RESEARCH

Open Access

Robust phylogenetic tree-based microbiome association test using repeatedly measured data for composition bias

Kangjin Kim[1] and Sungho Won[2,3,4*]

*Correspondence:
won1@snu.ac.kr

[1] Department of Applied Statistics, Gachon University, Seongnam, South Korea
[2] Institute of Health and Environment, Seoul National University, Seoul, South Korea
[3] Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, South Korea
[4] Interdisciplinary Program for Bioinformatics, College of Natural Science, Seoul National University, Seoul, South Korea

Abstract

Background: The effects of microbiota on the host phenotypes can differ substantially depending on their age. Longitudinally measured microbiome data allow for the detection of the age modification effect and are useful for the detection of microorganisms related to the progression of disease whose identification change over time. Moreover, longitudinal analysis facilitates the estimation of the within-subject covariate effect, is robust to the between-subject confounders, and provides better evidence for the causal relationship than cross-sectional studies. However, this method of analysis is limited by compositional bias, and few statistical methods can estimate the effect of microbiota on host diseases with repeatedly measured 16S rRNA gene data. Herein, we propose mTMAT, which is applicable to longitudinal microbiome data and is robust to compositional bias.

Results: mTMAT normalized the microbial abundance and utilized the ratio of the pooled abundance for association analysis. mTMAT is based on generalized estimating equations with a robust variance estimator and can be applied to repeatedly measured microbiome data. The robustness of mTMAT against compositional bias is underscored by its utilization of abundance ratios.

Conclusions: With extensive simulation studies, we showed that mTMAT is statistically relatively powerful and is robust to compositional bias. mTMAT enables detection of microbial taxa associated with host diseases using repeatedly measured 16S rRNA gene data and can provide deeper insights into bacterial pathology.

Keywords: Tree-based microbiome association test, Microbiota, Microbiome data, MTMAT

Background

Recent advancements in high-throughput technologies such as microarrays and next-generation sequencing have significantly elucidated the microbial world. For instance, microbiota has been found to play essential roles in the host by influencing energy homeostasis, body adiposity, blood sugar control, insulin sensitivity, hormone secretion and metabolic profiles. Moreover, ongoing research suggests potential associations between microbiota and complex diseases, including asthma, atopic dermatitis, obstructive sleep apnea, inflammatory bowel disease, and type-2 diabetes [1–6]. However, the composition of the microbiota varies from subject to subject, and the abundances of microbial taxa are often sparse with excessive zeros. Furthermore, the microbiota is affected by various factors, such as age and sex, which cause its abundance to vary considerably. Such sparsity in the data makes controlling type-1 and type-2 errors in statistical analyses difficult, and the inference of causal relationships through statistical analysis of microbiota data should be performed cautiously.

Repeatedly measured microbiota studies are useful for the detection of microorganisms related to the progression of disease and change identification over time, and provide more evidence for the causal relationship than cross-sectional studies [7]. Furthermore, the estimation of within-subject covariate effects is robust to between-subject confounders, and repeatedly measured microbiome data allow for the robust identification of microbiota effects on the risk of diseases in the host. Statistical analyses with repeatedly observed 16S rRNA gene require the adjustment of similarity among the measurements of the same subjects. However, only a limited number of methods are applicable in the repeated observation of 16S rRNA gene data, and the development of statistical methods for longitudinal studies is required for investigating the association between the human microbiome and diseases.

Xia et al. [8] comprehensively reviewed statistical methods of longitudinal data analysis in microbiome studies. These methods can be categorized into several categories: (1) standard longitudinal model, (2) overdispersed and zero-inflated longitudinal models, and (3) multivariate distance/kernel-based longitudinal models. First, the standard longitudinal model includes the linear mixed effect model (LMM) with generalized estimation equation (GEE) and generalized linear mixed effect model (GLMM) among others. LMM provides a standardized and flexible approach toward modeling both fixed and random effects. However, operational taxonomic unit (OTU) or amplicon sequence variant (ASV) abundances cannot address the sparsity issue and should be transformed or normalized to avoid the violation of distribution assumptions. Second, overdispersed and zero-inflated longitudinal models include the zero-inflated Gaussian (ZIG) mixture model, extensions of negative binomial mixed-effects (NBMM) [9], and zero-inflated negative binomial models (FZINBMM) [10]. The two-part zero-inflated beta regression model with random effects (ZIBR) extends the zero-inflated beta regression model to longitudinal data settings [11]. FZINBMM and ZIBR can analyze overdispersed and zero-inflated longitudinal metagenomics data. Finally, the multivariate distance/kernel-based longitudinal model includes the correlated sequence kernel association test (cSKAT) for continuous outcome, and the generalized linear mixed model and its data-driven adaptive test (GLMM-MiRKAT) for non-normally distributed outcome such as binary traits [12].

However, despite such developments, longitudinal microbiome data suffers from compositional bias, and only a few methods are robust to this problem. The magnitude of the sequence depth differs from subject to subject in metagenomics data, and the sums of absolute abundances for each subject are substantially different; therefore, relative abundances are generally used. However, this generates compositional bias, especially in longitudinally observed microbiome data [13], because relative abundances at each time point only provide a proportion of a taxon and it is therefore not possible to derive meaningful signals by comparing the relative abundance of the same subject over time. Furthermore, relative abundances are correlated among different taxa, and their adjustment is necessary if they were utilized as a response variable [14].

Several methods have been proposed to adjust the compositional bias. For instance, additive log-ratio (ALR) that uses a reference abundance for its denominator and centered log-ratio (CLR) that uses geometric mean can be considered. Network analysis, including sparse correlations for compositional data and sparse inverse covariance estimation for ecological association inference, can also be considered, modeling the whole community with a statistical model. However, they can only be applied to cross-sectional data, and none of them can be applied to analyzing longitudinally observed datasets [15].

In this article, we propose the phylogenetic tree-based microbiome association test for repeatedly measured data (mTMAT), which pools the abundance of OTU/ASVs based on the phylogenetic distance, thus resolving the problem of zero-inflation, making it robust to compositional bias. Through extensive simulation and real data analyses, we prove its robustness to compositional bias and its improved statistical power compared to that of other methods.

Methods

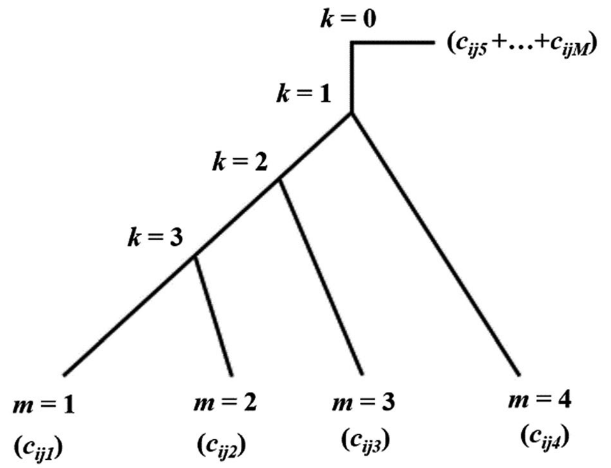
Phylogenetic tree

We apply similar notations and assumptions as in TMAT [16]. For this analysis, OTU/ASV profiling is performed across all subjects and time points simultaneously, with a rooted binary phylogenetic tree provided for these OTU/ASVs.

