Compositional knockoff filter for high-dimensional regression analysis of microbiome data

Arun Srinivasan[1], Lingzhou Xue[1,*], and Xiang Zhan[2,]**

[1]Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

[2]Department of Public Health Sciences, Pennsylvania State University, Hershey, PA 17033, U.S.A.

**email:* lzxue@psu.edu

***email:* xyz5074@psu.edu

SUMMARY: A critical task in microbiome data analysis is to explore the association between a scalar response of interest and a large number of microbial taxa that are summarized as compositional data at different taxonomic levels. Motivated by fine-mapping of the microbiome, we propose a two-step compositional knockoff filter (CKF) to provide the effective finite-sample false discovery rate (FDR) control in high-dimensional linear log-contrast regression analysis of microbiome compositional data. In the first step, we employ the compositional screening procedure to remove insignificant microbial taxa while retaining the essential sum-to-zero constraint. In the second step, we extend the knockoff filter to identify the significant microbial taxa in the constrained sparse regression model for compositional data. Thereby, a subset of the microbes is selected from the high-dimensional microbial taxa as related to the response under a pre-specified FDR threshold. We study the asymptotic properties of the proposed two-step procedure including both sure screening and effective false discovery control. We demonstrate the finite-sample properties in simulation studies, which show the gain in the empirical power while controlling the nominal FDR. We also illustrate the usefulness of the proposed method with application to an inflammatory bowel disease dataset to identify microbial taxa that influence host gene expressions.

KEY WORDS: Compositional constraint; Compositional screening; FDR control; Knockoff filter; Log-contrast model; Microbiome.

1. Introduction

The human microbiome refers to all the microbes that live in and on the human body with their collected genome, which has been linked to many human health and disease conditions (Cho and Blaser, 2012; Morgan et al., 2015; Wang and Jia, 2016; Mitchell et al., 2017). The advent of next-generation sequencing technologies enables studying the microbiome composition via direct sequencing of microbial DNA without the need of laborious isolation and cultivation, which largely boosts research interests in the human microbiome (Turnbaugh et al., 2007). Due to the varying amount of DNA yielding materials across different samples, the count of sequencing reads can vary greatly from sample to sample. As a result, it is a common practice to normalize the raw sequencing read counts to relative abundances making the microbial abundances comparable across samples (Li, 2015; Weiss et al., 2017). Besides the compositional constraint, the increasing availability of massive human microbiome datasets, whose dimensionality is much larger than its sample size, also poses new challenges to statistical analysis (Li, 2015).

A central goal in microbiome analysis is fine-mapping of the microbiome to identify microbial taxa that are associated with a certain response of interest (e.g., body mass index, disease/environmental exposure status, host genomic/genetic feature). In general, existing methods of fine-mapping the microbiome fall into two main categories: marginal approach and joint approach. The marginal approach usually casts the fine-mapping problem into the microbiome-wide multiple testing framework by examining marginal association between each microbial taxon and the response followed by multiple testing corrections (Wang and Jia, 2016; Xiao, Chao, and Chen, 2017), and taxa with adjusted p-values below a certain FDR threshold are identified as important ones that influence the response. This marginal approach is often limited for high-dimensional microbiome compositional data due to the following two reasons. First, it tends to have low detection power due to the heavy burden of

multiple testing adjustment inherent from the high-dimensional nature of microbiome data (Li, 2015). Second, it fails to account for the simplex nature of compositional data and may suffer from spurious negative correlations imposed by the fact that relative abundances across all taxa must sum to one within a given microbiome sample. As a consequence, traditional FDR control procedures (Benjamini and Hochberg, 1995) may not work for microbiome-wide multiple testing (Hawinkel et al., 2017).

On the other hand, a joint microbiome selection approach usually models all taxa collectively using penalized regression (Chen and Li, 2013; Lin et al., 2014). These joint approaches achieve fine-mapping of the microbiome via variable selection, yet they have no guarantee on the false discoveries among the selected microbiome features. This is probably because the number of microbial features in the joint regression model is much larger than the sample size and it is difficult to obtain a p-value measuring the significance of association between the outcome and each microbial feature. Yet, a canonical FDR control approach in general needs to plug p-values into a certain multiple testing procedure (Benjamini and Hochberg, 1995). Without FDR control, existing joint microbiome fine-mapping methods can produce less reliable discoveries and would probably lead to costly and fruitless downstream validation and functional studies (Wang and Jia, 2016; Hawinkel et al., 2017).

To address the potential limitations in existing marginal and joint microbiome fine-mapping approaches, we propose a new method in a joint regression framework to select microbial taxa under FDR control. In literature, the FDR control can be achieved via the knockoff filter framework (Barber and Candès, 2015; Candès et al., 2018). To facilitate FDR-controlled variable selection, the essence of knockoff filter lies in construction of a dummy copy of the original design matrix (also known as the knockoff matrix), which has the same underlying correlation structure as the original covariate matrix. However, the existing knockoff filter framework does not take into account the compositional structure of microbiome data. In the

literature of many other statistical inference (e.g., regression-based modelling, two-sample testing and statistical casual mediation analysis), it has been observed that applying classic statistical methods to analyze microbiome composition data is usually underpowered and sometimes can render inappropriate results (Aitchison, 2003; Shi, Zhang, and Li, 2016; Cao, Lin, and Li, 2017; Sohn and Li, 2019; Lu, Shi and Li, 2019; Zhang et al., 2019). Thus, new methods are desired rather than directly applying knockoff filter to microbiome data.

Following the pioneering work of Aitchison and Bacon-shone (1984), we model all taxa jointly in a linear log-contrast model to address the compositional nature of data and propose a two-step regression-based FDR control procedure to identify response-associated taxa. To deal with high-dimensional microbiome data, we follow the philosophy of recycled fixed-X knockoff (Barber and Candès, 2016). In the first step, we introduce the compositional screening procedure as a new method of variable screening for high-dimensional microbiome data subject to the compositional constraint. In the second step, we extend the recycled fixed-X knockoff procedure to the linear log-contrast model with compositional microbiome data. Both theoretical properties of the compositional screening procedure and the compositional knockoff filter are investigated. Using numerical studies, we demonstrate that the proposed method can jointly assess the significance of microbial covariates while also theoretically ensuring finite-sample FDR control.

