf{Y}, \boldsymbol{\pi}) &= \sum_{i=1}^{S} \log f(\boldsymbol{y}_i, \pi_i | \boldsymbol{p}^{(r)}, \boldsymbol{p}^{(o)}, a_{1,i}, a_{2,i}) \\ &= \sum_{i=1}^{S}\sum_{g=1}^{G} y_{ig} \log(\pi_i p_g^{(r)} + (1-\pi_i)p_g^{(o)}) \\ &\quad + \sum_{i=1}^{S}(a_{1,i}-1)\log(\pi_i) + \sum_{i=1}^{S}(a_{2,i}-1)\log(1-\pi_i) \\ &\quad - \sum_{i=1}^{S}\log\left(B(a_{1,i}, a_{2,i})\right) + C, \end{aligned} \tag{15}$$

where C is some constant, $\mathbf{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_S)^\top$, and $\boldsymbol{\pi} = (\pi_1,$

Algorithm 3. Estimating $\{p_g^{(o)}, \beta_0, \boldsymbol{\beta}, \gamma_0, \boldsymbol{\gamma}\}$: an EM approach

Require: $y_{ig}, \mathbf{x}_i$, an integer K, an integer M, a small number σ (e.g. $\sigma = 0.0001$)
1: **Initialize:** $\{p_g^{(o)}, \beta_0, \boldsymbol{\beta}, \gamma_0, \boldsymbol{\gamma}\} = \{p_g^{(o)}, \beta_0, \boldsymbol{\beta}, \gamma_0, \boldsymbol{\gamma}\}^0$
2: **for** $k = 1, \ldots, K$ **do**
3: (E-step) Generate a random sample $\tilde{\boldsymbol{\pi}}_m^{(k)}$ from the conditional distribution of $\boldsymbol{\pi}$ given $\boldsymbol{Y}$ and $\boldsymbol{\theta}^{(k-1)}$ using the Metropolis–Hastings (MH) algorithm. Approximate the conditional expectation $Q_k(\boldsymbol{\theta}) = E[l_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{\pi}) | \boldsymbol{\theta}^{(k-1)}, \mathbf{y}_i : i = 1, \ldots, S]$ using $\tilde{Q}_k(\boldsymbol{\theta}) = M^{-1} \sum_{m=1}^{M} l_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \tilde{\boldsymbol{\pi}}_m^{(k)})$,
4: (M-step) Obtain $\boldsymbol{\theta}^{(k)}$ by maximizing the objective function $\tilde{Q}_k(\boldsymbol{\theta})$.
5: **if** $\sum \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\| \leq \sigma$ **then**
6: **return** $\hat{\boldsymbol{\theta}} = \{p_g^{(o)}, \beta_0, \boldsymbol{\beta}, \gamma_0, \boldsymbol{\gamma}\}^k$
7: **end if**
8: **end for**
9: **return** $\hat{\boldsymbol{\theta}} = \{p_g^{(o)}, \beta_0, \boldsymbol{\beta}, \gamma_0, \boldsymbol{\gamma}\}^K$

