Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detecting Sparse Microbial Association Signals Adaptively From Longitudinal Microbiome Data Based on Generalized Estimating Equations

Han Sun
Central China Normal University
Xiaoyun Huang
Central China Normal University
Ban Huo
Central China Normal University
Yuting Tan
Central China Normal University
Tingting He
Central China Normal University
Xingpeng Jiang (xpjiang@mail.ccnu.edu.cn)
Central China Normal University https://orcid.org/0000-0002-8848-9300

Methodology

Keywords: Microbiome-based association test, Longitudinal microbiome data, Higher criticism, Generalized estimating equations, Sparse microbial association signals, Phylogenetic information

Posted Date: October 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1002100/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

RESEARCH

Detecting sparse microbial association signals adaptively from longitudinal microbiome data based on generalized estimating equations

Han Sun[1,2], Xiaoyun Huang[2,3], Ban Huo[2,4], Yuting Tan[1,2], Tingting He[2,4,5] and Xingpeng Jiang[2,4,5*]

*Correspondence:
xpjiang@mail.ccnu.edu.cn
[1]School of Mathematics and Statistics, Central China Normal University, Wuhan, China
Full list of author information is available at the end of the article

Abstract

Background: The relationship between the compositions of microbial communities and various host phenotypes is an important research topic. Microbiome association research addresses multiple domains, such as human disease, diet and medicine. Statistical methods for testing microbiome-phenotype associations have been studied recently to determine their ability to assess longitudinal microbiome data. However, existing methods fail to detect sparse association signals in longitudinal microbiome data.

Methods: In this paper, we developed a novel method, namely, aGEEMiHC, which is a data-driven adaptive microbiome higher criticism analysis based on generalized estimating equations, to detect sparse microbial association signals from longitudinal microbiome data. aGEEMiHC adopts a generalized estimating equations framework that fully considers the correlation among different observations from the same cluster (individuals) in longitudinal data, and it integrates multiple microbiome higher criticism analyses based on generalized estimating equations by setting different working correlation structures. Thus, the proposed method is robust to diverse correlation structures for longitudinal data.

Results: The proposed method shows a stable performance for diverse association patterns in both sparsity levels and phylogenetic relevance. Extensive simulation experiments demonstrate that it can control the type I error correctly and achieve superior performance according to a statistical power comparison. In our simulation, we applied aGEEMiHC to longitudinal microbiome data with various types of host phenotypes to demonstrate the stability of our method. aGEEMiHC is also utilized for real longitudinal microbiome data, and we found a significant association between the gut microbiome and Crohn's disease.

Conclusions: aGEEMiHC is a statistical method that facilitates association testing for sparse microbial association signals from longitudinal microbiome data, and it can be applied to situations in which the true underlying correlations among different observations from the same cluster in longitudinal data are unknown. It is worth noting that our method also ranks the significant factors associated with the host phenotype to provide potential biomarkers. The R package **GEEMiHC** is available at https://github.com/xpjiang-ccnu/GEEMiHC.

Keywords: Microbiome-based association test; Longitudinal microbiome data; Higher criticism; Generalized estimating equations; Sparse microbial association signals; Phylogenetic information

Background

Microbial research has demonstrated that microbes have an important influence on human health [1]. The dysbiosis of microbial ecology can promote the incidence of diseases [2],

such as inflammatory bowel diseases [3], obesity [4] and rheumatoid arthritis [5]. An imbalance of the gut microbiome may also promote the production of chronic inflammation and carcinogenic metabolites, thus leading to the formation of tumors [6]. In addition, changes in the gut microbiome induced by genetics and the environment may increase the risk of neurodegenerative diseases and neurological disorders [7, 8]. Moreover, regulating the microbial community by microbiota-directed therapeutic foods [9] or probiotics [10] is beneficial to disease remission and even the eradication of certain diseases. These studies can provide a better understanding of the associations among the microbiome, disease, diet, etc.

16S rRNA gene amplicon sequencing is often used to study the characteristics of the microbiome and the influence of these characteristics on human diseases [11]. Genetic similarity sequences from the sequencing results are grouped and divided into operational taxonomic units (OTUs). Characteristic OTU information is used as a feature to annotate and analyze different species [12, 13]. The microbiome association test detects the association between the microbiome and a host phenotype by using feature abundance matrices. These matrices are often high-dimensional and phylogenetically relevant, and traditional methods [14] are not applicable for such scenarios (i.e., small n and large p).

Recently, many statistical methods have been developed to address this difficulty, and they are mainly summarized based on two categories according to their null hypothesis, namely, univariate or multivariate testing. For univariate testing, it is common to project higher-dimensional microbiome data onto a random effect based on the species diversity index, such as the within-sample diversity (i.e., α-diversity) and between-sample diversity (i.e., β-diversity index). Then, only the variance for random effects is the null hypothesis. More specifically, the optimal microbiome regression-based kernel association test (OMiRKAT) [15] adopts β-diversity via a kernel machine framework to project microbiome data. The adaptive microbiome α-diversity-based association test (aMiAD) [16] utilizes multiple candidate α-diversity metrics, including nonphylogenetic metrics (i.e., richness, Shannon, and Simpson) and phylogenetic metrics (i.e., phylogenetic diversity, phylogenetic entropy, and phylogenetic quadratic entropy). For multivariate testing, it is common to make a global null hypothesis for all effects of the features (OTUs) from microbiome data. For example, researchers have proposed the microbiome-based sum of powered score (MiSPU) [17] and optimal microbiome-based association test (OMiAT), which combines the sum of powered score tests (SPU) and OMiRKAT [18].

These methods apply to cases where a high percentage of OTUs are associated with the host phenotype (i.e., association signals with low sparsity levels) [19, 20]. They are less effective for association signals with high sparsity levels (i.e., only one or a small number of OTUs associated with the host phenotype) [21, 22]. Thus, microbiome higher criticism analysis (MiHC) is introduced based on the higher criticism (HC) test [23]. This multivariate testing method is a data-driven test that combines the unweighted HC test, weighted HC test and Simes test, and it is applied to different high sparsity levels and phylogenetic relevance levels. However, all of the above methods adopt a linear model or generalized linear model (GLM) framework, which has one assumption in common: the samples are independent. This assumption is invalid for longitudinal data in which an individual usually has several observations at different time points.

Longitudinal studies can provide a more profound description of microbial community interdependencies, periodic patterns and temporal variations [24] than cross-sectional studies.

Longitudinal data can also reduce potential unmeasured confounding effects by repeating measurements for each subject [25, 26]. As an extension of OMiRKAT [15] from cross-sectional analysis to longitudinal studies, the correlated sequence kernel association test (CSKAT) is proposed to address correlated data, including paired data and longitudinal data [27]. The disadvantage of CSKAT is that it can only handle the outcomes with a Gaussian distribution (e.g., body mass index). Subsequently, researchers further proposed a distance-based kernel association test based on the generalized linear mixed model (aGLMMMiRKAT) [28], which is adapted for data with outcomes that include Binomial or Poisson distributions.

Both CSKAT and aGLMMMiRKAT have less power in detecting sparse association signals from longitudinal microbiome data. To overcome this difficulty, we propose the aGEEMiHC method by integrating multiple generalized estimating equations (GEEs) with different working correlation structures [29] to conduct a higher criticism analysis [30]. Compared with MiHC [23], aGEEMiHC adopts multiple GEEs instead of a GLM and thus can explore different working correlation structures to accommodate various correlations among different observations for the same cluster. The proposed method regards each subject from longitudinal data as a cluster and fully considers the information for each cluster, and it can be applied to data with outcomes that include an exponential distribution, such as binary data and count data.

Through numerous simulation experiments, we find that the type I error of aGEEMiHC is always controlled close to the significance level of 5%. Compared with previous methods, aGEEMiHC has almost the best performance for different association patterns (i.e., sparsity levels and phylogenetic relevance) and longitudinal data with unknown correlations among different observations for the same cluster. Finally, we apply aGEEMiHC to diverse types of microbiome datasets to explain the stability of our methods.

Methods and materials

In this section, we introduce GEEMiHC and aGEEMiHC in detail, including the models for our method and the calculation of the p value. GEEMiHC adopts the GEE-based [29] higher criticism (HC) test [30] to detect sparse association signals from longitudinal microbiome data. aGEEMiHC, an optimal test, integrates three GEEMiHCs with different working correlation structures (Fig. 1).

