Published online 14 January 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqab120

Adaptive and powerful microbiome multivariate association analysis via feature selection

Kalins Banerjee [1,*], Jun Chen[2] and Xiang Zhan [3,*]

[1]Department of Public Health Sciences, Pennsylvania State University, 500 University Drive, Hershey, PA 17033, USA, [2]Division of Biomedical Statistics and Informatics, Mayo Clinic, 1216 2nd Street SW, Rochester, MN 55902, USA and [3]Department of Biostatistics, School of Public Health and Beijing International Center for Mathematical Research, Peking University, 38 Xueyuan Rd, Haidian District, Beijing 100191, China

Received May 30, 2021; Revised November 13, 2021; Editorial Decision December 10, 2021; Accepted December 24, 2021

ABSTRACT

The important role of human microbiome is being increasingly recognized in health and disease conditions. Since microbiome data is typically high dimensional, one popular mode of statistical association analysis for microbiome data is to pool individual microbial features into a group, and then conduct group-based multivariate association analysis. A corresponding challenge within this approach is to achieve adequate power to detect an association signal between a group of microbial features and the outcome of interest across a wide range of scenarios. Recognizing some existing methods' susceptibility to the adverse effects of noise accumulation, we introduce the Adaptive Microbiome Association Test (AMAT), a novel and powerful tool for multivariate microbiome association analysis, which unifies both blessings of feature selection in high-dimensional inference and robustness of adaptive statistical association testing. AMAT first alleviates the burden of noise accumulation via distance correlation learning, and then conducts a data-adaptive association test under the flexible generalized linear model framework. Extensive simulation studies and real data applications demonstrate that AMAT is highly robust and often more powerful than several existing methods, while preserving the correct type I error rate. A free implementation of AMAT in R computing environment is available at https://github.com/kzb193/AMAT.

INTRODUCTION

Many microbiome studies often aim to investigate the statistical association between human microbiome compositions and an outcome of interest, such as a disease status. These studies can not only improve our understanding of the non-genetic components of complex traits and diseases, but also lead to potential development of preventive or therapeutic strategies targeted at the disease-associated microbial taxa (1–3). Next generation sequencing technology, with its recent progress, has increasingly begun to distinguish among strains or exact sequence variants during taxonomic profiling (4), which provides researchers the possibilities to answer clinical and biological questions that have eluded scientific efforts for decades. On the other hand, these research opportunities on new organisms of higher resolution have brought in new statistical challenges in microbiome association analysis. The first challenge is data sparsity or zero inflation. The higher the taxonomic resolution (e.g. low taxonomic ranks such as species or strain), the sparser the data, which makes it more difficult to detect association signals. The second challenge in the analysis is the curse of dimensionality. At a higher taxonomic resolution, there are more taxa available for association analysis, which usually indicates a heavier multiple testing correction burden. Consequently, it becomes more difficult to identify associations under family-wise statistical significance level. A naive approach would be to aggregate low-rank taxa belonging to the same high-rank category, and then perform a univariate association analysis between the high-rank taxon and the outcome. This approach can both address the sparsity issue and reduce the number of tests. However, it suffers from substantial power loss when low-rank taxa have opposite directions of effects, which get cancelled out during aggregation. Thus, new powerful and robust statistical association analysis methods are desired.

Researchers have frequently encountered a similar scenario (of variables with high dimensionality and low frequency) in genetic association analysis, where millions of rare variants have been genotyped in a typical whole genome sequencing study. A consensus among statistical geneticists is to group multiple rare genetic variants by genes or genomic regions to perform a set-based multivariate association analysis (5–7). In comparison to univariate

*To whom correspondence should be addressed. Tel: +86 10 62744132; Fax: +86 10 62744134; Email: zhanx@bjmu.edu.cn
Correspondence may also be addressed to Kalins Banerjee. Email: kbanerjee@pennstatehealth.psu.edu

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

association analyses, multivariate approaches considering multiple variants simultaneously, in general, enjoy higher statistical powers by combining weak association signals and by reducing multiple testing burden. Following the same spirit, several multivariate microbiome association analysis methods and tools have been proposed recently (8–14). In this paper, we follow the multivariate association analysis research line to investigate new robust and powerful statistical methods for testing association between a microbial community/clade of multiple taxa and an outcome of interest.

Microbial epidemiologists have increasingly recognized that not all taxa within a clade are equally functional, where functional can be understood as outcome-associated within the context of microbiome association analysis considered in this paper. In fact, it is likely that most microbial taxa within a clade are not associated with the outcome of interest (9,13). Hence, a big challenge in multivariate microbiome association analysis is to achieve enough power to detect the association signal amid noises. One possible approach to achieve more powerful results within the context of multivariate association analysis is to assign larger weights to more important taxa. The weighting idea in multivariate association analysis first stems from rare-variant genetic association analysis (5), where each rare variant is weighted based on its minor allele frequency under the assumption that rarer variants, in general, have larger impacts on the phenotype. The resulting test might not be optimal if the underlying true association patterns are against this assumption. Therefore, an assumption-free and data-adaptive weighting strategy would be preferred for more robust multivariate statistical association analysis (9,13,15). One such example is to assign weights to each variant according to the score statistics between the outcome and each individual variant (9,16,17). By doing this, the problem can be alleviated to some extent, but cannot be fully addressed as there still exist small non-zero weights to potential noises (i.e., taxa with smaller score statistics). The accumulation of such small noises can deteriorate the performance of the multivariate association test, especially when the number of taxa being tested is moderate or relatively large (18), a widely-observed phenomenon that has been termed as 'curse of dimensionality' in high-dimensional statistical literature (19).

The technique of shrinkage can be used to improve estimation of microbial associations (20). In the specific context of multivariate microbiome association analysis considered in this paper, the aforementioned problem of noise accumulation can also be mitigated by shrinkage. Specifically, we want to shrink the weights of potential noises to exactly zeroes, which can be achieved via statistical variable/feature selection. That is, we first select a subset of taxa and then construct a multivariate association testing statistic only using those selected taxa, which is equivalent to assigning a zero weight to each taxon that has not been selected. The selected subset of features is often referred to as the testing subset (21). Incorporation of feature selection can not only boost the power of the association test by mitigating the burden of accumulated noise features, but also increase the interpretability of the result in the sense that the testing subset can provide insight on the taxa that are more likely to drive the overall association. Feature selection has been extensively studied in the high dimensional statistical literature (19,22–27), and recently has been extended for microbiome data analysis (28–31). Such successful attempts in microbiome research further motivate us to incorporate feature selection into the framework of multivariate microbiome association analysis in order to develop a more powerful and robust analysis framework than existing ones.