$\ldots, \pi_S)^\top$. Thanks to the logarithm, Equation (15) is always well defined, regardless of how small $\{\pi_i p_g^{(r)} + (1 - \pi_i) p_g^{(o)}\}^{y_{ig}}$ becomes. However, we cannot directly maximize Equation (15) because the π_i values are not observed. To address this, we develop an EM algorithm (Dempster *et al.* 1977) for estimating $\boldsymbol{\theta}$ based on Equation (15). We begin with an initial value $\boldsymbol{\theta}_0$. In the kth E-step, we compute $Q_k(\boldsymbol{\theta}) = E[l_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{\pi}) | \boldsymbol{\theta}_{k-1}, \mathbf{y}_i : i = 1, \ldots, S]$, which is the conditional expectation of l_{joint} given $\mathbf{y}_i : i = 1, \ldots, S$ and the values of $\boldsymbol{\theta}$ from the previous $(k-1)$th step for $k \geq 1$. Unfortunately, $Q_k(\boldsymbol{\theta})$ does not have a closed form. Thus, we approximate $Q_k(\boldsymbol{\theta})$ using $\tilde{Q}_k(\boldsymbol{\theta}) = M^{-1} \sum_{m=1}^{M} l_{\text{joint}}(\boldsymbol{\theta}, \boldsymbol{Y}, \tilde{\boldsymbol{\pi}}_m^{(k)})$, where $\tilde{\boldsymbol{\pi}}_m^{(k)}$ is a random sample from the conditional distribution of $\boldsymbol{\pi}$ given $\boldsymbol{Y}$ and $\boldsymbol{\theta}_{k-1}$, generated using the Metropolis–Hastings (MH) algorithm (Hastings 1970). In the kth M-step, we then obtain $\boldsymbol{\theta}_k$ by maximizing $\tilde{Q}_k(\boldsymbol{\theta})$ with existing optimization tools such as the Nelder–Mead algorithm. The final estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{p}}^{(r)}, \hat{\boldsymbol{p}}^{(o)}, \hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\gamma}_0, \hat{\boldsymbol{\gamma}})^\top$ is obtained by iterating the EM algorithm between the E-step and M-step until convergence.Due to the additional variation introduced by the EM algorithm, Algorithm 2 is no longer suitable for generating p-values for testing $H_0 : \boldsymbol{p}^{(o)} = \boldsymbol{p}^{(r)}$ based on the EM estimates. We fix this issue by developing a permutation-based procedure that works seamlessly with the EM algorithm; see Algorithm 4. Recall that $y_{jg}^{(r)}$ denotes the absolute abundance for taxon g in the jth subject of the rectum dataset and $\mathbf{x}_j^{(r)}$ is the corresponding covariate. Letting $\boldsymbol{y}_j^{(r)} = (y_{j1}^{(r)}, \ldots, y_{jG}^{(r)})^\top$, we first randomly select S individuals from $\{\boldsymbol{y}_1^{(r)}, \ldots, \boldsymbol{y}_{S_r}^{(r)}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_S\}$ to form a permuted stool microbiome dataset, with the remaining S_r subjects forming a permuted rectum microbiome dataset. We also obtain the permuted covariates corresponding to the permuted stool sample. Next, we use the permuted rectum sample to estimate $\boldsymbol{p}^{(r)}$ and apply the EM algorithm to the permuted stool microbiome data and covariates to obtain an estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}_{\text{perm}} = (\hat{\boldsymbol{p}}_{\text{perm}}^{(r)}, \hat{\boldsymbol{p}}_{\text{perm}}^{(o)}, \hat{\beta}_{0,\text{perm}}, \hat{\boldsymbol{\beta}}_{\text{perm}}, \hat{\gamma}_{0,\text{perm}}, \hat{\boldsymbol{\gamma}}_{\text{perm}})^\top$. We then calculate the permuted test statistic $T_{\text{perm}} = \|\hat{\boldsymbol{p}}_{\text{perm}}^{(r)} - \hat{\boldsymbol{p}}_{\text{perm}}^{(o)}\|^2$. This permutation procedure is repeated B times to generate B permuted test statistics $T_{\text{perm}}^{(1)}, \ldots, T_{\text{perm}}^{(B)}$. Finally, the permutation p-value for testing $H_0 : \boldsymbol{p}^{(r)} = \boldsymbol{p}^{(o)}$ is defined as the proportion of $T_{\text{perm}}^{(1)}, \ldots, T_{\text{perm}}^{(B)}$ that are larger than or equal to the unpermuted test statistic $T = \|\hat{\boldsymbol{p}}^{(r)} - \hat{\boldsymbol{p}}^{(o)}\|^2$.Compared to Micro-DeMix-1, the proposed EM algorithm overcomes the numerical issues caused by rare microbes or large counts, giving it broader applicability. However, the EM algorithm is more computationally intensive due to its relatively slow convergence rate and the repeated use of the MH algorithm in the E-step. Additionally, the permutation test requires running the EM algorithm multiple times, making it less efficient than Algorithm 2, which relies on the asymptotic null distribution. Furthermore, as shown in the simulation studies below, Micro-DeMix-1 may have higher power than the EM algorithm in numerical applications.

Algorithm 4. Permutation method of testing $H_0 : \boldsymbol{p}^{(o)} = \boldsymbol{p}^{(r)}$

Require: $y_{jg}^{(r)}, y_{ig}, \mathbf{x}_i, \mathbf{x}_j^{(r)}$, a large integer B
1: Estimate $\hat{\boldsymbol{p}}^{(r)}$ and $\hat{\boldsymbol{p}}^{(o)}$ using Algorithm 3 in Section 2.3.
2: Compute $\hat{T}$ as in (11) using $\hat{\boldsymbol{p}}^{(o)}$ and $\hat{\boldsymbol{p}}^{(r)}$.
3: **for** $b = 1, \ldots, B$ **do**
4: Randomly permute data $\{(\boldsymbol{y}_j^{(r)}, \mathbf{x}_j^{(r)})\}_{j=1}^{S^{(r)}}, \{(\boldsymbol{y}_i, \mathbf{x}_i)\}_{i=1}^{S}$ to generate permuted rectum data and permuted stool data $\{(\tilde{\boldsymbol{y}}_j^{(r)}, \tilde{\mathbf{x}}_j^{(r)})\}_{j=1}^{S^{(r)}}, \{(\tilde{\boldsymbol{y}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^{S}$.
5: Estimate $\hat{\boldsymbol{p}}_{\text{perm}}^{(r)}$ and $\hat{\boldsymbol{p}}_{\text{perm}}^{(o)}$ using permutated data $\{(\tilde{\boldsymbol{y}}_j^{(r)}, \tilde{\mathbf{x}}_j^{(r)})\}_{j=1}^{S^{(r)}}, \{(\tilde{\boldsymbol{y}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^{S}$.
6: Compute $\hat{T}_{\text{perm}}^{(b)} = \|\hat{\boldsymbol{p}}_{\text{perm}}^{(r)} - \hat{\boldsymbol{p}}_{\text{perm}}^{(o)}\|^2$.
7: **end for**
8: Calculate the p-value:

$$\hat{p} \leftarrow \frac{1}{1+B} \left(1 + \sum_{b=1}^{B} 1\{\hat{T}_{\text{perm}}^{(b)} \geq \hat{T}\} \right).$$

9: **return** $\hat{p}$

3 Results

We first investigate the finite-sample performance of Micro-DeMix using simulation. We studied the accuracy in the estimation of $\boldsymbol{p}^{(o)}$ and the type-I error rate and power when testing for differential abundance between the gut microbiome in the rectum and other GI locations.

