



## 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; Wang and Jia, 2016; Mitchell et al., 2017; Schneider et al., 2020). The advent of next-generation sequencing technologies enables studying the microbiome composition via direct sequencing of microbial DNA without the need for 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 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 a specific response of interest (e.g., body mass index or a 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 the marginal association between each microbial taxon and the response followed by multiple testing corrections to identify taxa with adjusted p-values below a certain FDR threshold as important ones that influence the response (Wang and Jia, 2016; Xiao, Chao, and Chen, 2017). The marginal approach is often limited to microbiome data analysis 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

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

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

45 To address the potential limitations in existing marginal and joint approaches, a new  
46 method in a joint regression framework to select microbial taxa with finite-sample FDR  
47 control is desired. In the statistics literature, the FDR control can be achieved via the  
48 knockoff filter framework, in which a dummy knockoff copy of the original design matrix  
49 with the same covariance structure has been constructed and flagged as false positives to  
50 facilitate FDR-controlled variable selection (Barber and Candès, 2015). However, it has been  
51 observed, in the literature of many other statistical inference methods (e.g., regression-based  
52 modeling, two-sample testing, and statistical causal mediation analysis), that applying classic  
53 statistical methods to analyze microbiome composition data is usually underpowered and

54 sometimes can render inappropriate results (Aitchison, 2003; Shi, Zhang, and Li, 2016; Cao,  
55 Lin, and Li, 2017; Sohn and Li, 2019; Lu, Shi and Li, 2019; Zhang et al., 2019). Thus, new  
56 FDR-controlled variable selection methods are desired for microbiome compositional data.

57 Following the pioneering work of Aitchison and Bacon-shone (1984), we model all taxa  
58 jointly in a linear log-contrast model to address the compositional nature of data and  
59 propose a two-step regression-based FDR-controlled variable selection procedure named  
60 compositional knockoff filter (CKF) to identify response-associated taxa. In the first step,  
61 we introduce the compositional screening procedure (CSP) as a new method of variable  
62 screening for high-dimensional microbiome data subject to the compositional constraint.  
63 In the second step, we apply the fixed-X knockoff procedure (Barber and Candès, 2015)  
64 to the reduced model in the first screening step. The theoretical properties of the novel  
65 compositional screening procedure are investigated. Using numerical studies, we demonstrate  
66 that the proposed method can jointly assess the significance of microbial covariates while also  
67 theoretically ensuring finite-sample FDR control. The proposed method will greatly benefit  
68 downstream microbiome functional studies by enhancing the reproducibility and reliability  
69 of discovery results in microbiome association studies.

70 Our primary contributions are summarized as follows. First, we introduce the CSP to  
71 screen true signals from high-dimensional compositional data and theoretically verify that  
72 CSP attains the desirable sure screening property under mild assumptions. As demonstrated  
73 in thorough simulation, the newly proposed CSP yields a much higher likelihood of attaining  
74 all true signals compared to some existing methods that do not account for the compositional  
75 nature. Second, by leveraging the high-dimensional knockoff filter framework (Barber and  
76 Candès, 2019), we avoid the non-trivial sequential conditional independent pairs algorithm  
77 of model-X knockoffs (Candès et al., 2018) and provide an alternative CKF approach to  
78 ensure strong finite-sample FDR control for microbial taxa selection. Construction of model-

79 X knockoff features through methods such as the sequential conditional independent pairs  
80 algorithm (Candès et al., 2018) requires both complete knowledge of the joint distribution  
81 of the microbiome design matrix and repeated derivation of the conditional distributions,  
82 that are non-trivial for many non-Gaussian distributions such as Dirichlet-multinomial and  
83 logistic normal, which are frequently used in modeling microbiome data (Aitchison, 2003;  
84 Chen and Li, 2013; Lin et al., 2014; Tang and Chen, 2018; Harrison et al., 2020). While the  
85 development of methods to construct exact or approximate knockoff features for a broader  
86 class of distributions is a promising area of active research (Bates et al., 2019; Romano, Sesia,  
87 and Candès, 2019), the robustness of how model-X knockoff to the departure of the joint  
88 distribution from multivariate Gaussian is currently unknown. To this end, the proposed CKF  
89 with finite-sample FDR control guarantee is appealing through versatility for microbiome  
90 taxa selection.

91 The rest of this paper is organized as follows. We propose the methodology of composi-  
92 tional knockoff filter in Section 2. The theoretical properties of the compositional screening  
93 procedure are investigated in Section 3. The numerical properties are demonstrated through  
94 simulation studies in Section 4 and application to a microbiome data set collected from an  
95 inflammatory bowel disease study in Sections 5. Technical proofs and additional numerical  
96 evaluations are deferred to the online supporting information.

## 97 **2. Compositional Knockoff Filter**

98 This section presents the compositional knockoff filter to perform FDR-controlled variable  
99 selection analysis for microbiome compositional data. The proposed method aims to address  
100 the high-dimensional compositional nature of microbiome data (i.e.,  $p > n$ ). To this end, we  
101 follow the philosophy of recycled fixed-X knockoff procedure (Barber and Candès, 2019) to  
102 develop a new two-step procedure for high-dimensional compositional data, which consists  
103 of a compositional screening step and then a subsequent selection step. After introducing

104 the log-contrast model in Section 2.1, we will present the screening step in Section 2.2 and  
 105 the selection step in Section 2.3.

