HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2022 September 01.

Published in final edited form as:

Biometrics. 2021 September ; 77(3): 984–995. doi:10.1111/biom.13336.

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

Arun Srinivasan[1], **Lingzhou Xue**[1,*], **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.

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 propose a new 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 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 theoretical properties of the proposed two-step procedure, including both sure screening and effective false discovery control. We demonstrate these properties in numerical simulation studies to compare our methods to some existing ones and show power gain of the new method while controlling the nominal FDR. The potential usefulness of the proposed method is also illustrated with application to an inflammatory bowel disease dataset to identify microbial taxa that influence host gene expressions.

Keywords

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). The advent of next-generation sequencing technologies enables studying the microbiome composition via direct sequencing of microbial DNA without the

* lzxue@psu.edu . ** xyz5074@psu.edu .

Supporting Information

Web Appendices referenced in Section 2 to Section 4 are available with this article at the Biometrics website on Wiley Online Library. R code to implement the proposed methods is also available there.

need for laborious isolation and cultivation, which largely boosts research interest in the human microbiome. Due to the varying amount of DNA yielding materials across different samples, the count of sequencing reads can vary significantly 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 (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 an outcome of interest. In general, existing methods 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 the marginal association between each microbial taxon and the outcome followed by multiple testing correction to identify important taxa that are associated with the outcome (Wang and Jia, 2016; Xiao, Chao, and Chen, 2017). The marginal approach is often limited 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 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 it is difficult to obtain a p-value measuring the significance of the association between the outcome and each microbial feature given that the number of microbial features in the joint regression model is much larger than the sample size. 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 may often 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 approaches, a new method in a joint regression framework to select microbial taxa with finite-sample FDR control is desired. In the statistics literature, finite-sample FDR control can be achieved via the knockoff filter framework, in which a dummy knockoff copy of the original design matrix has been constructed and flagged as false positives to facilitate FDR-controlled variable selection (Barber and Candès, 2015). However, as observed in the literature of many other statistical inference methods (e.g., regression-based modeling, two-sample testing, and statistical causal mediation analysis), applying classic statistical methods to analyze compositional 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 FDR-controlled variable selection methods are desired for analysis of microbiome compositional 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-controlled variable selection procedure named compositional knockoff filter (CKF) to identify response-associated taxa under finite-sample FDR control. In the first step, we introduce the compositional screening procedure (CSP) as a new method of variable screening for high-dimensional microbiome data subject to the compositional constraint. In the second step, we apply the fixed-X knockoff filter procedure (Barber and Candès, 2015) to the reduced model from the first screening step. Using numerical studies, we demonstrate that the proposed CKF method can jointly assess the significance of microbial covariates while also theoretically ensuring finite-sample FDR control. 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 CSP to screen true signals from high-dimensional compositional data and theoretically verify that CSP attains the desirable sure screening property under mild assumptions. As demonstrated in simulations, CSP yields a much higher likelihood of attaining all true signals compared to some existing methods that do not account for the compositional nature. Second, by leveraging the high-dimensional knockoff filter framework (Barber and Candès, 2019), we avoid the non-trivial sequential conditional independent pairs algorithm of model-X knockoffs (Candès et al., 2018) and provide an alternative CKF approach to ensure finite-sample FDR control for microbial taxa selection. Construction of model-X knockoff features through methods such as the sequential conditional independent pairs algorithm (Candès et al., 2018) requires both complete knowledge of the joint distribution of the microbiome design matrix and repeated derivation of the conditional distributions, that are non-trivial for many non-Gaussian distributions such as Dirichlet-multinomial and logistic normal, which are frequently used for modeling microbiome data (Lin et al., 2014; Tang and Chen, 2019). While the development of methods to construct exact or approximate knockoff features for a broader class of distributions is a promising area of active research (Bates et al., 2020; Romano, Sesia, and Candès, 2019), the robustness of model-X knockoff performance to an arbitrary multivariate (non-Gaussian) distribution is currently unknown. To this end, the proposed CKF with finite-sample FDR control guarantee is appealing for taxa selection.

The rest of this paper is organized as follows. We propose the methodology of compositional knockoff filter in Section 2. Theoretical properties are investigated in Section 3. The numerical properties are demonstrated through simulation studies in Section 4 and applications to a microbiome data set collected from an inflammatory bowel disease study in Sections 5. The paper concludes with a brief discussion in Section 6. Technical proofs and additional numerical evaluations are deferred to the online supporting information.

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, 2019) to develop a new two-step procedure for high-dimensional microbiome 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 then the selection step in Section 2.3.