For rectum samples, we simulated the microbial counts $\mathbf{y}_j^{(r)}$ from a Multinomial$(N_j^{(r)}, \boldsymbol{p}^{(r)})$ distribution with $p_g^{(r)} = 1/G$ for $g = 1, \ldots, G$ and $j = 1, \ldots, S^{(r)}$. We next generated the stool microbiome data using the proposed beta-multinomial model. Specifically, we first set $(\beta_0, \beta, \gamma_0, \gamma) = (0.1, 0.1, 0.2, 0.1)$ and generated the covariate x_i from the standard normal distribution. We then randomly generated π_i from Beta$(a_{1,i}, a_{2,i})$, where $a_{1,i}$ and $a_{2,i}$ are defined in Equations (3)–(6). We assigned u_g at evenly spaced values from -1 to 1 for $g = 1, \ldots, G-1$ and let $u_G = -\sum_{g=1}^{G-1} u_g$. We computed $p_g^{(o)}$ from u_g according to Equation (10). Finally, we simulated the stool microbiome data $\mathbf{y}_i$ from a Multinomial$(N_i, \boldsymbol{p}_i)$ distribution where $\boldsymbol{p}_i$ is defined in Equation (1) for $i = 1, \ldots, S$.

Table 1. Mean and standard deviation of RSE obtained under each setting while estimating $\boldsymbol{p}^{(o)}$ in Micro-DeMix.

	G = 5			G = 10		
Library size	500	1000	5000	500	1000	5000
Mean	0.00049	0.00035	0.00014	0.00557	0.00501	0.00364
Stand. Dev.	0.00045	0.00035	0.00027	0.00261	0.00201	0.00111

3.1 Simulation: Estimation

We assessed the performance of Micro-DeMix in estimating $\boldsymbol{p}^{(o)}$ under six settings, where we considered $S = S^{(r)} = 100$, $N_j^{(r)} = N_i = 500, 1000, 5000$ for all i and j, and $G = 5, 10$. Given the small scale of this simulation and the efficiency of the Micro-DeMix-1 algorithm, we only implemented the Micro-DeMix-1 algorithm in this section. We reported relative squared errors (RSE) to quantify the discrepancy between our Micro-DeMix estimators $\hat{p}^{(o)}$ and the true parameters according to

$$RSE = \frac{\sum_g (p_g^{(o)} - \hat{p}_g^{(o)})^2}{\sum_g (p_g^{(o)})^2}.$$

The mean and standard deviation of RSE under all six settings are presented in Table 1. It was observed that as the library size increases, both the mean and standard deviation of the RSE tend to decrease. This indicates that the accuracy of estimation improves as we increase the library size. Furthermore, it is noticeable that the RSEs for $G = 10$ are greater than those for $G = 5$. This phenomenon occurs because the true proportions become smaller as the total number of taxa increases. Specifically, by design, $p_g^{(o)} = \{0.059, 0.116, 0.225, 0.439, 0.161\}$ when $G = 5$, and $p_g^{(o)} = \{0.031, 0.039, 0.051, 0.065, 0.083, 0.107, 0.137, 0.177, 0.227, 0.083\}$ when $G = 10$. This makes estimation more challenging and contributes to a decrease in relative accuracy.

We also examined the performance of Micro-DeMix-EM using a larger-scale simulation study with $G = 100$, as detailed in the Supplementary Material.

3.2 Simulation: Type-I error rate and power

In this subsection, we assessed the empirical type-I error rate and power of both Micro-DeMix testing procedures to detect differential abundance between rectal microbes and microbes from other GI locations. Accordingly, we increased the total number of microbes to $G = 10, 20$. The data-generating process was the same as that in the previous section except for a few modifications to facilitate the assessment of the type-I error rate and power. Specifically, we set $(\beta_0, \beta, \gamma_0, \gamma) = (0.1, 0.1, 0.7, 0.2)$ and generated the covariate x_i from a normal distribution $N(10, 0.1)$. For each G, we generated $\boldsymbol{p}^{(r)}$ by allowing each $p_g^{(r)}$ and $p_g^{(o)}$ to differ by δ: $p_g^{(r)} = p_g^{(o)} + \delta$ for $g = 1, \ldots, G/2$, and $p_g^{(r)} = p_g^{(o)} - \delta$ for $g = G/2 + 1, \ldots, G$. This process ensures that $\sum_{g=1}^{G} p_g^{(r)} = 1$ for any δ. We considered relatively weak signals, i.e. $\delta = 0, 0.0004, 0.0008, \ldots, 0.002$, for $G = 10$, and further decreased the signal strength by considering $\delta = 0, 0.0002, 0.0004, \ldots, 0.001$ for $G = 20$. To ensure a large enough sample size and library size for signal detection, we considered $S = S^{(r)} = 100$, $N_j^{(r)} = N_i = 30\,000$ for $G = 10$ and $N_j^{(r)} = N_i = 35\,000$ for $G = 20$.

We compared the performance of our method with seven existing differential abundance tools for microbiome data, including ANCOM-BC (Mandal *et al.* 2015), DESeq2 (Love *et al.* 2014), edgeR (Robinson *et al.* 2010), limma voom (Ritchie *et al.* 2015), MaAsLin2 (Mallick *et al.* 2020), *t*-test, and Wilcoxon test, using simulated datasets with two sample groups. The Wilcoxon test and *t*-test were conducted using the R functions `wilcox.test` and `t.test`, respectively. ANCOM-BC, which analyzes microbiome compositions, was performed using the `ancombc` function from the R package `ANCOMBC`. EdgeR, DESeq2, limma voom, and MaAsLin2 were implemented based on microbial counts using the R packages `edgeR`, `DESeq2`, `limma`, and `Maaslin2`, respectively.

