Microbiome

METHODOLOGY Open Access

A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals

Hyunwook Koh and Ni Zhao[*]

Abstract

Background: In human microbiome studies, it is crucial to evaluate the association between microbial group (e.g., community or clade) composition and a host phenotype of interest. In response, a number of microbial group association tests have been proposed, which account for the unique features of the microbiome data (e.g., high-dimensionality, compositionality, phylogenetic relationship). These tests generally fall in the class of aggregation tests which amplify the overall group association by combining all the underlying microbial association signals, and, therefore, they are powerful when many microbial species are associated with a given host phenotype (i.e., low sparsity). However, in practice, the microbial association signals can be highly sparse, and this is especially the situation where we have a difficulty to discover the microbial group association.

Methods: Here, we introduce a powerful microbial group association test for sparse microbial association signals, namely, microbiome higher criticism analysis (MiHC). MiHC is a data-driven omnibus test taken in a search space spanned by tailoring the higher criticism test to incorporate phylogenetic information and/or modulate sparsity levels and including the Simes test for excessively high sparsity levels. Therefore, MiHC robustly adapts to diverse phylogenetic relevance and sparsity levels.

Results: Our simulations show that MiHC maintains a high power at different phylogenetic relevance and sparsity levels with correct type I error controls. We also apply MiHC to four real microbiome datasets to test the association between respiratory tract microbiome and smoking status, the association between the infant's gut microbiome and delivery mode, the association between the gut microbiome and type 1 diabetes status, and the association between the gut microbiome and human immunodeficiency virus status.

Conclusions: In practice, the true underlying association pattern on the extent of phylogenetic relevance and sparsity is usually unknown. Therefore, MiHC can be a useful analytic tool because of its high adaptivity to diverse phylogenetic relevance and sparsity levels. MiHC can be implemented in the R computing environment using our software package freely available at https://github.com/hk1785/MiHC.

Keywords: Microbiome association studies, Microbial ecology, Adaptive association analysis, Higher criticism, Sparse microbial associations, Phylogenetics

* Correspondence: nzhao10@jhu.edu
Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe Street, Office E3622, Baltimore, MD 21205, USA

Background

The recent advent of next-generation sequencing has enabled unbiased microbiome profiling for all microbes inhabiting in different organs of the human body. The two major sequencing platforms for microbiome profiling are the targeted polymerase chain reaction amplicons for the 16S ribosomal RNA (rRNA) gene [1, 2] and the shotgun metagenomics for the whole microbial genome [3]. These sequencing platforms produce various types of metagenomic information, such as microbial abundance, gene content, and metabolic capacity [4]. Among those, here we focus on a common type of the microbiome data for the microbial composition with relative abundances and phylogenetic relationships. We also consider the operational taxonomic unit (OTU) as a surrogate of microbial species and the smallest unit of the microbial biomarkers nested in different microbial assemblages (e.g., communities (bacteria, fungi, viruses), upper- or lower-level taxa (phyla, classes, orders, families, genera)). The roles of the microbiome on human health or disease have been intensely studied throughout all different microbial assemblages. For example, the community of bacteria has been primarily studied on the disparity in microbial diversity among different populations (e.g., diseased vs. non-diseased, treatment vs. placebo) [5–8]. While the communities of fungi or viruses have been less studied, they are gaining more and more attention [9, 10]. Moreover, investigators have intensely studied the disparity in microbial taxon composition throughout a breadth of hierarchical taxonomic classifications (e.g., phyla to genera) [11, 12].

Here, we refer, in general, the study on the association between any microbial group (e.g., community or clade) composition and a host phenotype (or any other health/disease-related factor) as a microbial group association study. In response to the popularity of such studies, researchers have proposed a number of microbial group association tests while incorporating the unique features of the microbiome data (e.g., high-dimensionality, compositionality, phylogenetic relationship) into their proposed tests. The most popular approaches are the association tests using α- or β-diversity indices [5, 6]. α-diversity measures within-sample diversity, by which the high-dimensional microbiome information can be projected into a single diversity variable. We can then easily test the association between an α-diversity index and a host phenotype using a traditional statistical method (e.g., generalized linear models), or we can jointly consider multiple α-diversity indices and conduct an omnibus microbial diversity association analysis using the adaptive microbiome α-diversity-based association test (aMiAD) [13]. On the other hand, β-diversity measures between-sample diversity (i.e., dissimilarity or distance), by which the high-dimensional microbiome information can be projected into a full rank similarity matrix via a kernel machine framework [14]. We can then test the association using the ANOVA-type association test known as the permutational multivariate analysis of variance (PERMANOVA) [15–17] or the regression-type association test known as the regression-microbiome regression-based kernel association test (MiRKAT) [14], while they result in a similar performance [14]. Researchers have also proposed diverse microbial group association tests to amplify the overall group association by combining all the underlying microbial association signals (e.g., the microbiome sum of powered score tests (MiSPU) [18]).

All the above tests generally fall in the class of aggregation tests as all the underlying microbial association signals are aggregated into the α- or β-diversity or the overall group association statistic [19]. Therefore, they are powerful when a large number of OTUs are associated with a host phenotype (i.e., low sparsity) [19]. However, in practice, it is possible that only few OTUs are associated with a host phenotype (i.e., high sparsity), and, as an extreme case, even a single OTU can cause human disease (e.g., a small influx of *Escherichia coli* O157:H7 can cause food poisoning [20]). However, it is questionable if the current methods can powerfully discover the microbial group association for the high sparsity situation. For example, it is so obvious that there is a huge disparity in a variety of host phenotypes between the normal and germ-free mice because of the huge disparity in their microbiomes (i.e., presence vs. absence of microbiome) [21], and, in this low sparsity situation, any of the current methods can powerfully discover the microbial group association with no need for any additional method development. Thus, here we instead move our focus onto the high sparsity situation, in which only a small portion of the OTUs are associated with a host phenotype and the pressing issue of powerfully discovering the disparity in a host phenotype driven by the sparse association signals.

We notice that the group association test, known as higher criticism (HC) test, is powerful at high sparsity levels because its test statistic reflects only the single largest association signal among underlying individual association signals [22]. While the use of the higher criticism test has been extended to genome-wide association studies [23, 24], it has not been well-appreciated for the microbiome group association analysis. This might be because of the unique features of the microbiome data and the resulting need for more sophisticated analysis procedures. Thus, here we further tailor the higher criticism test for microbial group association analysis by incorporating phylogenetic information and modulating sparsity levels, as follows. First, we notice that phylogenetically relevant species share similar genetic components and evolutionary histories and, as a result, they are likely to have similar functional effects on a host phenotype [25]. Thus, to improve power when the OTUs associated with a host phenotype are phylogenetically relevant, we introduce a weighted higher criticism test which gives a higher weight to the OTUs whose phylogenetically relevant OTUs have larger association signals. Second, the original higher criticism test is powerful at a high sparsity level but rapidly loses power as the sparsity level decreases. Thus, to improve power for lower sparsity levels, we introduce a modulated higher criticism test which flexibly reflects the single or multiple largest association signal(s) among underlying individual association signals. In addition, we notice that the Simes test [26] is also powerful at high sparsity levels because it requires only a single strong association signal among underlying individual association signals which is significant even after the multiple testing correction. We heuristically, but not theoretically, found that the Simes test is more powerful at excessively high sparsity levels than the higher criticism test while the Simes test more rapidly loses power as the sparsity level decreases than the higher criticism test (see the "Simulation results" section).

Here, the dilemma in reality is that the OTUs associated with a host phenotype can be phylogenetically relevant or not, and they can be highly sparse or less sparse. Yet, unfortunately, we cannot presume which specific association pattern underlies our study in advance because of the lack of prior knowledge. Thus, here we introduce a data-driven omnibus test, namely, microbiome higher criticism analysis (MiHC), which robustly adapts to diverse association patterns. To achieve the robust adaptivity, we first construct multiple candidate tests by combining the principles of the original, weighted and modulated higher criticism tests, and the Simes test, in which each of the individual candidate tests suits some specific association pattern. Then, we use the minimum *p* value among those candidate tests as the test statistic of MiHC with the aim of closely reaching the highest power among those candidate tests. Finally, we use a residual-based permutation approach based on the minimum *p* value statistic to calculate the *p* value for MiHC. Here, the residual-based permutation approach enables to preserve OTU-by-OTU correlations [27], which are inherent in the microbiome data because of the compositional constraint (also known as unit sum constraint), phylogenetic relevance, and other potential sources.

Our extensive simulations show that MiHC robustly maintains a high power at different phylogenetic relevance and sparsity levels with correct type I error controls at the significance level of 5%. We also apply MiHC to four real microbiome datasets to test the association between respiratory tract microbiome and smoking status [28], the association between the infant's gut microbiome and delivery mode [29], the association between the gut microbiome and type 1 diabetes status [30], and the association between the gut microbiome and human immunodeficiency virus (HIV) status [31].

Methods and materials

This section is devoted to describe methodological details (i.e., models, notations, test statistics, and computational procedures) for our proposed methods. Since we use many notations, we organize them in a summary table for easy follow-up (Additional file 1: Table S1).