MATERIALS AND METHODS

Notation and model

Suppose the data include n subjects, an outcome of interest, p microbiome features (e.g. abundances of p operational taxonomic units/OTUs), and q covariates that are potential confounders, such as age and gender. For the ith subject, let Y_i be the outcome, $\boldsymbol{Z}_{i\cdot} = (Z_{i1}, \ldots, Z_{ip})'$ be the vector of OTU abundances, and $\boldsymbol{X}_{i\cdot} = (X_{i1}, \ldots, X_{iq})'$ be the vector of covariates. Correspondingly, we define $\boldsymbol{Y}_{n\times 1} = (Y_1, \ldots, Y_n)'$, $Z_{n\times p} = (\boldsymbol{Z}_{1\cdot}, \ldots, \boldsymbol{Z}_{n\cdot})'$, and $X_{n\times q} = (\boldsymbol{X}_{1\cdot}, \ldots, \boldsymbol{X}_{n\cdot})'$. Additionally, let $(\boldsymbol{Z}_{\cdot 1}, \ldots, \boldsymbol{Z}_{\cdot p})$ be the columns of Z. To link the outcome of interest with microbial features and clinical coviariates, we consider a linear model (Equation 1) for a continuous outcome, and a logistic model (Equation 2) for a binary outcome:

$$Y_i = \beta_0 + \sum_{j=1}^{p} Z_{ij}\beta_j + \sum_{k=1}^{q} X_{ik}\alpha_k + \epsilon_i, \tag{1}$$

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

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'$ are regression coefficients and ϵ_i's are random errors that independently follow a Normal distribution with mean 0 and variance σ^2. Our goal is to test for possible association between OTUs and the outcome, i.e. to test H_0 : $\boldsymbol{\beta} = \boldsymbol{0}$ against H_1: atleastone $\beta_j \neq 0$, $(j = 1, \ldots, p)$, which has been extensively studied in literature (8–14).

Adaptive multivariate association analysis

A major challenge in achieving adequate statistical power to test the multivariate null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ is that, many underlying features are truly null (i.e., $\beta_j = 0$ for many Z_j's). The increased degrees of freedom paid to these noise variables Z_j's can deteriorate the power of the multivariate association test of H_0: $\beta_1 = \cdots = \beta_p = 0$, especially when the number of variables p is relatively large. To alleviate the accumulated noise effects, one feasible approach is the sum of powered score (SPU) test, which weights each variable differently according to the score vector (9,12,16). Let $\boldsymbol{U} = (U_1, \ldots, U_p)'$ be the score vector of $\boldsymbol{\beta}$ evaluated under the null model. Then, the SPU statistic (characterized by a tuning parameter γ) is defined as:

$$T_{\text{SPU}(\gamma)} = \begin{cases} \sum_{j=1}^{p} U_j^{\gamma} & , \text{if } \gamma = 1, 2, 3, \ldots, \\ Max\,\{|U_j| : j = 1, \ldots, p\} & , \text{if } \gamma = \infty. \end{cases}$$

The SPU test can be viewed as a weighted multivariate score test, which assigns a weight of $U_j^{\gamma-1}$ to score U_j. One

potential limitation of the SPU test is that, its power largely depends on the choice of γ. However, the optimal choice of γ relies on the true underlying outcome-microbiome association pattern, which stays largely unknown (9,12,13). Consequently, the adaptive SPU (aSPU) test, which combines multiple SPU tests, has been developed (16,17), and the corresponding test statistic is given as $T_{\text{aSPU}} = Min\{P_{\text{SPU}(\gamma)}: \gamma \in \Gamma\}$, where $P_{\text{SPU}(\gamma)}$ denotes the P-value of $T_{\text{SPU}(\gamma)}$. In practice, researchers have observed that $\Gamma = \{1, 2, \ldots, 8, \infty\}$ often suffices, with $\gamma = 1$ usually producing low powers when directions of individual effects are opposite, and $\gamma = 8$ often providing almost identical results as those with $\gamma = \infty$ (9). The unification of a wide range of SPU tests by taking the minimum allows the aSPU test to be data-adaptive and robust in terms of maintaining relatively high power across a wide range of scenarios, since at least one SPU test is likely to be powerful for the true association mechanism underlying the data.

A new powerful and adaptive multivariate association analysis via feature selection

We observe that the weighted sum $\sum_{j=1}^{p} U_j^{\gamma}$ of all p variables in $T_{\text{SPU}(\gamma)}$ may suffer from low power due to the high degrees of freedom paid to all variables, especially when most variables are not associated with the outcome. One way to circumvent this scenario of low statistical power is to reduce the degrees of freedom by constraining the potential noise variables to have zero weights. In other words, we will use a new multivariate association test statistic of the form $\sum_{j\in S} U_j^{\gamma}$, where $S \subset \{1, 2, \ldots, p\}$ is a collection of taxa that are more likely to be outcome-associated. In this paper, we have determined the testing subset S using statistical feature selection methods. Specifically, motivated by a previous publication (32) in mediation analysis involving high-dimensional microbial features, we have utilized the statistical framework of distance correlation (DC) learning (26,33) to determine S in the multivariate microbiome association analysis considered here.

We first present a synopsis of DC, which quantifies the degree of dependence between two random variables (33). Let $\{(V_i, W_i): i = 1, \ldots, n\}$ be a random sample of size n from the population (V, W), with $V \in \mathcal{R}$, and $W \in \mathcal{R}$. We define, $\hat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|V_i - V_j||W_i - W_j|$, $\hat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|V_i - V_j|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|W_i - W_j|$, and $\hat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|V_i - V_l||W_j - W_l|$. Then, $\widehat{\text{dcov}}(V, W) = \sqrt{\hat{S}_1 + \hat{S}_2 - 2\hat{S}_3}$ is an estimate of the distance covariance between V and W, and $\widehat{\text{dcorr}}(V, W) = \frac{\widehat{\text{dcov}}(V,W)}{\sqrt{\widehat{\text{dcov}}(V,V)\,\widehat{\text{dcov}}(W,W)}}$ is defined as the corresponding sample DC, which is bounded within [0,1]. A remarkable property of DC is that it is zero, if and only if the two underlying variables are independent.

One important application of DC is for feature screening. Distance correlation based sure independence screening or DC-SIS (26), is a highly robust and attractive feature selection procedure due to its following two characteristics. First, it is completely model-free in the sense that it allows for arbitrary regression relationship between the predictors and the outcome/ response (continuous/discrete/categorical), regardless of whether it is linear or non linear. Second, it possesses the sure screening property which ensures that, all active features can be selected with probability approaching one as the sample size increases (26). Both properties guarantee that association signals between the microbial features and the outcome would be amplified after screening. As a result, it would be easier to detect the signals for the new method to be proposed in this paper. To this end, our new statistical association analysis strategy proceeds by first screening and then association testing.

We defer the details of our two-stage procedure and focus on the screening stage first. The fundamental idea of DC-SIS is that, given a response and a set of predictors, at first the DCs between each predictor and the response are computed, and then predictors with DCs above a threshold are selected. Even though there exist few proposals regarding the choice of this threshold (26), we have implemented a new data-driven thresholding strategy for selecting a testing subset S. Since the data involve potential confounders, at first we obtain the adjusted response $\boldsymbol{r} = (r_1, \ldots, r_n)'$ as the residuals of regressing the outcome $\boldsymbol{Y}$ on covariates X. Let the sample DC between $\boldsymbol{r}$ and the jth column of Z be denoted as dc_j. We now present in Algorithm 1, a DC-based data-adaptive procedure for feature selection:

Algorithm 1: Data-adaptive feature selection via distance correlation learning

Input: Response (adjusted for additional covariates) $\boldsymbol{r}_{n\times 1}$, and predictors $Z_{n\times p}$.

Output: Set of indices $S \subset \{1, \ldots, p\}$ denoting selected features.