The first M_1 OTU/ASVs are categorized under a taxonomy of interest t, (where $t = 1$, … T), for studying associations with host diseases, while the remaining M–M_t OTU/ASVs are classified into different taxonomies. Within the genus of interest, there are M_t – 1 internal nodes and M_1 leaf nodes. Internal nodes are denoted by k (where $k = 1, \ldots, M_\mathrm{t} - 1$) and leaf nodes by m (where $m = 1, \ldots, M$). Each leaf node corresponds to a single OTU/ASV. If $m \leq M_\mathrm{t}$, it represents a leaf node within the genus of interest, otherwise m belongs to a different genus. The absolute abundance of OTU/ASV m of subject i at time point j is denoted by $c_{ijm}^{(t)}$. Assuming mutations occur during transmission from an internal node k to its child nodes, the relative abundance of leaf nodes of the left child node increases for cases where the mutation occurs during transmission to the left child, and decreases if it does so during transmission to the right child node. When assessing the association between an internal node k and a host disease, k and its associated leaf nodes are considered as the test nodes and test leaf nodes, respectively. The left and right test leaf nodes correspond to the leaf nodes of the left and right child nodes of k, and are denoted as L_k and R_k, respectively. Figure 1 provides a visual representation of theses definitions.

Quasi-likelihood

The log-transformed counts per million (CPM) $r_{ijm}^{(t)}$, which is used for the edgeR package (version 3.16.5) is expressed as follows.



$$M_l = 4, k \in \{0,1,2,3\}, m \in \{1,2,3,4\}$$
$$k = 0 \Rightarrow L_0 = \{m|m = 1,2,3, \text{or } 4\}, R_0 = \{m|m = 5,\ldots,M\}$$
$$k = 1 \Rightarrow L_1 = \{m|m = 1,2, \text{or } 3\}, R_1 = \{m|m = 4\}$$
$$k = 2 \Rightarrow L_2 = \{m|m = 1, \text{or } 2\}, R_2 = \{m|m = 3\}$$
$$k = 3 \Rightarrow L_3 = \{m|m = 1\}, R_3 = \{m|m = 2\}$$

Fig. 1 Examples of rooted binary phylogenetic trees

$$r_{ijm}^{(t)} = log_2\left(\frac{c_{ijm}^{(t)} + \frac{c_{ij\cdot}^{(t)}}{2}}{\sum_{m=1}^{M} c_{ijm}^{(t)} + c_{ij\cdot}^{(t)}} \times 10^6 + 1\right).$$

$x_{ij}^{(k)}$, where k=1, ..., M_1−1, is expressed as follows

$$x_{ij}^{(k)} = log\left(\frac{C_{ij}^{(k)}}{D_{ij}^{(k)}}\right), C_{ij}^{(k)} = \sum_{m=1}^{M_t} r_{ijm}^{(t)} \cdot I(m \in L_k), D_{ij}^{(k)} = \sum_{m=1}^{M_t} r_{ijm}^{(t)} \cdot I(m \in R_k).$$

Here, I(A) indicate the indicator function. As all OTU/ASVs in taxonomy t can be associated with the host disease, x_i^0 for such a case is expressed as follows:

$$x_{ij}^{(0)} = log\left(\frac{E_{ij}^{(t)}}{G_{ij}}\right), E_{ij}^{(t)} = \sum_{m=1}^{M_t} r_{ijm}^{(t)}, G_{ij} = \sum_{m=M_t+1}^{M} r_{ijm}^{(t)}.$$

To facilitate comprehension, we assume that$x^{(0)}$, ..., and $x^{(M_1-1)}$ are ordered according to the sequential sequence of internal nodes commencing from the root node, as illustrated in Fig. 1. $x^{(0)}$ is used to test the association of all OTU/ASVs belonging to the genus of interest by pooling them, and $x^{(1)}$ is used for testing the root node of the phylogenetic tree. The phenotype of subject i at time point j is denoted by y_{ij} and is coded as 1 and 0 for cases and controls, respectively. To address potential arguments about how to best represent all other microbial abundances relative to the genus of interest, we considered two additional options for the value ofG_{ij}, both of which account for compositional bias. The first one uses $E_{ij}^{(t)}$ value of a reference taxon, while the second one uses the geometric mean $\left(\prod_{t=1}^{T} E_{ij}^{(t)}\right)^{\frac{1}{T}}$ for all taxonomies. These two options correspond to the ALR and CLR transformations, respectively. The vectors and matrices for testing the association of the genus of interest are expressed as follows:

$$\mathbf{x}_i^{(k)} = \begin{pmatrix} x_{i1}^{(k)} \\ \vdots \\ x_{iN_i}^{(k)} \end{pmatrix}, \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iN_i} \end{pmatrix}$$

The stacked observations across all subjects are represented as

$$\mathbf{x}^{(k)} = \begin{pmatrix} \mathbf{x}_1^{(k)} \\ \vdots \\ \mathbf{x}_N^{(k)} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{x}^{(0)} & \cdots & \mathbf{x}^{(M_1-1)} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}.$$

Based on such matrix notations, we provide the quasi-likelihood for repeatedly observed 16S rRNA gene data. If we denote R_i and σ_{kk} as a working correlation matrix and overdispersion parameter, respectively, and define D_{ik} as a diagonal matrix with its diagonal entries being $var(x_{ij}^{(k)}), j=1,...,N_i$, the covariance matrix for the observations of subject i is expressed as follows:

$$\boldsymbol{\Sigma}_i^{(k)} = \sigma_{kk} \boldsymbol{D}_{ik}^{1/2} \boldsymbol{R}_i \boldsymbol{D}_{ik}^{1/2}.$$

Then, the covariance matrix $\mathbf{\Sigma}^{(k)}$ can be expressed as follows:

$$\boldsymbol{\Sigma}^{(k)} = \begin{pmatrix} \boldsymbol{\Sigma}_1^{(k)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_N^{(k)} \end{pmatrix}.$$

If we let $\mathbf{Z}$ be a design matrix for p covariates, including the intercept, we assume

$$E\left(\boldsymbol{x}^{(k)}|\boldsymbol{Z},\boldsymbol{y}\right) = \boldsymbol{Z}\boldsymbol{\alpha}_k + \boldsymbol{y}\beta_k, var\left(\boldsymbol{x}^{(k)}|\boldsymbol{Z},\boldsymbol{y}\right) = \mathbf{\Sigma}^{(k)}, k = 0,\ldots,M_1-1.$$

Therefore, quasi-score functions for $\boldsymbol{\alpha}_k$ and β_k can be expressed as follows:

$$U(\boldsymbol{\alpha}_k,\beta_k) = \begin{pmatrix} \boldsymbol{U}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_k,\beta_k) \\ U_\beta(\boldsymbol{\alpha}_k,\beta_k) \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}^t\left(\mathbf{\Sigma}^{(k)}\right)^{-1}\left(x^{(k)} - \boldsymbol{Z}\boldsymbol{\alpha}_k - \boldsymbol{y}\beta_k\right) \\ \boldsymbol{y}^t\left(\mathbf{\Sigma}^{(k)}\right)^{-1}\left(x^{(k)} - \boldsymbol{Z}\boldsymbol{\alpha}_k - \boldsymbol{y}\beta_k\right) \end{pmatrix}.$$

