Bioinformatics, 2017, 1–9
doi: 10.1093/bioinformatics/btx311
Advance Access Publication Date: 13 May 2017
Original Paper

Genetics and population analysis

False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing

Jian Xiao[1,†], Hongyuan Cao[2,†] and Jun Chen[1,*]

[1]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, U.S.A. and [2]Department of Statistics, University of Missouri, Columbia, MO, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Oliver Stegle

Received on October 31, 2016; revised on April 13, 2017; editorial decision on May 7, 2017; accepted on May 12, 2017

Abstract

Motivation: Next generation sequencing technologies have enabled the study of the human microbiome through direct sequencing of microbial DNA, resulting in an enormous amount of microbiome sequencing data. One unique characteristic of microbiome data is the phylogenetic tree that relates all the bacterial species. Closely related bacterial species have a tendency to exhibit a similar relationship with the environment or disease. Thus, incorporating the phylogenetic tree information can potentially improve the detection power for microbiome-wide association studies, where hundreds or thousands of tests are conducted simultaneously to identify bacterial species associated with a phenotype of interest. Despite much progress in multiple testing procedures such as false discovery rate (FDR) control, methods that take into account the phylogenetic tree are largely limited.
Results: We propose a new FDR control procedure that incorporates the prior structure information and apply it to microbiome data. The proposed procedure is based on a hierarchical model, where a structure-based prior distribution is designed to utilize the phylogenetic tree. By borrowing information from neighboring bacterial species, we are able to improve the statistical power of detecting associated bacterial species while controlling the FDR at desired levels. When the phylogenetic tree is mis-specified or non-informative, our procedure achieves a similar power as traditional procedures that do not take into account the tree structure. We demonstrate the performance of our method through extensive simulations and real microbiome datasets. We identified far more alcohol-drinking associated bacterial species than traditional methods.
Availability and implementation: R package *StructFDR* is available from CRAN.
Contact: chen.jun2@mayo.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The human microbiome is now known to be intricately linked to overall human health. Specific changes in microbiome composition have been associated with many human diseases such as obesity, inflammatory bowel disease and various cancers through the direct sequencing of microbial DNA (Gilbert *et al.*, 2016). One popular sequencing strategy targets the 16S rRNA gene, which carries evolutionary information, to profile the taxonomic content of the microbiome (Kuczynski *et al.*, 2012). In this approach, the sequenced 16S tags are first clustered into small units, called operational taxonomic units (OTU), based on sequence divergence. At 97% similarity level, these OTUs are assumed to correspond to biological species, though the exact correspondence may not always hold. The representative sequences of OTUs can then be used to infer a phylogenetic tree, which provides important prior knowledge about how these OTUs are related evolutionarily. Environmental and disease conditions have a tendency to affect bacterial clades, a group of closely related biological species, at different phylogenetic depth (Martiny *et al.*, 2015). Thus, the phylogenetic tree can be potentially used to increase the efficiency and power of microbiome data analysis by borrowing the states of neighboring OTUs. Indeed, many tree-based statistical methods have been demonstrated to be superior to those non-tree-based counterparts (Chen *et al.*, 2012, 2013; Purdom, 2011; Silverman *et al.*, 2017).

A common goal in sequencing-based microbiome analysis is the identification of OTUs or taxa associated with a phenotype of interest (e.g. body mass index, smoking status, disease subtypes) (Gilbert *et al.*, 2016). Consequently, hundreds or thousands of tests are conducted simultaneously. False discovery rate (FDR) control (Benjamini and Hochberg, 1995) is one of the most commonly used approaches for multiple comparison adjustment in microbiome-wide association studies. By definition, FDR is the expected proportion of falsely rejected null hypotheses among all of the rejected null hypotheses. There is a vast literature on FDR control and its applications in various fields, such as microarray analysis, astronomical surveys and brain imaging, among others (Efron *et al.*, 2001; Tusher *et al.*, 2001; Dudoit *et al.*, 2002; Schwartzman *et al.*, 2008).

The predominant framework in FDR control is via individual analysis—testing each hypothesis separately and declaring statistical significance if the *P*-value is less than a certain threshold (Benjamini and Hochberg, 1995) or the test statistic falls into the rejection region (Cao and Kosorok, 2011). Although this approach works well for phenotypes that depend on strong effects from a few variants (genotypes/OTUs), it is less suitable for complex phenotypes that are influenced by weak effects from many different variants. In fact, it has now been recognized that a majority of biological phenotypes manifest from a complex interaction among different variants. Thus, FDR control without considering the underlying biological structure can lead to very low statistical power. For microbiome data, the biological structure is encoded in the phylogenetic tree, and closely related OTUs are expected to respond to the environmental perturbations in similar manners. When an OTU is associated with a disease, it is likely that the neighboring OTUs are also disease associated.

Multiple testing under dependence is a challenging problem. Although it has been shown that the Benjamini and Hochberg procedure remains valid under various dependence structure, such as positive dependence (Benjamini and Yekutieli, 2001; Wu, 2008), extensive evidence shows that there will be a substantial power loss without considering the actual dependence (Owen, 2005; Efron, 2007; Conneely and Boehnke, 2007). To incorporate the dependence structure, Sun and Cai (2009) proposed a hidden Markov Model, Leek and Storey (2008) and Friguet *et al.* (2009) used a factor model approach and Fan *et al.* (2012) integrated principal component analysis and a factor model. These methods require strong assumptions and are not applicable to microbiome data. Recently, there have been multiple testing methods that leverage prior biological knowledge to increase the power of FDR control (Yekutieli, 2008; Kang *et al.*, 2009; Hu *et al.*, 2012; Ignatiadis *et al.*, 2016). For microbiome data, Sankaran and Holmes (2014) proposed to apply the group Benjamini-Hochberg (GBH) procedure of Hu *et al.* (2012) and the hierarchical false discovery rate (HFDR) controlling procedure of Yekutieli (2008) to utilize the hierarchical tree structure. In GBH, the proportion of alternative hypothesis within each group is estimated and overall *P*-values of that group are re-weighted incorporating the proportion of alternative hypothesis within the group to improve power. HFDR arranges families of related hypotheses along a tree and restricting attention to particular subtrees that are more likely to contain alternative hypotheses. GBH requires the specification of the groups while HFDR uses the tree topology. Both of them are not able to fully exploit the tree structure information, which contains both the topology and branch lengths.

In this paper, we aim to boost the statistical power of microbiome-wide multiple testing by proposing a new method that leverages the relationships among different OTUs through the phylogenetic tree. Our method combines an empirical Bayes approach to incorporate prior correlation structure (Wei and Li, 2007; Li *et al.*, 2010) and a permutation approach for FDR control (Xie *et al.*, 2005). It has several unique distinctive features. First, a working hierarchical model that incorporates prior biological knowledge allows us to borrow information across different tests. Second, an algorithm based on permutation is used to control FDR. Permutation retains the dependence among test statistics and has the flexibility to adjust to the unknown null distribution of the test statistics, even when the working model is mis-specified. Third, our procedure has substantial power gain when the phylogenetic tree accurately describes the dependence structure and achieves a similar power as traditional procedures when the phylogenetic tree is mis-specified or non-informative. Simulation studies and real data applications show that our procedure has favorable performances compared with conventional approaches.

