Koh et al. Microbiome (2017) 5:45
DOI 10.1186/s40168-017-0262-x

Microbiome

METHODOLOGY Open Access

A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping

Hyunwook Koh[1], Martin J. Blaser[2] and Huilin Li[1*]

Abstract

Background: The role of the microbiota in human health and disease has been increasingly studied, gathering momentum through the use of high-throughput technologies. Further identification of the roles of specific microbes is necessary to better understand the mechanisms involved in diseases related to microbiome perturbations.

Methods: Here, we introduce a new microbiome-based group association testing method, optimal microbiome-based association test (OMiAT). OMiAT is a data-driven testing method which takes an optimal test throughout different tests from the sum of powered score tests (SPU) and microbiome regression-based kernel association test (MiRKAT). We illustrate that OMiAT efficiently discovers significant association signals arising from varying microbial abundances and different relative contributions from microbial abundance and phylogenetic information. We also propose a way to apply it to fine-mapping of diverse upper-level taxa at different taxonomic ranks (e.g., phylum, class, order, family, and genus), as well as the entire microbial community, within a newly introduced microbial taxa discovery framework, microbiome comprehensive association mapping (MiCAM).

Results: Our extensive simulations demonstrate that OMiAT is highly robust and powerful compared with other existing methods, while correctly controlling type I error rates. Our real data analyses also confirm that MiCAM is especially efficient for the assessment of upper-level taxa by integrating OMiAT as a group analytic method.

Conclusions: OMiAT is attractive in practice due to the high complexity of microbiome data and the unknown true nature of the state. MiCAM also provides a hierarchical association map for numerous microbial taxa and can also be used as a guideline for further investigation on the roles of discovered taxa in human health and disease.

Keywords: Microbial association test, Microbial group analysis, Upper-level taxa, Taxonomic structure, Phylogenetic tree, Comprehensive association mapping

Background

The human microbiota is the set of all microorganisms inhabiting the human body [1]. Recently, their roles in human health and disease have been highlighted [2–9]. Advancement in studies of the microbiota has gathered momentum due to the advent of high-throughput sequencing technologies which enable microbiota profiling [10–12]. Since raw sequences preprocessed by these platforms include highly variable regions to be used as unique markers for each microbe, diverse microbes can be identified based on the sequence similarity and then assigned to operational taxonomic units (OTUs) [13]. These OTUs are characterized by their quantity, read count, or relative abundance, and the difference in microbial abundances may be associated with health or disease status [14, 15]. The phylogenetic tree illustrates taxonomical and evolutionary relationships among diverse microbes [13, 16, 17], and its related microbial complexity provides further insights about possible health and disease etiologies [18, 19]. Further identification of microbial taxa

* Correspondence: huilin.li@nyumc.org
[1]Department of Population Health and Environmental Medicine, New York University School of Medicine, New York, NY 10016, USA
Full list of author information is available at the end of the article

is needed to better understand the relationship of particular microbiota with human health and disease. It is also common that recent studies report discovered upper-level taxa at a high taxonomic rank (e.g., phylum, class, order, family, and genus) along with the dynamics of the entire microbial community complexity instead of enumerating individual microorganisms. The upper-level taxa can be considered as groups of various nested lineages. Hence, likewise the entire microbial community, numerous statistical challenges can arise to analyze them properly [20]. Nevertheless, a conventional ecological method, referenced as the aggregate-based method in this paper, is most commonly used for association testing [8, 9, 21–24]. The aggregate-based method is based on a univariate analysis, using aggregates of microbial abundances in a lower-level lineage per sample as a single predictor variable. It is also regarded as an approach equipped with the popular methods, the linear discriminant analysis effect size (LEfSe) [21], STAMP [22], DESeq2 [23], and metagenomeSeq-fit Zig [24]. The major problem of this approach is its underlying assumption that associated OTUs nested at each upper-level taxon are all in the same effect direction. Any violation of this assumption can lead to a substantial loss of power.

As a counterpart to the aggregate-based method, we investigate two existing methods, microbiome regression-based kernel association test (MiRKAT) [25] and microbiome-based sum of powered score tests (MiSPU) [26], and propose a new method, optimal microbiome-based association test (OMiAT), for more sophisticate microbial association testing. Recently, MiRKAT has been spotlighted in the literature because of its comprehensive capability to incorporate diverse distance-based measures, including the unique fraction (UniFrac) distance [27–29] and the Bray-Curtis dissimilarity, into its kernel machine regression framework [30]. The distance-based measures integrate different relative contributions from microbial abundance and phylogenetic tree information, and thus, they suit different association patterns, respectively [25, 27–29]. In practice, any strong biological evidence inclined to a particular distance-based measure is usually absent; hence, the data-driven approach of MiRKAT, Optimal MiRKAT, which uses an optimal test among different distance-based measure trials, is highly attractive. On the other hand, MiSPU is constructed on the sum of powered score tests (SPU) framework [31] based on a newly defined measure, generalized taxon proportion [26]. Similar to unweighted and weighted UniFrac distances [27, 28], Wu et al. [26] describe that two different versions of the generalized taxon proportion, unweighted and weighted generalized taxon proportion, are suitable for discovering rare and common/abundant taxa, respectively. Wu et al. [26] also insist that the data-driven approach of MiSPU, adaptive MiSPU (aMiSPU), is robust and powerful by taking a highly adaptive test, utilizing the variable selection/weighting of the SPU framework based on the two generalized taxon proportions, comprehensively. However, we argue that the unweighted generalized proportion might not be sufficient to account for varying microbial abundances because it is based on the presence or absence of microbial taxa with no further microbial abundance information incorporation. In addition, since the generalized taxon proportion weights microbial taxa by their branch lengths [26], its weighting scheme might be efficient only when associated microbial taxa have relatively large branch lengths.

Our proposed method, OMiAT, is a data-driven testing method which takes an optimal test through diverse tests from both SPU and MiRKAT. To avoid confusion, we explain here that the SPU used for OMiAT is different from MiSPU in that it is not based on the generalized taxon proportion but implemented on standardized compositional data with no phylogenetic information incorporation. We have first been convinced that MiRKAT is suitable to modulate relative contributions from microbial abundance and phylogenetic information by the use of diverse distance-based measures. However, we emphasize that SPU is advantageous over MiRKAT to modulate different association patterns arising from highly imbalanced microbial abundances, utilizing its wide range of power value choices [31]. Consequently, OMiAT is highly efficient to discover significant association signals from diverse underlying association patterns and thus attractive in practice due to the high complexity of microbiome data and the unknown true nature of the state. Our extensive simulations and real data analyses also demonstrate more robust and powerful performance of OMiAT, compared with other competing methods.

We also introduce a microbial taxa discovery framework, namely, microbiome comprehensive association mapping (MiCAM), which uses different configurations to fine-map diverse microbial taxa throughout all taxonomic ranks, comprehensively. MiCAM tests all upper- and lower-level taxa and applies multiple testing correction per taxonomic rank. MiCAM incorporates OMiAT as a group analytic method for assessing upper-level taxa. MiCAM discovers significantly associated taxa and controls false discovery rate at 5% per taxonomic rank. Testing numerous microbial taxa individually may lead to a huge computational burden. Thus, we apply a combined permutation-based algorithm to MiCAM to obtain stable outcomes (e.g., P values) in a computationally manageable manner. A newly introduced visualization approach for MiCAM also helps to organize discovered microbial taxa hierarchically.

