Ph.D. Dissertation of Public Health

Statistical Method Development of 16S rRNA Metagenomics-based Association Analysis and its Application

16S rRNA 메타 유전체 연관성 분석
통계 방법론의 개발과 적용

August 2021

Graduate School of Public Health
Seoul National University
Public Health Major

Kangjin Kim

16S rRNA 메타 유전체 연관성 분석 통계 방법론의 개발과 적용

지도교수 원 성 호

이 논문을 보건학박사 학위논문으로 제출함
2021년 8월

서울대학교 보건대학원
보건학과 보건통계학 전공
김 강 진

김강진의 보건학박사 학위논문을 인준함
2021년 8월

위 원 장 ______ 성 주 헌 ______

부위원장 ______ 박 태 성 ______

위　　원 ______ 이 우 주 ______

위　　원 ______ 이 하 나 ______

위　　원 ______ 원 성 호 ______

Abstract

Statistical Method Development of 16S rRNA Metagenomics-based Association Analysis and its Application

Background:

Increased availability of affordable sequencing technology and advances in throughput technology have led to the birth and widespread development of a new scientific discipline, metagenomics that includes large-scale analysis of microbial communities. However, analysis with metagenomics data suffers from compositional bias and zero-inflated problems, and the statistical methods available for association analysis with 16S rRNA data is very limited, especially for the repeatedly observed 16S rRNA data. Therefore investigation on the statistical method and software development is necessary.

Objective:

The main goal is (1) to develop new methods with cross-sectional and repeatedly observed 16S rRNA data that correct for the problems including compositional bias, zero-inflation and package implementation that can unify the preprocessing procedures; (2) to identify microorganisms which can be affect type-2 diabetes (T2D)-related traits with repeatedly observed 16S rRNA data.

Methods:

To consider the characteristics of microbiome data and correct compositional bias and zero-inflated problem, the phylogenetic tree based method, TMAT, and its extension to the repeatedly observed 16S rRNA measurement, mTMAT, were developed. I also implemented a new package that can generate both statistics, and conduct OTU clustering with different databases. This package also allows the comparison of different statistics. Furthermore, association analysis of microorganisms with T2D were conducted by using repeatedly measured EV in urine samples. EV-derived metagenomic (N = 393), clinical (N = 5032), and metabolite (N = 574) data were observed for a prospective and longitudinal Korean community-based cohort (KARE) three times and genetic data was available. They were analyzed with generalized linear mixed model to identify microbes associated with T2D and their interaction with metabolites.

Results and Conclusions:

The proposed phylogenetic tree-based microbiome association test (TMAT) normalized microbial abundances and pooled abundances based on the phylogenetic tree structure was utilized for association analysis. Results from simulation studies showed that TMAT correctly controls type-1 error rates, and statistically more powerful. Second, I also implemented all-inclusive microbiome association analysis (AMAA) package. AMAA package provides the analysis result of various methods including TMAT under a unified preprocessing and allows comparison of the results based on different databases or clustering methods. Third, mTMAT which is the extended version of TMAT for repeatedly measured 16S rRNA data was developed. It uses generalized estimating equations with robust variance estimator and can be applied to repeated measured samples. Statistical power of mTMAT was superior to existing methods in terms of controlling the type-1 error and minimizing the type-2 error, and it is robust against the compositional bias. Fourth, from the association analysis with repeatedly measured EV-based metagenome data, it was found that *GU174097_g*, an uncultured *Lachnospiraceae*, was associated with T2D (β = −189.13; p = 0.00006). These results indicates that *GU174097_g* may decrease the HbA1c level and the risk of T2D.

Keyword: statistical method, microbiome association test, longitudinal data analysis, multi-omics.

Student Number: 2018-34334

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Study Background

Increased availability of affordable sequencing technology and advances in throughput technology have led to the birth and widespread development of a new scientific discipline, metagenomics that includes large-scale analysis of microbial communities [1]. Recent investigation has identified pivotal roles of the bacterial community in human diseases, including diabetes, obesity, Crohn's disease, and irritable bowel syndrome (IBS). However, even though various associations have been successfully discovered between microbial ecological patterns and host diseases, the characteristic of microbiome data such as zero-inflated problem and compositional bias complicates association analysis of microbiome. These issues, in addition, make the modeling of longitudinal analyzes more difficult, where complex correlations must be considered within repeated measurements. Heterogeneous result of metagenome analysis due to preprocessing, OTU filtering, database and clustering methods also can complicates association researches of microbiome.

Zero inflation

Metagenomic data have high variability [2]. The composition of the microbial community greatly differs from person to person even for microbial communities that function the same. Technical variability induced by insufficient sequencing depth, sequencing errors, or calculation errors in gene quantification is substantial [3]. Furthermore, microbial community consists of many species, and the small sample size. In consequence, OTUs shares across individuals, which makes microbial data very sparse and their statistical analysis complicated.

Compositional bias

In microbial data, the size of the sequencing depth varies from subject to subject, and the total absolute abundance collected for each subject substantially differs. Thus, relative abundance is generally utilized, but statistical analyses with relative abundance suffer from several problems.

First, the compositionality effects can introduce false positive associations and this bias stem primarily from their compositional characteristics [4] . I have fixed the sum of the abundances of each microorganism in each subject. If the absolute abundance of one taxon increases, the other taxa becomes decreased even though their abundance still remains same. Therefore longitudinally observed relative abundance of the same subject cannot be compared, and unless this so-called compositional bias is correctly adjusted, a false negative correlation can occur [5].

Second, biological insight is often related to absolute abundance. For example, absolute abundance of fecal microbiota in patients with Crohn's disease correlates bacterial load with disease phenotype. However the association disappeared when using relative abundance data [6, 7].

However, many association studies still do not adequately handled the compositional bias and are affected by the limitations of the relative abundance data and are potentially leading to false association results [7].

Heterogeneity of microbiome data

There are several reasons that make microbiome data heterogamous. The abundances of microbial taxa are often sparse with excessive zeros at species level. In detail, it is rare that any given taxa observed in all samples. Most of microbiota were observed in a small proportion of samples. This makes heterogeneity between samples and further, heterogeneity of dataset. Sample collection and storage can be aimportant source of heterogeneity [8]. Preprocessing steps such as OTU clustering, choice of OTU filtering threshold, rarefying can be another source of heterogeneity. 16S rRNA database contains the sequence of various taxa and those sequences were utilized for taxonomic assignment. Thus, the results and interpretation of the association analysis between taxa and host phenotypes can be affected by these pre-processing before the construction of microbial count table [9].

Importance of longitudinally observed microbiome data

The gut microbiota substantially becomes changed along the host age, and the effect of gut microbiota on the host phenotypes can be affected by age. Their risk on the host phenotypes can substantially differ by his/her ages and longitudinally measured microbiome data enables detecting their effect modification by age. Moreover, the estimation of within-subject covariate effect is robust against the between-subject confounders. However, in spite of such efficiency and validity, the nature of sparseness and compositional bias of metagenomics data complicates the statistical method development. Furthermore there are some correlations among repeatedly observed measurements of the same subject. [10] Therefore statistical method which is robust against those problems needs to be developed.

1.2. Literature Review

Statistical methods of cross-sectional 16S rRNA association analysis

OTUs have the high inter-subject variation, and their sparsity prevented application of linear/logistic regression. Many statistical methods have been suggested for statistical analysis with OTUs. For instance, OTUs belonging to the same genus or phylum can be pooled, and their relative proportions can be compared between cases and controls. However, as such pooling does not consider the heterogeneity among OTUs, several phylogenetic tree-based statistics have been suggested to adjust for these differences. Standard pipelines such as QIIME and mothur [11, 12] are used to cluster the 16S rRNA gene sequences of microorganisms into OTUs. The phylogenetic distance between pairs of OTUs was weighted with the UniFrac distance, and their weighted sums can be compared between cases and controls. For instance, PERMANOVA calculates weighted UniFrac distances between pairs of subjects and compares the average phylogenetic distances between cases and controls [13]. MiRKAT calculates a kernel matrix based on one of the various distance matrices, including weighted and unweighted UniFrac distances, and uses it to weight generalized linear model-based score tests. Optimal MiRKAT (oMiRKAT) combines the results of different distance matrix choices. The adaptive microbiome-based sum of the powered score (aMiSPU) considers phenotypes as responses for regression and uses the sums of weighted proportions for multiple taxa as covariates [14]. Both methods can adjust for environmental effects by adding them as covariates. In particular, the practical choice of the statistic is usually unclear, and robust approaches, such as minimum p-values, have been proposed. Notably, the optimal microbiome-based association test (OMiAT) considered the minimum p-value between oMiRKAT and aMiSPU and was shown to perform better under various scenarios [15]. The minimum p-value among multiple statistics generated from different types of distance matrices can also be useful owing to the uncertainty regarding the merit of the most efficient phylogenetic distances [16].

Statistical methods of longitudinal 16S rRNA association analysis

Longitudinal analysis for microbiome can be categorized into several parts. One is the standard statistical models including generalized linear mixed model (GLMM) or generalized estimating equations (GEE) and another is about zero-inflated mixture models such as ZINBMM and ZIBR [17]. Recently a kernel based longitudinal association test method GLMM-MiRKAT is also developed [18].

Methods to correct compositional bias

Compositional data is constrained to sum to a constant, naïve traditional statistical methods cannot be used for the compositional data [19]. Taking the logarithm of microbial abundance is a transformation of the constituent data that can preserve much of the usefulness of traditional statistical analysis in situations where library size needs to be considered, such as relative abundance [19]. Taking the logarithms has its problem in the choice of denominator. Additive log-ratio (alr) uses a reference abundance for its denominator and centered log-ratio (clr) uses geometric mean and both are well-known approach to consider compositional bias problem. Network analysis including SPARCC and SPIEC-EASI are also can be considered modeling the whole community in a statistical modeling.

Characteristics of 16S rRNA Databases and OTU picking methods

Many methods for defining OTUs have been proposed and they can be divided into closed-reference methods and de novo methods. In closed-reference method, each reference sequence in the database defines an associated closed-reference OTU. The input sequence is aligned to a reference sequence, and this reference sequence becomes a centroid of the OTU clusters. The taxonomy of OTU can be obtained from the information in this reference sequence [20]. If the same database is used, closed-reference OTU assignments from independently processed data sets can be validly compared, a property referred to as consistent labeling. Therefore, if the same reference database is used, the consistent labels can be pertained for independently processed datasets. Then the OTUs can be validly compared. However, the input sequences that failed to align to the reference database are lost. Therefore, the alpha-diversity will be greatly affected by the database.

De novo method uses the distance between sequences to cluster sequences into OTUs rather than the distance to a reference database. Therefore, there are no loss of input sequences as long as the sequences are clustered at a given level of similarity. This makes de novo method can estimate the maximal variance of microbiome organisms and correctly estimate alpha diversity compared to closed-reference method. However, the boundaries and members of clusters depend on a defined dataset. As a result, it is conceptually impossible to compare de novo OTUs defined on two different data sets [21, 22].

Amplicon sequence variants (ASVs) is recently suggested as an alternative approach for taxonomy assignment. ASVs are inferred by the Poisson-based process which assume biological sequences are more likely to be repeatedly observed than sequences containing observational errors [23, 24]. The ASV can be inferred for each individual, and the ASV for each sample was consistent. This allows ASVs to be used with consistent labeling [23]. Furthermore, ASV is not dependent on databases and there is no loss of sequences depending on the choice of databases.

However, in spite of flexibility of ASV, it has some limitation. ASV calculates the likelihood based on biological distance. However biological distance between the sequences is confounded by other environmental factors or batch effects, and their distance cannot consider such confounding. Also, ASVs cannot be merged if the underlying sequence was derived from the different region of 16S rRNA gene [23].

Taxonomy assignment depends on the database including ExTaxon, Silva, Greengene and significant differences of taxonomy assignment among databases have been reported [25, 26].

To examine the accuracy of the three public databases, the known taxonomies for 60 strains from the mock community were compared to the outcome of taxonomic assignment for each databases. To simplify the comparison, no sequencing error or missing strain was assumed. The accuracy of each database was evaluated with the number of true-positive, false-positive and false-negative taxa. These measures can be affected by the number of reference sequences. ExTaxon contained the smallest number of sequences among the compared databases and this can deflate the number true-positives and false-positives and inflate false-negatives. ExTaxon database found to be the most accurate database with the most highest number of true positive and smaller number of false-positives and false-negatives than those of other databases [26].

However the result is depend on the simulated mock community dataset which consist of only 60 strains that were uniformly distributed. Most microbiome communities are composed of more than thousands of species, and their constitution is not uniform. The most accurate database can vary according to the samples analyzed [26, 27].

1.3. Purpose of Research

The main purpose of my research is as follows.

1) Development of a statistical method for association analyses with cross-sectionally observed microbiome data and package implementation.

2) Development of a statistical method for association analysis with longitudinally observed microbiome data.

3) Identifying microorganisms associated with the type-2 diabetes with longitudinally measured microbiome data.

Chapter 2. Phylogenetic Tree-based Microbiome Association Test and Package Development for Microbiome Analysis

2.1. Introduction

The explosive accumulation of research data using advanced high-throughput technologies such as microarrays and next-generation sequencing (NGS) has greatly improved our understanding of the microbial world and has greatly improved our understanding of biological research. Provided the underlying idea [28].

However, even though various associations have been successfully discovered between microbial ecological patterns and host diseases, high inter-subject variation complicates association analyses with microbiomes. For instance, most operational taxonomic units (OTUs) are observed only in a few subjects, and the absolute abundances of many OTUs are often 0, making the assumption of asymptotic normality of the observed abundances unlikely. Thus, associations of OTUs with host diseases are often tested with non-parametric approaches, such as the Mann−Whitney *U*-test and Wilcoxon rank-sum test [29]. However, non-parametric statistics use the ranks of observed relative abundances for statistical inferences instead of the observed relative abundances themselves, and the degree of difference between cases and controls is neglected. Such information loss can increase the false-negative rates of non-parametric statistics. Alternatively, the observed relative abundances can be subjected to an arcsine-root transformation, but this has been shown not to correctly control the type-1 error rates for low-abundance species [30].

Many analysis strategies have been suggested to adjust for the sparsity of

OTUs induced by the high inter-subject variation observed. For instance, OTUs belonging to the same genus or phylum can be pooled, and their relative proportions can be compared between cases and controls. However, as such pooling does not consider the heterogeneity among OTUs, several phylogenetic tree-based statistics have been suggested to adjust for these differences. Standard pipelines such as QIIME and mothur [11, 12] are used to cluster the 16S rRNA gene sequences of microorganisms into OTUs, and the phylogenetic distances between pairs of OTUs can be calculated and weighted with the UniFrac distance, allowing their weighted sums to be compared between cases and controls. For instance, PERMANOVA calculates weighted UniFrac distances between pairs of subjects and compares the average phylogenetic distances between cases and controls [13]. MiRKAT calculates a kernel matrix based on one of the various distance matrices, including weighted and unweighted UniFrac distances, and uses it to weight generalized linear model-based score tests, and Optimal MiRKAT (oMiRKAT) combines the results of different distance matrix choices. The adaptive microbiome-based sum of the powered score (aMiSPU) considers phenotypes as responses for regression and uses the sums of weighted proportions for multiple taxa as covariates [14]. Both methods can adjust for environmental effects by adding them as covariates. In particular, the practical choice of the statistic is usually unclear, and robust approaches, such as minimum p-values, have been proposed. Notably, the optimal microbiome-based association test (OMiAT) considered the minimum p-value between oMiRKAT and aMiSPU and was shown to perform better under various scenarios [15]. The minimum p-value among multiple statistics generated from different types of distance matrices can also be useful owing to the uncertainty regarding the merit of the most efficient phylogenetic distances [16].

Multiple investigations have found that different species or strains within the same genus can differentially affect diseases, and the importance of intra-genus mutations has been repeatedly highlighted [31]. However, because relative abundances at the species level are often very sparse, association analyses at this level have been limited. In this article, I propose the phylogenetic tree-based microbiome association test (TMAT) to identify OTUs associated with host diseases. TMAT considers the log-transformed read count per million (CPM) as the response, and the log CPM is assumed to follow the normal distribution. TMAT tests whether each internal node of a phylogenetic tree is associated with a host disease, and the resulting statistics are combined into a single statistic. By the nature of the proposed statistics, node statistics are independent, and internal nodes associated with host diseases can be detected by aggregating those statistics. Here, I define the proposed TMAT statistics and describe both real data and *in silico* experiments. The superiority of the proposed methods over existing methods is demonstrated through extensive simulations based on metagenomics datasets for colorectal carcinoma (CRC) and myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). TMAT was applied to these datasets, and significantly associated OTUs were identified. Lastly, the distinctive features of TMAT and the main reason for its superiority over existing methods are discussed.

In addition, All-inclusive Microbiome Association Analysis (AMAA), a package that envelope a pipeline building microbial count tables based on different databases and clustering method and the methods for metagenome-wide association analysis will be introduced. It provides the convenient use of various methods for microbiome association analysis under a unified preprocessing and comparison the results based on different databases or clustering method.

2.2. Materials and Methods

Ethics statement

The protocol used in this study was approved by the Institutional Review Board (IRB No. E2108/001-001) in Seoul National University.

Phylogenetic tree

Let us assume that N subjects are sequenced, and M OTUs are observed. I assume that a rooted binary phylogenetic tree is provided for these OTUs, and the first M_1 OTUs belong to a genus of interest for the analysis of its association with host diseases, while the other $M - M_1$ OTUs belong to different genera. For the genus with the first M_1 OTUs, there are $M_1 - 1$ internal nodes and M_1 leaf nodes. Internal nodes are denoted by k, where k =1,..., $M_1 - 1$. Leaf nodes are denoted by m, where $m = 1, \dots , M$. For each leaf node there is a corresponding single OTU; if $m = 1, \dots$, or M_1, m is the leaf node of the genus of interest, and otherwise m belongs to a different genus. I assume that mutations that affect host diseases occur during transmission from the internal node k to its left (or right) child node. These mutations may be transmitted from the left (right) child node to all of its leaf nodes, and the relative abundances of OTUs corresponding to those leaf nodes should significantly differ between cases and controls. Under this assumption, the relative proportion of leaf nodes of the left child node increases for cases if the mutation occurs during transmission to its left child node, and it decreases if it does so during transmission to the right child node. If the association of an internal node *k* with a host disease of interest is tested, let the internal node *k* and its leaf nodes represent a test node and test leaf nodes, respectively. The left and right test leaf nodes further represent the leaf nodes of the left and right child nodes of a test node, respectively.

For internal node *k* in the genus with M_1 OTUs, let L_k and R_k be the sets of its left and right leaf nodes, respectively. Figure 2.1A and B illustrates these definitions.

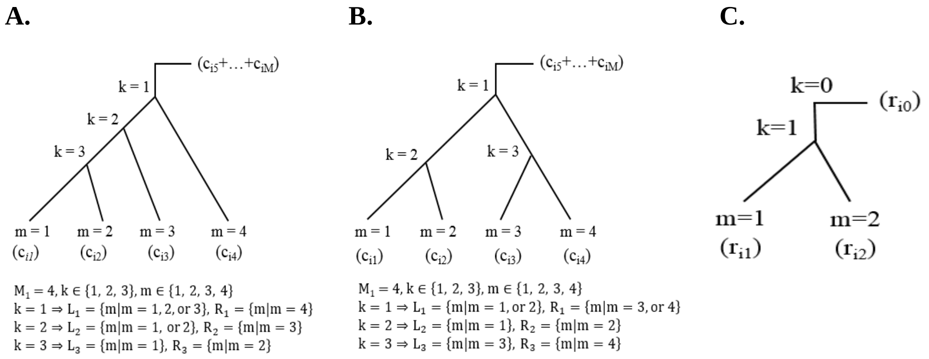

A.
$(c_{i5}+\ldots+c_{iM})$
k = 1
k = 2
k = 3
m = 1 (c_{i1})
m = 2 (c_{i2})
m = 3 (c_{i3})
m = 4 (c_{i4})
$M_1 = 4, k \in \{1, 2, 3\}, m \in \{1, 2, 3, 4\}$
$k = 1 \Rightarrow L_1 = \{m|m = 1, 2, \text{or } 3\}, R_1 = \{m|m = 4\}$
$k = 2 \Rightarrow L_2 = \{m|m = 1, \text{ or } 2\}, R_2 = \{m|m = 3\}$
$k = 3 \Rightarrow L_3 = \{m|m = 1\}, R_3 = \{m|m = 2\}$
B.
$(c_{i5}+\ldots+c_{iM})$
k = 1
k = 2
k = 3
m = 1 (c_{i1})
m = 2 (c_{i2})
m = 3 (c_{i3})
m = 4 (c_{i4})
$M_1 = 4, k \in \{1, 2, 3\}, m \in \{1, 2, 3, 4\}$
$k = 1 \Rightarrow L_1 = \{m|m = 1, \text{or } 2\}, R_1 = \{m|m = 3, \text{or } 4\}$
$k = 2 \Rightarrow L_2 = \{m|m = 1\}, R_2 = \{m|m = 2\}$
$k = 3 \Rightarrow L_3 = \{m|m = 3\}, R_3 = \{m|m = 4\}$
C.
k=0
(r_{i0})
k=1
m=1 (r_{i1})
m=2 (r_{i2})

Figure 2.1. Examples of rooted binary phylogenetic trees.