2 Materials and methods

2.1 Background and notation

Consider a typical microbiome dataset with m OTUs from n individual samples, and further suppose that a phenotype of interest is recorded for each sample. Such data can be represented in an $m \times n$ matrix, with rows corresponding to individual OTUs and columns corresponding to individual samples. The total number of samples n is usually in the order of tens or hundreds, and the number of OTUs is in the order of hundreds or thousands. In addition, a phylogenetic tree relates all the OTUs, which can be represented as an $m \times m$ distance matrix $\mathbf{D}$, containing the pairwise patristic distances between OTUs (the sum of the lengths of the branches that link two OTUs in the tree). Note that the $m \times m$ distance matrix is independent of the observations in the $m \times n$ data matrix.

Suppose that we are interested in the association between the phenotype and OTUs. This can be casted into a multiple testing problem—the simultaneous tests of the null hypothesis H_j of no association between the OTU j and the phenotype:

$$H_{0j}: \text{ the } j\text{th OTU is not associated with the phenotype }, \quad j = 1, \ldots m$$

to see which OTUs are correlated with the phenotype.

A by far classic approach for this problem is the FDR control procedure pioneered by Benjamini and Hochberg (Benjamini and Hochberg, 1995). The algorithm works as follows. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered observed P-values of m hypotheses. Define $k = \max\{i : P_{(i) \leq i\alpha/m}\}$ and reject $H_{0(1)}, \ldots H_{0(k)}$, where α is a pre-specified error rate. If no such k exists, reject no hypothesis. This approach is shown to be valid under independence or positive dependence among the P-values. Fan *et al.* (2012) imposed a factor model on the observed data to model the dependence structure and utilized the principal component analysis of the variance covariance matrix. Instead of directly modeling the dependence, we develop a hierarchical modeling strategy to leverage prior structure information by an empirical Bayesian approach. We work with the z-value, which can be conveniently transformed from the P-value by

$$z_j = \Phi^{-1}(1 - P_j), j = 1, \ldots, m, \tag{1}$$

where Φ denotes the cumulative distribution function of the standard normal random variable. If $P_j < \epsilon$ or $P_j > 1 - \epsilon$, where ϵ is a very small number (e.g. 10^{-15}), we set $P_j = \epsilon$ or $1 - \epsilon$. To accommodate the direction of the effect, z-value can be obtained by

$$z_j = \Phi^{-1}(1 - P_j/2) \quad \text{or} \quad \Phi^{-1}(P_j/2), \tag{2}$$

depending on the sign of the effect. For microbiome-based applications, we focus on the transformation (2). P-values can be obtained through various test statistics or regression models.

2.2 Phylogeny-induced prior correlation

To facilitate the incorporation of the phylogenetic tree into FDR control, we define a tree-based correlation structure based on the distance matrix $\mathbf{D}$. Following the trait evolution model (Martin and Hansen, 1997), the correlation of the traits between OTU i and OTU j, $\mathbf{C}_{ij}$, can be modeled using an exponential function

$$\mathbf{C}_{ij} = \exp(-2\rho \mathbf{D}_{ij}), \ \ i, j = 1, \ldots, m,$$

where the parameter $\rho \in (0, \infty)$ characterizes the evolutionary rate. If $\rho = 0$, $\mathbf{C}_{ij} = 1$ for $\forall i, j$, meaning that all the traits are the same and there is maximal phylogenetic relationship. On the other extreme, if $\rho \to \infty$, $\mathbf{C}_{ij} = 0$ for $\forall i \neq j$, meaning that all the traits are not related and they can evolve independently. Statistically, ρ can also be interpreted as a parameter governing the grouping effect. If $\rho = 0$, all the OTUs are grouped together, and if $\rho \to \infty$, all the OTUs are independent. Thus by tuning the parameter ρ, we can achieve different phylogenetic resolution, which has a similar effect of grouping the OTUs at different phylogenetic depth or taxonomic ranks. From an informational perspective, ρ determines the scale of the neighborhood for information borrowing, with a larger ρ indicating a smaller neighborhood so that OTUs of more close proximity contribute the information. We denote $\mathbf{C}_\rho$ as the phylogeny-induced correlation matrix of m OTUs.

2.3 Hierarchical model

With $\mathbf{z} = (z_1, \ldots, z_m)^T$ and $\mathbf{C}_\rho$, we can define a hierarchical model. We assume that, conditional on the mean, $\mathbf{z}$ follows a multivariate normal distribution

$$\mathbf{z}|\boldsymbol{\mu} \sim \text{MVN}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \tag{3}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix, σ^2 is the unknown variance and the vector $\boldsymbol{\mu}$ is of main interest, where we want to draw posterior inference. We further assume that $\boldsymbol{\mu}$ follows a prior multivariate normal distribution

$$\boldsymbol{\mu} \sim \text{MVN}(\gamma \mathbf{1}, \tau^2 \mathbf{C}_\rho), \tag{4}$$

where 1 is a vector of 1s and γ, τ^2 and ρ are hyperparameters. The hierarchical modeling allows the incorporation of prior correlation structure and the benefit of using conjugate prior is that a closed form posterior can be obtained.

We obtain the marginal distribution of $\mathbf{z}$ by the formula as follows:

$$\mathbf{z} \sim \text{MVN}(\gamma \mathbf{1}, \tau^2 \mathbf{C}_\rho + \sigma^2 \mathbf{I}). \tag{5}$$

Therefore, the prior structure induces dependence among the z-values. We use an empirical Bayesian approach to estimate hyperparamters from available data.

To be noted, the proposed hierarchical model does not assume mixture components as more commonly used in FDR literature (Xie *et al.*, 2011) and neither does it model the data correlation $(\mathbf{z}|\mu)$. Therefore it is a **working** model per se. The purpose of this simple hierarchical model is to derive an efficient moderated test statistic as described in the next section.