The methodological aspects of OMiAT and MiCAM can be found in the following "Methods" section. Extensive simulation experiments and real data analyses are addressed in the "Results" section.

Methods

Models and notations

This section is devoted to describe the methodological aspects of OMiAT and how its performance is affected in microbial group analysis. Since OMiAT is based on two existing methods, SPU [31] and MiRKAT [25], we start with the descriptions of SPU and MiRKAT. OMiAT shares some of the useful features of SPU and MiRKAT, as follows. OMiAT is based on a generalized linear model framework so that different types of outcome traits, such as continuous and binary responses, with potential covariate adjustments, can be handled. OMiAT is also based on score tests which do not require any statistical estimation for the parameters of major interest [32].

Suppose the data include n subjects, p OTUs, and q covariates (e.g., environmental factors) and the subscripts, i, j, and k, indicate a subject, an OTU, and a covariate, respectively. Then, an $n \times 1$ vector, Y, for the outcome response, either as a form of continuous or binary traits, is marked asY_i for $i = 1, \ldots, n$, an $n \times p$ matrix, Z, for the OTUs in a microbial group is marked as Z_{ij} for $i = 1,\ldots, n$ and $j = 1,\ldots, p$, and an $n \times k$ matrix, X, for the covariates is marked as X_{ik} for $i = 1,\ldots, n$ and $k = 1, \ldots, q$.

To relate OTUs with an outcome response while adjusting for covariates, we consider a multiple regression model (Eq. 1) for continuous traits and a multiple logistic regression model (Eq. 2) for binary traits.

$$Y_i = \beta_0 + \sum_{k=1}^{q} X_{ik}\alpha_k + \sum_{j=1}^{p} Z_{ij}\beta_j + \in_i, \quad (1)$$

where $\in_i$ is an error term which is independently and identically distributed with a mean zero and a variance of σ^2.

$$\text{logit } P(Y_i = 1) = \beta_0 + \sum_{k=1}^{q} X_{ik}\alpha_k + \sum_{j=1}^{p} Z_{ij}\beta_j. \quad (2)$$

Then, the score vector to test the null hypothesis of no association between OTUs and an outcome trait, $H_0 : \beta = (\beta_1, \ldots, \beta_p)' = 0$, is given as in Eq. 3.

$$U = (U_1, \ldots, , U_p)' = \left(\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_{i,0}\right)Z_{i1}, \ldots, \sum_{i=1}^{n}\left(Y_i - \hat{\mu}_{i,0}\right)Z_{ip}\right)', \quad (3)$$

where $\hat{\mu}_{i,0}$ is the predicted value of Y_i under H_0 which can be estimated as $\hat{\beta}_0 + \sum_{k=1}^{q} X_{ik}\hat{\alpha}_k$ for continuous traits or $\text{logit}^{-1}\left(\hat{\beta}_0 + \sum_{k=1}^{q} X_{ik}\hat{\alpha}_k\right)$ for binary traits, where $\hat{\beta}_0$ and $\hat{\alpha}_k$ are maximum likelihood estimates under H_0 [33–35].

SPU [31]

Pan et al. [31] formulated their method, SPU, with its test statistic as in Eq. 4 to obtain a generalized framework to sum individual score components to be powered with diverse γ value choices ($\gamma \geq 1$, integer).

$$T_{SPU(\gamma)} = \sum_{j=1}^{p} U_j^{\gamma} \quad (4)$$

SPU was originally proposed for gene- or region-based association testing in genome-wide association studies. SPU in our proposed method, OMiAT, is implemented on the standardized compositional data (i.e., starting from the form of compositional data (i.e., percentages), for each OTU, we subtract its mean from individual raw percentages and then divide the difference by its standard deviation) because of varying total reads per sample.

In microbial group analysis, the SPU test using an odd value of γ is suitable when associated OTUs have the same effect direction, while the SPU test using an even value of γ is more suitable when those are in mixed effect directions [31, 33, 36]. To explain, when γ is an odd number, the score components in the SPU test can be canceled out in its final summing stage by the existence of opposite directional components, which results in a significant loss of power in the mixed effect situations. Instead, when γ is an even number, those in the SPU test can be protected, which result in a powerful performance in the mixed effect directions. As γ increases, relatively larger score components will gradually be weighted more, while relatively small score components will gradually be ignored [31]. We can see that the score statistic value is affected by OTU abundance as in Eq. 3. Thus, a situation in which abundant OTUs are associated indicates that their corresponding components in the score vector are large, and thus, SPU using a large value of γ is more suitable by weighting them more and removing noisy small signals from the others. In contrast, when associated OTUs are rare in abundance, indicating small score components, SPU using a small value of γ can be more suitable by preserving them in the final aggregate. Since, in most microbiomes, OTU abundance is highly imbalanced across individual OTUs, the high receptivity of SPU to a wide range of γ value choices must be maximized in microbial group analyses.

Therefore, the performance of SPU differs according to γ and the true underlying association patterns. Since we cannot predict which situation is related to our study in advance, the adaptive SPU (aSPU) test (Eq. 5) which

takes the minimum P value among different γ value trials can be importantly considered [31].

$$T_{aSPU} = min_{\gamma \epsilon \Gamma} P_{SPU(\gamma)} \quad (5)$$

The γ value can take any natural number (i.e., $\Gamma = \mathbb{N}$), but we used the candidate set, $\Gamma = \{1,2,3,4,5,6,7,8, \infty\}$, where $T_{SPU(\infty)} = \max\{|U_j|, j = 1,\ldots\ldots, p\}$ [31], in our simulations and real data analyses and it was sufficient.

MiRKAT [25]

MiRKAT has recently been introduced to the microbiome research community for microbial community-level association testing. Here, we describe its key formula and ideas and refer to the original paper [25] for more details. MiRKAT is built on the kernel machine regression framework with the kernel formula, Eq. 6 [25, 30], to incorporate diverse distance-based measures, such as UniFrac distances [27–29] and Bray-Curtis dissimilarity.

$$K = -\frac{1}{2}\left(I - \frac{11'}{n}\right) D^2 \left(I - \frac{11'}{n}\right), \quad (6)$$

where D is the $n \times n$ pairwise distance matrix and D^2 is its element-wise square, I is the $n \times n$ identity matrix, and 1 in $11'$ is the column vector of n ones. We can specify the pairwise distance matrix, D, in this kernel formula, choosing among diverse distance-based measures. Then, using the resulting kernel, the variance component score statistic can be formulated with Eq. 7 below.

$$Q_{MiRKAT(k)} = \frac{1}{2\Phi}\left(Y - \hat{\mu}_0\right)' K_{(k)} \left(Y - \hat{\mu}_0\right), \quad (7)$$

where Φ is the dispersion parameter which can be estimated as $\Phi = \hat{\sigma}_0^2$, where $\hat{\sigma}_0^2$ is the estimated residual variance under H_0, for continuous traits, and as $\Phi = 1$ for binary traits; $Y - \hat{\mu}_0$ is the vector, $\left(Y_1 - \hat{\mu}_{1,0}, \ldots, Y_n - \hat{\mu}_{n,0}\right)$; and k is an index for a particular kernel based on a chosen distance-based measure.