Quasi-Score test statistic for TMAT

I denote the absolute abundance of OTU m in subject i by c_{im}, the log-transformed CPM, and r_{im}, which is used for the edgeR package (version 3.16.5), is defined by

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

Here, $c_{i\cdot}$ is a pseudocount that is proportional to the total read count for subject i, and it is calculated using the same method as is used in the edgeR package. $\frac{c_{im}+\frac{c_{i\cdot}}{2}}{\sum_{m=1}^{M} c_{im}+c_{i\cdot}} \times 10^6$ can be less than 1, and in such case its logarithm becomes negative. Thus, I add 1 to make r_{im} positive. Log-CPM transformation is widely used in RNA sequencing data analyses [32]. Then, x_i^k $(i \in L_k)$, where $k = 1, \ldots, M_1 - 1$, is defined by

$$x_i^k = log\left(\frac{C_i^k}{D_i^k}\right), C_i^k = \sum_{m=1}^{M_1} r_{im} \cdot I(m \in L_k), D_i^k = \sum_{m=1}^{M_1} r_{im} \cdot I(m \in R_k).$$

As all OTUs in the genus can be associated with the host disease, x_i^0 for such case is defined by

$$x_i^0 = log\left(\frac{C_i^0}{D_i^0}\right), C_i^0 = \sum_{m=1}^{M_1} r_{im}\,, D_i^0 = \sum_{m=M_1+1}^{M} r_{im}\,.$$

The phenotype of subject i is denoted by y_i and is coded as 1 and 0 for cases and controls, respectively. Their vectors and matrices for testing the association of the genus of interest are defined by

$$\boldsymbol{x}^k = \begin{pmatrix} x_1^k \\ \vdots \\ x_N^k \end{pmatrix}, \boldsymbol{X} = (\boldsymbol{x}^0 \quad \cdots \quad \boldsymbol{x}^{M_1-1}), \ \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

Here, I assume that $\boldsymbol{x}^0$, … , and $\boldsymbol{x}^{M_1-1}$ are ordered according to the depth of the internal nodes. $\boldsymbol{x}^0$ is used for testing the association of all OTUs belonging to the genus of interest by pooling them, and $\boldsymbol{x}^1$ is for testing the root node of the phylogenetic tree. If I denote an N×N identity matrix as $\boldsymbol{I}_N$ and let $\boldsymbol{Z}$ be a design matrix for p covariates including the intercept,

I assume that

$$E(\boldsymbol{x}^k|\boldsymbol{Z},\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{\alpha}_k + \boldsymbol{y}\boldsymbol{\beta}_k, Var(\boldsymbol{x}^k|\boldsymbol{Z},\boldsymbol{y}) = \sigma_{kk}\boldsymbol{I}_N, k = 0,\dots,M_1 - 1.$$

Quasi-score functions for $\boldsymbol{\alpha_k}$ and $\boldsymbol{\beta}_k$ can be obtained by

$$U_{\boldsymbol{\alpha}}(\boldsymbol{\alpha_k},\boldsymbol{\beta}_k) = \frac{1}{\sigma_{kk}}\boldsymbol{Z}^t(\boldsymbol{x}^k - \boldsymbol{Z}\boldsymbol{\alpha}_k - \boldsymbol{y}\boldsymbol{\beta}_k),$$

$$U_{\boldsymbol{\beta}}(\boldsymbol{\alpha_k},\boldsymbol{\beta}_k) = \frac{1}{\sigma_{kk}}\boldsymbol{y}^t(\boldsymbol{x}^k - \boldsymbol{Z}\boldsymbol{\alpha}_k - \boldsymbol{y}\boldsymbol{\beta}_k).$$

Under the null hypothesis H_0: $\boldsymbol{\beta}_k = 0$, $\widehat{\boldsymbol{\alpha}}_k$ is estimated by

$$\widehat{\boldsymbol{\alpha}}_k = (\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{x}^k.$$

If I let $\mathbf{A} = \mathbf{I}_N - \mathbf{Z}(\mathbf{Z}^t\mathbf{Z})^{-1}\boldsymbol{Z}^t$, the quasi-profile score for $\boldsymbol{\beta}_k$ becomes

$$\boldsymbol{s}_k = \boldsymbol{y}^t(\boldsymbol{I}_N - \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{Z})^{-1}\boldsymbol{Z}^t)\boldsymbol{x}^k = \boldsymbol{y}^t\boldsymbol{A}\boldsymbol{x}^k,$$

and it can be used for testing the null hypothesis. The covariance matrix of $\boldsymbol{s}_k$

can be obtained by

$$var(\boldsymbol{s}_k) = var\left(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{x}^k\right) = \sigma_{kk}\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{y} = \sigma_{kk}\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y}.$$

σ_{kk} is estimated by

$$\hat{\sigma}_{kk} = \frac{1}{N-p}\boldsymbol{x}^{k^t}\boldsymbol{A}\boldsymbol{x}^k.$$

Therefore, the score test statistic of $\boldsymbol{\beta}_k$ for the test node *k* can be defined as

$$T_k = \frac{1}{\hat{\sigma}_{kk}}\boldsymbol{s}_k^t(\boldsymbol{y}^t\boldsymbol{A}\boldsymbol{y})^{-1}\boldsymbol{s}_k \sim \ \chi^2(df = 1)\ \ under\ \ H_0.$$

If the sample size is small, normality of T_k under H_0 may not be achieved, and the assumption of the quasi-score test can be violated. If I apply the inverse normal transformation to $x_1^k, \dots, x_N^k$, then the same statistics can be obtained. This is denoted by T_k^{INT}. Rank-based inverse normal transformation with adjust parameter 0.5 is used for the transformation and data with tie values were mapped to a same value in the transformed data [33].

Statistics for $H_0{:}\beta_k = 0$ can be combined to test $H_0{:}\beta_0 = \beta_1 = \cdots = \beta_{M_1-1} = 0$ using the minimum p-value. If p-values for Tk are denoted by pTk, the proposed statistics, $TMAT_M$ and $TMAT_{IM}$, are defined by

$$TMAT_M = min\{pT_0, \quad \cdots \quad, pT_{M_1-1}\},$$

$$TMAT_{IM} = min\left\{pT_0^{INT}, \quad \cdots \quad, pT_{M_1-1}^{INT}\right\}.$$

In particular, T_k and T_{k+1} are sufficient and ancillary statistics for β_k, respectively, and $T_0, \dots, T_{M_1-1}$ are shown to be independent (see Supplementary Text 2). Therefore,

$$TMAT_M, TMAT_{IM}\ \sim\ beta(1, M_1)\ \ under\ \ H_0.$$

Independence of score test statistics

I assume that there is a single internal node. In such a case, there are two leaf nodes, and phylogenetic tree becomes the one in Figure 2.1C. I let the observed absolute read counts of an j^{th} observation of subject i be $r_{\mathrm{i}1}$ and $r_{\mathrm{i}2}$, respectively. Then, test statistics T_0 and T_1 are functions of $r_{i1}/(r_{i1}+r_{i2})$, and $r_{i1}+r_{i2}$. Let the observed absolute read counts of subject i be $r_{\mathrm{i}1}$ and $r_{\mathrm{i}2}$, respectively. Then, test statistics T_0 and T_1 are functions of $r_{i1}/(r_{i1}+r_{i2})$, and $r_{i1}+r_{i2}$. Let $f_{\boldsymbol{x},\boldsymbol{y}}(x,y)$ be the joint probability density function (PDF) of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $f_{\boldsymbol{x}}(x)$ and $f_{\boldsymbol{y}}(y)$ be their two marginal PDFs. I assume that r_{i1} and r_{i2} independently follow a Poisson distribution with parameters μ_1 and μ_2, respectively. If I set $\rho_1=\mu_1/(\mu_1+\mu_2)$, then $r_{i1}|(r_{i1}+r_{i2})\sim\mathrm{B}(r_{i1}+r_{i2},\rho_1)$. Therefore, then the joint distribution of r_{i1} and r_{i2} is equivalent to

$$\log f_{\boldsymbol{r}_{i1},\boldsymbol{r}_{i2}}(r_{i1},r_{i2};\mu_1,\mu_2)=\log f_{\boldsymbol{r}_{i1}|\boldsymbol{r}_{i1}+\boldsymbol{r}_{i2}}(r_{i1}|r_{i1}+r_{i2};\rho_1)$$
$$+\log f_{\boldsymbol{r}_{i1}+\boldsymbol{r}_{i2}}(r_{i1}+r_{i2};\mu_1+\mu_2).$$

$r_{i1}/(r_{i1}+r_{i2})$ and $r_{i1}+r_{i2}$ are maximum likelihood estimators of ρ_1 and $\mu_1+\mu_2$, respectively; and $\partial l^2/\partial\rho_1\partial(\mu_1+\mu_2)=0$. $\boldsymbol{x}^0$ and $\boldsymbol{x}^1$ are functions of $r_{i1}/(r_{i1}+r_{i2})$ and $r_{i1}+r_{i2}$, respectively. Thus, I can conclude that $\boldsymbol{x}^0$ and $\boldsymbol{x}^1$ are asymptotically independent, which indicates that T_0 and T_1 that are functions of $\boldsymbol{x}^0$ and $\boldsymbol{x}^1$ respectively are also asymptotically independent.

I assume that there are M_1 internal nodes, and statistics for those, $T_0,\ldots,T_{M_1}$, are asymptotically independent.

I consider the phylogenetic tree with M_1+1 internal nodes. I assume that internal nodes are sorted in the ascending order of their depth. Then internal node M_1+1 has the largest depth. I decompose those into the first M_1 internal nodes and the last internal node M_1+1. By the assumption (2), I can assume that $T_0, \dots, T_{M_1}$ are asymptotically independent. Let the leaf nodes of internal node M_1+1 be r_{iM_1+1} and r_{iM_1+2}, and the other leaf nodes be r_{i1} , … , r_{iM_1}. I assume that r_{iM_1+1} and r_{iM_1+2} independently follow a Poisson distribution with parameters μ_{M_1+1} and μ_{M_1+2}, respectively. Then if I let $\rho_{M_1+1} = \mu_{M_1+1}/(\mu_{M_1+1} + \mu_{M_1+2})$, I can show that

$$\begin{aligned} &\log f_{\boldsymbol{r}_{iM_1+1}, \boldsymbol{r}_{iM_1+2}}\left(r_{iM_1+1}, r_{iM_1+2}; \mu_{M_1+1}, \mu_{M_1+2}\right) \\ &= \log f_{r_{iM_1+1} | r_{iM_1+1} + r_{iM_1+2}}\left(r_{iM_1+1} | r_{iM_1+1} + r_{iM_1+2}; \rho_{M_1+1}\right) \\ &\quad + \log f_{\boldsymbol{r}_{i1} + \boldsymbol{r}_{i2}}\left(r_{iM_1+1} + r_{iM_1+2}; \mu_{M_1+1} + \mu_{M_1+2}\right). \end{aligned}$$

Similar to what is shown above (1), I can conclude that $r_{iM_1+1}/(r_{iM_1+1} + r_{iM_1+2})$ and $r_{iM_1+1} + r_{iM_1+2}$ are approximately independent, which indicates the approximate independence between r_{iM_1+1}/r_{iM_1+1} and $r_{iM_1+1} + r_{iM_1+2}$. Therefore $(T_0, \dots, T_{M_1})$ and T_{M_1+1} are asymptotically independent.

Constructing phylogenetic trees for TMAT and quality control

Statistical analysis with TMAT requires the construction of a phylogenetic tree, and databases such as Silva (release 128) [34] and EzTaxon [35] were used for all taxonomic assignments in the CRC and ME/CFS datasets. The Silva database was used to generate reference trees, which were then used to calculate phylogenetic distances. The EzTaxon database does not generate phylogenetic trees, and these were therefore obtained through the SINA method [36] using the reference sequences available from the EzTaxon database. I used a de-novo picking method for taxonomic assignment. The de-novo picking method can assign different OTUs to the same species. OTUs assigned to the same species were considered the same OTU, and their absolute abundances were pooled. Characteristics of the microbiome community may be affected by filtering conditions. For each OTU, I calculated its relative proportion among all OTUs and determined the mean value across all subjects. If the resulting value was smaller than 0.001, the OTU was excluded from the analysis [37].

Fecal microbiota data for early-stage detection of colorectal cancer

A microbiome profiling study conducted by Zeller et al. [38] examined the potential utility of the fecal microbiota for early-stage detection of colorectal carcinoma (CRC). The 16S rRNA amplicon sequencing data from the study are available from the European Nucleotide Archive (ENA) database under project accession number PRJEB6070. The paired-end sequence pairs for 225 individuals targeted the V4 region of the bacterial 16S rRNA gene. The primers F357 (5'-CCTACGGGAGGCAGCAG-3') and R519 (5'-GTNTTACNGCGGCKGCTG-3'), which are widely used for amplifying the V4 region, were detected and removed using CUTADAPT software with a minimum overlap of 11, maximum error rate of 10%, and a minimum length of 10. Sequences were merged using CASPER software with a mismatch ratio of 0.27, resulting in sequences 230−270 base pairs in length [39]. After merged sequences were dereplicated, chimeric sequences were detected and removed using VSEARCH software with the Silva Gold reference database for chimeras. A *de novo* picking method was used to obtain the resulting OTU table with a 97% sequence identity threshold. Information about the disease status was missing for 109 subjects, who were excluded from the study, resulting in 41 CRC patients and 75 controls being considered for following simulation studies and real data analyses.

Fecal microbiota data for myalgic encephalomyelitis

Giloteaux et al. [40] used 16S rRNA gene sequencing to examine the microbiome profiles of subjects with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). Data were downloaded under the ENA project accession number PRJEB13092. The V4 region of the bacterial 16S rRNA gene was targeted. The methods described above for the CRC data were applied for primer detection, merging, length filtering, dereplication, and OTU picking. The final dataset consisted of 49 ME/CFS patients and 39 controls, which were used for simulations and real data analyses.

Simulation studies

I conducted extensive simulations to evaluate the performance of TMAT with two datasets; one with 41 CRC patients and 75 controls, and the other with 49 ME/CFS patients and 39 controls. Detailed description for both datasets is provided in Supplementary Text 3. For the simulation studies, the disease status of the subjects was permuted, and specific numbers of cases and controls were randomly selected. Then, I randomly selected a single test node from the internal nodes, and from their test leaf nodes, either a single OTU, 50% of OTUs, or 90% of OTUs were randomly selected as causal OTUs. These were denoted by p = 1 OTU, 50%, and 90%, respectively. It should be noted that p = 1 indicates that there is a single OTU associated with the host disease, and thus, the phylogenetic tree structure does not provide any useful information for TMAT. If I let the sample variances of c_{im} for causal OTUs be $\hat{\sigma}_{mm}$, $\delta = \beta\hat{\sigma}_{mm}$ was added to the observed absolute abundances of the selected causal OTUs for only affected subjects, and the absolute abundances of the other OTUs were used without any modification. β was set to 0, 0.01, 0.05, 0.1, or 0.2. $\beta = 0$ was considered for estimation of empirical type-1 error rates, and the others were used for estimating statistical power. Type-1 error rates were estimated at the 0.05, 0.01, 0.005, and 0.001 significance levels with 20,000 replicates. Empirical power was estimated at the 0.05 significance level with 2,000 replicates.

For the sake of comparison with TMAT, oMiRKAT (version 0.02), MiSPU (version 1.0), OMiAT (version 5.1), ANCOM (version 1.1-3), edgeR (version 3.16.5), and the Wilcoxon test were considered. Association analyses were conducted at the genus level. Wilcoxon, ANCOM, and edgeR were applied by pooling all OTUs within each genus. Each genus consisted of multiple OTUs, and oMiRKAT, MiSPU, and OMiAT were applied to OTUs belonging to each genus.

MiSPU, OMiAT, and oMiRKAT use permutation-based p-values, and they were calculated with 5,000 and 20,000 permutated replicates for estimation of power and type-1 error rates, respectively. oMiRKAT offers several distance metrics, including Unifrac distance as a default choice, while MiSPU also uses Unifrac distance as the default option. I considered the default choices; however, Unifrac distance cannot be calculated if read counts are not observed. Thus, subjects with no read counts were excluded from oMiRKAT and MiSPU. Furthermore, none of these can analyze a genus with a single OTU; hence, such instances were not considered for statistical power estimations of such genera. The R package ANCOM provides a “Multcorr” option for ANCOM function, and it was set to 2, which indicates “less strict correction.” The negative binomial generalized log-linear model was used for edgeR. All other options were set to default values.

Package and software used for AMAA

AMAA provides three main analyses: metagenome sequence data processing to build microbiome count table, microbiome compositional analysis and metagenome-wide association study (Table 2.1).

For the first step in data processing of metagenome sequence, adaptor sequences are detected and removed using the CUTADAPT software (https://cutadapt.readthedocs.io) with a default option as a minimum overlap of 11, maximum error rate of 10%, and a minimum length of 10 [41]. Sequences will be merged using CASPER (http://best.snu.ac.kr/casper) with a mismatch ratio of 0.27 and filtered by the Phred (Q) score, resulting in sequences with a certain range of length according to the target region of 16S rRNA gene [42]. For example, 350–550 bp is the suitable range for V3-V4 region. After the merged sequences were dereplicated, chimeric sequences will be detected and removed using VSEARCH (https://github.com/torognes/vsearch) and the Silva Gold reference database for chimeras. With a single command line, either of open-reference using UCLUST (http://www.drive5.com/usearch), De-novo clustering, closed-reference, operational taxonomical unit (OTU) picking methods using VSEARCH [43] and detection of ASVs using DADA2 [24] or all of them depending on the chosen option, will be conducted. For closed-reference and open-reference method, choice of database can be Silva [44], Greengenes [45] and EzTaxon [46]. Lastly, taxonomies are assigned based on taxonomies of chosen database. The characteristic of each databases and clustering methods are described in Table 2.2 and Table 2.3 respectively.

In microbiome compositional analysis, rarefying step using GUniFrac package precede calculation of alpha or beta diversity. Rarefying step is needed for sample quality control and normalization of total read counts for each sample. In microbiome compositional analysis, normalized dataset can reduce the bias in alpha and beta-diversity measurements. Alpha-diversities such as ACE, Chao 1, Shannon and Simpson index will be calculated and described with boxplots or scatter plots depending on the proper type of traits. For the choice of beta-diversity, the Bray-curtis, weighted UniFrac and unweighted UniFrac distance are provided. With the calculated beta-diversity, overall microbial variance can be described with Non-metric Multi-dimensional Scaling (NMDS) plot and PERMANOVA determines the associated-traits. For repeated measured data, pldist is used for consideration of correlations in within-subject samples.

For metagenome-wide association test in AMAA, preprocessing step is conducted such as filtering extremely sparse OTUs, dividing OTUs into certain taxonomy group, rarefaction procedure if needed for testing, removing of the samples with zero abundance. After the preprocessing step, community-level association tests such as MiRKAT, MiSPU and OMiAT, non-parametric method, Wilcoxon test, ANCOM, EdgeR and TMAT will be provided for cross-sectional analysis. LMM with arcsine square root transformation, ZIBR, cSKAT and FZINBMM is available for repeated measured data.

Similarity measure for a pair of databases

The identifiers of specific microbial organisms vary from database to database. There are multiple database and OTU clustering results depends on the database. Thus, different OTUs can be generated according to the choice of database, and I consider compositional dissimilarity between a pair of samples can be evaluated with beta diversity. It was assumed that Unifrac distance that considers both of the existence or abundance of OTUs and phylogenetic tree information is available. Then if I let D_d and $D_{d'}$ be the Unifrac distances based on database d and d' respectively, similarity measure of D_d and $D_{d'}$ is constructed with a modified version of correlation matrix distance [47] as follows:

$$S_{dd'} = \frac{tr(S_d S_{d'})}{||S_d||_f \times ||S_{d'}||_f}$$

where $S_d = 1 - D_d$, $S_{d'} = 1 - D_d$.