Generalized linear models and marginal score statistics

We suppose that the data include n samples, m OTUs in a microbial group of interest (e.g., community or clade) and l covariates (e.g., age, gender). Let y_i denote a host phenotype (or any other health/disease-related factor) of interest, o_{ij} denote an OTU in relative abundance (i.e., proportion), and x_{ik} denote a covariate for $i = 1,\dots, n$, $j = 1,\dots, m$ and $k = 1,\dots, l$. To test the association between OTUs and a host phenotype adjusting for covariates, we consider a generalized linear model [32] (Eq. 1),

$$g\left(\mu_i\right) = x_i^T\alpha + o_i^T\beta, \qquad (1)$$

where $g(\cdot)$ is a canonical link function, $\mu_i = E(y_i \mid x_i, o_i)$, and $\alpha = (\alpha_0, \dots, \alpha_l)^T$ and $\beta = (\beta_1, \dots, \beta_m)^T$ are the regression coefficients for the covariates, $x_i = (1, x_{i1}, \dots, x_{il})^T$, and the OTUs, $o_i = (o_{i1}, \dots, o_{im})^T$, respectively. Here, y_i conditional on x_i and o_i is assumed to follow a distribution in the exponential dispersion family with the probability density/mass function (Eq. 2).

$$f(y_i| \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \theta_i)\right\}, \qquad (2)$$

where θ_i is the natural parameter, ϕ is the dispersion parameter, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are the known functions [32]. Let $b^{'}(\theta_i)$ and $b^{''}(\theta_i)$ denote the first two derivatives of $b(\theta_i)$ evaluated at θ_i; as such, $E(y_i \mid x_i, o_i) = b^{'}(\theta_i)$ and $\text{Var}(y_i \mid x_i, o_i) = a_i(\phi)b^{''}(\theta_i)$. Here, we are interested in testing the global null hypothesis of no association between OTUs and a host phenotype adjusting for covariates (Eq. 3).

$$\begin{aligned} &H_0 : \beta_j = 0 \text{ for all } j^{'}\text{sin } \{1, \dots, m\} \text{ vs. } . \\ &H_1 : \beta_j \neq 0 \text{ for some } j^{'}\text{sin } \{1, \dots, m\} \end{aligned} \qquad (3)$$

While we will soon address the above global hypothesis testing in the following sections, here we first delineate the marginal standardized score statistic for each OTU (Eq. 4) as it is the key component of the higher criticism test [24].

$$Z_j = \frac{o_j^T(y - \hat{\mu}_0)}{\sqrt{o_j^T P o_j}}, \tag{4}$$

where $o_j = (o_{1j}, \ldots, o_{nj})^T$, $y = (y_1, \ldots, y_n)^T$, $\hat{\mu}_0$ is the vector of the expected values of y_i's estimated under the null model of $g(\mu_i) = x_i^T \alpha$; $\hat{\mu}_0 = (\hat{\mu}_{1,0}, \ldots, \hat{\mu}_{n,0})^T = (g^{-1}(x_1^T \hat{\alpha}_0), \ldots, g^{-1}(x_n^T \hat{\alpha}_0))^T = (b^{'}(\hat{\theta}_{1,\ 0}), \ldots, b^{'}(\hat{\theta}_{n,\ 0}))^T$, and $P = W - WX(X^T WX)^{-1} X^T W$, where W is the diagonal matrix of the marginal variances of y_i's estimated under the null model of $g(\mu_i) = x_i^T \alpha$; $W = \mathrm{diag}(a_1(\hat{\phi}_0)b''(\hat{\theta}_{1,\ 0}), \ldots, a_n(\hat{\phi}_0)b''(\hat{\theta}_{n,\ 0}))$, and $X = (x_1, \ldots, x_n)^T$, for $j = 1, \ldots, m$. Here, the statistic Z_j tells the effect direction and size for the jth OTU, and we assume that Z_j follows the standard normal distribution $N(0,1)$ under the marginal null hypothesis of $\beta_j = 0$. Then, we can calculate the marginal p value for the jth OTU as $P(|Z_j| > N(0,1))$.

Unweighted and weighted higher criticism analyses

Donoho and Jin first derived the higher criticism test, motivated by an idea of the great statistician, John Wilder Tukey [22]. Then, the higher criticism test has been further developed by a few follow-up studies [23, 24, 33, 34]. While there are different forms of the test statistic, we use the simplest form of (Eq.5) based on [23].

$$\mathrm{uHC} = \max_{j \in \{1,\ldots,m\}} \left\{ \frac{r_j/m - p_j}{\sqrt{p_j(1-p_j)/m}} \right\}, \tag{5}$$

where uHC is the test statistic for the higher criticism test [23], p_j is the p value for the jth OTU, and r_j is the rank of p_j in the ascending order of p_j's for $j = 1, \ldots, m$. We denote this higher criticism test as the unweighted higher criticism (uHC) test in order to distinguish it from the forthcoming weighted higher criticism test. Here, a relatively large observed statistic value compared with null statistic values indicates a higher chance to discover the group association. Prior studies have found that this higher criticism test sensitively detects highly sparse association signals [22–24, 33]. The major rationale behind is that the test statistic (Eq. 5) focuses on the single largest deviation between the expected ($\frac{r_j/m}{\sqrt{p_j(1-p_j)/m}}$) and observed ($\frac{p_j}{\sqrt{p_j(1-p_j)/m}}$) quantiles of significance among all the m tests; as such, only a small number of association signals are sufficient to get a large statistic value [22–24].

In microbiome association studies, phylogenetically relevant species tend to have similar effects on a host phenotype because of their similarities in genetic components and evolutionary histories [25]. Thus, to improve power when the OTUs associated with a host phenotype are phylogenetically relevant, we introduce the weighted higher criticism (wHC) test (Eq. 6).

$$\mathrm{wHC} = \max_{j \in \{1,\ldots,m\}} \left\{ \frac{w_j(r_j/m - p_j)}{\sqrt{p_j(1-p_j)/m}} \right\}, \tag{6}$$

where wHC is the test statistic for the weighted higher criticism test and w_j is the weight for the jth OTU. To assign the weight to each OTU (i.e., w_j for $j = 1, \ldots, m$), we first partition the m OTUs into C phylogenetically close clusters based on OTUs' pairwise cophenetic distances [35], where the cophenetic distance of any two OTUs refers to the total length of the branches to their most common ancestor (i.e., the closest intersection) in the phylogenetic tree and we calculate it using the function, *cophenetic*, in the R package, *stats*. For this, we use the partitioning-around-medoids algorithm [36] based on the optimal number of clusters (C) which maximizes the average silhouette width searching up to 30 clusters [36]. Let $\zeta(j)$ denote a cluster anchored at the jth OTU among the C clusters. Then, we define w_j as (Eq. 7),

$$w_j = \frac{\sum_{j' \in \zeta(j) \backslash \{j\}} \frac{1}{D_{j,j'}} \mid Z_{j'} \mid}{\sum_{j' \in \zeta(j) \backslash \{j\}} \frac{1}{D_{j,j'}}} + 1, \tag{7}$$

where $D_{j,j'}$ is the cophenetic distance between jth and j'th OTUs, $j \in \zeta(j)$ and $j' \in \zeta(j) \backslash \{j\}$. w_j is designed to give a higher weight to the OTUs whose neighboring OTUs, with respect to closer phylogeny (see $\frac{1}{D_{j,j'}}$), have larger association signals (see $\mid Z_{j'} \mid$). Therefore, w_j amplifies the association signals from close phylogeny and hence can suit when the OTUs associated with a host phenotype are phylogenetically relevant.

Modulated higher criticism analyses for lower sparsity levels

Again, the higher criticism test is powerful for high sparsity levels, but it is underpowered for low sparsity levels [24]. In practice, the true associations are not always so highly sparse that the higher criticism can be underpowered. Thus, to improve power for lower sparsity levels, we make some modulations to the original test statistic as (Eq. 8),

$$\mathrm{uHC}_{(h)} = \frac{1}{h} \sum_{j'=1}^{h} \frac{r_{j'}/m - p_{j'}}{\sqrt{p_{j'}\left(1 - p_{j'}\right)/m}}, \tag{8}$$

where uHC$_{(h)}$ is the test statistic for the unweighted higher criticism test for a given h value, $\frac{r_{j'}/m-p_{j'}}{\sqrt{p_{j'}(1-p_{j'})/m}}$ is the j'th order statistics of $\frac{r_j/m-p_j}{\sqrt{p_j(1-p_j)/m}}$'s in the descending order for j = 1,..., m, and h needs to be pre-specified, $h \in \{1, 2, ..., m\text{-}1, m\}$. uHC$_{(h)}$ is the average of the first h largest deviations between the expected ($\frac{r_j/m}{\sqrt{p_j(1-p_j)/m}}$) and observed ($\frac{p_j}{\sqrt{p_j(1-p_j)/m}}$) quantiles of significance among all the m tests (Eq. 8). uHC$_{(h)}$ is also a generalization of the original higher criticism test (Eq. 5) because when h = 1, uHC$_{(h)}$ becomes the original higher criticism test (i.e., uHC$_{(1)}$). uHC$_{(1)}$ relies on the single largest deviation and hence can suit high sparsity levels. As h increases, uHC$_{(h)}$ considers more deviations to the next level association signals and hence can suit lower sparsity levels. When h=m, uHC$_{(h)}$ becomes uHC$_{(m)}$. uHC$_{(m)}$ considers all the m deviations and hence can suit the least sparsity level. Without loss of generality, we can apply the same modulations to the weighted higher criticism (wHC) test (Eq. 9).

$$\mathrm{wHC}_{(h)} = \frac{1}{h}\sum_{j'=1}^{h} \frac{w_{j'}\left(r_{j'}/m-p_{j'}\right)}{\sqrt{p_{j'}\left(1-p_{j'}\right)/m}}, \tag{9}$$

where wHC$_{(h)}$ is the test statistic for the weighted higher criticism test for a given h value, and $\frac{w_{j'}(r_{j'}/m-p_{j'})}{\sqrt{p_{j'}(1-p_{j'})/m}}$ is the j'th order statistics of $\frac{w_j(r_j/m-p_j)}{\sqrt{p_j(1-p_j)/m}}$'s in the descending order for j = 1,..., m. We calculate the p values for the individual unweighted (uHC$_{(h)}$'s) and weighted (wHC$_{(h)}$'s) tests based on a permutation method (see the "p value calculation" section).

Simes test

Simes (1986) introduced a simple modification of the Bonferroni procedure for multiple hypothesis testing and a group association test, known as the Simes test, that calculates the p value as the minimum p value among the marginal p values that are corrected by the Bonferroni procedure (i.e., multiplied by the number of tests) and weighted by the inverse of their ranks (i.e., multiplied by the inverse of their ranks) [26] (Eq. 10).

$$P_{\mathrm{Simes}} = T_{\mathrm{Simes}} = \min_{j\in\{1,\ldots,m\}}\left\{\frac{mp_j}{r_j}\right\}, \tag{10}$$

where P_{Simes} and T_{Simes} are the p value and the test statistic for the Simes test, p_j is the p value for the jth OTU, and r_j is the rank of p_j in the ascending order of p_j's for j = 1,..., m. To discover the group association, the Simes test requires only a single strong association signal which can produce a significant p value even after adjusting for multiple hypothesis testing. Thus, the Simes test is also powerful at highly sparsity levels. Our simulations demonstrate that the Simes test is more powerful at excessively high sparsity levels than the higher criticism test while the Simes test more rapidly loses power as the sparsity level decreases (see the "Simulation results" section).