Procedure:

1. Compute $\{dc_j: j = 1, \ldots, p\}$.
2. Randomly permute the elements in $\boldsymbol{r}$ B-times to obtain $\{\boldsymbol{r}^{(b)} : b = 1, \ldots, B\}$.
3. For each b, compute DCs between $\boldsymbol{r}^{(b)}$ and columns of Z to get $\{dc_j^{(b)} : j = 1, \ldots, p\}$; $b = 1, \ldots, B$.
4. For $j = 1, \ldots, p$, if $dc_j > Mean\{dc_j^{(b)} : b = 1, \ldots, B\}$, then $j \in S$.
5. If none of the columns of Z are selected in step 4, select the feature having maximum DC with $\boldsymbol{r}$.

In Algorithm 1, we have used the mean of $\{dc_j^{(b)} : b = 1, \ldots, B\}$ as threshold for the jth feature. It is an estimate of the average distance correlation value between the jth feature and the adjusted response when the former truly has no prediction power over the latter (i.e. under the null model). Thus, if the jth feature is a signal, we can expect dc_j to exceed this threshold most of the times. It is of note that Algorithm 1 is not the only method to obtain a testing subset. In fact, many classic statistical feature selection methods can achieve the same goal. To this end, we have conducted comprehensive numerical studies comparing Algorithm 1 with other existing methods from two aspects. One is the accuracy of selection in terms of precision and recall rates. The other is how the selection results would affect the performance of the statistical association testing procedure introduced in the next paragraph. The corresponding results (Supplementary Tables S1–S2 and Supplementary Figures S1–S2), presented in Section 1 of the online

Supplementary Data, have clearly demonstrated superiority of the proposed strategy described in Algorithm 1. We now introduce our new feature selection infused adaptive microbiome association testing (AMAT) procedure for more powerful multivariate association analysis in Algorithm 2.

Algorithm 2: Adaptive Microbiome Association Test (AMAT)

Input: A vector of continuous or binary outcome $\boldsymbol{Y}_{n\times 1}$, OTU abundance matrix $Z_{n\times p}$, and a matrix of additional covariates $X_{n\times q}$.

Output: A P-value for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus H_1: at least one $\beta_j \neq 0$, $(j = 1, \ldots, p)$.

Procedure:

1. Obtain the normalized OTU matrix Z^* and the adjusted response $\boldsymbol{r}$.
2. Use Algorithm 1 with $\boldsymbol{r}$ and Z as inputs to obtain S.
3. Compute the aSPU statistic based on Z^*_S. Denote this statistic as T_{AMAT}.
4. For each $b = 1, \ldots, B$, permute the elements of $\boldsymbol{r}$ to get $\boldsymbol{r}^{(b)}$, and repeat steps 2-3 (use $\boldsymbol{r}^{(b)}$ and Z as inputs in step 2). Denote the resulting aSPU statistics as $T^{(b)}_{\text{AMAT}}; b = 1 \ldots, B$.
5. The P-value is estimated as $P_{\text{AMAT}} = \frac{1}{B}\sum_{b=1}^{B} I[T^{(b)}_{\text{AMAT}} \leq T_{\text{AMAT}}]$, where $I[\,\cdot\,]$ is the indicator function.

Note that, we first follow the same normalization technique used in a previous SPU-based approach (12). Specifically, we transform the OTU abundance matrix Z into a compositional matrix (if Z contains counts), and then the OTU-wise proportions are standardized to have zero mean and unit variance (Z^* matrix as described in Algorithm 2). Let $S := \{j_1, \ldots, j_{|S|}\} \subset \{1, \ldots, p\}$ denote a collection of $|S|$ OTUs that are obtained via the aforementioned feature selection procedure described in Algorithm 1. Then, the microbial design matrix used to examine the multivariate null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is $Z^*_S = (\boldsymbol{Z}^*_{\cdot j_1}, \ldots, \boldsymbol{Z}^*_{\cdot j_{|S|}})$, which is very different from existing similar approaches (9,12) that consider all p variables in the test statistic. Since we are conducting the test with microbial features that are more likely to be outcome-associated, permutations are used to control the type I error. One can also view AMAT as a weighted multivariate association test, which assigns zero weights to all variables that are not selected. Although we use permutations in AMAT for calculating both the P-values of inherent SPU statistics (i.e., $P_{\text{SPU}(\gamma)}$'s) and the final P-value P_{AMAT}, the computational cost is greatly reduced as the same set of null statistics are used to serve both purposes (12). The computational details are provided in Section 2 of the online Supplementary Data.

RESULTS

We have used both numerical simulation studies and applications to multiple real data sets to illustrate the performance of the proposed method AMAT. We demonstrate the usefulness of our new approach by comparing it to other well-established methods in the literature.

Simulation design

A comprehensive simulation study has been conducted to compare AMAT with five existing multivariate microbiome association tests: adaptive microbiome-based sum of powered score test or aMiSPU (9), the optimal microbiome regression-based kernel association test or OMiRKAT (8), optimal microbiome-based association test or OMiAT (12), linear decomposition model or LDM (14), and microbiome higher criticism analysis or MiHC (13). The aMiSPU test first uses OTU abundances and branch lengths of a phylogentic tree to compute a separate variable called generalized taxon proportions, and then uses those to conduct a set of SPU tests which are finally combined via the minimum P-value approach (9). Unlike the SPU framework, which combines multiple taxa via weighted linear combination of individual score statistics, the MiRKAT combines taxa via beta-diversity induced kernel metrics (e.g., Bray-Curtis kernel, weighted UniFrac kernel, unweighted UniFrac kernel, and generalized UniFrac kernel with parameter 0.5) (8). Then, OMiRKAT takes the minimum of these MiRKAT P-values as its test statistic (8). The OMiAT further merges a set of MiRKATs and SPU tests by using the minimum of their P-values as its test statistic (12). The LDM method examines microbiome associations using decomposition of linear models. Finally, the MiHC method uses the same minimum P-value approach to combine the Simes test and two modified versions of the higher criticism test (13). The default settings in the corresponding software packages were used to implement these aforementioned competing methods. Note that, all of these tests (AMAT, aMiSPU, LDM, MiHC, OMiAT, and OMiRKAT) use permutations to establish statistical significance. Particularly in our simulations, we used $B = 500$ permutations and set $\Gamma = \{2, 3, 4, 8\}$ for AMAT. It is of note that aMiSPU examines a slightly different hypothesis (association between the outcome and generalized taxon proportions computed at p leaf nodes as well as at the internal nodes) from the other tests, which only test for association between p OTUs and the outcome.

We followed simulation settings used in prior microbiome association analyses (8,9,12) to first generate the OTU table that mimicked a real throat microbiome data set with 856 OTUs (34), which is also analyzed later in this paper. The procedure is described in the following steps:

- Based on the throat microbiome data set (34), the estimated OTU proportions $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{856})$ as well as the estimated over dispersion parameter $\hat{\theta}$ were obtained via the method of maximum likelihood (35).
- For sample i, the observed OTU proportions were randomly generated from a Dirichlet distribution: $(p_{1i}, p_{2i}, \ldots, p_{856i}) \sim \text{Dirichlet}(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{856}, \hat{\theta})$.
- The total count of OTUs for sample i, say n_i, was randomly drawn from a negative binomial distribution with mean 1000 and size 25.
- For sample i, the observed OTU counts were randomly generated from a multinomial distribution: $(Z_{i1}, Z_{i2}, \ldots, Z_{i856}) \sim \text{Multinomial}(n_i; p_{1i}, p_{2i}, \ldots, p_{856i})$.