Generalized estimating equations and marginal score statistics

We assume that the dataset has collected n subjects (i.e., clusters), and that each cluster contains response variables (i.e., outcomes), m OTUs and l covariates. In the experiment, the observations are repeated k_i times for each subject at different points in time. We assume that when the ith cluster is observed for the tth time, we obtain the response variable y_{it}, which represents the host phenotype of interest (e.g., health/disease-related factor), the abundance data of m OTUs $\mathbf{o}_{it} = (o_{it1}, o_{it2}, \ldots, o_{itm})^T$ and $(l+1)$-dimension vectors of covariates $\mathbf{X}_{it} = (1, x_{it1}, \ldots, x_{itl})^T$, where $i = 1, 2, \ldots, n$; $t = 1, 2, \ldots, k_i$ (please refer to the Input Data in Fig. 1). It is common to assume that the observations from different clusters are independent. However, different observations for the same clusters are considered to have certain correlations [31]. The true intracluster correlation is usually unknown [14].

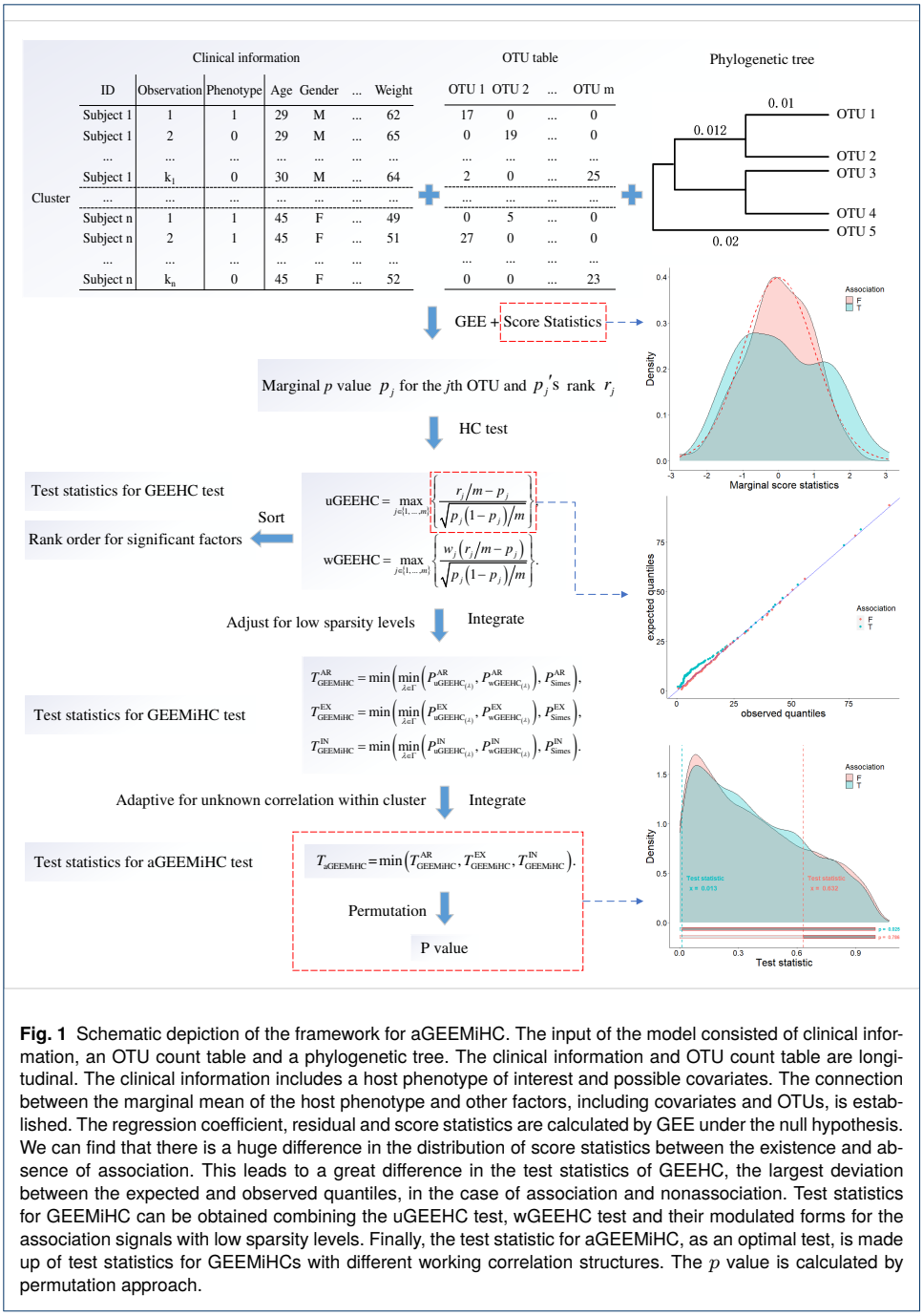

Fig. 1 Schematic depiction of the framework for aGEEMiHC. The input of the model consisted of clinical information, an OTU count table and a phylogenetic tree. The clinical information and OTU count table are longitudinal. The clinical information includes a host phenotype of interest and possible covariates. The connection between the marginal mean of the host phenotype and other factors, including covariates and OTUs, is established. The regression coefficient, residual and score statistics are calculated by GEE under the null hypothesis. We can find that there is a huge difference in the distribution of score statistics between the existence and absence of association. This leads to a great difference in the test statistics of GEEHC, the largest deviation between the expected and observed quantiles, in the case of association and nonassociation. Test statistics for GEEMiHC can be obtained combining the uGEEHC test, wGEEHC test and their modulated forms for the association signals with low sparsity levels. Finally, the test statistic for aGEEMiHC, as an optimal test, is made up of test statistics for GEEMiHCs with different working correlation structures. The p value is calculated by permutation approach.

Let $\mathrm{E}(y_{it}|\mathbf{X}_{it}, \mathbf{o}_{it}) = \mu_{it}$; we establish the connection between the marginal mean and other factors (i.e., covariates and OTUs) by specifying that

$$g(\mu_{it}) = \mathbf{X}_{it}^T\boldsymbol{\alpha} + \mathbf{o}_{it}^T\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_l)^T$ denotes the intercept and regression coefficient of $\mathbf{X}_{it}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T$ denotes the regression coefficient of $\mathbf{o}_{it}$. $g(\cdot)$ is a link function, and the variance of response variable y_{it} is $\mathrm{Var}(y_{it}|\mathbf{X}_{it}, \mathbf{o}_{it}) = \dot{g}^{-1}(\gamma_{it})$, where $\gamma_{it} = \mathbf{X}_{it}^T\boldsymbol{\alpha} + \mathbf{o}_{it}^T\boldsymbol{\beta}$.

Then, $\dot{g}^{-1}(\gamma_{it})$ is the first partial derivative of $g^{-1}(\gamma_{it})$ with respect to γ_{it}. For the binary response variable, the variance $\mathrm{Var}(y_{it}|\mathbf{X}_{it}, \mathbf{o}_{it})$ of response variable y_{it} is $\mu_{it}(1-\mu_{it})$. When $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, let $\mathbf{Y}_i = (y_{i1}, \ldots, y_{ik_i})^T$ and $\boldsymbol{\mu}_i(\boldsymbol{\theta}) = (\mu_{i1}(\boldsymbol{\theta}), \ldots, \mu_{ik_i}(\boldsymbol{\theta}))^T$ represent the vector of responses and expectation for all observations of the ith cluster, respectively. Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ik_i})^T$ denote the $k_i \times (l+1)$ matrix of covariates and $\mathbf{o}_i = (\mathbf{o}_{i1}, \mathbf{o}_{i2}, \ldots, \mathbf{o}_{ik_i})^T$ represent the $k_i \times m$ abundance matrix of OTUs. Furthermore, let $\mathbf{A}_i(\boldsymbol{\theta})$ represent an $k_i \times k_i$ diagonal matrix, with its tth diagonal element representing the variance for response variable y_{it}. Then, the estimator $\boldsymbol{\theta}$ is calculated by solving the following:

$$\begin{aligned}\mathbf{S}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} \begin{pmatrix} \mathbf{X}_i^T \\ \mathbf{o}_i^T \end{pmatrix} \mathbf{A}_i(\boldsymbol{\theta}) \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) = 0, \end{aligned} \tag{2}$$