Of importance is that different distance-based measures suit different association patterns, respectively [25, 27–29]. The UniFrac distances are constructed on the basis of phylogenetic tree information and modulate the extent of microbial abundance to be incorporated by different weighting schemes [27–29]. Thus, the UniFrac distances are suitable when associated OTUs are phylogenetically related. Then, the unweighted UniFrac distance [27] is suitable for considering rare lineages, while the weighted UniFrac distance [28] can be used for studying common/abundant lineages. The generalized UniFrac distance [29] is a compromise version; thus, its use can also be modulated according to its parameter value. When associated OTUs are not phylogenetically related, the Bray-Curtis dissimilarity can be best because it is constructed based solely on microbial abundance not incorporating phylogenetic tree information. In terms of relative contribution from microbial abundance and phylogenetic tree information, we can also understand that the Bray-Curtis dissimilarity is most inclined to microbial abundance, and then, the weighted, generalized, and unweighted UniFrac distances follow in the name ordered. In practice, prior knowledge about the true underlying association patterns of numerous OTUs is likely to be absent; thus, Zhao et al. [25] have proposed the data-driven approach of MiRKAT, Optimal MiRKAT, which takes the minimum P value among multiple distance-based measure trials as in Eq. 8.

$$Q_{OMiRKAT} = min_{k \epsilon \{1, \ldots, l\}} P_{MiRKAT(k)} \quad (8)$$

where $P_{MiRKAT(k)}$ is the P value of the MiRKAT test based on $Q_{MiRKAT(k)}$. We used seven candidate distance-based measures, Bray-Curtis dissimilarity, unweighted UniFrac, weighted UniFrac, four different generalized UniFrac measures with parameter values, 0, 0.25, 0.5, and 0.75, respectively, in our simulations and real data analyses.

OMiAT

OMiAT takes the minimum P value from all the score tests for SPU (i.e., $T_{SPU(\gamma)}$ in Eq. 4) and MiRKAT (i.e., $Q_{MiRKAT(k)}$ in Eq. 7) as its test statistic and can be simply expressed in Eq. 9.

$$M_{OMiAT} = min\{T_{aSPU}, Q_{OMiRKAT}\} \quad (9)$$

Consequently, OMiAT is highly robust and powerful by taking an optimal test from all different tests for varying microbial abundances by SPU and for different relative contributions from microbial abundance and phylogenetic information by MiRKAT. Of course, we do not use the genuine minimum P value, M_{OMiAT}, to be reported as a final outcome P value, but it is a test statistic to be used for estimating a P value. As with Pan et al. [31] and Zhao et al. [25], we also use a permutation-based method [37] to calculate P values for the test statistics, $T_{SPU(\gamma)}$, T_{aSPU}, $Q_{MiRKAT(k)}$, $Q_{OMiRKAT}$, and M_{OMiAT}. Detailed information on it is addressed in Additional file 1: Table S1.

Aggregate-based method

There exist different aggregate-based methods, LEfSe [21], STAMP [22], DESeq2 [23], and metagenomeSeq-fit Zig [24]. LEfSe and STAMP employ the non-parametric Kruskal-Wallis test [21, 22, 38] as a univariate analytic method. This non-parametric method is designed for one-way layout data structure, and thus, it is difficult to handle covariate adjustments (e.g., environmental

factors). Moreover, this method cannot analyze continuous outcome traits; hence, its usability is limited. For DESeq2 and metagenomeSeq-fit Zig, the main assumption of their parametric methods may not be validated due to the issue of relative abundance, which can result in inflated type I error rates [16, 39]. For these reasons, we do not consider these existing machineries, as they are, for the aggregate-based method. Instead, we employ a semi-parametric approach which is based on a score test and a permutation-based method for the aggregate-based method to be investigated in our simulations and real data analyses. It begins with the standardized compositional data and aggregates it per sample. Then, using a resulting single predictor variable for the aggregates and an outcome variable, we estimate P values based on the score test statistic, U, in Eq. 3 and a permutation-based method [37].

MiCAM

In this section, we illustrate a new microbial taxa discovery framework, MiCAM, to fine-map diverse microbial taxa from the highest (e.g., kingdom/the entire community) to the lowest (e.g., species) taxonomic rank. MiCAM tests all microbial taxa using different configurations for the assessment of upper-level taxa, taxa in the species taxonomic rank, and taxa that include only one OTU, respectively, and applies multiple testing correction per taxonomic rank. We also describe its testing algorithm and graphical representation.

Assessment of upper-level taxa

The upper-level taxa at different taxonomic ranks (e.g., kingdom, phylum, class, order, family, and genus) are a group of individual OTUs except for a few which include only one OTU. Thus, we apply OMiAT to assess the upper-level taxa by sub-grouping OTUs and pruning a phylogenetic tree for the ones nested in each of the upper-level taxa. For small upper-level taxa, including only a few OTUs, the UniFrac distances may not be computed when there is no phylogenetic disparity for any pairwise sample comparison. For this case, the Optimal MiRKAT part in Eq. 9 is replaced with the MiRKAT based on a single kernel for the Bray-Curtis dissimilarity.

Assessment of taxa in the species taxonomic rank

The species taxonomic rank may not be regarded as a microbial group but as individual microbes. However, we include this species rank to be analyzed because testing separately for individual species is also of interest. Although it might be considered ideal to have one-to-one correspondence between OTUs and species, in reality, some species include multiple OTU. As such, we can consider an OTU as the smallest unit and thus those species as a group so that OMiAT is applied. Alternatively, users can consider any species as the smallest unit, by properly combining the relevant OTUs per species.

Assessment of taxa that include only one OTU

Group analytic tools lose their efficiency when the taxa surveyed include only one OTU. For this case, we employ a simple semi-parametric approach which is based on a score test, U (Eq. 3), and a permutation-based method [37].

Control of false discovery rate

Importantly, since there can exist multiple tests for multiple taxa at a given taxonomic rank, a multiple testing correction procedure is needed to suppress increased type I error rate. We apply the Benjamini-Hochberg procedure to control false discovery rate at 5% per taxonomic rank [40–42] as it is valid robustly whenever the multiple tests are independent or correlated in various scenarios [43]. Thus, the error probability applies to a family of inferences at each taxonomic rank.

A combined permutation-based algorithm

One issue we have encountered is a large computational burden to test numerous microbial taxa throughout all different taxonomic ranks. Although the score-based test using a permutation-based method is efficient in computation to test a small number of microbial taxa, testing all existing microbial taxa can require enormous computational time. Moreover, to reach sufficient convergence in all the outcomes (e.g., P values) consistently, computational needs can be even greater.

Our experiences have also revealed that outcomes are sensitive to different implementation specifications (e.g., different numbers of permutations). Accordingly, especially when the P values are close to 0.05, their discovery status can even be reversed. Therefore, we apply a combined permutation-based algorithm which shares the same vector and permuted vectors of residuals for every microbial taxon assessment; hence, we can avoid repeating such procedures. We have found that this combined permutation-based algorithm produces stable outcomes (e.g., P values) with better correspondence/convergence than individual permutation-based tests for numerous assessments using a relatively moderate number of permutations (e.g., 50,000). In contrast, individual permutation-based tests may produce highly irregular outcomes unless an extremely high number of permutations (e.g., 500,000) are specified.