## 106 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 compositional data (Aitchison, 2003; 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 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 the log-ratio transformation of  $\mathbf{X}$ , where  $Z_{ij}^p = \log(X_{ij}/X_{ip})$  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  $(\beta_1, \beta_2, \dots, \beta_{p-1})$  and error  $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ . 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,

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

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

## 2.2 Compositional Screening Procedure

As the fixed-X knockoff requires that  $n \geq 2p$ , 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  $\hat{S}_0 \subset \{1, \dots, p\}$  such that  $|\hat{S}_0| \leq \frac{n_1}{2}$ , where  $|\hat{S}_0|$  denotes the cardinality of set  $\hat{S}_0$ . Throughout this paper, we always assume  $|\hat{S}_0| \leq \frac{n_1}{2}$  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 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:

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

133 The proposed compositional screening problem (2) can also be viewed as maximizing the log-  
 134 likelihood  $\ell_n(\boldsymbol{\beta})$  under the sparsity constraint  $\|\boldsymbol{\beta}\|_0 \leq k$  (Xu and Chen, 2014). The choice of  $k$   
 135 is a fundamental question in many high-dimensional screening procedures. Practical domain  
 136 knowledge may provide information on how sparse one believes the underlying signal is.  
 137 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,  
 138 2008; Li et al., 2012). However, as noted by Li et al. (2012), the choice of screening set size  
 139 can be viewed as a tuning parameter within the model and concrete means to determine the  
 140 screening set size are an area of future development. Although (2) is a NP-hard problem,  
 141 the mixed integer optimization (MIO) allows us to approximately solve the global solution  
 142 of the nonconvex optimization problem (2) in an efficient manner (Konno and Yamamoto,  
 143 2009; Bertsimas, King and Mazumder, 2016). Finally, we demonstrate in the Section 3 that  
 144 the computed solution of (2) by MIO attains the desirable sure screening guarantees.

145 After screening, the model reduces to  $y_i = \sum_{j \in \hat{S}_0} Z_{ij} \beta_j^r + \varepsilon_i$  subject to  $\sum_{j \in \hat{S}_0} \beta_j^r = 0$ .  
 146 Comparing it to the original log-contrast model (1), the regression coefficients in the reduced  
 147 model  $\beta_j^r$  does not necessarily match  $\beta_j$  in the original model. To solve this discrepancy,  
 148 we propose 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  
 149 of notation, we still use  $Z_{ij} = \log(X_{ij}^*)$  to denote the design matrix to be used in the  
 150 subsequent selection step. Details about this normalization is available at Section S.1 of the  
 151 online supporting information.

### 152 2.3 Controlled Variable Selection

Let  $\mathbf{Z}_{\hat{S}_0}^{(1)} \in \mathbb{R}^{n_1 \times |\hat{S}_0|}$  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  $\tilde{\mathbf{Z}}_{\hat{S}_0}^{(1)}$  is constructed using  $\mathbf{Z}_{\hat{S}_0}^{(1)}$  following the fixed-X knockoff framework (Barber and Candès, 2015). The primary assumption of the fixed-X framework is that the burden of knowledge is placed on the design and  $\mathbf{Z}^{(1)}$  is assumed to be fixed and the response is generated via a linear Gaussian model. Notably, the fixed-X knockoff places no assumptions on knowing the noise level. Thus, a key appeal of the knockoff filter is the relative lack of strong assumptions needed for theoretical finite-sample control to hold. We refer to Barber and Candès (2015) for a review of the construction of knockoff matrix and for a deeper study into the assumptions needed by the knockoff filter. 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 one of size at most  $\frac{n_1}{2}$ . Thus  $\mathbf{Z}_{\hat{S}_0}^{(1)}$  is of dimension at most  $n_1 \times \frac{n_1}{2}$ . As the knockoff matrix is constructed on  $\mathbf{Z}_{\hat{S}_0}^{(1)}$  alone, this satisfies the dimensionality requirements for the construction of the fixed-X knockoff matrix  $\tilde{\mathbf{Z}}_{\hat{S}_0}^{(1)}$ . To further boost the power of the procedure, we followed data recycling mechanism outlined in Barber and Candès (2019), we construct the recycled knockoff matrix as

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

153 Note that we treat  $(\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)})$  as fixed after the screening step and the first part of knockoff  
 154 copies are exact copies under the recycling scheme (Barber and Candès, 2019).

We now 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 knockoff

copies. With this new augmented design matrix, we solve the following Lasso problem:

$$\bar{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbb{Z}_{\hat{S}_0} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \quad (3)$$

155 where  $\bar{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$  is a vector appending the coefficients of original features and knockoff  
 156 features. Other penalties such as the folded concave penalties (Fan and Li, 2001; Fan et  
 157 al., 2014) may also be used for the purpose of variable selection. For ease of presentation,  
 158 we focus on the Lasso problem (3), as existing methods (Lin et al., 2014) do not provide  
 159 a rigorous FDR control on the selected variables. Comparing to previous problems (1) and  
 160 (2), we no longer require a sum-to-zero constraint in our augmented Lasso problem (3).  
 161 This is because, by adding  $|\hat{S}_0|$  knockoff features in the augmented design matrix  $\mathbb{Z}_{\hat{S}_0}$ , the  
 162 corresponding microbiome data matrix  $\mathbb{X}_{\hat{S}_0} = \exp(\mathbb{Z}_{\hat{S}_0})$  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  $\{\bar{\boldsymbol{\beta}}(\lambda)\} = \{(\hat{\boldsymbol{\beta}}(\lambda), \tilde{\boldsymbol{\beta}}(\lambda))\}$ . 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  $j$ :