Quasi-fisher information can be expressed as follows:

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{U}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \boldsymbol{U}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \\ \boldsymbol{U}_{\boldsymbol{\beta}\boldsymbol{\alpha}} & \boldsymbol{U}_{\boldsymbol{\beta}\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}^t\left(\mathbf{\Sigma}^{(k)}\right)^{-1}\boldsymbol{Z} & \boldsymbol{Z}^t\left(\mathbf{\Sigma}^{(k)}\right)^{-1}\boldsymbol{y} \\ \boldsymbol{y}^t\left(\mathbf{\Sigma}^{(k)}\right)^{-1}\boldsymbol{Z} & \boldsymbol{y}^t\left(\mathbf{\Sigma}^{(k)}\right)^{-1}\boldsymbol{y} \end{pmatrix}.$$

Score test with small sample adjustment

The null hypothesis can be expressed as follows:

$$H0:\mathbf{L}\begin{bmatrix}\boldsymbol{\alpha}_k\\ \beta_k\end{bmatrix}=0$$

where $\mathbf{L}$ is a matrix of linear constraints with c rows and number of columns equal to the length of $\begin{bmatrix}\boldsymbol{\alpha}_k\\ \beta_k\end{bmatrix}$ and $\mathbf{0}$ is the zero vector of matching dimension. To test the null hypothesis, H_0: $\beta_k=0$, the generalized score statistics by Boos can be applied [17] by setting $\mathbf{L}$ = $\begin{bmatrix}0^t & 1\end{bmatrix}$ as follows [18]:

$$T_k=\boldsymbol{U}(\boldsymbol{\alpha}_k,0)^t\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^t\left(\boldsymbol{L}\widetilde{\boldsymbol{H}}^{-1}\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^t\right)^{-1}\boldsymbol{L}\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{U}(\boldsymbol{\alpha}_k,0)\sim\chi^2(df=1)$$

where

$$\tilde{\boldsymbol{H}}=\sum_{i=1}^{N}-\boldsymbol{D}_i^t\sum_{i}^{-1}\boldsymbol{D}_i,\boldsymbol{D}_i=\begin{bmatrix}\boldsymbol{Z}_i & \boldsymbol{y}_i\end{bmatrix},$$

$$\tilde{\mathbf{B}}=\sum_{i=1}^{N}\mathbf{U}_i(\boldsymbol{\alpha}_k,0)\mathbf{U}_i(\boldsymbol{\alpha}_k,0)^t.$$

To adjust the small sample bias, $\widetilde{\mathbf{B}}$ is further updated as follows [19]:

$$\widetilde{\mathbf{B}}_{adj}=\sum\nolimits_{i=1}^{N}\mathbf{D}_i^t\sum\nolimits_{i}^{-1}\left(\mathbf{I}_i-\widetilde{\mathbf{P}}_{ii}\right)^{-1}\mathbf{S}_i\mathbf{S}_i^t\left(\mathbf{I}_i-\widetilde{\mathbf{P}}_{ii}^t\right)^{-1}\sum\nolimits_{i}^{-1}\mathbf{D}_i$$

where

$$\mathbf{S}_i=\mathbf{x}_i^{(k)}-\left(\mathbf{Z}_i\boldsymbol{\alpha}_k+\mathbf{y}_i\beta_k\right)$$

$$\widetilde{\mathbf{P}}_{ii}=\mathbf{D}_i\left(\mathbf{I}-\widetilde{\mathbf{H}}^{-1}\mathbf{L}^t\left(\mathbf{L}\widetilde{\mathbf{H}}^{-1}\mathbf{L}^t\right)^{-1}\mathbf{L}\right)\widetilde{\mathbf{H}}^{-1}\mathbf{D}_i^t\sum\nolimits_{i}^{-1}$$

Wald test with small sample adjustment

The Wald statistic with sandwich estimator with correction of small sample bias was considered [18]. β_k can be estimated by solving the estimating equation, $U_\beta(\boldsymbol{\alpha}_k,\beta_k)=0$, as follows:

$$\hat{\beta}_k=\left(\mathbf{y}^t\left(\hat{\Sigma}^{(k)}\right)^{-1}\mathbf{y}\right)^{-1}\left(\mathbf{y}^t\left(\hat{\Sigma}^{(k)}\right)^{-1}\left(\mathbf{x}^{(k)}-\mathbf{Z}\hat{\alpha}_k\right)\right).$$

For the estimation of the variance of $\widehat{\beta}_k$, we consider the robust variance estimator with a small sample adjustment as follows

$$\widehat{V}_k=\left(\sum\nolimits_{i=1}^{N}\mathbf{y}_i^t\left(\widehat{\boldsymbol{\Sigma}}_i^{(k)}\right)^{-1}\mathbf{y}_i\right)^{-1}\widehat{B}_{adj}\left(\sum\nolimits_{i=1}^{N}\mathbf{y}_i^t\left(\widehat{\boldsymbol{\Sigma}}_i^{(k)}\right)^{-1}\mathbf{y}_i\right)^{-1}$$

where

$$\widehat{B}_{adj} = \sum_{i=1}^{N} \mathbf{y}_i^t \left(\widehat{\boldsymbol{\Sigma}}_i^{(k)}\right)^{-1} \widehat{Cov\left(\mathbf{x}_i^{(k)}\right)}_{robust} \left(\widehat{\boldsymbol{\Sigma}}_i^{(k)}\right)^{-1} \mathbf{y}_i.$$

$$\widehat{Cov\left(\mathbf{x}_i^k\right)}_{robust} = \left(\mathbf{I}_{N_i} - \widehat{\mathbf{P}}_{ij}\right)^{-1}\left(\mathbf{x}_i^{(k)} - \mathbf{Z}_i\widehat{\boldsymbol{\alpha}}_k\right)\left(\mathbf{x}_i^{(k)} - \mathbf{Z}_i\widehat{\boldsymbol{\alpha}}_k\right)^t\left(\mathbf{I}_{N_i} - \widehat{\mathbf{P}}_{ij}\right)^{-1}$$

$$\hat{\mathbf{P}}_{ij} = \mathbf{y}_i \left(\sum_{i=1}^{N} \mathbf{y}_i^t \left(\hat{\Sigma}_i^{(k)}\right)^{-1} \mathbf{y}_i\right)^{-1} \mathbf{y}_j^t \left(\hat{\Sigma}_j^{(k)}\right)^{-1}.$$

Based on this result, the robust Wald statistic of β_k for test node k is expressed as follows:

$$T_{k,wald} = \hat{\beta}_k^t \left(\hat{V}_{\mathbf{k}}\right)^{-1} \hat{\beta}_k = \frac{\hat{\beta}_k^2}{\hat{V}_{\mathbf{k}}} \sim \chi^2(df = 1)\ under\ H_0.$$

mTMAT

Following the approach described in Kim et al. (2020), we combined statistics to test the null hypothesis $H_0 : \beta_0 = \beta_1 = \cdots = \beta_{M_1-1} = 0$ using the minimum p-value method. If p-values for T_k are denoted by pT_k, the proposed mTMAT$_M$ statistics are defined as:

$$mTMAT_M = min\left\{ pT_0, \cdots, pT_{M_1-1} \right\}.$$

The p-values $pT_0, \cdots, pT_{M_1-1}$ are asymptotically independent, as shown in previous studies [16]. Therefore, under H_0 the distribution of mTMAT$_M$ follows a beta distribution with parameters (1, M$_1$).