Table 2.1. Analyses available in AMAA

Main analyses	Sub-catecory	Description	Reference and software
Sequence data processing	Removal of adaptor sequences	Primers	Martin et al. (2011) - Cutadapt
	Merging of sequences	Paired ends	Kwon et al. (2014) - Casper
	Filtering unqualified sequences	Phred (Q) score/Sequence length filtering	Bokulich et al. (2013)
	Remove chimeric sequences	Using Silva Gold reference database	Rognes et al. (2016) -VSEARCH
	Building microbial count table	Open-reference OTU picking	Edgar et al. (2010) - UCLUST
		Closed-reference OTU picking	Rognes et al. (2016) -VSEARCH
		De-novo OTU picking	Rognes et al. (2016) -VSEARCH
		Amplicon sequence variants (ASVs)	Callahan et al. (2016) - DADA2
	Taxonomy assignment	Choice of database	Edgar et al. (2010) - UCLUST
Microbial composition analysis	Rarefying	Normalized table	Chen et al. (2018) - GUniFrac
	Alpha-diversity	ACE, Chao 1, Shannon, Simpson index	Oksanen et al. (2007) - vegan
	Beta-diversity	Bray-curtis, Unifrac	Oksanen et al. (2007) - vegan Chen et al. (2018) - GUniFrac
	NMDS plot, PCA, Kernel PCA	Visualization of beta-diversity	Oksanen et al. (2007) - vegan / skikit-lean
	Pldist	Beta-diviersity for repeated data	Plantinga et al. (2019) - pldist
	PERMANOVA	Find traits that explains microbial variance	Anderson et al. (2013) - PERMANOVA
Metagenome-wide association analysis	TMAT	Phylogenetic tree-based	Kim et al. (2020)
	ANCOM	Compositional bias correction	Mandal et al. (2015)
	MiRKAT	Kernel based regression	Wilson et al. (2020)
	MiRKAT-s	Extension of MiRKAT for survival data.	Plantinga et al. (2017)
	aMiSPU	Generalized taxon proportion	Wu et al. (2016)
	OMiAT	Combines oMiRKAT & aMiSPU	Koh et al. (2017)
	OMiSA	Combines MiSALN & MiRKAT-s.	Robinson et al. (2010)
	EdgeR	Method for RNA expression data	Bauer et al. (1972)
	Wilcoxon rank sum test	Non-parametric approach	
	GLMM-MiRKAT	Extension of MiRKAT for longitudinal data.	Koh et al (2019)
	ZIBR	Zero-inflated beta random effect (Only balanced data)	Chen et al. (2016)
	cSKAT	Small-sample kernel association test for correlated data	Zhan et al. (2018)
	FZINBMM	Zero-inflated negative binomial mixed model	Zhang et al. (2020)
	LMM (arcsine square root, log)	Linear mixed model with arcsine root or log transformation.	

Table 2.2. Databases available in AMAA

Database	Country	Sequences	Lowest rank	Description
Greengenes	U.S.A (Berkeley)	99,000	Species (Limited)	Approximately 5% have species-level names
Silva	German (Max Plank)	190,000	Species	Most of the strains are unclassified.
EzTaxon	Korea (Chunlab)	63,000	Species	All the strains have species-level names.

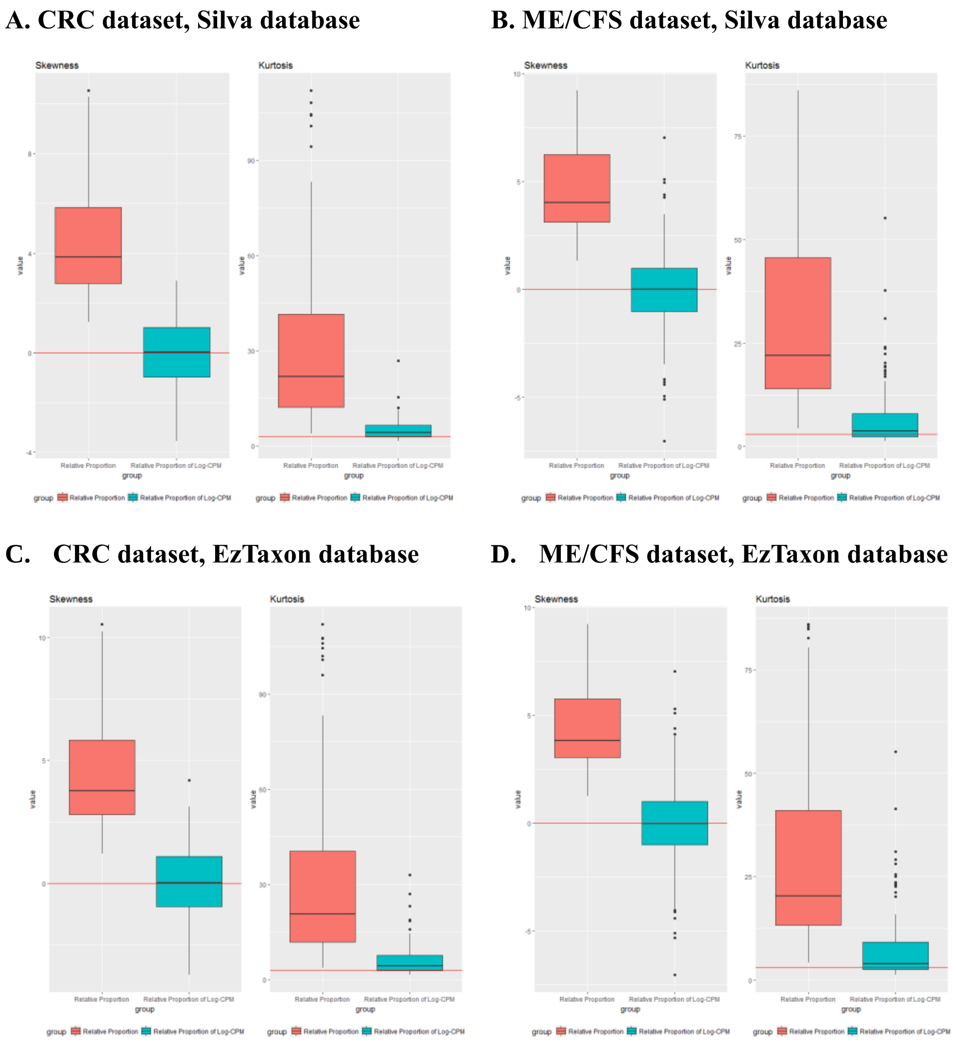
Table 2.3. Characteristics of clustering methods available in AMAA.

	Validity of diversity measurement	Robust to change of references	Application across environments	Guaranteed observation & replication	Meta-analysis	Computational costs
De novo	Y	Y	Y	N	N	High
Closed reference	N	N	N	Y	Y	Low
Open reference	Y	N	N	N	N	High
ASVs	Y	Y	Y	Y	Y	High

2.3. Results

Distribution of the relative proportion of log CPM

Proposed statistic assume that x_i^k are normally distributed and that the statistical scores for internal nodes are independent. To verify the normality of x_i^k, their skewness and kurtosis were calculated using the OTUs in the CRC and ME/CFS datasets. Figure 2.2A and 2.2B show boxplots of skewness and kurtosis for x_i^k for the OTUs in the CRC and ME/CFS datasets after using the Silva database for OTU clustering. The results showed that the median values of skewness and kurtosis of the relative abundances were substantially greater than 0. However, the medians of skewness and kurtosis of x_i^k were much closer to 0, and thus I can conclude that the distribution of x_i^k is much closer to a normal distribution. Figure 2.3 shows scatter plots of each pair of score statistics for internal nodes. The scatter plots in Figure 2.3A and 2.3B do not show any significant patterns, and the correlations were 0.0250 (p-value = 0.4966) and 0.0189 (p-value = 0.553). The results based on EzTaxon are shown in Figure 2.3C and 3.3D. Therefore, the statistical scores for internal nodes are approximately independent.



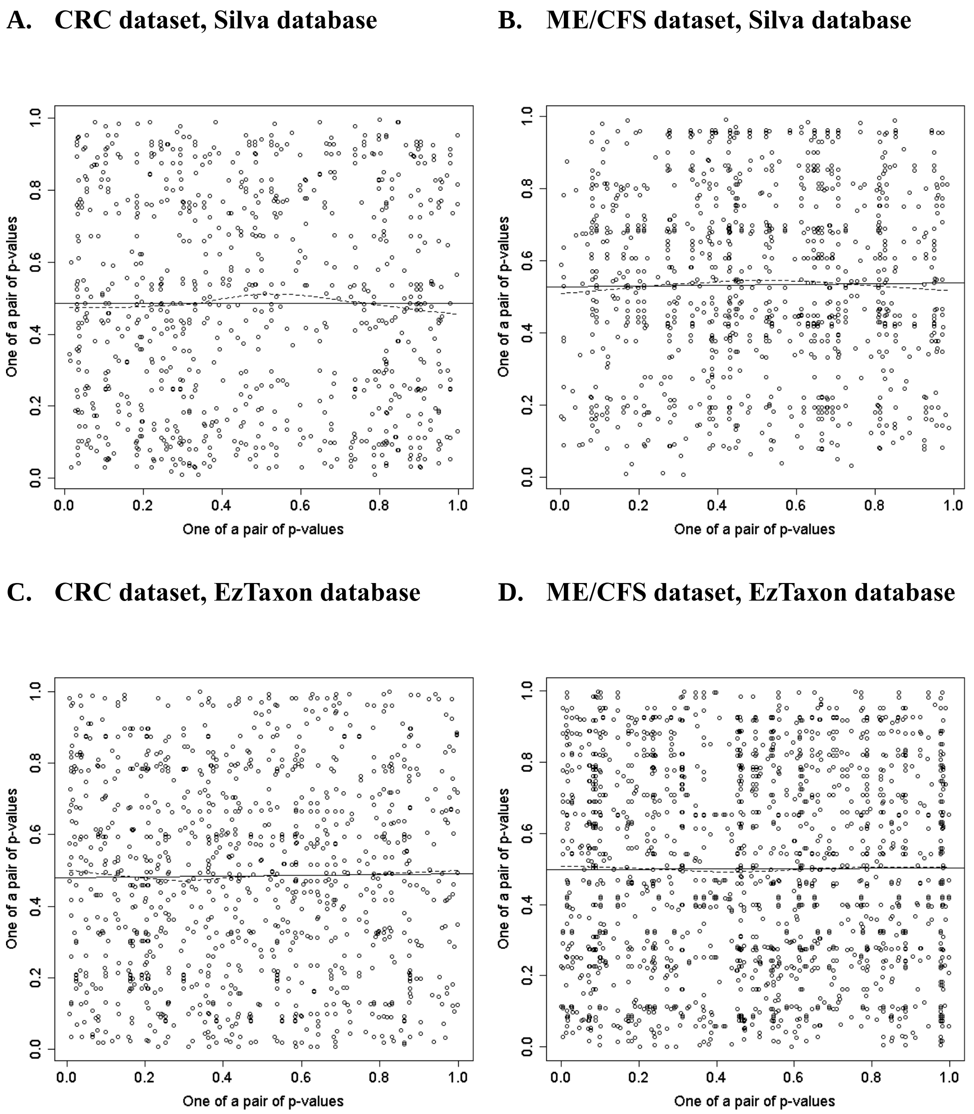
A. CRC dataset, Silva database

B. ME/CFS dataset, Silva database

C. CRC dataset, EzTaxon database

D. ME/CFS dataset, EzTaxon database

Figure 2.2. Skewness and kurtosis of relative abundances and x_{Nm}^{k}. Skewness and kurtosis were calculated for relative abundances and x_{Nm}^{k} in the CRC and ME/CFS datasets. Skewness and kurtosis for normal distributions are 0, represented by vertical lines. Results are based on Silva and EzTaxon databases.



A. CRC dataset, Silva database

B. ME/CFS dataset, Silva database

C. CRC dataset, EzTaxon database

D. ME/CFS dataset, EzTaxon database

Figure 2.3. Scatter plots for pairs of p-values of internal nodes. Each pair of p-values for each internal node was used to create scatter plots. Solid and dashed lines represent a simple linear regression line and LOWESS smooth line [48], respectively. LOWESS smooth line was fitted with the function 'lowess' in R with default options. Results are shown for CRC and ME/CFS with OTUs clustered by Silva and EzTaxon databases.

Results from simulated data

I first calculated the empirical type-1 error estimates using simulated data based on the CRC and ME/CFS datasets. Table 2.4 shows the characteristics of each dataset, including species richness and Pielou's evenness [49].

I generated simulated data by modifying the CRC and ME/CFS datasets and conducted extensive simulations to evaluate the performance of TMAT. I used the Silva and EzTaxon databases for OTU clustering, and the results from EzTaxon database is provided in Table 2.5. These results indicated that the results from edgeR were substantially inflated for both databases. TMAT and the other methods except edgeR preserved the nominal type-1 error rate at the 0.05, 0.01, 0.005, and 0.001 significance levels if the sample sizes were more substantial than or equal to 50. However, the proposed methods became more conservative as the sample size decreased, and a inverse normal transformation made them less conservative. No significant differences in OTU clustering were observed between the Silva and EzTaxon databases.

I also considered the effect of library size on type-1 error rates, and the results are provided in Table 2.6. Results showed that the TMAT statistics were not affected by the total read counts. The type-1 error rates of edgeR tended to increase as the library sizes increased, but other approaches were robust to the library size. Upon evaluating the effect of the number of leaf nodes (Table 2.7), results showed that $\mathrm{TMAT_M}$ became slightly conservative if the number of leaf nodes was larger than 30 but that $\mathrm{TMAT_{IM}}$ was less affected. Table 2.8 shows the effect of sparsity on the type-1 error rate. For each genus, I calculated its sparsity, defined as the proportion of subjects with no abundance, and type-1 error rates were calculated. Results showed that the type-1 error rates of edgeR were the most inflated and that some inflation was observed for OMiAT when the mean sparsity is greater than 10%. Slight inflation was shown for oMiRKAT when the mean sparsity was larger than 30%. oMiRKAT is based on the permutation, and the permutation-based p-value is generally robust to the non-normality. However, if there exists heteroscedasticity, its statistical validity can be impaired. A substantial amount of sparsity may induce the heteroscedasticity, which may explain the type-1 error inflation. The type-1 error rates for ANCOM became deflated. Some deflation was also observed for $\mathrm{TMAT}_{\mathrm{M}}$, but rates for $\mathrm{TMAT}_{\mathrm{IM}}$ were less deflated.

Table 2.4. Characteristics of CRC and ME/CFS datasets. Species richness and Pielou's evenness were calculated using the R package vegan 2.4-3 [49].

Dataset	Database	Number of OTUs	Richness (mean ± sd)	Evenness (mean ± sd)
CRC	Silva	129	115.0259 ± 7.1608	0.6522 ± 0.1017
	EzTaxon	152	137.7586 ± 7.1913	0.6567 ± 0.0990
ME/CFS	Silva	118	91.6092 ± 10.4139	0.6121 ± 0.1108
	EzTaxon	133	107.0920 ± 10.7418	0.6138 ± 0.1086

Table 2.5. Type-1 error estimates with OTUs clustered by EzTaxon database. The numbers of cases and controls were assumed to be the same. The total sample size is denoted by N, and I considered N = 20, 30, 50, and 70. All subjects were selected without replacement. Type-1 error estimates were calculated with 2,000 replicates at the significance level 0.05, 0.01, 0.005 and 0.001.

Data	Method	N = 20				N = 30				N = 50				N = 70			
		α = 0.05	α = 0.01	α = 0.005	α = 0.001	α = 0.05	α = 0.01	α = 0.005	α = 0.001	α = 0.05	α = 0.01	α = 0.005	α = 0.001	α = 0.05	α = 0.01	α = 0.005	α = 0.001
CRC	TMAT_{M}	0.03029	0.00237	0.00075	0.00003	0.03468	0.00424	0.00158	0.00012	0.04153	0.00612	0.00265	0.00030	0.04499	0.00730	0.00323	0.00041
	TMAT_{IM}	0.03715	0.00384	0.00110	0.00002	0.04054	0.00558	0.00218	0.00016	0.04498	0.00730	0.00323	0.00042	0.04673	0.00827	0.00378	0.00053
	Wilcoxon	0.04324	0.00892	0.00390	0.00070	0.04463	0.00962	0.00475	0.00089	0.04977	0.00983	0.00492	0.00101	0.04971	0.00976	0.00488	0.00091
	oMiRKAT	0.05360	0.01497	0.00803	0.00440	0.04897	0.00972	0.00492	0.00098	0.04951	0.00992	0.00497	0.00104	0.05096	0.01052	0.00530	0.00104
	OMiAT	0.06426	0.02164	0.01589	0.01059	0.06320	0.01758	0.01110	0.00504	0.05208	0.01205	0.00682	0.00248	0.05169	0.01162	0.00673	0.00245
	aMiSPU	0.05611	0.02399	0.01928	0.01400	0.05576	0.01734	0.01274	0.00187	0.04351	0.00938	0.00458	0.00108	0.04604	0.00916	0.00488	0.00098
	edgeR	0.24591	0.14590	0.12365	0.09419	0.12881	0.04785	0.02892	0.01259	0.21783	0.12601	0.11151	0.08692	0.23542	0.13926	0.11501	0.07729
	ANCOM	0.03185	0.00515	0.00088	0.00000	0.06246	0.01005	0.00377	0.00021	0.06238	0.01539	0.00763	0.00086	0.06759	0.01724	0.00845	0.00114
ME/CFS	TMAT_{M}	0.02884	0.00223	0.00065	0.00003	0.03485	0.00372	0.00124	0.00009	0.03959	0.00558	0.00218	0.00022	0.04340	0.00712	0.00305	0.00047
	TMAT_{IM}	0.03547	0.00362	0.00103	0.00002	0.04070	0.00545	0.00204	0.00015	0.04370	0.00712	0.00298	0.00038	0.04547	0.00790	0.00358	0.00053
	Wilcoxon	0.04283	0.00819	0.00365	0.00068	0.04574	0.00965	0.00470	0.00090	0.04901	0.00918	0.00456	0.00085	0.04870	0.00967	0.00478	0.00089
	oMiRKAT	0.07647	0.03454	0.03008	0.02372	0.05255	0.01237	0.00678	0.00280	0.05149	0.01285	0.00771	0.00374	0.05043	0.01064	0.00567	0.00170
	OMiAT	0.07933	0.02646	0.01947	0.01237	0.06282	0.02025	0.01448	0.00902	0.05212	0.01113	0.00586	0.00136	0.06041	0.01266	0.00682	0.00158
	aMiSPU	0.03857	0.00737	0.00397	0.00087	0.05228	0.00977	0.00516	0.00178	0.04294	0.00856	0.00419	0.00093	0.04541	0.00947	0.00540	0.00095
	edgeR	0.25105	0.13340	0.11478	0.07147	0.18830	0.09199	0.06348	0.01936	0.18910	0.09792	0.07946	0.04741	0.17870	0.07611	0.05310	0.02313
	ANCOM	0.06383	0.00804	0.00226	0.00000	0.07943	0.01518	0.00486	0.00026	0.05472	0.01073	0.00432	0.00037	0.05799	0.01185	0.00453	0.00033

Table 2.6. Effect of total read counts on type-1 error estimates. Subjects were sorted by total read counts and categorized into four groups, and for each group, the type-1 error rates were calculated separately. G1 consisted of subjects below the 25th percentile for total read counts, and G2, G3, and G4 consisted of those in the next three quartiles. I generated simulation data based on read counts from the CRC dataset clustered by the Silva database and assumed total sample size (N) is equal to 50.

Method	Group by total read counts	Significance level			
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
TMAT$_{\text{M}}$	G1	0.03617	0.00403	0.00144	0.00009
	G2	0.03881	0.00455	0.00161	0.00016
	G3	0.03615	0.00411	0.00151	0.00009
	G4	0.03716	0.00431	0.00159	0.00014
TMAT$_{\text{IM}}$	G1	0.04178	0.00572	0.00223	0.00016
	G2	0.04170	0.00566	0.00221	0.00022
	G3	0.04166	0.00590	0.00234	0.00020
	G4	0.04231	0.00603	0.00238	0.00020
Wilcoxon	G1	0.04597	0.00893	0.00414	0.00084
	G2	0.04590	0.00888	0.00427	0.00083
	G3	0.04635	0.00900	0.00436	0.00091
	G4	0.04637	0.00928	0.00445	0.00096
OMiAT	G1	0.06126	0.01568	0.00928	0.00340
	G2	0.06122	0.01674	0.01074	0.00504
	G3	0.05944	0.01602	0.01007	0.00495
	G4	0.05685	0.01841	0.01338	0.00894
oMiRKAT	G1	0.05312	0.01382	0.00876	0.00405
	G2	0.04814	0.00960	0.00501	0.00107
	G3	0.05129	0.01129	0.00592	0.00142
	G4	0.05415	0.01451	0.00965	0.00535
aMiSPU	G1	0.04253	0.00866	0.00431	0.00093
	G2	0.04244	0.00898	0.00408	0.00094
	G3	0.04235	0.00907	0.00449	0.00088
	G4	0.04334	0.00881	0.00423	0.00083
edgeR	G1	0.12434	0.04056	0.02528	0.01122
	G2	0.16231	0.06906	0.04895	0.02219
	G3	0.27072	0.18366	0.16007	0.11443
	G4	0.27333	0.15392	0.10788	0.05577
ANCOM	G1	0.04185	0.00753	0.00192	0.00010
	G2	0.03449	0.00613	0.00237	0.00007
	G3	0.03382	0.00526	0.00157	0.00011
	G4	0.03700	0.00592	0.00250	0.00006

Table 2.7. Effect of numbers of leaf nodes on type-1 error estimates. Families were categorized into four different groups according to the number of leaf nodes, and for each taxon, type-1 error rates were estimated. Simulation data were generated with read counts from the CRC and ME/CFS datasets. OTUs were clustered by the Silva database, and I assumed the total sample size (N) is equal to 50.