where $\mathbf{V}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\theta}) \mathbf{R}(\boldsymbol{\tau}) \mathbf{A}_i^{1/2}(\boldsymbol{\theta})$ is the working covariance matrix and $\mathbf{R}(\boldsymbol{\tau})$ is the working correlation matrix, which can be understood as a weight matrix [32]. $\boldsymbol{\tau}$ is a disturbance parameter. The working covariance matrix $\mathbf{V}_i$ is proposed to solve the main problem of variance bias due to the correlation of longitudinal data in repeated measurements. We consider three frequently used working correlation structures: autoregressive (AR), exchange (EX) and independence (IN) [33], which are formulated as follows: $\mathbf{R}^{\mathrm{AR}} = \mathbf{M}_1 + \rho \mathbf{M}_2 + \ldots + \rho^{k-1} \mathbf{M}_k$, $\mathbf{R}^{\mathrm{EX}} = \mathbf{M}_1 + \rho \mathbf{M}_2 + \ldots + \rho \mathbf{M}_k$, and $\mathbf{R}^{\mathrm{IN}} = \mathbf{M}_1$, where ρ is the correlation coefficient [34]. As a symmetric matrix, k'th main diagonal elements of $\mathbf{M}_{k'}$ are 1 while the others are 0, where $k' = 1, \ldots, k$. GEE fully considers correlations among different observations for the same cluster by flexibly using the working correlation structure. The choice of these structures usually depends on different scenarios. When the IN structure is adopted, the GEE estimator is reduced to GLMs. In addition, it is worth noting that the estimators for GEE with different working correlation structures are consistent for cross-sectional data but different for longitudinal data.

Here, we are interested in examining whether there is an association between OTUs and the host phenotype. Therefore, the null hypothesis can be formulated as H_0: $\beta_j = 0$ for $\forall j \in \{1, 2, \ldots, m\}$, and its corresponding alternative is H_1: $\beta_j \neq 0$ for $\exists j \in \{1, 2, \ldots, m\}$. The GEE-based marginal score statistic is as follows:

$$Z_j^{\mathrm{GEE}} = \frac{\mathbf{o}_{\cdot j}^T (\mathbf{Y} - \hat{\boldsymbol{\mu}}_0)}{\sqrt{\mathbf{o}_{\cdot j}^T \mathbf{P} \mathbf{o}_{\cdot j}}}, \tag{3}$$

where $\mathbf{o}_{\cdot j} = (o_{11j}, o_{12j}, \ldots, o_{nk_nj})^T$ and $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)^T$. The vector of the expected values $\hat{\boldsymbol{\mu}}_0$ consists of the estimator of μ_{it} based on Eq. 1 under the null hypothesis, and $\mathbf{P} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$, where $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$ and $\mathbf{W}$ is the diagonal matrix with the jth diagonal element $\mathrm{Var}(y_{it}|\mathbf{X}_{it}, \mathbf{o}_{it})$. Then, Z_j^{GEE} is the direction and magnitude of the jth OTU, which is the standard normal distribution [35]. Because $\mathbf{P}(|Z_j^{\mathrm{GEE}}| > N(0,1))$, the marginal p value for the jth OTU p_j can be calculated.

Higher criticism analyses and rank order for significant factors

Donoho and Jin [30] proposed the higher criticism test, which was sensitive to sparse association signals, and this test was subsequently widely researched [36, 37] and applied [35]. In microbiome association studies, Koh and Zhao [23] introduced the weighted HC (wHC) test, whose weighted factor is based on the cophenetic distance of the phylogenetic tree:

$$w_j = \frac{\sum_{j' \in \zeta(j)/\{j\}} \frac{1}{D_{j,j'}} |Z_{j'}^{\mathrm{GEE}}|}{\sum_{j' \in \zeta(j)/\{j\}} \frac{1}{D_{j,j'}}} + 1, \tag{4}$$

where $D_{j,j'}$ denotes the cophenetic distance and $\zeta(j)$ is a cluster with jth OTU by using the partitioning-around-medoids algorithm (PAM) [38]. Here, we consider association signals with close phylogenetic relevance to be amplified by adding the weighted factor w_j to the GEE-based HC test (i.e., GEEHC test). Following the HC test and MiHC, the test statistics for the original and weighted GEE-based HC test (uGEEHC, wGEEHC) are as follows:

$$\mathrm{uGEEHC} = \max_{j \in \{1,2,\ldots,m\}} \frac{r_j/m - p_j}{\sqrt{p_j(1-p_j)/m}}, \tag{5}$$

$$\mathrm{wGEEHC} = \max_{j \in \{1,2,\ldots,m\}} \frac{w_j(r_j/m - p_j)}{\sqrt{p_j(1-p_j)/m}}, \tag{6}$$

where r_j is the descending rank of p_j and w_j is the weight of the jth OTU; $\frac{r_j/m}{\sqrt{p_j(1-p_j)/m}}$ denotes the expected quantiles of significance; and $\frac{p_j}{\sqrt{p_j(1-p_j)/m}}$ denotes the observed quantiles of significance.

The higher criticism test regards the largest deviation between the expected and observed quantiles as test statistics and is sensitive to association signals with high sparsity levels. Thus, we consider the factor represented by the largest deviation to be the most significant factor (OTU) associated with the host phenotype at a high sparsity level. Let uGEEHC_j and wGEEHC_j represent the jth largest uGEEHC and wGEEHC (Eq. 5, 6). We make lists r_u and r_w according to the corresponding OTU names of uGEEHC_j and wGEEHC_j. Moreover, we calculate the corresponding test statistics for uGEEHC and wGEEHC based on GEE with different structures and establish the rank order for significant factors.

Adjustments for low sparsity levels

In fact, the associations between microbes and the host phenotype are not unique in microbiome studies. As previously discussed, there are cases where a small number of OTUs are associated with a host phenotype and where multiple OTUs act together. With the decrease in sparsity level, a higher criticism test gradually loses power [35]. Some adjustments can be made, and the modulated test statistics are integrated to adapt to different scenarios [23]. We consider the single largest deviation between expected and observed quantiles of significance to be adjusted to the weighted average of the first λ maximum deviations (i.e., $\mathrm{uGEEHC}_1, \ldots, \mathrm{uGEEHC}_\lambda$), where $\lambda \in \Gamma$ and Γ is a subset of $\{1, \ldots, m\}$. The models for

tests after adjustments for low sparsity levels are as follows:

$$\mathrm{uGEEHC}_{(\lambda)} = \sum_{j'=1}^{\lambda} \lambda_{j'} * \mathrm{uGEEHC}_{j'}, \tag{7}$$

$$\mathrm{wGEEHC}_{(\lambda)} = \sum_{j'=1}^{\lambda} \lambda_{j'} * \mathrm{wGEEHC}_{j'}, \tag{8}$$

where $\lambda_{j'}$ is the j'th weight and $\sum_{j'=1}^{\lambda} \lambda_{j'} = 1$. For extremely high sparsity levels, we take $\lambda_1 = 1$, that is, the original higher criticism test. For low sparsity levels, we can make appropriate adjustments, for example, an increase of λ and reasonable assignment of $\lambda_{j'}$ (i.e., the increase of $\lambda_{j'}$ with a larger j'). To facilitate comparisons, we take $\lambda_{j'} = \frac{1}{\lambda}$ and $\Gamma = \{1, 3, 5, 7, 9\}$.

Microbiome higher criticism analysis for longitudinal data

It is worth exploring whether the microbiome associated with the host phenotype has phylogenetic relevance. To apply this unknown situation, the GEEMiHC a high-efficiency method is proposed, which is in the following form:

$$T_{\mathrm{GEEMiHC}} = \min(\min_{\lambda \in \Gamma}(P_{\mathrm{uGEEHC}_{(\lambda)}}, P_{\mathrm{wGEEHC}_{(\lambda)}}), P_{\mathrm{Simes}}), \tag{9}$$

where $P_{\mathrm{uGEEHC}_{(\lambda)}}$ and $P_{\mathrm{wGEEHC}_{(\lambda)}}$ are the p values for the test statistics $\mathrm{uGEEHC}_{(\lambda)}$ and $\mathrm{wGEEHC}_{(\lambda)}$, respectively; and P_{Simes} is the p value for the Simes test [39], whose model is as follow:

$$P_{\mathrm{Simes}} = T_{\mathrm{Simes}} = \min_{j \in \{1,2,\ldots,m\}} \{\frac{mp_j}{r_j}\}. \tag{10}$$

For the multiple hypothesis test, the Simes test is much better than the classic Bonferroni in finding several highly correlated signals. Because the Simes test is sensitive to sparse association signals, we obtain the test statistic for GEEMiHC by synthesizing the above tests. It is worth noting that the minimum p value in the above tests is regarded as the test statistic for GEEMiHC and not the final p value for GEEMiHC. It is common to minimize the p value as the test statistic [15, 16, 17]. Let GEEMiHC(AR), GEEMiHC(EX) and GEEMiHC(IN) denote GEEMiHC with corresponding working correlation structures. Of course, MiHC based on the GLM framework is only a special case of GEEMiHC (i.e., GEEMiHC(IN)). In addition, GEE with a predetermined structure may have a good fit for longitudinal data with similar structures, thus leading to good performance of GEEMiHC for detecting microbial association signals. However, GEEMiHC may not be stable for longitudinal data with diverse structures.