Table 1. Empirical type I error rates with a continuous outcome. Under the null model of scenario I, covariate X_2 is associated with a randomly selected set of OTUs, and D denotes the corresponding signal density. Under the null model of scenario II, covariate X_2 is associated with a set of OTUs that are phylogenetically related, and under the null model of scenario III, covariate X_2 is associated with a set of abundant OTUs. n denotes the sample size

n	Scenario	AMAT	aMiSPU	LDM	MiHC	OMiAT	OMiRKAT
100	I, D=3%	0.0488	0.0456	0.0508	0.0442	0.0464	0.0470
	I, D=10%	0.0476	0.0504	0.0464	0.0444	0.0460	0.0452
	I, D=20%	0.0434	0.0516	0.0474	0.0458	0.0494	0.0502
	I, D=30%	0.0484	0.0498	0.0474	0.0414	0.0464	0.0484
	II	0.0476	0.0474	0.0442	0.0468	0.0426	0.0422
	III	0.0518	0.0490	0.0462	0.0390	0.0536	0.0470
200	I, D=3%	0.0490	0.0496	0.0540	0.0542	0.0506	0.0510
	I, D=10%	0.0524	0.0532	0.0484	0.0546	0.0544	0.0466
	I, D=20%	0.0496	0.0510	0.0460	0.0506	0.0446	0.0502
	I, D=30%	0.0452	0.0464	0.0448	0.0542	0.0450	0.0498
	II	0.0470	0.0504	0.0462	0.0532	0.0456	0.0486
	III	0.0522	0.0480	0.0466	0.0532	0.0522	0.0472

Table 2. Empirical type I error rates with a binary outcome. Under the null model of scenario I, covariate X_2 is associated with a randomly selected set of OTUs, and D denotes the corresponding signal density. Under the null model of scenario II, covariate X_2 is associated with a set of OTUs that are phylogenetically related, and under the null model of scenario III, covariate X_2 is associated with a set of abundant OTUs. n denotes the sample size

n	Scenario	AMAT	aMiSPU	LDM	MiHC	OMiAT	OMiRKAT
100	I, D=3%	0.0482	0.0498	0.0446	0.0210	0.0486	0.0462
	I, D=10%	0.0496	0.0502	0.0468	0.0224	0.0510	0.0478
	I, D=20%	0.0514	0.0486	0.0492	0.0252	0.0512	0.0474
	I, D=30%	0.0518	0.0478	0.0468	0.0250	0.0552	0.0496
	II	0.0538	0.0534	0.0494	0.0284	0.0534	0.0430
	III	0.0530	0.0500	0.0498	0.0364	0.0488	0.0532
200	I, D=3%	0.0514	0.0484	0.0484	0.0196	0.0516	0.0446
	I, D=10%	0.0502	0.0494	0.0496	0.0176	0.0490	0.0430
	I, D=20%	0.0518	0.0492	0.0538	0.0216	0.0496	0.0480
	I, D=30%	0.0516	0.0474	0.0518	0.0200	0.0500	0.0470
	II	0.0506	0.0488	0.0470	0.0206	0.0470	0.0426
	III	0.0544	0.0522	0.0488	0.0354	0.0454	0.0454

Then, continuous and binary outcomes were generated under the linear model (Equation 3) and the logistic model (Equation 4) respectively as,

$$Y_i = 0.5\ \text{scale}(X_{1i} + X_{2i}) + \sum_{j\in\mathcal{A}} \beta_j\ \text{scale}(Z_{ij}) + \epsilon_i, \quad (3)$$

$$\text{logit P}(Y_i = 1) = 0.5\ \text{scale}(X_{1i} + X_{2i}) + \sum_{j\in\mathcal{A}} \beta_j\ \text{scale}(Z_{ij}), \quad (4)$$

where X_{1i} and X_{2i} were the covariates to be adjusted for, the error $\epsilon_i \sim$ N(0, 1) independently, $\mathcal{A}$ was the set of indices for outcome-associated OTUs, and the 'scale' function was used for standardization (mean 0 and standard deviation 1) across different samples. X_{1i}'s were generated from a Bernoulli distribution with success probability 0.5, and X_{2i}'s were generated to be correlated with the OTUs as, $X_{2i} = \sum_{j\in\mathcal{A}} \text{scale}(Z_{ij}) + \text{N}(0, 1)$.

Under the null model we set $\beta_j = 0$, for all $j \in \mathcal{A}$, and under the alternative model we studied three different scenarios: (I) the outcome was associated with a randomly selected set of OTUs, (II) the set of associated OTUs were phylogenetically related, (III) the outcome was associated with some abundant OTUs: Under the first scenario, we considered four different signal density (say D) levels: 3%, 10%, 20% and 30%. Under the second scenario, the OTUs were partitioned into a number of clusters based on the cophenetic distances in the real phylogenetic tree (36). For this purpose, we used the Partitioning Around Medoids (PAM) algorithm based on the optimal number of clusters, which maximized the average silhouette width in a search up to 30 clusters (13,37). Then, we randomly assigned the clusters into each iteration of the simulations as the signal-set. This was done to overcome the bias of specifying arbitrary cluster(s) as the signal-set throughout (12). Under the third scenario, we randomly picked 10 OTUs from the top 100 most abundant OTUs as the association signals. Additional simulations studies that include generating data under different schemes, with different library sizes, and from a different distribution such as negative binomial are presented in Section 3.1– Section 3.3 of the online Supplementary Data (Supplementary Tables S3– S6 and Supplementary Figures S3–S9). The regression coefficients $\{\beta_j : j \in \mathcal{A}\}$ were simulated from Uniform$(-1, 1)$ distribution to represent mixed effect directions. We used 5000 replicates to evaluate the empirical type I error rate and 1000 replicates to evaluate the empirical power. For each of the 1000 replicates under the alternative model, the set of causal OTUs were randomly selected. We considered $n = 100, 200$ as sample sizes, and set the nominal level of significance $\alpha = 0.05$ throughout this simulation.

Simulation results

The empirical type I error rates with a continuous outcome are reported in Table 1, and those with a binary outcome are reported in Table 2. All tests seem to have well controlled type I error rates across all configurations, except for MiHC,

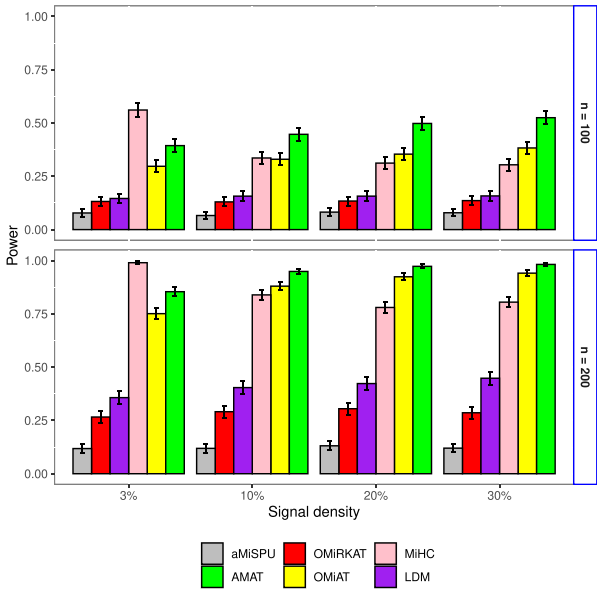

Figure 1. Empirical powers and the corresponding 95% confidence intervals obtained with a continuous outcome under scenario I.