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

163 A large and positive  $W_j$  would suggest strong evidence that the original feature is significantly  
 164 outcome-associated as an important feature tends to remain longer in lasso path as  $\lambda$   
 165 increases. Similarly, a negative or zero  $W_j$  value would indicate that the covariate tends to  
 166 be noise. Thus,  $W_j$  is used to calculate the data-dependent knockoff thresholds that ensure  
 167 finite sample FDR-controlled variable selection. The choice of an  $\ell_1$ -penalization problem  
 168 (3) used in the selection step is driven by the need to accurately compute a solution path  
 169 for construction of the knockoff statistic. As a comparison, the solution of  $\ell_0$ -penalization

170 problem (2) via MIO does not yield a solution path and is unsuitable for knockoff test statistic  
 171 construction. Finally, both the standard knockoff and knockoff+ thresholds are considered  
 172 for the purpose of selection:

KNOCKOFF THRESHOLD:

$$T = \min \left\{ t \in \mathcal{W} : \frac{|\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|} \leq q \right\}, \quad (5)$$

KNOCKOFF+ THRESHOLD:

$$T = \min \left\{ t \in \mathcal{W} : \frac{1 + |\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|} \leq q \right\}, \quad (6)$$

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

### 179 3. Theoretical Properties

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

---

**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:  $\mathbb{Z}_{\hat{S}_0} = [\mathbf{Z}_{\hat{S}_0} \quad \tilde{\mathbf{Z}}_{\hat{S}_0}]$ .
    - (b) Solve equation (3) to calculate the coefficients  $\bar{\beta}(\lambda)$ .
    - (c) Calculate knockoff statistics  $W_j$  from  $\bar{\beta}_j(\lambda)$ .
    - (d) Use the knockoff or knockoff+ threshold (5) and (6) to calculate  $T$  from  $\mathcal{W}$ .
    - (e) Determine the knockoff or knockoff+ selection set as  $\hat{S} = \{j : W_j \geq T\}$ .
- 

### 189 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  $\{1, \dots, p\}$  corresponding to a sub-model with coefficients  $\beta_s$ , and  $S^*$  be the true model with  $p^*$  nonzero coefficients, with corresponding true coefficient vector  $\beta^*$ . 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  $\mathbf{S}_+^k = \{s : S^* \subset s\}$ .

$s; \|s\|_0 \leq k\}$  denote the set of all overfit models and  $\mathbf{S}_-^k = \{s : S^* \not\subset s; \|s\|_0 \leq k\}$  denote the set of underfit models. We will show that the compositional screening procedure does not miss true signals with high probability. That is:

$$P(S^* \subset \hat{S}_0) \rightarrow 1 \text{ as } n \rightarrow \infty. \quad (7)$$

190 For the technical aspects of our sure-screening proof to hold, we make the following assump-  
191 tions (1-4), encompassing requirements on the dimension, signal strength and microbiome  
192 design matrix:

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

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

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

199 ASSUMPTION 4: There exist constants  $c_2 > 0$  and  $c_3 > 0$  such that  $|Z_{ij}| \leq c_2$  and  
200  $\max_{1 \leq j \leq p} \max_{1 \leq i \leq n} \left\{ \frac{Z_{ij}^2}{\sum_{i=1}^n Z_{ij}^2 \sigma_i^2} \right\} \leq c_3 n^{-1}$  when  $n$  is sufficiently large, where  $\sigma_i^2 = \text{Var}(\mathbf{Y}|\mathbf{Z})$ .

201 Assumption 1 places a weak restriction on  $p$  and  $n$  of the data, which is very likely to be  
202 met in many microbiome studies (Wang and Jia, 2016). Assumption 2 places a restraint on  
203 the minimum strength of true signals, such that they are discoverable. This assumption is  
204 common for statistical screening and variable selection methods (Fan and Lv, 2008; Fan and  
205 Song, 2010; Lin et al., 2014). Assumption 3 corresponds to the UUP condition (Candes and  
206 Tao, 2007) which controls the pairwise correlations between the columns of  $\mathbf{Z}$ . This condition  
207 is prevalent across many high-dimensional variable selection methods such as the Dantzig  
208 selector (Candes and Tao, 2007), SIS-DS (Fan and Lv, 2008), forward regression (Wang,

209 2009), and the sparse MLE (Xu and Chen, 2014). We have conducted numerical studies to  
 210 evaluate the applicability of Assumption 3 in the context of microbiome data settings in  
 211 Section S.2 of the online supporting information. Finally, as noted by Xu and Chen (2014)  
 212 and Chen and Chen (2012), Assumption 4 likely will hold for a wide class of design matrices  
 213 as long  $\sigma_i^2$  is not degenerate. In Section S.2 of the online supporting information, we illustrate  
 214 the validity of Assumption 4 on the mucosal microbiome data analyzed later in this paper.  
 215 Further details on these assumptions are available in Section S.2 of the online supporting  
 216 information. Under Assumptions 1–4, Theorem 1 shows that the proposed compositional  
 217 screening procedure attains the sure screening property. The proof of Theorem 1 relies on  
 218 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  $\mathbf{S}_+^k = \{s : S^* \subset s; \|s\|_0 \leq k\}$ . Assume that Assumptions 1–4 hold and  $\tau_1 + \tau_2 < \frac{(1-m)}{2}$ . Then:*