Method	Number of leaf nodes	Significance level			
		α = 0.05	α = 0.01	α = 0.005	α = 0.001
TMAT_M	2-9	0.04205	0.00666	0.00271	0.00036
	10-19	0.03468	0.00445	0.00193	0.00018
	20-29	0.03028	0.00415	0.00153	0.00013
	>=30	0.02198	0.00228	0.00073	0.00003
TMAT_IM	2-9	0.04389	0.00738	0.00323	0.00043
	10-19	0.03685	0.00550	0.00258	0.00030
	20-29	0.03458	0.00520	0.00220	0.00020
	>=30	0.03240	0.00463	0.00155	0.00010
Wilcoxon	2-9	0.04991	0.00964	0.00492	0.00099
	10-19	0.04990	0.00973	0.00523	0.00105
	20-29	0.04990	0.00940	0.00465	0.00108
	>=30	0.04955	0.00993	0.00455	0.00100
oMiRKAT	2-9	0.05012	0.01011	0.00513	0.00095
	10-19	0.05005	0.00970	0.00523	0.00130
	20-29	0.05088	0.01020	0.00533	0.00083
	>=30	0.04975	0.00950	0.00523	0.00113
OMiAT	2-9	0.05099	0.01083	0.00567	0.00159
	10-19	0.06615	0.01535	0.00938	0.00228
	20-29	0.07028	0.01773	0.01020	0.00275
	>=30	0.03670	0.00648	0.00283	0.00060
aMiSPU	2-9	0.04649	0.00921	0.00523	0.00124
	10-19	0.05090	0.00900	0.00503	0.00110
	20-29	0.05005	0.01090	0.00588	0.00138
	>=30	0.05143	0.01008	0.00528	0.00088
edgeR	2-9	0.17653	0.07416	0.05214	0.02360
	10-19	0.15088	0.04013	0.02175	0.00518
	20-29	0.01125	0.00090	0.00043	0.00005
	>=30	0.09413	0.02580	0.01235	0.00160
ANCOM	2-9	0.03831	0.00878	0.00393	0.00058
	10-19	0.03335	0.00790	0.00345	0.00055
	20-29	0.03178	0.00665	0.00325	0.00065
	>=30	0.02893	0.00883	0.00378	0.00058

Table 2.8. Effect of sparsity on type-1 error estimates. For each genus, I calculated its sparsity as the proportion of subjects with no abundance. Genera were sorted by their sparsity and categorized into three different groups, and for each taxon, type-1 error rates were estimated. Simulation data were generated by using read counts from the CRC dataset. OTUs were clustered by the Silva database, and I assumed the total sample size (N) is equal to 50.

Method	Mean sparsity level of genera	Mean number of leaf nodes	Significance level			
			$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
TMAT$_{\mathrm{M}}$	0-10%	2.58	0.03983	0.00540	0.00223	0.00021
	10-30%	4.00	0.04207	0.00635	0.00265	0.00033
	30-100%	2.75	0.04339	0.00666	0.00313	0.00030
TMAT$_{\mathrm{IM}}$	0-10%	2.58	0.04524	0.00751	0.00347	0.00044
	10-30%	4.00	0.04366	0.00744	0.00326	0.00043
	30-100%	2.75	0.04576	0.00744	0.00294	0.00044
Wilcoxon	0-10%	2.58	0.04908	0.00913	0.00467	0.00095
	10-30%	4.00	0.04844	0.00914	0.00459	0.00090
	30-100%	2.75	0.04799	0.00901	0.00440	0.00084
oMiRKAT	0-10%	2.58	0.05064	0.01052	0.00524	0.00125
	10-30%	4.00	0.04963	0.00980	0.00484	0.00086
	30-100%	2.75	0.05645	0.01158	0.00599	0.00188
OMiAT	0-10%	2.58	0.05332	0.01432	0.00917	0.00458
	10-30%	4.00	0.06285	0.01481	0.00802	0.00193
	30-100%	2.75	0.06271	0.01508	0.00844	0.00279
aMiSPU	0-10%	2.58	0.04346	0.00900	0.00476	0.00096
	10-30%	4.00	0.04444	0.00897	0.00474	0.00092
	30-100%	2.75	0.04205	0.00813	0.00398	0.00079
edgeR	0-10%	2.58	0.19448	0.11181	0.09343	0.07026
	10-30%	4.00	0.25196	0.13918	0.11599	0.07272
	30-100%	2.75	0.15773	0.07364	0.05633	0.03068
ANCOM	0-10%	2.58	0.03254	0.00689	0.00265	0.00021
	10-30%	4.00	0.03498	0.00727	0.00279	0.00022
	30-100%	2.75	0.01501	0.00306	0.00116	0.00009

I also calculated statistical power estimates with 2,000 replicates at the 0.05 significance levels, and these were compared with those of other statistical methods. I considered genera consisting of two or more OTUs. OTU clustering was conducted using the EzTaxon database (Figure 2.4).

In summary, I confirmed that TMAT is generally the most efficient among the available methods in the simulations. TMAT considers phylogenetic tree structures and uses log CPM transformation, which may lead to its superiority over other methods. OMiAT is the second most powerful, but its power substantially decreases if a genus has a single OTU. Furthermore, while OMiAT, oMiRKAT, and aMiSPU are based on permutation approaches, which can be computationally very intensive if the significance level is small, TMAT utilizes a distribution-based p-value and is therefore computationally fast (Figure 2.5).

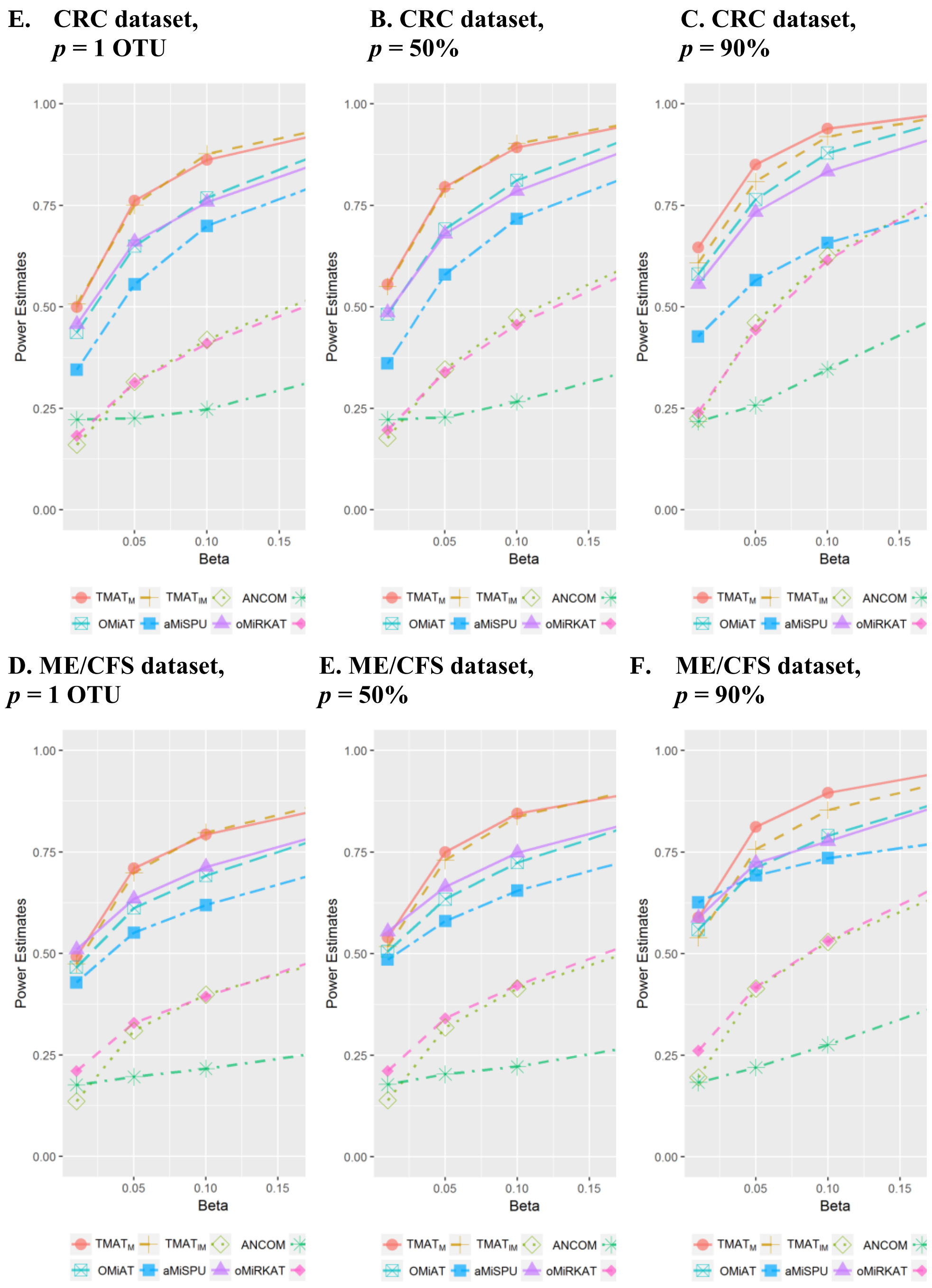
Real data analysis

The CRC and ME/CFS datasets were analyzed with TMAT and the other methods. The CRC dataset includes the age, sex, and body mass index (BMI) of the subjects, and the ME/CFS dataset contains the age and gender. It has often been shown that OTUs are affected by factors such as age and gender and that such factors consequently affect disease outcomes. Table 2.9 show that TMAT_M resulted in the greatest number of significant genera: *Fusobacterium, Lysinibacillus, Anaerostipes*, and *Streptococcus*. Figure 2.6 shows the internal nodes for *Fusobacterium* and their relative proportions. *Fusobacterium* has a single internal node with two different leaf nodes. The data showed that the relative abundances of leaf node m = 0 were much higher in controls and that those of m = 1 were higher in specific cases.

Fusobacterium nucleatum has been reported to be an opportunistic and commensal anaerobe related to periodontitis and appendicitis, and it is significantly associated with CRC. Interestingly, a recent study indicated that the relative ratio of *F. nucleatum* to probiotics plays an essential role in the detection of CRC [50]. Figure 2.7A shows that *Lysinibacillus* has a single internal node and two different leaf nodes. The results showed that the significance of *Lysinibacillus* was driven by internal node k = 1, with controls tending to have much higher abundances of m = 1 than cases. The antimicrobial potential of *Lysinibacillus* has been reported, and its bacteriocin can be used in foods to protect against cancer-inducing food-borne bacterial and fungal pathogens [51]. Thus, the negative correlation of *Lysinibacillus* with CRC is a credible result. These results confirm that the genera identified using TMAT may be associated with CRC. Thus, it can be concluded that TMAT successfully detected genera associated with host diseases.

Table 2.9. Association analysis results of CRC dataset with genera clustered by Silva database. OTUs were clustered with the Silva database, and associations of genera with CRC were tested. Results for genera significantly associated with at least one method at the FDR-adjusted 0.05 significance level were summarized.

Genus	TMAT_{M}	TMAT_{IM}	oMiRKAT	OMiAT	aMiSPU
Fusobacterium	0.00169	0.01200	NA	0.69875	NA
Lysinibacillus	0.01288	0.01884	NA	0.28178	NA
Anaerostipes	0.04380	0.04540	NA	0.14308	NA
Roseburia	0.04380	0.09218	0.06752	0.11131	0.10439
Streptococcus	0.05028	0.04540	0.06752	0.19117	0.10149



E. CRC dataset, *p* = 1 OTU

B. CRC dataset, *p* = 50%

C. CRC dataset, *p* = 90%

D. ME/CFS dataset, *p* = 1 OTU

E. ME/CFS dataset, *p* = 50%

F. ME/CFS dataset, *p* = 90%

Figure 2.4. Power estimates for genera consisting of more than one OTU clustered by EzTaxon database. Power estimates at the significance level of 0.05 were calculated with 2,000 replicates. I generated simulation data based on read counts from CRC and ME/CFS datasets and considered genera with more than one OTU. OTUs were clustered by EzTaxon database.

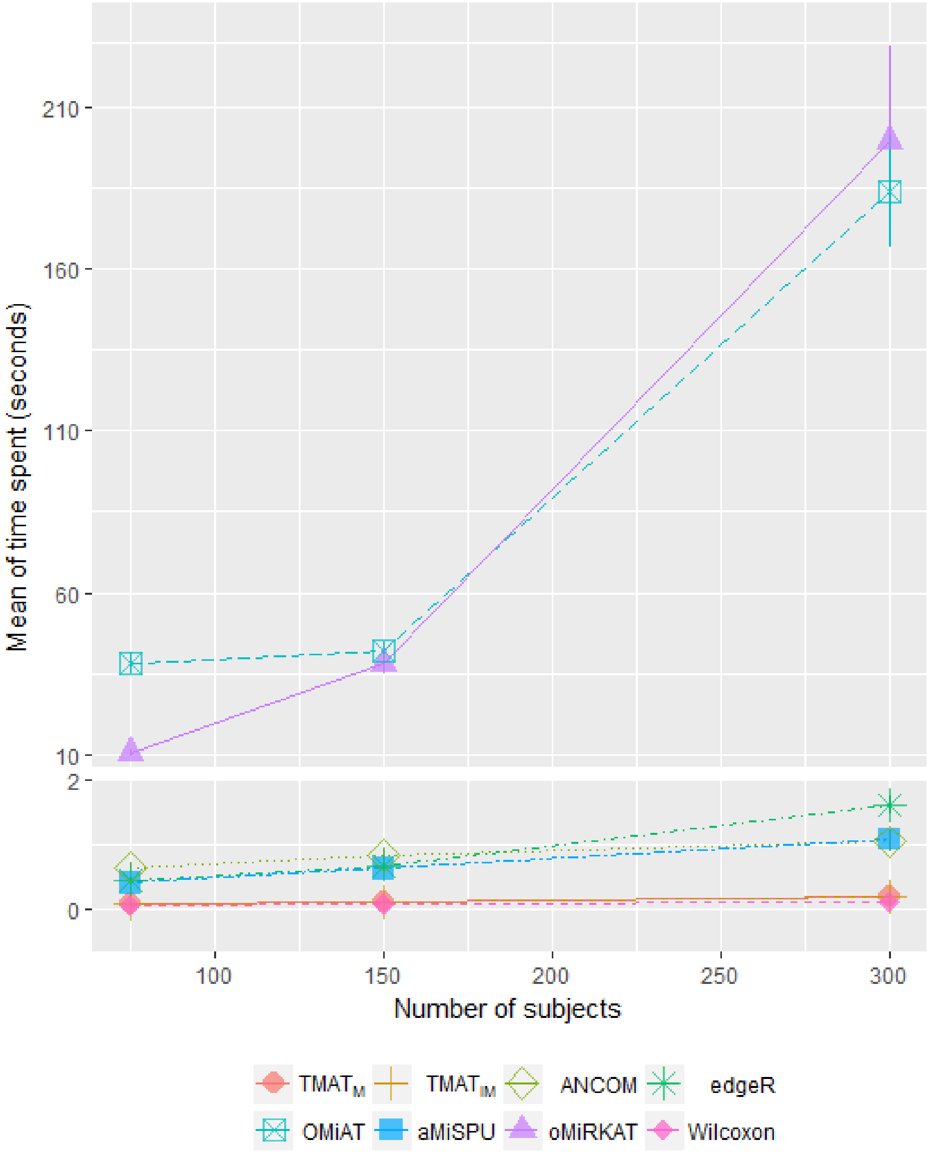


Figure 2.5. Comparison of computational costs. For CRC dataset, computational costs of TMAT, optimal oMiRKAT, aMiSPU, OMiAT, Wilcoxon, edgeR, and ANCOM were compared. The number of permutation-based p-values for MiRKAT, aMiSPU, and OMiAT was set to 1,000. Means of the computational time required for 15 different CRC datasets and their 95% confidence intervals are provided. Analyses were conducted using the R package microbenchmark (version 1.4-4) with an Intel(R) Xeon(R) CPU E3-1230 v3 @ 3.30GHz processor.

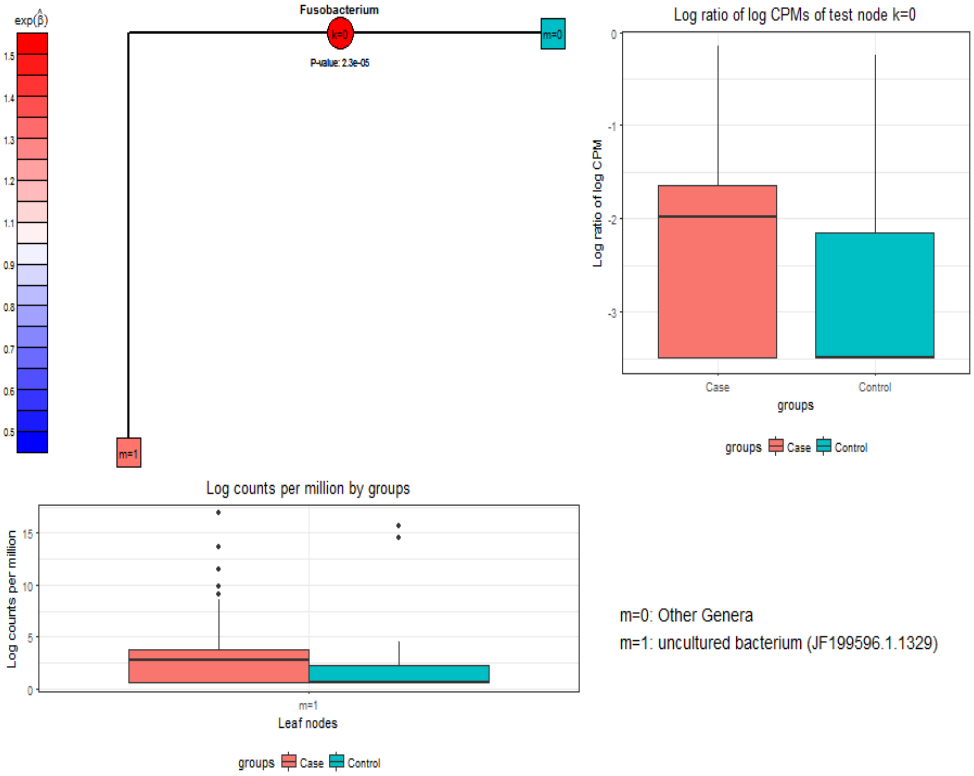
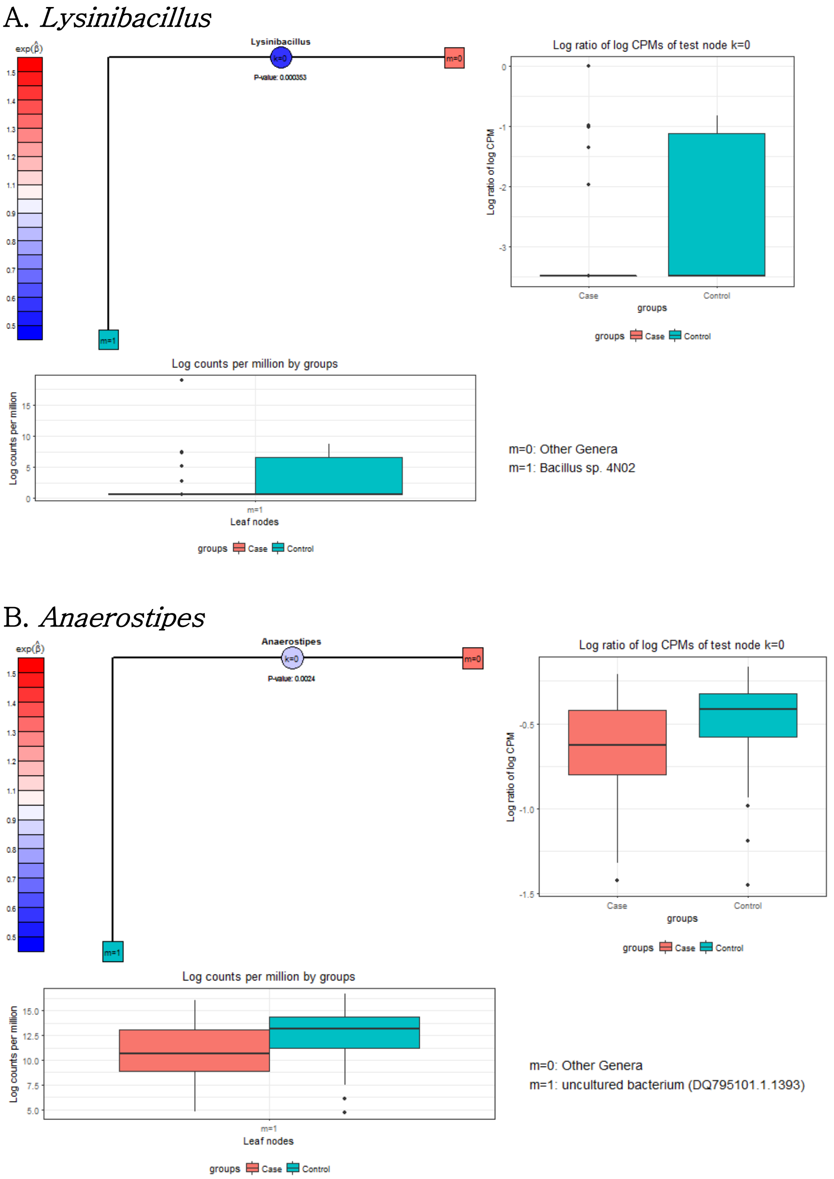


Figure 2.6. OTU distributions of *Fusobacterium*. Relative proportions of OTUs belonging to *Fusobacterium* were calculated. The blue internal node indicates that OTUs are more abundant in cases than controls. Each OTU has its corresponding leaf node, and leaf nodes in green and red indicate that they are more frequently observed in cases and controls, respectively. For exp($\widehat{\boldsymbol{\beta}}$), $\widehat{\boldsymbol{\beta}}$ indicates the maximum likelihood estimate by the quasi-likelihood method, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^k_i/D^k_i between cases and controls after adjusting for covariates. OTUs were clustered by the Silva database.