However, all these existing methods perform separate univariate tests for individual taxa, while our proposed Micro-DeMix focuses on global testing for a group of microbes. To facilitate a meaningful comparison, we applied Bonferroni correction to the univariate *p*-values generated by each existing method to control the family-level type-I error rate, rejecting the global null hypothesis if at least one adjusted *p*-value was below the significance level of $\alpha = 0.05$.

We implemented the proposed Micro-DeMix testing procedure (detailed in Algorithm 2) to produce a *p*-value for testing $H_0 : \boldsymbol{p}^{(r)} = \boldsymbol{p}^{(o)}$ based on 10 000 simulated datasets from the null distribution; i.e. $B = 10,000$ in Algorithm 2. Meanwhile, we also applied the Micro-DeMix-EM testing procedure as outlined in Algorithm 4 to obtain a *p*-value with $B = 100$. A smaller value of B was chosen for the permutation procedure to reduce the computational burden of Algorithm 4.

Based on 100 independent replications and the significance level $\alpha = 0.05$, for $\delta = 0$ and $\delta > 0$, the rejection rate in Fig. 1 represents the type-I error rate and the power, respectively. The value t on x-axis represents the total signal, and some calculations yield that $t = ||\boldsymbol{p}^{(r)} - \boldsymbol{p}^{(o)}||_2^2 = G\delta^2$. The points at 0 on x-axis representing the scenario that $\boldsymbol{p}^{(r)} = \boldsymbol{p}^{(o)}$.

As shown in Fig. 1, the power increases as the total signal grows. Micro-DeMix-EM exhibits lower power than Micro-DeMix, likely due to the slow convergence rate of the EM algorithm and/or the reduced number of permutations. Nonetheless, both Micro-DeMix procedures outperform all existing methods, except for ANCOM-BC in terms of power, but ANCOM-BC exhibits the highest type-I error rate. This result highlights that Micro-DeMix offers the best balance between type-I error rate and power among the methods evaluated. The reason most existing methods lack power is that they overlook the fact that the stool microbiome is a mixed population of rectum microbes and microbes from other GI locations. ANCOM-BC's distinct performance warrants further benchmarking studies to better understand its unique behavior.

We next revisit the iHMP IBD cohort, which is briefly discussed in Section 1, to demonstrate the effectiveness and efficiency of the proposed Micro-DeMix in elucidating the composition of real fecal microbiome data. We are particularly interested in understanding the heterogeneity of the stool microbiome, how the stool microbiome differs from the rectum microbiome in IBD populations, and how that difference may be related to the pathology of IBD.

As discussed in Section 1, the IBD cohort was a longitudinal study, collecting taxonomic data on the microbiome from 132 individuals (104 IBD patients) by analyzing 1785 stool samples and 651 intestinal biopsies over the course of 1 year

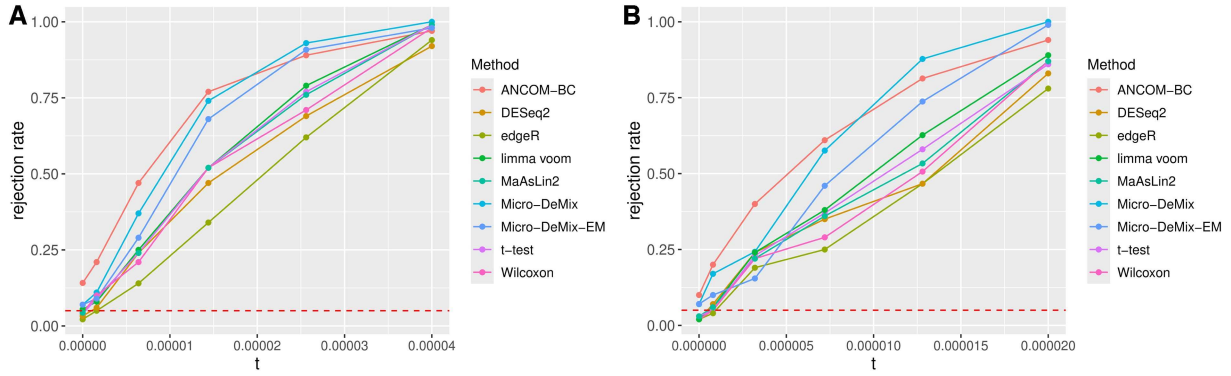

Figure 1. Type-I error rates and powers obtained at various signal strengths from the simulation study. (A) Power curves for $G = 10$. (B) Power curves for $G = 20$. A horizontal dashed line is shown at 0.05; t on the x-axis denotes the total signal. Rejection rates were obtained using both Micro-DeMix procedures and seven existing tools.

(Integrative HMP *et al.* 2019). This IBD cohort was part of the iHMP, which was designed to explore host–microbiome interplay to gain a holistic view of host–microbe interactions (Integrative HMP *et al.* 2019). The IBD IBDMDB research team collected microbiome data from different sites within the human gut of individuals with IBD to understand how the microbiome impacts human health and disease. For a fair comparison, our analysis focused on stool and rectum samples collected at the baseline, resulting in 13 samples for stool and 38 samples for rectum data in IBD patients. After excluding subjects with missing covariates of interest, such as age and gender, the stool group was narrowed down to 12 subjects.