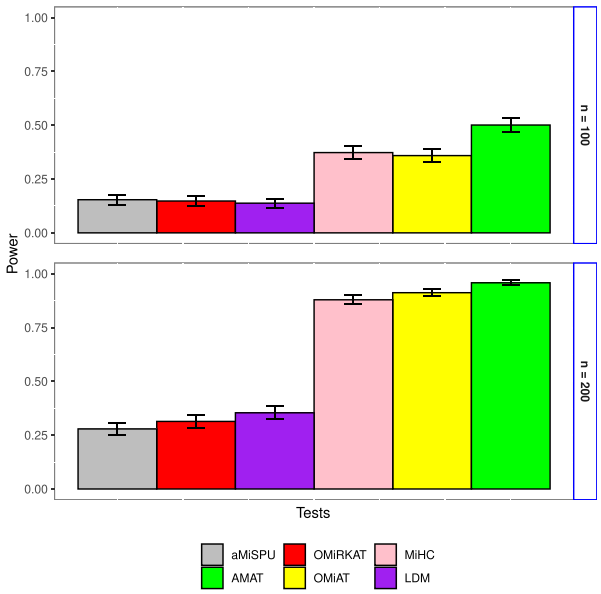

Figure 2. Empirical powers and the corresponding 95% confidence intervals obtained with a continuous outcome under scenario II.

which tends to be conservative especially when the outcome is binary.

The empirical powers with a continuous outcome are presented in Figure 1 (Scenario I), Figure 2 (Scenario II), and Figure 3 (Scenario III). We observe that AMAT has the best performance in most cases. Under Scenario I, only MiHC was able to outperform AMAT in the sole setting where the density of association signals was extremely sparse (i.e. 3%). MiHC quickly lost its superiority as the signal density increased, and AMAT was the most powerful test thereafter. Considering sample size $n = 100$ with scenario I, where the

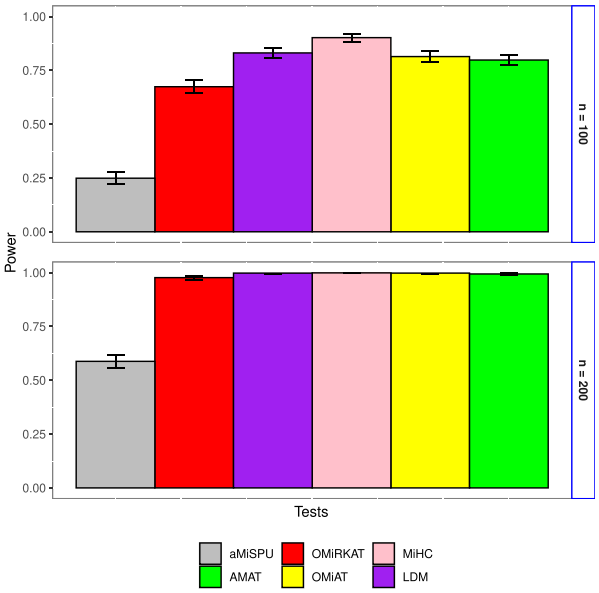

Figure 3. Empirical powers and the corresponding 95% confidence intervals obtained with a continuous outcome under scenario III.

signal-set consisted of a randomly selected set of OTUs, as an example, AMAT had powers of 44.7%, 49.8% and 52.5% under signal densities 10%, 20% and 30%, respectively, while the corresponding powers of the first runner-ups were only 33.6%, 35.4% and 38.3%. Under Scenario II, where the signal-set was characterized by phylogenetic relationships, the power of AMAT under sample size $n = 100$ was 50.1%, which was 34.3% higher than that of the first runner-up. Under Scenario III, i.e. when association signals were abundant OTUs, most tests were powerful, and MiHC was slightly better than the rest given the fact that association signals were relatively sparse (10 abundant OTUs out of 856 OTUs).

The empirical powers with a binary outcome are presented in Figure 4 (Scenario I), Figure 5 (Scenario II), and Figure 6 (Scenario III). In this case, AMAT had the best performance throughout all scenarios being considered. Under Scenario I and II, OMiAT was the second most powerful test, and under Scenario III, LDM became powerful and had powers similar to AMAT. As observed in Table 2, MiHC tends to be conservative when the outcome variable is binary. Correspondingly, MiHC tends to less powerful with a binary outcome even when the signal density is low in Figures 4 and 6.

To summarize, the proposed method AMAT tends to be the most powerful statistical association analysis method than many existing methods under most scenarios considered in our comprehensive numerical studies, except for the only scenario where the association signal density is extremely sparse (e.g. 3%). The MiHC test is specifically designed to tackle this scenario of very sparse signal densities. However, the performance of MiHC with a binary outcome variable often is inadequate. Overall, the simulation results clearly depict the highly robust and powerful performance

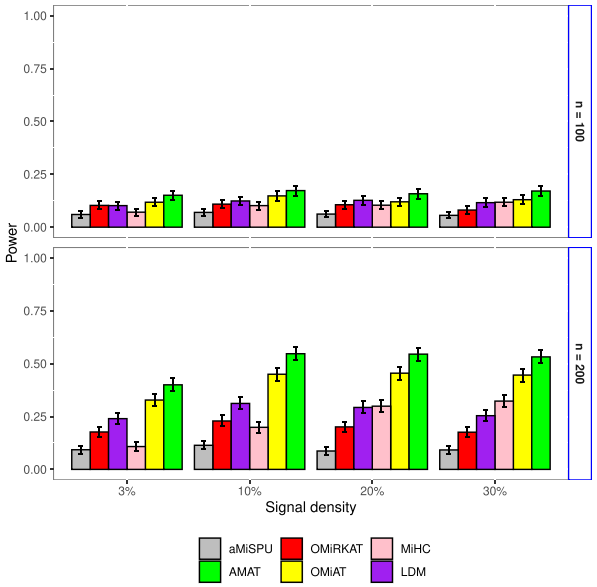

Figure 4. Empirical powers and the corresponding 95% confidence intervals obtained with a binary outcome under scenario I.

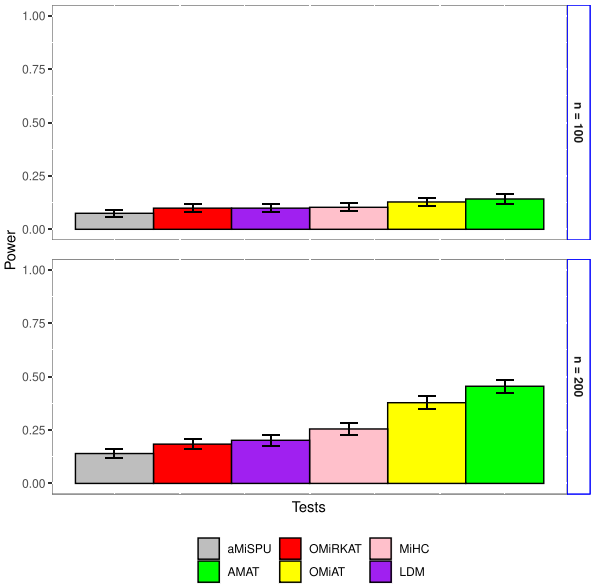

Figure 5. Empirical powers and the corresponding 95% confidence intervals obtained with a binary outcome under scenario II.

of AMAT, and thus it can be considered as an efficient tool for microbiome association analysis.