In microbiome studies, the correlation structure among different observations for the same cluster is uncertain. Despite GEEMiHCs strong power for detecting sparse association signals from longitudinal microbiome data, the GEE-based approach suffers from an important challenge in terms of selecting the best working correlation structure. The setting for the

structure has a certain influence on the GEE estimator [40], and the underlying correlation structure among different observations from the same cluster varies and is hard to establish. When the cluster size is large enough (i.e., "large n"), the estimators for the GEE approach with different structures may be consistent [31]. However, it might be a little different for a small sample. Actually, the subject size of most microbiome studies is insufficient. The GEE approach does not provide an effective way to determine the setting of the working correlation structure [33]. To accommodate these uncertain correlation structures, it is valuable to analyze various related structures to obtain more accurate conclusions. We further propose the adaptive GEEMiHC, i.e., aGEEMiHC, which integrates multiple GEEMiHCs with different working correlation structures:

$$T_{\text{aGEEMiHC}} = \min(T^{\text{AR}}_{\text{GEEMiHC}}, T^{\text{EX}}_{\text{GEEMiHC}}, T^{\text{IN}}_{\text{GEEMiHC}}), \tag{11}$$

where $T^{\text{AR}}_{\text{GEEMiHC}}$, $T^{\text{EX}}_{\text{GEEMiHC}}$ and $T^{\text{IN}}_{\text{GEEMiHC}}$ represent test statistics for GEEMiHC(AR), GEEMiHC(EX) and GEEMiHC(IN), respectively (Eq. 9). For longitudinal data with an unknown underlying correlation within clusters, GEEMiHCs cannot always have robust stability, but aGEEMiHC can perform well. It is common to perform an optimal test based on multiple tests [18, 41].

P values calculation

The asymptotic distribution was utilized to accurately calculate the p value for uGEEHC [30]. However, it requires a large test and is independent of each other. The scale of OTUs in the microbiome makes it difficult to achieve convergence conditions. Therefore, we adopt the permutations approach to calculate the p value, which has been widely used [15, 18, 23, 42]. The details of the calculation processes of the p value can be found in Additional file 1.

Previous methods

To our knowledge, CSKAT [27] and aGLMMMiRKAT [28] are the methods of performing association analyses for longitudinal microbiome data. Since the latter is an extension of the former and can be applied to data with response variables that shown an exponential distribution, only the latter is selected here for comparison. In consideration of the finite of statistical methods for longitudinal microbiome data, we also compared the methods for testing microbiome-phenotype associations from cross-sectional microbiome data, including MiHC [23], aMiAD [16], aMiSPU [17], OMiAT [18] and OMiRKAT [15]. Unfortunately, these methods present a disadvantage in analyzing cross-sectional data because they do not consider the differences among individuals. Although the fitting effect is great, the method may lead to conclusions that run counter to reality.

Results

To show the power of our methods on different association patterns (i.e., sparsity levels and phylogenetic relevance), we considered diverse association patterns and designed corresponding experiments. Through several comparative methods (please see the "Previous methods" section), we demonstrate the potent power of our methods. To illustrate the stability of GEEMiHC, experiments were performed using datasets of diverse types.

Simulation study

The simulation is based on previous research [17, 23, 28]. We selected the 100 most abundant OTUs from real throat microbiome data [43] and generated an OTU count table based on the Dirichlet-multinomial model. Then, we utilized the R package *ape* [44] to generate a phylogenetic tree. Considering the influence of sample size, we selected two sample sizes with cluster sizes of n = 20 and 40, respectively, and they are divided into two equal parts. Without a loss of generality, we assume that the number of repeated observations k_i is 2 for the former part and k_i is 3 for the latter part. The OTUs for the tth observation in the ith cluster are denoted as $\mathbf{O}_{it} = (O_{it1}, \ldots, O_{itm})^T$. Considering that there may be certain correlations among microbial communities observed at different times, we update the value for subsequent observations by using the random disturbance function [28], that is, $\mathbf{O}_{it} = \frac{1}{2}(\mathbf{O}_{i(t-1)} + \mathbf{O}_{it})$, where $t = 2, 3$. We generate continuous response variable by specifying that

$$y_{it} = 0.5 \cdot \sum_{l=1}^{2} \text{scale}(x_{itl}) + \beta \cdot \sum_{j \in \Lambda} \text{scale}(O_{itj}) + \varepsilon_{it}, \tag{12}$$

where β is the effect size; Λ is a subset of $\{1, 2, \ldots, m\}$, which consists of OTUs related to a host phenotype; x_{it1} and x_{it2} are cluster-specific (e.g., gender) and non-cluster-specific covariate (e.g., age), respectively, which are generated from the Bernoulli distribution B(1, 0.5) and the standard normal distribution $N(0, 1)$; and ε_{it} is an error term from the standard normal distribution. Here, we modulate covariate x_{it2} to perform different observations correlated with the R package *mvtnorm* [45]. To apply to the correlation that generally occurs within clusters, we consider an unstructured correlation structure with the fewest restrictions, namely

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{pmatrix}$$

for the number of repeated observations k_i = 3. $\rho \in [-1, 1]$ is the correlation coefficient. The "scale" function is the standardization function.

To estimate type I error rates, we assume $\beta = 0$ under the condition that hypothesis H_0 is true. The significance level $\alpha = 0.05$ is controlled, and the simulation is repeated 5000 times. Considering the impact of effect size β on power (Fig. 2), we assign $\beta = 1, 0.5$ for n = 20, 40. Based on Eq. 12, the corresponding response variable y_{it} is generated and the simulation is repeated 2000 times to estimate its power. To show the performance of our methods with different association patterns (i.e., different sparsity levels and phylogenetic relevance), we adopted two benchmarks to select the microbial association signals (i.e., OTUs associated with the host phenotype). (1) Random microbial association signals or signals with phylogenetic relevance are selected. (2) Based on different sparsity levels, we selected 2%, 4%, 6%, 8%, 10% and 12% of OTUs associated with the response variable. To illustrate that the rank order for significant factors has a certain reference value for the host phenotype, we also performed many simulation experiments. We considered a constant OTU ID and selected the fixed part associated with the host phenotype. In addition, the simulation is repeated 1000 times. To ensure a fair and effective comparison, all the simulated data in this section are used uniformly and available at https://github.com/SunHan5/GEEMiHC-SupplementaryMaterial.

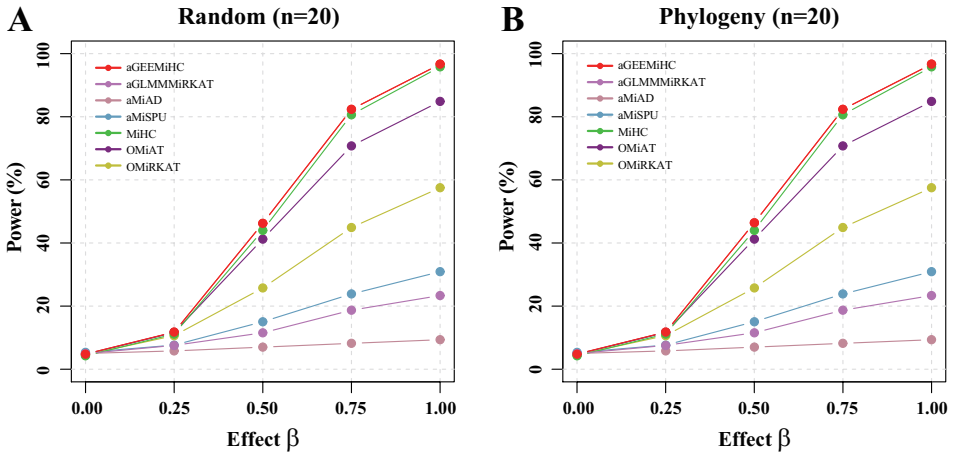

Fig. 2 Comparison of the type I error rates and power between our method (aGEEMiHC) and previous methods, including aGLMMMiRKAT and aMiAD, under diverse association patterns (n = 20). (**A**) Random selection of 2% of microbial signals. (**B**) Phylogenetic relevance-based selection of 2% of microbial signals.