The operational taxonomic unit (OTU)-level data for both groups are highly zero-inflated. More than 90% of the entries in the OTU table derived from the stool samples are zero. Therefore, in the results below, we took a hierarchical approach to investigate the microbial composition of the fecal microbiome as compared to the rectum microbiome at higher taxonomic levels to mitigate the potential influences of excess zeros. We also noticed significant variations in total reads (library size) across samples for both groups, specifically, ranging from 1499 to 79 228 for stool samples and from 34 to 31 781 for rectum samples. This nonbiological variation can result in samples with disproportionately high read counts dominating the analysis under the multinomial model, potentially masking true biological patterns due to this technical factor. We mitigated this issue by performing rarefaction, using the R function `rarefy_even_depth` from package `phyloseq` as in the exiting literature (McMurdie and Holmes 2013). In microbial studies, such rarefying has been used for normalization by randomly subsampling the data to mitigate the potential influence of varying library sizes (Hughes and Hellmann 2005, Koren *et al.* 2013, Navas-Molina *et al.* 2013). We also acknowledge that there is an ongoing debate in the research community regarding the use of rarefaction; see McMurdie and Holmes (2014) for more details. Given this, we suggest that the decision to perform rarefaction in microbiome data analysis should be carefully considered based on the data characteristics, research goals, and biological interpretations. We also applied the Micro-DeMix-EM algorithm to the IBD data without performing rarefaction, and the results are presented in the Supplementary Material.

3.3 iHMP IBD data: Phylum-level analysis

We aggregated the absolute counts from the OTU level to the phylum level, resulting in eight phyla, leading to a much lower percentage of zeros compared to the OTU-level data (50.9% versus 93.6%). We used 1499 as the rarefaction level for stool samples, as it is the minimum library size in the stool group. For a fair comparison, we applied the same rarefaction level to rectum samples, resulting in 30 samples in rectum group. Since this procedure is based on random subsampling without replacement, we repeated the procedure 100 times and took the average to reduce variability for both groups. Samples with library sizes less than 1499 were removed from the analysis.

We computed the relative abundance of these eight phyla in both the rectum and stool samples, and the results revealed a clear difference in the microbiome composition (Fig. 2).

Phylum Firmicutes was highly abundant in both the rectum and stool. It was the major phylum detected, representing more than 50% of the total microbial relative abundance in both groups. Overall, the relative abundance of Firmicutes in the stool (∼61.4%) was slightly higher than that in the rectum (∼52.0%). When compared to the microbial profile in rectum, the stool samples exhibited a considerable increase in the relative abundance of phylum Proteobacteria (from 7.1% to 30.1%), while the relative abundance of Bacteroidetes decreased from 38.1% to 6.7%. In addition, we observed an elevated relative abundance of the Verrucomicrobia phylum in the stool microbiome compared to the rectum, with an increase from 0.31% to 1.5%. Conversely, the phylum Fusobacteria showed a decrease from over 1.9% to a lower level. The remaining three phyla were nearly imperceptible in rectum and stool.

To further understand the composition of the fecal microbiome in IBD populations, we fitted the proposed Micro-DeMix model, incorporating gender and age as covariates. This analysis revealed estimates of the relative abundance of selected phyla from the other GI locations (Fig. 2). Compared to the rectum microbiome, the microbial profile of other GI locations revealed a higher relative abundance of phylum Firmicutes (∼64.3%), Proteobacteria (∼33.7%), and Verrucomicrobia (∼1.7%). Phylum Bacteroidetes was undetectable in other GI locations. The remaining phyla also exhibited relatively low abundance (<1%). Based on 10 000 simulated datasets, the Micro-DeMix test suggested significant differences in the taxonomic composition of the eight phyla between the rectum and other GI locations in IBD populations (p-value $< 1e-4$).

Our findings are consistent with recent studies on the gut microbiome in IBD populations. Specifically, the rectum and fecal microbiome of IBD patients is dominated by three major

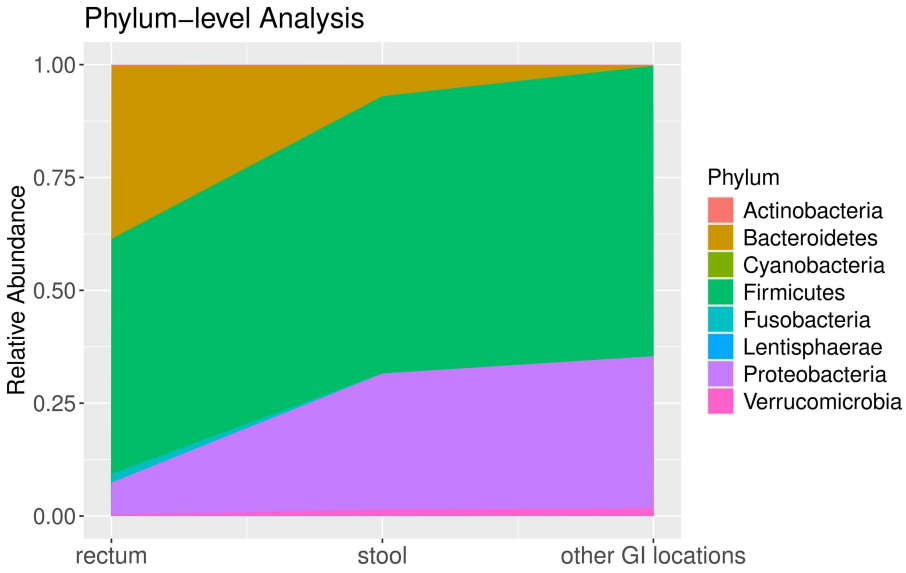

Figure 2. Relative abundance of microbial populations at the phylum level in the rectum, stool, and other GI locations.