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

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

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

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

223 Lemma 2 demonstrates that the computed solution of the compositional screening proce-

224 dure through mixed integer optimization converges to the global solution of the constrained  
 225 sparse maximum likelihood problem with high probability. By combining Lemma 1 and  
 226 Lemma 2, it follows that the computed solution attains the sure screening property. This  
 227 result 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  $p^* < |\hat{S}_0| < p$ . Then:*

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

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

### 231 3.2 FDR Control Properties of Compositional Knockoff Filter

232 In this section, we briefly outline the FDR control properties of CKF. In order to control  
 233 FDR, the knockoff statistic must obey the *anti-symmetry* and *sufficiency* properties while the  
 234 design matrix and response must satisfy both the *Pairwise Exchangeability for the Response*  
 235 *Lemma* and *Pairwise Exchangeability for the Features Lemma* (Barber and Candès, 2015).  
 236 In this paper we primarily focus on the LSM knockoff statistic (4) which has been shown to  
 237 satisfy the *anti-symmetry* and *sufficiency* properties (Barber and Candès, 2019). Therefore,  
 238 the FDR control properties of CKF are a direct consequence of the FDR control theory  
 239 outlined in Barber and Candès (2015) and Barber and Candès (2019) and we reiterate the  
 240 FDR control property here for posterity. As we have validated that the CSP attains the  
 241 sure-screening property, the compositional knockoff+ threshold ensures finite sample FDR  
 242 control as stated in the following Theorem 2.

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

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

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

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

REMARK 1: Given the above exchangeability results and the previous theorems, the stan-  
 dard knockoff threshold controls a modified form of false discovery rate (Barber and Candès,  
 2015). In particular, for  $q \in [0, 1]$ , the knockoff method ensures:

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

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

## 257 4. Simulation Studies

258 We conducted two sets of simulation studies (screening simulation and selection simulation)  
 259 to evaluate numerical performance of the proposed CKF methods. In the screening simu-  
 260 lation, we evaluated the sure screening property of the proposed CSP. We compared CSP



261 to two other popular statistical screening procedures in literature: one based on Pearson  
262 correlation/PC (Fan and Lv, 2008) and the other based on distance correlation/DC (Li et al.,  
263 2012). We considered a sample size of  $n_0 = 100$  and screening set size  $|\hat{S}_0| = 40 \approx 2 \lfloor \frac{n_0}{\log(n_0)} \rfloor$ .  
264 In the selection simulation, we evaluated the selection performance of CKF methods. For  
265 comparison, we also consider other methods that are widely used for microbial taxa selection.  
266 One is the compositional Lasso (Lin et al., 2014) and the other is the marginal method  
267 which examines one taxon at a time followed by the Benjamini-Hochberg procedure for FDR  
268 control (Benjamini and Hochberg, 1995; Paulson et al., 2013; Parks et al., 2014). We also  
269 compared the proposed CKF to the original model-X knockoff filter method (Candès et al.,  
270 2018) in this selection simulation. To mimic a real dataset analyzed later in this paper, we  
271 considered sample size  $n = 250$  and number of microbiome covariates  $p = 400$  in the selection  
272 simulations. Among these  $n = 250$  samples, a randomly selected sub-sample with  $n_0 = 100$   
273 observations were used in the first screening step and the rest  $n_1 = 150$  observations were  
274 used for the selection step.

275 Two schemes have been used to generate the microbiome compositional design matrix used  
276 in both simulations. The first scheme was to generate microbiome counts from the Dirichlet-  
277 multinomial (DM) distribution, whose parameters were estimated from a real microbiome  
278 data set following previous designs (Zhao et al., 2015). The library size of each sample was  
279 randomly simulated from a negative binomial distribution with a mean parameter of 10000  
280 and dispersion parameter of 25. Raw zero counts were first replaced by a pseudo count of 0.5,  
281 as commonly suggested in microbiome data analysis (Lin et al., 2014; Cao et al., 2017; Weiss  
282 et al., 2017; Lu et al., 2019; Zhang et al., 2019) and then counts were transformed to relative  
283 abundances. The second scheme for generating microbiome compositional data was to use  
284 the logistic normal (LN) distribution, which is also widely used to generate compositional  
285 data (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  $i, j = 1, \dots, p$ . 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  $i = 1, \dots, n, j = 1, \dots, p$ .

Next, we varied the sparsity levels  $|S^*| \in \{15, 20, 25, 30\}$  and set the first 30 entries  $\boldsymbol{\beta}_{1:30}$  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 regression  
coefficients  $\boldsymbol{\beta}_{31:400}$  were all set to be 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  $\boldsymbol{\beta}_{S^*} = \boldsymbol{\beta}_{1:|S^*|}$ ,  $|S^*| \in \{15, 20, 25, 30\}$  and  $\varepsilon \sim \mathcal{N}(0, I)$ .

#### 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,  $|\hat{S}_0 \cap S^*|/|S^*|$ , 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.

[Table 1 about here.]

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 one based on Table

311 1. It is of note that the performance of screening is crucial to the subsequent selection  
 312 inference. To show this, we have further conducted additional numerical studies to compare  
 313 the performance of CKF and CKF+ with three different screening procedures at a target  
 314 nominal FDR of 0.1. The results are reported in Table S4 and Table S5 in Section S.4 of the  
 315 online Supporting Information.

#### 316 4.2 Selection Simulation

In this section, we compared CKF/CKF+ to some existing methods including the model-X knockoff filter (KF) methods (Candès et al., 2018), compositional lasso (CL) method (Lin et al., 2014) and Benjamini-Hochberg (BH) procedure. 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., 2019). Therefore, we employ the use of Gaussian model-X knockoffs as used previously (Candès et al., 2018) in this simulation. For the CL method, the optimal  $\lambda$  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 selection methods, empirical FDR and empirical power were calculated.