A. *Lysinibacillus*

B. *Anaerostipes*

Figure 2.7. OTU distributions of significantly associated genera for CRC dataset based on Silva database. Relative proportions of OTUs belonging to significantly associated genera according to TMAT$_M$ were calculated. The blue internal node indicates that OTUs are more abundant in cases than in controls. Each OTU has its corresponding leaf node, and leaf nodes in green and red indicate that they are more frequently observed in cases and controls, respectively. For exp($\hat{\boldsymbol{\beta}}$), $\hat{\boldsymbol{\beta}}$ indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\hat{\boldsymbol{\beta}}$) indicates the mean difference of C^k_i/D^k_i between cases and controls after adjusting for covariates. OTUs were clustered by the Silva database. Roseburia was omitted for the resulting plot.

Input and output of AMAA

AMAA supports two types of data input files: fastq files, and microbial count table. When fastq file is given, only four mandatory inputs are needed. A path to the fastq folder, the adaptor sequences, a range of length for 16S rRNA gene target region and meta file. Other options for Cutadapt, Casper, VSEARCH, UCLUST and DADA2 will be synchronized with analysis environment file, env.ini, and can be easily edited. The analysis can be separately conducted with the option of range of sub-analysis to be conducted. Output of processing sequence data is OTU/ASV table.

For microbiome compositional analysis, output will be microbial compositional plot, alpha- and beta- diversity output, result of PERMANOVA and plots based on NMDS, PCA and Kernel PCA. Metagenome-wide association analysis will be conducted at genus level as a default option and can be changed. The final output will be p-values for each method, effect size if available, and FDR adjusted p-values. The result file will be generated for each combination of clustering method and database.

Specification of strength and weakness of methods in microbiome association analysis

Metagenome-wide association study in AMAA, provides the different methods for microbiome association analyses depending on the types of trait data (binary or continuous), robustness to compositional bias, availability of covariate adjustment, set based analysis and OTU level analysis. By integrating different types of association methods designed for specific research characteristics into a common interface, AMAA enables more extensive and systematic microbiome association analyses. In Tables 2.10, I summarized the characteristics of each methods and proper situation to use for the microbiome association test available in AMAA. Detailed procedure including input file achievement user can get for each analysis step is described in Figure 2.8.

Similarity of databases based on CRC dataset

OTU count table were generated based on Silva, EzTaxon and Greengene databases using CRC dataset. For each OTU, the relative proportion was calculated and determined the mean value across all subjects. If the resulting value was smaller than the cutoff value, the OTU was excluded for the calculation of similarity. All the OTU count tables were rarefied with the minimum library size across all the samples of all the count tables. Similarity was calculated between any possible pairs of database, and is shown in Figure 2.9. The cutoff, 10^{-3}, which is often used for quality control of OTUs is shown with vertical black dashed line [18, 37, 52]. The similarity for every pairs of databases monotonically increased when $\log_{10}$ cutoff increases from -7 to -3.30 (from 10^{-7} to 5×10^{-4} for raw cutoff value). The mean value of the number of remained OTUs decreased according to cutoff values and the similarities started to up and down irregularly when cutoff is more than -2.70 (cutoff = 2×10^{-3}) and the $\log_{10}$ mean number of OTUs is less than 1.83 (The mean number of OTUs ≈ 67) (Figure 2.9). The similarity of the databases were maximized near the cutoff, 10^{-3}, when the mean number of OTUs is more than 67.

Table 2.10. Availability of different variable types, covariates, information and analyses support in association analyses in AMAA.

	Trait type			Covariate adjustment	Single OTU analysis	Effect size and direction provided	Compositional bias corrected	Repeatedly measured	Reference and software
	binary	continuous	censored time-to-event						
TMAT	Y	Y	N	Y	Y	Y	Y	Y	Kim et al. (2020)
ANCOM	Y	N	N	Y	Y	N	Y	N	Mandal et al. (2015)
MiRKAT	Y	Y	N	Y	N	N	N	Y	Wilson et al. (2020)
MiRKAT-s	Y	Y	Y	Y	N	N	N	N	Plantinga et al. (2017)
aMiSPU	Y	Y	N	Y	N	N	N	N	Wu et al. (2016)
OMiAT	Y	Y	N	Y	Y	N	N	N	Koh et al. (2017)
OMiSA	Y	Y	Y	Y	Y	N	N	N	Koh et al. (2018)
EdgeR	Y	Y	N	Y	Y	Y	N	N	Robinson et al. (2010)
Wilcoxon	Y	N	N	N	Y	Y	N	N	Bauer et al. (1972)
GLMM-MiRKAT	Y	Y	N	Y	N	N	N	Y	Koh et al. (2019)
ZIBR	Y	Y	N	Y	Y	Y	N	Y	Chen et al. (2016)
cSKAT	N	Y	N	Y	Y	Y	N	Y	Zhan et al. (2018)
FZINBMM	Y	Y	N	Y	Y	Y	N	Y	Zhang et al. (2020)

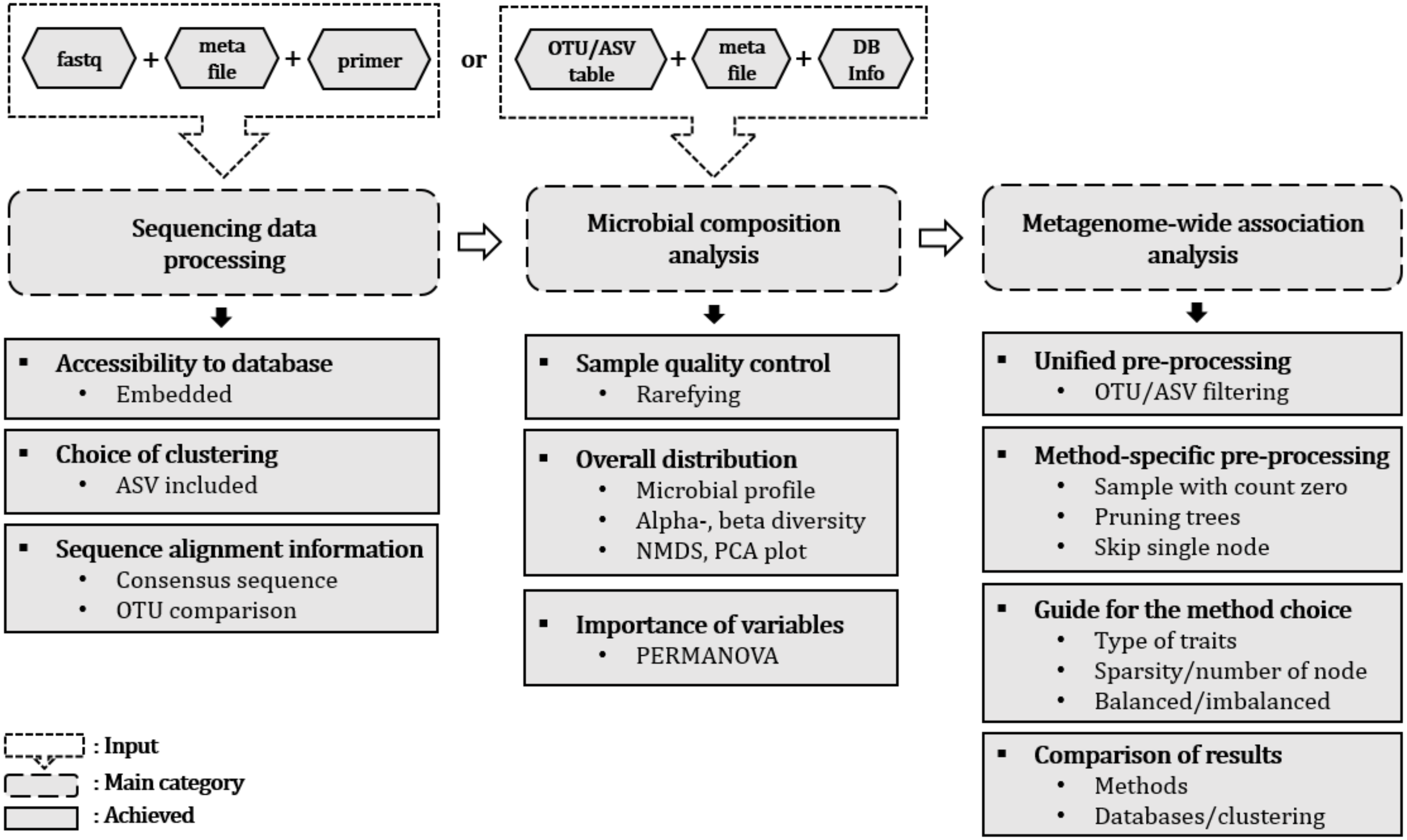


Figure 2.8 Flow chart of analysis using AMAA

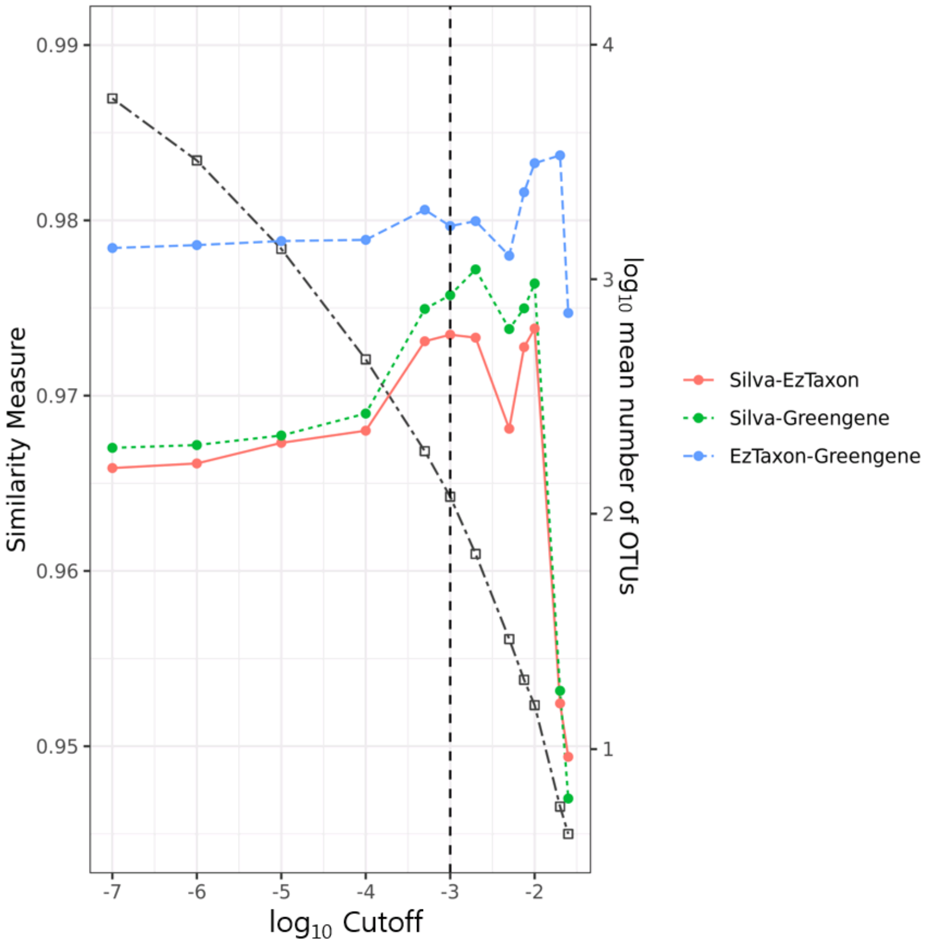


Figure 2.9. Similarity between each pair of databases. Similarity measure were calculated for each database pair, Silva and EzTaxon, Silva and Greengene and EzTaxon and Greengene. Calculated similarities according to different $\log_{10}$ cutoff are shown. The vertical line shows the cutoff value 10^{-3}. For each database, the number of OTUs remaining after filtering for the corresponding cutoff is calculated. The value obtained by taking the $\log_{10}$ of the average of the number of OTUs across all the databases is indicated by black squared dotted dash line.

2.4. Discussion

The importance of microbiome-host interactions has been known for more than a century [53], and it has been shown that the occurrence of many human diseases is related to bacterial communities. However, the abundances of many OTUs are very low, and inter-subject variation is high, complicating statistical analyses. Here, I propose a new method for detecting OTUs associated with host diseases. TMAT statistics are based on quasi-scores for internal nodes in a phylogenetic tree, and those statistics are combined into a single statistic with a minimum p-value. By using such quasi-score statistics, TMAT can identify differences among OTUs significantly associated with host diseases, while existing statistical methods, such as aMiSPU, OMiAT, and oMiRKAT, cannot. Furthermore, by the nature of the proposed statistics, the statistical scores for internal nodes are independent, and the minimum p-value can be directly calculated. I compared the performance of TMAT with those of oMiRKAT, aMiSPU, and OMiAT under various simulation scenarios. According to the results, TMAT correctly controlled the nominal type-1 error rate and was statistically the most powerful method for detecting associations with host diseases in the simulation studies. Furthermore, TMAT is computationally less intensive than the other methods, allowing the completion of statistical analyses within a few minutes. It should be noted that previous methods, such as OMiAT, can require several days when the sample size is larger than 1,000. I implemented TMAT using the R package with multiple functions for association analyses, and this implementation is available at http://healthstat.snu.ac.kr/software/tmat.

However, despite the flexibility of TMAT, the proposed method has several limitations. First, there are multiple software programs available for OTU clustering and multiple 16S rRNA gene sequence databases, and the statistical properties of TMAT ultimately depend on which method is used. The statistical power of TMAT should be maximized when the OTU clustering results are the most accurate, but the best strategy for OTU clustering remains unclear. Multiple studies have compared the accuracy of databases by using a mock community whose microbial composition is known, and the EzTaxon database was reported to be the most accurate among the existing databases, including the Silva and Greengenes databases [25, 26]. Thus, I considered the EzTaxon database for the simulation. However, the most accurate database can vary according to the samples analyzed. This issue can be handled statistically with a simple modification of the proposed statistics. For instance, OTUs can be clustered under multiple conditions, and the statistics for association analyses with OTU clustering can be combined with Fisher's combination method or minimum p-value approaches. Second, I showed that TMAT outperforms existing methods using extensive simulations. If a single OTU is associated with disease status, a single statistic corresponding to the OTU is expected to be significant, while the others should not be significant. If most of the OTUs are associated with disease status, T_0 or T_1 is expected to be significant, and the others not significant. In both scenarios, a single p-value is significant, and the minimum p-value method is known to be the most powerful for combining p-values. This is why I considered the minimum p-value method. However, the results depend on the simulation settings and cannot be generalized to different simulation scenarios. For instance, the minimum p-value approach is less efficient when most of T_k within the same genus are significant. For Figure 1 B, I assume that two OTUs, m = 1 and 3, affect the host diseases while the other two OTUs do not. Next, the score statistics for the internal nodes for k = 2, 3 become significant. In this case, Fisher's method is expected to be more powerful and can be a valuable alternative. Further extensive simulation studies are still necessary. Third, statistical power and type-1 error of TMAT are affected by the number of leaf nodes. In case the latter is significant, TMAT becomes conservative and loses statistical power. In the simulation studies, I found that TMAT is uniquely conservative, but it can be adjusted by using permutation used for other methods. Power loss is observed for all methods. If the number of nodes is large, the effect of OTUs on host diseases can be heterogeneous, and statistical analyses should be conducted carefully. Fourth, I assumed that the absolute read count for each leaf node follows Poisson distribution and showed that scores for internal nodes are independent. I found that the independence is preserved in the real datasets. However, this can be violated in some scenarios. In such case, I suggest to test their independence by using Kolmogorov Smirnov test, etc, and if it is violated, I recommend to use a robust method such as permutation. Lastly, 16S rRNA gene sequencing clustering enables the identification of taxa associated with host diseases at the genus or phylum levels, but the accuracy of OTUs at the species level remains controversial. Besides, recent improvements in sequencing technology have enabled the detection of functional genes using metagenomics shotgun sequencing, and several approaches have been proposed to handle such data. However, the current version of TMAT cannot be applied to the detection of functional genes in metagenomics data. Future work will aim to further develop the method to extend its applicability to such data.

AMAA enables a researcher to perform many of the microbiome association analysis with different choice of input file, database, clustering methods and test statistics. AMAA supports both of two common types of data input files, fastq and microbial count table. It provides a unified preprocessing procedures for association methods and a rich choice of methods for association analysis based on different database and clustering methods. This enables a researcher can get comprehensive information for viable associations depending on the choice of database and clustering methods. Similarity measure between a pair of databases can evaluate the similarity between a pair of databases based on a certain dataset. AMAA also has a limitation. As reference sequences of microbial clusters are different based on the choice of clustering methods and databases, comparison of the results across different clustering methods and databases are limited. Similarity measure of databases is limited in that it is based on beta diversity rather than similarity of reference sequences, and results may vary across datasets.

Over the last decades, it has been expected that bacterial communities may be associated with many disease conditions in humans; however, association analyses have not met with expectations owing to the absence of a standard analysis toolset with efficient and reproducible statistics. The proposed TMAT methods and AMAA package allow non-experts to efficiently conduct statistical analyses with small computational costs, which may lead to an improved understanding of the complex interplay between bacterial communities and hosts.

Chapter 3. Longitudinal Microbiome Association Test based on Phylogenetic Tree

3.1. Introduction

Recent advance of high-throughput technologies such as microarrays and next-generation sequencing has greatly increased our understanding of the microbial world. For instance, it has been shown that intestinal microbiota plays essential roles in host by affecting energy homeostasis, body adiposity, blood sugar control, insulin sensitivity and hormone secretion [54-56]. However, the abundances of microbial taxa are often sparse with excessive zeros, and taxa observed in all samples are usually rare. Most of microbiota were observed in a small proportion of samples and this makes statistical testing hard to control the type-1 and type-2 errors. In addition, microbiota are highly variable because they are affected by various factors, such as age and sex. Therefore, caution should be taken when inferring causal relationships through statistical analysis of microbiota data.

Longitudinal microbiota studies are useful to detect microorganisms related to the progression of disease and identify the change along the time, and provides more evidence for the causal relationship than cross-sectional studies [57]. Furthermore the estimation of within-subject covariate effects is robust against between-subject confounders, and longitudinally measured microbiome data enable the robust identification of microbiota effects on the risk of diseases in the host. Statistical analyses with repeatedly observed 16S rRNA requires the adjustment of similarity among the measurements of the same subjects. However existing methods which can be applied to repeatedly observed 16S rRNA data are limited, and statistical method development for longitudinal studies are needed to investigate the association between the human microbiome and diseases.

Statistical methods of longitudinal data analysis in microbiome studies has been comprehensively reviewed by Xia et al [58]. Those can be categorized into several categories: (1) standard longitudinal model, (2) overdispersed and zero-inflated longitudinal models (3) multivariate distance/kernel-based longitudinal models. First, standard longitudinal model includes such as linear mixed effect model (LMM) with generalized estimation equation (GEE) and generalized linear mixed effect model (GLMM). Class LMM method provide a standardized and flexible approach to model both fixed and random effects. However, OTU abundances should be transformed or normalized to avoid the violation of distribution assumptions and cannot address the sparsity issue. Second, overdispersed and zero-inflated longitudinal models include zero-inflated Gaussian (ZIG) mixture model, extensions of negative binomial mixed-effects (NBMM) [59] and zero-inflated negative binomial models (FZINBMM) [60]. Two-part zero-inflated beta regression model with random effects (ZIBR) extends zero-inflated beta regression model to longitudinal data setting [10]. FZINBMM and ZIBR can analyze overdispersed and zero-inflated longitudinal metagenomics data. Last, multivariate distance/kernel-based longitudinal model includes correlated sequence kernel association test (cSKAT) for continuous outcome and generalized linear mixed model and its data-driven adaptive test (GLMM-MiRKAT) for non-normally distributed outcome such as binary traits. However all of those methods are vulnerable to compositional bias [18, 61].

In this article, I propose longitudinal microbiome association test based on phylogenetic tree (mTMAT) which is an extended version of TMAT. mTMAT pools the abundance of OTUs based on the phylogenetic distance which corrected zero-inflated problems. With extensive simulation and real data analyses, I prove its robustness against compositional bias and misclassified variance covariance structures, and statistical power improvement compared to other methods.