2.1 Log-Contrast Model

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 one. Thus $\mathbf{X}$ is not of full rank, leading to identifiability issues for the regression parameters. In order to account for this structure, the linear log-contrast model is often used for compositional data (Aitchison and Bacon-shone, 1984; Lin et al., 2014). We assume that $X_{ij} > 0$ by replacing the zero proportions by a tiny pseudo positive value as routinely performed in microbiome data analysis (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 the log-ratio transformation of $\mathbf{X}$, where $Z_{ij}^p = \log\left(X_{ij}/X_{ip}\right)$ and p denotes the reference covariate. The linear log-contrast model is formulated as $\mathbf{Y} = \mathbf{Z}^p\boldsymbol{\beta}_{\setminus p} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}_{\setminus p}$ is the vector of $(p-1)$ coefficients $(\beta_1, \beta_2, \ldots, \beta_{p-1})$ and error $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$. To avoid picking a reference component for better model interpretability, the log-contrast model is often reformulated into a symmetric form with a sum-to-zero constraint (Lin et al., 2014). That is,

$$y_i = \sum_{j=1}^{p} Z_{ij}\beta_j + \varepsilon_i \text{ subject to } \sum_{j=1}^{p} \beta_j = 0, \tag{1}$$

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

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 be used 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 $\mathbf{Z}^{(0)}$ 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 $\hat{S}_0 \subset \{1, \ldots, p\}$ 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 estimated screening set to contain all relevant features. It is popular to perform screening using Pearson correlation (Fan and Lv, 2008; Xue and Zou, 2011) or distance correlation (Li, Zhong and Zhu, 2012). Despite enjoying 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 account for the compositional structure, we introduce the novel compositional screening procedure to improve the efficiency of 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:

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

The proposed compositional screening problem (2) can also be viewed as maximizing the log-likelihood $\ell_n(\boldsymbol{\beta})$ under the sparsity constraint $\|\boldsymbol{\beta}\|_0 \leqslant k$ (Xu and Chen, 2014). The choice of k is a fundamental question in many high-dimensional screening procedures. Practical domain knowledge may provide information on how sparse one believes the underlying signal is. Common choices for screening set size are often $k = c\lfloor \frac{n_0}{\log(n_0)} \rfloor$ for some $c > 0$ (Fan and Lv, 2008; Li et al., 2012). However, as noted by Li et al. (2012), the choice of screening set size can be viewed as a tuning parameter within the model and concrete means to determine the screening set size are an area of future development. Although (2) is a NP-hard problem, the mixed integer optimization (MIO) allows us to approximately solve the global solution of the nonconvex optimization problem (2) 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 (2) by MIO attains the desirable sure screening guarantees.

After screening, the model reduces to $y_i = \sum_{j \in \hat{S}_0} Z_{ij}\beta_j^r + \varepsilon_i$ subject to $\sum_{j \in \hat{S}_0} \beta_j^r = 0$. Comparing it to the original log-contrast model (1), the regression coefficients in the reduced model β_j^r does not necessarily match β_j in the original model. To solve this discrepancy, we implement a normalization procedure $X_{ij}^* = X_{ij} / \sum_{j \in \hat{S}_0} X_{ij}$ for $j \in \hat{S}_0$ and

for an abuse of notation, we still use $Z_{ij} = log\left(X^*_{ij}\right)$ to denote the design matrix to be used in the subsequent selection step. Rationale of this normalization is available at Section S.1 of the online supporting information.

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 from the computed solution of (2), and we delineate this from the selection set from the global solution of (2) which we instead denote as $\tilde{S}_0$. The knockoff matrix $\widetilde{\mathbf{Z}}^{(1)}_{\hat{S}_0}$ is constructed using $\mathbf{Z}^{(1)}_{\hat{S}_0}$ following the fixed-X knockoff framework (we refer to Barber and Candès (2015) for a review of the construction of knockoff matrix under the fixed-X design). Notably, the fixed-X knockoff requires $n \geqslant 2p$ but places no assumptions on the distribution of $\mathbf{Z}^{(1)}$. Thus, a key appeal of the knockoff filter is the relative lack of strong assumptions on the design matrix needed for theoretical finite-sample FDR control to hold. The use of the screening step allows us to apply the fixed-X knockoff framework in the high-dimensional setting. While fixed-X knockoffs traditionally require a low-dimensional regime, the screening step first reduces the effective dimension to at most $\frac{n_1}{2}$ (i.e., $\mathrm{Z}^{(1)}_{\hat{S}_0}$ has at most $\frac{n_1}{2}$ columns). As the knockoff matrix is constructed on $\mathrm{Z}^{(1)}_{\hat{S}_0}$ alone, this satisfies the dimensionality requirements for the construction of the fixed-X knockoff matrix $\widetilde{\mathbf{Z}}^{(1)}_{\hat{S}_0}$. To further boost the selection power, we follow the data recycling mechanism outlined in Barber and Candès (2019) to construct the recycled knockoff matrix as $\widetilde{\mathbf{Z}}_{\hat{S}_0} = \left(\mathbf{Z}^{T(0)}_{\hat{S}_0}, \widetilde{\mathbf{Z}}^{T(1)}_{\hat{S}_0}\right)^T$. Note that we treat ($\mathbf{Z}^{(0)}$, $\mathbf{Y}^{(0)}$) as fixed after the screening step and the first part of knockoff copies are exact copies under the recycling scheme.