2.4 Prior-structure moderated test statistic

Given the true values of the hyperparameters ρ_0, γ_0, σ_0^2 and τ_0^2, the posterior density function of $\boldsymbol{\mu}$ can be obtained from the Bayesian formula

$$\begin{aligned} f(\boldsymbol{\mu}|\mathbf{z}) &= \frac{f(\mathbf{z}|\boldsymbol{\mu}) f(\boldsymbol{\mu})}{f(\mathbf{z})} \\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}^*)^T \mathbf{K}^{-1} (\mu - \boldsymbol{\mu}^*)\}, \end{aligned}$$

where

$$\mathbf{K} = (\sigma_0^{-2} \mathbf{I} + \tau_0^{-2} \mathbf{C}_{\rho_0}^{-1})^{-1}$$

and

$$\begin{aligned} \boldsymbol{\mu}^* &= \mathbf{K}(\sigma_0^{-2} \mathbf{z} + \tau_0^{-2} \mathbf{C}_{\rho_0}^{-1} \gamma_0 1) \\ &= (\mathbf{I} + k \mathbf{C}_{\rho_0}^{-1})^{-1} (k \mathbf{C}_{\rho_0}^{-1} \gamma_0 1 + \mathbf{z}), \quad k = \sigma_0^2 / \tau_0^2. \end{aligned} \tag{6}$$

Statistical inference is based on $\boldsymbol{\mu}^*$ (the posterior mean of $\boldsymbol{\mu}|\mathbf{z}$).

Based on the posterior distribution, μ^* is the estimator that minimizes the mean squared error. We use $\boldsymbol{\mu}^*$ as the prior-structure moderated test statistic for permutation-based significance assessment.

There are some interesting observations from the moderated statistic. Fix ρ_0, if $k \to 0$, the shrinkage estimator $\boldsymbol{\mu}^* \to \mathbf{z}$. In this case, either the prior $\boldsymbol{\mu}$ is not informative with a large variance τ^2 or the variability of the conditional distribution of $\mathbf{z}|\boldsymbol{\mu}$ is extremely small. Consequently, z-values are our best estimate. On the other hand, if $k \to \infty$, the shrinkage estimator $\boldsymbol{\mu}^* \to \gamma_0 1$. In this case, prior structure shrinks the posterior estimator to a common stable value γ_0. It indicates that the prior structure provides overwhelming information. Thus, the parameter k balances information from the data and the prior correlation structure. The parameter ρ_0 has a similar effect. Through the combination of ρ_0 and k, various degrees of moderation can be achieved. Thus, the resulting $\boldsymbol{\mu}^*$ provides prior-structure adjusted ranking of the importance of different OTUs.

We employ an empirical Bayesian idea to obtain the estimates of hyperparameters from available data. In fact, the parameter estimation under model (5) has been well studied in the statistics literature (Draper and Smith, 1998). In our implementation, we used MLE to estimate the hyperparameters as implemented in the R function *gls*. The values of the estimates indicate the informativeness of the phylogenetic tree, with large $\widehat{\rho}$ or small $\widehat{k}$ suggests limited use of the tree. After obtaining the estimates $\widehat{\gamma}$, $\widehat{\rho}$, $\widehat{k}$, we plug them into $\boldsymbol{\mu}^*$ to obtain $\widehat{\boldsymbol{\mu}}^* = (\mathbf{I} + \widehat{k} \mathbf{C}_{\widehat{\rho}}^{-1})^{-1} (\widehat{k} \mathbf{C}_{\widehat{\rho}}^{-1} \widehat{\gamma} 1 + \mathbf{z})$.

2.5 Permutation-based FDR control algorithm

In this section, we use permutation to develop an FDR control procedure based on $\widehat{\boldsymbol{\mu}}^*$. The advantage of using permutation is that it adjusts to the unknown null distribution of the test statistic, and is robust against different types of dependence, such as the correlations in the data $(\mathbf{z}|\mu)$, even when the working models (3) and (4) are misspecified. As mentioned earlier, the posterior mean estimate $\widehat{\boldsymbol{\mu}}^*$ provides a new ranking of the importance of different OTUs. Our goal is to find a threshold such that the FDR is controlled at a desired level. When the hierarchical model is used, the distribution of the statistics for OTUs under the null will be affected by OTUs under the alternative due to information pooling. To derive the correct null distribution through permutation, the signals need to be preserved in the permutation. We thus confine our permutation to the OTUs with P-values larger than the median. When calculating the moderated statistics under the permutation, we use P-values from all OTUs including those with P-values less than the median. Specifically, we use the following algorithm

1. Sort the $|\widehat{\boldsymbol{\mu}}^*|$ in an increasing order and denote $d_1 = |\widehat{\boldsymbol{\mu}}^*_{(1)}|, d_2 = |\widehat{\boldsymbol{\mu}}^*_{(2)}|, \ldots, d_m = |\widehat{\mu}^*_{(m)}|$, where $|\cdot|$ means absolute value function.
2. Create two sets containing the indices of the OTUs with the original P-values $\le$ and $>$ the median P-values, respectively. Denote the two sets as $\mathcal{A}^1$ and $\mathcal{A}^0$.
3. Confine the permutation to OTUs in $\mathcal{A}^0$. Permute B times (e.g. $B = 20$) to recalculate P-values and z-values for these OTUs. For OTUs in $\mathcal{A}^1$, P-values and z-values remain as original. For the bth permuted dataset, compute $\widehat{\mu}^*_b = \{\widehat{\mu}^*_{b,1}, \ldots, \widehat{\mu}^*_{b,m}\}$ based on (6) for all OTUs from $\mathcal{A}^1$ and $\mathcal{A}^0$. For a given $t \in \{d_1, \ldots, d_m\}$, we can calculate the number of positives (P) as

$$\mathrm{P}(t) = \#\{i : |\widehat{\boldsymbol{\mu}}^*_i| > t\} \tag{7}$$

and estimate the number of false positive (FP) based on OTUs from $\mathcal{A}^0$ as

$$\widehat{\mathrm{FP}}(t) = \frac{2}{B}\sum_{b=1}^{B} \#\{i \in \mathcal{A}^0 : |\widehat{\boldsymbol{\mu}}^*_{b,i}| > t\}. \tag{8}$$

Consequently, FDR is estimated as

$$\widehat{\mathrm{FDR}}(t) = \frac{\widehat{\mathrm{FP}}(t)}{\mathrm{P}(t) \vee 1}. \tag{9}$$

4. For a pre-specified error rate α, the threshold k is defined as

$$k = \max_i\{\widehat{\mathrm{FDR}}(d_i) \le \alpha \quad \text{and} \quad \widehat{\mathrm{FDR}}(d_{i+1}) > \alpha\}. \tag{10}$$

Reject all $H_{0,(j)} forj \le k$, where $H_{0,(j)}, j = 1, \ldots, k$ denote the hypotheses corresponding to the order statistics $\mu^*_{(1)}, \ldots, \boldsymbol{\mu}^*_{(k)}$.