To the best of our knowledge, our method is the first one to consider FDR controlled variable selection of microbiome compositional covariates in a joint regression framework. The proposed method will greatly benefit downstream microbiome functional studies by enhancing the reproducibility and reliability of discovery results in microbiome association studies. Our primary contributions are summarized as follows. First, we introduce the compositional screening procedure to screen true signals from high-dimensional compositional data. As demonstrated in thorough simulation, the newly proposed compositional screening

procedure yields a much higher likelihood of attaining all true signals compared to other commonly used methods, which do not account for the compositional nature. Further, we theoretically prove that the compositional screening procedure attains the desirable sure screening property under mild assumptions. The second main contribution of this paper is demonstrating that the proposed compositional knockoff filter (CKF) provides strong finite sample FDR control for microbial taxa selection. CKF uses fixed-X knockoff filter with recycling and accounts for the nature of microbiome data through the use of compositional constraint. In this high-dimensional microbiome covariates setting, we demonstrate that the proposed approach is more appropriate than the original model-X formulation (Candès et al., 2018), which requires complete knowledge of the conditional distribution of the design matrix to accommodate high dimensional design matrix . In microbiome studies, commonly assumed microbiome data distributions such as Dirichlet-multinomial (Chen and Li, 2013) is too complicated for the model-X formulation and yield poor control over FDR. Our proposed procedure is better suited for microbiome data analysis and achieves superior FDR control and power compared to other existing methods.

The rest of this paper is organized as follows. We propose the methodology of compositional knockoff filter in Section 2. The theoretical properties of the compositional knockoff filter are investigated in Section 3. The numerical properties are demonstrated through simulation studies in Section 4 and application to a microbiome data collected from an inflammatory bowel disease study in Sections 5. Technical proofs and additional numerical evaluations are deferred to the online supplementary materials.

2. Compositional Knockoff Filter

This section presents the compositional knockoff filter to perform FDR-controlled variable selection analysis for microbiome compositional data. The proposed method aims to address the high-dimensional compositional nature of microbiome data (i.e., $p > n$). To this end, we

follow the philosophy of recycled fixed-X knockoff procedure (Barber and Candès, 2016) to develop a new two-step procedure for high-dimensional compositional data, which consists of a compositional screening step and then a subsequent selection step. After introducing the log-contrast model in Section 2.1, we will present the screening step in Section 2.2 and the selection step in Section 2.3.

2.1 *Log-Contrast Model*

We use the log-contrast model (Aitchison and Bacon-shone, 1984) for joint microbiome regression analysis. Let $\mathbf{Y} \in \mathbb{R}^n$ denote the response vector and $\mathbf{X} \in \mathbb{R}^{n\times p}$ denote a matrix of microbiome compositions. By structure of the microbiome compositional components, each row of $\mathbf{X}$ must individually sum to 1. Thus $\mathbf{X}$ is not of full rank, leading to identifiability issues for the regression parameters. In order to account for this structure, the log-linear contrast model is often used for microbiome data (Lin et al., 2014; Shi et al., 2016). Without loss of generality, we assume that $X_{ij} > 0$ by replacing the zero proportions by a tiny pseudo positive value as routinely performed in practice (Lin et al., 2014; Shi et al., 2016; Cao et al., 2017; Lu et al., 2019; Zhang et al., 2019). Let $\mathbf{Z}^p \in \mathbb{R}^{n\times(p-1)}$ be a log-ratio transformation of the matrix $\mathbf{X}$, where $Z^p_{ij} = \log(X_{ij}/X_{ip})$ and p denotes the reference covariate. The linear log-contrast model is formulated as $\mathbf{Y} = \mathbf{Z}^p\boldsymbol{\beta}_{\setminus p} + \varepsilon$, where $\boldsymbol{\beta}_{\setminus p}$ is the vector of $(p-1)$ coefficients $(\beta_1, \beta_2, ..., \beta_{p-1})$ and the error vector $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. To avoid choosing a reference component towards a better interpretability, the linear log-contrast model is often reformulated into a symmetric form with a sum-to-zero constraint (Lin et al., 2014). That is,

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \varepsilon \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j = 0, \tag{1}$$

where $\mathbf{Z}$ is the $n \times p$ log-composition matrix with $Z_{ij} = \log(X_{ij})$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ are the regression coefficients for microbiome covariates. For ease of presentation, model (1) does not explicitly include other covariates, but all the results in the rest of this article still hold with other covariates.

We can use the ℓ_1-penalty to perform variable selection subject to the sum-to-zero constraint by solving the following compositional Lasso problem (Lin et al., 2014):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ ||\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j = 0. \tag{2}$$

Other penalties such as the folded concave penalties (Fan and Li, 2001; Fan et al., 2014) may also be used for the purpose of variable selection. For ease of presentation, we only focus on the ℓ_1-penalization problem (2), where existing methods (Lin et al., 2014) do not provide a rigorous FDR control on the selected variables.

2.2 *Compositional Screening Procedure*

As the fixed-X knockoff requires that $n \geqslant 2p$, screening the predictor set to a low-dimensional setting is necessary for the analysis of high-dimensional compositional data. Let n_0 denote the number of samples to use for screening and n_1 denote the remaining observations, where $n = n_0 + n_1$. We randomly split the original data $(\mathbf{Z}, \mathbf{Y})$ into $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$ and $(\mathbf{Z}^{(1)}, \mathbf{Y}^{(1)})$, where $\mathbf{Z}^{(0)} \in \mathbb{R}^{n_0 \times p}$, $\mathbf{Y}^{(0)} \in \mathbb{R}^{n_0}$, $\mathbf{Z}^{(1)} \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{Y}^{(1)} \in \mathbb{R}^{n_1}$. By ensuring that $\mathbf{Z}^{(0)}$ and $\mathbf{Z}^{(1)}$ are disjoint, we are able to implement a recycling step to reuse the original screening data $\mathbf{Z}^{(0)}$, in order to increase the selection power. To this end, we first use the sub-data $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$ to perform the screening and obtain a subset of features $\hat{S}_0 \subset \{1, ..., p\}$ such that $|\hat{S}_0| \leqslant \frac{n_1}{2}$, where $|\hat{S}_0|$ denotes the cardinality of set $\hat{S}_0$. Throughout this paper, we always assume $|\hat{S}_0| \leqslant \frac{n_1}{2}$ to ensure that we are able to construct the fixed-X knockoffs (Barber and Candès, 2015) for data $(\mathbf{Z}^{(1)}, \mathbf{Y}^{(1)})$ in the subsequent selection step.