We next run the knockoff regression procedure using $\mathrm{Z}_{\hat{S}_0}$, $\widetilde{\mathbf{Z}}_{\hat{S}_0}$, and $\mathbf{Y}$. In particular, we first append the screened original and knockoff matrices to create an augmented design matrix $\mathbb{Z}_{\hat{S}_0} = \left[\mathrm{Z}_{\hat{S}_0}, \widetilde{\mathbf{Z}}_{\hat{S}_0}\right]$, with 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 knockoff copies. With this new augmented design matrix, we solve the following Lasso problem:

$$\overline{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \left\| \mathbf{Y} - \mathbb{Z}_{\hat{S}_0} \boldsymbol{\beta} \right\|_2^2 + \lambda \parallel \boldsymbol{\beta} \parallel_1 \right\}. \tag{3}$$

where $\overline{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}})$ is a vector appending the coefficients of original features and knockoff features. Comparing to previous problems (1) and (2), we no longer require a sum-to-zero constraint in our augmented Lasso problem (3). This is because, by adding $|\hat{S}_0|$ knockoff

features in the augmented design matrix $\mathbb{Z}_{\hat{S}_0}$, the corresponding microbiome data matrix $\mathbb{X}_{\hat{S}_0} = \exp\left(\mathbb{Z}_{\hat{S}_0}\right)$ is no longer compositional in nature.

The above optimization problem (3) is performed over the entire Lasso path and provides a set of Lasso coefficients denoted by $\left\{\overline{\boldsymbol{\beta}}(\lambda)\right\} = \left\{\widehat{\boldsymbol{\beta}}(\lambda), \widetilde{\boldsymbol{\beta}}(\lambda)\right\}$. Based on $\left\{\overline{\boldsymbol{\beta}}(\lambda)\right\}$, 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 $\mathbb{Z}_{\hat{S}_0, j}$ denote original covariate j and $\widetilde{\mathbf{Z}}_{\hat{S}_0, j}$ denote knockoff covariate j:

$$W_j = \left(\max \lambda \mathbf{Z}_{\hat{S}_0, j} \text{ or } \widetilde{\mathbf{Z}}_{\hat{S}_0, j} \text{ enters lasso path}\right) \times \begin{cases} 1 \text{ if } \mathbf{Z}_{\hat{S}_0, j} \text{ enters before } \widetilde{\mathbf{Z}}_{\hat{S}_0, j} \\ -1 \text{ if } \widetilde{\mathbf{Z}}_{\hat{S}_0, j} \text{ enters before } \mathbf{Z}_{\hat{S}_0, j} \end{cases} \quad (4)$$

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. Finally, we consider both the knockoff threshold $T = \min\left\{t \in \mathscr{W} : \left(\left|\left\{j : W_j \leqslant -t\right\}\right|\right) / \left(1 \vee \left|\left\{j : W_j \geqslant t\right\}\right|\right) \leqslant q\right\}$ and the knockoff+ threshold $T = \min\left\{t \in \mathscr{W} : \left(1 + \left|\left\{j : W_j \leqslant -t\right\}\right|\right) / \left(1 \vee \left|\left\{j : W_j \geqslant t\right\}\right|\right) \leqslant q\right\}$, where $q \in [0, 1]$ is the user-specified nominal FDR level, $\mathscr{W} = \{|W_j| : j \in \hat{S}_0\} \setminus \{0\}$ are the unique non-zero values of $|W_j|$'s ($T = +\infty$ if $\mathscr{W}$ is empty) and $a \vee b$ denotes the maximum of a and b. Once this threshold has been calculated, we select covariates $\hat{S} = \{j : W_j \geqslant T\}$. Depending on the threshold being used, we term this FDR-control variable selection procedure as either compositional knockoff filter (CKF) or compositional knockoff filter+ (CKF+). For completeness, we summarize the proposed CKF procedures in Algorithm 1.

3. Theoretical Properties

In this section, we first present the theoretical properties of CSP and show that the computed solution from solving the constrained sparse maximum likelihood problem (2) via the mixed integer optimization attains the desired sure screening property. We then summarize the theoretical properties of the proposed CKF methods. Leveraging the framework of high-dimensional knockoff filter (Barber and Candès, 2019), we verify that CKF/CKF+ attains finite sample FDR control. The main results are presented in this main text and details on the proofs to establish these theoretical properties is available through Section S.3 of the online supporting information.