A hierarchical visualization

A graphical representation is introduced to summarize discovered and undiscovered microbial taxa in a

hierarchical taxonomic structure. To explain, we stack all individual OTUs vertically and enumerate taxonomic ranks from highest to lowest horizontally with each OTU consistently belonging to their upper-level taxa. Then, using color, we highlight microbial taxa according to their discovery status to overview multiple discovery statuses comprehensively. In addition, on the right end line on each graph, we enumerate the effect directions for each OTU by calculating the score test statistic, U (Eq. 3), for each OTU, assigning "+", if it is ≥0 and "−" if it is <0. Related outcomes are presented in the "Real data analysis" section. Detailed information on the exact taxonomic names and their P values matched with each OTU ID can also be found in separate tables using our software facility, OMiAT.

Other methods in MiCAM

Although we propose OMiAT to be used as a group analytic method in MiCAM, for the purpose of comparison in our real data analyses, we have also integrated other competing group analytic methods, Optimal MiRKAT, aMiSPU, and the aggregate-based method, respectively, into the MiCAM framework.

Results

Simulations

We have conducted extensive simulations to evaluate different methods, OMiAT, Optimal MiRKAT, aMiSPU, and the aggregate-based method in terms of type I error and statistical power. For simplicity, we use the entire community as a microbial group of interest. In practice, any subgroup for different upper-level taxa can be considered.

Simulation design

The simulation design used is based on the prior studies [24, 25, 30]. We first simulated OTU counts for 100 subjects from the Dirichlet-multinomial distribution with total reads per sample to be randomly sampled from a negative binomial distribution with mean 300 and size 10. The dispersion parameter and proportion means to be inserted into the Dirichlet-multinomial distribution were estimated from the early childhood antibiotics and the microbiome (ECAM) project's intestinal microbiome data [7]. The ECAM data includes 2261 OTUs for 43 infants, but as a demonstration, we selected 32 infants aged from 30 to 40 days of life and applied a filtering rule that retains only OTUs with a proportion mean $>10^{-3}$, as such, 71 OTUs were included in the analysis. Then, continuous and binary outcome traits were generated under the linear model (Eq. 10) and the logistic model (Eq. 11), respectively.

$$y_i = 0.5 * \mathrm{scale}(X_{1i} + X_{2i}) + \sum_{j=1}^{p} \beta_j \mathrm{scale}(Z_{ij}) + \epsilon_i \tag{10}$$

$$\mathrm{logit}\, P(y_i = 1) = 0.5 * \mathrm{scale}(X_{1i} + X_{2i}) + \sum_{j=1}^{p} \beta_j \mathrm{scale}(Z_{ij}), \tag{11}$$

where ϵ_i is an error term with $\epsilon_i \sim N(0,1)$, X_{1i} and X_{2i} are two covariates, Z_{ij} is an OTU count, and the "scale" function is for the standardization to have mean 0 and standard deviation (SD) 1 and is further defined as $\mathrm{scale}(X_{1i} + X_{2i}) = \frac{X_{1i} + X_{2i} - \mathrm{mean}(X_{11} + X_{21},\ X_{12} + X_{22},\ \ldots,\ X_{1n} + X_{2n})}{\mathrm{SD}(X_{11} + X_{21},\ X_{12} + X_{22},\ \ldots,\ X_{1n} + X_{2n})}$ and $\mathrm{scale}(Z_{ij}) = \frac{Z_{ij} - \mathrm{mean}(Z_{1j},\ Z_{2j},\ \ldots,\ Z_{nj})}{\mathrm{SD}(Z_{1j},\ Z_{2j},\ \ldots,\ Z_{nj})}$, for subjects $i = 1,\ldots, n$ and OTUs $j = 1,\ldots, p$. X_{1i}'s were generated to be independent with OTUs from the Bernoulli distribution with success probability 0.5. X_{2i}'s were generated in two different ways: one to be correlated with OTUs as $X_{2i} = \sum_{j \in \Lambda} \mathrm{scale}(Z_{ij}) + N(0, 1)$, where Λ is a set of indices for associated OTUs, and the other to be independent with OTUs as $X_{2i} = N(0,1)$.

To estimate type I error rates, outcome traits were generated from the null model by setting $\beta = (\beta_1, \ldots, \beta_p)' = 0$. To estimate statistical powers, we first selected a set of associated OTUs with four different simulation scenarios: (1) OTUs in upper 10% in abundance, (2) a random 10% of OTUs, (3) OTUs in lower 10% in abundance, and (4) OTUs in the selected cluster using the partitioning-around-medoids (PAM) algorithm [44]. The fourth scenario is for a situation when associated OTUs are phylogenetically related. For this, we first partitioned all OTUs into five clusters using the PAM algorithm based on the cophenetic distances in the real phylogenetic tree [45]. Then, we randomly assigned all these five clusters (which contain 20.9, 21.6, 32.5, 15.3, and 9.7% of total abundance, respectively) into each iteration in our simulations. This is to overcome the arbitrariness of the choice of clusters and opposite to working on a single or a couple of chosen cluster(s) as conducted in prior studies [24, 26, 30]. If we work on simulations with some particularly chosen clusters, it would not be a fair comparison because those clusters can be favorable to a particular testing method. Especially, adaptive methods are needed to be tested from diverse simulation environments (e.g., differing microbial abundances and phylogenetic relationships using different associated clusters evenly) to evaluate their adaptivity.

Because the fourth scenario combines both microbial abundance and phylogenetic information, it is believed to be more realistic than the first three scenarios. However, the first three scenarios are useful to check whether each method discovers abundant or rare microbial taxa,

equivalently. We denote Λ as a set of indices for the associated OTUs. Then, $\beta_{j\in\Lambda}$ is a vector of coefficients corresponding to associated OTUs. For each experimental setting, $\beta_{j\in\Lambda}$ are simulated with three different continuous uniform distributions, Uniform(0,1), Uniform(0,2), and Uniform(0,3), for the same effect direction and with another three continuous uniform distributions, Uniform(−1,1), Uniform(−2,2), and Uniform(−3,3), for mixed effect directions, separately.

Simulation results

For presentation, we include only the outcomes for the adaptive methods (with the exception of the aggregate-based method) and for the logistic models, moving all the other outcomes to additional material (Additional file 2: Figure S1 reports complete type I error estimates, Additional files 3 and 4: Figure S2 and S3 report complete power estimates for the linear models, and Additional files 5 and 6: Figure S4 and S5 report complete power estimates for the logistic models).

Type I error First, we observe mostly well-controlled type I error rates (≤~5%) across all methods [Table 1, Additional file 2: Figure S1]. Therefore, any discovered microbial taxa using any of these methods are from statistically valid approaches.

Power We observe that with the increase of effect size, the power increases for all methods under any simulation scenario [Figs. 1 and 2, Additional files 3, 4, 5, and 6: Figure S2–S5]. We also observe that power is generally higher for linear than logistic models [Additional files 3, 4, 5, and 6: Figure S2–S5], but the relative performance among different methods within the linear or logistic model remains similar. Whether the covariate, X_2, is independent or correlated with OTUs does not strongly alter the relative performance.