As the selection step further reduces the feature set after screening, we must ensure that true signals are not lost before the selection step. For this reason, we desire screening methods that attain the sure screening property (Fan and Lv, 2008). That is, with high probability, we desire the selection set estimated by the screening method of choice to contain all relevant features. It is popular to perform screening using Pearson correlation (Fan and Lv, 2008; Fan and Song, 2010; Xue and Zou, 2011) or distance correlation (Li, Zhong and Zhu, 2012).

Despite that both marginal correlations-based screening methods enjoy the sure screening property asymptotically, these methods do not account for the compositional nature of microbiome data, which might lead to inefficient inference. We will further demonstrate this issue in the simulation studies of Section 4.1.

To effectively account for the compositional structure, we introduce the novel compositional screening procedure to improve the efficiency for screening microbiome compositional covariates. In general, best-subset selection is often used to identify the optimal k best features (Beale, Kendall and Mann, 1967). In our log-contrast model, the best-subset selection problem can be expressed as a constrained sparse least-squares estimation problem as follows:

$$\min_{\beta} \frac{1}{2}||\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}||_2^2 \quad \text{s.t.} \quad ||\boldsymbol{\beta}||_0 \leqslant k \quad \text{and} \quad \sum_{j=1}^{p} \beta_j = 0. \tag{3}$$

The proposed compositional screening procedure (3) can also be viewed as maximizing the log-likelihood $\ell_n(\boldsymbol{\beta})$ under the sparsity constraint that $||\boldsymbol{\beta}||_0 \leqslant k$ (Xu and Chen, 2014). Note that at most k features are retained after screening. As the screening is followed by a controlled variable selection step, a relatively lax choice of k can be used in the screening step to retain as many signals as possible for the subsequent selection step.

Although (3) is a NP-hard problem in general, the mixed integer optimization allows us to approximately solve the global solution of the nonconvex optimization problem (3) in an efficient manner (Konno and Yamamoto, 2009; Bertsimas, King and Mazumder, 2016). Finally, we demonstrate in the Section 3 that the computed solution of (3) by the mixed integer optimization attains the desirable sure screening guarantees.

2.3 *Controlled Variable Selection*

Let $\mathbf{Z}^{(1)}_{\hat{S}_0} \in \mathbb{R}^{n_1 \times |\hat{S}_0|}$ denote the columns of $\mathbf{Z}^{(1)}$ corresponding to $\hat{S}_0$, the selected set at the screening step. The knockoff matrix $\tilde{\mathbf{Z}}^{(1)}_{\hat{S}_0}$ is constructed using $\mathbf{Z}^{(1)}_{\hat{S}_0}$. We refer to Barber and Candès (2015) for a review of the construction of knockoff matrix. To increase selection

power, we construct the recycled knockoff matrix as

$$\tilde{\mathbf{Z}}_{\hat{S}_0} = \begin{bmatrix} \mathbf{Z}^{(0)}_{\hat{S}_0} \\ \tilde{\mathbf{Z}}^{(1)}_{\hat{S}_0} \end{bmatrix}$$

and then run the knockoff regression procedure using $\mathbf{Z}_{\hat{S}_0}$, $\tilde{\mathbf{Z}}_{\hat{S}_0}$, and $\mathbf{Y}$. In particular, we first append the screened original and knockoff matrices to create an augmented design matrix $\mathbb{Z}_{\hat{S}_0} = [\mathbf{Z}_{\hat{S}_0} \ \tilde{\mathbf{Z}}_{\hat{S}_0}]$. This augmented design matrix is of dimension $\mathbb{Z}_{\hat{S}_0} \in \mathbb{R}^{n \times 2|\hat{S}_0|}$ where the first $|\hat{S}_0|$ features are the original covariates and the remaining $|\hat{S}_0|$ features are the associated knockoff covariates. With this new augmented design matrix, we reformulate (2) as below:

$$\bar{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \left\{ ||\mathbf{Y} - \mathbb{Z}_{\hat{S}_0}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \right\} \quad \text{subject to} \quad \sum_{j=1}^{2|\hat{S}_0|} \beta_j = 0, \tag{4}$$

where $\bar{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$ is a vector appending the coefficients of original features and knockoff features. We consider a new microbiome community consists of both original microbes and their knockoff copies, and thus apply the sum-to-zero constraint $\sum_{j=1}^{2|\hat{S}_0|} \beta_j = 0$ to both the original and knockoff coefficients jointly.

The above optimization problem is performed over the entire Lasso path and provides a set of Lasso coefficients denoted by $\{\bar{\boldsymbol{\beta}}(\lambda)\} = \{(\hat{\boldsymbol{\beta}}(\lambda), \tilde{\boldsymbol{\beta}}(\lambda))\}$. Based on $\{\bar{\boldsymbol{\beta}}(\lambda)\}$, we next calculate the knockoff statistic W_j, which measures evidence against the null hypothesis $\beta_j = 0$ for each $j \in \hat{S}_0$. For the scope of this paper we use the Lasso signed lambda max statistic (LSM). Let $\mathbf{Z}_{\hat{S}_0,j}$ denote original covariate j and $\tilde{\mathbf{Z}}_{\hat{S}_0,j}$ denote knockoff covariate j:

$$W_j(\lambda) = (\max \lambda \text{ such that } \mathbf{Z}_{\hat{S}_0,j} \text{ or } \tilde{\mathbf{Z}}_{\hat{S}_0,j} \text{ enter lasso path}) \times \begin{cases} 1 \text{ if } \mathbf{Z}_{\hat{S}_0,j} \text{ enters before } \tilde{\mathbf{Z}}_{\hat{S}_0,j} \\ -1 \text{ if } \tilde{\mathbf{Z}}_{\hat{S}_0,j} \text{ enters before } \mathbf{Z}_{\hat{S}_0,j} \end{cases} \tag{5}$$

A large and positive W_j would suggest strong evidence that the original feature is significantly outcome-associated as an important feature tends to remain longer in lasso path as λ increases. Similarly, a negative or zero W_j value would indicate that the covariate tends to be noise. Thus, W_j is used to calculate the data-dependent knockoff thresholds that ensure

finite sample FDR-controlled variable selection. In this paper, both the standard knockoff and knockoff+ thresholds are considered:

KNOCKOFF THRESHOLD:

$$T = \min\left\{t \in \mathcal{W} : \frac{|\{j : W_j \leqslant -t\}|}{1 \vee |\{j : W_j \geqslant t\}|} \leqslant q\right\}, \tag{6}$$

KNOCKOFF+ THRESHOLD:

$$T = \min\left\{t \in \mathcal{W} : \frac{1 + |\{j : W_j \leqslant -t\}|}{1 \vee |\{j : W_j \geqslant t\}|} \leqslant q\right\}, \tag{7}$$

where $q \in [0, 1]$ is the user-specified nominal FDR level, $\mathcal{W} = \{|W_j| : j \in \hat{S}_0\}\backslash\{0\}$ are the unique non-zero values of $|W_j|$'s ($T = +\infty$ if $\mathcal{W}$ is empty) and $a \vee b$ denotes the maximum of a and b. Once this threshold has been calculated, we select covariates $S = \{j : W_j \geqslant T\}$. Depending on the threshold being used, we term this FDR-control variable selection procedure as either compositional knockoff filter or compositional knockoff filter+, whose properties will be studied in Section 3. For completeness, we summarize the proposed compositional knockoff filter procedures in Algorithm 1.

3. Theoretical Properties

In this section, we show the theoretical properties of both compositional screening procedure and compositional knockoff filter. Firstly, we show that the computed solution from solving the constrained sparse maximum likelihood problem (3) via the mixed integer optimization attains the desired sure screening property. Then, we demonstrate that the knockoff thresholds attain finite sample FDR control under the compositional constraint. The proof to establish these theoretical properties is available through the online supplementary materials.

3.1 *Theoretical Properties of Compositional Screening*

We will show in this section that the compositional screening procedure attains the sure screening property. For ease of presentation, some notation is introduced first. Let s denote an arbitrary subset of $\{1, \ldots, p\}$ corresponding to a sub-model with coefficients β_s, and S^*

Algorithm 1 Compositional Knockoff Filter (CKF)

Input: log-compositional matrix $\mathbf{Z}$, response $\mathbf{Y}$, FDR threshold q, screening sample size n_0 and screening set size $|\hat{S}_0|$

Output: knockoff selection set S

Procedure:

(1) Randomly split the data $(\mathbf{Z}, \mathbf{Y})$ into disjoint $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$ and $(\mathbf{Z}^{(1)}, \mathbf{Y}^{(1)})$.

(2) **Screening Step:**

(a) Run the compositional screening procedure method on $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$ to identify $\hat{S}_0$.

(3) **Selection Step:**

(a) Generate the recycled knockoff matrix $\tilde{\mathbf{Z}}_{\hat{S}_0}$ and construct the augmented design matrix: $\mathbb{Z}_{\hat{S}_0} = [\mathbf{Z}_{\hat{S}_0} \ \tilde{\mathbf{Z}}_{\hat{S}_0}]$.

(b) Solve equation (4) to calculate the coefficients $\bar{\boldsymbol{\beta}}(\lambda)$.

(c) Calculate knockoff statistics W_j from $\bar{\beta}_j(\lambda)$.

(d) Use the knockoff or knockoff+ threshold (6) and (7) to calculate T from $\mathcal{W}$.

(e) Determine the knockoff or knockoff+ selection set as $S = \{j : W_j \geqslant T\}$.

be the true model with p^* nonzero coefficients, with corresponding true coefficient vector $\boldsymbol{\beta}^*$. Let $\hat{S}_0$ denote the computed screened sub-model after applying the compositional screening procedure. Assume that $\hat{S}_0$ retains at most k features with $p^* < k < p$. Let $\mathbf{S}_+^k = \{s : S^* \subset s; ||s||_0 \leqslant k\}$ denote the set of all overfit models and $\mathbf{S}_-^k = \{s : S^* \not\subset s; ||s||_0 \leqslant k\}$ denote the set of underfit models. We will show that the compositional screening procedure does not miss true signals with high probability. That is:

$$P(S^* \subset \hat{S}_0) \to 1 \text{ as } n \to \infty. \tag{8}$$

For the technical aspects of our proof to hold, we make the following assumptions (1-4), encompassing requirements on the dimension, signal strength and microbiome design matrix:

ASSUMPTION 1: $\log(p) = O(n^m)$ for some $0 \leqslant m < 1$.

ASSUMPTION 2: There exists $w_1 > 0$ and $w_2 > 0$ and non-negative constants τ_1 and τ_2 such that $\min_{j \in S^*} |\beta_j^*| \geqslant w_1 n^{-\tau_1}$ and $p^* < k \leqslant w_2 n^{\tau_2}$.

ASSUMPTION 3: There exist constants $c_1 > 0$ and $\delta_1 > 0$ such that for sufficiently large n such that $\lambda_{min}[n^{-1} \sum_{i=1}^n \mathbf{Z}_{is} \mathbf{Z}_{is}^t] \geqslant c_1$ for $s \in \mathbf{S}_+^{2k}$ and $||\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*||_2 \leqslant \delta_1$, where λ_{min} denotes the smallest eigenvalue of the matrix and $\mathbf{Z}_{is} = (Z_{ij})_{j \in s}$.

ASSUMPTION 4: There exist constants $c_2 > 0$ and $c_3 > 0$ such that $|Z_{ij}| \leqslant c_2$ and $\max_{1 \leqslant j \leqslant p} \max_{1 \leqslant i \leqslant n} \{ \frac{Z_{ij}^2}{\sum_{i=1}^n Z_{ij}^2 \sigma_i^2} \} \leqslant c_3 n^{-1}$ when n is sufficiently large.

Assumption 1 places a weak restriction on p and n of the data, which is very likely to be met in many microbiome studies, where p is on the order of thousands and n is on the order of hundreds (Wang and Jia, 2016). Assumption 2 places a restraint on the minimum strength of true signals, such that they are discoverable. This assumption is common for statistical screening and variable selection methods (Fan and Lv, 2008; Fan and Song, 2010; Lin et al., 2014). Both Assumption 3 and Assumption 4 place constraints on the microbiome design matrix $\mathbf{Z}$ and are more technical. Using examples of both simulated microbiome data sets and the mucosal microbiome data analyzed in Section 5, we illustrate that both Assumption 3 and 4 are very realistic for microbiome data. Details are available through the online supplementary materials. Under Assumptions 1–4, Theorem 1 shows that the proposed compositional screening procedure attains the sure screening property. The proof of Theorem 1 relies on two key lemmas which will be presented first.