3.1 Theoretical Properties of Compositional Screening

We first introduce notations before showing that CSP attains the sure screening property. Let s denote an arbitrary subset of $\{1, \ldots, p\}$ corresponding to a sub-model with coefficients $\boldsymbol{\beta}_s$,

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

$$P\left(S^* \subset \hat{S}_0\right) \to 1 \text{ as } n \to \infty . \tag{5}$$

For the technical aspects of our sure-screening proof to hold, we make the following assumptions encompassing requirements on the 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 $\min_{j \in S^*} \left|\beta_j^*\right| \geqslant w_1 n^{-\tau_1}$ 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 $\lambda_{\min}\left[n^{-1}\sum_{i=1}^n \mathbf{Z}_{is}\mathbf{Z}_{is}^t\right] \geqslant c_1$ for $s \in \mathrm{S}_+^{2k}$ and $\left\|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\right\|_2 \leqslant \delta_1$, where $\lambda_{min}[M]$ denotes the smallest eigenvalue of the matrix M, 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 $\max_{1 \leqslant j \leqslant p} \max_{1 \leqslant i \leqslant n} \left\{\frac{Z_{ij}^2}{\sum_{i=1}^n Z_{ij}^2 \sigma_i^2}\right\} \leqslant c_3 n^{-1}$ when n is sufficiently large, where $\sigma_i^2 = Var(\mathbf{Y} \mid \mathbf{Z})$.

Algorithm 1 Compositional Knockoff Filter (CKF)

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

Output: knockoff selection set $\hat{S}$

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 $\hat{S}_0$.

(b) Apply the normalization procedure $X_{ij}^* = X_{ij} / \sum_{j \in \hat{S}_0} X_{ij}$ for $j \in \hat{S}_0$ and calculate $Z_{ij} = log(X_{ij}^*)$ as the design matrix to be used in the subsequential selection step.

(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 (3) to calculate coefficients $\tilde{\boldsymbol{\beta}}(\lambda)$ and then W_j from $\tilde{\beta}_j(\lambda)$ using (4).

(c) Calculate the selection set $\hat{S} = \{j : W_j \geqslant T\}$, where T is either knockoff threshold or knockoff+ threshold.

Assumption 1 places a weak restriction on p and n of the data, which is very likely to be met in many microbiome studies (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; Lin et al., 2014). Assumption 3 corresponds to the UUP condition (Candes and Tao, 2007) which controls the pairwise correlations between the columns of **Z**. This condition is prevalent across many high-dimensional variable selection methods such as the Dantzig selector (Candes and Tao, 2007), SIS-DS (Fan and Lv, 2008), forward regression (Wang, 2009) and sparse MLE (Xu and Chen, 2014). We have conducted numerical studies to evaluate Assumption 3 in the context of microbiome data analysis in Section S.2 of the online supporting information. Based on the results presented there, Assumption 3 typically holds for microbiome data and it might be problematic under some extreme scenarios such as when very highly correlated taxa or very rare taxa are included in the design matrix. Therefore, we would suggest removing some taxa out of a highly correlated taxa cluster and very rare taxa before analysis. As shown in the numerical studies presented later in this paper, the proposed method can successfully capture a majority of the true underlying signals, which suggests that Assumption 3 may not be a concern for microbiome data analysis. Finally, as noted by Xu and Chen (2014) and Chen and Chen (2012), Assumption 4 likely will hold for a wide class of design matrices as long as σ_i^2 is not degenerate. In Section S.2 of the online supporting information, we illustrate the validity of Assumption 4 on the mucosal microbiome data analyzed later in this paper. Finally, 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 are presented first.

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

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

Lemma 1 ensures that the model selected by the solution of problem (2) 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 $\widehat{\boldsymbol{\beta}}_{MIO}$ denote the computed coefficients of the constrained sparse maximum likelihood problem selected by the compositional screening procedure through mixed integer optimization and $\widetilde{\boldsymbol{\beta}}$ denote the coefficients of the global solution of the constrained sparse maximum likelihood problem. Given $\varepsilon > 0$, then:

$$P\left(\|\widehat{\boldsymbol{\beta}}_{MIO} - \widetilde{\boldsymbol{\beta}}\|_\infty < \varepsilon\right) \to 1$$

Lemma 2 demonstrates that the computed solution of problem (2) 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, which is presented in 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 $p^{\star} < |\hat{S}_0| < p$. Then:

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

Theorem 1 allows us to claim that CSP 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 FDR Control Property of Compositional Knockoff Filter