Let π_0 be the proportion of true null hypotheses among m hypotheses tests. The proposed algorithm is conservative due to the overestimation of FP. A more accurate estimate of FP would be

$$\frac{2\pi_0}{B}\sum_{b=1}^{B} \#\{i \in \mathcal{A}^0 : |\widehat{\boldsymbol{\mu}}^*_{b,i}| > t\}.$$

Since π_0 is unknown and is usually close to 1, we set $\pi_0 = 1$ in FP estimation. Various methods were developed to estimate π_0 to improve power, which is certainly warranted but beyond the scope of current paper (Storey and Tibshirani, 2003; Xie *et al.*, 2005). Our algorithm essentially is a ranking and thresholding approach. However, the ranking of $\widehat{\boldsymbol{\mu}}^*$ leverages prior structure information contained in the phylogenetic tree and the proposed method has substantial power gain as demonstrated in our simulation studies. As we use permutation, our approach is robust to the mis-specification of the working models (3) and (4). We did not explicitly state the permutation scheme since it is problem-dependent. For a simple two-sample comparison, we can shuffle the group labels. For a more complicated design with covariates, we may use a residual permutation approach. Also note that the permutation approach is heuristic and the FDR control is validated by simulations.

2.6 Computation and implementation

We implement our method in the R package 'StructFDR'. The main function 'TreeFDR' requires the OTU counts, the phylogenetic tree, the testing function, which produces the association P-values and the signs of the effect, and a permutation function, which permutes the data in a user-defined way. We also provide a more generic function 'structFDR', which accepts a distance matrix among features instead of a tree, to perform structure-based FDR control for other genomic data as long as a distance metric is defined between features. The computation time depends on the efficiency of the testing function and the number of permutations. Estimation of the hyperparameters is much faster compared to the actual statistical tests. If Wilcoxon rank-sum tests are used, computation usually takes several minutes on a desktop for a typical dataset of ≈100 samples and ≈1,000 OTUs (B = 100). The R package includes a vignette demonstrating the use of the method.

3 Results

3.1 Simulation studies

We conduct comprehensive simulation studies to demonstrate the superior performance of our method over traditional methods that do not incorporate the phylogenetic tree structure. We consider the case-control design, examine five different scenarios, three of which have informative phylogeny and two of which have non-informative phylogeny. For all the scenarios, we simulate $m = 400$ OTUs and each sample group has 50 observations. Without loss of generality, the first 50 samples are from the control. All results are obtained based on 200 replications.

Specifically, we use R package *rcoal* to generate a random coalescent tree in each replication to achieve various clustering patterns of the 400 OTUs. R package *pam* is used to partition the generated tree into clusters. We next generate the OTU counts. First, we simulate the sequencing depth (total counts) of the samples by a negative binomial distribution with mean 10 000 and size 25. Second, a Dirichlet distribution, where the parameter values were estimated from a real throat microbiome dataset (Chen *et al.*, 2012) (included in the 'StructFDR' package), is used to generate 100×400 composition matrix. Third, a multinomial distribution is used to generate count data based on the sequencing depth and 100×400 composition matrix. For the case samples, we multiply the counts with a fold change vector $\{\exp(\beta_1), \ldots, \exp(\beta_{400})\}$.

We consider scenarios with informative phylogeny and non-informative phylogeny. With informative phylogeny, we study three scenarios, each with associated clusters of balanced sizes accounting for approximately 10–20% of OTUs. Let $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_R$ contain the indices of the OTUs from the associated clusters (C1 and C2 in Fig. 1) and the rest clusters. Scenario one (S1) considers the perfect case where signals within each of the two clusters have the same sign and strength ($\beta_{i\in\mathcal{A}_1} = 4$, $\beta_{i\in\mathcal{A}_2} = -4$ and $\beta_{i\in\mathcal{A}_R} = 0$). Scenario two (S2) considers the case where signals in each cluster have the same sign but varied strength ($\beta_{i\in\mathcal{A}_1} \overset{iid}{\sim} N(4, 2)$, $\beta_{i\in\mathcal{A}_2} \overset{iid}{\sim} N(-4, 2)$, and $\beta_{i\in\mathcal{A}_R} = 0$). Information still can be borrowed from neighboring OTUs. Scenario three (S3) is similar to S1 but 10 small clusters (out of 100) are associated with the outcome. We next study two scenarios, where the phylogeny is non-informative. Scenario four (S4) corresponds to the case where signals can be positive or negative within each cluster and thus canceled out ($\beta_{i\in\mathcal{A}_1} \overset{iid}{\sim} N(0, 2)$, $\beta_{i\in\mathcal{A}_2} \overset{iid}{\sim} N(0, 2)$ and $\beta_{i\in\mathcal{A}_R} = 0$). This scenario is contradictory to our assumption that closely related OTUs have similar effects. Scenario five (S5) corresponds to the case where 40

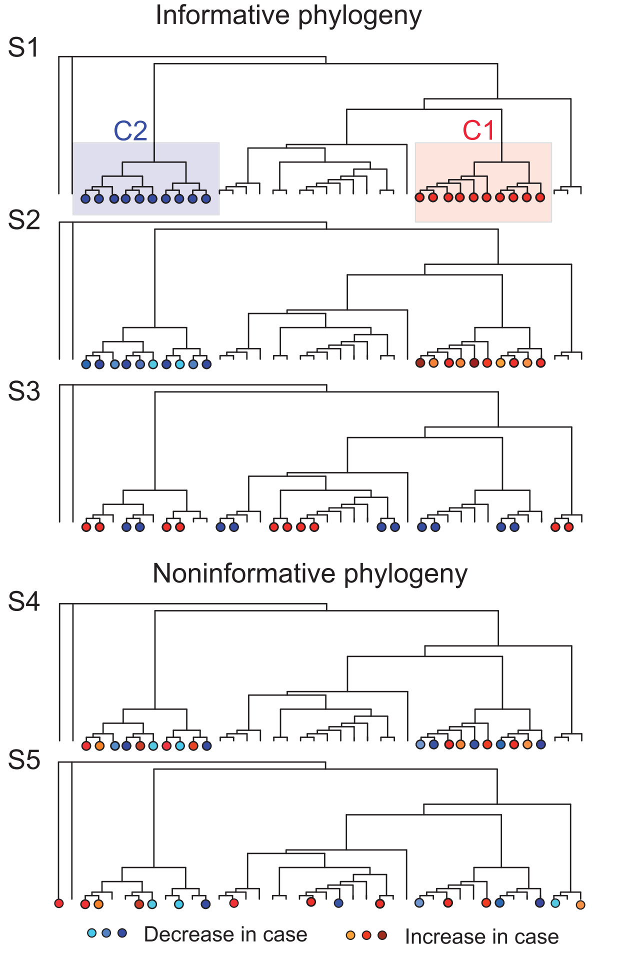

Fig. 1. Simulation configuration, where red circles represent positive values and blue circles represent negative values

associated OTUs are randomly distributed on the tree with different signs and varied strength ($\beta_i \overset{iid}{\sim} N(0, 4)$). Clearly, in this scenario, the tree does not provide any useful information as how the associated signals are correlated. S4 and S5 are used to study the robustness of the proposed method. These five scenarios are illustrated in Figure 1.