LEMMA 1: *Let $\tilde{S}_0$ denote the set of screened features from the global solution of the constrained sparse maximum-likelihood estimation problem* (3)*, where $|\tilde{S}_0| = k$. Let $\mathbf{S}_+^k = \{s : S^* \subset s; ||s||_0 \leqslant k\}$. Assume that Assumptions 1-4 hold and $\tau_1 + \tau_2 < \frac{(1-m)}{2}$. Then:*

$$P(\tilde{S}_0 \in \mathbf{S}_+^k) \to 1 \text{ as } n \to \infty$$

Lemma 1 ensures that the model selected by the solution of the constrained sparse maximum-likelihood estimation will be in the set of overfit models with high-probability. Thus, this ensures no signals are lost during screening. In other words, the global solution of the constrained sparse maximum-likelihood estimation problem attains the sure screening property.

LEMMA 2: *Let $\hat{\boldsymbol{\beta}}_{MIO}$ denote the computed coefficient magnitudes of the model selected by the compositional screening procedure through mixed integer optimization and $\tilde{\boldsymbol{\beta}}$ denote the coefficients of the global solution of the constrained sparse maximum likelihood problem. Given $\varepsilon > 0$, then:*

$$P(||\hat{\boldsymbol{\beta}}_{MIO} - \tilde{\boldsymbol{\beta}}||_\infty < \varepsilon) \to 1$$

Lemma 2 demonstrates that the computed solution of the compositional screening procedure through mixed integer optimization converges to the global solution of the constrained sparse maximum likelihood problem with high probability. By combining Lemma 1 and Lemma 2, it follows that the computed solution attains the sure screening property. This result is presented as Theorem 1.

THEOREM 1: *Given we have n independent observations with p possible features. Assume that Assumptions 1-4 hold and $\tau_1 + \tau_2 < \frac{(1-m)}{2}$. Let $\hat{S}_0$ denote the computed screened set from the compositional screening procedure where $|\hat{S}_0| = k$. Then:*

$$P(S^* \subset \hat{S}_0) \to 1 \text{ as } n \to \infty$$

Theorem 1 allows us to claim that the compositional screening procedure will not lose any signals during screening with high probability. In summary, the compositional screening procedure accounts for the compositional constraint and also ensures the screening power.

3.2 *Theoretical Properties of Compositional Knockoff Filter*

In order to control FDR, the knockoff statistic must obey the *anti-symmetry* and *sufficiency* properties while the design matrix and response must satisfy both the *Pairwise Exchangeabil-*

ity for the Response Lemma and *Pairwise Exchangeability for the Features Lemma* (Barber and Candès, 2015). In this paper we primarily focus on the LSM knockoff statistic (5) which has been shown to satisfy the *anti-symmetry* and *sufficiency* properties (Barber and Candès, 2016). This result is unchanged under the addition of the sum-to-zero constraint. Therefore, the main focus of the theory relating to the selection step is to show that even with the additional sum-to-zero constraint on the $\boldsymbol{\beta}$ coefficients, the two exchangeability results still hold. Given the exchangeability results outlined in the online supplementary materials, the compositional knockoff+ threshold attains finite sample FDR control as stated in the following theorem.

THEOREM 2: *For $q \in [0,1]$, the compositional knockoff+ method with data-recycling ensures:*

$$\mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } j \in S\}|}{|S| \vee 1} \,\middle|\, E\right] \leqslant q$$

where S denotes the set of selected coefficients through the compositional knockoff+ procedure, E denotes the event $\{S^* \subset \hat{S}_0\}$. The expectation is over the Gaussian noise vector ε and $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are fixed.

Theorem 2 demonstrates that compositional knockoff+ procedure controls the FDR at a user-specified level q, after conditioning on the results of the screening procedure. Following the argument in Theorem 2 of Barber and Candès (2016), if a proper screening procedure which attains the sure screening property (such as our proposed compositional screening procedure through mixed integer optimization) is implemented in the screening step, FDR is controlled even without conditioning on E.

REMARK 1: Given the above exchangeability results and the previous theorems, the standard compositional knockoff threshold controls a modified form of false discovery rate (Barber

and Candès, 2015). In particular, for $q \in [0, 1]$, the compositional knockoff method ensures:

$$\mathbb{E}\left[\frac{|\{j : \beta_j = 0 \text{ and } j \in S\}|}{|S| + q^{-1}}\middle| E\right] \leqslant q.$$

Compared with the formula in Theorem 2, the additional q^{-1} in the denominator sometimes favors a larger selected set S in CKF compared to CKF+. But when the selected set S is large or when the nominal FDR threshold q is relatively large, the difference between CKF and CKF+ vanishes as q^{-1} has little effect compared to $|S|$ under such scenarios.

4. Simulation Studies

We conducted two sets of simulation studies to evaluate numerical performance of the proposed CKF methods. In our first simulation study, we evaluated the sure screening property of the proposed compositional screening procedure (CSP). The compositional screening procedure was implemented by modifying the methods in the *bestsubset* package (Hastie, Tibshirani and Tibshirani, 2017). We compared CSP to two other popular statistical screening procedures in literature: one based on Pearson correlation/PC (Fan and Lv, 2008) and the other based on distance correlation/DC (Li et al., 2012). In the second set of simulations, we evaluated the selection performance of CKF methods. For comparison, we also consider other methods that are widely used for microbial taxa selection. One is the compositional Lasso (Lin et al., 2014) and the other is the marginal method which examines one taxon at a time followed by the Benjamini-Hochberg procedure for FDR control (Benjamini and Hochberg, 1995; Paulson et al., 2013; Parks et al., 2014). Additional numerical simulations comparing the CKF and original model-X knockoff filter (KF) methods are available through the online supplementary materials. The *zeroSum* R package (Altenbuchinger et al., 2017; Rehberg, 2017) was used to perform the sum-to-zero constrained optimization in this simulation.