We observe that OMiAT is clearly more powerful than the other methods under most of the scenarios [Figs. 1 and 2]. Exceptions include situations where abundant OTUs are associated, in which Optimal MiRKAT is most powerful [Figs. 1 and 2a, b], and where rare OTUs are associated, in which the aggregate-based method is most powerful [Figs. 1 and 2e, f]. However, even then, OMiAT is highly comparable. Based on the first three scenarios, we can observe that SPU using a high γ value (≥4) is powerful when abundant OTUs are associated [Additional files 3, 4, 5, and 6: Figure S2–S5: A, B], SPU using a medium γ value (~4) is powerful when random OTUs are associated [Additional files 3, 4, 5, and 6: Figure S2–S5: C, D], and SPU using a low γ value (≤4) is powerful when rare OTUs are associated [Additional files 3, 4, 5, and 6: Figure S2–S5: E, F]. In contrast, the major drawback of Optimal MiRKAT occurs when rare or random OTUs are associated, resulting in low power values [Figs. 1c–f and 2c–f]. Consequently, we can observe that OMiAT reaches the highest power considerably beyond Optimal MiRKAT for the fourth scenario [Figs. 1g, h and 2g, h], as explained by the assistance from diverse SPU tests within its machinery.

The aggregate-based method is highly underpowered when associated OTUs are in mixed effect directions [Figs. 1b, d, f, h and 2b, d, f, h], as explained by the violation of its underlying assumption that all associated OTUs are in the same effect direction. Moreover, the aggregate-based method is less powerful than the other methods under most of the other scenarios [Figs. 1a, c, g and 2a, c, g]. The only exception is when rare OTUs are associated and they are in the same effect direction [Figs. 1e and 2e], which can be explained similarly with the situation where the SPU test using a small value of γ outperforms.

aMiSPU is not observed to be as powerful as Optimal MiRKAT as well as OMiAT [Figs. 1 and 2], and it is opposite to the simulation outcomes reported in Wu et al. [26]. For the reasons, we can further observe two related simulation outcomes as follows. Firstly, the MiSPU tests based on the unweighted generalized taxon proportion are mostly underpowered [Additional files 3, 4, 5, and 6: Figure S2–S5]. This may be because it is solely based on the presence or absence of microbial taxa with no further microbial abundance incorporation. Secondly, the MiSPU tests based on the weighted generalized taxon proportion are less powerful than the MiRKAT tests based on different UniFrac distances [Additional files 3, 4, 5, and 6: Figure S2–S5]. This may be because the

Table 1 Type I error rate estimates in percent for both linear and logistic models

		OMiAT	Optimal MiRKAT	aMiSPU	Aggregate-based
Linear model	Independent X_2	4.98	5.10	5.13	4.84
	Correlated X_2	4.93	4.92	5.15	4.92
Logistic model	Independent X_2	5.09	4.98	5.09	5.10
	Correlated X_2	5.26	5.21	4.94	4.99

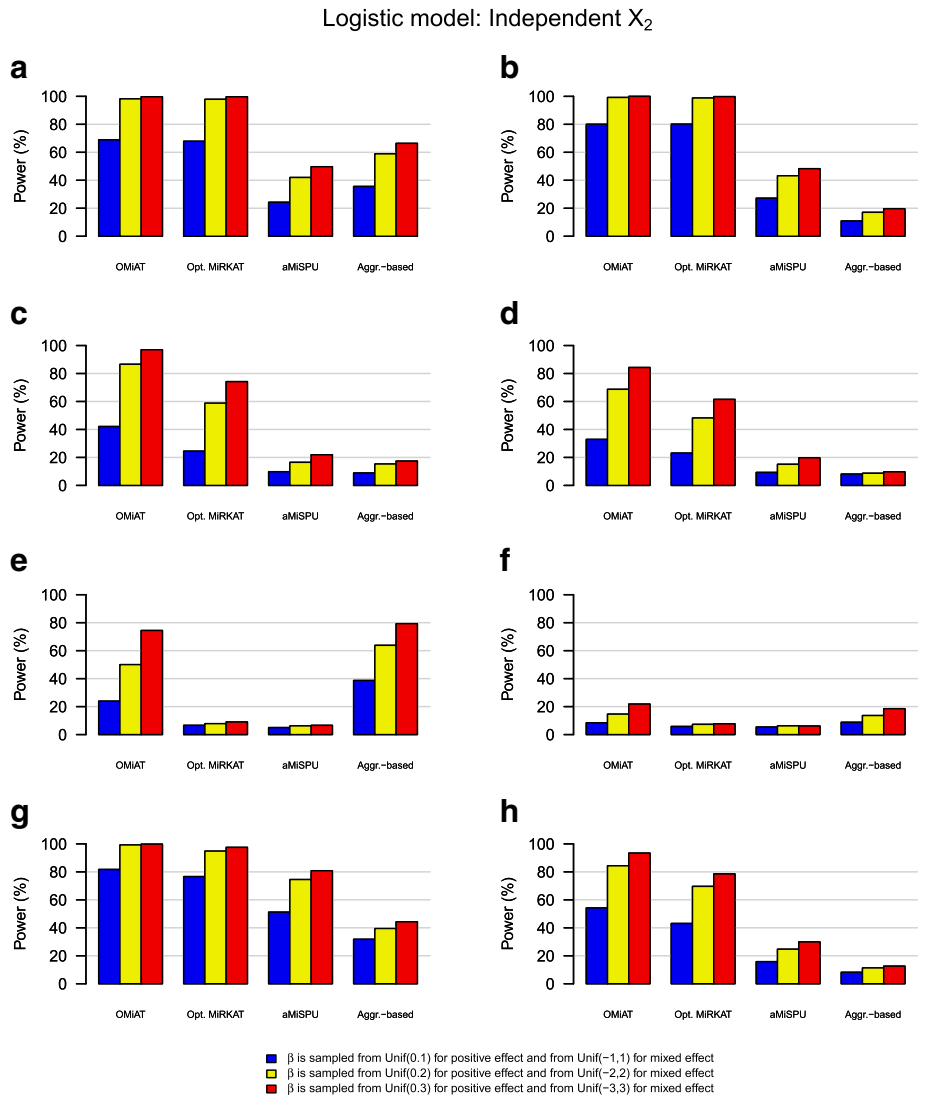

Fig. 1 Power estimates for the logistic model when the covariate X_2 is independent with OTUs. **a** Positive effect: OTUs in upper 10% in abundance. **b** Mixed effect: OTUs in upper 10% in abundance. **c** Positive effect: a random 10% of OTUs. **d** Mixed effect: a random 10% of OTUs. **e** Positive effect: OTUs in lower 10% of abundance. **f** Mixed effect: OTUs in lower 10% of abundance. **g** Positive effect: OTUs in the cluster. **h** Mixed effect: OTUs in the cluster

generalized taxon proportion weights microbial taxa by their branch lengths [26], and thus, it is efficient only when associated microbial taxa have relatively large branch lengths, but not in general. In addition, in Wu et al. [26], a limited number of candidate distance-based measures were surveyed for different MiRKAT tests, which can lead to a lower power for Optimal MiRKAT [26].

Real data analysis

Here, we apply the methods, OMiAT, Optimal MiRKAT, aMiSPU, and the aggregate-based method, respectively, to the MiCAM framework to assess existing microbial taxa throughout all different taxonomic ranks from kingdom to species using two real data sets [6, 7]. Along with the simulation results, we also compare different methods by the extent of discovered microbial taxa from our real data analyses.