bacterial phyla: Firmicutes, Bacteroidetes, Proteobacteria, and to a lesser degree Verrucomicrobia and Actinobacteria (Frank *et al.* 2007, Lo Presti *et al.* 2019). Our results are also consistent with recent studies that described distinct changes in the composition of the microbiota along the GI tract in IBD populations. The lower GI tract exhibited a larger abundance of Firmicutes and Bacteroidetes, whereas the upper GI tract was predominated by Proteobacteria and Firmicutes (Vuik *et al.* 2019). Compared to healthy individuals, the reduction in the relative abundance of Firmicutes and an elevation in Bacteroidetes have been observed in IBD patients (Khan *et al.* 2019, Lo Presti *et al.* 2019, Vuik *et al.* 2019, Zhang *et al.* 2022). Thus, our results shed light on the scientific premise of developing an effective microbiome-based treatment approach for IBD, such as the therapeutic supplementation of probiotics, prebiotics, and symbiotics, and fecal microbiota transplantation.

3.4 iHMP IBD data: Microbial composition within Proteobacteria and Firmicutes

In Section 4.1, we identified Proteobacteria and Firmicutes as the most abundant phyla in stool samples. To further understand the microbes in these two phyla, we conducted analyses of microbial composition at lower taxonomic levels (class, order, family) within Proteobacteria and Firmicutes phylum (Fig. 3).

Similar to the analyses in Section 4.1, we computed the relative abundance of all microbial classes and orders within phylum Proteobacteria for both the stool and rectum samples, and implemented the Micro-DeMix model to elucidate the difference between the rectum microbiome and the microbes in other GI locations. We followed the rarefying procedure that described in Section 4.1 to process the stool OTU data. In stool samples, the minimum library size is 13, which is not big enough to be included in the study. Therefore, we used 369, the second minimum, as the rarefaction level for both stool and rectum group. The outcomes of our analyses are visually represented in Fig. 3A and B.

Fig. 3A reveals that within phylum Proteobacteria, the major classes in rectum and stool are Betaproteobacteria and Gammaproteobacteria. However, the rectum microbiome is distinguished by a higher relative abundance of class Deltaproteobacteria ($\sim$11.1%) and Alphaproteobacteria ($\sim$ 3.0%), which are significantly low in abundance in stool samples. There is a low relative abundance of Epsilonproteobacteria (<1%) in both stool and rectum. As seen in Fig 3A, Gammaproteobacteria ($\sim$ 99.6%) is the most dominant class among the selected classes in other GI locations, while the other four classes are nearly undetectable. Although this finding may appear extreme, it is aligned with the existing knowledge that IBD patients have an altered gut microbiota marked by an increased relative abundance of Gammaproteobacteria class (Scales *et al.* 2016).

We proceeded to analyze the microbial composition of the phylum Proteobacteria at the order level, revealing marked differences between the rectum and stool samples (Fig. 3B). While both showed high abundance in orders Burkholderiales, Enterobacteriales, and Pasteurellales, the rectum samples were mainly characterized by the orders Desulfovibrionales and Neisseriales. We utilized the Micro-DeMix method to assess the microbiome composition in other GI locations. Our analysis revealed a dominance of Enterobacteriales ($\sim$81.1%), along with a lower proportion of Pasteurellales ($\sim$ 18.83%). This finding is supported by a recent study showing that Enterobacteriaceae exhibit increased abundance in the low-pH GI location and ileum (Morgan *et al.* 2012).

The Micro-DeMix test indicates significant differences in microbial profiles between the rectum and other GI locations (p-value $< 1e-4$) within the phylum Proteobacteria at both the class and order levels.

We further explored the microbial composition of the Firmicutes phylum. In the stool samples, three classes were identified: Bacilli, Clostridia, and Erysipelotrichi. At the original OTU level, class Clostridia dominated with a percentage of 97.1%, while class Erysipelotrichi was nearly undetectable. Given these findings, we narrowed our focus to lower taxonomic levels, specifically, order and family within the Firmicutes phylum. Similarly, we employed the minimum library size, 493, as the rarefaction level and processed both stool and rectum OTU data. The corresponding results are shown in Fig. 3C and D.