The FDR control property of the CKF is a consequence of the FDR control theory outlined in knockoff framework Barber and Candès (2015, 2019), as the compositional nature of the design matrix after screening and normalization does not affect argument in the proof of FDR control of fixed-X knockoff framework. More details is provided in Section S.3 of the online supporting information and we reiterate the FDR control property here for posterity. As we have validated that the CSP attains the sure-screening property, the compositional knockoff+ threshold ensures finite sample FDR control as stated in the following:

Theorem 2: For q ∈ [0, 1], the proposed CKF+ method ensures:

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

and the proposed CKF method controls a modified form of FDR:

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

where $\hat{S}$ denotes the index set of selected coefficients through CKF+/CKF, E denotes the event $\{S^* \subset \hat{S}_0\}$. The expectation is over the Gaussian noise vector ε and $(\mathbf{Z}, \widetilde{\mathbf{Z}})$ are fixed.

Theorem 2 demonstrates that CKF+ controls the FDR at a user-specified level q, after conditioning on the results of the screening procedure. By the argument in Theorem 2 of Barber and Candès (2019), if a proper screening procedure which attains the sure screening property (such as the proposed CSP through mixed integer optimization) is implemented in the screening step, FDR is controlled even without conditioning on E.

4. Simulation Studies

We conducted two sets of simulation studies (screening simulation and selection simulation) to evaluate numerical performance of the proposed CKF methods. In the screening simulation, we evaluated the sure screening property of CSP and compared it 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). We considered a sample size of $n_0 = 100$ and screening set size $|\hat{S}_0| = 40 \approx 2\lfloor \frac{n_0}{\log(n_0)} \rfloor$. In the selection simulation, we evaluated the selection performance of CKF methods. For comparison, we also considered some other methods in the selection simulation: 1) compositional Lasso (Lin et al., 2014); 2) a marginal method which examines one taxon at a time followed by correction (Benjamini and Hochberg, 1995); 3) the original model-X knockoff (Candès et al., 2018). To mimic the real dataset analyzed later in this paper, we considered sample size $n = 250$ and number of microbiome covariates $p = 400$ in the selection simulation. Among these $n = 250$ samples, a randomly selected sub-sample with $n_0 = 100$ observations were used in the first screening step and the rest $n_1 = 150$ observations were used for the selection step.

Two schemes were used to generate the microbiome compositional design matrix in both simulations. The first scheme was to generate microbiome counts from the Dirichlet-multinomial (DM) distribution following a previous design (Zhan et al., 2017). The library size of each sample was randomly simulated from a negative binomial distribution with a mean parameter of 10000 and dispersion parameter of 25. Raw zero counts were first replaced by a pseudo count of 0.5 and then transformed to relative abundances. The second scheme for generating microbiome compositional data was to use the logistic normal (LN) distribution (Aitchison, 2003; Lin et al., 2014; Cao et al., 2017). Following a previous design (Lin et al., 2014), we first simulated an intermediate $n \times p$ data matrix $\mathbf{M}$ from multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mu_i = 1$ and $\Sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$. Then, we calculated the log-composition design matrix as $Z_{ij} = \log\left(\frac{\exp\{M_{ij}\}}{\sum_{j=1}^{p}\exp\{M_{ij}\}}\right)$ for $i = 1, \ldots, n, j = 1, \ldots, p$.

Next, we varied the sparsity levels $|S^*| \in \{15, 20, 25, 30\}$ and set the first 30 entries of the whole regression coefficient vector $\boldsymbol{\beta}_{1:400}$ as: $\boldsymbol{\beta}_{1:30} = (-3, 3, 2.5, -1, -1.5; 3, 3, -2, -2, -2; 1, -1, 3, -2, -1; -1, 1, 2, -1, -1; 3, 3, -3, -2, -1; 3, 3, -3, -2, -1)$. The remaining coefficients $\boldsymbol{\beta}_{31:400}$ were all zeros. We constructed the regression coefficients in the aforementioned way such that $\sum_{j=1}^{|S^*|}\beta_j = 0$, for each $|S^*| \in \{15, 20, 25, 30\}$. 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 $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta}_{S^*} + \varepsilon$, where $\boldsymbol{\beta}_{S^*} = (\boldsymbol{\beta}_{1:|S^*|}, \mathbf{0})$ for $|S^*| \in \{15, 20, 25, 30\}$ and $\varepsilon \sim \mathcal{N}(0, I)$.

4.1 Screening Simulation

We first evaluate the screening accuracy of three methods (CSP, PC, DC) by calculating the proportion of true features being selected in the screened set, $|\hat{S}_0 \cap S^*|/|S^*|$, where $\hat{S}_0$ is the screening set and S^* is the set of covariates with true non-zero coefficients in the log-contrast model. The results on screening accuracy of different methods are summarized in Table 1.