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

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

320 As observed from Table 2, CKF+, KF+ and BH can control the nominal FDR level, which  
 321 is desired. CKF and KF yield slightly inflated FDR levels above the nominal rate, but this  
 322 is expected as both KF and CKF are only guaranteed to control a modified version of the  
 323 FDR (Remark 1 of Theorem 2). Finally, CL has a high empirical false discovery rate across  
 324 all scenarios. The Lasso method has proven to be a versatile tool with appealing estimation  
 325 and selection properties in the asymptotic setting (Tibshirani, 1996; Lin et al., 2014). Yet, its  
 326 performance under finite sample setting is not guaranteed. Our results on CL is consistent  
 327 with the fact that a relatively large number of false positives are reported in Table 1 of Lin  
 328 et al. (2014). Despite being able to guarantee model selection consistency, CL tends to select  
 329 more unnecessary variables to recover the true model.

330 Since CL has an extremely inflated FDR, it is not meaningful to compare its power to the  
 331 other methods that can control FDR and hence power of CL is not reported. Comparing  
 332 the empirical power of methods with FDR control in Table 2, both CKF+ and KF+ are  
 333 much more powerful than BH. This power gap is likely due to the fact that CKF+ and KF+  
 334 analyze the microbial covariates jointly, and the effectiveness of the marginal BH method  
 335 deteriorates when the dimension (or multiple correction burden) is relatively high. Under DM  
 336 distribution, KF+ achieves as higher power than CKF+ in sparse setting ( $|S^*| = 15$  or  $20$ ).  
 337 However, CKF+ becomes more powerful over KF+ as the signal becomes denser ( $|S^*| = 25$  or  
 338  $30$ ). On the other hand, under LN distribution, the effectiveness of KF+ quickly deteriorates  
 339 and CKF+ is much more powerful than KF+ based on Table 2. The KF+ method generated  
 340 knockoffs based on an underlying Gaussian assumption on the covariates, and therefore its  
 341 performance under the microbiome setting (e.g., Dirichlet-multinomial and logistic normal  
 342 distributions considered in our simulations) may not be guaranteed. As a comparison, the  
 343 proposed CKF+ method avoids the assumption on the joint distribution of design matrix and  
 344 therefore is more robust to potential model misspecification. We limited the aforementioned

345 discussions to CKF+ and KF+, while the same conclusions also apply when comparing CKF  
346 and KF.

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

353 To summarize, the proposed CSP enjoys the sure screening property, which is crucial to  
354 guarantee a high power of the downstream selection analysis. Our CKF methods successfully  
355 control the FDR of selecting outcome-associated microbial features in a regression-based  
356 manner which jointly analyzes all microbial covariates, while having the highest power  
357 detecting outcome-associated microbes. CKF methods are more robust than KF methods  
358 modelling the microbiome compositional data and are more powerful than the corresponding  
359 KF methods under most scenarios. Compared to CKF methods, Other existing methods may  
360 either be underpowered (BH) or render inappropriate results (CL) by having an inflated FDR  
361 than the nominal threshold.

## 362 5. Real Data Example

363 To further demonstrate the usefulness of our method, we apply it to a real data set ob-  
364 tained from a study examining the association between host gene expression and mucosal  
365 microbiome using samples collected from patients with inflammatory bowel disease (Morgan  
366 et al., 2015). The abundances of 7000 OTUs from  $n = 255$  samples were measured using  
367 16S rRNA gene sequencing and most to these 7000 species-level OTUs were in extremely  
368 low abundances with a large proportion of OTUs being simply singletons. As suggested in  
369 literature (Li, 2015), we aggregated these OTUs to genus and perform the analysis in the

370 genus level, which may be more robust to potential sequencing errors. These 7000 OTUs  
 371 belonged to  $p = 303$  distinct genera, whose abundances were the microbial covariates of  
 372 interest in our analysis.

373 It has been previously found that microbially-associated host transcript pattern is enriched  
 374 for complement cascade genes, such as genes CFI, C2, and CFB (Morgan et al., 2015).  
 375 Moreover, principal component-based enrichment analysis shows that host gene expression  
 376 is inversely correlated with taxa *Sutterella*, *Akkermansia*, *Bifidobacteria*, *Roseburia* abun-  
 377 dance and positively correlated with *Escherichia* abundance under the nominal FDR of 0.25  
 378 (Morgan et al., 2015). In this analysis, we took the expression values of complement cascade  
 379 genes (CFI, C2, and CFB) as the outcomes of interest, and applied the proposed CKF and  
 380 CKF+ method to detect host gene expression-associated genera for each outcome under the  
 381 FDR threshold of 0.25. For the initial screening step, we fixed the screening sample size  
 382  $n_0 = 100$  with screening set size  $|\hat{S}_0| = 40$  as done in simulations. As the data-splitting is  
 383 random, we repeated the CKF algorithm 10 times with different splits and report those taxa  
 384 that appeared in more than one of the splits. By using multiple split matrices, we were more  
 385 likely able to identify any possible signals under the desired FDR level.

386 [Table 3 about here.]