The Wilcoxon rank sum test is used to conduct differential abundance tests and obtain the *P*-values. The *P*-values are then transformed to get the *z*-values through

$$z_j = \begin{cases} \Phi^{-1}(1 - p_j/2) & \text{If } \bar{X}_{\text{case},j} \le \bar{X}_{\text{control},j} \\ \Phi^{-1}(p_j/2) & \text{If } \bar{X}_{\text{case},j} > \bar{X}_{\text{control},j} \end{cases}$$

where $\bar{X}_{\text{control},j}$ and $\bar{X}_{\text{case},j}$ are mean abundances for control and case samples, respectively.

We compare the performance of our method with the procedures proposed in Benjamini and Hochberg (1995) (BH), Storey and Tibshirani (2003) (ST), Hu *et al.* (2012) (GBH) and Yekutieli (2008)(HFDR). The ST procedure fixes a threshold value *t*, estimates the FDR and chooses *t* so that the estimated FDR is no larger than α. In ST procedure, the proportion of null hypothesis is estimated through a conservative estimator $\widehat{\pi}_0(\lambda) = \#\{p_i > \lambda; i = 1, \ldots, m\}/m(1 - \lambda)$, where λ is a tuning parameter. R package *q-value* is used to obtain results based on the ST procedure. Since GBH depends on the group specification, we create different numbers of groups/clusters (10, 20, 40 and 100, denoted as GBH-10, GBH-20, GBH-40 and GBH-100) using *pam* based on the matrix of pairwise distances between OTUs. For HFDR, we choose an alpha level that controls the upper bound of the FDR at the tree tips (Yekutieli, 2008; Sankaran and Holmes, 2014). GBH and HFDR are performed using the R package *structSSI* (Sankaran and Holmes, 2014). Actual FDR and power are used to measure the performance of different procedures at the nominal FDR level $0.01, 0.05$ and 0.1. The power is defined as the expectation of the ratio of correctly rejected hypotheses and total alternative hypotheses. Specifically, actual FDR and power are estimated based on the average over 200 replications.

The results are summarized in Tables 1 and 2. Table 1, we can see that our procedure is conservative and FDR is controlled under the pre-specified level for most scenarios. BH procedure is conservative and ST procedure has actual FDRs closer to the nominal ones than BH procedure as the proportion of null hypothesis has been incorporated in ST procedure. From Table 2, we can see that our procedure has substantial power gain when phylogeny is informative at all pre-specified FDR levels. When phylogeny is non-informative (S4 and S5), our procedure has similar power as BH and ST procedure, indicating the robustness of our method. For GBH, the performance depends on the number of pre-specified groups and GBH-100 does not control the FDR properly (Supplementary Table S1). When the signals form large clusters (S1–S2), GBH is more powerful than BH and ST and slightly less powerful than or similar to our procedure. When the signals form small clusters (S3), GBH becomes much less powerful than our procedure. S4 favors GBH procedure since the signals have a group structure. As the signals are more scattered (S5), GBH breaks down: it fails to control the FDR and is less powerful than BH, ST and TreeFDR. In contrast, HFDR has very conservative FDR control when the signals are clustered (S1–S4), but is slightly anti-conservative in S5, when the signals are randomly distributed. Regarding the statistical power, it is generally less powerful than GBH and TreeFDR, especially for S5.

We perform additional simulations to study the robustness of our method to clade-inconsistent associations. We more densely sample the continuum of clade-consistent versus -inconsistent associations by simulating randomly associated OTUs (clade-inconsistent associations) on the basis of S1. Details of the simulations are included in the Supplementary file and the results are summarized in Supplementary Tables S3 and S4. The basic observation is that our method is robust to clade-inconsistent signals to a large degree and is generally more powerful than the other procedures compared while controlling the FDR.

It is important to point out that the higher power of our procedure is not at the price of a higher FDR level; this is illustrated by ROC curves in Figure 2. The true positive rate is calculated as the average proportions of correctly identified OTUs and the false positive rate is calculated as the average proportions of falsely identified OTUs. We vary the significant threshold and calculate corresponding true positive rates and false positive rates. We can see that our procedure significantly outperforms BH and ST procedure for scenario 1–3, and has comparable performance with BH and ST procedure for scenario 4 and 5. To summarize, when the phylogenetic tree is informative, our procedure dominates BH and ST procedure, and when the tree is not informative, our procedure is comparable to BH and ST procedure.

3.2 Real data application

We demonstrate our method using a real microbiome dataset from a study of long-term dietary effects on the human gut microbiome

Table 1. The actual FDR at the nominal FDR level 0.01, 0.05, and 0.10

Scenario	Methods	Level 0.01	Level 0.05	Level 0.10
1	TreeFDR	0.002 (0.008)	0.016 (0.045)	0.034 (0.065)
	BH	0.007 (0.037)	0.030 (0.053)	0.068 (0.095)
	ST	0.007 (0.037)	0.035 (0.058)	0.077 (0.100)
	GBH-20	0.003 (0.017)	0.009 (0.023)	0.029 (0.084)
	HFDR	0 (0)	0.000 (0.004)	0.001 (0.007)
2	TreeFDR	0.007 (0.071)	0.026 (0.114)	0.054 (0.143)
	BH	0.005 (0.015)	0.035 (0.105)	0.074 (0.132)
	ST	0.010 (0.072)	0.040 (0.106)	0.084 (0.147)
	GBH-20	0.011 (0.100)	0.028 (0.129)	0.046 (0.136)
	HFDR	0 (0)	0 (0)	0.000 (0.006)
3	TreeFDR	0.006 (0.017)	0.030 (0.042)	0.065 (0.061)
	BH	0.005 (0.017)	0.035 (0.045)	0.079 (0.062)
	ST	0.005 (0.017)	0.041 (0.049)	0.087 (0.066)
	GBH-20	0.017 (0.030)	0.069 (0.074)	0.133 (0.099)
	HFDR	0.000 (0.005)	0.004 (0.018)	0.008 (0.023)
4	TreeFDR	0.004 (0.019)	0.039 (0.092)	0.071(0.141)
	BH	0.004 (0.027)	0.031 (0.089)	0.070 (0.140)
	ST	0.005 (0.021)	0.034 (0.090)	0.075 (0.141)
	GBH-20	0.008 (0.072)	0.044 (0.166)	0.087 (0.222)
	HFDR	0 (0)	0.001 (0.012)	0.001 (0.010)
5	TreeFDR	0.011 (0.031)	0.043 (0.056)	0.095 (0.076)
	BH	0.006 (0.022)	0.036 (0.050)	0.085 (0.075)
	ST	0.007 (0.023)	0.041 (0.055)	0.091 (0.074)
	GBH-20	0.018 (0.045)	0.082 (0.101)	0.183 (0.139)
	HFDR	0.018 (0.059)	0.074 (0.095)	0.118 (0.111)

Note: GBH-20 is the GBH procedure based on 20 pre-specified groups. Results are averaged over 200 replications and standard deviation (sd) are given in the parentheses.