To mimic a real dataset analyzed later in this paper, we considered sample size $n = 250$ and number of microbiome covariates $p = 400$ in both simulations. The number of observations

used in screening was set to $n_0 = 60$, corresponding to roughly 25% of the data, and the rest $n_1 = 190$ observations were used for the selection step. The magnitude of the screening set was fixed at $|\hat{S}_0| = 50$. We first generated the microbiome counts from the Dirichlet-multinomial distribution following previous designs (Zhao et al., 2015; Zhan et al., 2017a,b). Zero counts were first replaced by a pseudo count of 0.5, as commonly suggested in microbiome data analysis (Lin et al., 2014; Cao et al., 2017; Weiss et al., 2017; Lu et al., 2019; Zhang et al., 2019), and then microbiome counts were transformed to relative abundances. Next, we varied the sparsity levels $k = [10\ 15\ 20\ 25]$ and set the first 25 entries of the coefficient vector as: $\boldsymbol{\beta} = (-6, 6, 5, -2, -3, 5, 5, -3, -3, -4, 5, 3, -4, -2, -2, -4, 4, 2, 2, -4, -7, 3, 4, -3, 3)$. The remaining 375 entries were set to be 0. For each $k \in [10\ 15\ 20\ 25]$ we constructed the coefficients $\boldsymbol{\beta}_k$ by combining the first k entries of $\boldsymbol{\beta}$ and the remaining $p-k$ entries of zeroes. Under this scheme, it is easy to check that the coefficient vector always satisfies the sum-to-zero constraint under each of the four sparsity levels. Finally, we simulated the response vector $\mathbf{Y}$ from $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta}_k + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$. Both response and predictors were centered so that the intercept was omitted. To illustrate the potential usefulness and robustness of our CKF methods, we also considered another scheme to generate the microbiome composition data besides the Dirichlet-multinomial distribution used here. The corresponding simulation results are presented in the online supplementary materials.

4.1 *Screening Simulation*

We first applied the three screening methods (CSP, PC, DC) to the simulated data to evaluate the screening accuracy by calculating the proportion of true features being selected in the screened set. The performance of screening is crucial to the subsequent selection inference. To see this, we compared the performance of CKF and CKF+ with three different screening procedures at a target nominal FDR of 0.1. To measure their performance, empirical FDR and empirical power were calculated. Let S denote the final empirical selection set of either

CKF or CKF+, S^* denote the set of true non-zero coefficients and $\boldsymbol{\beta}$ be the true model coefficients:

$$\widehat{\text{FDR}} = \mathbb{E}_N\left[\frac{|\{j : \beta_j = 0 \text{ and } j \in S\}|}{|S| \vee 1}\right]; \qquad \widehat{\text{Power}} = \mathbb{E}_N\left[\frac{|\{j : \beta_j \neq 0 \text{ and } j \in S\}|}{|S^*|}\right],$$

where $\mathbb{E}_N$ denotes the empirical average over $N = 200$ replicates.

[Table 1 about here.]

The results of screening accuracy are summarized in Table 1. The proposed compositional screening procedure has much better performance than the other two competing methods (Fan and Lv, 2008; Li et al., 2012), which have been widely used in statistic literature. This is another example that classic statistical methods may be inefficient for microbiome data without accounting for the compositional nature (Lin et al., 2014; Shi et al., 2016; Cao et al., 2017; Lu et al., 2019; Zhang et al., 2019). By incorporating the compositional constraint, the proposed CSP achieves the sure screening property for microbiome data as the proportion of true features retained in the screened set is always one based on Table 1.

[Table 2 about here.]

We further calculate the empirical FDR and power of CKF/CKF+ with different screening methods and report the results in Table 2. As observed, both CKF and CKF+ with all screening methods can control FDR under the nominal level. Based on the below part of Table 2, we see that a CSP-based CKF/CKF+ method has a much higher power than its counterparts that are built on either PC-based or DC-based screening. Combining this with Table 1, we see the importance of correctly screening relevant features on the first step. If true signals are missed in the original screening step, then they will never be identified in the downstream selection step. This ripple effect has important consequences on the power of the downstream selection procedure. Therefore, a screening procedure (such as CSP) with sure screening property is crucial to guarantee the power of CKF/CKF+. Based on these

results, CSP was set as the default screening procedure in CKF/CKF+ and was the only one evaluated in the rest of simulations.

4.2 *Selection Simulation*

In this section, we compared CKF/CKF+ to two other taxa selection methods including compositional Lasso (CL) and Benjamini-Hochberg (BH) procedure. For the CL method, the optimal λ used in the compositional Lasso was determined through 10-fold cross-validation. As the number of microbial features is typically larger than the sample size in microbiome association studies, it is difficult to obtain joint association p-values for each microbial feature. We examined the association between the outcome and each microbial feature marginally and applied the Benjamini-Hochberg (BH) procedure to these marginal p-values to identify features significant under FDR of 0.1. The performance of these methods are reported in Table 3.

[Table 3 about here.]

As observed from Table 3, CKF, CKF+ and BH can control the nominal FDR level, while CL has an extremely high empirical false discovery rate. Lasso has proven to be a versatile tool with appealing estimation and selection properties in the asymptotic setting (Tibshirani, 1996). Yet, its performance under finite sample setting is not guaranteed. It is also not surprising that CL may have inflated empirical FDR given that the original CL method is developed for variable selection and does not necessarily guarantee on the FDR control of the selected variables. The power of each procedure is summarized in the bottom half of Table 3. As the CL has an extremely inflated FDR, it is not meaningful to compare its power to the other methods that can control FDR and hence power of CL is not reported. From Table 3, both CKF and CKF+ are much more powerful than the marginal BH method especially when the signal is dense ($k = 20, 25$). This is likely due to the fact

that CKF methods analyze the microbial covariates jointly, and in the dense signal cases, the effectiveness of the marginal method deteriorates.

To summarize, the proposed compositional screening procedure enjoys the sure screening property, which is crucial to guarantee a high power of the downstream selection analysis. Our CKF methods successfully control the FDR of selecting outcome-associated microbial features in a regression-based manner which jointly analyzes all microbial covariates, while having the highest power detecting outcome-associated microbes under the nominal FDR threshold. As a comparison, existing methods may either be underpowered (BH) or render inappropriate results (CL) by having an inflated FDR than the nominal threshold.