If the sample size is small, the normality of T_k under H_0 may not be achieved, and the assumption of the quasi-score test can be violated. If we apply the inverse normal transformation to $x_{ij}^{(k)}$, then the same statistics can be obtained. This is denoted by T_k^{INT}. Rank-based inverse normal transformation with an adjustment parameter of 0.5 was used for the transformation, and data with tie values were mapped to the same value in the transformed data [20]. Then, mTMAT$_{IM}$ is expressed as follows:

$$mTMAT_{IM} = min\left\{ pT_0^{INT}, \cdots, pT_{M_1-1}^{INT} \right\} \sim beta(1, M_1) under H_0.$$

KARE cohort data

The Korea Association REsource (KARE) cohort is a prospective study cohort involving subjects from the rural community of Ansung and the urban community of Ansan in South Korea. It began in 2001 as part of the Korean Genome Epidemiology study [21]; we used data from 2,072 urine samples from 691 participants in 2013, 2015, and 2017. Their 16S rRNA gene amplicon sequencing data used in the study were obtained from the NCBI Sequence Read Archive database under project accession number PRJNA716550 [22]. For paired-end sequencing of the V3-V4 region of the bacterial 16S rRNA gene, the widely used primers 16S_V3_F (5'- TCGTCGGCAGCGTCAGATGTG

TATAAGAGACAG-CCTACGGGNGGCWGCAG-3') and 16S_V4_R (5'-GTCTCG TGGGCTCGGAGATGTGTATAAG-AGACAGGACTACHVGGGTATCTAATCC-3') were used. Adaptor sequences were detected and removed using the CUTADAPT software (version 4.4) with a minimum overlap of 11 bp, maximum error rate of 10%, and a minimum length of 10 bp [23]. Sequences were merged using CASPER (version 0.8.2) with a mismatch ratio of 0.27 and filtered by the Phred (Q) score, resulting in sequences of 350–550 bp in length [24, 25]. After the merged sequences were dereplicated, chimeric sequences were detected and removed using VSEARCH (version 2.3.4) and the Silva Gold reference database for chimeras [26]. The open-reference Operational Taxonomic Unit (OTU) picking was based on the EzTaxon database using UCLUST (version 1.2.22) with a 97% sequence identity threshold [27, 28]. The phylogenetic trees based on EzTaxon database were obtained through the SINA method [29] using the reference sequences available from the EzTaxon database. We calculated the proportion of each species among the total taxa and determined the mean value across all subjects. If the resulting value was < 0.001, the species was excluded [30]. A histogram of the read count distribution is shown in Additional file 1: Fig. S1. Among the 691 subjects, those with a read count of < 3,000 or for whom genomic data were not available in any phase were excluded. As a result, 1179 samples from 393 subjects, including 70 genera, were used for the simulation analysis.

Simulation studies

We conducted extensive simulations to evaluate the performance of mTMAT with two types of datasets. One dataset with 393 subjects who participated in all three phases from KARE cohort and the generative dataset based on microbiomeDASim [31]. For the KARE cohort dataset, we defined disease status based on a creatinine level threshold of 1.15, with values above 1.15 as cases and others as controls; however, this specific threshold was not critical, as the focus was on maintaining an appropriate case–control ratio. We followed a methodology similar to that described in Kim et al. (2020) to evaluate the effect of disease status permutation on the identification of causal taxa, which in this context refers to taxa designated as associated with disease within the simulation framework. The case-to-control ratios were assumed to be 1:1 or 1:3 depending on the simulation scenario, and for the 1:3 ratio, cases were rounded down, and controls were rounded up to maintain the total sample size. Briefly, the disease status of subjects was permuted, and a single test node was randomly selected from the internal nodes of the phylogenetic tree. From the leaf nodes descending from this test node, causal taxa were chosen randomly at three different levels: a single taxon, 50% of the taxa, or 90% of the taxa, corresponding to p = 1 taxon, 50%, and 90%, respectively. Notably, when p = 1, indicating a single taxon associated with the disease, the phylogenetic structure does not offer additional information for mTMAT.

For the simulation, the sample variances of $c_{ijm}^{(t)}$ for causal taxa were denoted by $\widehat{\sigma}_{mm}^{(t)}$. For affected subjects, an increment $\delta = \beta\widehat{\sigma}_{mm}^{(t)}$ was added to the observed absolute abundances of the selected causal taxa, where β values were set to 0, 0.01, 0.02, or 0.04. The absolute abundances of non-causal taxa remained unchanged. The case with $\beta = 0$ was used to estimate empirical type-1 error rates, while non-zero β values were employed for statistical power estimation. Type-1 error rates were assessed at significance levels of 0.1, 0.05, 0.01, and 0.005 using 5,000 replicates, while empirical power was evaluated at the 0.05 significance level with 500 replicates. Unless described otherwise, these replicates were used throughout the study. Type-2 error, defined as 1 − power, was therefore indirectly addressed through the power analysis.

For comparison with $\mathrm{mTMAT_M}$ and $\mathrm{mTMAT_{IM}}$, GLMM-MiRKAT (version 1.2), FZINBMM (version 1.0), linear mixed model (LMM) with arcsine square root transformation (LMM-arcsine), and LMM with log transformation (LMM-log) with nlme package (version 3.1) were considered. Additionally, phylogenetic tree-based microbiome association test (TMAT, version 1.01), optimal microbiome regression-based kernel association test (oMiRKAT, version 0.02), adaptive microbiome-based sum of powered score (aMiSPU, version 1.0), and the Wilcoxon test were included to compare how cross-sectional methods perform in the context of repeatedly measured data. This allows for benchmarking $\mathrm{mTMAT_M}$ and $\mathrm{mTMAT_{IM}}$ against methods that do not explicitly account for temporal correlations, providing insights into the importance of considering longitudinal structures in microbiome association studies. Association analyses were conducted at the genus level. FZINBMM, LMM models, and Wilcoxon were applied by pooling all species within each genus. Each genus consisted of multiple species, and oMiRKAT and aMiSPU were applied to species belonging to each genus.

For $\mathrm{mTMAT_M}$ and $\mathrm{mTMAT_{IM}}$, robust wald and score statistics with four different choices of working correlation matrix: identity, compound symmetry (CS), first-order autoregressive (AR1) and unstructured (UN), were considered. For aMiSPU and oMiRKAT, we followed the same parameters to Kim et al. (2020) using 500 and 5,000 permuted replicates to assess power and type-1 error rates, respectively. For GLMM-MiRKAT, Unifrac distance which is default choice is used for its distance metrics. Unifrac distance requires observed read counts, and therefore, samples with zero read counts were excluded from the analyses involving GLMM-MiRKAT, oMiRKAT, and aMiSPU. Additionally, these methods cannot handle genera consisting of a single taxon, so such cases were omitted from the statistical power analyses for these genera.

Following the simulation with the generated dataset with microbiomeDASim, Identity, CS, and AR1 with different parameter values are assumed for the simulation, and type-1 error estimates were compared for different uses of working correlation matrices for $\mathrm{mTMAT_{IM}}$. The mean value of relative abundance and proportion of zero count samples were estimated from the KARE cohort study for all the genera; the genera with first quantile, median, and third quantile sparsity level were selected for the simulation. The values were 52%, 64%, and 73%.