Table 1 Empirical type I error rates for previous methods (aGLMMMiRKAT, aMiAD, aMiSPU, OMiAT, OMiRKAT and MiHC) and our methods (GEEMiHC and aGEEMiHC) at a significance level of 5%.

	Category	Method	n = 20	n = 40
Previous Methods	Non-HC tests	aGLMMMiRKAT	0.050	0.049
		aMiAD	0.050	0.050
		aMiSPU	0.053	0.052
		OMiAT	0.047	0.051
		OMiRKAT	0.045	0.048
	HC test	MiHC	0.043	0.045
Present Methods	Single structure	GEEMiHC(AR)	0.048	0.045
		GEEMiHC(EX)	0.050	0.046
		GEEMiHC(IN)	0.043	0.045
	Multiple structures	aGEEMiHC	0.048	0.046

Type I error

We find that the type I error rates of our methods and previous methods can be satisfactorily controlled at the significance level of 5% (Table 1). For individual GEEHC tests (i.e., uGEEHC$_{(\lambda)}$ and wGEEHC$_{(\lambda)}$) and omnibus GEEMiHC tests (i.e., GEEMiHC and aGEEMiHC), type I error rates are mostly well controlled (Additional file 1: Table S1).

Power

Here, power comparisons are presented: (1) power comparisons for individual GEEHC tests (i.e., uGEEHC$_{(\lambda)}$ and wGEEHC$_{(\lambda)}$) and the Simes test (Additional file 1: Figure S1–S6 (AR, EX, IN, n = 20, 40)); (2) power comparisons between aGEEMiHC and GEEMiHCs (Fig. 3 (n = 20) and Additional file 1: Figure S7 (n = 40)) and (3) power comparisons for aGEEMiHC with previous methods (Fig. 4 and Additional file 1: Figure S8 (n = 40)).

For the individual GEEHC tests, uGEEHC$_{(\lambda)}$ and wGEEHC$_{(\lambda)}$ with the EX or IN structure possess strong power for low λ at the high sparsity levels, although the power for the Simes tests declines rapidly with the decrease of sparsity levels. The results of individual tests with AR structures are disappointing; however, the power for the Simes test is relatively

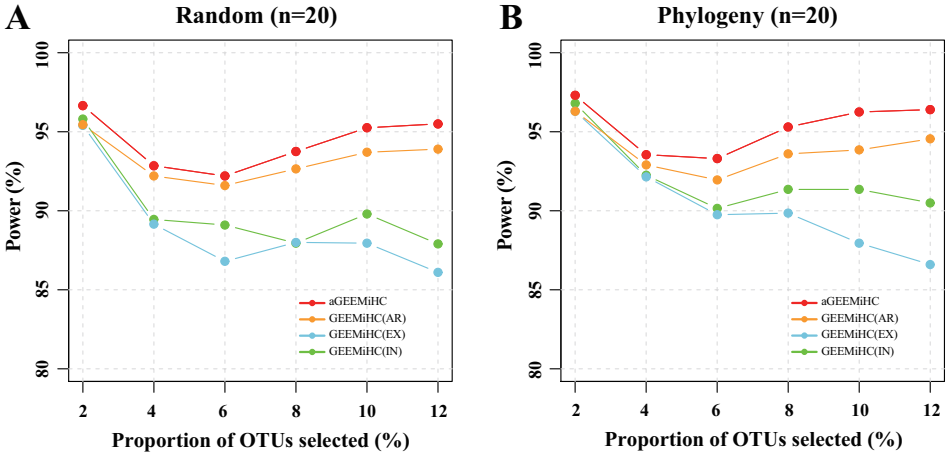

Fig. 3 Comparison of the power for aGEEMiHC and GEEMiHCs with different working correlation structures (i.e., GEEMiHC(AR), GEEMiHC(EX), GEEMiHC(IN)) under diverse association patterns (n = 20). (**A**) Random selection of microbial association signals with diverse sparsity levels. (**B**) Phylogenetic relevance-based selection of microbial association signals with diverse sparsity levels.

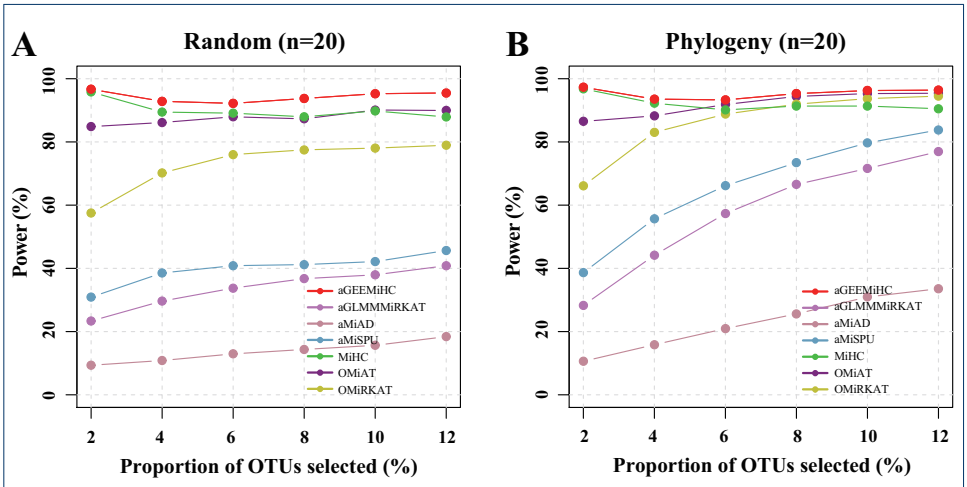

Fig. 4 Comparison of the power for the aGEEMiHC and previous methods under diverse association patterns (n = 20). (**A**) Random selection of microbial association signals with diverse sparsity levels. (**B**) Phylogenetic relevance-based selection of microbial association signals with diverse sparsity levels.

stable (Additional file 1: Figure S1–S6). Compared with GEEMiHC, the power of the aGEEMiHC always shows specific improvements at any association pattern (Fig. 3 and Additional file 1: Figure S7). Compared with other methods, the power of aGEEMiHC is nearly the highest (Fig. 4 (n = 20) and Additional file 1: Figure S8 (n = 40)). The differences in the association patterns for selecting microbial association signals randomly is quite significant. In addition, the gap between aGEEMiHC and MiHC becomes more evident as the sparsity level decreases (Fig. 4).

Accuracy

We selected OTUs so that they were associated with the host phenotype based on different association patterns. To illustrate that the selected OTUs have large deviations between expected and observed quantiles (Eq. 5, 6), we selected the corresponding number of OTUs from the rank order for significant factors. There was a frequency of OTU occurrences when 2% ("711", "277") and 6% of OTUs were selected randomly ("165", "223", "67", "847", "598", "480") (Additional file 1: Figure S9A, S10A). For the former case (i.e., 2%), these associated OTUs appeared at the top of most lists for each method. Compared with the former case, the results of the latter case (i.e., 6%) were not as obvious, although the corresponding OTUs can still be found. Combining different sparsity levels and phylogenetic relevance, we noticed that the preselected OTUs associated with the host phenotype could be accurately found in the case of association signals with high sparsity levels (Additional file 1: Figures S9, S12, S15, S18). As the sparsity level decreased, preselected OTUs could also be found through GEEHC. Unfortunately, the effect was not as significant as before (Additional file 1: Figure S9–S20). Compared with the other structures, the result of GEEHC with the AR structure is not satisfactory. The selection pattern (i.e., random or phylogenetic) had little effect on the results. In contrast, a larger sample size corresponded to more obvious results.