The proposed CSP has much better performance than the other two competing methods PC and DC (Fan and Lv, 2008; Li et al., 2012), which have been widely used in the statistical 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 close to one based on Table 1. To show the importance of the screening performance to the subsequent selection inference, we have further conducted additional numerical studies to compare the performance of CKF and CKF+ with three different screening procedures. The results are reported in Table S4 and Table S5 in Section S.4 of the online supporting information.

4.2 Selection Simulation

In this section, we compared CKF/CKF+ to some existing methods including compositional lasso (CL) method (Lin et al., 2014), Benjamini-Hochberg (BH) procedure and model-X knockoff filter (KF). The KF method places the burden of knowledge on knowing the complete conditional distribution of $\mathbf{Z}$, and there is no algorithm that can generate model-X knockoffs for general distributions efficiently (Bates et al., 2020). Therefore, we employ the default design used previously (Candès et al., 2018) in this simulation. 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. To measure performance of different methods, empirical FDR and empirical power were calculated.

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

where $\mathbb{E}_N$ denotes the empirical average over $N = 200$ replicates. The results of empirical FDR and empirical power are reported in Table 2.

As observed from Table 2, CKF+, KF+ and BH can control the nominal FDR level, which is desired. CKF and KF yield slightly inflated FDR levels above the nominal rate, but this is expected as both KF and CKF are only guaranteed to control a modified version of FDR.

Finally, CL has a high empirical false discovery rate across all scenarios. The Lasso method has proven to be a versatile tool with appealing estimation and selection properties in the asymptotic setting (Tibshirani, 1996). Yet, its finite-sample performance is not guaranteed. Our results on CL is consistent with the fact that a relatively large number of false positives are reported in Table 1 of Lin et al. (2014). Despite being able to guarantee model selection consistency, CL tends to select more unnecessary variables to recover the true model.

Since CL has an extremely inflated FDR, it is not meaningful to compare its power to the other methods that can control FDR and hence is omitted in power simulations. Comparing empirical powers reported in Table 2, both CKF+ and KF+ are much more powerful than BH. This power gap is likely due to the fact that CKF+ and KF+ analyze the microbial covariates jointly, and the effectiveness of the marginal BH method deteriorates when the dimension (or multiple correction burden) is relatively high. Under DM distribution, KF+ achieves as higher power than CKF+ in sparse setting ($|S^*| = 15$ or 20). However, CKF+ becomes more powerful over KF+ as the signal becomes denser ($|S^*| = 25$ or 30). On the other hand, under LN distribution, the effectiveness of KF+ quickly deteriorates and CKF+ is much more powerful than KF+ based on Table 2. The KF+ method generated knockoffs based on an underlying Gaussian assumption on the covariates, and therefore its performance under the microbiome setting (e.g., Dirichlet-multinomial and logistic normal distributions considered in our simulations) is not guaranteed. As a comparison, the proposed CKF+ method avoids the assumption on the joint distribution of design matrix and therefore is more robust to potential misspecifications. We limited the aforementioned discussions to CKF+ and KF+, while the same conclusions also apply when comparing CKF and KF.

Finally, we note that theoretically, the CKF method is only guaranteed to control a modified version of FDR and usually has a higher FDR level than CKF+. In exploratory settings where FDR control is not at a premium, we suggest using CKF as the default for maximal power across all sparsity levels. In non-exploratory settings where users wish to have rigorous FDR control, we suggest the use of CKF+ as the default since CKF+ ensures theoretical finite sample FDR control and still attains high power in a majority of settings.

To summarize, the proposed CSP 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 most scenarios. Compared to CKF methods, other existing methods may either be underpowered (BH and KF methods) or render inappropriate results (CL) by having an inflated FDR than the nominal threshold.

5. Real Data Example

To further demonstrate the usefulness of our method, we applied 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. Thus, we aggregated these OTUs to genus and performed the analysis in the genus level, which may be more robust to potential sequencing errors (Li, 2015). These 7000 OTUs belonged to $p = 303$ distinct genera, whose abundances were the microbial features 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 these three genes (CFI, C2 and CFB) as the outcomes of interest, and applied the proposed CKF method to detect host gene expression-associated taxa for each outcome under the FDR threshold of 0.25. For the initial screening step, we fixed the screening sample size $n_0 = 100$ with screening set size $|\hat{S}_0| = 40$ as done in simulations. As the data-splitting is random, we repeated the CKF algorithm 10 times with different splits and reported those taxa that appeared in the selected set of more than one of the splits. By using multiple split matrices, we were more likely able to identify any possible signals under the desired FDR level.