Sustained effects on intestinal microbiota by early-life low-dose penicillin exposure [6]

Cox et al. [6] have conducted a microbiome profiling study to examine whether the intestinal microbiota altered during maturation by low-dose antibiotic, low-dose penicillin (LDP) induces sustained effects on body composition (e.g., tendency to obesity). Here, we

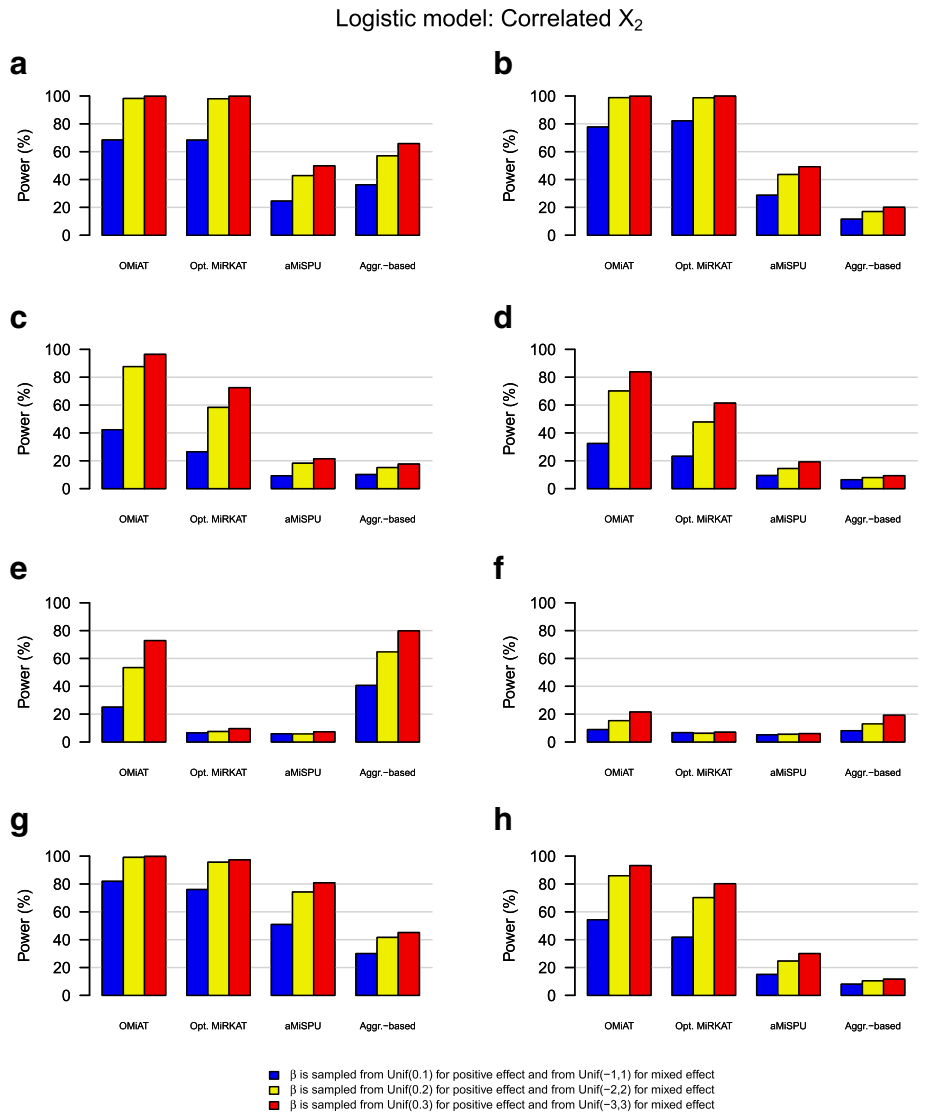

Fig. 2 Power estimates for the logistic model when the covariate X_2 is correlated with OTUs. **a** Positive effect: OTUs in upper 10% in abundance. **b** Mixed effect: OTUs in upper 10% in abundance. **c** Positive effect: a random 10% of OTUs. **d** Mixed effect: a random 10% of OTUs. **e** Positive effect: OTUs in lower 10% of abundance. **f** Mixed effect: OTUs in lower 10% of abundance. **g** Positive effect: OTUs in the cluster. **h** Mixed effect: OTUs in the cluster

re-examine a small portion of its original analyses, which address whether the LDP-affected microbial compositions are recovered after its cessation. For this, cecal microbiota were transferred from control-microbiota recipients (CR1) to seven germ-free mice (CR2) and LDP-microbiota recipients (PR1) to eight germ-free mice (PR2). Fecal specimens from these 15 recipient mice were collected 23 days after the transfer, and their DNA samples were analyzed by targeting the V4 region of the bacterial 16S rRNA gene. Using the QIIME pipeline [13] to quantify OTUs and construct a phylogenetic tree, 424 OTUs were observed, but after filtering OTUs with a proportion mean $\leq 10^{-3}$, 28 OTUs were analyzed.

We examined whether there is any disparity in microbial profiles between two groups (CR2 and PR2). No covariate adjustment was made assuming that other potential confounding factors were already well controlled in the randomized experimental design.

To summarize the results [Fig. 3, Additional file 7: Table S2], while many upper-level taxa were discovered consistently by the three methods, OMiAT, Optimal MiRKAT, and aMiSPU, the aggregate-based method discovered apparently less. Since many of the OTUs trend in opposite directions, the weakness of the aggregate-based method likely originates from the violation of its assumption of same effect directions of all associated