Microbiome higher criticism analysis

In reality, the true microbial associations can be phylogenetically relevant or not, and they can be highly sparse or less sparse, yet we do not know the true underlying association pattern in advance. Thus, to robustly adapt to the unknown phylogenetic relevance and sparsity levels, we propose a data-driven omnibus test, namely, microbiome higher criticism (MiHC) analysis (Eq. 11).

$$T_{\mathrm{MiHC}} = \min\left(\min_{h\in\Gamma}\left(P_{\mathrm{uHC}(h)}, P_{\mathrm{wHC}(h)}\right), P_{\mathrm{Simes}}\right), \tag{11}$$

where $P_{\mathrm{uHC}(h)}$'s are the p values based on the uHC$_{(h)}$ tests, $P_{\mathrm{wHC}(h)}$'s are the p values based on the wHC$_{(h)}$ tests for h's in a set Γ (e.g., Γ = {1, 3, 5, 7, 9}), and P_{Simes} is the p value based on the Simes test. T_{MiHC} is the minimum p value among all the uHC$_{(h)}$ (Eq. 8) and wHC$_{(h)}$ (Eq. 9) tests for h's in Γ and the Simes test (Eq. 10). Of course, we do not use this minimum p value as the final p value for MiHC, but we instead use it as the test statistic of MiHC. We calculate the p value for MiHC based on a permutation method (see the "p value calculation" section). This kind of the minimum p value statistic approach has also been widely used in many prior association tests [13, 14, 18, 37–39]. The set (Γ) can be spanned up to the union set of {1, 2,..., m-1, m}. However, it is a huge computational burden to survey all the h values in the union set because of the high-dimensionality of the microbiome data. Thus, we use a candidate set of Γ = {1, 3, 5, 7, 9} and it was sufficient in our simulations and real data applications. The use of the minimum p value statistic allows MiHC to closely approach the most powerful test among all the candidate tests in Γ and the Simes test. Therefore, compared with the original higher criticism test (which is only for h = 1) or the Simes test, our candidate set always gives a similar or higher power. Our extensive simulation experiments demonstrate the high adaptivity of MiHC to various phylogenetic relevance and sparsity levels while robustly maintaining a high power with well-controlled type I error rates (see the "Simulation results" section).

By the same logic, we can also consider two local omnibus tests, namely, uHC$_A$ (Eq. 12) and wHC$_A$ (Eq. 13), that are taken within each of the two sub-domains: (1) the unweighted higher criticism tests (i.e., $\mathrm{uHC}_{(h)}$tests for h's in Γ) and the Simes test and (2) the weighted higher criticism tests (i.e., $\mathrm{wHC}_{(h)}$ tests for h's in Γ) and the Simes test.

$$T_{\mathrm{uHC}_A} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{uHC}(h)}\right), P_{\mathrm{Simes}}\right), \tag{12}$$

$$T_{\mathrm{wHC}_A} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{wHC}(h)}\right), P_{\mathrm{Simes}}\right), \tag{13}$$

T_{uHC_A} is the minimum p value among $\mathrm{uHC}_{(h)}$ (Eq. 8) tests and the Simes test (Eq. 10) while T_{wHC_A} is the minimum p value among $\mathrm{wHC}_{(h)}$ (Eq. 9) tests and the Simes test (Eq. 10). These two local omnibus tests are distinguished from the global omnibus test, MiHC, that is taken within the global domain of all the unweighted and weighted higher criticism tests (i.e., all the $\mathrm{uHC}_{(h)}$ and $\mathrm{wHC}_{(h)}$ tests for h's in Γ) and the Simes test. T_{uHC_A} and T_{wHC_A} are the test statistics of uHC_A and wHC_A, respectively, and we calculate the p value based on a permutation method (see the "*p* value calculation" section). By the formula, we can infer that uHC_A and wHC_A can modulate sparsity levels through h's in Γ and the Simes test for excessively high sparsity levels, while uHC_A suits the low phylogenetic relevance, but wHC_A suits the high phylogenetic relevance. Although the global omnibus test (i.e., MiHC) (Eq. 11) is our major proposal for microbial group association analysis, we introduce these two local omnibus tests (i.e., uHC_A and wHC_A) especially because uHC_A is useful to modulate sparsity levels when the phylogenetic information is not available (e.g., microbial functional studies for genetic/metabolic content).

p value calculation

There have been different approaches to calculate the p value for the higher criticism test [22–24, 33, 34]. The analytical approaches based on an asymptotic distribution proposed in [22, 33, 34] have the advantage of producing a closed-form p value in a computationally efficient manner. However, the analytical approaches assume independent tests and/or rely on asymptotics in m which requires m as large as a million for valid statistical inferences [23]. In microbiome association studies, the independence assumption can be easily violated because of the inherent compositional constraint and phylogenetic relevance. Furthermore, the microbiome data do not usually include a million OTUs so that the slow convergence rate to asymptotics in m can lead to invalid statistical inferences [23]. Thereafter, Barnett et al. proposed an exact p value calculation which releases the independence assumption and the huge m requirement. However, its computational burden increases exponentially as m increases; hence, it can handle only a small number of OTUs. Therefore, instead of using the asymptotic or the exact method, we use a permutation method to calculate the p value for our proposed method. In particular, we use the following procedures.

1. Fit the null generalized linear model and estimate the residuals as $\hat{e}_0 = \left(y_1 - g^{-1}\left(x_1^T\hat{\alpha}_0\right), \ldots, y_n - g^{-1}\left(x_n^T\hat{\alpha}_0\right)\right)^T$.
2. Calculate the marginal score statistics as $Z_j = o_j^T\hat{e}_0 / \sqrt{o_j^T P o_j}$ (Eq. 4) and the marginal p values as $p_j = P(|Z_j| > N(0,1))$ for $j = 1,\ldots, m$. Calculate the observed statistics, $\mathrm{uHC}_{(h)}$ (Eq. 8) and $\mathrm{wHC}_{(h)}$ (Eq. 9), for each $h \in \Gamma$, and the p value for the Simes test, P_{Simes} (Eq. 10).
3. Permute the estimated residuals $\hat{e}_0$ multiple times (say, B times) and denote each permuted residual vector as $e_b^{'}$ for $b = 1,\ldots, B$. Repeat step 2 B times, replacing $\hat{e}_0$ with each $e_b^{'}$, and calculate the null statistics, $\mathrm{uHC}_{(h)(b)}$ (Eq. 8) and $\mathrm{wHC}_{(h)(b)}$ (Eq. 9), for each $h \in \Gamma$ and for each $b \in \{1,\ldots, B\}$ and the null statistics for the Simes test, $T_{\mathrm{Simes}(b)} = \min_{j\in\{1,\ldots,m\}}\left\{\frac{mp_{j(b)}}{r_{j(b)}}\right\}$ for each $b \in \{1,\ldots, B\}$.
4. Calculate $P_{\mathrm{uHC}(h)} = \sum_{b=1}^{B}\left[I\left(\mathrm{uHC}_{(h)(b)} > \mathrm{uHC}_{(h)}\right)+1\right]/(B+1)$ and $P_{\mathrm{wHC}(h)} = \sum_{b=1}^{B}\left[I\left(\mathrm{wHC}_{(h)(b)} > \mathrm{wHC}_{(h)}\right)+1\right]/(B+1)$ for each $h \in \Gamma$. Calculate the observed statistics, $T_{\mathrm{uHC}_A} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{uHC}(h)}\right), P_{\mathrm{Simes}}\right)$ (Eq. 12), $T_{\mathrm{wHC}_A} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{wHC}(h)}\right), P_{\mathrm{Simes}}\right)$ (Eq. 13) and $T_{\mathrm{MiHC}} = \min\left(\min_{h\varepsilon}\Gamma\left(P_{\mathrm{uHC}(h)}, P_{\mathrm{wHC}(h)}\right), P_{\mathrm{Simes}}\right)$ (Eq. 11).
5. Calculate $P_{\mathrm{uHC}(h)(b)} = \sum_{b'\neq b}\left[I\left(\mathrm{uHC}_{(h)(b')} > \mathrm{uHC}_{(h)(b)}\right) + 1\right]/(B+1)$ and $P_{\mathrm{wHC}(h)(b)} = \sum_{b'\neq b}\left[I\left(\mathrm{wHC}_{(h)(b')} > \mathrm{wHC}_{(h)(b)}\right) + 1\right]/(B+1)$ for each $h \in \Gamma$, and $P_{\mathrm{Simes}(b)} = \sum_{b'\neq b}\left[I\left(T_{\mathrm{Simes}(b')} < T_{\mathrm{Simes}(b)}\right) + 1\right]/(B+1)$, where $b \in \{1,\ldots, B\}$ and $b^{'} \in \{1,\ldots, B\}$. Calculate the null statistics, $T_{\mathrm{uHC}_{A(b)}} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{uHC}(h)(b)}\right), P_{\mathrm{Simes}(b)}\right)$ (Eq. 12), $T_{\mathrm{wHC}_{A(b)}} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{wHC}(h)(b)}\right), P_{\mathrm{Simes}(b)}\right)$ (Eq. 13) and $T_{\mathrm{MiHC}(b)} = \min\left(\min_{h\epsilon\Gamma}\left(P_{\mathrm{uHC}(h)(b)}, P_{\mathrm{wHC}(h)(b)}\right), P_{\mathrm{Simes}(b)}\right)$ (Eq. 11), for $b = 1,\ldots, B$.
6. Calculate the p values for uHC_A as $P_{\mathrm{uHC}_A} = \sum_{b=1}^{B}\left[I\left(T_{\mathrm{uHC}_A(b)} < T_{\mathrm{uHC}_A}\right) + 1\right]/(B+1)$, wHC_A as $P_{\mathrm{wHC}_A} = \sum_{b=1}^{B}\left[I\left(T_{\mathrm{wHC}_A(b)} < T_{\mathrm{wHC}_A}\right) + 1\right]/(B+1)$ and MiHC as $P_{\mathrm{MiHC}} = \sum_{b=1}^{B}\left[I\left(T_{\mathrm{MiHC}(b)} < T_{\mathrm{MiHC}}\right) + 1\right]/(B+1)$.

Importantly, our permutation method can robustly account for any correlation structure among the m tests using the same permuted residual vectors repeatedly for each test (i.e., residual-based permutation) [27]. Moreover, since MiHC is based on the score test (Eq. 4) which is computationally efficient and the null model needs to be fitted only once, our method is computationally manageable.

Visualization

Here, we introduce simple Q-Q plots to demonstrate influential OTUs in each of the two sub-domains (i.e., uHC and wHC) for MiHC. First, we draw Q-Q plots between the expected ($\frac{r_j/m}{\sqrt{p_j(1-p_j)/m}}$) and observed ($\frac{p_j}{\sqrt{p_j(1-p_j)/m}}$) quantiles for uHC (Eq. 5) and between the expected ($\frac{w_j(r_j/m)}{\sqrt{p_j(1-p_j)/m}}$) and observed ($\frac{w_j p_j}{\sqrt{p_j(1-p_j)/m}}$) quantiles for wHC (Eq. 6), respectively, for $j = 1,\ldots, m$. Here, we use (blue) dots to represent individual OTUs and a (red) diagonal line with intercept 0 and slope 1 to represent no influential points; as such, the OTUs that fall along the diagonal line have no influence on the host phenotype while the OTUs that have larger deviations from the diagonal line are more influential on the host phenotype. Then, we report the 10 most influential OTUs corresponding to the 10 largest deviations from the diagonal line with respect to uHC and wHC, respectively. We use darker to lighter vertical lines to represent more to less influential OTUs in rank order among the 10 most influential OTUs. Example visualizations can be found later in the "Real data applications" section.