Simulation experiments for datasets of diverse types

We utilize three sets of simulation data based on a prior study [28] to indicate that it can be used for different types of longitudinal microbiome data, which can be found in the R package *GLMMMiRKAT*. In this study, 20 clusters are included, and the measurement times of each cluster are not all the same. A total of 59 samples are included in each dataset, and they differ in that their response variables are from Gaussian, Binomial and Poisson distributions. We observe that there is a slight disparity among GEEMiHCs, although aGEEMiHC invariably discovers significant association signals for different datasets. In contrast, except for aGLMMMiRKAT, the other comparison methods cannot always test association signals, e.g., MiHC for dataset C (Table 2). Moreover, other methods do not work for all datasets, such as OMiAT and OMiRKAT. Compared with MiHC, GEEMiHC(AR) and GEEMiHC(IN), GEEMiHC(EX) has more significant results (i.e., lower p value) on binary data (Table 2 dataset B). Thus, we further designed a comparative study to explore the difference. The dynamic change for the 20 most influential OTUs selected from the rank order is shown in Fig. 5 and Additional file 1: Figure S21–S27. An OTU named "2700" is considered the most important OTU for the host phenotype by GEEHC with an EX structure. It has a larger value for observations confirmed as patient relative to nonpatient within each cluster, e.g., clusters of green, gray, light green, among others. Unfortunately, it does not appear in the list of GEEHC with other structures and MiHC.

Real data applications

Microbial group association study for cross-sectional microbiome data

To illustrate that aGEEMiHC still maintains efficiency for cross-sectional data, we conducted real data experiments to study the association between smoking status and nasopharyngeal and oropharyngeal microbial communities. Charlson et al. gathered the nasopharyngeal and oropharyngeal microbiome data of the subjects [43]. A total of 856 OTUs were collected from 60 samples, including 32 nonsmokers and 28 smokers, and they were already available in the R package *MiSPU* [17]. A total of 273 OTUs with a mean relative abundance over

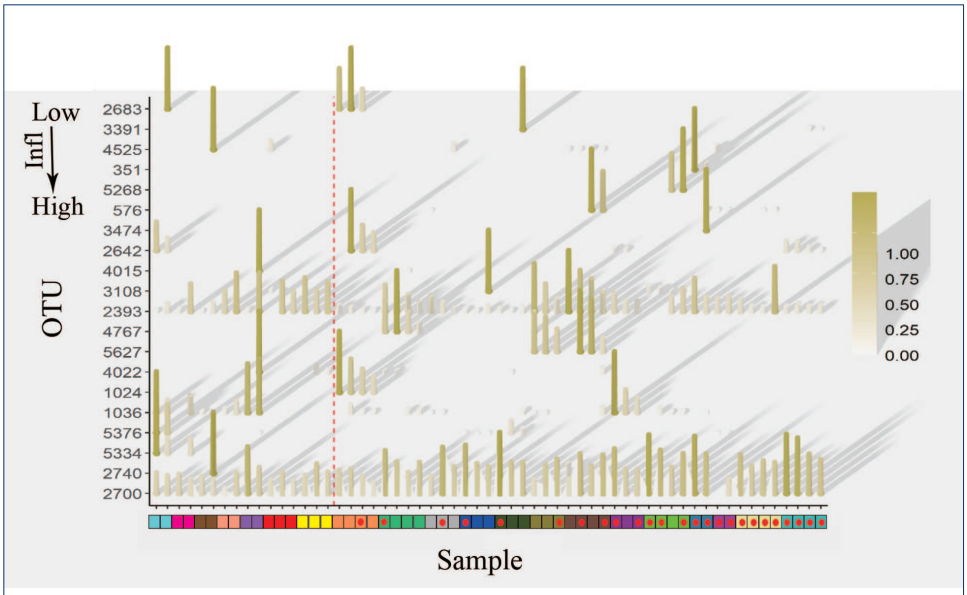

Fig. 5 3D histogram for the abundance of the 20 most influential OTUs selected from the rank order r_u^{EX} for significant factors. We transformed the abundance value (i.e., the min-max normalization). The left side of the dotted red line is full of nonpatients, and the right side represents that at least one observation of all clusters is a patient. The color differences in the blocks at the bottom of the image denote different clusters, and the block with a red dot denotes a sample of the patient.

Table 2 P values for previous and present methods for datasets of diverse types. A, B, and C denote the datasets of response variables obeying Gaussian, Binomial and Poisson distributions, respectively. - indicates that the method cannot handle the data type, and * indicates significant p values.

	Category	Method	A	B	C
Previous Methods	Non-HC tests	aGLMMMiRKAT	**0.005***	**0.002***	**0.013***
		aMiAD	**0.015***	0.443	0.182
		aMiSPU	**0.001***	**0.008***	-
		OMiAT	**<0.001***	**0.005***	-
		OMiRKAT	**0.002***	**0.002***	-
	HC test	MiHC	**<0.001***	**0.013***	0.063
Present Methods	Single structure	GEEMiHC(AR)	**0.042***	**0.012***	**0.050***
		GEEMiHC(EX)	**<0.001***	**0.001***	**0.037***
		GEEMiHC(IN)	**<0.001***	**0.013***	0.063
	Multiple structures	aGEEMiHC	**<0.001***	**0.002***	**0.041***

10^{-4} were further singled out, which can be obtained in the R package *MiHC* [23]. For the application of GEEMiHC with different structures to the processed data, we can obtain consistent conclusions (i.e., consistent p values for GEEMiHC with different structures, Additional file 1: Table S2). A significant relation is also detected by previous methods except aMiSPU (Table S3). In addition, the 10 most influential OTUs detected by MiHC can be discovered in the top 10 list of GEEMiHC (Additional file 1: Figure S28).

Microbial group association study for longitudinal microbiome data

To explore the volatile microbial signatures of patients with Crohn's disease (CD) from longitudinal data, Vázquez-Baeza et al. collected stool samples from 31 subjects, which included 19 CD subjects and 12 control subjects [46]. The raw data can be found at the

Table 3 P values for previous and present methods for the longitudinal microbiome datasets of different scenarios. $*$ indicates significant p values.

	Category	Method	Scenario 1	Scenario 2
Previous Methods	Non-HC tests	aGLMMMiRKAT	0.867	0.399
		aMiAD	0.130	0.120
		aMiSPU	0.248	0.297
		OMiAT	**0.021***	**0.008***
		OMiRKAT	**0.020***	**0.001***
	HC test	MiHC	**0.013***	0.056
Present Methods	Single structure	GEEMiHC(AR)	**0.005***	**0.006***
		GEEMiHC(EX)	**0.004***	**0.010***
		GEEMiHC(IN)	**0.013***	0.056
	Multiple structures	aGEEMiHC	**0.006***	**0.008***

European Bioinformatics Institute (EBI: ERP104742), and data processed by using the QIIME pipeline [47] can be found in Qiita (https://qiita.ucsd.edu/, ID:2538). Since many repeated values are produced by multiple measurements, only a portion of the samples is included in our experiment. According to the time span of data acquisition, we select different observations (at least once and at most 5 times) for each cluster. A total of 81 samples (Scenario 1) were eventually included in our experiment. We further preprocessed the data and selected 71 samples (Scenario 2) from 81 samples to prove the robust stability. Then, we singled out 372 and 376 OTUs under the condition that the relative abundance was not less than 10^{-4} for Scenarios 1 and 2, respectively. For covariates, we selected age and smoking status, which may have a stronger effect than other factors, including gender [48]. Finally, we used MEGA7 [49] to construct a corresponding phylogenetic tree according to the dataset selected.

In practice, we discovered a significant association between CD and the gut microbiome. For scenario 1, all individual and omnibus GEEMiHC tests showed a significant association, while half of the comparison methods (e.g., aGLMMMiRKAT) could not achieve this association (Table 3). Moreover, over half of the individual unweighted tests (i.e., uGEEHC) and MiHC did not show a significant association in scenario 2 (Additional file 1: Table S4). This finding implies that these tests either establish an inconsistent association mode or ignore the differences among individuals, resulting in biased conclusions and insufficient robustness in processing longitudinal data. In contrast, aGEEMiHC maintains stable p value. According to the p values for individual tests, we can find a high probability that a variety of OTUs are associated with CD (i.e., low sparsity level). We also generated Q-Q plots between the expected and observed quantiles and compiled lists of the 10 most influential OTUs (Additional file 1: Figure S29–S31). In addition, we depict the differences in the 50 most important OTUs between CD patients and nonpatients in scenario 1 (Additional file 1: Figure S32–S37). We can conclude that the gut microbiome is more abundant in nonpatients than in patients. Among the top 30 OTUs, some can be detected for most of the nonpatients, including *Lachnospiraceae* (OTU ID 3586, 7929, 9498, et al.), *Ruminococcaceae* (1784, 4760, 594, et al.) and others, while an empty space is observed for all patients. Previous research has ever reported differences in these bacteria between CD patients and healthy people [50]. The processed data and mapping file are available at R package *GEEMiHC* and https://github.com/SunHan5/GEEMiHC-SupplementaryMaterial, respectively.