We also evaluate the robustness of the proposed method to the compositional bias. Different choices of G_{ij} corresponding to ALR and CLR are also considered. The KARE dataset was simulated 2000 times with simulation parameters N and the ratios of the cases and controls were 50 and 1:3, respectively. Then a genus containing more than one species was selected and assumed to be associated with a phenotype with $\beta = 0.15$ and $p = 50\%$. Additionally, a species outside the selected genus was randomly chosen and assumed to be associated with the same phenotype using the same β value. To evaluate how changes in the abundance of this external taxon influence the bias estimate for the selected genus, its abundance was increased by its standard deviation multiplied by factors of 0, 1, 5, 10, 50, and 100. Then the mean and interquartile range of bias estimate of the selected genus was calculated and compared with a different value of the multiplier.

Pregnant microbiome data

We used publicly available datasets from Romero [32], who conducted a retrospective case–control longitudinal study that included non-pregnant women (n = 32) and pregnant women who delivered at term (38 to 42 weeks) without complications (n = 22) using pplacer and version 0.2 of the vaginal community 16S rRNA gene database [33] for the taxonomic classification. The neighbor joining method based on the Bray–Curtis dissimilarity index was used to obtain a phylogenetic tree [34]. The pregnant dataset includes data on the race, days after the first visit (GDColl), household income, maternal education, and baby gender.

Results

Results from simulated data

The performances of $\mathrm{mTMAT}_{\mathrm{M}}$ and $\mathrm{mTMAT}_{\mathrm{IM}}$ were evaluated using simulated data. Additional file 1: Fig. S2 shows the overall distribution of microbial composition. Results for $\mathrm{mTMAT}_{\mathrm{M}}$ are presented in Additional file 1: Table S1, which indicates that inflation of type-1 error rates occurred as the number of case samples and total samples increased. Conversely, $\mathrm{mTMAT}_{\mathrm{IM}}$ adequately preserved the nominal type-1 error with a slight inflation when unstructured correlation (Table 1 and Additional file 1: Table S2).

GLMM-MiRKAT, FZINBMM and LMM models are designed to be used as longitudinal microbiome data and can be compared with $\mathrm{mTMAT}_{\mathrm{IM}}$ and $\mathrm{mTMAT}_{\mathrm{M}}$. FZINBMM and GLMM-MiRKAT could not preserve type-1 error rates with extremely high type-1 error estimates for FZINBMM. GLMM-MiRKAT suffered singular matrix problem during calculating the test statistics (Table 1). In this case, the resulting p-values were excluded for the estimation of type-1 error rate and power.

We also evaluated the effect of the number of leaf nodes (Additional file 1: Table S3), and the results showed that $\mathrm{mTMAT}_{\mathrm{M}}$ became slightly conservative when the number of leaf nodes exceeded 5, but that $\mathrm{mTMAT}_{\mathrm{IM}}$ was less affected. The result with more than 15 leaf nodes can be dependent on specific genera chosen with a small number of genera.

Additional file 1: Table S6 evaluated the effect of sparsity on the type-1 error rate using an approach same to that described in Kim et al. (2020). The results indicate that the type-1 error rates for FZINBMM were the most inflated, with some inflation also observed for GLMM-MiRKAT when mean sparsity exceeded 20%. Although GLMM-MiRKAT employs permutation-based p-values, which are generally robust to non-normality, their validity can be affected by heteroscedasticity. A high degree of sparsity may introduce heteroscedasticity, contributing to the observed type-1 error inflation. While there was some inflation noted for $\mathrm{mTMAT}_{\mathrm{M}}$, the type-1 error rates for $\mathrm{mTMAT}_{\mathrm{IM}}$ were well-controlled.

The effect of the assumed correlation matrix for different scenarios was evaluated with the use of microbiomeDASim [31]. Identity, CS, and AR1 with different parameter

Table 1 Type-1 error estimates of mTMAT$_{IM}$ and other statistical methods from repeatedly measured microbiome data at three time points at the significance levels 0.1, 0.05

Method	Working correlation matrix	α = 0.1						α = 0.05					
		Case: Control = 1: 1			Case: Control = 1: 3			Case: Control = 1: 1			Case: Control = 1: 3		
		N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100
mTMAT$_{IM}$	Identity	0.0884	0.0963	0.0933	0.0999	0.0921	0.1041	0.0413	0.0493	0.0451	0.0463	0.0458	0.0495
mTMAT$_{IM}$	CS	0.0882	0.0957	0.0915	0.1004	0.0938	0.1039	0.0421	0.0503	0.0454	0.0446	0.0451	0.0498
mTMAT$_{IM}$	AR1	0.0893	0.0959	0.0923	0.0990	0.0943	0.1036	0.0413	0.0499	0.0454	0.0460	0.0449	0.0498
mTMAT$_{IM}$	UN	0.0901	0.0954	0.0905	0.1003	0.0941	0.1039	0.0397	0.0478	0.0447	0.0476	0.0453	0.0499
GLMM-MiRKAT		0.1320	0.1377	0.1301	0.1450	0.1466	0.1347	0.0914	0.0889	0.0875	0.0922	0.0974	0.0846
FZINBMM		0.4882	0.4548	0.4770	0.4778	0.4625	0.4411	0.4209	0.3915	0.4162	0.4055	0.3908	0.3786
LMM-arcsin		0.1094	0.1328	0.1482	0.0926	0.0982	0.0994	0.0537	0.0681	0.0804	0.0449	0.0497	0.0467
LMM-log		0.1021	0.1115	0.1223	0.0922	0.0934	0.0905	0.0466	0.0529	0.0613	0.0421	0.0475	0.0436
TMAT$_{IM}$		0.0988	0.1085	0.1235	0.1342	0.0927	0.0929	0.0489	0.0580	0.0660	0.0613	0.0448	0.0456
TMAT$_{M}$		0.0978	0.1132	0.1289	0.1336	0.1042	0.0944	0.0463	0.0564	0.0642	0.0660	0.0505	0.0448
Wilcoxon		0.1034	0.1174	0.1337	0.1320	0.0947	0.0938	0.0496	0.0558	0.0635	0.0631	0.0488	0.0449
oMiRKAT		0.1041	0.1141	0.1298	0.1308	0.0950	0.0945	0.0509	0.0514	0.0585	0.0678	0.0527	0.0436
aMiSPU		0.0839	0.0973	0.1108	0.1056	0.0800	0.0691	0.0435	0.0471	0.0536	0.0568	0.0426	0.0335

The ratios between cases and controls were assumed to be 1:1 and 1:3. The total sample size is denoted by N, and we considered N = 30, 50, and 100. For a 1:3 ratio, cases were rounded down and controls were rounded up to maintain the total sample size. All subjects were selected without replacement. Type-1 error estimates were calculated with 5,000 replicates at the significance levels 0.1, 0.05

values are assumed for the simulation with different uses of working correlation matrix, robust score statistic and for mTMAT$_{IM}$ (Additional file 1: Table S4). The result shows that mTMAT$_{IM}$ preserved type-1 error for all the scenarios.