Simulation results

We conducted simulation experiments to compared MiHC with the prior tests, Simes test [26], higher criticism (HC) test (i.e., uHC (Eq. 5)) [22], aMiAD [13], adaptive MiSPU (aMiSPU) [18], and Optimal MiRKAT (OMiRKAT) [14]. Our simulation design is based on prior studies [14]. We first estimated the proportions and dispersion for the 100 most abundant OTUs from the real respiratory tract microbiome data [28] based on the Dirichlet-multinomial model [40]. Then, we iteratively generated an OTU count table using the Dirichlet-multinomial model with the estimated proportions and dispersion and a rooted phylogenetic tree with 100 leaves using the function, *rtree*, in the R package, *ape* [41]. Here, we fixed the total reads per sample as 1000 to mimic the compositional constraint and considered two different sample sizes, n = 50 and n = 100, respectively. To illustrate fits of the simulated data, we generated histograms of the relative abundances for the 100 most abundant OTUs of the real respiratory tract microbiome data and the simulated data based on the Dirichlet-multinomial model (Additional file 2: Figure S1). For the relative abundances of the simulated data, we took averages across 100 simulated data sets that were iteratively generated based on the Dirichlet-multinomial model for n = 50 and n = 100, respectively. To estimate type I error rates and powers, we generated Gaussian responses based on the linear regression model below.

$$y_i = 0.5 \times \mathrm{scale}(x_{1i}) + 0.5 \times \mathrm{scale}(x_{2i}) + \beta \times \sum_{j \in \Lambda} \mathrm{scale}(o_{ij}) + \varepsilon_i,$$

where x_{i1} are x_{i2} are the covariates generated from the Bernoulli distribution with success probability 0.5 and the standard normal distribution $N(0, 1)$, respectively, Λ is a set of OTUs that are associated with the host phenotype y_i, ε_i is an error term generated from the standard normal distribution $N(0, 1)$, and scale is the standardization function to have mean 0 and standard deviation 1.

To estimate type I error rates, we assigned β = 0 to reflect the null hypothesis of no association for all OTUs (Eq. 3). To estimate powers, we assigned β = 1 for n = 50 and β = 0.5 for n = 100, while choosing the set of associated OTUs (Λ) based on two different scenarios: (1) we randomly selected 2%, 4%, 6%, 8%, 10%, or 12% of the OTUs to be associated with the host phenotype and (2) we selected 2%, 4%, 6%, 8%, 10%, or 12% of the OTUs which are phylogenetically close to be associated with the host phenotype. We regard the second scenario more realistic because the phylogenetic relevance likely to give shared functional attributes. In particular, for the second scenario, we randomly selected one OTU as a seed OTU and then included 2%, 4%, 6%, 8%, 10%, or 12% of the OTUs that are closest to the seed OTU (including the seed OTU) with respect to cophenetic distance [35]. For both of the scenarios, 2%, 4%, 6%, 8%, 10%, and 12% reflect from high to low sparsity levels.

Results

Simulation results

Fits of the simulated data

Additional file 2: Figure S1 reports the histograms of the relative abundances for the real respiratory tract microbiome data [28] and the simulated data based on the Dirichlet-multinomial model [40]. We can observe that the simulated data approximate to the real data in shape while including high proportions for rare OTUs (Additional file 2: Figure S1). This indicates that the Dirichlet-multinomial model is useful to simulate microbiome data.

Type I error

Table 1 reports empirical type I error rates at the significance level of 5% for all the surveyed methods. We can observe correct type I error controls (i.e., the empirical type I error rates close to the significance level of 5%) for all the individual (i.e., $\mathrm{uHC}_{(h)}$'s and $\mathrm{wHC}_{(h)}$'s) and omnibus (i.e., uHC_A, wHC_A, and MiHC) higher criticism tests and the Simes test and also for all the other competing tests (i.e., aMiAD, aMiSPU, and OMiRKAT) (Table 1). Therefore, all the surveyed tests are valid in hypothesis testing.

Power

Here, we report the power comparisons in the order of (i) the comparison for the individual (i.e., $\mathrm{uHC}_{(h)}$'s and $\mathrm{wHC}_{(h)}$'s) and local omnibus (i.e., uHC_A and wHC_A) higher criticism tests and the Simes test (Fig. 1 (n = 50) and Additional file 3: Figure S2 (n = 100)); (ii) the comparison for the local omnibus (i.e., uHC_A and wHC_A) and global omnibus (i.e., MiHC) higher criticism tests (Fig. 2 (n = 50) and Additional file 4: Figure S3 (n = 100)); and (iii) the comparison for MiHC with the prior tests (i.e., Simes, HC, aMiAD, aMiSPU, and OMiRKAT) (Fig. 3 (n = 50) and Additional file 5: Figure S4 (n = 100)).

Table 1 Empirical type I error rates at the significance level of 5% for the individual (i.e., $\mathrm{uHC}_{(h)}$'s and $\mathrm{wHC}_{(h)}$'s for $h \in \{1, 3, 5, 7, 9\}$) and omnibus (i.e., uHC_A, wHC_A, and MiHC) higher criticism tests, the Simes test, and the other competing tests (i.e., aMiAD, aMiSPU, and OMiRKAT)

Category		Method	n=50	n=100
Individual HC tests	Unweighted tests	$\mathrm{uHC}_{(1)}$	0.051	0.049
		$\mathrm{uHC}_{(3)}$	0.050	0.052
		$\mathrm{uHC}_{(5)}$	0.049	0.049
		$\mathrm{uHC}_{(7)}$	0.050	0.050
		$\mathrm{uHC}_{(9)}$	0.050	0.051
	Weighted tests	$\mathrm{wHC}_{(1)}$	0.047	0.047
		$\mathrm{wHC}_{(3)}$	0.048	0.049
		$\mathrm{wHC}_{(5)}$	0.048	0.049
		$\mathrm{wHC}_{(7)}$	0.049	0.049
		$\mathrm{wHC}_{(9)}$	0.049	0.049
Omnibus HC tests	Local omnibus tests	uHC_A	0.049	0.049
		wHC_A	0.049	0.049
	Global omnibus tests	MiHC	0.050	0.052
Non-HC tests		Simes	0.048	0.049
		aMiAD	0.050	0.051
		aMiSPU	0.051	0.051
		OMiRKAT	0.050	0.050

The individual (i.e., $\mathrm{uHC}_{(h)}$ and $\mathrm{wHC}_{(h)}$) tests are more powerful using a smaller h value for higher sparsity levels, while they are more powerful using a larger h value for lower sparsity levels (Fig. 1 and Additional file 3: Figure S2), which is explained by the modulation scheme (Eq. 8 and Eq. 9). The Simes test is powerful at high sparsity levels but rapidly loses power as the sparsity level decreases (Fig. 1 and Additional file 3: Figure S2). The Simes test is more powerful at high sparsity levels (i.e., ≤ 2% or ≤ 4%) even than the original high criticism test (i.e., $\mathrm{uHC}_{(1)}$), but it is less powerful at low sparsity levels (i.e., ≥ 4% or ≥ 6%) than any individual higher criticism tests (Fig. 1 and Additional file 3: Figure S2). uHC_A closely approaches the most powerful test among the individual unweighted tests (i.e., $\mathrm{uHC}_{(h)}$'s) and the Simes test (Fig. 1a, c and Additional file 3: Figure S2:A,C), while wHC_A closely approaches the most powerful test among the individual weighted tests (i.e., $\mathrm{wHC}_{(h)}$'s) and the Simes test (Fig. 1b, d and Additional file 3: Figure S2:B,D), which is explained by the adaptivity of the minimum p value statistic (Eq. 12 and Eq. 13). The unweighted tests (i.e., $\mathrm{uHC}_{(h)}$'s and uHC_A) are more powerful than the weighted tests (i.e., $\mathrm{wHC}_{(h)}$'s and wHC_A) when randomly selected OTUs are associated with the host phenotype (Fig. 1a > Fig. 1b, Additional file 3: Figure S2A > Additional file 3: Figure S2B, and Fig. 2a), while the weighted tests are more powerful than the unweighted tests when phylogenetically relevant OTUs are associated with the host phenotype (Fig. 1c < Fig. 1d, Additional file 3: Figure S2C < Additional file 3: Figure S2D, and Fig. 2b), which is explained by the weighting scheme for phylogenetic relevance (Eq. 7). In addition, the unweighted tests (i.e., $\mathrm{uHC}_{(h)}$'s and uHC_A) are almost equally powerful when either randomly selected OTUs (Fig. 1a) or phylogenetically relevant OTUs (Fig. 1c) are associated with the host phenotype (Fig. 1a ≅ Fig. 1c). This is because the unweighted tests do not utilize any phylogenetic information; hence, they treat either randomly selected OTUs or phylogenetically relevant OTUs all equally as randomly selected OTUs.