387 In Table 3, we report taxa that were identified as host gene expression associated in  
 388 our analysis. Taxa in bold were also identified in the original paper (Morgan et al., 2015)  
 389 using marginal method to control the FDR at 0.25. For the coefficient column of Table  
 390 3, we fit the reduced linear regression models with predictors of both selected taxa and  
 391 clinical variables including disease subtype, antibiotic use, tissue location and inflammatory  
 392 score, as done previously (Morgan et al., 2015). These clinical variables were included  
 393 in the model to adjust for potential confounding effects and to obtain a more accurate  
 394 estimate of the microbiome effect on host gene expression. The sign of a taxon coefficient

395 reflects the direction of association (activation or inhibition). Recall that five taxa *Sutterella*,  
396 *Akkermansia*, *Roseburia*, *Bifidobacterium* and *Escherichia* were detected in the original  
397 principal component-based marginal analysis (Morgan et al., 2015). All these five except  
398 *Roseburia* were identified in our analysis in more than one split. Moreover, we further see  
399 that the coefficient signs for each taxa of interest are consistent with the expected direction  
400 posited by Morgan et al. (2015). In other words, we correctly identify a majority of taxa  
401 of interest function as inhibitors (negative coefficient) or activators (positive coefficient) for  
402 each cascade gene expression.

403 We also observe that the taxa set identified for each cascade gene are different, which  
404 suggests that specific taxa play key roles on individual gene expression. *Escherichia* and  
405 *Sutterella* appear in all gene sets, and *Escherichia* in particular was noted by Morgan et al.  
406 (2015) to be hugely influential in patients with inflammatory bowel issues. Despite that we  
407 missed taxa *Roseburia* compared to the original analysis, many new taxa were identified as  
408 complement cascade gene expression-associated in our CKF analysis. For example, *Epulop-*  
409 *iscium* appears in the selection sets for both the CFB and CFI as an inhibitor which may  
410 be of particular interest. Likewise, *Lactobacillus* appears in both the CFB gene and C2 gene  
411 acting as an activator. On the other hand, *Lachnospira* appears to be an activator for both  
412 C2 and CFB but an inhibitor for CFI. The mechanism of how these new taxa affect the  
413 host transcript pattern warrants further laboratory investigation.

414 To conclude, the proposed CKF is more powerful in detecting significant taxa than the  
415 original principal component-based marginal analysis (Morgan et al., 2015) under the same  
416 nominal FDR of 0.25. Our new method not only provides additional statistical support to  
417 results obtained from the original analysis but also gains new biological and biomedical  
418 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, the proposed CKF method is more powerful than the marginal BH procedure, and can achieve false discovery control compared to the compositional lasso method. Further numerical study demonstrates a gain in power through employing CKF over the original model-X knockoffs under the logistic normal setting and denser signal scenarios under the Dirichlet-multnomial settings. CKF is extremely useful for microbiome compositional data analysis, as it may be more natural to place the burden of knowledge on the response 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 analysis. Finally, the application our method to the host-microbiome data not only identifies most of gene expression-associated taxa that were identified 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. Methods such as the debiased Lasso (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2018; van de Geer, 2019) and the MOCE (Wang et al., 2019) method are not guaranteed to retain the compositional constraint on  $\beta$  under the log-contrast model after the debiasing step. However, it is of future interest to study debiasing methods that retain the sum constraint. The CKF



444 procedure is in the class of "screen and clean" methods such as Wasserman and Roeder (2009)  
445 and Meinshausen et al. (2009). However, Meinshausen et al. (2009) does not account for  
446 the underlying sparsity assumption in high-dimensional microbiome compositional analysis.  
447 Further, these methods do not employ recycling which can lead to a reduction of power,  
448 which is especially pronounced in Wasserman and Roeder (2009) which relies on a three-way  
449 sample split. Finally, the aforementioned methods do not ensure finite sample FDR control  
450 which is a key benefit of the CKF procedure.

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

463

#### ACKNOWLEDGEMENTS

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

## REFERENCES

- 468  
469 Aitchison, J., and Bacon-shone, J. (1984). Log contrast models for experiments with  
470 mixtures. *Biometrika* **71**, 323–330.
- 471 Aitchison, J. (2003). The statistical analysis of compositional data. Caldwell, New Jersey:  
472 Blackburn Press.
- 473 Barber, R., and Candès E. (2015). Controlling the false discovery rate via knockoffs. *The*  
474 *Annals of Statistics* **43**, 2055–2085.
- 475 Barber, R., and Candès E. (2019). A knockoff filter for high-dimensional selective inference.  
476 *The Annals of Statistics* *47*(5), 2504-2537.
- 477 Bates, S., Candès E., Janson, L., and Wang W. (2019) Metropolized knockoff sampling.  
478 <https://arxiv.org/pdf/1903.00434.pdf>.
- 479 Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in  
480 multivariate analysis. *Biometrika* **54**, 357-366.
- 481 Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and  
482 powerful approach to multiple testing. *Journal of the royal statistical society. Series B*  
483 **57**, 289-300.
- 484 Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern  
485 optimization lens. *The Annals of Statistics* **44**, 813-852.
- 486 Cao, Y., Lin, W., and Li, H. (2017). Two-sample tests of high-dimensional means for  
487 compositional data. *Biometrika* **105**, 115–132.
- 488 Candès, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much  
489 larger than n. *The Annals of Statistics* *35*, 2313-2351.
- 490 Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for  
491 high dimensional controlled variable selection. *Journal of the Royal Statistical Society:*  
492 *Series B* **80**, 551-577.