3.2. Materials and Methods

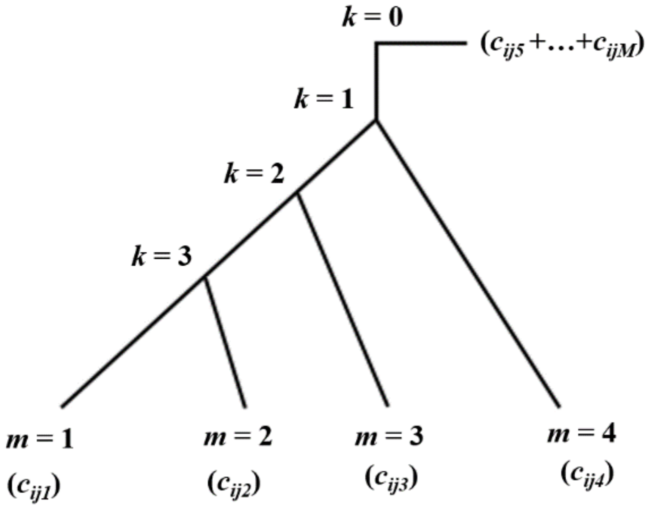
Ethics statement

The protocol used in this study was approved by the Institutional Review Board (IRB No. E2108/001-001) in Seoul National University.

Phylogenetic tree

The same notations and assumptions as TMAT was implemented [52]. Let us denote that the absolute abundance of OTU m of subject i at time point j as c_{ijm}, where $i = 1, \dots ,$ $N, j = 1, \dots , Ni$, m = 1, … , M. I assumed that OTUs are clustered by profiling sequence of all subjects at all the time points simultaneously and a rooted binary phylogenetic tree was provided for these OTUs. The first M_1 OTUs belong to a taxonomy of interest for the analysis of its association with host diseases, while the other $M - M_1$ OTUs belong to other taxonomy. For the genus with the first M1 OTUs, there are $M_1 - 1$ internal nodes and M1 leaf nodes. Internal nodes are denoted by k, where k =1, …, $M_1 - 1$. Leaf nodes are denoted by m, where $m = 1, \dots , M$. For each leaf node there is a corresponding single OTU; if m = 1, …, or M_1, m is the leaf node of the genus of interest, and otherwise m belongs to a different genus. The absolute abundance of OTU m of subject i at time point j is denoted by c_{ijm}. Under the assumption that mutations be transmitted from the left (right) child node to all of its leaf nodes, the relative proportion of leaf nodes of the left child node increases for cases if the mutation occurs during transmission to its left child node, and decreases if it does so during transmission to the right child node. If the association of an internal node k with a host disease of interest is tested, I let the internal node k and its leaf nodes represent a test node and test leaf nodes, respectively. The left and right test leaf nodes further represent the leaf nodes of the left and right child nodes of a test node, respectively.

For internal node k in the genus with M_1 OTUs, I let L_k and R_k be the sets of its left and right leaf nodes, respectively. Figure 3.1 illustrates these definitions.



$$M_1 = 4, k \in \{0, 1, 2, 3\}, m \in \{1, 2, 3, 4\}$$
$$k = 0 \Rightarrow L_0 = \{m|m = 1, 2, 3, \text{or } 4\}, R_0 = \{m|m = 5, \ldots, M\}$$
$$k = 1 \Rightarrow L_1 = \{m|m = 1, 2, \text{or } 3\}, R_1 = \{m|m = 4\}$$
$$k = 2 \Rightarrow L_2 = \{m|m = 1, \text{or } 2\}, R_2 = \{m|m = 3\}$$
$$k = 3 \Rightarrow L_3 = \{m|m = 1\}, R_3 = \{m|m = 2\}$$

Figure 3.1. Examples of rooted binary phylogenetic trees.

Quasi-likelihood

The log-transformed CPM *r*r_{ijm}, which is used for the edgeR package (version 3.16.5), is defined by as follows.

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

x_{ij}^k, where $k = 1, \ldots, M_1 - 1$, is defined by

$$x_{ij}^k = log\left(\frac{C_{ij}^k}{D_{ij}^k}\right), C_{ij}^k = \sum_{m=1}^{M_1} r_{ijm} \cdot I(m \in L_k), D_{ij}^k = \sum_{m=1}^{M_1} r_{ijm} \cdot I(m \in R_k).$$

As all OTUs in the genus can be associated with the host disease, x_i^0 for such case is defined by

$$x_{ij}^0 = log\left(\frac{C_{ij}^0}{D_{ij}^0}\right), C_{ij}^0 = \sum_{m=1}^{M_1} r_{ijm}\, , D_{ij}^0 = \sum_{m=M_1+1}^{M} r_{ijm}.$$

The phenotype of subject *i* at time point *j* is denoted by y_{ij} and is coded as 1 and 0 for cases and controls, respectively. Their vectors and matrices for testing the association of the genus of interest are defined by

$$\mathbf{x}_i^k = \begin{pmatrix} x_{i1}^k \\ \vdots \\ x_{iN_i}^k \end{pmatrix}, \ \mathbf{x}^k = \begin{pmatrix} x_{11}^k \\ \vdots \\ x_{1N_1}^k \\ \vdots \\ x_{NN_N}^k \end{pmatrix}, \quad \boldsymbol{X} = (\boldsymbol{x}^0 \quad \cdots \quad \boldsymbol{x}^{M_1-1}),$$

$$\boldsymbol{y}_{\mathrm{i}} = \begin{pmatrix} y_{\mathrm{i}1} \\ \vdots \\ y_{\mathrm{i}N_i} \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1N_1} \\ \vdots \\ y_{NN_N} \end{pmatrix}.$$

Here, I assume that $\boldsymbol{x}^0, \dots$, and $\boldsymbol{x}^{M_1-1}$ are ordered according to the depth of the internal nodes. $\boldsymbol{x}^0$ is used for testing the association of all OTUs belonging to the genus of interest by pooling them, and $\boldsymbol{x}^1$ is for testing the root node of the phylogenetic tree.

If I denote $\boldsymbol{R}_i$ and σ_{kk} as working correlation matrix and over dispersion parameter and define $\boldsymbol{D}_{ik}$ as diagonal matrix with its diagonal entries are $\mathrm{var}(\mathbf{x}_{ij}^k)$ *j = 1,…,* N$_{\mathrm{i}}$, the covariance matrix for the observations of subject i is defined by

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

Then the covariance matrix $\boldsymbol{\Sigma}^k$ can be defined as

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

If I let Z be a design matrix for *p* covariates including the intercept, I assume

$$E(\boldsymbol{x}^k|\boldsymbol{Z},\boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{\alpha}_k + \boldsymbol{y}\boldsymbol{\beta}_k, \qquad var(\boldsymbol{x}^k|\boldsymbol{Z},\boldsymbol{y}) = \boldsymbol{\Sigma}^k, \qquad \mathrm{k} = 0,\dots,\mathrm{M}_1 - 1$$

Therefore, quasi-score functions for $\boldsymbol{\alpha_k}$ and $\boldsymbol{\beta}_k$ can be denoted by

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

Quasi-fisher information can be denoted by

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

Score test with small sample adjustment

The null hypothesis can be denoted as

$$H_0\text{:}\ \boldsymbol{L}\begin{bmatrix}\boldsymbol{\alpha}_k \\ \beta_k\end{bmatrix} = \boldsymbol{0}$$

where $\boldsymbol{L}$ is a matrix of linear constraints with c rows and number of columns equal to the length of $\begin{bmatrix}\boldsymbol{\alpha}_k \\ \beta_k\end{bmatrix}$ and $\boldsymbol{0}$ is the zero vector of matching dimension. To test the null hypothesis H_0: $\beta_k = 0$, the generalized score statistics by Boos can be provided [62] by setting $\boldsymbol{L} = [\boldsymbol{0}^t \quad 1]$ as follows:

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

where

$$\widetilde{\boldsymbol{H}} = \sum\nolimits_{i=1}^{N} -\boldsymbol{D}_i^t\Sigma_i^{-1}\boldsymbol{D}_{\boldsymbol{i}},\ \boldsymbol{D}_{\boldsymbol{i}} = [\boldsymbol{Z}_i \quad \boldsymbol{y}_i],$$

$$\widetilde{\boldsymbol{B}} = \sum\nolimits_{i=1}^{N} U_i(\boldsymbol{\alpha}_{\boldsymbol{k}}, 0)U_i(\boldsymbol{\alpha}_{\boldsymbol{k}}, 0)^t \ \text{[63].}$$

To adjust the small sample bias, $\widetilde{\boldsymbol{B}}$ is further updated by

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

where

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

$$\widetilde{\boldsymbol{P}}_{ii} = \boldsymbol{D}_{\boldsymbol{i}}\left(\boldsymbol{I} - \widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^t\left(\boldsymbol{L}\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{L}^t\right)^{-1}\boldsymbol{L}\right)\widetilde{\boldsymbol{H}}^{-1}\boldsymbol{D}_i^t\Sigma_i^{-1} \ \text{[64].}$$

Wald test with small sample adjustment

The Wald statistic with sandwich estimator with correction of small sample bias was considered [63]. $\boldsymbol{\beta}_k$ can be estimated by solving the estimating equation $U_{\boldsymbol{\beta}}(\boldsymbol{\alpha_k}, \boldsymbol{\beta}_k) = 0$ as

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

For the estimation of variance of $\widehat{\boldsymbol{\beta}}_k$ I consider robust variance estimator with small sample adjustment as

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

where

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

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

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

Therefore, the robust Wald statistic of $\boldsymbol{\beta}_k$ for the test node k is defined as

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

Model selection with quasi-information criterion

Quasi-information criterion (QIC) for generalized estimating equation [65] can be used to find the best working correlation matrix by achieving the minimum QIC, and is defined as

$$QIC(R) = -2\Psi\left(\hat{\beta}(R); I\right) + 2trace\left(\hat{\Omega}_I^k \hat{V}_{LZ}^k\right)$$

where $\Psi\left(\hat{\beta}(R); I\right)$ and $\hat{\Omega}_I^k$ are the sum of quasi likelihood and a variance component under the independence working correlation assumption and $\hat{V}_{LZ}^k$ is a covariance matrix of $\hat{\beta}$ under the hypothesized working correlation $\boldsymbol{R}_i$ defined by [66].

mTMAT

Statistics for $H_0: \beta_k = 0$ can be combined to test $H_0: \beta_0 = \beta_1 = \cdots = \beta_{M_1-1} = 0$ using the minimum p-value. If p-values for T_k are denoted by pT_k, the proposed statistics, mTMAT$_M$, is defined by

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

It should be noted that $pT_0, \quad \cdots \quad , pT_{M_1-1}$ are asymptotically independence [52]. Therefore, I can conclude

$$mTMAT_M \sim beta(1, M_1) \ under \ H_0.$$

If the sample size is small, normality of T_k under H_0 may not be achieved, and the assumption of the quasi-score test can be violated. If I apply the inverse normal transformation to $x_{11}^k, \dots, x_{NN_N}^k$, then the same statistics can be obtained. This is denoted by T_k^{INT}. Rank-based inverse normal transformation with adjust parameter 0.5 is used for the transformation and data with tie values were mapped to a same value in the transformed data [33]. Then, mTMAT$_{IM}$ is defined by

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

KARE Cohort data

The KARE cohort is a prospective study cohort involving subjects from the rural community of Ansung and the urban community of Ansan in South Korea. It began in 2001 as part of the Korean Genome Epidemiology study [67], and I used data from 2,072 urine samples from 691 subjects participated in 2013, 2015, and 2017. Their 16S rRNA amplicon sequencing data from the study were available from the NCBI Sequence Read Archive database under project accession number PRJNA716550. Paired-end sequencing of the V3-V4 region of the bacterial 16S rRNA gene used the widely used primers 16S_V3_F (5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTAC-GGGNGGCWGCAG -3') and 16S_V4_R (5'- GTCTCGTGGGCTCGGAGATGTGTAT-AAGAGACAGGACTACHVGGGTATCTAATCC-3'). Adaptor sequences were detected and removed using the CUTADAPT software with a minimum overlap of 11, maximum error rate of 10%, and a minimum length of 10 [41]. Sequences were merged using CASPER with a mismatch ratio of 0.27 and filtered by the Phred (Q) score, resulting in sequences 350–550 bp in length [42, 68]. After the merged sequences were dereplicated, chimeric sequences were detected and removed using VSEARCH and the Silva Gold reference database for chimeras [43]. The open-reference Operational Taxonomic Units (OTU) picking was conducted based on the EzTaxon database using UCLUST [46, 69]. Phylogenetic trees based on EzTaxon database were obtained through the SINA method [36] using the reference sequences available from the EzTaxon database. For each OTU, I calculated its proportion among all OTUs and determined the mean value across all subjects. If the resulting value was <0.001, the OTU was excluded [70]. Among the 691 subjects, those with a read count <3,000 or for whom genomic data were not available in any phase were excluded. As a result, 1179 samples from 393 subjects, including 70 genera, were used for the simulation analysis.

Simulation studies

I conducted extensive simulations to evaluate the performance of mTMAT with two datasets; one with 393 subjects participated in all the three phases from KARE cohort and generative dataset based on microbiomeDASim [71]. The disease status of the subjects was permuted, and certain numbers of cases and controls were assumed to be missing to identify the effect of the unbalancedness. The randomly selected a single test node from the internal nodes, and from their test leaf nodes, either a single OTU, 50% of OTUs, or 90% of OTUs were randomly selected as causal OTUs. These were denoted by p = 1 OTU, 50%, and 90%, respectively. It should be noted that p = 1 indicates that there is a single OTU associated with the host disease, and thus, the phylogenetic tree structure does not provide any useful information for mTMAT. If I let the sample variances of c_{im} for causal OTUs be $\hat{\sigma}_{mm}$, the observed absolute abundances of the selected causal OTUs for only affected subjects was assumed to be $\delta = \beta\hat{\sigma}_{mm}$ where $\beta =$ 0, 0.01, 0.02, or 0.04, and the absolute abundances of the other OTUs were used without any modification. $\beta = 0$ was considered for estimation of empirical type-1 error rates, and the others were used for estimating statistical power. Type-1 error rates were estimated at the 0.1, 0.05, 0.01 and 0.005 significance levels with 5,000 replicates. Empirical power was estimated at the 0.05 significance level with 500 replicates.

For the comparison with mTMAT$_{\mathrm{M}}$ and mTMAT$_{\mathrm{IM}}$, GLMM-MiRKAT (version 1.2), FZINBMM (version 1.0), linear mixed model (LMM) with arcsine square root transformation (LMM-arcsine) and LMM with log transformation (LMM-log) with nlme package (version 3.1) were considered. TMAT (version 1.01), oMiRKAT (version 0.02), MiSPU (version 1.0) and the Wilcoxon test were also considered for the comparison with cross-sectional methods. Association analyses were conducted at the genus level. FZINBMM, LMM models and Wilcoxon were applied by pooling all OTUs within each genus. Each genus consisted of multiple OTUs, and oMiRKAT and MiSPU were applied to OTUs belonging to each genus.

For mTMAT$_{M}$ and mTMAT$_{IM}$, robust wald and score statistics with four different choices of working correlation matrix, identity, compound symmetry (CS), autoregressive with order 1 (AR1) and unstructured, were considered. MiSPU and oMiRKAT use permutation-based p-values, and they were calculated with 500 and 5,000 permutated replicates for estimation of power and type-1 error rates, respectively. GLMM-MiRKAT and oMiRKAT offer several distance metrics, including Unifrac distance as a default choice, while MiSPU also uses Unifrac distance as the default option. I considered the default choices; however, Unifrac distance cannot be calculated if read counts are not observed. Thus, subjects with no read counts were excluded from GLMM-MiRKAT, oMiRKAT and MiSPU. Furthermore, none of these can analyze a genus with a single OTU; hence, such instances were not considered for statistical power estimations of such genera.

With the simulation with the generated dataset with microbiomeDASim, Identity, CS and AR1 with different value of parameter is assumed for the simulation and type-1 error estimates were compared for different use of working correlation matrices for mTMAT$_{IM}$. Mean value of relative abundance and proportion of zeros count samples were estimated from KARE cohort study for all the genera and the genera with first quantile, median and third quantile sparsity level were chosen for the simulation. The value was 52%, 64% and 73%.

I also evaluate the robustness of the proposed method against the compositional bias. KARE dataset was simulated 2000 times with simulation parameters N and the ratio of cases and controls equal to 50 and 1:3, respectively. Then a genus containing more than one OTU is chosen and assumed to be associated with phenotype with $\beta = 0.15$ and p =50%. Then an OTU that is not contained in the chosen genus was selected and set to be associated to phenotype with the same β. Then, the abundance of the selected OTU that is not in the chosen genus was added by its standard deviation multiplied by the multiplier 0, 1, 5, 10, 50 and 100. Then the power estimate of the chosen genus was compared with different value of the multiplier.

Pregnant microbiome data

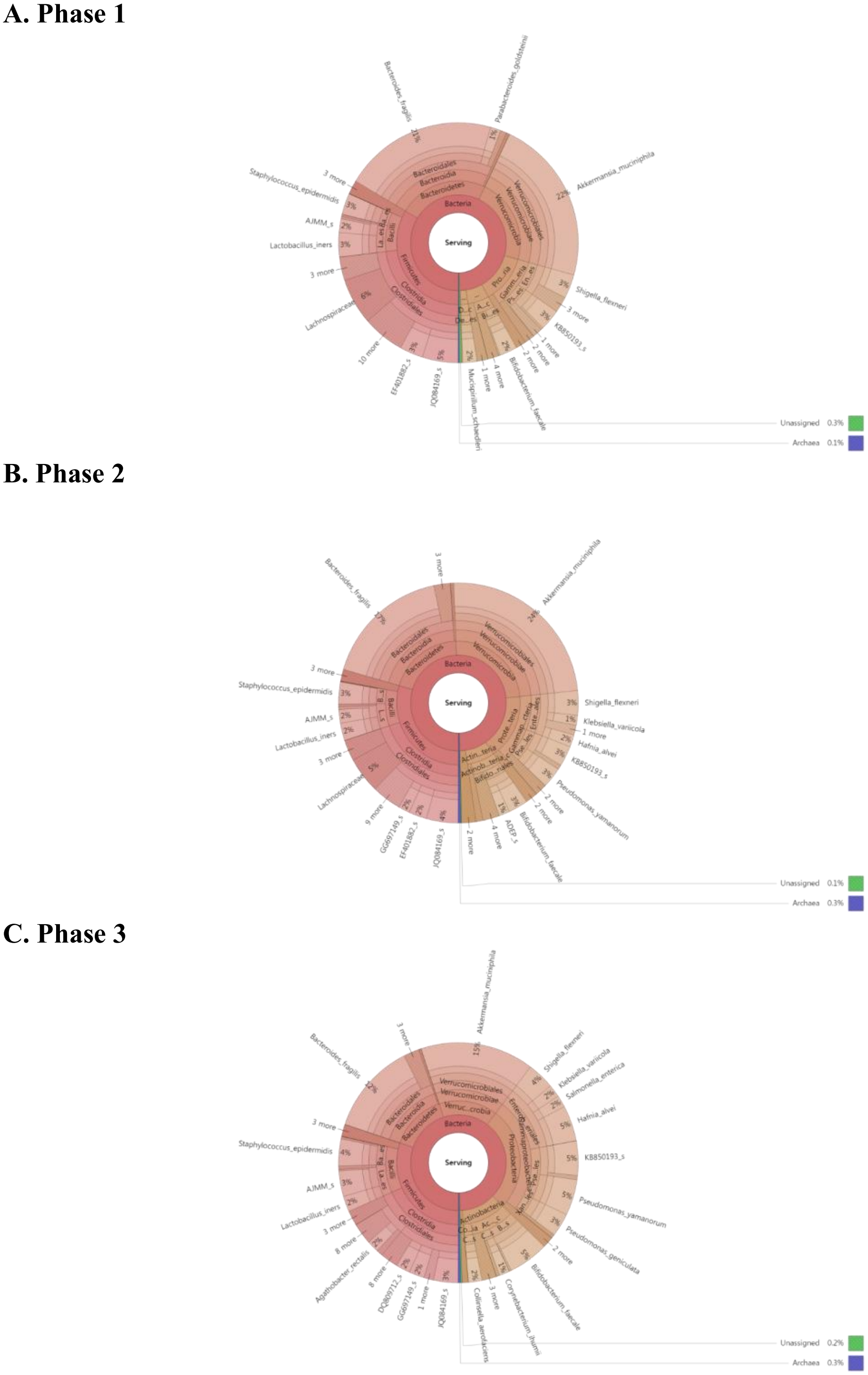
I used a publicly available datasets from Romero [72]. It is a retrospective case–control longitudinal study was designed and included non-pregnant women (n = 32) and pregnant women who delivered at term (38 to 42 weeks) without complications (n = 22) using pplacer and version 0.2 of the vaginal community 16S rRNA gene reference tree [73] for the taxonomically classification and phylogenetic tree [72]. The pregnant dataset includes the race, days after the first visit (GDColl), house hold income, maternal education, gender of baby.

3.3. Results

With simulated data, the performance of $\mathrm{mTMAT}_{\mathrm{M}}$ and $\mathrm{mTMAT}_{\mathrm{IM}}$ were evaluated. Figure 3.2 shows the overall distribution of microbial composition. Table 3.1 shows that inflation of type-1 error was observed when the number of case sample and total sample size increases. There seems no notable difference of type-1 error rates for the choice of working correlation matrix. On the other hand, $\mathrm{mTMAT}_{\mathrm{IM}}$ preserved nominal type-1 error well with a slight inflation when unstructured correlation and robust Wald statistic is used (Table 3.2). Robust Wald statistic tend to have higher type-1 error rates than robust score statistic no matter what correlation structure or type of mTMAT is used. A slight inflation is observed when 10% of samples were randomly excluded comparing to the complete dataset.