In Table 3, we report taxa that were identified as host gene expression associated in our analysis. 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 3, we fit the reduced 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 in more than one split. 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). 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. *Escherichia* and *Sutterella* appear in all gene sets, and *Escherichia* in particular was noted by Morgan et al. (2015) to be hugely influential in patients with inflammatory bowel issues. 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, *Epulopiscium* 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 activator. 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 FDR control. Traditional methods usually cast this problem into a multiple testing framework and examine each microbiome feature individually followed by certain multiple testing procedures to control the FDR. To avoid potential heavy multiple adjustment burden, we alternatively adopt a joint regression approach and achieve FDR control via applying the compositional knockoff filter to the regression. As shown in the numerical studies, the proposed CKF method is more powerful than the marginal procedure and can achieve FDR control compared to the compositional lasso method. Furthermore, numerical studies demonstrate a gain in power through employing CKF over the original model-X knockoffs under most settings for microbiome compositional data analysis. It may be more natural to place the burden of knowledge on the response (as in CKF) instead of the features as we have yet been able to develop means to efficiently construct model-X knockoff features for common distributions used for microbiome data analysis. Finally, the CKF application to the host-microbiome data not only identifies most of gene expression-associated taxa found in the original study (Morgan et al., 2015), but also leads to new discoveries, which may provide new biological insights with further laboratory investigation.

As noted by a referee, a wide array of penalized methods have been proposed for the analysis of high-dimensional regression problems. However, methods such as the debiased Lasso (Javanmard and Montanari, 2018; van de Geer, 2019) and the MOCE (Wang et al., 2019) method are not guaranteed to retain the compostional constraint on $\boldsymbol{\beta}$ under the log-contrast model after the debiasing step. The CKF procedure is in the class of "screen and clean" methods (Wasserman and Roeder, 2009; Meinshausen, Meier and Bühlmann, 2009). However, these methods do not account for the underlying sparsity assumption in high-dimensional microbiome compositional analysis and they also do not employ recycling which can lead to a reduction of power. Finally, all these aforementioned methods do not ensure finite sample FDR control which is a key benefit of the CKF procedure.

Currently, CKF can only identify taxa that are associated with a single continuous outcome variable. It is of future interest to extend CKF to more complicated models to accommodate microbiome studies with more complicated designs. In the traditional research vein of p-values based procedures with FDR control for microbiome fine-mapping, there has been a wealth of research interest to utilize additional information (e.g., phylogenetic information) of microbiome data to increase the power of detection while maintaining FDR control (Xiao

et al., 2017; Hu et al., 2018). It is of future interest to incorporate such information in CKF framework to further boost the detection power of controlled variable selection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors wish to thank the editor, associate editor and three referees for their insightful comments and suggestions that have improved the paper. This research was partially supported by the National Institutes of Health grants R21AI144765 and T32GM102057, and National Science Foundation grants DMS-1811552, DMS-1953189 and CCF-2007823.

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.

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

Barber R, and Candès E (2019). A knockoff filter for high-dimensional selective inference. The Annals of Statistics 47(5), 2504–2537.

Bates S, Candès E, Janson L, and Wang W (2020) Metropolized knockoff sampling. Journal of the American Statistical Association DOI: 10.1080/01621459.2020.1729163.

Beale EML, Kendall MG, and Mann DW (1967). The discarding of variables in multivariate analysis. Biometrika 54, 357–366. [PubMed: 6063999]

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. The 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.

Candes E, and Tao T (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35, 2313–2351.

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 Chen Z (2012). Extended BIC for small-n-large-P sparse GLM. Statistica Sinica 22, 555–574.

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

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

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.

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 MJ, and Li H (2018). A two-stage microbial association mapping framework with advanced FDR control. Microbiome 6, 131. [PubMed: 30045760]

Javanmard A, and Montanari A (2017). Debiasing the lasso: optimal sample size for Gaussian designs. The Annals of Statistics 46, 593–2622.

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. [PubMed: 25249709]

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. [PubMed: 30039859]

Meinshausen N, Meier L, Bühlmann P (2009). p-Values for high-dimensional regression Journal of the American Statistical Association 104(488), 1671–1681.

Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, 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. [PubMed: 25887922]

Romano Y, Sesia M, and Candès E (2019). Deep knockoffs. Journal of the American Statistical Association DOI: 10.1080/01621459.2019.1660174.

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

Sohn MB, and Li H (2019). Compositional Mediation Analysis for Microbiome Studies. The Annals of Applied Statistics 13, 661–681.

Tang Z and Chen G (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. Biostatistics, 20(4), 698–713. [PubMed: 29939212]

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

van de Geer S (2019). On the asymptotic variance of the debiased Lasso Electronic Journal of Statistics 13(2), 2970–3008.

Wang H (2009). Forward regression for ultra-high dimensional variable screening Journal of the American Statistical Association 104(488), 1512–1524.