To facilitate easier comparison, Fig. 2 and Additional file 4: Figure S3 report estimated powers only for the local omnibus (i.e., uHC_A and wHC_A) and global omnibus (i.e., MiHC) higher criticism tests. Here again, uHC_A is more powerful than wHC_A when randomly selected OTUs are associated with the host phenotype (Fig. 2a and Additional file 4: Figure S3A), while wHC_A is more powerful than uHC_A when phylogenetically relevant OTUs are associated with the host phenotype (Fig. 2b and Additional file 4: Figure S3B), which is explained by the weighting scheme for phylogenetic relevance (Eq. 7). Importantly, we can observe that MiHC closely approaches the most powerful test between uHC_A and wHC_A (Fig. 2 and Additional file 4: Figure S3), which is

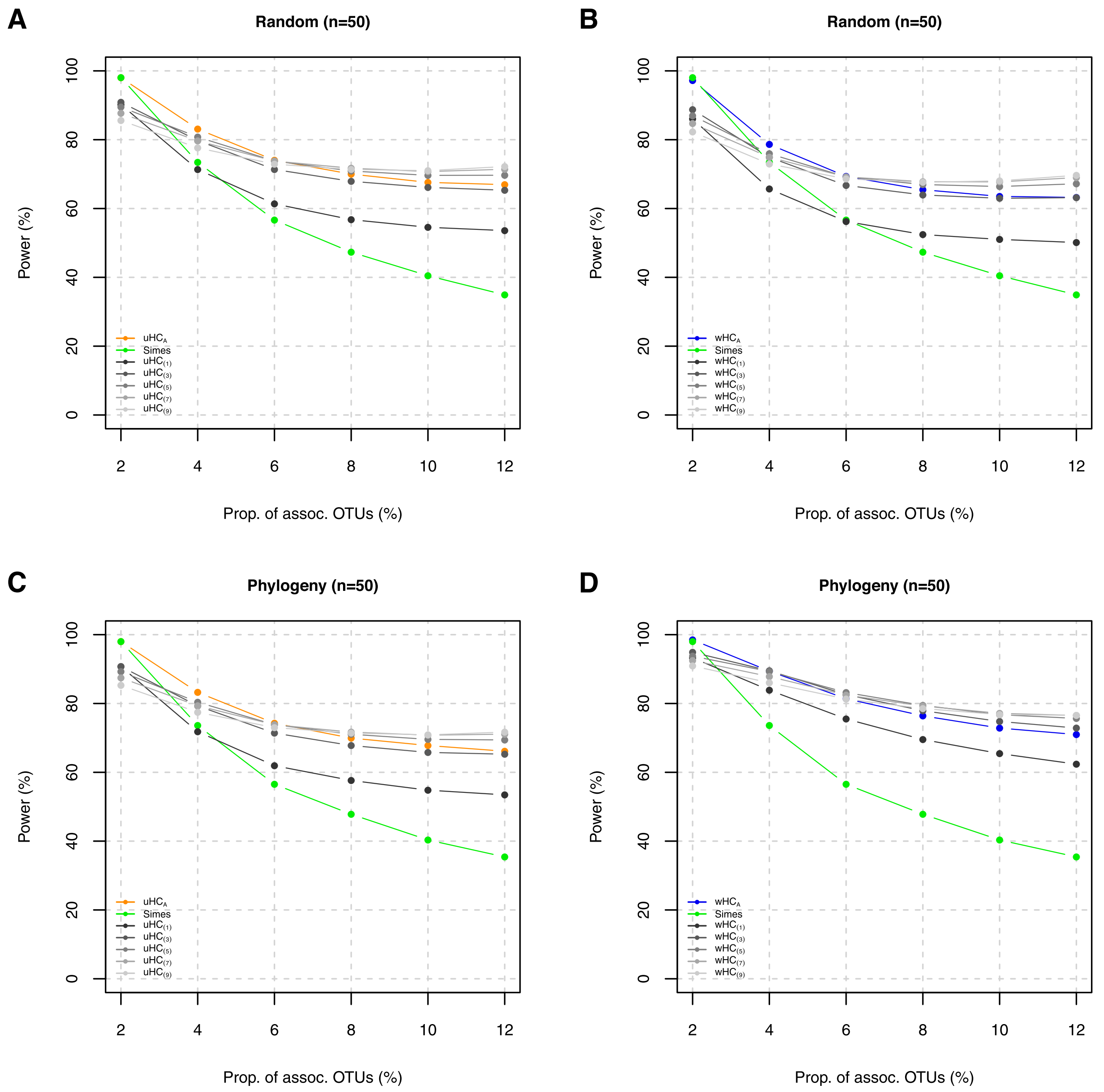

Fig. 1 Power estimates for the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s for $h \in \{1, 3, 5, 7, 9\}$) and local omnibus (uHC$_A$ and wHC$_A$) higher criticism tests (n = 50) (unit, %). **a** uHC$_{(h)}$'s and uHC$_A$ for the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}). **b** wHC$_{(h)}$'s and wHC$_A$ for the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs}). **c** uHC$_{(h)}$'s and uHC$_A$ for the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}). **d** wHC$_{(h)}$'s and wHC$_A$ for the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% 12% phylogenetically close OTUs})

explained by the adaptivity of the minimum p value statistic in the entirety (Eq. 11). This indicates that MiHC maintains a high power throughout different phylogenetic relevance and sparsity levels, while the individual or the local omnibus tests are limitedly powerful only for some specific phylogenetic relevance and sparsity levels (Figs. 1 and 2 and Additional files 3 and 4: Figure S2-S3). Thus, we suggest to use MiHC especially in respond to the unknown phylogenetic relevance and sparsity levels in practice.

Here, we also compare MiHC with the prior tests, Simes, HC, aMiAD, aMiSPU, and OMiRKAT. MiHC, Simes, and HC are powerful for high sparsity levels, while they lose power gradually for lower sparsity levels (Fig. 3 and Additional file 5: Figure S4). However, the power decay is slower for MiHC than Simes and HC (Fig. 3 and Additional

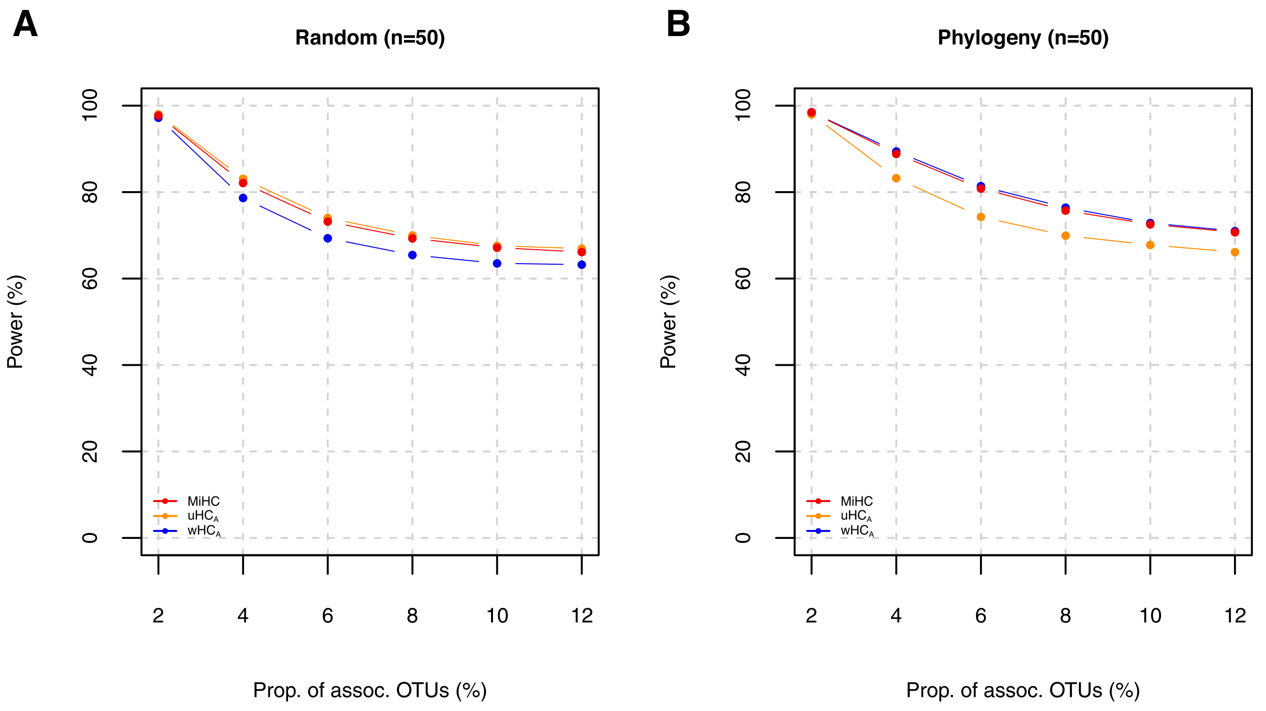

Fig. 2 Power estimates for the omnibus (uHC$_A$, wHC$_A$, and MiHC) higher criticism tests (n = 50) (unit, %). **a** For the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}. **b** For the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs})

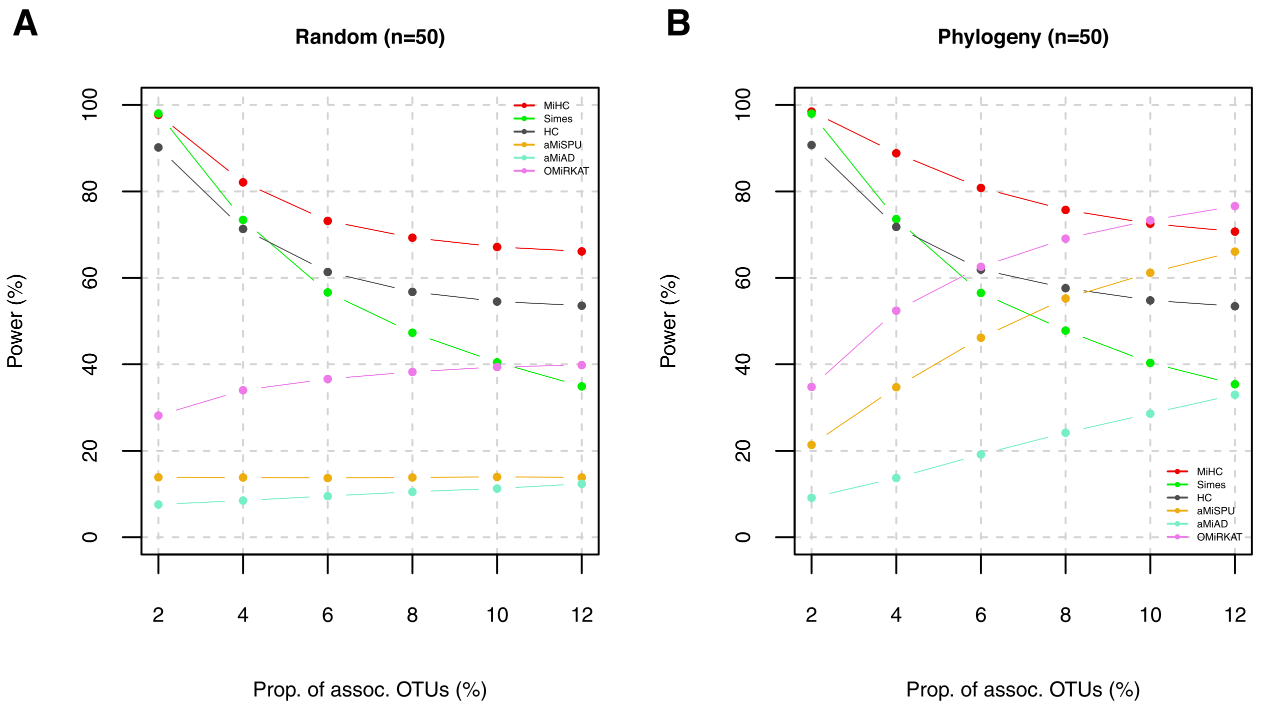

Fig. 3 Power estimates for MiHC compared with the prior tests, HC, aMiAD, aMiSPU, and OMiRKAT (n = 50) (unit, %). **a** For the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}. **b** For the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs})