Table 2. The power at the nominal FDR level 0.01, 0.05, and 0.10

Scenario	Methods	Level 0.01	Level 0.05	Level 0.10
1	TreeFDR	0.454 (0.343)	0.533 (0.362)	0.586 (0.356)
	BH	0.353 (0.316)	0.388 (0.323)	0.411 (0.326)
	ST	0.356 (0.316)	0.392 (0.325)	0.415 (0.328)
	GBH-20	0.434 (0.339)	0.510 (0.341)	0.544 (0.341)
	HFDR	0.381 (0.310)	0.452 (0.308)	0.506 (0.298)
2	TreeFDR	0.372 (0.325)	0.471 (0.351)	0.533 (0.358)
	BH	0.318 (0.296)	0.358 (0.305)	0.380 (0.310)
	ST	0.320 (0.297)	0.360 (0.306)	0.384 (0.311)
	GBH-20	0.403 (0.324)	0.478 (0.339)	0.509 (0.341)
	HFDR	0.353 (0.287)	0.421 (0.292)	0.468 (0.292)
3	TreeFDR	0.485 (0.177)	0.536 (0.179)	0.574 (0.177)
	BH	0.391 (0.177)	0.430 (0.180)	0.454 (0.179)
	ST	0.394 (0.178)	0.434 (0.181)	0.459 (0.178)
	GBH-20	0.381 (0.184)	0.423 (0.187)	0.449 (0.187)
	HFDR	0.342 (0.193)	0.399 (0.199)	0.436 (0.198)
4	TreeFDR	0.193 (0.222)	0.226 (0.245)	0.243 (0.255)
	BH	0.198 (0.240)	0.229 (0.257)	0.245 (0.264)
	ST	0.200 (0.241)	0.231 (0.258)	0.248 (0.265)
	GBH-20	0.263 (0.277)	0.321 (0.300)	0.361(0.315)
	HFDR	0.147 (0.185)	0.179 (0.203)	0.196 (0.208)
5	TreeFDR	0.286 (0.077)	0.322 (0.078)	0.345 (0.081)
	BH	0.285 (0.077)	0.319 (0.077)	0.342 (0.081)
	ST	0.286 (0.078)	0.322 (0.078)	0.345 (0.082)
	GBH-20	0.182 (0.083)	0.203 (0.086)	0.217 (0.089)
	HFDR	0.156 (0.067)	0.177 (0.068)	0.188 (0.070)

Note: GBH-20 is the GBH procedure based on 20 pre-specified groups. Results are averaged over 200 replications and standard deviation (sd) are given in the parentheses.

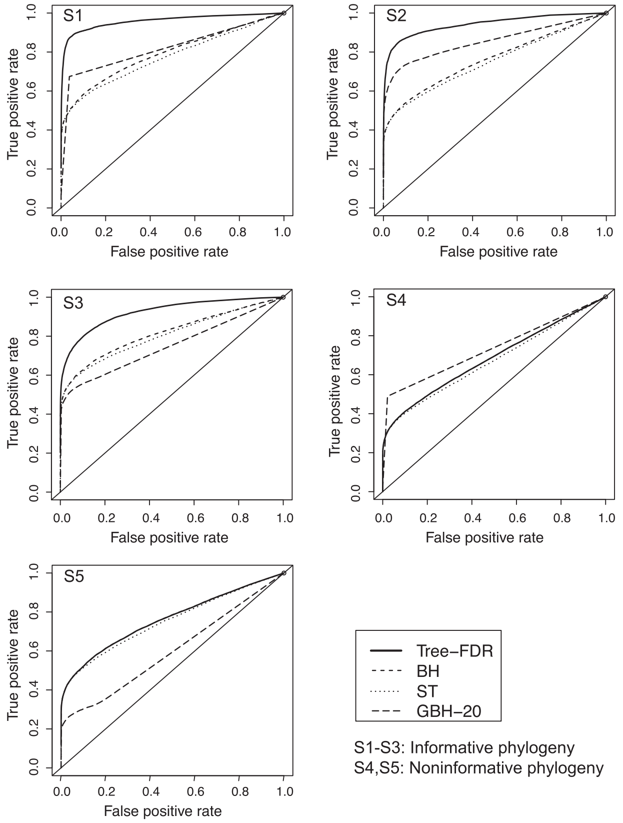

Fig. 2. ROC curves for different scenarios. ROC curves are generated based on the adjusted *P*-values. HFDR does not output adjusted *P*-values, thus is not included in the ROC analysis

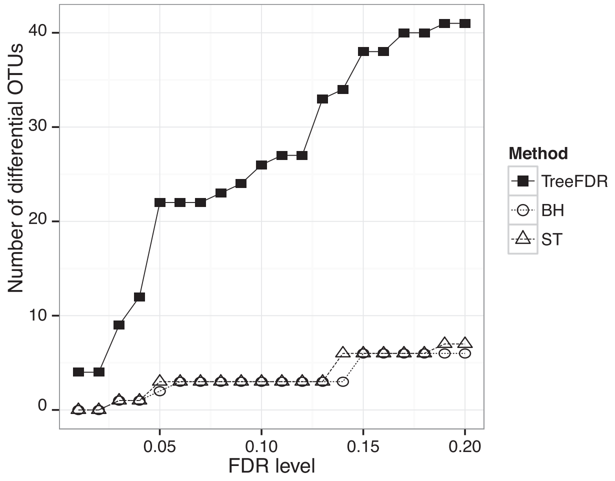

Fig. 3. Number of identified OTUs as a function of FDR level

(Wu *et al.*, 2011). The main scientific question of this study is to investigate the association of dietary and environmental variables with the gut microbiota. The specific variable we investigate is the alcohol intake. It has previously been shown that alcohol has an important influence on the human microbiome (Engen *et al.*, 2015; Chen *et al.*, 2016). In current analysis, we would like to identify OTUs associated with alcohol consumption, which is obtained from food frequency questionnaire and energy-adjusted (Willet *et al.*, 1997).

The dataset includes 98 samples and non-singleton 6, 674 OTUs. We exclude OTUs with occurrence frequency <0.1 and base our analysis on the remaining 949 relatively common OTUs. The OTU counts are normalized against library sizes before tests. To reduce the influence of measurement errors, continuous measure of alcohol consumption is converted to high intake and low intake, based on the median. The Wilcoxon rank sum test, which is a non-parametric test of the null hypothesis that the OTU abundance from high intake alcohol and low intake alcohol groups are the same against the alternative hypothesis that they are different, is used to obtain *P*-values. The phylogenetic tree is constructed based on the FastTree algorithm (Price *et al.*, 2010).