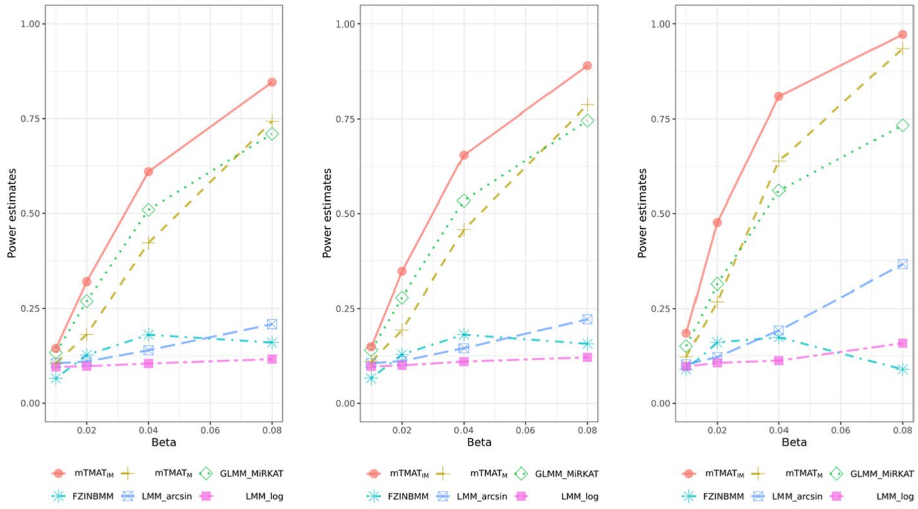
To further assess the type-1 error performance of mTMAT$_{IM}$ with larger sample sizes, we conducted additional simulations for $N=100$, 200 and 400. These simulations confirmed that type-1 error rates remain well-controlled and do not exhibit inflation as N increases (Additional file 1: Table S5).

We also calculated statistical power estimates with 2,000 replicates at the 0.05 significance level and compared them with those of other statistical methods. The significance levels for each method were adjusted based on the statistics from the simulation to calculate type-1 error for a valid performance comparison. The threshold is determined as the percentiles of the p-values calculated in the type-1 error simulation under null hypothesis. We considered genera comprising two or more taxa. In Fig. 2, mTMAT$_{IM}$ usually outperformed the other methods. The performance of GLMM-MiRKAT was comparable with mTMAT$_{M}$. FZINBMM and LMM-log exhibited significantly less power than other methods.

Additional file 1 Fig. S3 shows the results of genera comprising one or more taxa. GLMM-MiRKAT can only be calculated if more than one taxon is available. Therefore, it was excluded from this comparison. mTMAT$_{M}$ and mTMAT$_{IM}$ can be applied in such scenarios, and the results showed that the proposed method was the most efficient.

Comparison with methods for cross-sectional analysis (Additional file 1: Fig. S4) shows that TMAT$_{M}$, TMAT$_{IM}$, and mTMAT$_{IM}$ showed high statistical power. aMiSPU had the highest power estimate when beta was 0.02.

As shown in Additional file 1: Fig. S5, we evaluated the statistical power in relation to the number of leaf nodes. Keeping the number of causal taxa constant, we observed



(A) p = 1 OTU/ASV B) p = 0.5 (C) p = 0.9

Fig. 2 Power estimates for genera consisting of more than one taxon. Power estimates at the significance level 0.05 were calculated with 500 replicates. We generated simulation data based on read counts from datasets and considered genera with more than one taxon. For statistical methods whose type-1 errors are violated, their P-values were adjusted with the simulated data, which makes all statistical methods preserve the nominal significance level. We assumed that the total sample size (N) is equal to 50, the ratio of cases and controls is set to be 1:3. Identity working correlation matrix and robust score statistic were used for mTMAT

that the statistical power estimates decreased as the number of leaf nodes increased. Optimal performance was observed for mTMAT$_{\mathrm{IM}}$; however, GLMM-MiRKAT exhibited a high level of type-1 error rate with the leaf node ranging between 6 and 15. The sparsity levels were also considered, and statistical power estimates were compared. Power estimates were maximized at the middle-group sparsity level (Additional file 1: Fig. S6). The type-1 error and the power of group and time and interaction effect are also estimated using simulation dataset with microbiomeDASim package. The type-1 error rates for the group, time (Additional file 1: Table S7), and interaction effects (Additional file 1: Table S8) are well controlled for mTMAT$_{\mathrm{IM}}$ and mTMAT$_{\mathrm{M}}$, while LMM-arcsine and LMM-log exhibit type-1 error inflation. We verified that mTMAT$_{\mathrm{IM}}$ and mTMAT$_{\mathrm{M}}$ can successfully capture group and time effects with the other included as a covariate. Both methods also detected interaction effects with adequate control of type-1 error rates. Although LMM-arcsine and LMM-log showed higher power than mTMAT$_{\mathrm{IM}}$ and mTMAT$_{\mathrm{M}}$ in this simulation scenario, their reliability is limited due to the observed type-1 error inflation.

Additional file 1: Fig. S7 shows the effect of compositional bias. mTMAT$_{\mathrm{IM}}$ and FZINBMM had a smaller bias than other methods when the level of compositional bias was high. All three types of mTMAT$_{\mathrm{IM}}$ had smaller interquartile ranges of bias than FZINBMM, indicating that mTMAT$_{\mathrm{IM}}$ successfully mitigated compositional bias compared to other methods. Furthermore, the default G_{ij} option for mTMAT$_{\mathrm{IM}}$ consistently outperformed the ALR and CLR options in terms of reducing both bias and variability, highlighting its robustness in addressing compositional bias in the dataset.

In summary, we confirmed that mTMAT$_{\mathrm{IM}}$ is generally the most efficient method among those available in our simulations. mTMAT$_{\mathrm{IM}}$ considers phylogenetic tree structures, uses log CPM transformation, correction of compositional bias by taking a proportion among taxa, and considers correlations among repeatedly measured samples, which makes it superior to other methods. The overall power comparison results for cross-sectional methods are consistent among previous studies on TMAT [16]; however, the type-1 error rate for TMAT was inflated with longitudinal microbiome data. GLMM-MiRKAT is the second most powerful but failed to preserve type-1 rates and cannot be applied in analysis with a single taxon. Furthermore,

GLMM-MiRKAT is based on oMiRKAT, and they both used the kernel method and permutation approaches, which can be very computationally intensive, especially if the sample size increases [16].