file 5: Figure S4), which is explained by the modulation scheme (Eq. 8 and Eq. 9) and the adaptivity of the minimum p value statistic (Eq. 11). We can also observe that the power gap from MiHC to Simes or HC is larger when phylogenetically relevant OTUs are associated with the host phenotype (Fig. 3 and Additional file 5: Figure S4), which is explained by the weighting scheme for phylogenetic relevance (Eq. 7) and the adaptivity of the minimum p value statistic (Eq. 11). Therefore, MiHC better suits the microbiome association studies with multifarious phylogenetic relevance and sparsity levels. On the contrary, aMiAD, aMiSPU, and OMiRKAT are underpowered for high sparsity levels, yet they gain power gradually for lower sparsity levels (Fig. 3 and Additional file 5: Figure S4). This is because they amplify the overall group association by aggregating underlying microbial association signals in the sense of requiring as many association signals as possible. Especially, OMiRKAT is most powerful when phylogenetically relevant OTUs are associated with the host phenotype at low sparsity levels (i.e., ≥ 10%) (Fig. 3b), and we do not discourage the use of aMiAD, aMiSPU, and OMiRKAT for lower sparsity levels. MiHC is more powerful than aMiAD, aMiSPU, and OMiRKAT for many sparsity levels in our simulations (Fig. 3 and Additional file 5: Figure S4). We developed MiHC, from a different perspective, for the powerful discovery from high to low sparsity levels, which was especially challenging by the prior tests.

Real data applications

The association between the respiratory tract microbiome and smoking status

Charlson et al. have collected swab samples from the upper respiratory tract to survey the effect of cigarette smoking on the respiratory tract microbiome [28]. The microbiome data for the OTU abundance table and phylogenetic tree are publicly available in the R package, *GUniFrac* [42], where the raw sequence data had been processed using the QIIME pipeline [2] by targeting the V1–2 region of the 16S ribosomal RNA (rRNA) gene (refer to [28] for more detailed sampling/data processing procedures) and the phylogenetic tree had been constructed by using FastTree [43, 44]. The microbiome data include 273 OTUs with mean relative abundance ≥ 10^{-4} for 60 samples (28 smokers and 32 non-smokers). Here, we test the association between respiratory tract microbial composition and smoking status while adjusting for gender and antibiotic use within the last 3 months.

We found the significant association between respiratory tract microbial composition and smoking status throughout all the individual and omnibus higher criticism tests and the Simes test (Fig. 4a). We can also observe only a small difference between the unweighted (i.e., $\text{uHC}_{(h)}$'s and uHC_{A}) and weighted (i.e., $\text{wHC}_{(h)}$'s and wHC_{A}) tests (Fig. 4a), indicating that the OTUs associated with smoking status might have only mild phylogenetic relevance. We can also confirm it in visualization with similar graphical patterns between uHC and wHC (Fig. 4a). We also report the 10 most influential OTUs with respect to uHC and wHC, respectively (Fig. 4a). For the other competing methods, aMiAD and OMiRKAT find the significant association while aMiSPU does not (Table 2 A). MiHC finds the significant association (p value, 0.017) (Table 2 A and Fig. 4a).

The association between the infant's gut microbiome and delivery mode

Bokulich et al. have conducted a microbiome profiling study to survey the effect of the early life factors (e.g., delivery mode, infant nutrition, antibiotic use) on the infant's gut microbiome [29]. As a demonstration, we test the association between the infant's gut microbiome and delivery mode (i.e., vaginal or cesarean birth) while adjusting for gender and predominant diet (breastfeeding vs. formula). The microbiome data include 310 OTUs with mean relative abundance ≥ 10^{-4} for 32 infants (11 infants by cesarean delivery and 21 infants by vaginal delivery) [29, 37], where the raw sequence data had been processed using the QIIME pipeline [2] by targeting the V4 region of the 16S rRNA gene (refer to [29, 37] for more detailed sampling/data processing procedures) and the phylogenetic tree had been constructed by using FastTree [43, 44].

All the weighted tests (i.e., $\text{wHC}_{(h)}$'s and wHC_{A}) except for $\text{wHC}_{(1)}$ found the significant association between the infant's gut microbiome and delivery mode, while none of the unweighted tests (i.e., $\text{uHC}_{(h)}$'s and uHC_{A}) and the Simes test found it (Fig. 4b), indicating that the OTUs associated with delivery mode might have strong phylogenetic relevance. We can also confirm it in visualization with larger deviations between the expected and observed quantiles for wHC than uHC (Fig. 4b). We can also observe that the individual tests (i.e., $\text{uHC}_{(h)}$ and $\text{wHC}_{(h)}$) using a larger h value find smaller p values (Fig. 4b), indicating that many OTUs might be associated with delivery mode (i.e., low sparsity). We can also confirm it in visualization that many OTUs have some deviations between the expected and observed quantiles (Fig. 4b). We also report the 10 most influential OTUs with respect to uHC and wHC, respectively (Fig. 4b). For the other competing methods, OMiRKAT finds the significant association while aMiAD and aMiSPU do not (Table 2 B). MiHC finds the significant association (p value, 0.044) (Table 2 B and Fig. 4b).

The association between the gut microbiome and T1D status

Livanos et al. have conducted a microbiome profiling study to survey the roles of the gut microbiome on T1D onset through mouse experiments [30]. As a demonstration, we test the association between gut microbial composition and T1D status. For this, 19 mice were exposed to therapeutic-dose pulsed antibiotic (PAT)

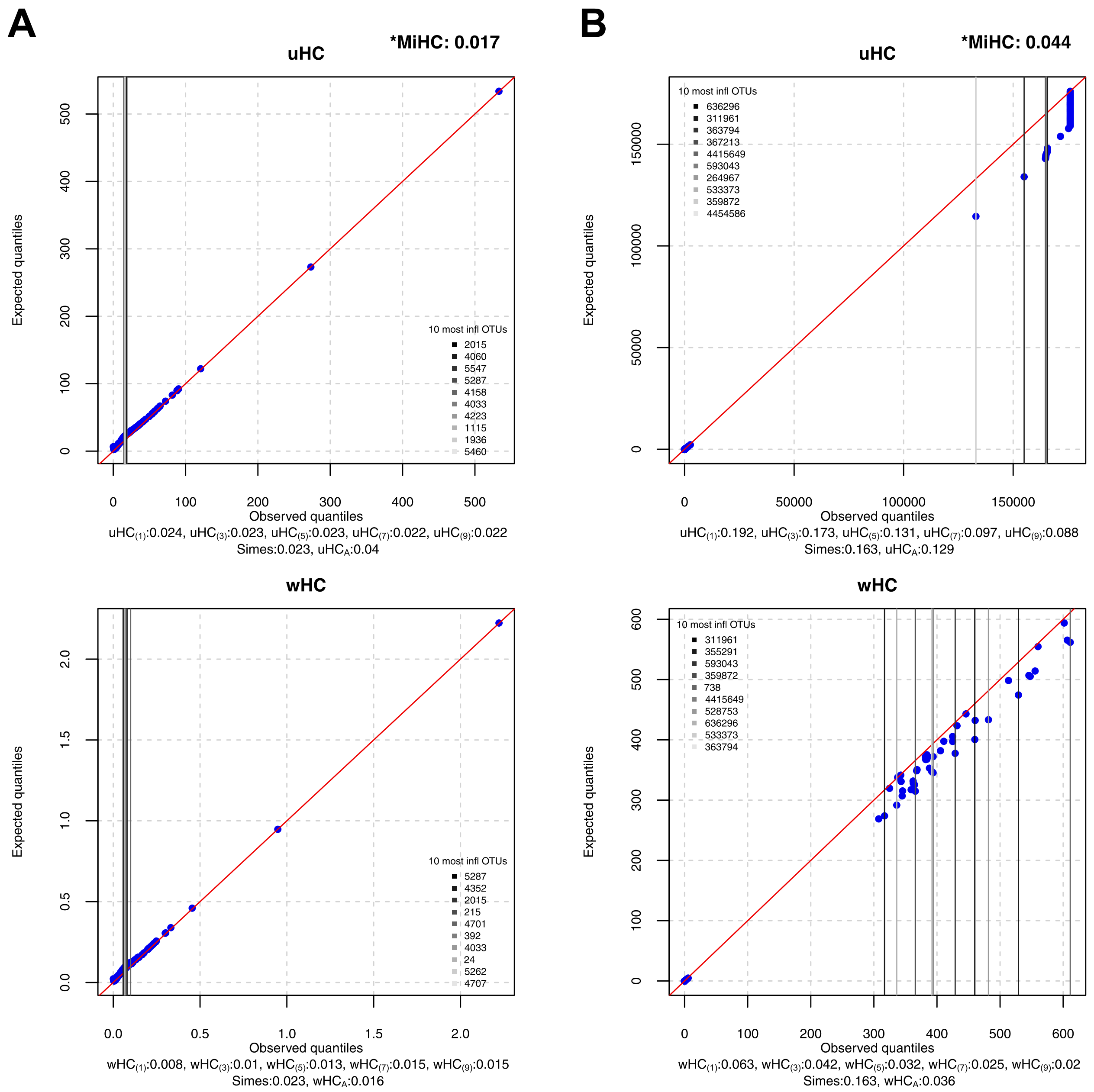

Fig. 4 The Q-Q plots between the expected and observed quantiles for uHC and wHC, respectively. **a** The association between the respiratory tract microbiome and smoking status. **b** The association between the infant's gut microbiome and delivery mode. Blue dots represent individual OTUs and a red diagonal line represents no influential points. Darker to lighter vertical lines represent more to less influential OTUs in rank order among the 10 most influential OTUs that correspond to the 10 largest deviations from the red diagonal line. The asterisk represents the *p* values for all the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s for $h \in \{1, 3, 5, 7, 9\}$) and omnibus (i.e., uHC$_{A}$, wHC$_{A}$, and MiHC) higher criticism tests and the Simes test

treatment at 6 weeks of age and then followed up for 30 weeks. The microbiome data include 120 OTUs with mean relative abundance ≥ 10^{-4} for the 19 mice at 30 weeks of the follow-up (9 T1D-free mice and 10 T1D-onset mice), where the raw sequence data had been processed using the QIIME pipeline [2] by targeting the V4 region of the 16S rRNA gene (refer to [30] for more detailed sampling/data processing procedures) and the phylogenetic tree had been constructed by using FastTree [43, 44].