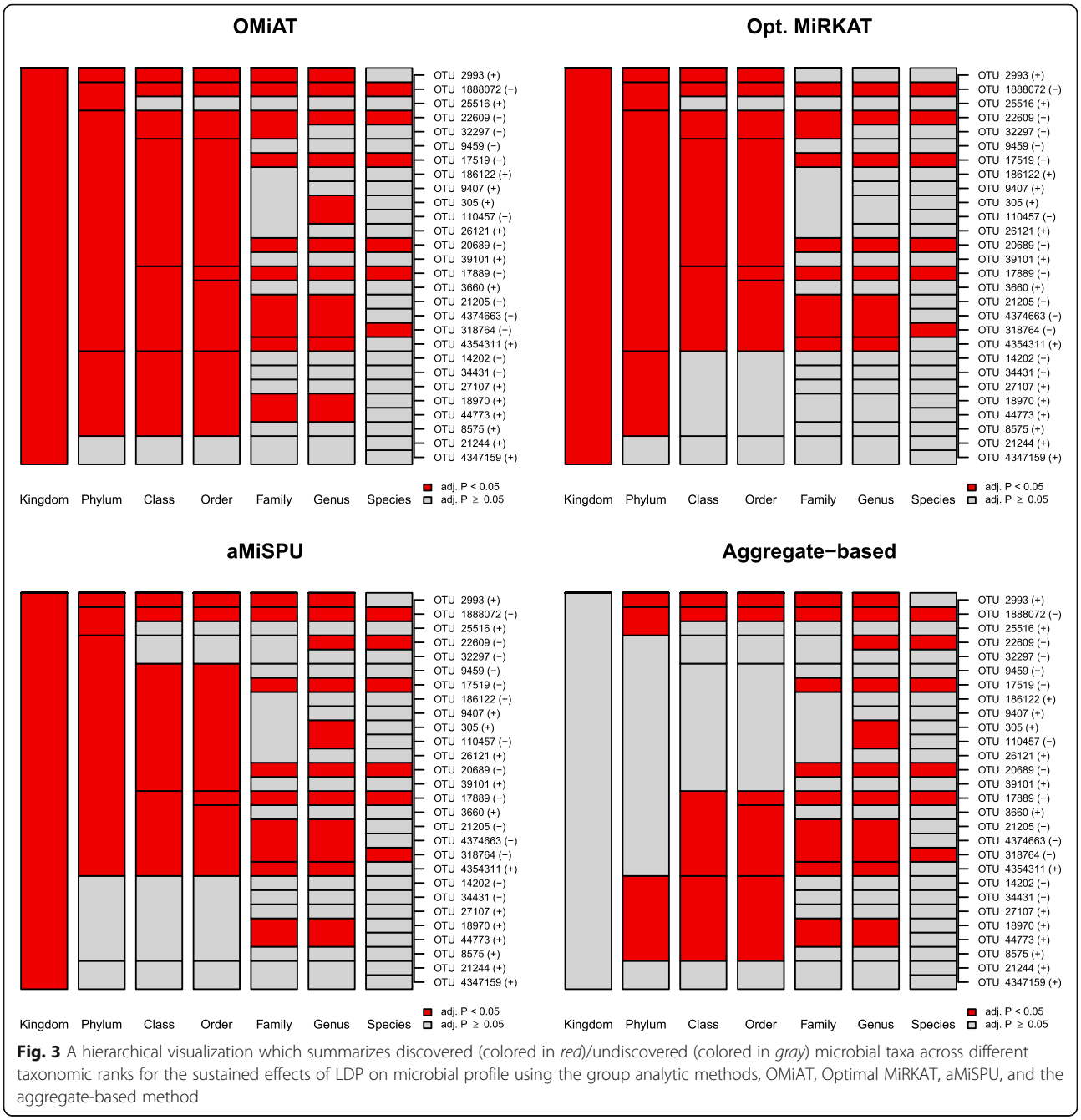

Fig. 3 A hierarchical visualization which summarizes discovered (colored in *red*)/undiscovered (colored in *gray*) microbial taxa across different taxonomic ranks for the sustained effects of LDP on microbial profile using the group analytic methods, OMiAT, Optimal MiRKAT, aMiSPU, and the aggregate-based method

OTUs. OMiAT discovered the greatest number of taxa, which is consistent with our simulations. The *P* values for testing the entire microbial community level were estimated as <0.001 for OMiAT, <0.001 for Optimal MiRKAT, <0.001 for aMiSPU, and 0.518 for the aggregate-based method.

Effects on intestinal microbiota by the early-life factor, vaginal or cesarean birth [7]

The early childhood antibiotics and the microbiome (ECAM) project is a longitudinal microbiome profiling study to examine the hypotheses that early life factors, such as delivery mode (e.g., vaginal or cesarean birth), infant nutrition (breast milk or formula predominance), and antibiotic usage, influence microbial community development, resulting in sustained states. Among 32 infants studied, 21 and 11 were delivered by vaginal and cesarean delivery, respectively. None had received antibiotics, and two covariate adjustments, predominant diet and sex, were included in our analyses.

The fecal samples from these infants were also assessed for the bacterial 16S rRNA V4 region, and

Fig. 4 A hierarchical visualization which summarizes discovered (colored in *red*)/undiscovered (colored in *gray*) microbial taxa across different taxonomic ranks for the effects of delivery method on microbial profile using the group analytic methods, OMiAT and Optimal MiRKAT

Fig. 5 A hierarchical visualization which summarizes discovered (colored in *red*)/undiscovered (colored in *gray*) microbial taxa across different taxonomic ranks for the effects of delivery method on microbial profile using the group analytic methods, aMiSPU, and the aggregate-based method

OTUs were determined, and a phylogenetic tree was constructed [13]. There were 2261 OTUs in the original, but after filtering with a proportion mean $\leq 10^{-3}$, 71 OTUs were analyzed.

We found that the aggregate-based method discovered apparently fewer microbial taxa than the other methods, since many of the OTUs had opposite effect directions [Figs. 4 and 5, Additional file 8: Table S3]. While many microbial taxa were consistently discovered by OMiAT and aMiSPU, many taxa do not overlap with Optimal MiRKAT [Figs. 4 and 5, Additional file 8: Table S3]. Some of the discovery statuses for the use of Optimal MiRKAT were also highly irregular by different specifications of the number of permutations since their P values were too close to 0.05. Here, again, OMiAT discovered the greatest number of taxa. The P values for testing the entire microbial community level were estimated as 0.005 for OMiAT, <0.001 for Optimal MiRKAT, 0.023 for aMiSPU, and 0.495 for the aggregate-based method.

Discussion

The computational performance of MiCAM still needs to be improved. Although we could obtain stable outcomes efficiently for our real data analyses by the use of combined permutation-based algorithm, analyzing a big data may pose huge computational challenges in practice. To illustrate, as the number of microbial taxa increases (e.g., using a less stringent filtering rule), its computational burden can increase. MiCAM is written in R to facilitate the use of existing R functions, but in case of such huge computation, the use of a lower-level language can be required.

We have described the use of different group analytic methods and the MiCAM framework focusing on microbiota profiles via target sequencing for the 16S rRNA gene [10]. However, as long as a data includes OTU abundance and a phylogenetic tree in groups of interest, similar approaches can apply. Therefore, the extension to the shotgun metagenomic data for the whole microbial genomes [11, 46] is also highly feasible.

Although MiRKAT and SPU cover a wide range of association patterns, the candidate tests in the search space of OMiAT in Eq. 9 are not limited to those two sets of tests. If one finds other tests which suit other association patterns which are not covered by MiRKAT and SPU tests, one can include them into the search space to yield extra power.

As an extension, OMiAT can also be implemented into a hierarchical multiple testing scheme [47] to identify which microbes are associated with the phenotype of interest in the lowest taxonomic rank. By utilizing the taxonomic tree structure, one can test the lower-level lineages only when their upper-level taxon is significant. In this way, the number of tests can be reduced and smaller penalty due to multiple testing correction is needed.

Conclusions

In this paper, we investigated two existing methods, MiRKAT and MiSPU, and a new method, OMiAT, that can be used as a counterpart to aggregate-based methods [21–24] in microbiome-based association studies. Due to the lack of knowledge about true underlying association patterns of numerous OTUs, the data-driven approaches (OMiAT, Optimal MiRKAT, and aMiSPU) are highly attractive in practice. We confirmed that they are all statistically valid approaches with well-controlled type I error rates. Among those, we observed that our proposed method, OMiAT, is most robust and powerful through extensive simulations and real data analyses. The high performance of OMiAT comes from its high adaptivity to suit two unique features of microbiome data, the high imbalance in microbial abundance and phylogenetic information.

The newly proposed microbial taxa discovery framework, MiCAM, organizes different configurations to test microbial taxa through a breadth of taxonomic ranks, and it is especially efficient for the assessment of upper-level taxa by integrating OMiAT as a group analytic method. Of importance is that MiCAM produces statistically significantly associated microbial taxa with a well-defined false discovery rate criterion. Its hierarchical visualization also helps rapidly overview multiple discovery statuses. Consequently, we can obtain a hierarchical association map for numerous microbial taxa, and this can also be used as a guideline for further investigation on the roles of discovered microbial taxa in human health and disease.