- 493 Chen, J., and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica*  
494 *Sinica* **22**, 555-574.
- 495 Chen, J., and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression  
496 with an application to microbiome data analysis. *The Annals of Applied Statistics* **7**,  
497 418-442.
- 498 Cho, I., and Blaser, M. J. (2012). The human microbiome: at the interface of health and  
499 disease. *Nature Reviews Genetics* **13**, 260-270.
- 500 Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its  
501 oracle properties. *Journal of the American statistical Association* **96**, 1348-1360.
- 502 Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature  
503 space. *Journal of the Royal Statistical Society: Series B* **70**, 849-911.
- 504 Fan, J., and Song, R. (2010). Sure independence screening in generalized linear models with  
505 NP-dimensionality. *The Annals of Statistics* **38**, 3567-3604.
- 506 Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized  
507 estimation. *The Annals of Statistics* **42**, 819-849.
- 508 Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear  
509 models via coordinate descent. *Journal of Statistical Software* **33**(1), 1-22.
- 510 Harrison, J.G., Calder, W.J., Shastry, V. and Buerkle, C.A. (2020). Dirichletmultinomial  
511 modelling outperforms alternatives for analysis of microbiome and other ecological count  
512 data. *Molecular Ecology Resources* Accepted Author Manuscript. doi:10.1111/1755-  
513 0998.13128
- 514 Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2017). A broken promise: microbiome  
515 differential abundance methods do not control the false discovery rate. *Briefings in*  
516 *bioinformatics* **20**, 210-221.
- 517 Hu, J., Koh, H., He, L., Liu, M., Blaser, M. J., and Li, H. (2018). A two-stage microbial

- 518 association mapping framework with advanced FDR control. *Microbiome* **6**, 131.
- 519 Jiang, L., Amir, A., Morton, J. T., Heller, R., Arias-Castro, E., and Knight, R. (2017). Dis-  
520 crete False-Discovery Rate Improves Identification of Differentially Abundant Microbes.  
521 *MSystems* **2**, e00092-17.
- 522 Javanmard, A., and Montanari A. (2017). Debiasing the lasso: optimal sample size for  
523 Gaussian designs. *The Annals of Statistics* **46**(6A), 593–2622.
- 524 Konno, H., and Yamamoto, R. (2009). Choosing the best set of variables in regression analysis  
525 using integer programming. *Journal of Global Optimization* **44**, 273-282.
- 526 Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning.  
527 *Journal of the American Statistical Association* **107**, 1129-1139.
- 528 Li, H. (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Anal-  
529 ysis. *Annual Review of Statistics and Its Application* **2**, 73–94.
- 530 Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with composi-  
531 tional covariates. *Biometrika* **101**, 785–797.
- 532 Lu, J., Shi, P., and Li, H. (2019). Generalized linear models with linear constraints for  
533 microbiome compositional data. *Biometrics* **75**, 235-244.
- 534 Meinshausen, N., Meier, L., Bühlmann, P. (2009). p-Values for high-dimensional regression  
535 *Journal of the American Statistical Association* **104**(488),1671–1681.
- 536 Mitchell, C. M., Srinivasan, S., Zhan, X., Wu, M. C., Reed, S. D., Guthrie, K. A., et al.  
537 (2017). Vaginal microbiota and genitourinary menopausal symptoms: a cross-sectional  
538 analysis. *Menopause* **24**, 1160–1166.
- 539 Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R.,  
540 et al. (2015). Associations between host gene expression, the mucosal microbiome, and  
541 clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome*  
542 *Biology* **16**, 67.

- 543 Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical  
544 analysis of taxonomic and functional profiles. *Bioinformatics* **30**, 3123-3124.
- 545 Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance  
546 analysis for microbial marker-gene surveys. *Nature Methods* **10**, 1200.
- 547 Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R. R. and Wu, M. C. (2017). MiRKAT-  
548 S: a community-level test of association between the microbiota and survival times.  
549 *Microbiome* **5**, 17.
- 550 Romano, Y., Sesia, M., and Candès E. (2019). Deep knockoffs. *Journal of the American*  
551 *Statistical Association* DOI: 10.1080/01621459.2019.1660174.
- 552 Schneider, A. M., Cook, L. C., Zhan, X., Banerjee, K., Cong, Z., Imamura-Kawasawa, Y.,  
553 et al. (2020). Loss of Skin Microbial Diversity and Alteration of Bacterial Metabolic  
554 Function in Hidradenitis Suppurativa. *Journal of Investigative Dermatology*, **140**, 716–  
555 720.
- 556 Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data.  
557 *The Annals of Applied Statistics* **10**, 1019–1040.
- 558 Sohn, M. B., and Li, H. (2019). Compositional Mediation Analysis for Microbiome Studies.  
559 *The Annals of Applied Statistics* **13**, 661-681.
- 560 Tang, Z. and Chen, G. (2015). Zero-inflated generalized Dirichlet multinomial regression  
561 model for microbiome compositional data analysis. *Genome Biology* **16**, 67. 698–713
- 562 Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal*  
563 *Statistical Society. Series B* **58**, 267–288.
- 564 Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I.  
565 (2007). The human microbiome project. *Nature* **449**, 804.
- 566 Van de Geer, S., Bhlmann, P., Ritov, Y. A., and Dezeure, R. (2014). On asymptotically op-  
567 timal confidence regions and tests for high-dimensional models. *The Annals of Statistics*