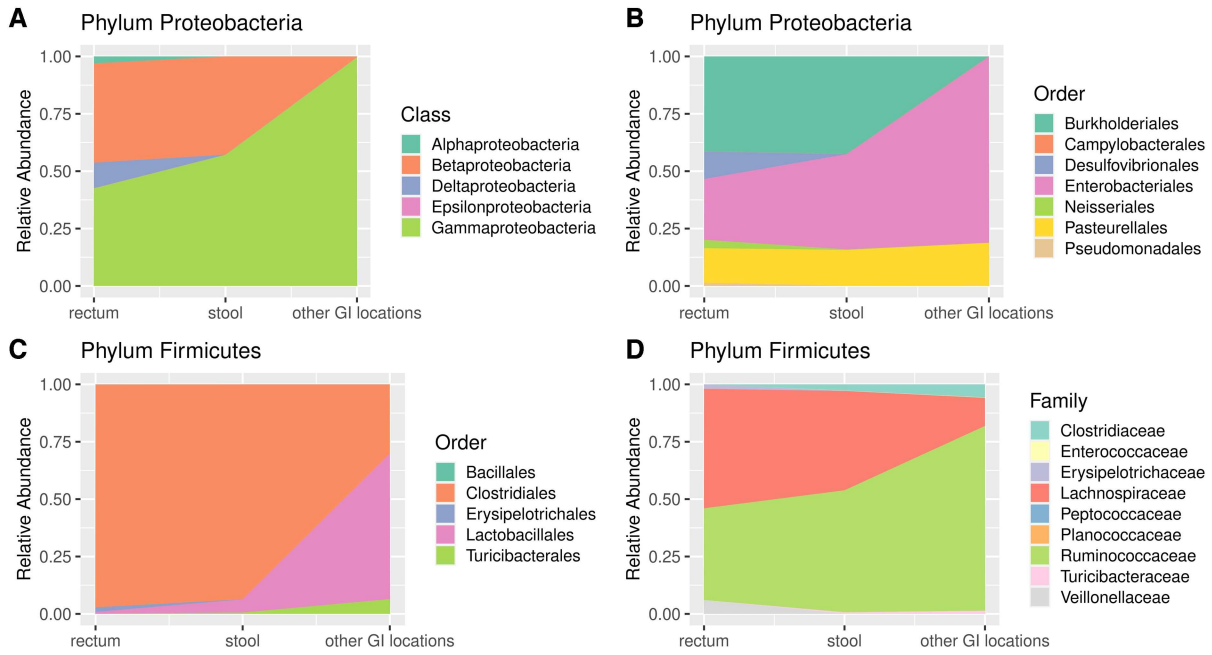

Figure 3. Relative abundance of microbial populations at various taxonomic levels in the rectum, stool, and other GI locations. (A, B) Microbial composition at the class and order levels within the phylum Proteobacteria. (C, D) Microbial composition at the order and family levels within the phylum Firmicutes.

Notably, at the order level (Fig. 3C), the phylum Firmicutes in the rectum and stool was predominantly represented by the order Clostridiales (97.1% in rectum and 93.6% in stool of all Firmicutes). Our results demonstrate the presence of order Turicibacterales in the stool samples, while it is absent in the rectum samples, which suggest that order Turicibacterales in stool samples may originate from other GI locations. The relative abundance of the order Lactobacillales increased in stool compared to rectum from 0.86% to 5.7%. In contrast, order Erysipelotrichales is found to be present (2.0%) in the rectum but remains at lower levels in the stool. The results of Micro-DeMix analysis show a p-value $< 1e-4$ for testing differential abundance between the rectum and other GI locations. Meanwhile, it highlights a significantly higher relative abundance of Lactobacillales ($\sim$63.4%) in other GI locations compared to the rectum and stool. The analysis estimates that order Turicibacterales ($\sim$6.4%) is also more abundant in other GI locations. These findings indicate a distinct microbial composition in other GI locations, with increased prevalence of orders Lactobacillales and Turicibacterales.

At the family level within the Firmicutes phylum (Fig. 3D), major families of the rectum microbiota are Ruminococcaceae, Lachnospiraceae, and Veillonellaceae. Comparisons between stool and rectum samples reveal an increase in Ruminococcaceae and a decrease in Lachnospiraceae and Veillonellaceae, aligning with findings from other studies (Lo Presti *et al.* 2019). One significant finding is the presence of the Clostridiaceae, Enterococcaceae, Planococcaceae, and Turicibacteraceae families in stool samples but not in the rectum. This observation suggests that these families detected in stool originate from other GI microbiome community. In other GI locations, our analysis shows a higher abundance of the Clostridiaceae and Ruminococcaceae families. In contrast, the family Lachnospiraceae exhibits a comparatively lower abundance. Furthermore, the p-value obtained through Micro-DeMix hypothesis testing is low ($< 1e-4$). These findings imply distinctive microbial profiles among the rectum, stool, and other GI locations, indicating variations in the relative abundance of specific taxa.

4 Conclusion

In this paper, we introduced Micro-DeMix, a beta-multinomial model designed for deconvoluting microbial abundance in stool samples. We proposed two parameter estimation procedures for the model, each with its strengths in specific types of applications. Additionally, we developed hypothesis testing procedures for detecting differential abundance, which has proven more effective compared to existing methods. We applied Micro-DeMix to rectum and stool microbiome datasets from the iHMP IBD cohort in elucidating the composition of real fecal microbiome data in IBD populations.

In the analyses of the iHMP IBD data, we performed rarefying on stool and rectum datasets to address the large variations in library sizes across samples. While rarefaction has been widely adopted in real microbiome applications, a limitation of this approach is that the selection of rarefaction level is data-dependent, and different rarefaction levels may lead to slightly different results. Due to the limited sample size, our strategy for selecting the rarefaction level is to keep as many samples as possible while removing samples with unreliably low library sizes. Different strategies may be considered for other datasets.

As mentioned earlier, the current paper focuses on integrating rectum samples and fecal samples due to data availability and biological relevance. However, the proposed method can seamlessly be used for integration of fecal samples with microbiome data collected from other GI locations. Furthermore, with microbiome data available from two or more GI locations, we could extend Micro-DeMix to have a refined decomposition of the fecal microbiome. Specifically, following work could

consider the stool microbiome as a J-component mixture of gut microbes, i.e. we let $p_{ig} = \pi_i^{(1)} p_g^{(1)} + \cdots + \pi_i^{(J-1)} p_g^{(J-1)} + (1 - \pi_i^{(1)} - \cdots - \pi_i^{(J-1)}) p_g^{(J)}$ in Equation (1), where $\boldsymbol{\pi}_i$ follows a Dirichlet $(\boldsymbol{a}_i)$ distribution. We leave this fruitful area for future research.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

Y.W. is supported in part by the funding R01GM145772 from the National Institute of Health. D.C. is supported by NSF Grant DMS-2220523 and Simons Foundation Collaboration Grant 854127.