We found the significant association between gut microbial composition and T1D status throughout all the individual and omnibus higher criticism tests but not through the Simes test (Fig. 5a). We can also observe

Table 2 The p values for the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s for $h \in \{1, 3, 5, 7, 9\}$) and omnibus (i.e., uHC$_A$, wHC$_A$, and MiHC) higher criticism tests, the Simes test, and the other competing tests (i.e., aMiAD, aMiSPU, and OMiRKAT). (A) The association between respiratory tract microbiome and smoking status. (B) The association between the infant's gut microbiome and delivery mode. (C) The association between the gut microbiome and type 1 diabetes status. (D) The association between the gut microbiome and human immunodeficiency virus status. * represents significant p values

Category		Method	A	B	C	D
Individual HC tests	Unweighted tests	uHC$_{(1)}$	**0.024***	0.192	**0.025***	**0.009***
		uHC$_{(3)}$	**0.023***	0.173	**0.024***	**0.008***
		uHC$_{(5)}$	**0.023***	0.131	**0.024***	**0.007***
		uHC$_{(7)}$	**0.022***	0.097	**0.024***	**0.007***
		uHC$_{(9)}$	**0.022***	0.088	**0.024***	**0.007***
	Weighted tests	wHC$_{(1)}$	**0.008***	0.063	**0.026***	**0.009***
		wHC$_{(3)}$	**0.010***	**0.042***	**0.024***	**0.008***
		wHC$_{(5)}$	**0.013***	**0.032***	**0.023***	**0.007***
		wHC$_{(7)}$	**0.015***	**0.025***	**0.023***	**0.007***
		wHC$_{(9)}$	**0.015***	**0.020***	**0.024***	**0.007***
Omnibus HC tests	Local omnibus tests	uHC$_A$	**0.040***	0.129	**0.029***	**0.013***
		wHC$_A$	**0.016***	**0.036***	**0.029***	**0.012***
	Global omnibus tests	MiHC	**0.017***	**0.044***	**0.030***	**0.012***
Non-HC tests		Simes	**0.023***	0.163	0.191	0.170
		aMiAD	**0.035***	0.449	0.346	**0.012***
		aMiSPU	0.063	0.170	**0.017***	0.181
		OMiRKAT	**0.005***	**0.001***	**0.016***	0.150

only a small difference between the unweighted (i.e., uHC$_{(h)}$'s and uHC$_A$) and weighted (i.e., wHC$_{(h)}$'s and wHC$_A$) tests (Fig. 5a). This might indicate that the OTUs associated with T1D status have only mild phylogenetic relevance. We also report the 10 most influential OTUs with respect to uHC and wHC, respectively (Fig. 5a). For the other competing methods, aMiSPU and OMiRKAT find the significant association while aMiAD does not (Table 2 C). MiHC finds the significant association (p value, 0.03) (Table 2 C and Fig. 5a).

The association between the gut microbiome and HIV status

Pinto-Cardoso et al. have conducted a microbiome profiling study to survey the effect of antiretroviral therapy (ART) on the gut microbiome of HIV-positive individuals [31]. As a demonstration, we test the association between gut microbial composition and HIV status while adjusting for age. For this, 33 HIV-infected individuals on ART and 10 HIV-uninfected individuals from Mexico were included in the analysis [31]. The microbiome data include 422 OTUs with mean relative abundance ≥ 10^{-4} for the 44 individuals, where the raw sequence data had been processed using the Resphera Insight [45] by targeting the V3 and V4 regions of the 16S rRNA gene (refer to [31, 46] for more detailed sampling/data processing procedures) and the phylogenetic tree had been constructed by using PyNAST [47].

We found the significant association between gut microbial composition and HIV status throughout all the individual and omnibus higher criticism tests, but not through the Simes test (Fig. 5b). We can also observe only a small difference between the unweighted (i.e., uHC$_{(h)}$'s and uHC$_A$) and weighted (i.e., wHC$_{(h)}$'s and wHC$_A$) tests (Fig. 5b), indicating that the OTUs associated with T1D status might have only mild phylogenetic relevance. We can also confirm it in visualization with similar graphical patterns between uHC and wHC (Fig. 5b). We also report the 10 most influential OTUs with respect to uHC and wHC, respectively (Fig. 5b). For the other competing methods, aMiAD finds the significant association while aMiSPU and OMiRKAT do not (Table 2 D). MiHC finds the significant association (p value, 0.012) (Table 2 D and Fig. 5b).

Discussion and conclusions

In this paper, we introduced a data-driven omnibus test, MiHC, to evaluate the association between microbial group (e.g., community or clade) composition and a host phenotype of interest. Our simulations demonstrated that MiHC robustly maintains a high power for both random and phylogenetic association patterns at different high sparsity levels with correct type I error controls.

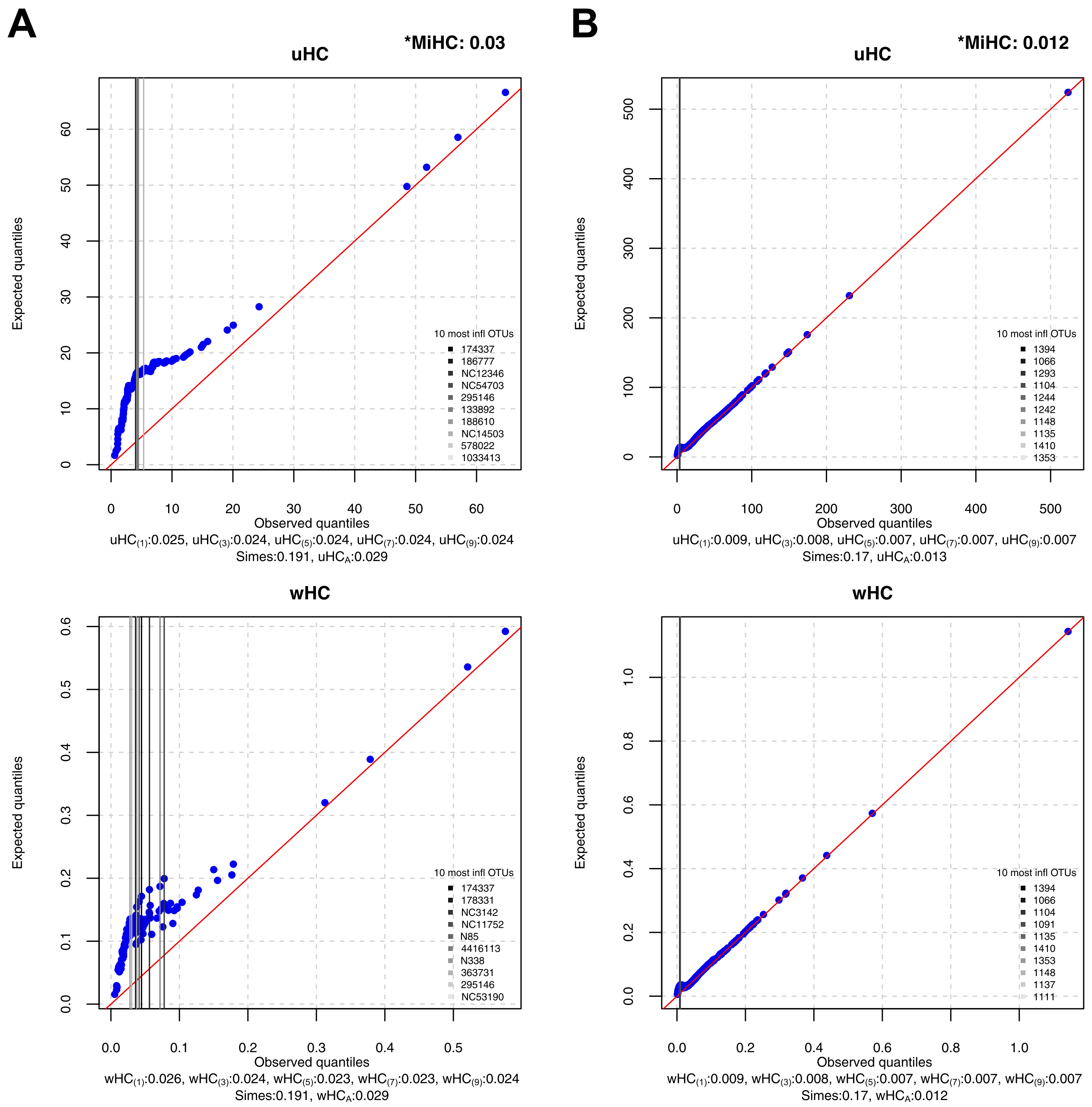

Fig. 5 The Q-Q plots between the expected and observed quantiles for uHC and wHC, respectively. **a** The association between the gut microbiome and type 1 diabetes status. **b** The association between the gut microbiome and human immunodeficiency virus status. Blue dots represent individual OTUs and a red diagonal line represents no influential points. Darker to lighter vertical lines represent more to less influential OTUs in rank order among the 10 most influential OTUs that correspond to the 10 largest deviations from the red diagonal line. The asterisk represents the *p* values for all the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s for $h \in \{1, 3, 5, 7, 9\}$) and omnibus (i.e., uHC$_{A}$, wHC$_{A}$, and MiHC) higher criticism tests and the Simes test

We also applied MiHC to four different real microbiome datasets and observed that MiHC finds stably low *p* values while the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s) higher criticism tests and the Simes test find differing *p* values depending on the underlying phylogenetic relevance and sparsity levels. Thus, MiHC is a useful analytic tool in practice because of the unknown phylogenetic relevance and sparsity levels.

We considered the optimal number of clusters which maximizes the average silhouette width searching up to 30 clusters and the candidate set of $\Gamma = \{1, 3, 5, 7, 9\}$, instead of the union set of $\Gamma = \{1, 2,..., m\text{-}1, m\}$, for the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s) higher criticism tests to avoid the exhaustive search and huge computation. However, any other upper limit to fine the optimal number of clusters and any other smaller or larger candidate set can alternatively be considered by the researcher's choice through the options in our software package. For example, you may believe that the candidate set of $\Gamma = \{1, 3, 5, 7, 9\}$ is too much tailored to high sparsity levels; hence, you can include larger values in the candidate set for lower sparsity levels. Moreover, a number of microbiome data normalization procedures have been proposed [48], but there is no consensus on which procedure is the best and such debate is beyond the scope of this paper. We did not survey any further normalization procedure except for using relative abundances (i.e., proportions), instead of absolute abundances (i.e., read counts), to control differing total read counts per sample. However, MiHC is compatible with any other normalization procedure (e.g., centered log-ratio transformation [49]), which can be considered by the researcher's choice. We set up all the implementation procedures described in this paper as a default in our software package, yet we do not strictly force to use it. Instead, we give researches some user options in our software package to make the best use of it.