Wang J, and Jia H (2016). Metagenome-wide association studies: fine-mining the microbiome. Nature Reviews Microbiology 14, 508. [PubMed: 27396567]

Wang F, Zhou L, Tang L, and Song P (2019). Method of contraction-expansion (MOCE) for simultaneous unference in linear models. https://arxiv.org/abs/1908.01253

Wasserman L, and Roeder K (2009). High-dimensional variable selection The Annals of Statistics 37, 2178–2201. [PubMed: 19784398]

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5, 27. [PubMed: 28253908]

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. [PubMed: 28505251]

Xu C, and Chen J (2014). The sparse MLE for ultrahigh-dimensional feature screening. Journal of the American Statistical Association 109, 1257–1269. [PubMed: 25382886]

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

Zhan X, Plantinga A, Zhao N, and Wu MC (2017). A fast small-sample kernel independence test for microbiome community-level association analysis. Biometrics 73, 1453–1463. [PubMed: 28295177]

Zhang H, Chen J, Li Z, and Liu L (2019). Testing for Mediation Effect with Application to Human Microbiome Data. Statistics in Biosciences DOI: 10.1007/s12561-019-09253-3

Table 1

Average screening proportions of true signals based on 200 replicates under the Dirichlet-multnomial (DM) distribution and logistic normal (LN) distribution.

Distribution	Screening Method	$\|S^*\| = 15$	$\|S^*\| = 20$	$\|S^*\| = 25$	$\|S^*\| = 30$
DM	CSP	1.000	1.000	1.000	1.000
	PC	0.599	0.497	0.495	0.447
	DC	0.561	0.462	0.464	0.413
LN	CSP	0.994	0.991	1.000	1.000
	PC	0.663	0.577	0.480	0.442
	DC	0.653	0.566	0.467	0.425

Table 2

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

Distribution	Metric	Method	$\lvert S^* \rvert = 15$	$\lvert S^* \rvert = 20$	$\lvert S^* \rvert = 25$	$\lvert S^* \rvert = 30$
DM	$\widehat{\text{FDR}}$	CKF	0.097	0.124	0.113	0.098
		CKF+	0.048	0.082	0.080	0.076
		KF	0.122	0.117	0.110	0.108
		KF+	0.068	0.084	0.079	0.084
		CL	0.814	0.783	0.670	0.620
		BH	0.106	0.095	0.100	0.102
	$\widehat{\text{Power}}$	CKF	0.939	0.944	0.955	0.960
		CKF+	0.897	0.914	0.946	0.958
		KF	0.999	0.998	0.953	0.931
		KF+	0.990	0.974	0.881	0.851
		BH	0.626	0.547	0.445	0.385
LN	$\widehat{\text{FDR}}$	CKF	0.132	0.107	0.102	0.102
		CKF+	0.073	0.064	0.070	0.075
		KF	0.101	0.115	0.101	0.090
		KF+	0.064	0.070	0.062	0.054
		CL	0.825	0.797	0.778	0.765
		BH	0.094	0.108	0.097	0.087
	$\widehat{\text{Power}}$	CKF	0.954	0.961	0.968	0.968
		CKF+	0.881	0.907	0.946	0.935
		KF	0.849	0.755	0.691	0.577
		KF+	0.730	0.582	0.555	0.457
		BH	0.521	0.426	0.475	0.409

Table 3

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

Gene	Taxa	Coefficient	Gene	Taxa	Coefficient
CFI	***Escherichia***	0.0312	C2	***Escherichia***	0.0376
	Sutterella	−0.0362		***Sutterella***	−0.0285
	Akkermansia	−0.0108		*Turicibacter*	−0.0212
	Bifidobacterium	−0.0189		*Lachnospira*	0.0332
	Clostridium	−0.0199		*Veillonella*	0.0293
	Prevotella	−0.0140		*Brevundimonas*	0.0424
	C. Clostridium	−0.0257		*Anaerococcus*	−0.0246
	L. Clostridium	−0.0257		*Bulleidia*	−0.0336
	R. Clostridium	0.0234		*Rhodoplanes*	0.0434
	Epulopiscium	0.0062		*Staphylococcus*	0.0198
	Dorea	−0.0118	CFB	***Escherichia***	0.0437
	Lachnospira	−0.0118		***Sutterella***	−0.0450
	Veillonella	0.0203		***Bifidobacterium***	−0.0144
	Actinomyces	−0.0264		*Epulopiscium*	0.0202
	Collinsella	−0.0073		*Lachnospira*	0.0195
	Staphylococcus	0.0449		*Collinsella*	−0.0167
	Brevundimonas	0.0731		*Eggerthella*	0.0809
	Finegoldia	−0.0336		*Enterococcus*	−0.0132
	R. Eubacterium	0.0506			
	E. Eubacterium	−0.1001			
	Enterococcus	−0.0061			
	Peptostreptococcus	0.0190			