GLMM-MiRKAT, FZINBMM and LMM models are designed to be used longitudinal microbiome data and can be compared with $\mathrm{mTMAT}_{\mathrm{IM}}$ and $\mathrm{mTMAT}_{\mathrm{M}}$. FZINBMM and GLMM-MiRKAT could not preserve type-1 error rates with extremely high type-1 error estimates for FZINBMM. GLMM-MiRKAT suffered singular matrix problem during calculating the test statistics (Table 3.3). In this case, the resulting p-value were excluded for the estimation of type-1 error rate and power.

The type-1 error estimates for cross-sectional methods were also compared to consider the effect of correlation within subjects on type-1 error rates (Table 3.3). When case is three times smaller than control, type-1 errors are well preserved for all the cross-sectional methods. However, inflation was observed when the sample size is large and the number of cases is the same as the control group.



A. Phase 1

B. Phase 2

C. Phase 3

Figure 3.2. Taxonomic composition. Krona plots for phases 1, 2, and 3 showing the mean relative abundances of bacterial taxa at different taxonomic level

Table 3.1. Type-1 error estimates of mTMAT$_M$ with genera from longitudinal dataset. The values 1:1 and 1:3 were assumed for the ratio of cases and controls. The total sample size is denoted by N, and I considered N = 30, 50, and 100. All subjects were selected without replacement. Type-1 error estimates were calculated with 2,000 replicates at the significance level 0.1, 0.05, 0.01 and 0.005.

			Balanced data (Missing rate = 0%)						Unbalanced data (Missing rate = 10%)					
	Methods	Working Correlation	Case : Control = 1 : 1			Case : Control = 1 : 3			Case : Control = 1 : 1			Case : Control = 1 : 3		
			N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100
α = 0.1	Robust Score	Identity	0.1061	0.1276	0.1621	0.1016	0.1181	0.1087	0.0993	0.1311	0.1473	0.1039	0.0984	0.0955
		CS	0.1034	0.1210	0.1518	0.1036	0.1159	0.1073	0.0945	0.1266	0.1409	0.1041	0.0970	0.0934
		AR1	0.1058	0.1277	0.1602	0.1009	0.1157	0.1082	0.0964	0.1314	0.1495	0.1045	0.0977	0.0948
		Unstructured	0.1007	0.1196	0.1443	0.1009	0.1154	0.1096	0.0968	0.1236	0.1375	0.1068	0.0977	0.0970
	Robust Wald	Identity	0.1106	0.1297	0.1634	0.1088	0.1232	0.1105	0.1318	0.1396	0.1496	0.1445	0.1347	0.1105
		CS	0.1159	0.1284	0.1563	0.1168	0.1261	0.1107	0.1426	0.1452	0.1493	0.1588	0.1410	0.1116
		AR1	0.1168	0.1356	0.1649	0.1159	0.1257	0.1121	0.1422	0.1478	0.1522	0.1538	0.1390	0.1119
		Unstructured	0.1291	0.1393	0.1567	0.1318	0.1346	0.1175	0.1722	0.1586	0.1566	0.1832	0.1552	0.1188
α = 0.05	Robust Score	Identity	0.0493	0.0662	0.0940	0.0540	0.0617	0.0587	0.0448	0.0652	0.0793	0.0552	0.0482	0.0439
		CS	0.0487	0.0644	0.0859	0.0535	0.0579	0.0587	0.0452	0.0655	0.0739	0.0552	0.0505	0.0443
		AR1	0.0490	0.0689	0.0939	0.0527	0.0590	0.0587	0.0448	0.0675	0.0795	0.0557	0.0493	0.0439
		Unstructured	0.0479	0.0613	0.0809	0.0530	0.0576	0.0562	0.0491	0.0655	0.0702	0.0577	0.0507	0.0452
	Robust Wald	Identity	0.0587	0.0722	0.0968	0.0656	0.0731	0.0628	0.0733	0.0812	0.0807	0.0904	0.0789	0.0596
		CS	0.0624	0.0732	0.0913	0.0700	0.0744	0.0624	0.0829	0.0855	0.0829	0.1010	0.0834	0.0615
		AR1	0.0623	0.0764	0.0987	0.0696	0.0747	0.0642	0.0804	0.0867	0.0856	0.1015	0.0837	0.0614
		Unstructured	0.0742	0.0787	0.0914	0.0807	0.0806	0.0642	0.1060	0.0953	0.0871	0.1248	0.0962	0.0662
α = 0.01	Robust Score	Identity	0.0085	0.0151	0.0266	0.0157	0.0166	0.0118	0.0084	0.0143	0.0180	0.0141	0.0102	0.0093
		CS	0.0084	0.0137	0.0222	0.0146	0.0153	0.0119	0.0080	0.0127	0.0145	0.0145	0.0107	0.0089
		AR1	0.0081	0.0156	0.0259	0.0143	0.0150	0.0117	0.0080	0.0139	0.0175	0.0132	0.0109	0.0091
		Unstructured	0.0084	0.0128	0.0199	0.0135	0.0151	0.0114	0.0084	0.0123	0.0141	0.0120	0.0100	0.0080
	Robust Wald	Identity	0.0158	0.0209	0.0310	0.0275	0.0247	0.0153	0.0225	0.0234	0.0218	0.0427	0.0282	0.0153
		CS	0.0179	0.0212	0.0273	0.0297	0.0255	0.0164	0.0252	0.0253	0.0204	0.0460	0.0311	0.0155
		AR1	0.0184	0.0223	0.0307	0.0288	0.0253	0.0160	0.0251	0.0249	0.0227	0.0462	0.0300	0.0159
		Unstructured	0.0230	0.0232	0.0259	0.0345	0.0284	0.0179	0.0370	0.0297	0.0207	0.0575	0.0351	0.0160
α = 0.005	Robust Score	Identity	0.0040	0.0080	0.0156	0.0080	0.0092	0.0064	0.0027	0.0075	0.0082	0.0055	0.0050	0.0043
		CS	0.0041	0.0074	0.0130	0.0074	0.0082	0.0064	0.0039	0.0064	0.0066	0.0052	0.0057	0.0039
		AR1	0.0043	0.0081	0.0150	0.0076	0.0084	0.0065	0.0032	0.0061	0.0077	0.0050	0.0055	0.0041
		Unstructured	0.0037	0.0072	0.0112	0.0070	0.0082	0.0061	0.0030	0.0052	0.0070	0.0043	0.0055	0.0039
	Robust Wald	Identity	0.0099	0.0126	0.0195	0.0210	0.0190	0.0085	0.0144	0.0153	0.0133	0.0321	0.0193	0.0092
		CS	0.0110	0.0128	0.0167	0.0233	0.0176	0.0094	0.0171	0.0153	0.0125	0.0347	0.0211	0.0100
		AR1	0.0104	0.0133	0.0193	0.0219	0.0181	0.0085	0.0152	0.0160	0.0130	0.0325	0.0210	0.0092
		Unstructured	0.0150	0.0139	0.0161	0.0263	0.0203	0.0102	0.0247	0.0184	0.0133	0.0430	0.0245	0.0092

Table 3.2. Type-1 error estimates of mTMAT_{IM} with genera from longitudinal dataset. The values 1:1 and 1:3 were assumed for the ratio of cases and controls. The total sample size is denoted by N, and I considered N = 30, 50, and 100. All subjects were selected without replacement. Type-1 error estimates were calculated with 2,000 replicates at the significance level 0.1, 0.05, 0.01 and 0.005.

			Balanced data (Missing rate = 0%)						Unbalanced data (Missing rate = 10%)					
	Methods	Working Correlation	Case : Control = 1 : 1			Case : Control = 1 : 3			Case : Control = 1 : 1			Case : Control = 1 : 3		
			N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100
α = 0.1	Robust Score	Identity	0.0884	0.0963	0.0933	0.0999	0.0921	0.1041	0.0914	0.1245	0.1409	0.0945	0.1023	0.0961
		CS	0.0882	0.0957	0.0915	0.1004	0.0938	0.1039	0.0920	0.1214	0.1375	0.0952	0.1000	0.0980
		AR1	0.0893	0.0959	0.0923	0.0990	0.0943	0.1036	0.0923	0.1243	0.1414	0.0957	0.0975	0.0959
		Unstructured	0.0901	0.0954	0.0905	0.1003	0.0941	0.1039	0.0925	0.1198	0.1325	0.0955	0.0995	0.0957
	Robust Wald	Identity	0.0933	0.0988	0.0945	0.1066	0.0959	0.1064	0.1226	0.1362	0.1400	0.1377	0.1362	0.1079
		CS	0.1014	0.1031	0.0959	0.1146	0.1016	0.1086	0.1338	0.1426	0.1407	0.1512	0.1421	0.1123
		AR1	0.1004	0.1022	0.0961	0.1146	0.1026	0.1090	0.1334	0.1442	0.1442	0.1497	0.1378	0.1105
		Unstructured	0.1177	0.1113	0.0990	0.1335	0.1131	0.1144	0.1626	0.1578	0.1475	0.1796	0.1592	0.1156
α = 0.05	Robust Score	Identity	0.0413	0.0493	0.0451	0.0463	0.0458	0.0495	0.0391	0.0618	0.0741	0.0473	0.0455	0.0427
		CS	0.0421	0.0503	0.0454	0.0446	0.0451	0.0498	0.0409	0.0593	0.0698	0.0436	0.0445	0.0443
		AR1	0.0413	0.0499	0.0454	0.0460	0.0449	0.0498	0.0398	0.0627	0.0732	0.0459	0.0436	0.0416
		Unstructured	0.0397	0.0478	0.0447	0.0476	0.0453	0.0499	0.0434	0.0595	0.0655	0.0457	0.0475	0.0443
	Robust Wald	Identity	0.0484	0.0533	0.0470	0.0578	0.0535	0.0536	0.0685	0.0770	0.0784	0.0832	0.0753	0.0562
		CS	0.0541	0.0576	0.0484	0.0626	0.0551	0.0555	0.0762	0.0823	0.0801	0.0916	0.0811	0.0597
		AR1	0.0539	0.0566	0.0482	0.0616	0.0560	0.0556	0.0773	0.0827	0.0815	0.0910	0.0805	0.0579
		Unstructured	0.0653	0.0610	0.0508	0.0769	0.0624	0.0602	0.1016	0.0948	0.0814	0.1160	0.0932	0.0653
α = 0.01	Robust Score	Identity	0.0057	0.0096	0.0092	0.0057	0.0061	0.0094	0.0084	0.0123	0.0141	0.0120	0.0100	0.0080
		CS	0.0065	0.0091	0.0091	0.0056	0.0063	0.0090	0.0066	0.0127	0.0159	0.0077	0.0075	0.0055
		AR1	0.0065	0.0097	0.0093	0.0054	0.0064	0.0088	0.0070	0.0107	0.0127	0.0068	0.0070	0.0055
		Unstructured	0.0060	0.0095	0.0090	0.0064	0.0066	0.0086	0.0073	0.0125	0.0161	0.0068	0.0086	0.0059
	Robust Wald	Identity	0.0119	0.0138	0.0108	0.0169	0.0131	0.0121	0.0189	0.0223	0.0193	0.0273	0.0222	0.0137
		CS	0.0138	0.0147	0.0112	0.0198	0.0141	0.0127	0.0219	0.0247	0.0181	0.0319	0.0240	0.0125
		AR1	0.0136	0.0150	0.0113	0.0197	0.0144	0.0126	0.0216	0.0241	0.0199	0.0319	0.0241	0.0136
		Unstructured	0.0184	0.0167	0.0124	0.0267	0.0177	0.0136	0.0359	0.0301	0.0199	0.0445	0.0311	0.0136
α = 0.005	Robust Score	Identity	0.0024	0.0040	0.0044	0.0022	0.0019	0.0037	0.0030	0.0052	0.0070	0.0043	0.0055	0.0039
		CS	0.0024	0.0035	0.0044	0.0024	0.0024	0.0037	0.0027	0.0048	0.0077	0.0034	0.0027	0.0030
		AR1	0.0024	0.0039	0.0046	0.0025	0.0024	0.0039	0.0027	0.0041	0.0075	0.0030	0.0030	0.0027
		Unstructured	0.0021	0.0038	0.0048	0.0025	0.0030	0.0037	0.0030	0.0050	0.0082	0.0027	0.0025	0.0032
	Robust Wald	Identity	0.0064	0.0074	0.0059	0.0094	0.0073	0.0064	0.0127	0.0142	0.0096	0.0189	0.0147	0.0067
		CS	0.0078	0.0083	0.0061	0.0113	0.0081	0.0066	0.0156	0.0153	0.0108	0.0221	0.0151	0.0066
		AR1	0.0082	0.0087	0.0061	0.0104	0.0079	0.0068	0.0155	0.0158	0.0112	0.0211	0.0145	0.0071
		Unstructured	0.0110	0.0100	0.0067	0.0160	0.0107	0.0073	0.0251	0.0200	0.0123	0.0314	0.0201	0.0073

Table 3.3. Type-1 error estimates of the methods for comparison with genera. The values 1:1 and 1:3 were assumed for the ratio of cases and controls. The total sample size is denoted by N, and I considered N = 30, 50 and 100. All subjects were selected without replacement. Type-1 error estimates were calculated with 2,000 replicates at the significance level 0.1, 0.05, 0.01 and 0.005. For unbalanced data, I applied 10% missing rate to phenotype for each time points.

			Balanced data (Missing rate = 0%)						Unbalanced data (Missing rate = 10%)					
	Methods	Working Correlation	Case : Control = 1 : 1			Case : Control = 1 : 3			Case : Control = 1 : 1			Case : Control = 1 : 3		
			N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100	N = 30	N = 50	N = 100
α = 0.1	Longitudinal	GLMM-MiRKAT	0.1320	0.1377	0.1301	0.1450	0.1466	0.1347	0.1383	0.1770	0.1815	0.1380	0.1495	0.1392
		FZINBMM	0.4882	0.4548	0.4770	0.4778	0.4625	0.4411	0.4580	0.4405	0.4616	0.4311	0.4207	0.4282
		LMM-arcsin	0.1094	0.1328	0.1482	0.0926	0.0982	0.0994	0.1074	0.1318	0.1502	0.0887	0.0994	0.1060
		LMM-log	0.1021	0.1115	0.1223	0.0922	0.0934	0.0905	0.1032	0.1118	0.1275	0.0918	0.0944	0.0991
	Cross-sectional	TMAT$_{IM}$	0.0988	0.1085	0.1235	0.1342	0.0927	0.0929	0.1030	0.1314	0.1466	0.0907	0.0957	0.0900
		TMAT$_{M}$	0.0978	0.1132	0.1289	0.1336	0.1042	0.0944	0.0961	0.1252	0.1466	0.1002	0.1016	0.0977
		Wilcoxon	0.1034	0.1174	0.1337	0.1320	0.0947	0.0938	0.0857	0.0978	0.0948	0.0926	0.0852	0.0800
		oMiRKAT	0.1041	0.1141	0.1298	0.1308	0.0950	0.0945	0.0955	0.1160	0.1285	0.1210	0.1035	0.0970
		aMiSPU	0.0839	0.0973	0.1108	0.1056	0.0800	0.0691	0.1009	0.1393	0.1611	0.0952	0.1045	0.1009
α = 0.05	Longitudinal	GLMM-MiRKAT	0.0914	0.0889	0.0875	0.0922	0.0974	0.0846	0.0958	0.1273	0.1286	0.0914	0.0988	0.0989
		FZINBMM	0.4209	0.3915	0.4162	0.4055	0.3908	0.3786	0.3866	0.3732	0.3934	0.3670	0.3516	0.3605
		LMM-arcsin	0.0537	0.0681	0.0804	0.0449	0.0497	0.0467	0.0585	0.0702	0.0774	0.0409	0.0492	0.0510
		LMM-log	0.0466	0.0529	0.0613	0.0421	0.0475	0.0436	0.0518	0.0566	0.0681	0.0414	0.0496	0.0521
	Cross-sectional	TMAT$_{IM}$	0.0489	0.0580	0.0660	0.0613	0.0448	0.0456	0.0516	0.0670	0.0780	0.0436	0.0493	0.0434
		TMAT$_{M}$	0.0463	0.0564	0.0642	0.0660	0.0505	0.0448	0.0468	0.0648	0.0814	0.0491	0.0482	0.0473
		Wilcoxon	0.0496	0.0558	0.0635	0.0631	0.0488	0.0449	0.0413	0.0543	0.0591	0.0517	0.0470	0.0426
		oMiRKAT	0.0509	0.0514	0.0585	0.0678	0.0527	0.0436	0.0515	0.0565	0.0705	0.0620	0.0500	0.0480
		aMiSPU	0.0435	0.0471	0.0536	0.0568	0.0426	0.0335	0.0509	0.0725	0.0893	0.0425	0.0514	0.0505
α = 0.01	Longitudinal	GLMM-MiRKAT	0.0539	0.0509	0.0491	0.0546	0.0607	0.0544	0.0684	0.0720	0.0750	0.0628	0.0665	0.0657
		FZINBMM	0.3182	0.2961	0.3162	0.2990	0.2990	0.2807	0.2902	0.2620	0.2855	0.2677	0.2495	0.2577
		LMM-arcsin	0.0115	0.0126	0.0221	0.0099	0.0109	0.0100	0.0125	0.0146	0.0231	0.0082	0.0089	0.0100
		LMM-log	0.0085	0.0100	0.0150	0.0104	0.0127	0.0113	0.0089	0.0109	0.0150	0.0082	0.0102	0.0116
	Cross-sectional	TMAT$_{IM}$	0.0085	0.0088	0.0100	0.0139	0.0073	0.0100	0.0100	0.0143	0.0177	0.0084	0.0105	0.0086
		TMAT$_{M}$	0.0077	0.0094	0.0107	0.0132	0.0097	0.0105	0.0105	0.0150	0.0173	0.0091	0.0116	0.0084
		Wilcoxon	0.0084	0.0099	0.0113	0.0130	0.0084	0.0078	0.0117	0.0152	0.0113	0.0143	0.0109	0.0100
		oMiRKAT	0.0082	0.0073	0.0083	0.0114	0.0077	0.0114	0.0125	0.0145	0.0180	0.0160	0.0140	0.0105
		aMiSPU	0.0057	0.0102	0.0116	0.0161	0.0070	0.0061	0.0109	0.0161	0.0241	0.0100	0.0091	0.0098
α = 0.005	Longitudinal	GLMM-MiRKAT	0.0504	0.0477	0.0446	0.0496	0.0552	0.0492	0.0661	0.0696	0.0724	0.0628	0.0665	0.0657
		FZINBMM	0.2864	0.2657	0.2812	0.2659	0.2703	0.2510	0.2593	0.2282	0.2525	0.2320	0.2207	0.2273
		LMM-arcsin	0.0051	0.0071	0.0121	0.0048	0.0048	0.0054	0.0057	0.0071	0.0128	0.0034	0.0046	0.0055
		LMM-log	0.0033	0.0044	0.0086	0.0051	0.0067	0.0063	0.0034	0.0039	0.0089	0.0039	0.0059	0.0055
	Cross-sectional	TMAT$_{IM}$	0.0047	0.0047	0.0054	0.0072	0.0042	0.0051	0.0030	0.0070	0.0109	0.0034	0.0052	0.0045
		TMAT$_{M}$	0.0041	0.0045	0.0052	0.0076	0.0058	0.0052	0.0061	0.0077	0.0086	0.0050	0.0055	0.0034
		Wilcoxon	0.0044	0.0035	0.0040	0.0071	0.0034	0.0045	0.0117	0.0148	0.0113	0.0139	0.0109	0.0100
		oMiRKAT	0.0045	0.0034	0.0038	0.0048	0.0045	0.0064	0.0035	0.0050	0.0060	0.0045	0.0050	0.0035
		aMiSPU	0.0026	0.0037	0.0043	0.0098	0.0017	0.0043	0.0048	0.0082	0.0148	0.0041	0.0055	0.0061

Sensitivity analysis of type-1 error rates was conducted to consider the effect of a violation of statistical assumption, the statistical characteristic of the abundance and phylogenetic tree for each genera. This includes investigating the effect of the number of leaf node, sparsity level, mean relative abundance level and assumed correlation matrix on type 1 error rates.

The effect of the number of leaf nodes (Table 3.4), results showed that mTMAT$_{M}$ became slightly conservative if the number of leaf nodes was larger than 5 but that mTMAT$_{IM}$ was less affected. The result with more than 15 number of leaf nodes can be dependent on specific genera chosen with small number of genera.

Table 3.5 shows the effect of sparsity on the type-1 error rate. For each genus, I calculated its sparsity, defined as the proportion of subjects with no abundance, and type-1 error rates were calculated. Results showed that the type-1 error rates of FZINBMM were the most inflated and that some inflation was observed for GLMM-MiRKAT when the mean sparsity is greater than 20%. GLMM-MiRKAT is based on the permutation, and the permutation-based p-value is generally robust to the non-normality. However, if there exists heteroscedasticity, its statistical validity can be impaired. A substantial amount of sparsity may induce the heteroscedasticity, which may explain the type-1 error inflation. Some inflation was also observed for TMAT$_{M}$, but rates for TMAT$_{IM}$ were preserved well.

