Supplementary Data

Adaptive and powerful microbiome multivariate association analysis via feature selection

Kalins Banerjee\textsuperscript{1}, Jun Chen\textsuperscript{2}, and Xiang Zhan\textsuperscript{3}

\textsuperscript{1}Department of Public Health Sciences, Pennsylvania State University, 500 University Drive, Hershey, PA 17033, USA, and
\textsuperscript{2}Division of Biomedical Statistics and Informatics, Mayo Clinic, 1216 2nd Street SW, Rochester, MN 55902, USA, and
\textsuperscript{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.

The content of this Supplementary Data are organized into three sections. In section 1, the optimum tool for testing subset selection (the intermediate procedure) in AMAT has been investigated. In Section 2, the computational aspects of AMAT are described in detail. In Section 3, we present some additional numerical studies to further compare AMAT with existing methods to get a fairer assessment of it, and also a better picture of the unique features of AMAT as compared to methods in literature.

1 Optimum tool for testing subset selection in AMAT

As pointed out in the main text, a testing subset can also be selected using feature selection methods other than the one we proposed (Algorithm 1). In order to find the optimum testing subset selection tool for AMAT, we have conducted two sets of comprehensive numerical studies with different feature selection methods under the simulation settings of microbiome association analysis described in the main text. In the first set of numerical studies, we have evaluated their screening performance, and in the second set of numerical studies, we have compared the performance of AMAT embedded with different feature selection methods.

Screening performance of different feature selection methods

We have conducted numerical studies to compare performance of the proposed testing subset selection technique (Algorithm 1) with that of some other popular methods using data generated under various alternative models considered in the main text. For ease of presen-
tation, we denote the proposed DC based adaptive feature selection technique described in Algorithm 1 as A-DC.

The first alternative technique we have considered is similar to A-DC except for the fact that it uses a quantile based thresholding strategy. In particular, if \( \zeta_q^{(b)} \) be the \( q^{th} \) quantile of \( \{ d_{c_j}^{(b)} : j = 1, \ldots, p \} \), then the median of \( \{ \zeta_q^{(b)} : b = 1, \ldots, B \} \) can be used as a common threshold, which allows for 100(1 − \( q \))% false selections, when \( r \) and the features are not related (Fan and Lv, 2017). We have used the notation DC\((q)\) to denote this alternative feature selection strategy. Note that, both A-DC and DC\((q)\) only utilize the marginal associations, and thus they might be susceptible to the limitations of independence learning (Fan and Lv, 2017). Consequently, we have considered as our second alternative, the Iterative Sure Independence Screening (ISIS) procedure (Fan and Lv, 2008; Saldana and Feng, 2018) which was primarily developed to address the potential drawbacks of independence learning.

Among the several variants of ISIS (Saldana and Feng, 2018), we have considered the Vanilla ISIS (VAN-ISIS) and the Permutation based ISIS (PERM-ISIS) for our study. The former is the default R implementation of ISIS, whereas the later uses a data-driven thresholding strategy. Similar to DC\((q)\), the threshold of PERM-ISIS\((q)\) allows only a (1 − \( q \)) proportion of inactive features to enter the model, when the features are truly not associated with the response. The third alternative is a new class of feature selection methods that use pseudo features. One such example is the permutation-assisted tuning procedure in lasso/plasso (Yang et al., 2020).

For A-DC, we used 100 permutations for computing the feature-specific thresholds, and for both DC\((q)\) and PERM-ISIS\((q)\) we set \( q = 0.95, 0.90 \). The performance of different testing subset selection procedures are evaluated based on the following three criteria: (1) \( \mathcal{P} \): the average precision (number of active features selected/number of features selected), (2) \( \mathcal{R} \): the average recall (number of active features selected/number of active features), and (3) \( \mathcal{F} \): the average F-score (harmonic mean of precision and recall). We used 500 replications to obtain these averages. Even though an ideal feature selection procedure is expected to be high in both precision and recall, there is usually an inverse relationship between them. Thus, the F-score is often used to combine these two measures. We are implementing a feature selection step to select a taxa-set for association testing, and in order to devise a powerful testing tool, we expect the taxa-set to contain as many active features as possible. Thus, we prioritize a testing subset selection procedure which retains a large proportion of the active features.

The results under scenario I with a continuous outcome and a binary outcome are presented in Table S1 and Table S2, respectively. Compared to most statistical feature selection papers (Saldana and Feng, 2018; Yang et al., 2020), the precision rate is really low. This is primarily because a much smaller sample size of \( n = 100 \) or 200 (as in microbiome association analysis) is considered in this numerical experiment, as compared to that in those previous papers (Saldana and Feng, 2018; Yang et al., 2020). Also, due to the same reason, recall rates of all methods reported in both Table S1 and Table S2 are smaller than the level of asymptotic sure independence screening. While both the precision and recall of A-DC might be improved by using some iterative procedures as done in VAN-ISIS and PERM-ISIS, we
do not pursue this in AMAT due to computational concern, as the same feature selection procedure would be required in each permutation when establishing significance (see next section for more details). Relatively speaking, we observed that the proposed A-DC procedure produced the best average recall rate among all methods across all configurations, and the best F-score for $D = 20\%, 30\%$. Besides, A-DC had a better F-score than others at $D = 10\%$ when the sample size was 100. There was no single method that had the optimum performance across scenarios where A-DC did not have the best F-score. In terms of precision, there was no single method that outperformed others uniformly. The results under scenario II also showed similar patterns, and thus were not reported. Note that, between DC(0.90) and DC(0.95), the former produced a much better recall always, and had a better F-score in most settings. Besides, all the variants of ISIS almost always had similar results. Consequently, we have compared the performance of AMAT based only on A-DC, DC(0.90), plasso, and VAN-ISIS in the following subsection.

**Performance of AMAT with different feature selection procedures**

The performance of AMAT based on A-DC, DC(0.90), plasso, and VAN-ISIS for a continuous outcome are presented in Figure S1 (scenario I) and Figure S2 (scenario II). We observed that A-DC based AMAT had the best overall performance. Even though DC(0.90) based AMAT had a slight advantage at signal density $D = 3\%$, it quickly lost power as the signal density increased. Thus, we have developed our association testing framework using A-DC as the tool for testing subset selection in AMAT.
Table S1: Performance of feature selection procedures with a continuous outcome under scenario I, where the outcome is associated with a randomly selected set of OTUs. \(n\) and \(D\), respectively denote sample size and signal density. Numbers listed without parentheses are average values over 500 replicates, and numbers listed within parentheses are the corresponding standard errors. For each \(n\), column wise maximum averages are marked in bold.

| \(n\) | Method      | \(D=3\%\) | \(P\) | \(R\) | \(F\) | \(D=10\%\) | \(P\) | \(R\) | \(F\) | \(D=20\%\) | \(P\) | \(R\) | \(F\) | \(D=30\%\) | \(P\) | \(R\) | \(F\) |
|-------|-------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
|      |             | \(n\) | \(0.066\) | \(0.787\) | \(0.122\) | \(0.016\) | \(0.599\) | \(0.247\) | \(0.263\) | \(0.514\) | \(0.347\) | \(0.362\) | \(0.479\) | \(0.514\) | \(0.347\) | \(0.362\) |
|      | A-DC        |      | \(0.067\) | \(0.886\) | \(0.125\) | \(0.168\) | \(0.701\) | \(0.272\) | \(0.276\) | \(0.6\)  | \(0.377\) | \(0.372\) | \(0.547\) | \(0.443\) | \(0.547\) | \(0.443\) |
|      | DC(0.95)    |      | \(0.271\) | \(0.503\) | \(0.37\)  | \(0.391\) | \(0.293\) | \(0.334\) | \(0.458\) | \(0.17\)  | \(0.248\) | \(0.526\) | \(0.13\)  | \(0.208\) | \(0.003\) | \(0.001\) |
|      | DC(0.90)    |      | \(0.175\) | \(0.681\) | \(0.279\) | \(0.308\) | \(0.401\) | \(0.347\) | \(0.397\) | \(0.263\) | \(0.316\) | \(0.475\) | \(0.211\) | \(0.292\) | \(0.002\) | \(0.001\) |
| 200  | plasso      |      | \(0.606\) | \(0.06\)  | \(0.106\) | \(0.31\)  | \(0.006\) | \(0.012\) | \(0.283\) | \(0.003\) | \(0.005\) | \(0.266\) | \(0.002\) | \(0.003\) | \(0.17\)  | \(0.011\) |
|      | VAN-ISIS    |      | \(0.548\) | \(0.671\) | \(0.594\) | \(0.309\) | \(0.356\) | \(0.33\)  | \(0.315\) | \(0.202\) | \(0.246\) | \(0.386\) | \(0.172\) | \(0.238\) | \(0.003\) | \(0.001\) |
|      | PERM-ISIS(0.95) |       | \(0.42\)  | \(0.679\) | \(0.506\) | \(0.356\) | \(0.356\) | \(0.356\) | \(0.338\) | \(0.209\) | \(0.258\) | \(0.403\) | \(0.187\) | \(0.255\) | \(0.002\) | \(0.001\) |
|      | PERM-ISIS(0.90) |       | \(0.502\) | \(0.675\) | \(0.561\) | \(0.354\) | \(0.372\) | \(0.362\) | \(0.337\) | \(0.208\) | \(0.257\) | \(0.401\) | \(0.177\) | \(0.246\) | \(0.002\) | \(0.001\) |
Table S2: Performance of feature selection procedures with a binary outcome under scenario I, where the outcome is associated with a randomly selected set of OTUs. \( n \) and \( D \), respectively denote sample size and signal density. Numbers listed without parentheses are average values over 500 replicates, and numbers listed within parentheses are the corresponding standard errors. For each \( n \), column wise maximum averages are marked in bold.

| \( n \) | Method       | \( D=3\% \) | \( D=10\% \) | \( D=20\% \) | \( D=30\% \) |
|-------|--------------|-------------|--------------|--------------|--------------|
|       |              | \( P \)     | \( R \)      | \( F \)      | \( P \)     | \( R \)     | \( F \)     | \( P \)     | \( R \)     | \( F \)     | \( P \)     | \( R \)     | \( F \)     |
| 100   | A-DC         | 0.05        | **0.669**    | 0.094        | 0.132        | **0.537**    | **0.212**   | 0.241        | **0.491**    | **0.323**   | 0.342        | **0.47**    | **0.396**   |
|       |              | (0)         | (0.01)       | (0)          | (0)          | (0.003)      | (0)         | (0)          | (0.001)      | (0)         | (0)          | (0)         | (0)         |
|       | DC(0.95)     | 0.103       | **0.128**    | 0.19         | **0.097**    | 0.128        | **0.073**    | **0.116**    | (0.004)      | (0)         | (0.005)      | (0.001)     | (0.001)     |
|       |              | (0.002)     | (0.005)      | (0.003)      | (0.001)      | (0.002)      | (0)         | (0)         | (0)          | (0)         | (0)          | (0)         | (0)         |
|       | DC(0.90)     | 0.081       | 0.26         | 0.124        | 0.169        | 0.168        | 0.271        | 0.133        | 0.178        | 0.365        | 0.122        | 0.182       |
|       |              | (0.001)     | (0.008)      | (0.002)      | (0.001)      | (0.002)      | (0)         | (0)         | (0)          | (0)         | (0)          | (0)         | (0)         |
| 200   | plasso       | 0.105       | 0.005        | **0.009**    | 0.083        | 0.001        | 0.002        | 0.077        | 0.001        | 0.001        | 0.07         | 0           | **0.001**   |
|       |              | (0.09)      | (0)          | (0.001)      | (0.071)      | (0)          | (0)          | (0.068)      | (0)          | (0)          | (0.062)      | (0)         | (0)         |
|       | VAN-ISIS     | 0.074       | 0.117        | 0.09         | 0.153        | 0.075        | 0.1          | 0.259        | 0.063        | 0.102        | 0.351        | 0.057       | 0.099       |
|       |              | (0.002)     | (0.004)      | (0.001)      | (0.003)      | (0.001)      | (0)          | (0.005)      | (0)          | (0)          | (0.006)      | (0)         | (0.001)     |
|       | PERM-ISIS(0.95) | 0.075     | 0.101        | 0.086        | 0.162        | 0.068        | 0.096        | 0.255        | 0.053        | 0.088        | 0.348        | 0.049       | 0.086       |
|       |              | (0.002)     | (0.004)      | (0.003)      | (0.004)      | (0.001)      | (0)          | (0.005)      | (0)          | (0)          | (0.006)      | (0)         | (0)         |
|       | PERM-ISIS(0.90) | 0.078     | 0.108        | 0.09         | 0.156        | 0.067        | 0.094        | 0.259        | 0.055        | 0.091        | 0.355        | 0.051       | 0.09        |
|       |              | (0.002)     | (0.004)      | (0.003)      | (0.004)      | (0.001)      | (0)          | (0.005)      | (0)          | (0)          | (0.006)      | (0)         | (0)         |
| 500   | plasso       | **0.32**    | **0.017**    | **0.032**    | 0.248        | 0.004        | 0.008        | 0.168        | 0.001        | 0.002        | 0.111        | 0.001       | 0.001       |
|       |              | (0.205)     | (0.001)      | (0.002)      | (0.171)      | (0)          | (0)          | (0.134)      | (0)          | (0)          | (0.096)      | (0)         | (0)         |
|       | VAN-ISIS     | 0.075       | 0.252        | 0.115        | 0.166        | 0.171        | 0.168        | 0.263        | 0.137        | 0.179        | 0.355        | 0.123       | 0.183       |
|       |              | (0.001)     | (0.008)      | (0.002)      | (0.002)      | (0.002)      | (0)          | (0.002)      | (0.001)      | (0)          | (0.002)      | (0)         | (0.001)     |
|       | PERM-ISIS(0.95) | 0.086     | 0.242        | 0.126        | 0.185        | 0.159        | 0.17         | 0.275        | 0.12         | 0.167        | 0.367        | 0.107       | 0.166       |
|       |              | (0.001)     | (0.008)      | (0.002)      | (0.002)      | (0.001)      | (0)          | (0.003)      | (0.001)      | (0)          | (0.003)      | (0)         | (0.001)     |
|       | PERM-ISIS(0.90) | 0.082     | 0.238        | 0.122        | 0.178        | 0.159        | 0.167        | 0.269        | 0.121        | 0.166        | 0.363        | 0.109       | 0.168       |
|       |              | (0.001)     | (0.007)      | (0.002)      | (0.002)      | (0.002)      | (0)          | (0.002)      | (0.001)      | (0)          | (0.003)      | (0)         | (0.001)     |
Figure S1: Empirical power of AMAT with different testing subset selection procedures and a continuous outcome. The results correspond to scenario I, where the outcome is associated with a randomly selected set of OTUs.

Figure S2: Empirical power of AMAT with different testing subset selection procedures and a continuous outcome. The results correspond to scenario II, where the outcome is associated with a set of OTUs that are phylogenetically related.
2 Computation details for AMAT

Here, we present the details of necessary computations for AMAT in a step by step manner. In the followings, for any \( n \times p \) matrix \( A \), \( A_j \) denotes its \( j^{th} \) column.

**Input:** A vector of continuous or binary outcome \( Y_{n \times 1} \), a matrix of OTU abundances \( Z_{n \times p} \), and a matrix of additional covariates \( X_{n \times q} \).

**Output:** p-value for AMAT.

**Steps:**

1. **Normalization**
   Obtain the normalized OTU matrix \( Z^*_{n \times p} \). For this, \( Z \) is first transformed into a compositional matrix (if \( Z \) contains counts), and then the OTU-wise proportions are standardized to have zero mean and unit variance.

2. **Computation of adjusted response**
   Obtain the predicted values \( \hat{\mu}_{n \times 1} \) under \( H_0 \). Note that,
   \[
   \hat{\mu}_i = \begin{cases} 
   \hat{\beta}_0 + \sum_{k=1}^{q} X_{ik} \hat{\alpha}_k, & \text{if the outcome is continuous} \\
   \logit^{-1}(\hat{\beta}_0 + \sum_{k=1}^{q} X_{ik} \hat{\alpha}_k), & \text{if the outcome is binary}
   \end{cases}
   \]
   and compute the corresponding working residuals \( r_{n \times 1} \), where
   \[
   r_i = \begin{cases} 
   Y_i - \hat{\mu}_i, & \text{if the outcome is continuous} \\
   \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i(1-\hat{\mu}_i)}, & \text{if the outcome is binary}
   \end{cases}
   \]

3. **Computation of thresholds for testing subset**
   - Compute \( \{dc_j : j = 1, 2, \ldots, p\} \). Recall that, \( dc_j \) is the sample DC between \( r \) and the \( j^{th} \) column of \( Z \).
   - Randomly permute the elements in \( r \) \( B \)-times to obtain \( \{r^{(b)} : b = 1, 2, \ldots, B\} \).
     - If the outcome is binary, use the same set of permuted indices to obtain \( \{(Y - \mu)^{(b)} : b = 1, 2, \ldots, B\} \). This sub-step will be utilized in steps 5 and 6.
   - For each \( b \), compute DCs between \( r^{(b)} \) and columns of \( Z \) to get \( \{dc_j^{(b)} : j = 1, 2, \ldots, p\}; b = 1, 2, \ldots, B \).
   - Compute the thresholds as, \( C_j = \text{Mean}\{dc_j^{(b)} : b = 1, 2, \ldots, B\}; j = 1, 2, \ldots, p \).

4. **Testing subset selection**
   - For \( j = 1, \ldots, p \), if \( dc_j > C_j \), then \( j \in S \). If none are selected, then \( S \) contains the index of the feature having maximum DC with \( r \). Let \( S = \{j_1, \ldots, j_{|S|}\} \). Then, the testing subset corresponding to the observed data is given as, \( Z^*_S = (Z^*_{j_1}, \ldots, Z^*_{j_{|S|}}) \).
For each \( b = 1, 2, \ldots, B \), repeat the step above with \( d e_j^{(b)} (j = 1, 2, \ldots, p) \), and construct a submatrix \( Z_{S(b)}^* \). Notice that, \( Z_{S(b)}^* \) actually corresponds to \( r^{(b)} \). Thus, we obtain \( (r^{(b)}, Z_{S(b)}^*)_{b=1}^B \) for a continuous outcome, and \( (r^{(b)}, (Y - \mu)^{(b)}, Z_{S(b)}^*)_{b=1}^B \) for a binary outcome.

5. Computation of observed AMAT statistic

- Calculate the SPU statistics \( T_{SPU(\gamma)} \), for \( \gamma \in \Gamma = \{2, 3, 4, 8\} \) using \( (Y - \mu) \) and \( Z_{S}^* \). Recall that, for continuous outcome \( r = (Y - \mu) \).

Note that, \( T_{SPU(\gamma)} = \sum_{j=1}^{S} U_j^\gamma \), for \( \gamma \in \Gamma \), where \( U_j = \frac{1}{\phi} (Y - \mu)' Z_{S,j}^* \); \( j = 1, 2, \ldots, |S| \), and

\[
\phi = \begin{cases} 
\hat{\sigma}^2 & \text{if the outcome is continuous} \\
1 & \text{if the outcome is binary}
\end{cases}
\]

with \( \hat{\sigma}^2 \) being estimate of \( \sigma^2 \) under \( H_0 \).

- For each \( b = 1, 2, \ldots, B \), repeat the above step with \( ((Y - \mu)^{(b)}, Z_{S(b)}^*) \), and denote the set of resulting statistics as \( \{T_{SPU(\gamma)}^{(b)} : \gamma \in \Gamma\} \).

- p-values of the SPU tests are obtained as,

\[
P_{SPU(\gamma)} = \frac{1}{B} \sum_{b=1}^{B} I( |T_{SPU(\gamma)}^{(b)}| \geq |T_{SPU(\gamma)}^{(b)}| ); \gamma \in \Gamma.
\]

- The observed AMAT statistic is \( T_{AMAT} = \text{Min}\{P_{SPU(\gamma)} : \gamma \in \Gamma\} \).

6. Estimation of the p-value

- For each \( b = 1, 2, \ldots, B \), compute the null AMAT statistic as

\[
T_{AMAT}^{(b)} = \text{Min}\{P_{SPU(\gamma)}^{(b)} : \gamma \in \Gamma\},
\]

where \( P_{SPU(\gamma)}^{(b)} = \frac{1}{B-1} \sum_{b_1 \neq b} I( |T_{SPU(\gamma)}^{(b_1)}| \geq |T_{SPU(\gamma)}^{(b)}| ) \).

- The p-value of AMAT is estimated as,

\[
P_{AMAT} = \frac{1}{B} \sum_{b=1}^{B} I[T_{AMAT}^{(b)} \leq T_{AMAT}], \text{ where } I[\cdot] \text{ is the indicator function.}
\]

3 Additional numerical results

3.1 Generating data under previous settings

To have a more comprehensive and fairer comparison between the new AMAT method and existing methods, we have considered two additional simulation scenarios that were used in the previous MiRKAT and aMiSPU publications (Zhao et al., 2015; Wu et al., 2016). Specifically, we followed the simulation design of MiRKAT and aMiSPU and partitioned the OTUs into 20 clusters for these two additional scenarios. In the first additional scenario (named Scenario A1 hereafter), we considered a cluster with 19.4% of the total OTU reads
as the signal set, and in the second additional scenario (named Scenario A2 hereafter), we considered a cluster with 6.7% of the total OTU reads as the signal set. Scenario A1 was considered in the MiRKAT paper and Scenario A2 was considered in the aMiSPU paper.

The empirical type I error rates of all tests under these two new scenarios A1 and A2 are reported in Table S3 and Table S4. As can be seen in the tables, all tests preserve the nominal significance level of 0.05 by having empirical type I error rates rates around 0.05, except for MiHC which is conservative when the outcome type is binary. The same phenomenon is also observed in the simulation results presented in the main text. The empirical powers of all tests are presented in Figures S3–S4 and Figures S5–S6 for Scenarios A1 and A2, respectively. The overall patterns are similar to those observed in the power figures presented in the main text.

Table S3: Empirical type I error rates under Scenario A1.

| n  | AMAT | aMiSPU | LDM   | MiHC | OMiAT | OMiRKAT |
|----|------|--------|-------|------|-------|---------|
|    |      |        |       |      |       |         |
|    |      |        |       |      |       |         |
| Continuous: | | | | | | |
| 100 | 0.0484 | 0.0432 | 0.0526 | 0.0422 | 0.0466 | 0.0432 |
| 200 | 0.0516 | 0.0458 | 0.0450 | 0.0486 | 0.0432 | 0.0432 |
| Binary: | | | | | | |
| 100 | 0.0536 | 0.044  | 0.0530 | 0.0244 | 0.0462 | 0.0442 |
| 200 | 0.0490 | 0.049  | 0.0456 | 0.0162 | 0.0454 | 0.0456 |

Table S4: Empirical type I error rates under Scenario A2.

| n  | AMAT | aMiSPU | LDM   | MiHC | OMiAT | OMiRKAT |
|----|------|--------|-------|------|-------|---------|
|    |      |        |       |      |       |         |
|    |      |        |       |      |       |         |
| Continuous: | | | | | | |
| 100 | 0.0514 | 0.0478 | 0.0484 | 0.0446 | 0.0476 | 0.0456 |
| 200 | 0.0500 | 0.0470 | 0.0488 | 0.0454 | 0.0490 | 0.0438 |
| Binary: | | | | | | |
| 100 | 0.0492 | 0.0438 | 0.0458 | 0.0270 | 0.0436 | 0.0442 |
| 200 | 0.0494 | 0.0480 | 0.0436 | 0.0268 | 0.0490 | 0.0466 |
Figure S3: Empirical powers and the corresponding 95% confidence intervals obtained with a continuous outcome under Scenario A1.

Figure S4: Empirical powers and the corresponding 95% confidence intervals obtained with a binary outcome under Scenario A1.
Figure S5: Empirical powers and the corresponding 95% confidence intervals obtained with a continuous outcome under Scenario A2.

Figure S6: Empirical powers and the corresponding 95% confidence intervals obtained with a binary outcome under Scenario A2.
3.2 Simulation studies with a larger library size

We have investigated the impact of sequencing depth on the performance of all association analysis methods evaluated in the main text. In the simulations studies presented in the main text, we randomly generated the sequencing depth of each sample from a negative binomial distribution with mean 1000 and size 25. For this section, we changed the mean parameter of the negative binomial distribution from 1000 to 10000 and kept other simulation settings the same as those described in the main text. To reduce redundancy, we report the simulation results under Scenario II of the main text (with a larger library size) in this section. Simulation results under Scenarios I and III with larger library size are similar and hence not reported.

The empirical type I error rates of different methods are reported in Table S5, where all tests tend to have an empirical type I error rate around the nominal significance level except for MiHC with a binary outcome, a phenomenon which has been observed multiple times under other simulations scenarios presented in the main text and also in Section 3.1 of this Supplementary Data. The empirical powers of different tests with larger library sizes are reported in Figures S7 and S8 which are similar to their counterparts in the main text (i.e., Figure 2 and Figure 5 in the main text). To summarize, the sequencing depth seems to have little impact on the performance of different tests evaluated in this paper.

| Table S5: Empirical type I error rate with a larger sequencing depth. |
|---|---|---|---|---|---|---|
| n | AMAT | aMiSPU | LDM | MiHC | OMiAT | OMiRKAT |
|---|---|---|---|---|---|---|
| Continuous: | | | | | | |
| 100 | 0.0470 | 0.0464 | 0.0470 | 0.0420 | 0.0416 | 0.0416 |
| 200 | 0.0492 | 0.0474 | 0.0416 | 0.0526 | 0.0444 | 0.0394 |
| Binary: | | | | | | |
| 100 | 0.0514 | 0.0444 | 0.0500 | 0.0284 | 0.0450 | 0.0422 |
| 200 | 0.0536 | 0.0516 | 0.0524 | 0.0236 | 0.0420 | 0.0380 |
Figure S7: Empirical powers and the corresponding 95% confidence intervals obtained with a continuous outcome under scenario II with a larger sequencing depth.

Figure S8: Empirical powers and the corresponding 95% confidence intervals obtained with a binary outcome under scenario II with a larger sequencing depth.
3.3 Generating OTUs counts using the negative binomial distribution

In this section, we consider another data generation scheme and investigate the performance of the proposed AMAT method when applied to data generated under this new scheme. Specifically, we consider the negative binomial distribution which has been widely used in microbiome data modelling (Zhang et al., 2017). We follow the simulation settings of a previous publication (Zhang et al., 2017) that used the negative binomial distribution to simulate microbiome data in this section.

Suppose the data include $n$ subjects, a binary outcome of interest, and $p$ microbial features (e.g., abundances of $p$ OTUs). For the $i^{th}$ subject, let $Y_i$ be the binary outcome and $Z_i = (Z_{i1}, \ldots, Z_{ip})'$ be the vector of OTU counts. In order to simulate the outcome, we first generated random samples from $N(0,1)$, and then used the quantile of 40% to transform them into binary observations. The OTU counts were randomly generated from a Negative Binomial distribution: $Z_{ij} \sim \text{NB}(\mu_{ij}, \theta)$, where $\mu_{ij}$ and $\theta$ were the mean and the shape parameter respectively ($i = 1, \ldots, n; j = 1, \ldots, p$). The logarithmic link function was used to relate $\mu_{ij}$ to the outcome of interest $y_i$: $\log(\mu_{ij}) = \log(T_i) + \mu + \beta_j y_i$, where $T_i$ and $\mu$ respectively denoted the sample specific total reads and the overall mean. We set $\mu = -7$, and randomly sampled $\theta$ and $\log(T_i)$ respectively from Uniform(0.1, 5) and Uniform(7.1, 10.5) (Zhang et al., 2017). Under the null hypothesis, we set $\beta_j = 0 \ \forall j$. To evaluate the power, we formed a signal set $A$ by randomly selecting $d$ many OTUs, and simulated $\{\beta_j : j \in A\}$ from Uniform(-0.5, 0.5). Finally, following the simulation settings described earlier, we considered $d \in \{3\%, 10\%\}$, $n \in \{100, 200\}$, and $p = 856$. The corresponding results are presented in Table S6 and Figure S9 where similar phenomena on empirical type I error rates and powers have been observed as in previous numerical studies.

Table S6: Empirical type I error rates with a binary outcome when OTU counts were simulated from the negative binomial distribution.

| n   | AMAT | aMiSPU | LDM  | MiHC | OMiAT | OMiRKAT |
|-----|------|--------|------|------|-------|---------|
| 100 | 0.0526 | 0.0554 | 0.0476 | 0.0456 | 0.0526 | 0.0480   |
| 200 | 0.0540 | 0.0500 | 0.0486 | 0.0490 | 0.0510 | 0.0530   |
3.4 Application to oral microbiome study on autism spectrum disorder

Autism spectrum disorder (ASD) is associated with several oropharyngeal abnormalities such as speech apraxia and salivary transcriptome alterations. A recent study (Hicks et al., 2018) investigated the possible perturbations of oral microbiome in children (2-6 years) with ASD. Here, we are interested in comparing the oral microbiota between two groups: the ASD group \( (n = 180) \) which was characterized using the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), and the typically developing (TD) group \( (n = 106) \) which included children with negative ASD screening and children who met typical developmental milestones on standardized physician assessment. The oral microbiome compositions were quantified with next generation sequencing, and sequence alignment with the k-mer method (Ainsworth et al., 2017) was used for comprehensive taxonomic classification. A filtering of taxonomic reads to include taxa with counts of at least 10 in at least 20% of the samples resulted in a community of 753 OTUs. For further details on data processing and data availability, we refer to an earlier publication (Hicks et al., 2018). As the phylogenetic tree was unavailable in this data, at first we randomly constructed a phylogenetic tree, which was required for implementing aMiSPU, the UniFrac components in OMiAT, and MiHC. However, it turned out that the corresponding results of aMiSPU and OMiAT were very sensitive to the choice of tree (data not shown). Therefore, we simply removed the aMiSPU test and the UniFrac components from OMiAT in our oral microbiome data analysis. A total of 10000 permutations were used to establish the statistical significance in this real
data application.

We tested for a possible change in the oral microbiome community between the ASD patients and the TD controls, and the p-values from AMAT, LDM, MiHC, OMiAT, and OMiRKAT were 0.0212, 0.1418, 0.1525, 0.0636, and 0.1124 respectively. Thus, only AMAT was able to detect a significant difference between the two groups. This conclusion from AMAT bolstered the findings of the original study (Hicks et al., 2018) which conducted a univariate differential analysis with the Mann Whitney test, and reported six species with significant differences at 5% FDR. Compared to other similar methods such as aMiSPU (Wu et al., 2016), OMiAT (Koh et al., 2017), and MiHC (Koh and Zhao, 2020) that do require a phylogenetic tree in their analysis pipeline, AMAT can serve as a useful and robust analysis tool for situations where a phylogenetic tree is not instantly available, and provide a beneficial supplement to the field.

References

Ainsworth, D., Sternberg, M. J., Raczy, C., and Butcher, S. A. (2017). k-slam: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic acids research* **45**, 1649–1656.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Fan, J. and Lv, J. (2017). Sure independence screening. *Wiley StatsRef: Statistics Reference Online* pages 1–8.

Hicks, S. D., Uhlig, R., Afshari, P., Williams, J., Chronoes, M., Tierney-Aves, C., Wagner, K., and Middleton, F. A. (2018). Oral microbiome activity in children with autism spectrum disorder. *Autism Research* **11**, 1286–1299.

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.

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**, 1–16.

Saldana, D. F. and Feng, Y. (2018). Sis: an r package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software* **83**, 1–25.

Wu, C., Chen, J., Kim, J., and Pan, W. (2016). An adaptive association test for microbiome data. *Genome medicine* **8**, 56.
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*.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics* 18, 1–10.

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. *The American Journal of Human Genetics* 96, 797–807.