Application to throat microbiome study on smoking

Cigarette smoking is associated with an increased risk of acute respiratory tract infections. In a study (34) investigating the effect of cigarette smoking on the upper airway bacterial communities, swab samples were collected from the right and left nasopharynx and oropharynx of 29 smoking and 33 non-smoking healthy asymptomatic adults. Denoised 16S rRNA gene sequences (region V1−V2) were

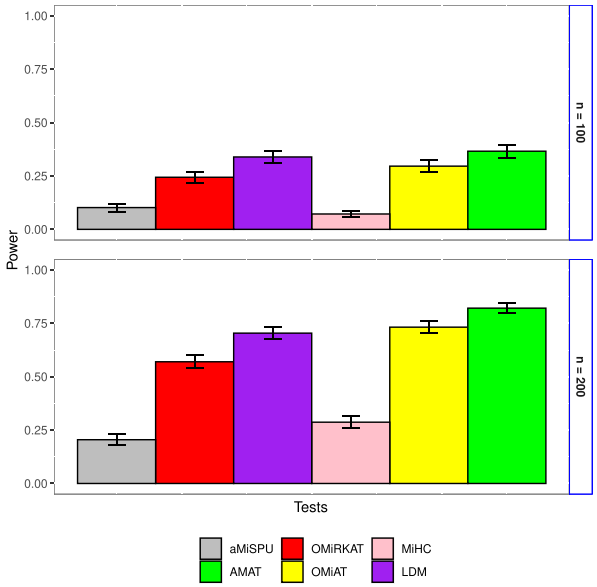

Figure 6. Empirical powers and the corresponding 95% confidence intervals obtained with a binary outcome under scenario III.

analysed using the QIIME pipeline (38) to construct the OTUs. Further details on data collection and processing can be found in the original paper (34). We used the left oropharyngeal samples to test the association between smoking status and microbial community composition. Potential confounders, which included gender and antibiotic usage within last 3 months, were adjusted for in our analysis. Quality control and data filtering steps resulted in an OTU table with 856 OTUs from 60 samples of which 28 were smokers. These 856 OTUs were further classified into different taxonomic levels (115 genera, 57 families, 27 orders, 16 classes and 11 phyla).

We first examined whether there was an overall shift in the composition of oropharyngeal microbiome community, consisting of the 856 taxa, between the smokers and the non-smokers. We used 10 000 permutations for all tests, and the corresponding *P*-values of AMAT, aMiSPU, LDM, MiHC, OMiAT and OMiRKAT were 0.0038, 0.0050, 0.0023, 0.4984, 0.0144 and 0.0070, respectively. Thus, all tests except MiHC showed that the association between microbiome profiles and smoking status was significant after adjusting for the potential confounders, which was consistent with the previous results (8,9). As mentioned earlier, besides being a powerful global test, AMAT can provide useful information on identifying taxa that are likely to be outcome-associated. The importance of members in the testing subset can be ranked based on the corresponding sample DCs. In this case, AMAT generated a testing subset of 317 OTUs, which was almost 63% reduction in the dimension of the feature space. Interestingly, almost 76% of OTUs from the testing subset belonged to only three phyla: *Firmicutes*, *Bacteroidetes* and *Proteobacteria*, which indicated that the overall community-level significant association with smoking might had been primarily driven by changes in these phyla.

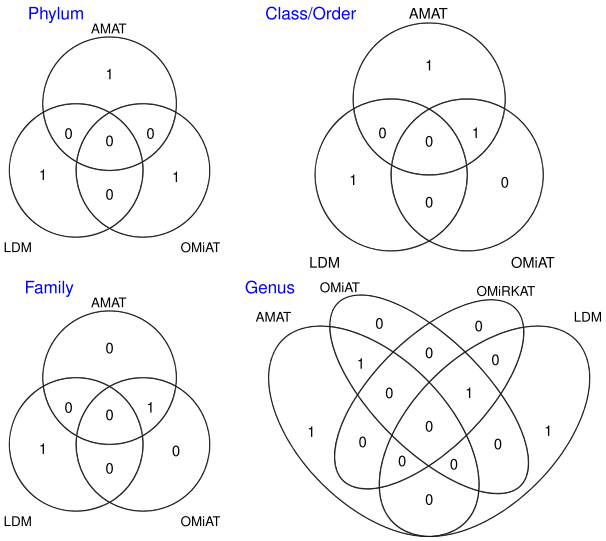

Figure 7. Venn diagram of detected associations between smoking status and oropharyngeal microbiota at different taxonomic ranks. Methods omitted at a rank indicates that no significance are detected at that rank.

Changes in the structure of the microbiota that are associated with the outcome of interest can occur at any taxonomic rank, or along any relevant branch of the phylogenetic tree. For example, it is well known that changes at the phylum level (e.g. *Firmicutes* and *Bacteroidetes*) are reported to be associated with obesity (39). On the other hand, strain level associations with the metabolism of drugs in human have been reported (40). Hence, it was also of interest to conduct a comprehensive association analysis among all taxonomic ranks to elucidate the relationship between smoking status and the oropharyngeal microbiota under consideration. Specifically, we conducted association analysis at the levels of genus, family, order, class and phylum with taxa-sets that contain at least five taxa. Consequently, 35 genera, 29 families, 17 orders, 14 classes and 9 phyla were subjected to association analysis. The family wise error rate (FWER) was controlled at 0.05 within each taxonomic rank via the Bonferroni correction. The results indicated that AMAT and OMiAT had a clear advantage in terms of identifying significant differences with AMAT providing maximum number of discoveries (see Figure 7). Overall, AMAT made eight discoveries of which four were unique to it (*Veillonella* genus, *Coriobacteriales* order, *Coriobacteriia* class and *Firmicutes* phylum), while four were also discovered by OMiAT. Several microbiome studies found *Firmicutes* to be associated with smoking (41,42). The discoveries made by LDM did not coincide with those of any other tests except for a case at the genus level. The original study (34), which conducted univariate association testing at both family level and genus level, found *Veillonellaceae* family along with *Megasphaera* and *Veillonella* genera from the *Firmicutes* phylum, to be included in the set of taxa that distinguished the left oropharyngeal microbial communities of smokers from nonsmokers (see Table 3 and Table S3 of (34)). Our multivariate analysis revealed *Veillonellaceae* as a significant family, which was identified only by AMAT and OMiAT. Besides, AMAT identified both *Veillonella* and *Megasphaera* genera, whereas only the latter was identified by OMiAT. No other findings from the competing tests matched with the discoveries of the original study. These results were consistent with our simulation studies in the sense that AMAT and OMiAT tend to be the two most powerful tests in most of the simulation settings. On the other hand, MiHC and aMiSPU failed to identify even a single significant taxon across all taxonomic ranks being considered, possibly due to power loss from a smaller sample size of this data set.

Application to gut microbiome study on body mass index

The human health is strongly affected by one's diet, which is known to partly modulate the gut microbial community. For instance, dysbiosis of the gut microbiome has been shown to be associated with obesity (39). In a previous research examining the relationship between dietary patterns and gut microbiome composition (43), fecal samples from 98 healthy volunteers were collected, the V1–V2 region of the 16S rRNA genes were sequenced, and the QIIME pipeline (38) was used to obtain the OTUs. We refer to the original paper (43) for further details. Here, our objective is to test for a possible association between the gut microbiota and body mass index (BMI). A filtering to include OTUs with counts of more than three in more than three samples resulted in a community of 557 OTUs, which were further taxonomically classified into 31 genera, 17 families, 7 orders, 8 classes and 6 phyla.

At the community-level association testing, which involved all 557 OTUs, all tests except aMiSPU and LDM were able to detect a significant association between BMI and the gut microbial community (*P*-values of AMAT, aMiSPU, LDM, MiHC, OMiAT and OMiRKAT were 0.0293, 0.0976, 0.1370, 0.0241, 0.0471 and 0.0345 respectively). The original study (43) also identified BMI to be significantly associated with the microbiome composition (see Table S1 of (43)). Next, we conducted association analysis at different taxonomic ranks with taxa-sets that contain at least five taxa. As before, the Bonferroni correction was used to control the FWER. The results showed that no test uniformly outperformed others across all taxonomic ranks. At the family level, both MiHC and OMiAT identified *Veillonellaceae*, whereas aMiSPU, LDM, and OMiRKAT identified *Lachnospiraceae*. But, AMAT showed superiority by identifying both *Veillonellaceae* and *Lachnospiraceae*, which had been found to be associated with BMI in other studies as well (44,45). On the other hand, AMAT did not detect any significant genera, while *Ruminococcaceae Incertae Sedis* was identified by MiHC, and *Lachnospiraceae Incertae Sedis* was identified by aMiSPU, LDM, and OMiRKAT . Furthermore, LDM, which failed to detect the community-level association, identified three additional taxa-sets (*Firmicutes* phylum, *Clostridia class* and *Clostridiales* order). But these findings were not consistent with the corresponding results from any of the other tests considered.

DISCUSSION AND CONCLUSION

In this paper, we have focused on the problem of statistical multivariate association analysis within the context of

microbiome studies and proposed the AMAT method, a new testing strategy that instills the benefits of feature selection into the multivariate association testing framework, which has not been attempted in many popular existing association analysis methods such as OMiRKAT (8), aMiSPU (9) and OMiAT (12). One major contribution of AMAT is that, by recognizing the existing multivariate tests' vulnerability to the adverse effects of accumulated noise features, AMAT introduces a novel perspective of utilizing dimension reduction techniques under the multivariate microbiome association analysis framework to achieve extremely robust and powerful results across a wide range of scenarios. The recently developed MiHC method has recognized the same phenomenon, but it only provides solution to scenarios with very sparse association signals (13). On the other hand, data-adaptive feature selection embedded in AMAT makes it more flexible and robust to different levels of association signal densities, and has been shown in our numerical studies to be much more powerful than MiHC except for few extremely sparse scenarios. Moreover, results of two real data application examples indicate that AMAT serves as a highly robust and powerful test for microbiome association analysis. A third example presented in Section 3.4 of the online Supplementary Data also supports this conclusion. Therefore, our new AMAT method fills an important gap in multivariate microbiome association analysis, and will be an extremely appealing tool for association analysis when the underlying signal density is neither too sparse nor too dense.

The performance of AMAT also depends on the efficiency of the intermediate feature selection procedure. We have utilized the distance correlation based sure independence screening framework which not only possesses the sure screening property, but also is highly robust to model misspecification (26). Additionally, we have implemented a new thresholding strategy that determines the dimension of the reduced feature space (say d) in a data-driven manner. Researchers often prefix d as $[n/log(n)]$ ([.] denotes the floor function) or $n-1$ (19,26), which have been shown to be versatile in classic high-dimensional inference. Such choices, however, may not be appropriate for the relatively small sample sizes frequently encountered in most current microbiome association analyses. Among some popular feature selection tools (19,26,27,46) evaluated in our numerical studies, the feature selection strategy outlined in Algorithm 1 had the optimum performance within the context of the current paper (see details in Section 1 of the online Supplementary Data). It is of future research interest to further boost the power of AMAT by sharping the intermediate feature selection tool embedded in it.

Unlike many previous microbiome community level association analysis tools, which have incorporated the microbial phylogenetic tree into association analysis (8,9,12,13), the role of phylogenetic tree has been downplayed in the current paper. One major concern against this idea is due to computational reasons. The intermediate feature selection step in each of the different permutations (used for establishing significance) results in different subsets of OTUs kept for association testing. This causes different pruning of the phylogenetic tree in the testing stage (if phylogenetic information is accommodated in association testing), which can be computationally expensive when either the number of OTUs or the number of permutations is large. Moreover, outcome-associated microbial changes can occur at any taxonomic ranks and/or along any relevant branch of the phylogenetic tree (2). When the analysis unit of AMAT is a group of OTUs belonging to some particular relatively low taxonomic rank (e.g. genus or family), all OTUs in the group being tested share relatively homologous phylogeny, and the phylogenetic tree is expected to play a less important role in this type of association analysis. Even when the tree plays an active role in microbiome-outcome association (e.g. simulation scenario II), our numerical studies have demonstrated the robustness of AMAT under such scenarios.

The size of data sets is exploding as metagenomic sequencing technologies keep evolving, and there is an even more pressing need to perform a more powerful microbiome association analysis so that true association signals can be detected amid a huge amount of background noises. Our research demonstrates that implementation of feature selection can improve the performance of microbiome association analysis, and correspondingly AMAT serves as a highly robust and powerful taxa-set based multivariate association testing tool. Furthermore, its testing subset can provide insights on the taxa that are more likely to drive the detected overall association, and thus can lead to cost effective downstream validation and functional studies. Finally, the good performance of AMAT comes at a price. Since the feature selection algorithm is implemented in each permutation used for P-value calculation, the computational cost of AMAT is typically larger than its competitors. Taking a simulation under Scenario II with $n = 100$ as an example, the average computation time of AMAT over 1000 replicates is around 100 s, while that of LDM and MiHC is around 70 s. OMiRKAT, the fastest among the tests considered, takes only a few seconds. Fortunately, the relatively small sample size of microbiome data makes AMAT still feasible from a computational perspective.

DATA AVAILABILITY

The R code to implement the proposed AMAT method is available at https://github.com/kzb193/AMAT. The throat microbiome data set and gut microbiome data set used to demonstrate AMAT in this study were accessed from previous publications (34) and (43), respectively.

SUPPLEMENTARY DATA

Supplementary data are available at NARGAB online.

ACKNOWLEDGEMENTS

The authors acknowledge comments and suggestions from the Editor, the Associate Editor and two anonymous reviewers.

FUNDING

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Virgin,H.W. and Todd,J.A. (2011) Metagenomics and personalized medicine. *Cell*, **147**, 44–56.
2. Gilbert,J.A., Quinn,R.A., Debelius,J., Xu,Z.Z., Morton,J., Garg,N., Jansson,J.K., Dorrestein,P.C. and Knight,R. (2016) Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, **535**, 94–103.
3. Surana,N.K. and Kasper,D.L. (2017) Moving beyond microbiome-wide associations to causal microbe identification. *Nature*, **552**, 244–247.
4. Callahan,B.J., McMurdie,P.J. and Holmes,S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.*, **11**, 2639–2643.
5. Wu,M.C., Lee,S., Cai,T., Li,Y., Boehnke,M. and Lin,X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
6. Lee,S., Abecasis,G.R., Boehnke,M. and Lin,X. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.
7. Zhan,X., Zhao,N., Plantinga,A., Thornton,T.A., Conneely,K.N., Epstein,M.P. and Wu,M.C. (2017) Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, **206**, 1779–1790.
8. Zhao,N., Chen,J., Carroll,I.M., Ringel-Kulka,T., Epstein,M.P., Zhou,H., Zhou,J.J., Ringel,Y., Li,H. and Wu,M.C. (2015) Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.*, **96**, 797–807.
9. Wu,C., Chen,J., Kim,J. and Pan,W. (2016) An adaptive association test for microbiome data. *Genome Med.*, **8**, 56.
10. Tang,Z.-Z., Chen,G. and Alekseyenko,A.V. (2016) PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, **32**, 2618–2625.
11. Plantinga,A., Zhan,X., Zhao,N., Chen,J., Jenq,R.R. and Wu,M.C. (2017) MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome*, **5**, 17.
12. Koh,H., Blaser,M.J. and Li,H. (2017) A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome*, **5**, 45.
13. Koh,H. and Zhao,N. (2020) A powerful microbial group association test based on the higher criticism analysis for sparse microbial association signals. *Microbiome*, **8**, 63.
14. Hu,Y.-J. and Satten,G.A. (2020) Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics*, **36**, 4106–4115.
15. Song,Y., Zhao,H. and Wang,T. (2020) An adaptive independence test for microbiome community data. *Biometrics*, **76**, 414–426.
16. Pan,W., Kim,J., Zhang,Y., Shen,X. and Wei,P. (2014) A powerful and adaptive association test for rare variants. *Genetics*, **197**, 1081–1095.
17. Pan,W., Kwak,I.-Y. and Wei,P. (2015) A powerful pathway-based adaptive test for genetic association with common or rare variants. *Am. J. Hum. Genet.*, **97**, 86–98.
18. Banerjee,K., Zhao,N., Srinivasan,A., Xue,L., Hicks,S.D., Middleton,F.A., Wu,R. and Zhan,X. (2019) An adaptive multivariate two-sample test with application to microbiome differential abundance analysis. *Front. Genet.*, **10**, 350.
19. Fan,J. and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. B*, **70**, 849–911.
20. Badri,M., Kurtz,Z.D., Bonneau,R. and Müller,C.L. (2020) Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genomics Bioinformatics*, **2**, lqaa100.
21. Fan,J. (1996) Test of significance based on wavelet thresholding and Neyman's truncation. *J. Am. Stat. Assoc.*, **91**, 674–688.
22. Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Met.*, **00**, 267–288.
23. Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.
24. Fan,J., Samworth,R. and Wu,Y. (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.
25. Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, **67**, 301–320.
26. Li,R., Zhong,W. and Zhu,L. (2012) Feature screening via distance correlation learning. *J. Am. Stat. Assoc.*, **107**, 1129–1139.
27. Yang,S., Wen,J., Eckert,S.T., Wang,Y., Liu,D.J., Wu,R., Li,R. and Zhan,X. (2020) Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning. *Bioinformatics*, **36**, 3811–3817.
28. Lin,W., Shi,P., Feng,R. and Li,H. (2014) Variable selection in regression with compositional covariates. *Biometrika*, **101**, 785–797.
29. Wang,T. and Zhao,H. *et al.* (2017) Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.*, **11**, 771–791.
30. Srinivasan,A., Xue,L. and Zhan,X. (2021) Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics*, **77**, 984–995.
31. Susin,A., Wang,Y., Lê Cao,K.-A. and Calle,M.L. (2020) Variable selection in microbiome compositional data analysis. *NAR Genomics Bioinformatics*, **2**, lqaa029.
32. Hamidi,B., Wallace,K. and Alekseyenko,A.V. (2019) MODIMA, a method for multivariate omnibus distance mediation analysis, allows for integration of multivariate exposure–mediator–response r. *Genes*, **10**, 524.
33. Székely,G.J., Rizzo,M.L., Bakirov,N.K. *et al.* (2007) Measuring and testing dependence by correlation of distances. *Ann. Stat.*, **35**, 2769–2794.
34. Charlson,E.S., Chen,J., Custers-Allen,R., Bittinger,K., Li,H., Sinha,R., Hwang,J., Bushman,F.D. and Collman,R.G. (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One*, **5**, e15216.
35. Chen,J., Bittinger,K., Charlson,E.S., Hoffmann,C., Lewis,J., Wu,G.D., Collman,R.G., Bushman,F.D. and Li,H. (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, **28**, 2106–2113.
36. Sneath,P.H. and Sokal,R.R. (1973) In: *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. Freeman, San Francisco, CA.
37. Reynolds,A.P., Richards,G., de la Iglesia,B. and Rayward-Smith,V.J. (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algorithms*, **5**, 475–504.
38. Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K., Fierer,N., Pena,A.G., Goodrich,J.K., Gordon,J.I. and et,al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335.
39. Ley,R.E., Turnbaugh,P.J., Klein,S. and Gordon,J.I. (2006) Human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.
40. Haiser,H.J., Gootenberg,D.B., Chatman,K., Sirasani,G., Balskus,E.P. and Turnbaugh,P.J. (2013) Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. *Science*, **341**, 295–298.
41. Wu,J., Peters,B.A., Dominianni,C., Zhang,Y., Pei,Z., Yang,L., Ma,Y., Purdue,M.P., Jacobs,E.J., Gapstur,S.M. *et al.* (2016) Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J.*, **10**, 2435–2446.
42. Lee,S.H., Yun,Y., Kim,S.J., Lee,E.-J., Chang,Y., Ryu,S., Shin,H., Kim,H.-L., Kim,H.-N. and Lee,J.H. (2018) Association between cigarette smoking status and composition of gut microbiota: population-based cross-sectional study. *J. Clin. Med.*, **7**, 282.
43. Wu,G.D., Chen,J., Hoffmann,C., Bittinger,K., Chen,Y.-Y., Keilbaugh,S.A., Bewtra,M., Knights,D., Walters,W.A., Knight,R. *et al.* (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science*, **334**, 105–108.
44. Duan,M., Wang,Y., Zhang,Q., Zou,R., Guo,M. and Zheng,H. (2021) Characteristics of gut microbiota in people with obesity. *Plos one*, **16**, e0255446.
45. Peters,B.A., Shapiro,J.A., Church,T.R., Miller,G., Trinh-Shevrin,C., Yuen,E., Friedlander,C., Hayes,R.B. and Ahn,J. (2018) A taxonomic signature of obesity in a large study of American adults. *Sci. Rep.-UK*, **8**, 9749.
46. Saldana,D.F. and Feng,Y. (2018) SIS: an R package for sure independence screening in ultrahigh dimensional statistical models. *J. Stat. Softw.*, **83**, 1–25.