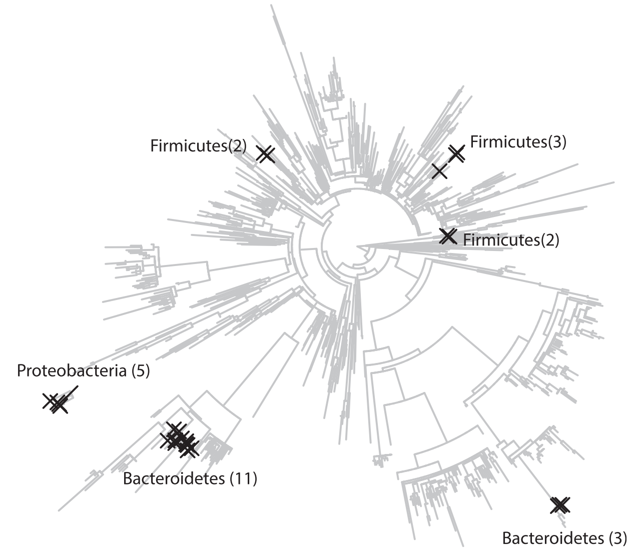

Fig. 4. OTUs (×) identified by TreeFDR procedure at an FDR of 10%. Numbers of OTUs in the clusters are indicated in parentheses

The results are summarized in Figures 3 and 4. Figure 3 plots the number of identified OTUs with different FDR levels based on our procedure, BH procedure and ST procedure. From the plot, we can see that our procedure identifies far more OTUs than the other two procedures at various FDR levels, which is consistent with the simulation studies. As the FDR level increases, the difference becomes even more striking due to the ability of our procedure to pick up weak clustered signals. Our method identified 22 and 26 alcohol-associated OTUs at an FDR of 5% and 10%, respectively. In contrast, BH and ST procedure only identified 2 and 3 OTUs at these levels. Many OTUs identified by our procedure form clusters (Fig. 4), indicating that the phylogenetic tree is informative in this case. The identified OTUs belong to six clusters from the Phylum of Bacteroidetes, Firmicutes and Proteobacteria. This implies that our procedure is effective in allowing information borrowing among neighboring OTUs, and the status of neighboring taxa is a very informative prior. By considering the *P*-values of neighboring OTUs, these OTUs with a marginally significant *P*-values will stand out and show statistical significance due to signal pooling. Interestingly, our procedure does not recover all the three OTUs identified by BH and ST at an FDR of 10%. This is due to the fact that our approach is powered to detect the clustered signals at some expense of randomly scattered signals when the phylogeny is informative. The associated taxonomies of the identified OTUs are also biologically interesting. For example, several OTUs are from the family *Firmicutes*; *Lachnospiraceae* and they have significantly increased abundance in high-alcohol takers. The bacteria from *Lachnospiraceae* family usually have alcohol dehydrogenase and can metabolize alcohol. Leclercq *et al.* (2014) reported a significant increase of *Lachnospiraceae* in alcohol-dependent groups.

Finally, we apply our method to another real dataset from a study of the microbiota after ileal pouch-anal anastomosis (IPAA) surgery for ulcerative colitis (UC) (Morgan *et al.*, 2015). Our procedure identifies significantly more UC-associated OTUs than BH and ST (See Supplementary file for details).

4 Discussion

We present a highly flexible, robust and computationally efficient method to the general problem of identifying differential features in the 'large p, small n' datasets with prior structure information that are becoming ubiquitous in biological experimentation (and elsewhere). Given that the parameter of the prior distribution can be inferred from the data, posterior likelihoods for diverse patterns of differential features can be inferred through an empirical Bayesian analysis. We describe here methods to infer the parameters of the prior distribution through maximum likelihood methods, but this is not essential; any method for inferring these parameters from the data might be applicable. Without incorporating the prior structural information, weak signals that jointly affect the phenotype cannot be picked up by prevailing methods of individual analysis. Cao and Wu (2015) also incorporates clustering effects, but from a different perspective.

The power of our approach depends on the assumption of 'clustered signals'. For microbiome data, association signals are often observed to cluster on the phylogenetic tree. This is due to the fact that environmental factors or disease conditions usually affect some microbial traits, and the traits of interest can be shared within bacterial clades and therefore largely consistent with the phylogenetic tree (Goberna and Verdu, 2016; Martiny *et al.*, 2015). With the development of high-resolution profiling method such as Metaphlan2 (Scholz *et al.*, 2016), we are able to obtain taxonomic resolution beyond species level. With such high-resolution data, we expect to see more clade-consistent associations.

Though the simulations demonstrated the robustness and power of our method across a wide range of scenarios, there are still scenarios where our method is less powerful than the traditional BH method. The power loss is most obvious when there are a large number of non-differential OTUs in a clade and the signal is strong. In such situation, there is sufficient phylogeny signal for 'information-borrowing', which leads to significantly inflated test statistic for these non-differential OTUs within that clade. To control the FDR at the desired level, power will be decreased accordingly. Additional simulations (Supplementary Figs S1 and S2, Additional Simulation 2) reveals that our method is robust to non-differential OTUs up to 20–40% across different signal levels. Beyond that limit, our method becomes less powerful than the BH procedure. To compensate the power loss, BH procedure could be used adaptively in such scenarios. One heuristic rule selects the BH procedure when our method identifies significantly less OTUs based on a Fisher's exact test. This approach controlled the FDR in the simulations and rescued the power to some extent (Additional Simulation 2). To optimally combine our method and other FDR methods to further increase the robustness of our approach warrants future investigation.

In simulation studies, when both positive and negative values appear in a cluster, their effects cancel each other out and we do not have any power gain. This makes sense for microbiome applications since we expect a similar effect for closely related species. However, for other applications such as gene expression data, both negative and positive correlations could be observed within the same network. In such case, the significance level is most informative and we are not concerned about the direction of effects. To accommodate such situations, we can use (1) to obtain *P*-values and use $\widehat{\boldsymbol{\mu}}^*$ instead of $|\widehat{\boldsymbol{\mu}}^*|$ as a test statistic. In addition, in (3), we may allow the variance to be heteroscedastic, that follows certain distributions and estimate the hyperparameters using an empirical Bayesian method. This is beyond the scope of current paper and future research is needed.

Our working model is designed to use the prior-induced correlation and the within-data correlation is not modeled in our framework. The data correlation has been shown to affect the performance of many FDR procedures including the popular BH procedure (Benjamini and Yekutieli, 2001). Here, we use permutation-based approach to have proper FDR control under data correlation since the permutation is assumed to preserve the correlation structure. To investigate the potential consequence of within-data correlation on our procedure, we performed additional simulations (Supplementary Tables S5 and S6, Additional Simulation 3), where the data were generated using multivariate normal with an AR correlation structure. We then repeated the simulations (S1-S5) and compared our procedure to BH and ST. We found that our procedure not only has conservative FDR control under data correlation but also has a much higher power than BH and ST for phylogeny-informative scenarios.