5. Real Data Example

To demonstrate the usefulness of our method, we further apply it to a real data set obtained from a study examining the association between host gene expression and mucosal microbiome using samples collected from patients with inflammatory bowel disease (Morgan et al., 2015). The abundances of 7000 OTUs from $n = 255$ samples were measured using 16S rRNA gene sequencing and most to these 7000 species-level OTUs were in extremely low abundances with a large proportion of OTUs being simply singletons, possibly due to a sequencing error. As suggested in literature (Li, 2015), we aggregated these OTUs to genus and perform a more robust analysis in the genus level. These 7000 OTUs belonged to $p = 303$ distinct genera, whose abundances were the microbial covariates of interest in our analysis.

It has been previously found that microbially-associated host transcript pattern is enriched for complement cascade genes, such as genes CFI, C2, and CFB (Morgan et al., 2015). Moreover, principal component-based enrichment analysis shows that host gene expression is inversely correlated with taxa *Sutterella*, *Akkermansia*, *Bifidobacteria*, *Roseburia* abundance and positively correlated with *Escherichia* abundance under the nominal FDR of 0.25 (Morgan et al., 2015). In this analysis, we took the expression values of complement cascade

genes (CFI, C2, and CFB) as the outcomes of interest, and applied the proposed CKF and CKF+ method to detect host gene expression-associated genera for each outcome under the FDR threshold of 0.25. For the initial screening step, we fixed the screening sample size $n_0 = 100$ and set size $|\hat{S}_0| = 40$. As the data-splitting is random, we repeated the CKF algorithm 10 times with different splits. By using multiple splits matrices, we were more likely able to identify any possible signals under the desired FDR level.

[Table 4 about here.]

In Table 4, we present taxa that were identified at least once across the ten runs. Taxa in bold were also identified in the original paper (Morgan et al., 2015) using marginal method to control the FDR at 0.25. For the coefficient column of Table 4, we fit the reduced log-contrast linear regression models with predictors of both selected taxa and clinical variables including disease subtype, antibiotic use, tissue location and inflammatory score, as done previously (Morgan et al., 2015). These clinical variables were included in the model to adjust for potential confounding effects and to obtain a more accurate estimate of the microbiome effect on host gene expression. The sign of a taxon coefficient reflects the direction of association (activation or inhibition). Recall that five taxa *Sutterella, Akkermansia, Roseburia, Bifidobacterium* and *Escherichia* were detected in the original principal component-based marginal analysis (Morgan et al., 2015). All these five except *Roseburia* were identified in our analysis. Moreover, we further see that the coefficient signs for each taxa of interest are consistent with the expected direction posited by Morgan et al. (2015) except for *Akkermansia*. In other words, we correctly identify a majority of taxa of interest function as inhibitors (negative coefficient) or activators (positive coefficient) for each cascade gene expression.

We also observe that the taxa set identified for each cascade gene are different, which suggests that specific taxa play key roles on individual gene expression. Despite that we

missed taxa *Roseburia* compared to the original analysis, many new taxa were identified as complement cascade gene expression-associated in our CKF analysis. For example, *Aeromonas* appears in the selection sets for both the CFB and CFI as an inhibitor which may be of particular interest. Likewise, *Lactobacillus* appears in both the CFB gene and C2 gene acting as an inhibitor. On the other hand, *Collinsella* appears to be performing as an inhibitor for CFB and C2, but as an activator for CFI. The mechanism of how these new taxa affect the host transcript pattern warrants further laboratory investigation.

To conclude, the proposed CKF is more powerful in detecting significant taxa than the original principal component-based marginal analysis (Morgan et al., 2015) under the same nominal FDR of 0.25. Our new method not only provides additional statistical support to results obtained from the original analysis but also gains new biological and biomedical insights on how taxa interact with host complement cascade gene expressions.

6. Discussion

In this paper, we consider the problem of identifying outcome-associated microbiome features under a pre-specified FDR. Traditional methods usually cast this problem into a multiple testing framework and examines each microbiome feature individually followed by certain multiple testing procedures to control the FDR. To avoid the potential heavy multiple adjustment burden, we alternatively adopt a joint approach which regresses the response on all microbiome features and achieve FDR control via applying the compositional knockoff filter to the regression. As shown in the numerical studies, our new methods are much more powerful (regarding detecting more true positives) than existing methods. Moreover, the application our method to the host-microbiome data not only identifies the same gene expression-associated taxa as in the original study (Morgan et al., 2015), but also leads to new discoveries, which may provide new biological insights with further laboratory investigation.

Currently, our method can only identify microbial taxa that are associated with a single

continuous outcome variable. It is of future interest to extend CKF to more complicated models such as survival models (Plantinga et al., 2017), multivariate-outcome models (Zhan et al., 2017a,b) and generalized linear models (Lu et al., 2019) to accommodate microbiome association studies with more complicated designs. The canonical approach of microbiome fine-mapping is to plug in marginal p-values into the BH procedure to identify outcome-associated taxa under FDR control (Paulson et al., 2013; Parks et al., 2014; Wang and Jia, 2016). Under this vein, there has been a wealth of research interests to utilize additional specific information (e.g., phylogenetic information) of microbiome data to increase the power of detection and maintain control of the FDR (Xiao et al., 2017; Jiang et al., 2017; Hu et al., 2018). It is of future interest to incorporate such information to our CKF framework to further boost the detection power while controlling the FDR at a certain threshold.

7. Supplementary Materials

Web Appendices referenced in Section 3 and Section 4 are available with this article. The supplemental materials include proofs of Lemmas 1 and 2, Theorems 1 and 2, notes on the assumptions of Theorem 1 in the context of microbiome data and additional simulation evaluations.

References

Aitchison, J., and Bacon-shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71**, 323–330.

Aitchison, J. (2003). The statistical analysis of compositional data. Caldwell, New Jersey: Blackburn Press.

Altenbuchinger, M., Rehberg, T., Zacharias, H. U., Staemmler, F., Dettmer, K., Weber, D., et al. (2017). Reference point insensitive molecular data analysis. *Bioinformatics* **33**, 219–226

Barber, R., and Candès E. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics* **43**, 2055–2085.

Barber, R., and Candès E. (2016). A knockoff filter for high-dimensional selective inference. https://arxiv.org/abs/1602.03574.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika* **54**, 357-366.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B* **57**, 289-300.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of statistics* **44**, 813-852.