To further illustrate that our ranking of OTUs had a certain significance, we divided 372 OTUs in scenario 1 into 6 groups on average based on uGEEHC with AR structure and utilized support vector machines (SVMs) to classify diseases. Due to the limited sample size, we adopted the 10-fold cross-validation method. Default values are selected for all parameters in SVM. We concluded that the accuracy of the group formed by the top OTUs reached a maximum value, indicating that they are most likely to be potential biomarkers of CD compared with other groups (Additional file 1: Table S5).

Discussion

In this paper, we adopt the GEE framework to analyze longitudinal data, rather than generalized linear mixed model. Because we consider that GEE can adapt to longitudinal data with different correlations between samples by setting corresponded working correlation structure, and we can further integrate these GEEs with diverse structures to detect sparse microbial association signals adaptively from longitudinal microbiome data. It should be noted that GEE is recently utilized in a different scenario to estimate both predictor effects and OTU correlations [51] in the analysis of microbiome data. In addition, aGEEMiHC achieves a superior and stable performance according to a statistical power. Considering the use of permutation-based method used to calculate the p-value, the computing performance of aGEEMiHC may encounter challenges when processing a big data. aGEEMiHC as well as other microbiome-based association tests focuses on the analysis of the association between the microbiome and a host phenotype, but they cannot evaluate the causal relationship.

Although we only design 16S rRNA gene sequencing data as real data experiments, aGEEMiHC can also be utilized to analyze metagenomic sequencing data. In the analysis of metagenomics, we consider that average nucleotide identity (ANI) [52, 53] can also be adopted to describe interspecies relationships. Furthermore, there is still much room for the development of methods to detect sparse microbial association signals from longitudinal microbiome data, for instance, the association between the microbiome and multiple host phenotypes [54] or a host phenotype with multiple categories, such as disease severity [55].

Conclusions

In this paper, we propose a novel method, aGEEMiHC, to test sparse microbial association signals from longitudinal microbiome data. We fully consider the information from correlations among different observations from the same cluster by utilizing a generalized estimating equations framework. Despite the good performance of GEEMiHC for different association patterns (i.e., sparsity levels and phylogenetic relevance), it is an important challenge to select the best working correlation structure. To adapt to diverse correlation structures, we further develop aGEEMiHC, which is a data-driven comprehensive test that integrates GEEMiHC with different structures. In contrast to GEEMiHC, aGEEMiHC is more powerful and stable for detecting association signals from datasets with diverse types. We utilized it on both real cross-sectional data and longitudinal data to detect the associations between smoking status and the nasopharyngeal and oropharyngeal microbial communities as well as the associations between Crohn's disease and the gut microbiome. It retained excellent performance over different types of longitudinal microbiome data. In addition, we also ranked the significant factors associated with the host phenotype. It can identify the influential OTUs on the host phenotype to some extent, and they can be considered potential disease biomarkers.

Additional Files

Additional file 1: PDF file includes calculation process for p value, supplemental tables (**Tables S1–S5**) and supplemental figures (**Figures S1–S37**). (PDF 10676 kb)

Abbreviations

aGEEMiHC: Adaptive microbiome higher criticism analysis based on generalized estimating equations; aGLMMMiRKAT: Adaptive distance-based kernel association test based on the generalized linear mixed model; aMiAD: Adaptive Microbiome α-diversity-based association test; aMiSPU: Adaptive microbiome sum of powered score tests; CD: Crohn's Disease; CSKAT: correlated sequence kernel association test; HC: Higher criticism; MiHC: Microbiome higher criticism analysis; OMiRKAT: Optimal microbiome regression-based kernel association test; OTU: Operational taxonomic unit; PAM: Partitioning-around-medoids; uGEEHC: Unweighted GEE-based higher criticism; wGEEHC: Weighted GEE-based higher criticism

Declarations

Ethical approval and consent to participate

This study only involves the analysis of existing data and all microbiome datasets used are publicly available. No ethics approval or consent to participate is required for this study.

Consent for publication

All microbiome datasets used are publicly available. No consent for publication is required for this study.

Availability of data and materials

We used two public microbiome datasets, including cross-sectional microbiome data and longitudinal microbiome data. These real microbiome data supporting the results of this article are available: (1) the cross-sectional microbiome data: the association between smoking status and the nasopharyngeal and oropharyngeal microbial communities (available in the R package, *MiHC*). (2) Longitudinal microbiome data: the association between Crohn's disease and the gut microbiome (available in the European Bioinformatics Institute (EBI) database under accession code ERP104742, https://www.ebi.ac.uk and in the Qiita database under accession code: 2538, https://qiita.ucsd.edu). The longitudinal microbiome data in our experiment (including phylogenetic tree and OTU table) can be found at R package *GEEMiHC*. All simulated data, mapping files and other sources are available at https://github.com/SunHan5/GEEMiHC-SupplementaryMaterial.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the National Natural Science Foundation of China [61872157, 61932008, and 61532008].

Authors' contributions

HS contributed to the methodological ideas for aGEEMiHC, performed the simulations and real data analyses, visualized the results, developed the software package, and wrote the manuscript. XH contributed to the real data curation and analyses, visualized the results, developed the software package and wrote the manuscript. BH contributed to the real data curation and analyses, visualized the results and wrote the manuscript. YT, TH and XJ contributed to the methodological ideas for aGEEMiHC, the biological insights and interpretations and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to show our deepest gratitude to all the code and data referenced in this article, especially the MiHC method written by Koh Hyunwook and Zhao Ni, which is a great inspiration for our work.

Author details

[1]School of Mathematics and Statistics, Central China Normal University, Wuhan, China. [2]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China. [3]Collaborative & Innovative Center for Educational Technology, Central China Normal University, Wuhan, China. [4]School of Computer, Central China Normal University, Wuhan, China. [5]National Language Resources Monitoring & Research Center for Network Media, Central China Normal University, Wuhan, China.

References

1. Huttenhower, C., Jacques, I.: Structure, function and diversity of the healthy human microbiome. Nature **486**, 207–214 (2012)
2. Helmink, B.A., Khan, M.A.W., Hermann, A., Gopalakrishnan, V., Wargo, J.A.: The microbiome, cancer, and cancer therapy. Nat Med **25**, 377–388 (2019)
3. McIlroy, J., Ianiro, G., Mukhopadhya, I., Hansen, R., Hold, G.L.: Review article: the gut microbiome in inflammatory bowel disease-avenues for microbial management. Aliment Pharmacol Ther **47**, 26–42 (2018)
4. Zupancic, M.L., Cantarel, B.L., Liu, Z., Drabek, E.F., Ryan, K.A., Cirimotich, S., Jones, C., Knight, R., Walters, W.A., Knights, D., Mongodin, E.F., Horenstein, R.B., Mitchell, B.D., Steinle, N., Snitker, S., Shuldiner, A.R., Fraser, C.M.: Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. PLoS ONE **7**, 43052 (2012)
5. Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., Huttenhower, C., Littman, D.R.: Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. eLife **2**, 01202 (2013)
6. Song, M., Chan, A.T., Sun, J.: Influence of the Gut Microbiome, Diet, and Environment on Risk of Colorectal Cancer. Gastroenterology **158**, 322–340 (2020)
7. Fang, P., Kazmi, S.A., Jameson, K.G., Hsiao, E.Y.: The Microbiome as a Modifier of Neurodegenerative Disease Risk. Cell Host & Microbe **28**, 201–222 (2020)
8. Cryan, J.F., O'Riordan, K.J., Sandhu, K., Peterson, V., Dinan, T.G.: The gut microbiome in neurological disorders. The Lancet Neurology **19**, 179–194 (2020)
9. Raman, A.S., Gehrig, J.L., Venkatesh, S., Chang, H.-W., Hibberd, M.C., Subramanian, S., Kang, G., Bessong, P.O., Lima, A.A.M., Kosek, M.N., Petri, W.A., Rodionov, D.A., Arzamasov, A.A., Leyn, S.A., Osterman, A.L., Huq, S., Mostafa, I., Islam, M., Mahfuz, M., Haque, R., Ahmed, T., Barratt, M.J., Gordon, J.I.: A sparse covarying unit that describes healthy and impaired human gut microbiota development. Science **365**, 4735 (2019)