- 568 **42**(3), 1166–1202.
- 569 van de Geer, S. (2019). On the asymptotic variance of the debiased Lasso *Electronic Journal*  
570 *of Statistics* **13**(2), 2970–3008.
- 571 Wang, H. (2009). Forward regression for ultra-high dimensional variable screening *Journal*  
572 *of the American Statistical Association* **104**(488),1512-1524.
- 573 Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the micro-  
574 biome. *Nature Reviews Microbiology* **14**, 508.
- 575 Wang, F., Zhou, L., Tang, L.,and Song, P. (2019). Method of contraction-expansion (MOCE)  
576 for simultaneous unference in linear models <https://arxiv.org/abs/1908.01253>
- 577 Wasserman, L., and Roeder, K. (2009). High-dimensional variable selection *The Annals of*  
578 *Statistics* **37**(5A), 2178–2201.
- 579 Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normal-  
580 ization and microbial differential abundance strategies depend upon data characteristics.  
581 *Microbiome* **5**, 27.
- 582 Xiao, J., Cao, H., and Chen, J. (2017). False discovery rate control incorporating phylogenetic  
583 tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* **33**,  
584 2873–2881.
- 585 Xu, C., and Chen, J. (2014). The sparse MLE for ultrahigh-dimensional feature screening.  
586 *Journal of the American Statistical Association* **109**, 1257-1269.
- 587 Xue, L., and Zou, H. (2011). Sure independence screening and compressed random sensing.  
588 *Biometrika*, **98**, 371-380.
- 589 Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M. C., and Chen, J. (2017a). A small-  
590 sample multivariate kernel machine test for microbiome association studies. *Genetic*  
591 *Epidemiology* **41**, 210–220.
- 592 Zhan, X., Plantinga, A., Zhao, N., and Wu, M. C. (2017b). A fast small-sample kernel

- 593 independence test for microbiome community-level association analysis. *Biometrics* **73**,  
594 1453–1463.
- 595 Zhan, X., Xue, L., Zheng, H., Plantinga, A., Wu, M. C., Schaid, D. J., Zhao, N., and Chen,  
596 J. (2017a). A smallsample kernel association test for correlated data with application to  
597 microbiome association studies. *Genetic Epidemiology* **42**, 772–782.
- 598 Zhang, C. H., and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters  
599 in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*  
600 (*Statistical Methodology*) **76**, 217–242.
- 601 Zhang, H., Chen, J., Li, Z., and Liu, L. (2019). Testing for Mediation Effect with Application  
602 to Human Microbiome Data. *Statistics in Biosciences*, 1–16.
- 603 Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., et al. (2015).  
604 Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based  
605 kernel association test. *American Journal of Human Genetics* **96**, 797–807.

## 606 7. SUPPORTING INFORMATION

607 Supporting Information referenced in Section 3 and Section 4 are available with this article at  
608 the *Biometrics* website on Wiley Online Library. It includes discussion on the normalization  
609 procedure after CSP, assumptions of Theorem 1 in the context of microbiome data, proofs  
610 of lemmas and theorems, additional numerical evaluations, and R code to implement the  
611 proposed methods.

**Table 1**

*Average screening proportions of true signals based on 200 replicates under the Dirichlet-multinomial (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	$ S^*  = 15$	$ S^*  = 20$	$ S^*  = 25$	$ S^*  = 30$
DM	$\widehat{\text{FDR}}$	CKF	0.132	0.107	0.102	0.102
		CKF+	0.073	0.064	0.070	0.075
		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	0.954	0.961	0.968	0.968
		CKF+	0.881	0.907	0.946	0.935
		KF	0.999	0.998	0.953	0.931
		KF+	0.990	0.974	0.881	0.851
LN	$\widehat{\text{FDR}}$	BH	0.626	0.547	0.445	0.385
		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
	$\widehat{\text{Power}}$	BH	0.094	0.108	0.097	0.087
		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	<b><i>Escherichia</i></b>	0.0312	C2	<b><i>Escherichia</i></b>	0.0376
	<b><i>Sutterella</i></b>	-0.0362		<b><i>Sutterella</i></b>	-0.0285
	<b><i>Akkermansia</i></b>	-0.0108		<i>Turicibacter</i>	-0.0212
	<b><i>Bifidobacterium</i></b>	-0.0189		<i>Lachnospira</i>	0.0332
	<i>Clostridium</i>	-0.0199		<i>Veillonella</i>	0.0293
	<i>Prevotella</i>	-0.0140		<i>Brevundimonas</i>	0.0424
	<i>C. Clostridium</i>	-0.0257		<i>Anaerococcus</i>	-0.0246
	<i>L. Clostridium</i>	-0.0257		<i>Bulleidia</i>	-0.0336
	<i>R. Clostridium</i>	0.0234		<i>Rhodoplanes</i>	0.0434
	<i>Epulopiscium</i>	0.0062		<i>Staphylococcus</i>	0.0198
	<i>Dorea</i>	-0.0118	CFB	<b><i>Escherichia</i></b>	0.0437
	<i>Lachnospira</i>	-0.0118		<b><i>Sutterella</i></b>	-0.0450
	<i>Veillonella</i>	0.0203		<b><i>Bifidobacterium</i></b>	-0.0144
	<i>Actinomyces</i>	-0.0264		<i>Epulopiscium</i>	0.0202
	<i>Collinsella</i>	-0.0073		<i>Lachnospira</i>	0.0195
	<i>Staphylococcus</i>	0.0449		<i>Collinsella</i>	-0.0167
	<i>Brevundimonas</i>	0.0731		<i>Eggerthella</i>	0.0809
	<i>Fingoldia</i>	-0.0336		<i>Enterococcus</i>	-0.0132
	<i>R. Eubacterium</i>	0.0506			
	<i>E. Eubacterium</i>	-0.1001			
<i>Enterococcus</i>	-0.0061				
<i>Peptostreptococcus</i>	0.0190				