Our method is quite flexible and users can specify the prior structure in terms of feature-to-feature distance matrix (e.g. sequence similarity) instead of using a tree inferred from it. Besides the phylogenetic tree structure, StructFDR may also be applicable to microbial species interaction networks (Faust and Raes, 2012). Beyond the microbiome applications, our method can be applied to other structure-rich genetic/genomic data. For example, genetic variants in linkage disequilibrium tend to be correlated. Gene networks describe various functional relationships between different genes, which can be positively or negatively correlated. DNA Methylation data also have local clustering phenomenon. Incorporating such information during the modeling stage can be challenging and computationally intensive. Our approach allows a substantial reduction in development time for information 'borrowing' analyses in 'large p, small n' setting. This reduction in development time is essential if statistical analysis methods are to keep pace with the rapid development of new technologies and new applications of those technologies that generate large volumes of biological data. The modulated statistic we propose in this paper efficiently incorporates the prior structure information and a tuning parameter is used to balance information in the prior structure and available data. This essentially produces a different ranking of features and FDR is used as a thresholding rule. In fact, other error rates can be used as well, such as *k*-familywise error rate or the tail probability of the false discovery proportion (Cao and Kosorok, 2011).

Funding

This work was supported by Gerstner Family Career Development Award and Center for Individualized Medicine at Mayo Clinic.

Conflict of Interest: none declared.

References

Benjamini,Y., and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Benjamini,Y., and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Cao,H., and Kosorok,M.R. (2011) Simultaneous critical values for *t*-tests in very high dimensions. *Bernoulli*, **17**, 347–394.

Cao,H., and Wu,W.B. (2015) Changepoint estimation: another look at multiple testing problems. *Biometrika*, **102**, 974–980.

Chen,J. *et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, **28**, 2106–2113.

Chen,J. *et al.* (2012) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–258.

Chen,J. *et al.* (2016) Impact of demographics on human gut microbial diversity in a US midwest population. *PeerJ*, **4**, e1514.

Conneely,K.N., and Boehnke,M. (2007) So many correlated tests, so little time! Rapid adjustment of *P* values for multiple correlated tests. *Am. J. Hum. Genet.*, **81**, 1158–1168.

Draper,N.R., and Smith,H. (1998) *Applied Regression Analysis*. 3rd edn, John Wiley and Sons, New York.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Efron,B. (2007) Correlation and large-scale simultaneous testing. *J. Am. Stat. Assoc.*, **102**, 93–103.

Engen,P.A. *et al.* (2015) The gastrointestinal microbiome: alcohol effects on the composition of intestinal microbiota. *Alcohol. Res.*, **37**, 223–236.

Fan,J. *et al.* (2012) Control of the false discovery rate under arbitrary covariance dependence. *J. Am. Stat. Assoc.*, **107**, 1019–1045.

Faust,K., and Raes,J. (2012) Microbial interactions: from networks to models. *Nat. Rev. Microbial.*, **10**, 538–550.

Ferreira,J., and Zwinderman,A. (2006) On the Benjamini–Hochberg Method. *Ann. Stat.*, **34**, 1827–1849.

Friguet,C. *et al.* (2009) A factor model approach to multiple testing under dependence. *J. Am. Stat. Assoc.*, **104**, 1406–1415.

Gilbert,J.A. *et al.* (2016) Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, **535**, 94–103.

Goberna,M., and Verdú,M. (2016) Predicting microbial traits with phylogenies. *ISME J*, **10**, 959–967.

Hu,J.X. *et al.* (2010) False discovery rate control with groups. *J. Amer. Stat. Assoc.*, **105**, 1215–1227.

Ignatiadis,N. *et al.* (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods*, **13**, 577–580.

Kang,G. *et al.* (2009) Weighted multiple hypothesis testing procedures. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–22.

Kuczynski,J. *et al.* (2011) Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.*, **13**, 47–58.

Leclercq,S. *et al.* (2014) Intestinal permeability, gut-bacterial dysbiosis, and behavioral markers of alcohol-dependence severity. *Proc. Natl. Acad. Sci. USA*, **111**, E4485–E4493.

Leek,J.T., and Storey,J.D. (2008) A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA*, **105**, 18718–18723.

Li,C. *et al.* (2010) Network-based empirical Bayes methods for linear models with applications to genomic data. *J. Biopharm. Stat.*, **20**, 209–222.

Martin,E.P., and Hansen,T.F. (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.*, **149**, 646–667.

Martiny,B.H. *et al.* (2015) Microbiomes in light of traits: a phylogenetic perspective. *Science*, **350**, aac9323–aac9323.

Miller,C. *et al.* (2001) Controlling the false-discovery rate in astrophysical data analysis. *Astronom. J.*, **122**, 3492–3505.

Morgan,X.C. *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 Biol.*, **16**, 67.

Owen,A.B. (2005) Variance of the number of false discoveries. *J. R. Stat. Soc. B*, **67**, 411–426.

Price,M.N. *et al.* (2010) FastTree 2: approximately maximum-likelihood tress for large alignments. *PLoS One*, **5**, e9490.

Purdom,E. (2011) Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.*, **5**, 2326–2358.

Sankaran,K., and Holmes,S. (2014) structSSI: simultaneous and selective inference for grouped or hierarchically structured data. *J. Stat. Softw.*, **59**, 1–21.

Scholz,M. *et al.* (2016) Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods*, **13**, 435–438.

Schwartzman,A. *et al.* (2008) False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.*, **2**, 153–175.

Silverman,J.D. *et al.* (2017) A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, **6**, e21887.

Storey,J. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

Storey,J. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B*, **66**, 187–205.

Storey,J., and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.

Sun,W., and Cai,T. (2009) Large-scale multiple testing under dependence. *J. R. Stat. Soc. B*, **71**, 393–424.

Tusher,V. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.

Wei,Z., and Li,H. (2007) A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**, 1537–1544.

Willet,W.C. *et al.* (1997) Adjustment of total energy intake in epidemiological studies. *Am. J. Clin. Nutr.*, **65**, 12205–12285.

Wu,W.B. (2008) On false discovery rate control under dependence. *Ann. Stat.*, **36**, 364–380.

Wu,G.D. *et al.* (2011) Linking long-term dietary pattern with gut microbial enterotypes. *Science*, **334**, 105–108.

Xie,Y. *et al.* (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, **21**, 4280–4288.

Xie,J. *et al.* (2011) Optimal false discovery rate control for dependent data. *Stat. Interface*, **4**, 417–430.

Yekutieli,D. (2008) Hierarchical false discovery rate? controlling methodology. *J. Amer. Stat. Assoc.*, **103**, 209–316.