Real data analysis

The pregnant datasets were analyzed with mTMAT, GLMM-MiRKAT, FZINBMM, and LMM with the arcsine square root transformation and LMM with log transformation to explore the association between microbiota and the pregnancy status of the women. The pregnant dataset includes the race, days after the first visit (GDColl), household income, maternal education, and gender of baby. The overall composition is presented in Additional file 1: Fig. S8; the overall composition change was clear after > 300 days. Additional file 1: Fig. S9 shows that the change can be related to the pregnancy state. PERMANOVA analysis result shows the associated phenotype that explained microbiome variability. Race was the covariate with the least p-value of 0.06 (Additional file 1:

Table 2 Association analysis results of the Pregnant dataset

Family taxonomy	Genus taxonomy	mTMAT_{IM}	mTMAT_{M}	GLMM-MiRKAT	FZINBMM	LMM arcsine	LMM log
Lactobacillaceae	*Lactobacillus*	0.01984	0.01549	0.00416	0.00000	0.00000	0.00000
Tissierellaceae	*Anaerococcus*	0.02110	0.02275	0.01247	0.00000	0.00000	0.00000
Tissierellaceae	*Peptoniphilus*	0.02599	0.02110	0.00416	0.00000	0.00000	0.00000
Veillonellaceae	*Dialister*	0.01984	0.02110	0.06983	0.00000	0.00000	0.00000
Campylobacteraceae	*Campylobacter*	0.01984	0.02275	NA	0.00000	0.00000	0.00000
Tissierellaceae	*Finegoldia*	0.04797	0.04647	NA	0.00000	0.00000	0.00001
Porphyromonadaceae	*Porphyromonas*	0.13031	0.10538	NA	0.00000	0.00000	0.00000
Streptococcaceae	*Streptococcus*	0.01984	0.02275	NA	0.00000	0.00000	0.00000
Actinomycetaceae	*Varibaculum*	0.55456	0.34057	NA	0.00001	0.28831	0.93412
Prevotellaceae	*Prevotella*	0.01984	0.02110	NA	0.00000	0.00000	0.00000
Tissierellaceae	*1–68*	0.84255	0.67380	NA	0.00006	0.00165	0.00171
Lactobacillaceae	*WAL_1855D*	0.09401	0.05409	NA	0.00000	0.00000	0.00000
Tissierellaceae	*Mobil uncus*	0.01984	0.02110	NA	0.00000	0.00000	0.00000
Tissierellaceae	*Atopobium*	0.01984	0.02522	NA	0.00000	0.00000	0.00000
Bifidobacteriaceae	*Gardnerella*	0.99925	0.79316	NA	0.00014	0.53031	0.59088
Actinomycetaceae	*Actinomyces*	0.01984	0.02110	NA	0.00000	0.00000	0.00000

GLMM-MiRKAT requires more than one taxon for calculation; therefore, genera with a single taxon show 'NA' for this method

Fig. S10). Table 2 shows that mTMAT_{IM} found 11 significant genera. FZINBMM, LMM-arcsine, and LMM-log found 16, 14, and 14 significant genera, respectively. As shown in the simulation study, most of the detected genera as significant only by FZINBMM can be false positives.

mTMAT_{IM} shared most of the significant genera with other methods. The most significant genera was *Lactobacillus*, which is consistent with the findings of the original study [32]. Additional file 1: Fig. S11 shows a Venn diagram comparing the numbers of significant genera implicated by the various applied methods. As LMM-arcsine and LMM-log differ only in their transformation, the methods shared all 16 detected genera. FZINBMM detected two more genera that were not detected by other methods. mTMAT_{IM} shared all the 12 detected genera with other methods. Results for genera significantly associated with at least one method at the FDR-adjusted 0.05 significance level are summarized.

Figure 3 shows the distribution of taxa under *Lactobacillus*. Lactobacillus has five leaf nodes and the relative abundances of all the leaf node m = 1, 2, 3, 4 and 5 were higher in the pregnant group. *Lactobacillus* was observed to be more abundant in the pregnant group than in the control group. This finding aligns with existing research indicating that *Lactobacillus* species are more abundant in pregnant women. Notably, a deficiency in vaginal *Lactobacillus* species has been linked to an increased risk of preterm delivery [32, 35]. Additional file 1: Fig. S12, 13, 14, 15, 16, 17 and 18 showed the

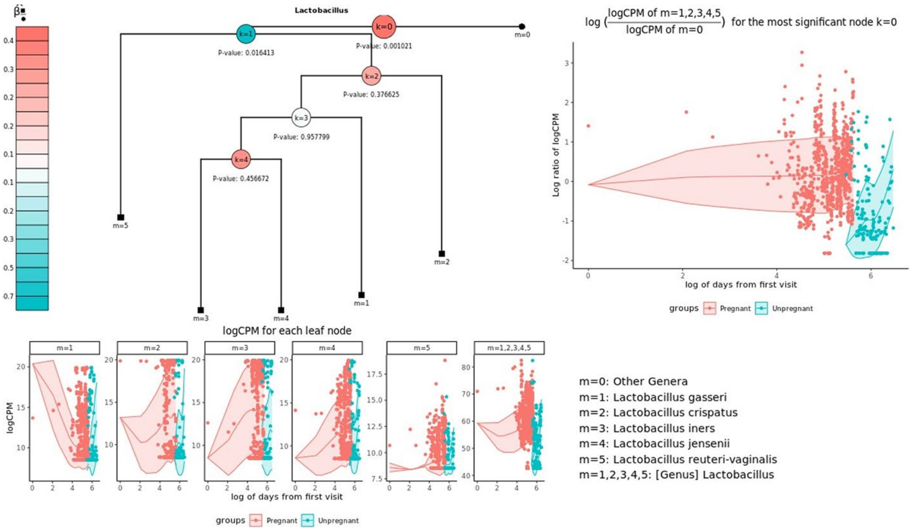


Fig. 3 Taxon distributions of significantly associated genus *Lactobacillus*. Relative proportions of taxa belonging to *Lactobacillus* at different time points were plotted. Each taxon has its corresponding leaf node; leaf nodes in black square and black circle indicate that they are in L_k and R_k , respectively. $\hat{\beta}_{\text{black square/black circle}}$ indicates the mean difference of $\log(C_{ij}^{(k)}/D_{ij}^{(k)})$ between pregnant and unpregnant subjects after adjusting for covariates, and the red internal node indicates that taxa in the left test leaf nodes are more abundant in pregnant subjects. The most significant node is enlarged

taxon distributions of other associated genera. These results confirm that the genera identified using mTMAT may be associated with delivery. Therefore, mTMAT successfully detected associated genera.

Discussion

The importance of microbiome-host interactions has been known for more than a century [36], and the occurrence of many human diseases has been found to be related to bacterial communities.

Microbiome data has phylogenetic structure and certain unique properties, including high dimensionality, rarity, and heterogeneity beyond composition. These properties cause multiple statistical problems when analyzing data across microbial composition and integrating multi-omics data such as large p and small n, dependencies, over-dispersion and zero inflation. The classical correlation and related methods throughout the microbial association study were applied in the actual study and used in the development of new methods. However, owing to those problems related to metagenomic analysis, traditional approaches are infrequently utilized for more complex models, such as longitudinal models including linear mixed models and generalized linear mixed models. Furthermore, those methods do not properly address compositional bias and may lead to erroneous results owing to limitations in relative abundance data [37].

In this study, we propose a novel approach, mTMAT, for identifying taxa associated with host diseases. This method is based on quasi-scores for internal nodes in a phylogenetic tree. It can account for various correlation structures and provides robust estimation for mis-specified correlation structures while maintaining statistical validity with small sample sizes and high statistical power. By leveraging log CPM transformation, integrating abundances, and taking ratios between integrated abundances based on the phylogenetic tree, mTMAT not only reduces sparsity issues and compositional bias but also provides insights into species-level associations by assessing patterns across closely related species.

This property is achieved by using log CPM transformation and integrating abundances based on the phylogenetic tree. Compositional bias correction is accomplished by taking ratios between two sets of integrated abundances. mTMAT leverages the phylogenetic tree structure, allowing us to incorporate the relationships between microbial species. This tree-based approach not only reduces sparsity issues but also provides insights into species-level associations by assessing patterns across closely related species.