10. Bhatt, A.P., Redinbo, M.R., Bultman, S.J.: The role of the microbiome in cancer development and therapy: Microbiome and Cancer. CA: A Cancer Journal for Clinicians **67**, 326–344 (2017)
11. Weinstock, G.M.: Genomic approaches to studying the human microbiota. Nature **489**, 250–256 (2012)
12. Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A.L., Madsen, K.L., Wong, G.K.-S.: Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front. Microbiol. **7**, 1002358 (2016)
13. Hamady, M., Knight, R.: Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Research **19**, 1141–1152 (2009)
14. Wang, L., Zhou, J., Qu, A.: Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics **68**, 353–360 (2012)
15. Zhao, N., Chen, J., Carroll, I., Ringel-Kulka, T., Epstein, M., Zhou, H., Zhou, J., Ringel, Y., Li, H., Wu, M.: Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. The American Journal of Human Genetics **96**, 797–807 (2015)
16. Koh, H.: An adaptive microbiome -diversity-based association analysis method. Sci Rep **8**, 18026 (2018)
17. Wu, C., Chen, J., Kim, J., Pan, W.: An adaptive association test for microbiome data. Genome Med **8**, 56 (2016)
18. Koh, H., Blaser, M.J., Li, H.: A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. Microbiome **5**, 45 (2017)
19. Jiang, H., Ling, Z., Zhang, Y., Mao, H., Ma, Z., Yin, Y., Wang, W., Tang, W., Tan, Z., Shi, J., Li, L., Ruan, B.: Altered fecal microbiota composition in patients with major depressive disorder. Brain, Behavior, and Immunity **48**, 186–194 (2015)
20. Bajaj, J.S., Betrapally, N.S., Hylemon, P.B., Heuman, D.M., Daita, K., White, M.B., Unser, A., Thacker, L.R., Sanyal, A.J., Kang, D.J., Sikaroodi, M., Gillevet, P.M.: Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy: HEPATOLOGY, Vol. XX, No. X, 2015. Hepatology **62**, 1260–1271 (2015)
21. Magruder, M., Edusei, E., Zhang, L., Albakry, S., Satlin, M.J., Westblade, L.F., Malha, L., Sze, C., Lubetzky, M., Dadhania, D.M., Lee, J.R.: Gut commensal microbiota and decreased risk for *Enterobacteriaceae* bacteriuria and urinary tract infection. Gut Microbes **12**, 1805281 (2020)
22. Mejía-León, M.E., Petrosino, J.F., Ajami, N.J., Domínguez-Bello, M.G., de la Barca, A.M.C.: Fecal microbiota imbalance in Mexican children with type 1 diabetes. Sci Rep **4**, 3814 (2015)
23. Koh, H., Zhao, N.: A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. Microbiome **8**, 63 (2020)
24. Secrier, M., Schneider, R.: Visualizing time-related data in biology, a review. Briefings in Bioinformatics **15**, 771–782 (2014)
25. Stewart, C.J., Ajami, N.J., OBrien, J.L., Hutchinson, D.S., Smith, D.P., Wong, M.C., Ross, M.C., Lloyd, R.E., Doddapaneni, H., Metcalf, G.A., Muzny, D., Gibbs, R.A., Vatanen, T., Huttenhower, C., Xavier, R.J., Rewers, M., Hagopian, W., Toppari, J., Ziegler, A.-G., She, J.-X., Akolkar, B., Lernmark, A., Hyoty, H., Vehik, K., Krischer, J.P., Petrosino, J.F.: Temporal development of the gut microbiome in early childhood from the TEDDY study. Nature **562**, 583–588 (2018)
26. Mayhew, D., Devos, N., Lambert, C., Brown, J.R., Clarke, S.C., Kim, V.L., Magid-Slav, M., Miller, B.E., Ostridge, K.K., Patel, R., Sathe, G., Simola, D.F., Staples, K.J., Sung, R., Tal-Singer, R., Tuck, A.C., Van Horn, S., Weynants, V., Williams, N.P., Devaster, J.-M., Wilkinson, T.M.A.: Longitudinal profiling of the lung microbiome in the AERIS study demonstrates repeatability of bacterial and eosinophilic COPD exacerbations. Thorax **73**, 422–430 (2018)
27. Zhan, X., Xue, L., Zheng, H., Plantinga, A., Wu, M.C., Schaid, D.J., Zhao, N., Chen, J.: A smallsample kernel association test for correlated data with application to microbiome association studies. Genet. Epidemiol. **42**, 772–782 (2018)
28. 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. **10**, 458 (2019)
29. Liang, K.-Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
30. Donoho, D., Jin, J.: Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. **32**, 962–994 (2004)
31. Wang, L.: GEE analysis of clustered binary data with diverging number of covariates. Ann. Statist. **39**, 389–417 (2011)
32. Chaganty, N.R., Joe, H.: Efficiency of generalized estimating equations for binary responses. J Royal Statistical Soc B **66**, 851–860 (2004)
33. Twisk, J.W.R.: Applied Longitudinal Data Analysis for Epidemiology, 2nd edn. Cambridge University Press, Cambridge (2013)
34. Li, D., Pan, J.: Empirical likelihood for generalized linear models with longitudinal data. Journal of Multivariate Analysis **114**, 63–73 (2013)
35. Barnett, I., Mukherjee, R., Lin, X.: The Generalized Higher Criticism for Testing SNP-Set Effects in Genetic Association Studies. Journal of the American Statistical Association **112**, 64–76 (2017)
36. Arias-Castro, E., Candès, E.J., Plan, Y.: Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. Ann. Statist. **39**, 2533–2556 (2011)
37. Barnett, I.J., Lin, X.: Analytical p-value calculation for the higher criticism test in finite-d problems. Biometrika **101**, 964–970 (2014)
38. Reynolds, A.P., Richards, G., de la Iglesia, B., Rayward-Smith, V.J.: Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. J Math Model Algor **5**, 475–504 (2006)
39. Simes, R.J.: An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**, 751–754 (1986)
40. Twisk, J.W.R.: Longitudinal Data Analysis. A Comparison Between Generalized Estimating Equations and Random Coefficient Analysis. Eur J Epidemiol **19**, 769–776 (2004)
41. Koh, H., Livanos, A.E., Blaser, M.J., Li, H.: A highly adaptive microbiome-based association test for survival traits. BMC Genomics **19**, 210 (2018)
42. Hall, P., Jin, J.: Innovated higher criticism for detecting sparse signals in correlated noise. Ann. Statist. **38**, 1686–1732 (2010)
43. Charlson, E.S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F.D., Collman, R.G.: Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE **5**, 15216 (2010)
44. Paradis, E., Claude, J., Strimmer, K.: APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics **20**, 289–290 (2004)
45. Mi, X., Miwa, T., Hothorn, T.: New Numerical Algorithm for Multivariate Normal Probabilities in Package mvtnorm. The R Journal **1**, 37 (2009)
46. Vázquez-Baeza, Y., Gonzalez, A., Xu, Z.Z., Washburne, A., Herfarth, H.H., Sartor, R.B., Knight, R.: Guiding longitudinal sampling in IBD cohorts. Gut **67**, 1743–1745 (2018)
47. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich,

J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R.: QIIME allows analysis of high-throughput community sequencing data. Nat Methods **7**, 335–336 (2010)

48. Torres, J., Mehandru, S., Colombel, J.-F., Peyrin-Biroulet, L.: Crohn's disease. The Lancet **389**, 1741–1755 (2017)
49. Kumar, S., Stecher, G., Tamura, K.: MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol **33**, 1870–1874 (2016)
50. Schirmer, M., Garner, A., Vlamakis, H., Xavier, R.J.: Microbial genes and pathways in inflammatory bowel disease. Nat Rev Microbiol **17**, 497–511 (2019)
51. Chen, B., Xu, W.: Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures. PLoS Comput Biol **16**, 1008108 (2020)
52. Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M.: DNADNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology **57**, 81–91 (2007)
53. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S.: High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun **9**, 5114 (2018)
54. Zhan, X., Tong, X., Zhao, N., Maity, A., Wu, M.C., Chen, J.: A small-sample multivariate kernel machine test for microbiome association studies: Zhan et al. Genet. Epidemiol. **41**, 210–220 (2017)
55. Clausen, M.-L., Agner, T., Lilje, B., Edslev, S.M., Johannesen, T.B., Andersen, P.S.: Association of Disease Severity With Skin Microbiome and Filaggrin Gene Mutations in Adult Atopic Dermatitis. JAMA Dermatol **154**, 293 (2018)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf