



### 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)



보건학 석사 학위논문

# Phylogenetic Tree-based Microbiome Association Test

계통수에 기반한 미생물과 질병 사이의  
연관성 검증 방법

2017년 08월

서울대학교 보건대학원  
보건학과 보건학전공  
김 강 진

# Phylogenetic Tree-based Microbiome Association Test

계통수에 기반한 미생물과 질병 사이의  
연관성 검증 방법

지도교수 원 성 호

이 논문을 보건학 석사 학위논문으로 제출함  
2017년 05월

서울대학교 보건대학원  
보건학과 보건학전공  
김 강 진

김강진의 석사학위논문을 인준함  
2017년 07월

위 원 장	천 종 식	(인)
부 위 원 장	고 광 표	(인)
위 원	원 성 호	(인)

## Abstract

# Phylogenetic Tree-based Microbiome Association Test

Kangjin Kim

Department of Public Health

Graduate School of Public Health

Seoul National University

**Background:** Microbial metagenomics data has large inter-subject variation and operational taxonomic units (OTU) for each species are usually very sparse. Because of these problems, non-parametric approaches such as Mann-Whitney U test and Wilcoxon rank-sum test have been utilized. However these approaches suffer from low statistical powers for association analyses and thus investigation on efficient statistical analyses is necessary.

**Objective:** Main goal in my thesis is to propose phylogenetic Tree-based Microbiome Association Test (TMAT) for association analyses between microbiome abundances of each OTU and disease phenotype.

**Method:** Phylogenetic tree reveals similarity between different OTUs, and thus was used to provide TMAT. TMAT calculates score test statistics for each node and test statistics for all nodes are combined into a single statistics by minimum p-value or Fisher's combining p-value method.

**Results:** TMAT was compared with existing methods with extensive simulations. Simulation studies show that TMAT preserves the nominal type-1 error and its statistical powers were usually much better than existing methods for considered scenarios. Furthermore it was applied to atopic diseases and found that community profiles of *Enterococcus* is associated.

**Keywords:** Microbiome, NGS, Statistical Method, Microbiome Association Test

**Student Number:** 2015-24003

## Contents

I. Introduction.....	7
II. Materials and Methods.....	9
1. Log-cpm transformation .....	9
2. Approach to use phylogenetic tree information .....	10
3. TMAT statistic.....	12
III. Results .....	15
1. Simulation Design .....	15
2. Simulation Results .....	18
3. Real Data Analysis.....	24
IV. Discussion .....	28
V. Conclusion.....	30
VI. References.....	31

## Lists of Figures

Figure 1. Data description of relative abundance data (left figure) and log-cpm transformed data (right figure). ....	10
Figure 2. Two examples of phylogenetic sub tree. ....	12
Figure 3. Power test of univariate test methods with third quantile selected node. ....	20
Figure 4. Power test of univariate test methods with median selected node .....	21
Figure 5. Power test of multivariate test methods with third quantile selected node. ....	22
Figure 6. Power test of multivariate test methods with median selected node.....	23
Figure 7. Test subgroups with the minimum p-value. ....	28

## List of Tables

Table 1. Type-1 error rates of TMAT .....	19
Table 2. Real data analysis with TMAT and univariate methods.....	25
Table 3. Real data analysis with TMAT and multivariate methods.....	26

## I. Introduction

There has been a variety attempts to try find information on the role of microbiome species or community in relation to variation in disease status. Although there are various known associations discovered between microbiome and human diseases such as diabetes, obesity, psoriasis and irritable bowel syndrome [2, 13, 14], there has been obvious drawbacks for widely-used statistical methods such for associations. This is mainly because the microbiome data has large inter-individual variation. Different combination of microbial species may have a similar role in human body. For this reason, a variety of species have zero values for many of the samples. This makes it hard to assume the distribution of microbiome abundance and non-parametric approach such as Mann-Whitney U test, Wilcoxon rank-sum test popular [5]. But, these approach has low power. Normalization techniques had been used to transform the data using arcsine-root transformation.(Morgan, Tickle et al. 2012) which results in poor type-1 error controls for low abundance species. With these difficulties, microbiome association tests had been conducted with upper-level taxa at high taxonomic ranks, genus or phylum, which conceptually sums up all of the counts or proportion of the species within the same genus or phylum with the same weights. However, OTUs are related not just by their taxonomy but by the phylogenetic tree structure. Even the species within a same genus can more or less related according to the sub-tree structure. So, different weights for different similarity of the pair would enhance the statistical power and the similarity can be described by phylogenetic information. There has been attempts to use phylogenetic information for a statistic. PERMANOVA one of the methods that used distance measure like Unifrac measure for a statistic [1][10]. A method called the microbiome regression-based kernel association test (MiRKAT) transforms phylogenetic distance matrix to a kernel. Then a kernel machine regression framework is

applied. Optimal MiRKAT integrates p-values of MiRKAT with different kind of distance matrices so that it has robust p-value for association [6]. Adaptive microbiome-based sum of powered score (aMiSPU) considered noise accumulation that is a vital problem for high-dimensional data. They described generalized taxon proportion and tried to handle the problem by assessing weight for scores functions. [11] A recent method optimal microbiome-based association test (OMiAT) took a minimum p-value of Optimal MirKAT and aMiSPU. They showed its higher performance for different kind of microbiome profile scenarios. [12]. But, they are for finding inferencing the effect of microbiome as a whole community. Even if the p-value from those methods are small enough, no one can say which species or genus has association with the outcome. For MirKAT and OMiAT, they used permutation approach to get the combined p-value. This obviously increases computational burden. TMAT (phylogenetic Tree-based Microbiome Association Test), a new statistical method for the inference of the association between microbiome abundances and disease phenotype is a novel strategy to solve such problems. TMAT can derive the exact distribution of combined p-value. So, it does not demand high computational cost. In addition to this, TMAT can find which subgroup of OTU are related to an outcome.

## II. Materials and Methods

The novel way comprises three steps: First, log-cpm transformation is applied to OTU absolute abundance data. Second, test statistics according to the subtree structure whose number is same with the number of OTUs are calculated. Third, TMAT combines the produced p-values from each statistics as minimum p-value so that it can be robust to actual type of disease-microbiome association.

### 1. Log-cpm transformation

It is hard to assume that relative abundance follows normal because the absolute abundance data is zero-inflated. To address this problem, log-cpm transformation is used. (Law, Chen et al. 2014). For the  $i^{th}$  subject,  $c_{im}$  denotes absolute abundance of  $m^{th}$  OTU and  $x_{im}$  denotes the log-cpm transformed value of  $m^{th}$  OTU where  $C_i = \sum_{m=1}^M c_{mi}$ .

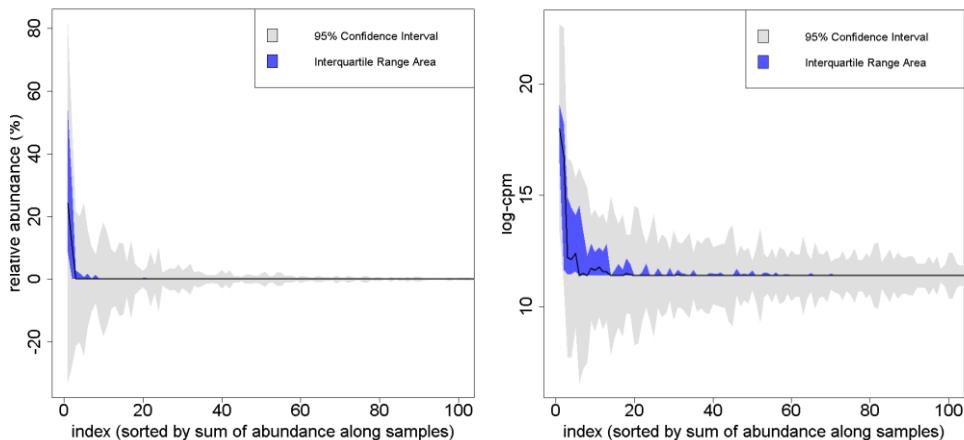
$$x_{im} = \log_2 \left( \frac{c_{im} + 0.5}{C_i + 1.0} \times 10^6 \right)$$

The absolute abundance of 125 samples with 558 species from a cohort study, COCOA (COhort for Childhood Origins of Asthma and Allergic Diseases) are used to generate a relative abundance dataset and log-cpm transformd dataset and they are described. (Figure 1.) Since the abundance is zero-inflated, inter quantile range of relative abundance is around zero for almost

all of the species. But, log-cpm transformed abundance has moderate change of variance and inter-quantile range along all the species, helping us to assume it to follow normal distribution.

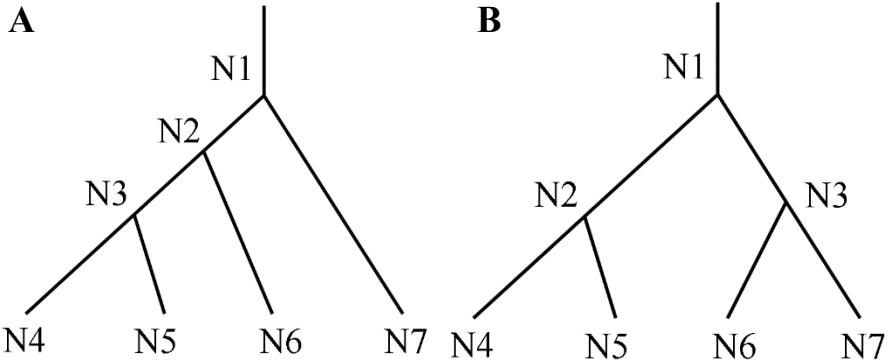
## 2. Approach to use phylogenetic tree information

Let's say that we have a rooted tree that contains OTUs as tips. A set that contains all the OTU tips to be tested is called a *test group* and a node that has only the test group as tips is called a *test node*. Let a set of tips contained in a child node of a test node be a *test subgroup*. Every



**Figure 1. Data description of relative abundance data (left figure) and log-cpm transformed data (right figure).** OTUs are sorted by its sum of abundances along samples from the most abundant OTU to the least abundant OTU. 95% confidence interval is a band with  $\pm 1.96 \cdot$  standard deviation of abundance of each OTUs. Interquartile range is the difference between 75<sup>th</sup> and 25<sup>th</sup> percentiles

internal node has two child nodes as long as the tree is rooted. This means that every test node has two test subgroups. So, for case A in Figure 2, the novel method put N1 be the first test node, then N4-7 are in the test group and N4-6 are in a same test subgroup and N7 is in the other test subgroup. Now, sum up the abundances of OTUs among each test subgroups and calculate relative abundance between two subgroups. Let say  $A_{Ni}$  be the sum of the absolute abundances of the OTU tips contained in the node  $Ni$ . Then the value from the data for a first test statistic is  $A_{N2}/(A_{N2} + A_{N7})$  or  $A_{N7}/(A_{N2} + A_{N7})$ . You can use either of them and will get the same results. The data for a second statistic is  $A_{N3}/(A_{N3} + A_{N6})$  or  $A_{N6}/(A_{N3} + A_{N6})$ . The data for a third statistic is  $A_{N4}/(A_{N4} + A_{N5})$  or  $A_{N5}/(A_{N4} + A_{N5})$ . For the last statistic, use the value of  $A_{N1}$ . For case B, the first one is  $A_{N2}/(A_{N2} + A_{N3})$  or  $A_{N3}/(A_{N2} + A_{N3})$ , the second one is  $A_{N4}/(A_{N4} + A_{N5})$  or  $A_{N5}/(A_{N4} + A_{N5})$ , the third one is  $A_{N6}/(A_{N6} + A_{N7})$  or  $A_{N7}/(A_{N6} + A_{N7})$  and the last one is  $A_{N1}$ . the sequence of second and third one can be changed. One of the benefits we can gain with the novel method is that the statistics are independent. For a certain sample, appearance of a read for each two test subgroups follows binomial distribution with the number of trial as total number of reads within the species contained in test node. For case the first statistic of case A, exclusively depends on  $A_{N7}$  and the second statistic exclusively depends on  $A_{N6}$  which represents that those two statistics are independent.



**Figure 2.** Two examples of phylogenetic sub tree.

### 3. TMAT statistic

**Notation** A statistic for  $k^{th}$  test node is produced in this section. Assume that  $N$  samples from a particular body site of each individuals are collected. Let  $M$  be the total number of OTUs and  $Q$  is that of phenotypes. For  $i^{th}$  subject,  $y_i$  denotes a phenotype,  $x_{im}$  is the log-cpm transformed value of  $m^{th}$  OTU.  $y_i$  is coded as 1 for affected individuals and 0 for unaffected individuals.  $w_{mk}$  is 1, if  $m^{th}$  OTU is the test group. Let's denote  $\delta_{mk}$  is 1, if  $m^{th}$  OTU is the test subgroup otherwise both are 0. Let's denote

$$\mathbf{x}^m = \begin{pmatrix} x_{1m} \\ \vdots \\ x_{Nm} \end{pmatrix}, \mathbf{X} = (\mathbf{x}^1 \quad \dots \quad \mathbf{x}^M),$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

$$\delta^k = \begin{pmatrix} \delta_{1k} \\ \vdots \\ \delta_{Mk} \end{pmatrix}, w^k = \begin{pmatrix} w_{1k} \\ \vdots \\ w_{Mk} \end{pmatrix}.$$

$$\mathbf{z}^k = [\mathbf{1}_M^t \delta^k]^{-1} \mathbf{X} w^k$$

$$\mathbf{z}_{abs}^k = \mathbf{X} w^k$$

Then  $\mathbf{z}^k$  represents the relative log-cpm transformed abundance between the two test subgroups for  $k^{th}$  test node

**Disease model and score test statistics for each test nodes** When we assume that variance-covariance matrix of  $\mathbf{z}^k$  as  $\sigma_k^2 \mathbf{I}_N$ , we can denote

$$var(\mathbf{z}^k) = \sigma_k^2 \mathbf{I}_N$$

In retrospective analysis, individuals are selected based on their phenotypes and then we compare microbial distributions between affected and unaffected individuals. Furthermore, the analysis is robust to non-normality of phenotype with retrospective analysis. Hence, the following can be assumed.

$$\mathbf{z}^k | \mathbf{y} = N(\alpha_k \mathbf{1}_N + \beta_k \mathbf{y}, \sigma_k^2 \mathbf{I}_N)$$

So, the score function is

$$\mathbf{Y}^t (\mathbf{z}^k - E(\mathbf{z}^k)).$$

To find  $\hat{\beta}_k$  that maximizes the profile likelihood,  $\hat{E}(\mathbf{z}^k)$  is estimated as MLE under null hypothesis.

$$\hat{E}(\mathbf{z}^k) = \mathbf{1}_N (\mathbf{1}_N^t \mathbf{1}_N)^{-1} \mathbf{1}_N^t \mathbf{z}^k.$$

If we let

$$\mathbf{A} = \mathbf{I}_N - \mathbf{1}_N (\mathbf{1}_N^t \mathbf{1}_N)^{-1} \mathbf{1}_N^t,$$

$\mathbf{A}$  is idempotent matrix and we have a score  $\mathbf{s}$  function

$$\mathbf{s} = \mathbf{y}^t (\mathbf{I}_N - \mathbf{1}_N (\mathbf{1}_N^t \mathbf{1}_N)^{-1} \mathbf{1}_N^t) \mathbf{z}^k = \mathbf{y}^t \mathbf{A} \mathbf{z}^k.$$

So, for  $k^{th}$  test node, the variance-covariance matrix of the score is

$$var(\mathbf{s}) = var(\mathbf{y}^t \mathbf{A} \mathbf{z}^k) = \sigma_k^2 \mathbf{y}^t \mathbf{A} \mathbf{A}^t \mathbf{y} = \sigma_k^2 \mathbf{y}^t \mathbf{A} \mathbf{y}.$$

Hence the score test statistic for  $k^{th}$  test node is

$$T_k = \frac{1}{\sigma_k^2 \mathbf{y}^t \mathbf{A} \mathbf{y}} \mathbf{s} \mathbf{s}^t \sim \chi_1^2$$

**Combining statistics as a minimum p-value** A minimum p-value of sequence of statistics yields a robust statistic with decent power regardless of the type of the structure of phylogenetic tree. Let  $K$  be the number of OTUs within  $k^{th}$  test node,  $pT_k$  be p-value of  $T_k$ . The number of test node is same with that of OTUs.

$$\begin{aligned} p_{min} &= \min\{pT_1, \dots, pT_K\} \\ P(p_{min} \leq p*) &= P(\min\{pT_1, \dots, pT_K\} \leq p*) \\ &= 1 - P(\min\{pT_1, \dots, pT_K\} \geq p*) = 1 - (1 - p*)^K \end{aligned}$$

The minimum p-value is TMAT statistic. When  $\mathbf{z}_{abs}^k$  is used in replacement of  $\mathbf{z}^k$ , we call the statistic as absolute TMAT statistic.

### III. Results

#### 1. Simulation Design

The absolute abundance of 89 samples with 6339 species from a cohort study, COCOA (COhort for Childhood Origins of Asthma and Allergic Diseases) are used to generate absolute abundance dataset. 97% OTUs were constructed with QIIME pipeline. OTUs are aligned to the reference sequence of the 16S rRNA Silva database (release 123) for QIIME. log-cpm transformation is conducted yielding a 89 by 6339 matrix with log-cpm transformed abundance values. OTUs that presents in fewer than 25% of the samples or whose abundance is less than 0.01% is filtered out producing a dataset with 89 samples and 558 species. The corresponding Silva reference tree is pruned to have only the filtered OTUs as tips. I got subtrees that has tree structure of each upper-level taxa (phylum, class, order, family and genus) from this pruned tree. For this simulation, the upper-level taxa is genus level. For each subtrees, the pruned tree is pruned again so that its root can contains all the species of a specific upper-level taxa in the tree and no other species as tips. Each subtrees tells us which nodes will be the test nodes and guides us to make a sequence of statistics.

Simulation study is performed to ensure that TMAT correctly controls type-1 error. Real dataset is used to conduct the simulation study and applied permutation approach to obtain empirical null distribution of test statistics. Labels of the datasets are shuffled yielding 2000 permuted datasets. At first, OTU tips that has association with outcome are selected. I call this pathogenic OTU or *pathogenic tips*. For each pathogenic OTUs, beta multiplied by standard deviation of abundances is added to abundances of each permuted datasets

Specifically,  $c_{im}$  is absolute abundance of  $m^{th}$  OTU for  $i^{th}$  sample and  $\mathbf{c}_m$  is N-length vector whose elements are  $c_{im}$ .  $\sigma_m$  denotes standard deviation of  $\mathbf{c}_m$ .  $I_{A_j}(m)$  is an indicator function. For each permutation  $j$ ,  $A_j$  is the indices of pathogenic OTUs and  $\mathbf{y}^j$  denotes permuted dichotomized or continuous phenotype information. Absolute abundance  $\mathbf{c}_m^j$  is simulated under the following model.

$$\mathbf{c}_m^j = \mathbf{c}_m + \beta \sigma_m I_{A_j}(m) \mathbf{y}^j$$

**The way to simulate  $A_j$**  For each subtree, TMAT is calculated based on the phylogenetic information of the subtree. An internal node is selected for each subtree. Let this node be *selected node*. Now, let's assume that some of the tips of the selected node are pathogenic tips.  $p$  is defined as the ratio of the number of pathogenic tips to the number of tips within the selected node. When  $p$  is multiplied by the number of tips of the selected node for each representative nodes, the number of pathogenic OTUs for each representative nodes is determined. If the multiplied value is not an integer, each simulation datasets takes a minimum integer that is less than the multiplied value or a maximum integer that is greater than the multiplied value so that the weighted average of the number of pathogenic OTUs in the 2000 repeating simulation dataset makes the correct  $p$ .

In detail, let say  $S_j$  is the \*indices of OTU tips under the selected node for permutation  $j$ . The cardinality of set A is denoted as  $|A|$ . The maximum integer that is less than a real number  $a$  is denoted as  $[a]$ .  $n_r$  is the number of replication. In this case,  $n_r = 2000$ . For any choice of  $p$  in simulation scenarios,  $n_r p$  has an integer value. Then for each permutation  $j$ , the elements of  $A_j$  are randomly selected among the tips of the selected node where

$$|A_j| = \begin{cases} [p|S|] & j = 1, \dots, n_r - n_r(p|S| - [p|S|]) \\ [p|S|] + 1 & j = n_r(p|S| - [p|S|]) + 1, \dots, n_r \end{cases}$$

Then these  $A_j$   $j = 1, \dots, n_r = 2000$  satisfies the following equation.

$$\frac{\sum_{j=1}^{2000} |A_j|}{n_r |S|} = p$$

**Simulation scenarios** It is plausible that the OTUs associated with the outcomes could be distributed close in phylogenetic trees. High  $p$  illustrates this scenario. In the real world, however, the pathogenic OTUs could also be distributed regardless of how close they are in the tree. Low  $p$  depicts this situation. To confirm the robustness of TMAT, the scenarios of  $p = 0.3$ ,  $p = 0.5$ ,  $p = 0.9$  are considered that reflect how pathogenic OTUs are related to phylogenetic tree structures. I also considered a situation where only one OTU in the subtree is associated with the outcome and assigned a zero value for  $p$  for this situation.

I also considered three different choice of selected nodes. The internal nodes of each subtree are sorted by the number of tips they have. The nodes with median and 75 percentile are chosen.

To verify that TMAT outperform other univariate tests, TMAT, aTMAT with combined p-values of Score tests and Wilcoxon test are evaluated for each scenarios and different choice of selected nodes. The performances of TMAT and aTMAT with multivariate test methods, Genus test, Optimal MirKAT, aMiSPU and OMiAT are also considered.

**The way to combine p-values of each OTUs for univariate tests** The p-values of TMAT, aTMAT, Score tests and Wilcoxon rank sum test are combined by false discovery rate [16], Bonferroni correction using `p.adjust` function in R statistical software with `fdr` and `bonferroni` method option. The way to calculate exact distribution of minimum p-value under the assumption of independence of each p-values is also considered.

**Optimal MirKAT, aMiSPU and OMiAT** Even if MiRKAT, MiSPU and OMiAT is designed for testing for overall composition of microbiome community, they describes how to use them to identify each individual taxa that makes the associations with phenotype. Each OTUs are grouped as their genus-level taxa and tested respectively. The null hypothesis of test with

MiRKAT, MiSPU and OMiAT is that there is no association between the microbiome communities of OTUs belonged to each genus and disease phenotype. Since MiRKAT, MiSPU and OMiAT are designed for testing for entire microbiome community, they assumes that all of the sample has more than one count. So, samples with zero abundance for each genus are eliminated before the tests. MiRKAT, MiSPU and OMiAT tests in R statistical software package are used. Version of them are 0.02, 1.0 and 5.1 respectively. For all of them, the option of the number of permutation are set to 1000 and all the other options are set to default. For MiRKAT, rarefying step is preceded the test as the manual of the package describes.

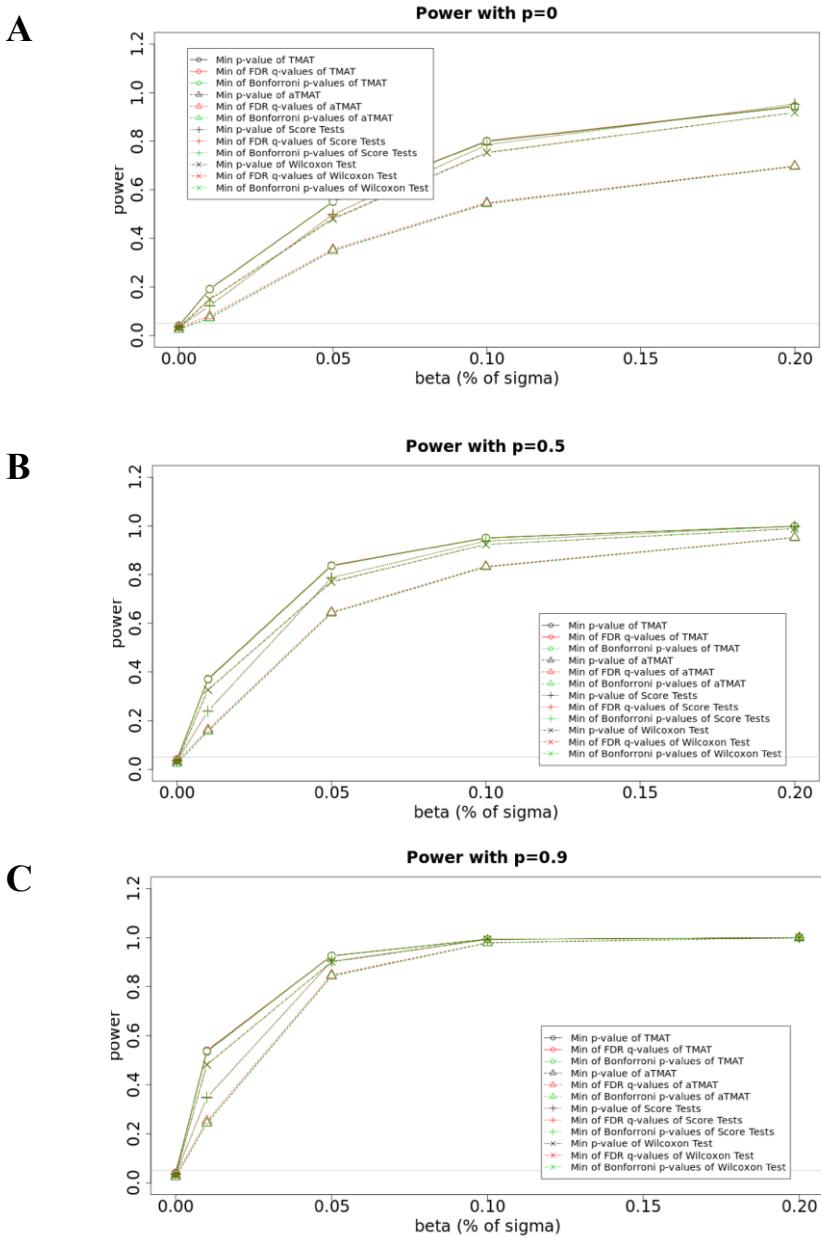
## 2. Simulation Results

Two thousands datasets with permutation of labels for type 1 error test for TMAT are used. Power test of TMAT, aTMAT, Score Test, Wilcoxon Test, MirKAT, aMiSPU and OMiAT are conducted. Two hundreds permuted datasets are used for the power tests because MirKAT, aMiSPU and OMiAT has their own permutation step making each calculation from permuted datasets heavy. Type-1 error rate of TMAT test of each genus is evaluated ratio of p-value below the significance level ( $\alpha$ ). As presented in Table.1, TMAT controls type-1 error well for all the genus-level OTUs with a variety choices of  $\alpha$ . In figure 3, TMAT out performs any other non-multivariate methods no matter what  $p$  is chosen. TMAT definitely outperformed the widely used univariate non-parametric method, Wilcoxon test. This can originate from the use of distribution assumption and phylogenetic tree information. Big difference between the power of TMAT and aTMAT illustrates the impact of the way we handle the tree information. Taking the relative proportion between two subgroups would enhance the statistical power. In comparison of figure 3 and 4, TMAT is robust to the choice of selected nodes. For the comparison with multivariate methods in figure 5, aMiSPU performs better than optimal

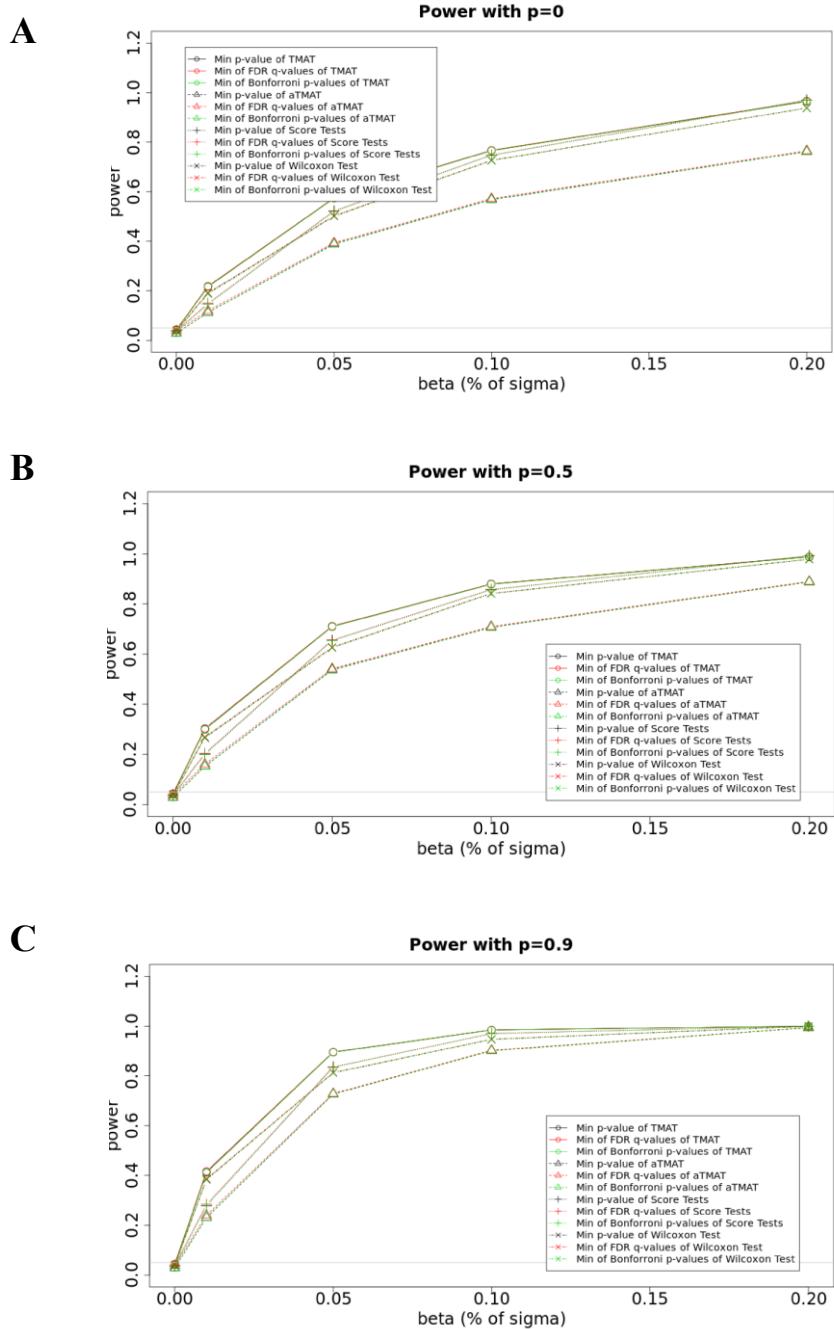
MiRKAT when many OTUs have small correlation with phenotype, in other words, when beta in not large enough and p is large. Optimal MiRKAT outperform aMiSPU when few OTUs have large association with outcome. OMiAT that combined these two methods is evaluated better than them no matter what scenario is set. Even if OMiAT used permutation approach and TMAT is much faster than OMiAT to calculate the statistic, there was no big difference in terms of statistical power between OMiAT and TMAT. There is also a tendency that TMAT gets better than OMiAT as beta gets larger. The result is robust to the choice of selected nodes in this case either.

**Table 1. Type-1 error rates of TMAT**

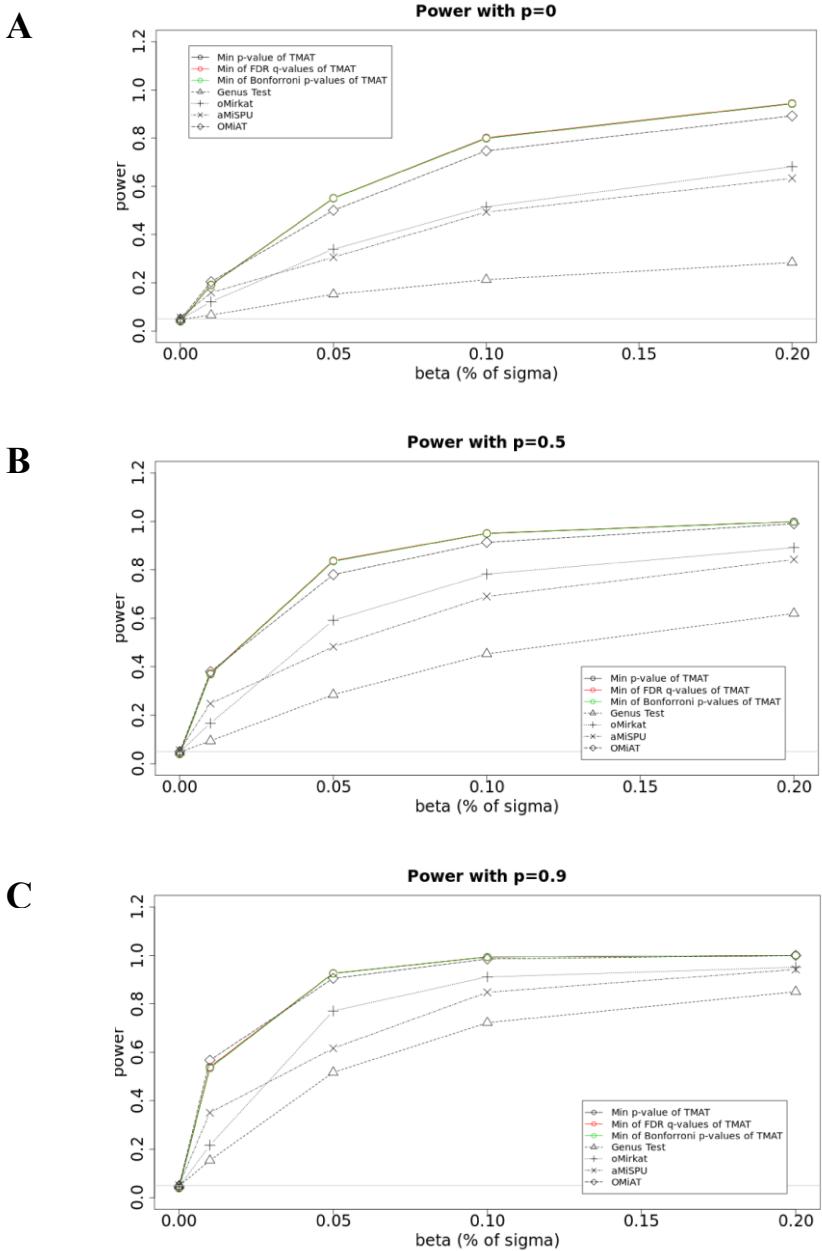
Genus Name	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.005$
<i>Bifidobacterium</i>	0.0870	0.0376	0.0076	0.0026
<i>Citrobacter</i>	0.0900	0.0482	0.0096	0.0052
<i>Cronobacter</i>	0.0886	0.0432	0.0098	0.0036
<i>Enterobacter</i>	0.0716	0.0344	0.0056	0.0032
<i>Enterococcus</i>	0.0742	0.0328	0.0046	0.0018
<i>Escherichia-Shigella</i>	0.0732	0.0336	0.0040	0.0018
<i>Klebsiella</i>	0.0856	0.0368	0.0060	0.0030
<i>Lactobacillus;</i>	0.0980	0.0488	0.0066	0.0016
<i>Pantoea</i>	0.1000	0.0462	0.0064	0.0040
<i>Raoultella</i>	0.1022	0.0498	0.0080	0.0042
<i>Rhodococcus</i>	0.0958	0.0478	0.0110	0.0048
<i>Streptococcus</i>	0.0718	0.0364	0.0070	0.0042
<i>uncultured bacterium</i>	0.0784	0.0368	0.0084	0.0040
<i>Veillonella</i>	0.0704	0.0328	0.0066	0.0024



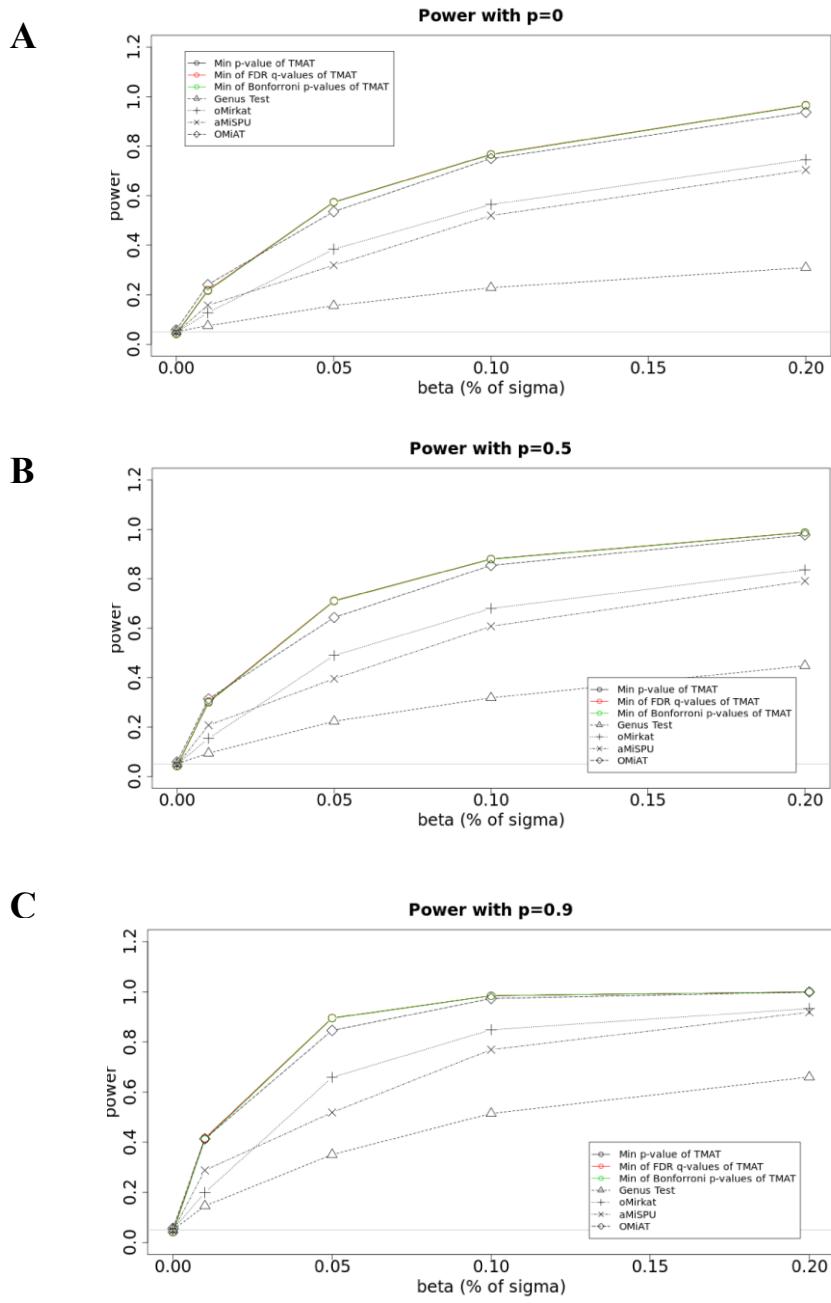
**Figure 3.** Power test of univariate test methods with third quantile selected node.



**Figure 4. Power test of univariate test methods with median selected node**



**Figure 5. Power test of multivariate test methods with third quantile selected node.**



**Figure 6. Power test of multivariate test methods with median selected node.**

### 3. Real Data Analysis

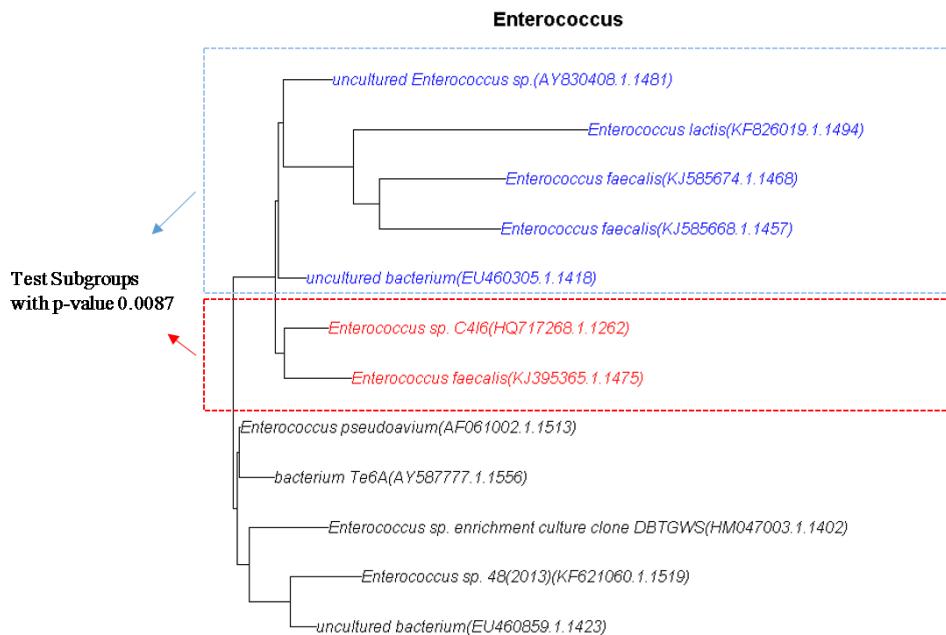
The COCOA data mentioned in simulation section is used for real data analysis. The purpose of the analysis is to test the association between microbiome and Atopic disease. The results in table 2 says the p-value trends of TMAT is almost same with the previous widely used univariate test methods which enhances the validity of TMAT. No significant association was found with all the univariate methods and TMAT. Significance of associations between each genus and atopic disease with multivariate approach are described in table 3. While TMAT does not give any significant p-value, optimal MiRKAT and OMiAT, the combined statistic with optimal MiRKAT and aMiSPU, give p-value that is lower than 0.05 for *Enterococcus*. But, note that MiRKAT, aMiSPU and OMiAT eliminates samples with zero count for each genus. This could bias the result of the analysis. Zero Prop section in the table 3 is calculated as the number of samples with zero count for the sum of abundances of OTUs under the genus over the total number of samples. Previous research found out that an increasing tendency of allergen-specific IgG2a level in mice after lysed *Enterococcus faecalis* FK-23 treatment for 21 days compared with controls. [15] Even though TMAT couldn't find significant correlation in this dataset, TMAT can find what OTUs exactly correlated with the outcome. The test groups with the minimum p-value can be described as Figure 7. This shows that the relative abundance of the sum of *Enterococcus lactis*, subspecies of *Enterococcus faecalis* and other uncultured sub species to that of a subspecies of *Enterococcus faecalis* and *Enterococcus sp. C416* has association with the outcome.

**Table 2. Real data analysis with TMAT and univariate methods**

Genus Name	Number of species	TMAT	FDR_Score_Tests	Wilcoxon
<i>Bifidobacterium</i>	7	0.2753	0.3441	0.4780
<i>Citrobacter</i>	4	0.2277	0.1432	0.0842
<i>Cronobacter</i>	4	0.2066	0.1748	0.4294
<i>Enterobacter</i>	28	0.9622	0.9061	0.9087
<i>Enterococcus</i>	12	0.0995	0.0681	0.0537
<i>Escherichia-Shigella</i>	47	0.1321	0.7502	0.6732
<i>Klebsiella</i>	9	0.3405	0.3609	0.6645
<i>Lactobacillus</i>	3	0.6992	0.4094	0.4065
<i>Pantoea</i>	7	0.5291	0.7635	0.7304
<i>Raoultella</i>	3	0.7919	0.6539	0.6575
<i>Rhodococcus</i>	3	0.5702	0.5701	0.5089
<i>Streptococcus</i>	21	0.9331	0.4790	0.5134
<i>uncultured bacterium</i>	15	0.9788	0.8152	0.7726
<i>Veillonella</i>	23	0.9898	0.7578	0.6997

**Table 3. Real data analysis with TMAT and multivariate methods**

Genus Name	Zero Prop	TMAT	GENUS	oMiRKAT	aMiSPU	OMiAT
<i>Bifidobacterium</i>	4.49%	0.2753	0.3075	0.3921	0.2657	0.4250
<i>Citrobacter</i>	20.22%	0.2277	0.7515	0.0644	0.0090	0.0050
<i>Cronobacter</i>	19.10%	0.2066	0.2822	0.2962	0.1359	0.2130
<i>Enterobacter</i>	5.62%	0.9622	0.8522	0.3352	0.6733	0.5730
<i>Enterococcus</i>	7.87%	0.0995	0.1727	0.0285	0.0799	0.0350
<i>Escherichia-Shigella</i>	11.24%	0.1321	0.5265	0.3826	0.2198	0.2930
<i>Klebsiella</i>	16.85%	0.3405	0.2890	0.1738	0.6124	0.7060
<i>Lactobacillus</i>	41.57%	0.6992	0.3299	0.1683	0.6024	0.5730
<i>Pantoea</i>	25.84%	0.5291	0.9257	0.8741	0.8831	0.9420
<i>Raoultella</i>	38.20%	0.7919	0.4074	0.8671	0.8651	0.8080
<i>Rhodococcus</i>	12.36%	0.5702	0.9456	0.4141	0.5075	0.7740
<i>Streptococcus</i>	4.49%	0.9331	0.1208	0.4076	0.7552	0.8720
<i>uncultured bacterium</i>	3.37%	0.9788	0.6215	0.9101	0.9770	0.8860
<i>Veillonella</i>	7.87%	0.9898	0.3293	0.6953	0.5355	0.9820



**Figure 7. Test subgroups with the minimum p-value.**

---

## IV. Discussion

As TMAT utilize phylogenetic tree information in a distinctive way, it has larger power than any other univariate methods and has same or higher level power with multivariate approaches. Well controlled type-1 error even in a variety of simulation scenarios verified the validity of TMAT also. In addition to this, TMAT is robust to actual type of disease-microbiome association as it uses minimum p-value as its statistic.

The computational cost of TMAT is much less than other multivariate methods. Optimal MiRKAT, aMiSPU and OMiAT uses permutation approaches to calculate the combined statistic. Even though TMAT is also an omnibus test, the exact distribution of the combined statistic can be calculated. Some other multivariate methods have slightly better performance than TMAT in some scenarios, but with limited time and resources TMAT can be a better choice.

Advantages of using prediction approaches such as Linear discriminant analysis Effect Size (LEfSe) and Random forests compared to multivariate approaches such as MiRKAT, aMiSPU and OMiAT are that prediction approaches can rank OTUs according to their contribution scores and intuitively displays the effect size. TMAT can also explain what groups of OTUs are related to the results and how they are associated. (Figure 7). TMAT can find the test node with minimum p-value and represents the OTUs contained in each test subgroups under the test node. Researchers and biologists can develop their intuitions and research plans based on the correlation between relative abundance of two test subgroups and outcome. The advantage over prediction methods is that the TMAT can produce an overall p-value with a statistically controlled error rate.

Microbiological studies are carried out at the species level and are further studied in subspecies units. It focuses on multivariate studies that take into account not only individual

microbes but also microbial associations. In addition to this, microbial studies are increasingly focused on functional gene studies using metagenomic shotgun sequencing. TMAT will be a powerful option in this trend as it utilizes the abundance and structure information of each species or subspecies in an effective way. Microbiome abundance data based on sequencing the gene 16S rRNA can easily be replaced with the data from metagenomic shotgun sequencing. The framework of TMAT can be generalized to multivariate phenotypes, repeated response variables and survival outcomes and covariate adjustment.

## V. Conclusion

There have been problems for widely-used statistical methods for finding the role of microbiome species or community in relation to variation in disease status. The power of the score test or non-parametric test approach of microbiome species can be enhanced when phylogenetic tree information is used. TMAT not only reflects phylogenetic tree information but also robust to actual type of disease-microbiome association. Since the sequence of statistics are set to be independent each other, the minimum p-value can be calculated explicitly which helps to save computational costs. I conducted studies to verify that type-1 error is well controlled and power is higher than any other univariate methods and similar with that of OMiAT which demands massive computational burden. The specific group of OTUs that has high correlation with outcome can be also detected. The relative proportion of two test subgroups that yielded minimum p-value can be considered to have association with phenotype. This can help biologists to conduct further studies or experiments regarding the results of the research with TMAT.

## VI. References

1. Chen, J., et al. (2012). "Associating microbiome composition with environmental covariates using generalized UniFrac distances." *Bioinformatics* 28(16): 2106-2113.
2. de Vos, W. M. and E. A. de Vos (2012). "Role of the intestinal microbiome in health and disease: from correlation to causation." *Nutrition reviews* 70(suppl 1): S45-S56.
3. Law, C. W., et al. (2014). "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts." *Genome biology* 15(2): 1.
4. Morgan, X. C., et al. (2012). "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment." *Genome biology* 13(9): 1.
5. Zhang, X., et al. (2015). "The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment." *Nature medicine*.
6. Zhao, N., et al. (2015). "Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test." *The American Journal of Human Genetics* 96(5): 797-807.
7. J. Si, et al (2015). "Genetic associations and shared environmental effects on the skin microbiome of Korean twins." *BMC Genomics* 16:992

8. Liang, Xue, Frederic D. Bushman, and Garret A. FitzGerald (2015). "Rhythmicity of the intestinal microbiota is regulated by gender and the host circadian clock." *Proceedings of the National Academy of Sciences* 112.33: 10479-10484.
9. D. Gevers, et al. (2015) "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host & Microbe* (2014) 15, 382–392ircadian clock. PNAS 112, 10479-10484
10. McArdle BH, Anderson MJ. Fitting multivariate data: a comment on distance-based redundancy analysis. *Ecology.models to community* 2001;82(1):290–7.
11. Wu, Chong, et al. "An adaptive association test for microbiome data." *Genome medicine* 8.1 (2016): 56.
12. Koh, Hyunwook, Martin J. Blaser, and Huilin Li. "A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping." *Microbiome* 5.1 (2017): 45.
13. Turnbaugh, Peter J., et al. "An obesity-associated gut microbiome with increased capacity for energy harvest." *nature* 444.7122 (2006): 1027.
14. Gao, Zhan, et al. "Substantial alterations of the cutaneous bacterial biota in psoriatic lesions." *PloS one* 3.7 (2008): e2719.
15. Shimada, T., et al. "Effects of lysed Enterococcus faecalis FK-23 on allergen-induced serum antibody responses and active cutaneous anaphylaxis in mice." *Clinical & Experimental Allergy*

34.11 (2004): 1784-1788.

16. Benjamini, Yoav, et al. "Controlling the false discovery rate in behavior genetics research."

Behavioural brain research 125.1 (2001): 279-284.

국문 초록

## 계통수에 기반한 미생물과 질병

### 사이의 연관성 검증 방법

김 강 진

서울대학교 보건대학원

보건학과 보건학전공

**연구배경:** 미생물 종 또는 군집이 질병 상태 변화에 어떠한 역할을 하는가를 연구하는 데 있어 기존의 통계적 방법론에 문제점이 있다. 이것은 주로 미생물 메타 지노믹 데이터의 특성에 기인한다. 미생물의 종류 및 존재량이 개인 간의 차이가 크고 이로 인해 미생물의 양에 대한 통계적 분포를 가정하기가 힘들다. 이러한 결과로 Mann-Whitney U test, Wilcoxon rank-sum test 등의 비모수적 접근법이 널리 쓰이고 있다.

**연구목적:** 이러한 비모수적 방법은 통계적 검정력이 낮은 문제점이 있고 따라서 미생물의 분포를 가정하기 위한 보완책으로써 arcsine-root transformation, log transformation 등의 여러 정규화 방법들이 사용되었지만 type-1 error를 control하는 데 있어 약점을 보이는 경향이 있다. 계통수에 담긴 정보가 통계적 검정력을 향상시킬 수 있음이 연구되었고 Unifrac과 같은 계통 발생정보를 사용

한 통계량들이 개발되었지만 이들은 전체 미생물의 군집의 분포와 질병상태의 관계를 추론하는 데 사용됩니다. 본 논문은 개별 미생물의 존재량과 질병 사이의 연관성의 추론 및 검정을 위한 계통수에 기반한 미생물과 질병 사이의 연관성 검증방법 Phylogenetic Tree-based Microbiome Association Test 을 제안합니다. (TMAT)

**연구방법:** 우리는 본 논문에서 계통수에 기반하여 통계량의 수열을 만들고 이를 최소 p 값으로 결합하는 방법을 제안한다. 통계량의 검정력이 낮은 문제를 해결하기 위한 기준의 방법으로는 같은 종이나 문에 속한 모든 생물 종의 양을 합치는 방법이 있다. 그러나 OTU 들은 그들의 분류법 뿐만 아니라 그들의 계통수의 구조와도 관련이 있다. TMAT 은 하위 트리 구조에 따라 통계량을 계산하고 각 통계량의 p 값을 최소 p 값으로 결합한다. 이전의 방법들과 달리 TMAT 은 최소 p 값의 정확한 분포를 도출할 수 있고 따라서 이전의 방법들 보다 계산 비용 측면에서 효율적이다. 이외에도 TMAT 은 유의하게 나타난 OTU 의 하위 그룹 또한 찾아낼 수 있다

**연구결과:** TMAT 은 계통수의 정보를 반영할 뿐만 아니라 질병과 미생물 간의 다양한 유형의 상관관계에도 robust 하다. 시뮬레이션 분석을 통해 TMAT 의 1 종 오류가 잘 제어되고 통계적 검정력이 대부분의 시뮬레이션 시나리오에서 기준에 사용되던 방법보다 높았다. 실제 데이터 분석 결과 *Enterococcus* 속이 아토피 질환과 관련되어 있음이 나타났고 이 결과는 기존의 타 논문 연구 결과와 일치합니다.

**주요어:** 미생물, NGS, 통계적 방법론, 미생물 연관성 연구

**학번:** 2015-24003