Additional files

Additional file 1: Table S1. The permutation-based method to estimate P values for the test statistics, $T_{SPU(\gamma)}$, T_{aSPU}, $Q_{MiRKAT(k)}$, $Q_{OMiRKAT}$, and M_{OMiAT} [25, 31, 37]. (DOCX 23 kb)

Additional file 2: Figure S1. Type I error rate estimates for both linear and logistic models and for using the covariate, X_2, as either correlated or independent with OTUs. (PDF 7 kb)

Additional file 3: Figure S2. Power estimates for the linear model using the covariate, X_2, as independent with OTUs. (PDF 14 kb)

Additional file 4: Figure S3. Power estimates for the linear model using the covariate, X_2, as correlated with OTUs. (PDF 14 kb)

Additional file 5: Figure S4. Power estimates for the logistic model using the covariate, X_2, as independent with OTUs. (PDF 15 kb)

Additional file 6: Figure S5. Power estimates for the logistic model using the covariate, X_2, as correlated with OTUs. (PDF 15 kb)

Additional file 7: Table S2. The names of the discovered microbial taxa using four methods to examine the sustained effects of LDP on microbial profiles. Discovered taxa without a name are excluded. (DOCX 15 kb)

Additional file 8: Table S3. The names of the discovered microbial taxa using four methods to examine the effects of birth mode on microbial profile. Discovered taxa without a name are excluded. (DOCX 14 kb)

Abbreviations
ECAM: Early childhood antibiotics and the microbiome; LDP: Low-dose penicillin; LEfSe: Linear discriminant analysis effect size; MiCAM: Microbiome comprehensive association mapping; MiRKAT: Microbiome regression-based kernel association test; MiSPU: Microbiome-based sum of powered score tests (aMiSPU: adaptive MiSPU); OMiAT: Optimal microbiome-based association test; OTU: Operational taxonomic unit; PAM: Partitioning-around-medoids; SD: Standard deviation; SPU: The sum of score powered tests (aSPU: adaptive SPU); UniFrac: Unique fraction

Acknowledgements
The authors are grateful to the anonymous reviewers for their insightful suggestions and comments.

Funding
The study was supported in part by NIH grants, R01 DK090989, R01DK110014, and U01CA182370, the Diane Belfer Program in Human Microbial Ecology, and the C&D Fund.

Availability of data and materials
We used two real microbiome datasets for the sustained effects on intestinal microbiota by early-life low-dose penicillin exposure [6] and the effects on intestinal microbiota by the early-life factor, vaginal or cesarean birth [7]. The former data is publicly available in the Sequence Read Archive (SRA) repository (https://www.ncbi.nlm.nih.gov/sra, SRA accession number: SRP042293), and the latter data is publicly available in the QIITA repository (https://qiita.ucsd.edu/, Study ID: 10249). The utilized synthetic data is the simulated OTU abundances as described in the "Simulation design" section in the "Results" section.
Our proposed methods, OMiAT and MiCAM, can be implemented in the software package, namely, OMiAT, in the R computing environment. Further information can be found at https://sites.google.com/site/huilinli09/software or https://github.com/hk1785/OMiAT.

Authors' contributions
HK contributed to the methodological ideas for OMiAT and MiCAM, performed the simulations and real data analyses, developed the software package, and wrote the manuscript. MJB contributed to the biological insights and interpretations. HL contributed to the methodological ideas for OMiAT and MiCAM, simulations, and real data analyses and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
This is not applicable: All utilized microbiome datasets are publicly available. No consent for publication was required for this study.

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.

Publisher's Note

Author details
[1]Department of Population Health and Environmental Medicine, New York University School of Medicine, New York, NY 10016, USA. [2]Department of Medicine and Microbiology, New York University Langone Medical Center, New York, NY 10010, USA.

Received: 11 October 2016 Accepted: 10 April 2017
Published online: 24 April 2017

References
1. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012;486(7402):215–21.
2. Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258–70.
3. Cho I, Yamanishi S, Cox L, Methé BA, Zavadil J, Li K, Gao Z, Mahana D, Raju K, Teitler I, Li H, Alekseyenko AV, Blaser MJ. Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature. 2012;488:621–6.
4. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490:55–60.
5. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L, et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Methods. 2013;19(5):576–85.
6. Cox LM, Yamanishi S, Sohn J, Alekseyenko AV, Leung JM, Cho I, Kim SG, Li H, Gao Z, Mahana D, et al. Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. Cell. 2013; 158(4):705–21.
7. Bokulich NA, Chung J, Battagila T, Henderson N, Jay M, Li H, D Lieber A, Wu C, Perez-Perez GI, Chen Y, Schweizer W, Zheng X, Contreras M, Dominguez-Bello MG, Blaser MJ. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med. 2016;8(343):343–82.
8. Mahana D, Trent CM, Kurtz ZD, Bokulich NA, Battaglia T, Chung J, Müller CL, Li H, Bonneau RA, Blaser MJ. Antibiotic perturbation of the murine gut microbiome enhances the adiposity, insulin resistance, and liver disease associated with high-fat diet. Genome Med. 2016;8(1):48.
9. Wu J, Peters BA, Dominianni C, Zhang Y, Pei Z, Yang L, Ma Y, Purdue MP, Jacobs EJ, Gapstur SM, Li H, Alekseyenko AV, Hayes RB, Ahn J. Cigarette smoking and the oral microbiome in a large study of American adults. ISME J. 2016;10(10):2435-46. doi:10.1038/ismej.2016.37.
10. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, Lewis BJ, Stahl D. Conservation of primary structure in 16S ribosomal RNA. Nature. 1975; 254:83–5.
11. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6(11):805–14.
12. Thomas T, Gilbert J, Meyer F. Metagenomics—a guide from sampling to data analysis. Microb Inform Exp. 2012;2:3.
13. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
14. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2016;444:1027–31.
15. Gao Z, Tseng C, Strober BE, Pei Z, Blaser MJ. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. PLoS One. 2008;3(7):e2719.
16. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakely BB, Parks DH, Robinson CJ, et al. Introducing mothur: open-source, platform independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.
17. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 2010;11:538.
18. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006;55(2):205–11.
19. Oakley BB, Fiedler TL, Marrazzo JM, Fredricks DN. Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis. Appl Environ Microbiol. 2008;74(15):4898–909.
20. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Appl. 2015;2:73–94.
21. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12:R60.
22. Parks DH, Tyson GW, Hugenholts P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics. 2014;30(21):3123–4.
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12):1–21.
24. Paulson JN, Stine OC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10(12):1200–2.
25. 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.

26. Wu C, Chen J, Kim J, Pan W. An adaptive association test for microbiome data. Genome Med. 2016;8(1):56. doi:10.1186/s13073-016-0302-3.
27. Lozupone CA, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12): 8228–35.
28. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007;73(5):1576–85.
29. 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.
30. Chen J, Li H. Kernel methods for regression analysis of microbiome composition data, Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association. New York: Springer; 1998. p. 191–201.
31. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014;197(4):1081–95.
32. Cox DR, Hinkley DV. Theoretical statistics. London: Chapman & Hall; 1974.
33. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009;33(6):497–507.
34. Basu B, Pan W. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol. 2011;35(7):606–19.
35. Lin D, Tang Z. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011;89(3):354–67.
36. Pan W, Han F, Shen X. Test selection with application to detecting disease association with multiple SNPs. Hum Hered. 2010;69:120–30.
37. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. Genetics. 1994;138(3):963–71.
38. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47(260):583–621.
39. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada D. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663.
40. Benjamini Y. Discovering the false discovery rate. J R Stat Soc B. 2010;70(4): 405–16.
41. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57(1):289–300.
42. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001;125:279–84.
43. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 2011;29(4):1165–88.
44. 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;5:474–504.
45. Sneath PHA, Sokal RR. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: Freeman; 1973.
46. Huson DH, Auch AF, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
47. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. F1000Research. 2016;5:1492.