Cao, Y., Lin, W., and Li, H. (2017). Two-sample tests of high-dimensional means for compositional data. *Biometrika* **105**, 115–132.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B* **80**, 551-577.

Chen, J., and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of applied statistics* **7**, 418–442.

Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13**, 260–270.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**, 1348-1360.

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* **70**, 849-911.

Fan, J., and Song, R. (2010). Sure independence screening in generalized linear models with

NP-dimensionality. *The Annals of Statistics* **38**, 3567-3604.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of statistics* **42**, 819-849.

Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2017). A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in bioinformatics* **20**, 210-221.

Hu, J., Koh, H., He, L., Liu, M., Blaser, M. J., and Li, H. (2018). A two-stage microbial association mapping framework with advanced FDR control. *Microbiome* **6**, 131.

Jiang, L., Amir, A., Morton, J. T., Heller, R., Arias-Castro, E., and Knight, R. (2017). Discrete False-Discovery Rate Improves Identification of Differentially Abundant Microbes. *MSystems* **2**, e00092-17.

Konno, H., and Yamamoto, R. (2009). Choosing the best set of variables in regression analysis using integer programming. *Journal of Global Optimization* **44**, 273-282.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129-1139.

Li, H. (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application* **2**, 73–94.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–797.

Lu, J., Shi, P., and Li, H. (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **75**, 235-244.

Mitchell, C. M., Srinivasan, S., Zhan, X., Wu, M. C., Reed, S. D., Guthrie, K. A., et al. (2017). Vaginal microbiota and genitourinary menopausal symptoms: a cross-sectional

analysis. *Menopause* **24**, 1160–1166.

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology* **16**, 67.

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**, 3123-3124.

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10**, 1200.

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R. and Wu, M. C. (2017). MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome* **5**, 17.

Rehberg, T. (2017). Elastic-net regularized regression with zerosum constraint https://github.com/rehbergT/zeroSum (accessed Feb 23, 2019).

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* **10**, 1019–1040.

Sohn, M. B., and Li, H. (2019). Compositional Mediation Analysis for Microbiome Studies. *Annals of Applied Statistics* **13**, 661-681.

Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267–288.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* **449**, 804.

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology* **14**, 508.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normal-

ization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27.

Xiao, J., Cao, H., and Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* **33**, 2873–2881.

Xu, C., and Chen, J. (2014). The sparse MLE for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association* **109**, 1257-1269.

Xue, L., and Zou, H. (2011). Sure independence screening and compressed random sensing. *Biometrika*, **98**, 371-380.

Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017a). A small-sample multivariate kernel machine test for microbiome association studies. *Genetic Epidemiology* **41**, 210–220.

Zhan, X., Plantinga, A., Zhao, N., and Wu, M. C. (2017b). A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics* **73**, 1453–1463.

Zhang, H., Chen, J., Li, Z., and Liu, L. (2019). Testing for Mediation Effect with Application to Human Microbiome Data. *Statistics in Biosciences*, 1–16.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *American Journal of Human Genetics* **96**, 797–807.

Table 1

Average screening proportions of true signals based on 200 replicates.

Screening Method	$k = 10$	$k = 15$	$k = 20$	$k = 25$
CSP	1.000	1.000	1.000	1.000
PC	0.745	0.601	0.511	0.431
DC	0.734	0.594	0.518	0.434

Table 2

Empirical FDR and power based on 200 replicates under the nominal FDR level of 0.1. The above half is empirical FDR and the below half is empirical power.

Selection Method	Screening Method	$k = 10$	$k = 15$	$k = 20$	$k = 25$
CKF	CSP	0.042	0.030	0.038	0.026
	PC	0.078	0.087	0.066	0.060
	DC	0.061	0.078	0.061	0.061
CKF+	CSP	0.020	0.012	0.019	0.016
	PC	0.013	0.033	0.023	0.019
	DC	0.015	0.020	0.020	0.018
CKF	CSP	1.000	0.989	0.981	0.943
	PC	0.724	0.538	0.429	0.317
	DC	0.700	0.531	0.428	0.311
CKF+	CSP	0.930	0.837	0.870	0.803
	PC	0.068	0.221	0.195	0.122
	DC	0.068	0.187	0.178	0.118

Table 3

Empirical FDR and power under nominal FDR of 0.1 based on 200 replicates.

Metric	Method	$k = 10$	$k = 15$	$k = 20$	$k = 25$
Empirical FDR	CKF	0.042	0.030	0.038	0.026
	CKF+	0.020	0.012	0.019	0.016
	CL	0.848	0.756	0.652	0.489
	BH	0.100	0.100	0.091	0.097
Empirical Power	CKF	1.000	0.989	0.981	0.943
	CKF+	0.930	0.837	0.870	0.803
	BH	0.846	0.712	0.609	0.493

Table 4

Taxa identified as host gene expression associated under the nominal FDR of 0.25.

Response Gene	Taxa Identified	Coefficient
CFI	***Escherichia***	0.0282
	Sutterella	-0.0280
	Akkermansia	0.0151
	Blautia	-0.0006
	Epulopiscium	0.0138
	Aeromonas	-0.0103
	Bulleidia	-0.0163
	Clostridium	-0.0085
	Eubacterium	-0.0040
	Collinsella	0.0027
C2	***Escherichia***	0.0217
	Bifidobacterium	-0.0185
	Sutterella	-0.0278
	Coprococcus	- 0.0016
	Veillonella	0.0330
	Collinsella	-0.0125
	Staphylococcus	0.0232
	Brevundimonas	0.0253
	Lactobacillus	-0.0330
	Anaerococcus	-0.0304
	Allobaculum	0.0600
	Bulleidia	-0.0528
	Rhodoplanes	0.0136
CFB	***Escherichia***	0.0243
	Sutterella	-0.0288
	Bifidobacterium	-0.0222
	Clostridium	-0.0120
	Coprococcus	0.0122
	Epulopiscium	0.0123
	Turicibacter	-0.0232
	Collinsella	-0.0006
	Eggerthella	0.0870
	Aeromonas	-0.0230
	Lactobacillus	-0.0247
	Anaerococcus	-0.0419
	Adlercreutzia	-0.1428
	Novosphingobium	0.0348
	Eubacterium	0.0511
	Bradyrhizobium	0.0550
	RFN20	-0.1535
	Anaeroglobus	0.1961