We developed MiHC based on the generalized linear models to handle exponential family responses with the linear predictor [32]. However, its application can be much broader, and, for example, the potential extensions to survival [38, 50], longitudinal [39, 51], or mediation [52] analysis need to be further studied.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40168-020-00834-9.

Additional file 1: Table S1. The summary of the notations.

Additional file 2: Figure S1. The histograms of the relative abundances (%). A. The real respiratory-track microbiome data. B. The simulated data based on the Dirichlet-multinomial model (n=50). C. The simulated data based on the Dirichlet-multinomial model (n=100).

Additional file 3: Figure S2. Power estimates for the individual (i.e., uHC$_{(h)}$'s and wHC$_{(h)}$'s for $h \in \{1, 3, 5, 7, 9\}$) and local omnibus (uHC$_A$ and wHC$_A$) higher criticism tests (n=100) (Unit: %). A. uHC$_{(h)}$'s and uHC$_A$ for the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}). B. wHC$_{(h)}$'s and wHC$_A$ for the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs}). C. uHC$_{(h)}$'s and uHC$_A$ for the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}). D. wHC$_{(h)}$'s and wHC$_A$ for the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs}).

Additional file 4: Figure S3. Power estimates for the omnibus (uHC$_A$, wHC$_A$ and MiHC) higher criticism tests (n=100) (Unit: %). A. For the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}. B. For the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs}).

Additional file 5: Figure S4. Power estimates for MiHC compared with the prior tests, HC, aMiAD, aMiSPU and OMiRKAT (n=100) (Unit: %). A. For the randomly selected OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% random OTUs}. B. For the phylogenetically relevant OTUs (i.e., Λ = {2%, 4%, 6%, 8%, 10% or 12% phylogenetically close OTUs}).

Abbreviations

aMiAD: Adaptive Microbiome α-diversity-based association test; aMiSPU: Adaptive microbiome sum of powered score tests; ART: Antiretroviral therapy; HC: Higher criticism; HIV: Human immunodeficiency virus; MiHC: Microbiome higher criticism analysis; MiRKAT: Microbiome regression-based kernel association test; MiSPU: Microbiome sum of powered score tests; OMiRKAT: Optimal microbiome regression-based kernel association test; OTU: Operational taxonomic unit; RNA: Ribosomal RNA; T1D: Type I diabetes; uHC: Unweighted higher criticism; wHC: Weighted higher criticism

Acknowledgements

The authors are grateful to the reviewers for their insightful observations and comments.

Authors' contributions

HK contributed to the methodological ideas, performed the simulations and real data analyses, developed the software package, and wrote the manuscript. NZ contributed to the methodological ideas, simulations, and real data analyses and wrote the manuscript. The authors read and approved the final manuscript.

Funding

This study was supported in part by NIH for the Environmental Influences of Child Health Outcomes (ECHO) Data Analysis Center (U24OD023382) and Johns Hopkins University Center for AIDS Research (1P30AI094189).

Availability of data and materials

We used four public microbiome datasets for (1) the association between respiratory tract microbiome and smoking status (available in the R package, *GUniFrac*, with three data objects, throat.otu.table, throat.tree, and throat.meta, https://cran.r-project.org/web/packages/GUniFrac/index.html), (2) the association between the infant's gut microbiome and delivery mode (available in the QIITA repository under accession code 10249, https://qiita.ucsd.edu), (3) the association between the gut microbiome and T1D status (available in the European Bioinformatics Institute (EBI) database under accession code ERP016357, https://www.ebi.ac.uk and in the Qiita database under accession code: 10508, https://qiita.ucsd.edu), and (4) The association between the gut microbiome and HIV status (available in the European Bioinformatics Institute (EBI) database under accession code PRJNA344791, https://www.ebi.ac.uk).
Our propose method can be implemented using the R package, MiHC, which is freely available at https://github.com/hk1785/MiHC. The detailed manual on the inputs, outputs, arguments, and options can also be found in the software webpage.

Ethics approval and consent to participate

All utilized microbiome datasets are publicly available. No ethics approval or consent to participate was required for this study.

Consent for publication

All utilized microbiome datasets are publicly available. No consent for publication was required for this study.

Competing interests

The authors declare that they have no competing interests.

Received: 18 June 2019 Accepted: 23 March 2020
Published online: 11 May 2020

References

1. Hamady M, Knight R. Microbial community profiling for human microbiome projects: tools, techniques. Genome Res. 2009;19(7):1141–52.
2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez Peña A, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.

3. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012;2:3.
4. Jovel J, Patterson J, Wang W, Hotte N, O'keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol. 2016;7:459.
5. Hill MO. Diversity and evenness: a unifying notation and its consequences. Ecology. 1973;54(2):427–32.
6. Tuomisto H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography. 2010;33(1):2–22.
7. Zhang Z, Li J, Krautkramer KA, Badri M, Battaglia T, Borbet TC, Koh H, Ng S, Sibley RA, Li Y, et al. Antibiotic-induced acceleration of type 1 diabetes alters maturation of innate intestinal immunity. eLife. 2018;7:e37816.
8. Liu M, Koh H, Kurtz ZD, Battaglia T, PeBenito A, Li H, Nazzal L, Blaser MJ. Oxalobacter formigenes-associated host features and microbial community structures examined using the American Gut Project. Microbiome. 2017;5:108.
9. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160(3):447–60.
10. Wang W, Jovel J, Halloran B, Wine E, Patterson J, Ford G, O'Keefe S, Meng B, Song D, Zhang Y, et al. Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria. Inflamm Bowel Dis. 2015;21(6):1419–27.
11. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12:R60.
12. Sohn MB, Du R, An L. A robust approach for identifying differentially abundant features in metagenomic sample. Bioinformatics. 2015;31(14):2269–75.
13. Koh H. An adaptive microbiome α-diversity-based association analysis method. Sci Rep. 2018;8:18026.
14. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet. 2015;96(5):797–807.
15. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26(1):32–46.
16. McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001;82(1):290–7.
17. Tang Z, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. Bioinformatics. 2016;32(17):2618–25.
18. Wu C, Chen J, Kim J, Pan W. An adaptive association test for microbiome data. Genome Med. 2016;8:56.
19. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
20. Kaper JB, Nataro JP, Mobley HL. Pathogenic Escherichia coli. Nat Rev Microbiol. 2004;2(2):123–40.
21. Bäckhed F, Manchester JK, Semenkovich CF, Gordon JI. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U.S.A. 2007;104(3):979–84.
22. Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. Ann Stat. 2004;32(3):962–94.
23. Barnett IJ, Lin X. Analytical p-value calculation for the higher criticism test in finite-d problems. Biometrika. 2014;101(4):964–70.
24. Barnett I, Mukherjee R, Lin X. The generalized higher criticism for testing SNP-set effects in genetic association studies. J Am Stat Assoc. 2017;112(517):64–76.
25. Martins EP, Hansen TF. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat. 1997;149(4):646–67.
26. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986;73(3):751–4.
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U.S.A. 2005;102(43):15545–50.
28. Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, Hwang J, Bushman FD, Collman RG. Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS One. 2010;5:12.
29. Bokulich NA, Chung J, Battagila T, Henderson N, Jay M, Li H, Lieber AD, Wu C, Perez-Perez GI, Chen Y, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med. 2016;8:343.
30. Livanos AE, Greiner TU, Vangay P, Pathmasiri W, Stewart D, McRitchie S, Li H, Chung J, Sohn J, Kim S, et al. Antibiotic-mediated gut microbiome perturbation accelerates development of type 1 diabetes in mice. Nat Microbiol. 2016;1:6140.
31. Pinto-Cardoso S, Lozupone C, Briceño O, Alva-Hernández S, Téllez N, Adriana A, Murakami-Ogasawara A, Reyes-Terán G. Fecal bacterial communities in treated HIV infected individuals on two antiretroviral regimens. Sci Rep. 2016;7:43741.
32. Agresti A. Foundations of linear and generalized linear models. Hoboken: Wiley; 2015.
33. Hall P, Jin J. Innovated higher criticism for detecting sparse signals in correlated noise. Ann Stat. 2010;38(3):1686–732.
34. Arias-Castro E, Candès E, Plan Y. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. Ann Stat. 2011;39(5):2533–56.
35. Sneath PHA, Sokal RR, Freeman WH. Numerical taxonomy: the principles and practice of numerical classification. Syst Zool. 1975;24(2):263–8.
36. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algorithms. 2006;5(4):474–504.
37. Koh H, Blaser MJ, Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. Microbiome. 2017;5:45.
38. Koh H, Livanos AE, Blaser MJ, Li H. A highly adaptive microbiome-based association test for survival traits. BMC Genom. 2018;19:210.
39. Koh H, Li Y, Zhan X, Chen J, Zhao N. A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. Front Genet. 2019;458:10.
40. Mosimann JE. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika. 1962;49(1/2):65–82.
41. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.
42. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012;28(16):2106–13.
43. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641–50.
44. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:3.
45. Grim CJ, Daquigan N, Lusk Pfefer TS, Ottesen AR, White JR, Jarvis KG. High-resolution microbiome profiling for detection and tracking of Salmonella enterica. Front Microbiol. 2017;8:1587.
46. Tuddenham SA, WLA K, Zhao N, White JR, Ghanem KG, Sears CL. HIV Microbiome Re-analysis Consortium. The impact of human immunodeficiency virus infection on gut microbiota α-diversity: an individual-level meta-analysis. Clin Infect Dis. 2019;ciz258.
47. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Anderson GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2020;26(2):266–7.
48. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázques-Baeza Y, Birmingham A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5:27.
49. Aitchison J. The statistical analysis of compositional data. J R Stat Soc B. 1982;44(2):139–77.
50. Plantinga A, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC. MiRKAT-S: a community-level test of association between the microbiota and survival times. Microbiome. 2017;5:17.
51. Zhan X, Xue L, Zheng H, Plantinga A, Wu MC, Schaid DJ, Zhao N, Chen J. A small-sample kernel association test for correlated data with application to microbiome association studies. Genet Epidemiol. 2018;42(8):772–82.
52. Sohn M, Li H. Compositional mediation analysis for microbiome studies. Ann Appl Stat. 2019;13(1):661–81.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.