The final statistic is formed by combining these quasi-scores to obtain a single p-value, which allows mTMAT to effectively detect taxa associated with disease status. Importantly, due to the independence of the statistical scores at internal nodes, the minimal p-value can be calculated directly (Kim et al., 2020). Comparative analyses against other methods like GLMM-MiRKAT, LMM-arcsine, LMM-log, and FZINBMM across various simulations demonstrate that mTMAT not only maintains the correct nominal type-1 error rate but also exhibits superior statistical power. Moreover, it is computationally more efficient than permutation-based methods such as GLMM-MiRKAT. While LEfSe is a useful tool for detecting differential abundance in certain microbiome datasets, it takes a different approach to statistical assessment, which may limit its ability to rigorously assess statistical significance.

However, despite its usefulness, mTMAT has several limitations. First, mTMAT combines the statistics for each internal node, and multiple comparisons occur when the number of leaf nodes is large. Second, we evaluated the performance of the proposed methods with extensive simulations, but this result cannot be generalized. Third, the choice of database, OTU clustering methods or use of ASV is likely to affect the statistical property of mTMAT. In our previous research, we showed that the effect of a misspecified phylogenetic tree is generally not substantial. However, further investigation is necessary, which we will undertake in our follow-up research.

Conclusions

mTMAT can help researchers easily perform fast and effective microbiome-wide association analysis, thereby comprehensively elucidating the interaction mechanism of the entire microbiota with the human body.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-06002-2.

Additional file 1.

Acknowledgements

Not applicable.

Author contributions

K.K. S.W.: conceptualization, methodology, software, and data curation; K.K. wrote the main manuscript. All authors contributed to the manuscript revision and approved the submitted version.

Funding
This work was supported by the National Research Foundation (NRF) grant (NRF-2021R1A5A1033157) funded by the Korean government (MSIT), by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Educa-tion (NRF-2023M3E5D3009803) and by the "Korea National Institute of Health" (KNIH) research project No.#2024ER210800.

Availability of data and materials
The 16S rRNA amplicon sequencing metagenomics datasets for Korea Association REsource cohort are accessible from the NCBI Sequence Read Archive database under project accession number PRJNA716550. mTMAT was implemented as an R package. Detailed information is available at https://healthstat.snu.ac.kr/software/mtmat.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 11 August 2024 Accepted: 27 November 2024
Published online: 06 March 2025

References

1. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nat Rev Microbiol. 2021;19(1):55–71.
2. Lee MJ, Kang MJ, Lee SY, Lee E, Kim K, Won S, et al. Perturbations of gut microbiome genes in infants with atopic dermatitis according to feeding type. J Allergy Clin Immunol. 2018;141(4):1310–9.
3. Nah G, Park SC, Kim K, Kim S, Park J, Lee S, et al. Type-2 diabetics reduces spatial variation of microbiome based on extracellur vesicles from gut microbes across human body. Sci Rep. 2019;9(1):20136.
4. Kim K, Lee Y, Won S. Relative contributions of the host genome, microbiome, and environment to the metabolic profile. Genes Genomics. 2022;44(9):1081–9.
5. Hong SN, Kim KJ, Baek MG, Yi H, Lee SH, Kim DY, et al. Association of obstructive sleep apnea severity with the composition of the upper airway microbiome. J Clin Sleep Med. 2022;18(2):505–15.
6. Kim YC, Choi S, Sohn KH, Bang JY, Kim Y, Jung JW, et al. Selenomonas: a marker of asthma severity with the potential therapeutic effect. Allergy. 2022;77(1):317–20.
7. VanderWeele TJ, Jackson JW, Li S. Causal inference and longitudinal data: a case study of religion and mental health. Soc Psychiatry Psychiatr Epidemiol. 2016;51(11):1457–66.
8. Xia Y, Sun J, Chen D-G. Introductory overview of statistical analysis of microbiome data. In: Xia Y, Sun J, Chen D-G, editors. Statistical analysis of microbiome data with R. Singapore: Springer; 2018. p. 43–75.
9. Zhang X, Pei YF, Zhang L, Guo B, Pendegraft AH, Zhuang W, et al. Negative binomial mixed models for analyzing longitudinal microbiome data. Front Microbiol. 2018;9:1683.
10. Zhang X, Yi N. Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. Bioinformatics. 2020;36(8):2345–51.
11. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016;32(17):2611–7.
12. Koh H, Li Y, Zhan X, Chen J, Zhao N. A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. Front Genet. 2019;10:458.
13. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9): e1002687.
14. Aitchison J, Egozcue J. Compositional data analysis: where are we and where should we be heading? Math Geol. 2005;37(7):829–50.
15. Zhan X, Xue L, Zheng H, Plantinga A, Wu MC, Schaid DJ, et al. A small-sample kernel association test for correlated data with application to microbiome association studies. Genet Epidemiol. 2018;42(8):772–82.
16. Kim KJ, Park J, Park SC, Won S. Phylogenetic tree-based microbiome association test. Bioinformatics. 2020;36(4):1000–6.
17. Boos DD. On generalized score tests. Am Stat. 1992;46(4):327–33.
18. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57(1):126–34.
19. Ristl R, Hothorn L, Ritz C, Posch M. Simultaneous inference for multiple marginal generalized estimating equation models. Stat Methods Med Res. 2020;29(6):1746–62.
20. Beasley TM, Erickson S, Allison DB. Rank-based inverse normal transformations are increasingly used, but are they merited? Behav Genet. 2009;39(5):580–95.
21. Kim Y, Han B-G. Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. Int J Epidemiol. 2017;46(2): e20.
22. Kim K, Lee S, Park SC, Kim NE, Shin C, Lee SK, et al. Role of an unclassified Lachnospiraceae in the pathogenesis of type 2 diabetes: a longitudinal study of the urine microbiome and metabolites. Exp Mol Med. 2022;54(8):1125–32.

23. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.j. 2011. https://doi.org/10.14806/ej.17.1.200.
24. Kwon S, Lee B, Yoon S. CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing. BMC Bioinformatics. 2014;15(Suppl 9):S10.
25. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods. 2013;10(1):57–9.
26. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4: e2584.
27. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol. 2017;67(5):1613–7.
28. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.
29. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28(14):1823–9.
30. Li K, Bihan M, Methé BA. Analyses of the stability and core taxonomic memberships of the human microbiome. PLoS ONE. 2013;8(5): e63139.
31. Williams J, Bravo HC, Tom J, Paulson JN. microbiomeDASim: simulating longitudinal differential abundance for microbiome data. F1000Res. 2019;8:1769.
32. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, Nikita L, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome. 2014. https://doi.org/10.1186/2049-2618-2-4.
33. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinform. 2010;11:538.
34. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
35. Farr A, Kiss H, Hagmann M, Machal S, Holzer I, Kueronya V, et al. Role of Lactobacillus species in the intermediate vaginal flora in early pregnancy: a retrospective cohort study. PLoS ONE. 2015;10(12): e0144181.
36. Dethlefsen L, McFall-Ngai M, Relman DA. An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature. 2007;449(7164):811–8.
37. Harrison JG, John Calder W, Shuman B, Alex BC. The quest for absolute abundance: the use of internal standards for DNA-based community ecology. Mol Ecol Resour. 2021;21(1):30–43.

Publisher's Note

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