Data availability

The IBD data analyzed in the paper can be accessed via HMP portal at https://portal.hmpdacc.org/. Code for implementing our methods is available in the R package `MicroDemix`, available at https://github.com/liuruoqian/MicroDemix.

References

Ahn J-S, Lkhagva E, Jung S *et al.* Fecal microbiome does not represent whole gut microbiome. *Cell Microbiol* 2023;**2023**:1.

Andoh A. Physiological role of gut microbiota for maintaining human health. *Digestion* 2016;**93**:176–81.

Aoun A, Darwish F, Hamod N. The influence of the gut microbiome on obesity in adults and the role of probiotics, prebiotics, and symbiotics for weight loss. *Prev Nutr Food Sci* 2020;**25**:113–23.

Bull MJ, Plummer NT. Part 1: the human gut microbiome in health and disease. *Integr Med (Encinitas, Calif.)* 2014;**13**:17–22.

Chunxi L, Haiyue L, Yanxia L *et al.* The gut microbiota and respiratory diseases: new evidence. *J Immunol Res* 2020;**2020**:2340670.

Cryan JF, O'Riordan KJ, Sandhu K *et al.* The gut microbiome in neurological disorders. *Lancet Neurol* 2020;**19**:179–94.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Series B (Methodol)* 1977;**39**:1–22.

Devaraj S, Hemarajata P, Versalovic J. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clin Chem* 2013;**59**:617–28.

Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* 2016;**14**:20–32.

Durack J, Lynch SV. The gut microbiome: relationships with disease and opportunities for therapy. *J Exp Med* 2019;**216**:20–40.

Frank DN, St. Amand AL, Feldman RA *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* 2007;**104**:13780–5.

Golub GH, Welsch JH. Calculation of gauss quadrature rules. *Math Comp* 1969;**23**:221–30.

Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;**57**:97–109.

Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods Enzymol* 2005;**397**:292–308.

Iatcu CO, Steen A, Covasa M. Gut microbiota and complications of type-2 diabetes. *Nutrients* 2021;**14**:166.

Integrative HMP, Proctor LM, Creasy HH *et al.* The integrative human microbiome project. *Nature* 2019;**569**:641–8.

Khan I, Ullah N, Zha L *et al.* Alteration of gut microbiota in inflammatory bowel disease (ibd): cause or consequence? Ibd treatment targeting the gut microbiome. *Pathogens* 2019;**8**:126.

Koren O, Knights D, Gonzalez A *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput Biol* 2013;**9**:e1002863.

Leite GGS, Weitsman S, Parodi G *et al.* Mapping the segmental microbiomes in the human small bowel in comparison with stool: a reimagine study. *Dig Dis Sci* 2020;**65**:2595–604.

Levitan O, Ma L, Giovannelli D *et al.* The gut microbiome–does stool represent right? *Heliyon* 2023;**9**:e13602.

Lo Presti A, Zorzi F, Del Chierico F *et al.* Fecal and mucosal microbiota profiling in irritable bowel syndrome and inflammatory bowel disease. *Front Microbiol* 2019;**10**:1655.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol* 2014;**15**:550–21.

Mallick H, Rahnavard A, McIver L. 2020. Maaslin 2: multivariable association in population-scale meta-omics studies. R/Bioconductor Package.

Mandal S, Van Treuren W, White RA *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 2015;**26**:27663.

Martin BD, Witten D, Willis AD. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann Appl Stat* 2020;**14**:94–115.

McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;**8**:e61217.

McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;**10**:e1003531.

Morgan XC, Tickle TL, Sokol H *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012;**13**:R79–18.

Navas-Molina JA, Peralta-Sánchez JM, González A *et al.* Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol* 2013;**531**:371–444.

Pellanda P, Ghosh TS, O'Toole PW. Understanding the impact of age-related changes in the gut microbiome on chronic diseases and the prospect of elderly-specific dietary interventions. *Curr Opin Biotechnol* 2021;**70**:48–55.

Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.

Scales BS, Dickson RP, Huffnagle GB. A tale of two sites: how inflammation can reshape the microbiomes of the gut and lungs. *J Leukoc Biol* 2016;**100**:943–50.

Shan Y, Lee M, Chang EB *et al.* The gut microbiome and inflammatory bowel diseases. *Annu Rev Med* 2022;**73**:455–68.

Sheth RU, Li M, Jiang W *et al.* Spatial metagenomic characterization of microbial biogeography in the gut. *Nat Biotechnol* 2019;**37**:877–83.

Tang WHW, Kitai T, Hazen SL *et al.* Gut microbiota in cardiovascular health and disease. *Circ Res* 2017;**120**:1183–96.

Vuik F, Dicksved J, Lam SY *et al.* Composition of the mucosa-associated microbiota along the entire gastrointestinal tract of human individuals. *United European Gastroenterol J* 2019;**7**:897–907.

Wu H-J, Wu E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes* 2012;**3**:4–14.

Zhang P. Influence of foods and nutrition on the gut microbiome and implications for intestinal health. *Int J Mol Sci* 2022;**23**:9588.

Zhang Y, Si X, Yang L *et al.* Association between intestinal microbiota and inflammatory bowel disease. *Animal Model Exp Med* 2022;**5**:311–22.

Bioinformatics, 2024, 40, 1–10
https://doi.org/10.1093/bioinformatics/btae667
Original Paper