The effect of mean relative abundance on the type-1 error rate result showed that mTMAT$_{IM}$, GLMM-MiRKAT, LMM-arcsin and LMM-log preserved type-1 error rates for all the genera groups with different mean relative abundance. Inflation of FZINBMM were more severe in the group with mean relative abundance less than 10% quantile for all genera (Table 3.6).

The effect of assumed correlation matrix for different scenarios was evaluated with the use of microbiomeDASim [71]. Identity, CS and AR1 with different value of parameter is assumed for the simulation with different use of working correlation matrix, robust score statistic and for $\mathrm{mTMAT}_{\mathrm{IM}}$ (Table 3.7). The result shows that $\mathrm{mTMAT}_{\mathrm{IM}}$ preserved type-1 error for the most of the scenarios. Slight inflation was observed for the statistic using identity working correlation was used when the assumed correlation is CS or AR1.

Table 3.4. Effect of numbers of leaf nodes on type-1 error estimates. Families were categorized into four different groups according to the number of leaf nodes, and for each taxon, type-1 error rates were estimated. Simulation data were generated with read counts from dataset. I assumed the total sample size (N) is equal to 50. The values 1:3 was assumed for the ratio of cases and controls. Identity working correlation matrix and robust score statistic are used for mTMAT.

Method	Number of leaf nodes	Number of Family	Significance level			
			α = 0.1	α = 0.05	α = 0.01	α = 0.005
mTMAT_{IM}	1	22	0.1050	0.0508	0.0102	0.0047
	2-5	12	0.0917	0.0411	0.0042	0.0014
	6-15	5	0.0733	0.0211	0.0000	0.0000
	>15	2	0.0700	0.0267	0.0017	0.0017
mTMAT_{M}	1	22	0.1147	0.0641	0.0182	0.0112
	2-5	12	0.1014	0.0442	0.0078	0.0039
	6-15	5	0.0833	0.0433	0.0044	0.0022
	>15	2	0.0700	0.0267	0.0033	0.0017
GLMM-MiRKAT	1	22	NA	NA	NA	NA
	2-5	12	0.0972	0.0519	0.0108	0.0047
	6-15	5	0.1011	0.0467	0.0100	0.0044
	>15	2	0.0850	0.0417	0.0033	0.0000
FZINBMM	1	22	0.5683	0.5092	0.4035	0.3717
	2-5	12	0.2806	0.2147	0.1228	0.0997
	6-15	5	0.1133	0.0633	0.0167	0.0111
	>15	2	0.1000	0.0333	0.0033	0.0000
LMM-arcsine	1	22	0.0549	0.0315	0.0067	0.0042
	2-5	12	0.0947	0.0450	0.0101	0.0054
	6-15	5	0.1211	0.0589	0.0133	0.0034
	>15	2	0.1633	0.1017	0.0267	0.0133
LMM-log	1	22	0.0920	0.0414	0.0118	0.0059
	2-5	12	0.991	0.0509	0.0095	0.0048
	6-15	5	0.0971	0.0426	0.0123	0.0045
	>15	2	0.1117	0.0687	0.0067	0.0050

Table 3.5. Effect of sparsity on type-1 error estimates. For each genus, I calculated its sparsity as the proportion of subjects with no abundance. Genera were sorted by their sparsity and categorized into three different groups, and for each taxon, type-1 error rates were estimated. Simulation data were generated by using read counts from dataset. I assumed the total sample size (N) is equal to 50. The values 1:3 was assumed for the ratio of cases and controls. Identity working correlation matrix and robust score statistic are used for mTMAT.

Method	Mean sparsity level of genera	Mean number of leaf nodes	Number of genus	Significance level			
				α = 0.1	α = 0.05	α = 0.01	α = 0.005
mTMAT$_{IM}$	<=20%	1	3	0.0989	0.0522	0.0056	0.0022
	20-50%	2.25	12	0.0883	0.0417	0.0061	0.0019
	>50%	1.66	58	0.1032	0.0496	0.0090	0.0039
mTMAT$_{M}$	<=20%	1	3	0.1200	0.0600	0.0111	0.0067
	20-50%	2.25	12	0.0989	0.0447	0.0072	0.0042
	>50%	1.66	58	0.1180	0.0620	0.0152	0.0093
GLMM-MiRKAT	<=20%	1	3	NA	NA	NA	NA
	20-50%	2.25	12	0.1225	0.0625	0.0100	0.0025
	>50%	1.66	58	0.1526	0.1047	0.0611	0.0584
FZINBMM	<=20%	1	3	0.2067	0.1333	0.0500	0.0367
	20-50%	2.25	12	0.2600	0.1900	0.1083	0.0858
	>50%	1.66	58	0.4961	0.4351	0.3297	0.2964
LMM-arcsine	<=20%	1	3	0.1013	0.0491	0.0045	0.0067
	20-50%	2.25	12	0.0973	0.0482	0.0132	0.0047
	>50%	1.66	58	0.0828	0.0381	0.0084	0.0051
LMM-log	<=20%	1	3	0.0956	0.0501	0.0043	0.0032
	20-50%	2.25	12	0.0962	0.0434	0.0078	0.0041
	>50%	1.66	58	0.0972	0.0499	0.0112	0.0063

Table 3.6. Effect of mean relative abundance on type-1 error estimates. Genera were categorized into four different groups according to the quantile of mean relative abundance, and for each taxon, type-1 error rates were estimated. $Q_{0.1}$, $Q_{0.5,}$ $Q_{0.9}$ represents 10%, 50%, 90% quantile $Q_{0.1} = 1.19\times10^{-3}$, $Q_{0.5} = 2.40\times10^{-3}$ and $Q_{0.9} = 7.41\times10^{-3}$) Simulation data were generated with read counts from dataset. I assumed the total sample size (N) is equal to 50. The values 1:3 was assumed for the ratio of cases and controls. Identity working correlation matrix and robust score statistic are used for mTMAT.

Method	Mean Relative Abundance	Mean number of leaf nodes	Significance level			
			α = 0.1	α = 0.05	α = 0.01	α = 0.005
mTMAT_{IM}	<=$Q_{0.1}$	1	0.1060	0.0480	0.0087	0.0040
	$Q_{0.1}$- $Q_{0.5}$	1.31	0.0879	0.0427	0.0075	0.0035
	$Q_{0.5}$- $Q_{0.9}$	2.35	0.0930	0.0418	0.0078	0.0042
	> $Q_{0.9}$	1.75	0.1011	0.0511	0.0056	0.0022
mTMAT_{M}	<=$Q_{0.1}$	1	0.0913	0.0453	0.0093	0.0047
	$Q_{0.1}$- $Q_{0.5}$	1.31	0.0902	0.0454	0.0083	0.0040
	$Q_{0.5}$- $Q_{0.9}$	2.35	0.1007	0.0474	0.0093	0.0045
	> $Q_{0.9}$	1.75	0.1145	0.0522	0.0089	0.0033
GLMM-MiRKAT	<=$Q_{0.1}$	1	0.1110	0.0670	0.0100	0.0060
	$Q_{0.1}$- $Q_{0.5}$	1.31	0.0843	0.0386	0.0086	0.0043
	$Q_{0.5}$- $Q_{0.9}$	2.35	0.1192	0.0667	0.0305	0.0266
	> $Q_{0.9}$	1.75	0.0940	0.0494	0.0090	0.0047
FZINBMM	<=$Q_{0.1}$	1	0.4050	0.3640	0.2910	0.2680
	$Q_{0.1}$- $Q_{0.5}$	1.31	0.2721	0.2174	0.1497	0.1303
	$Q_{0.5}$- $Q_{0.9}$	2.35	0.1616	0.1027	0.0469	0.0378
	> $Q_{0.9}$	1.75	0.2000	0.1514	0.0843	0.0800
LMM-arcsine	<=$Q_{0.1}$	1	0.0675	0.0338	0.0163	0.0050
	$Q_{0.1}$- $Q_{0.5}$	1.31	0.0890	0.0414	0.0093	0.0062
	$Q_{0.5}$- $Q_{0.9}$	2.35	0.0955	0.0485	0.0126	0.0072
	> $Q_{0.9}$	1.75	0.1018	0.0441	0.0088	0.0050
LMM-log	<=$Q_{0.1}$	1	0.0816	0.0415	0.0126	0.0063
	$Q_{0.1}$- $Q_{0.5}$	1.31	0.0937	0.0432	0.0090	0.0041
	$Q_{0.5}$- $Q_{0.9}$	2.35	0.1080	0.0524	0.0111	0.0068
	> $Q_{0.9}$	1.75	0.1075	0.0450	0.0075	0.0038

Table 3.7. Effect of assumed correlation structure on type-1 error estimates. For each taxon, type-1 error rates were estimated. Simulation data were generated by using read counts from simulation dataset with microbiomeDASim package. I assumed the total sample size (N) is equal to 50. Identity working correlation matrix and robust score statistic are used for $\mathrm{mTMAT}_{\mathrm{IM}}$. The number of time points was set to be 6.

Assumed Correlation Structure	Assumed rho	Working Correlation Structure	Significance level			
			α = 0.1	α = 0.05	α = 0.01	α = 0.005
Identity	0	Identity	0.1045	0.0473	0.0078	0.0033
		CS	0.1045	0.0503	0.0103	0.0035
		AR1	0.1035	0.0515	0.0100	0.0040
		Unstructured	0.1043	0.0515	0.0100	0.0040
CS	0.2	Identity	0.1020	0.0520	0.0095	0.0048
		CS	0.0998	0.0493	0.0085	0.0030
		AR1	0.0983	0.0488	0.0080	0.0030
		Unstructured	0.0995	0.0505	0.0070	0.0033
	0.5	Identity	0.1028	0.0455	0.0090	0.0053
		CS	0.0953	0.0463	0.0083	0.0043
		AR1	0.0970	0.0445	0.0078	0.0048
		Unstructured	0.0955	0.0453	0.0080	0.0045
	0.8	Identity	0.0988	0.0483	0.0095	0.0050
		CS	0.0980	0.0488	0.0073	0.0033
		AR1	0.0973	0.0493	0.0073	0.0033
		Unstructured	0.0958	0.0478	0.0075	0.0033
AR1	0.2	Identity	0.1013	0.0523	0.0123	0.0068
		CS	0.0983	0.0465	0.0073	0.0035
		AR1	0.0975	0.0463	0.0080	0.0035
		Unstructured	0.0958	0.0475	0.0070	0.0033
	0.5	Identity	0.1000	0.0510	0.0100	0.0045
		CS	0.0993	0.0505	0.0083	0.0035
		AR1	0.0993	0.0510	0.0090	0.0033
		Unstructured	0.1005	0.0480	0.0078	0.0033
	0.8	Identity	0.1015	0.0468	0.0060	0.0015
		CS	0.0913	0.0405	0.0083	0.0038
		AR1	0.0913	0.0395	0.0080	0.0035
		Unstructured	0.0923	0.0398	0.0065	0.0038

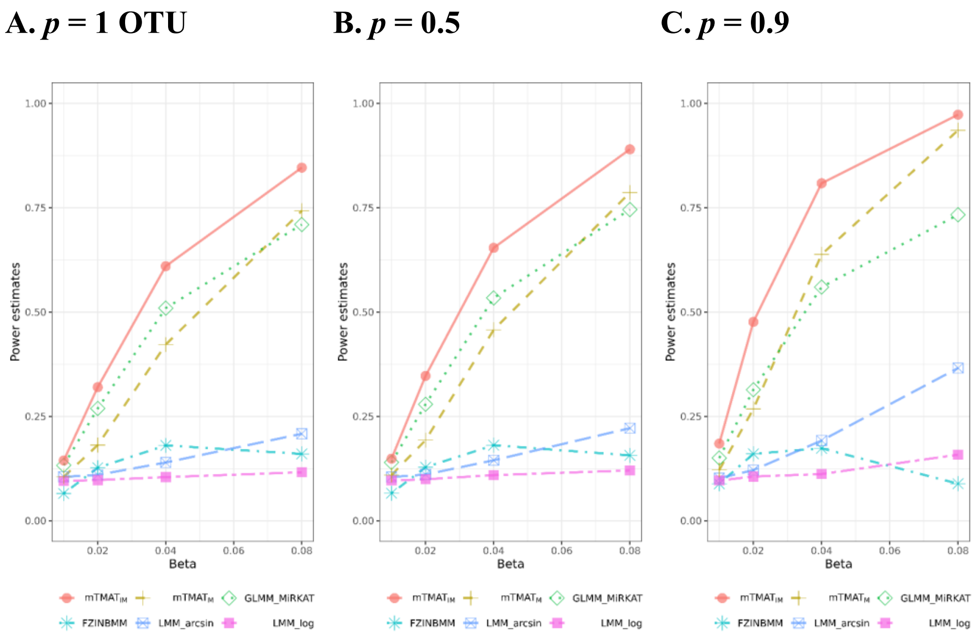
I also calculated statistical power estimates with 2,000 replicates at the 0.05 significance levels, and these were compared with those of other statistical methods. . The significance levels for each methods were adjusted based on the statistics from the simulation to calculate type-1 error to give a valid performance comparison. The threshold is determined as the percentiles of the p-values calculated in the type-1 error simulation under null hypothesis. I considered genera consisting of two or more OTUs. In Figure 3.3, $\mathrm{mTMAT}_{\mathrm{IM}}$ usually outperformed the other methods. The performance of GLMM-MiRKAT was comparable with $\mathrm{mTMAT}_{\mathrm{M}}$. FZINBMM and LMM-log had a much smaller power than other methods.

Figures 3.4 show the results when genera consisted of one or more OTUs. GLMM-MiRKAT can only be calculated if more than one OTU available. Thus, it was excluded from this comparison. $\mathrm{mTMAT}_{\mathrm{M}}$ and $\mathrm{mTMAT}_{\mathrm{IM}}$ can be applied in such scenarios, and the results showed that the proposed method was the most efficient.

The comparison with methods for cross-sectional analysis (Figure 3.5) shows that $\mathrm{TMAT}_{\mathrm{M}}$, TMATIM and $\mathrm{mTMAT}_{\mathrm{IM}}$ showed high statistical power. aMiSPU had highest power estimate when beta is 0.02.

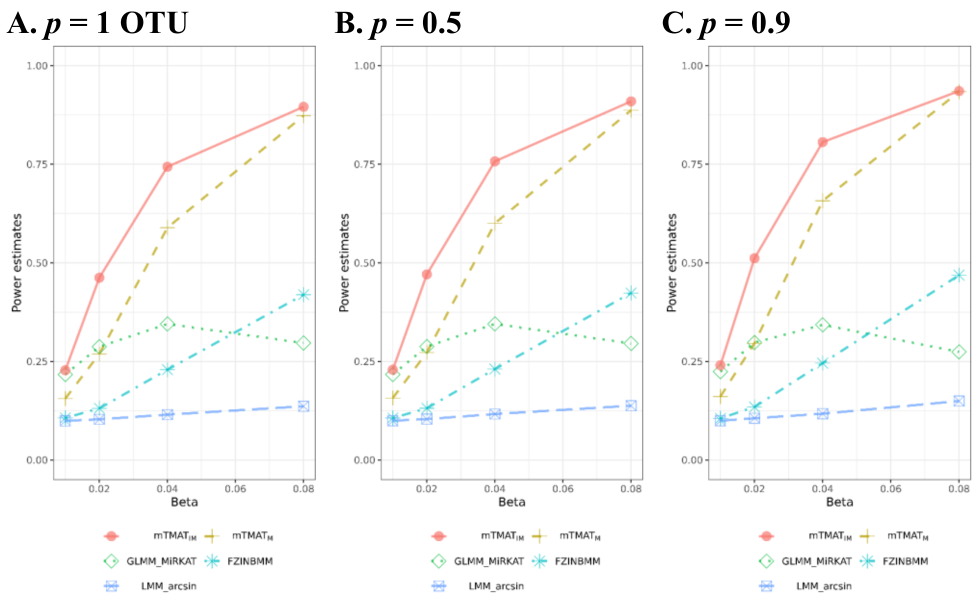
For the sensitivity analysis of power estimates, Figure 3.6 shows the statistical power estimates according to the number of leaf nodes. The number of causal OTUs was assumed to be the same, and statistical power estimates decreased according to the number of leaf nodes. The best performance was found for mTMATIM except that GLMM-MiRKAT had high level of type-1 error rate with the leaf node is between 6 and 15. I also evaluated the effect of sparsity on statistical power. For each genus, sparsity was defined by the proportion of subjects with no reads from that genus. As shown in Figure 3.6, $\mathrm{mTMAT}_{\mathrm{IM}}$ was always comparable with other methods that failed to preserve type-1 error rates. Sparsity level and the level of missing rate were also considered and statistical power estimates were compared. Power estimates were maximized in the middle-group sparsity level (Figure 3.7) and there was slight decrease of power estimate for overall methods when missing rate increases (Figure 3.8). In Figure 3.9, the effect of mean relative abundance on power estimate was evaluated. mTMAT$_{IM}$ and FZINBMM had the highest power estimate when mean relative abundance is lower than its 10% quantile. There were no consistent trend for the effect. Figure 3.10 shows the effect of compositional bias on the power estimate. For genera with more than one OTU, the power estimates of mTMAT$_{IM}$ and mTMAT$_{M}$ were not affected by compositional bias.

In summary, I confirmed that mTMAT$_{IM}$ is generally the most efficient among the available methods in the simulations. mTMAT$_{IM}$ considers phylogenetic tree structures, uses log CPM transformation, correction of compositional bias with taking a proportion among OTUs and consider correlations among repeatedly measured samples, which may lead to its superiority over other methods. Overall result for power comparison among the methods cross-sectional is consistent on previous paper for TMAT [52], but type-1 error rate for TMAT had inflated with correlated data. GLMM-MiRKAT is the second most powerful, but it failed to preserve type-1 rates and cannot be applied with analysis with single OTU. Furthermore, GLMM-MiRKAT is based on oMiRKAT and they both used kernel method and permutation approaches, which can be computationally very intensive especially if the number of sample size increases [52].



A. p **= 1 OTU** **B.** p **= 0.5** **C.** p **= 0.9**

Figure 3.3. Power estimates for genera consisting of more than one OTU. Power estimates at the significance level 0.05 were calculated with 500 replicates. I generated simulation data based on read counts from datasets, and considered genera with more than one OTU. The significance levels for each methods were adjusted based on the statistics from the simulation to calculate type-1 error to have the similar value of type-1 error. I assumed the total sample size (N) is equal to 50 and the ratio of cases and controls is set to be 1:3 at missing rate 10%. Identity working correlation matrix and robust score statistic are used for mTMAT.


A. *p* = 1 OTU

B. *p* = 0.5

C. *p* = 0.9

Figure 3.4. Power estimates for genera consisting of one or more OTUs. Power estimates at the significance level 0.05 were calculated with 500 replicates. I generated simulation data based on read counts from datasets, and results from GLMM-MIRKAT were excluded because they cannot be applied to genera consisting of a single OTU. The significance levels for each methods were adjusted to have the level of type-1 error rates based on the statistics from the simulation under null hypothesis. I assumed the total sample size (N) is equal to 50 and the ratio of cases and controls is set to be 1:3 at missing rate 10%. Identity working correlation matrix and robust score statistic are used for mTMAT.

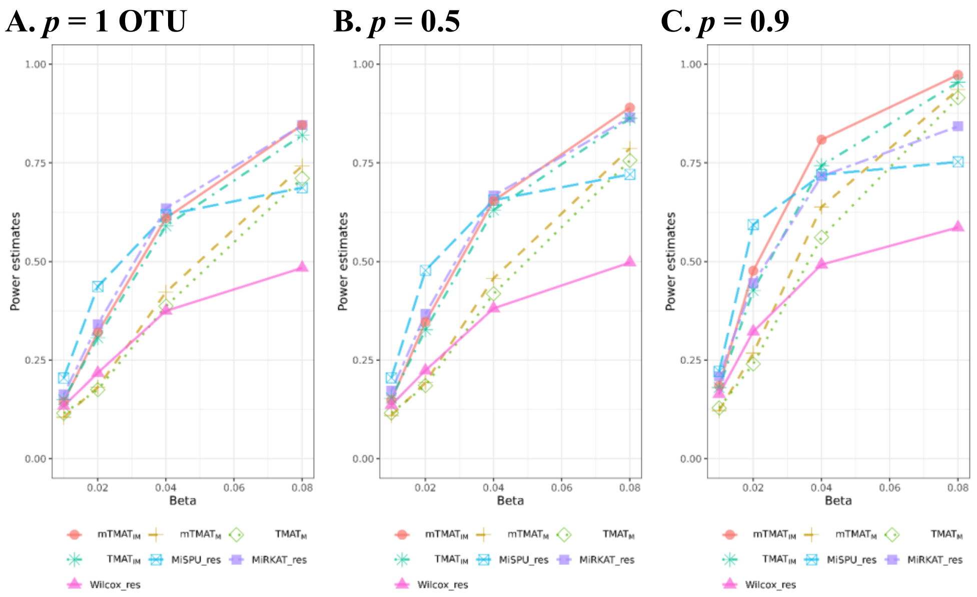


Figure 3.5. Power estimates comparison with the methods for cross-sectionally observed data. Power estimates at the significance level 0.05 were calculated with 500 replicates. I assumed the total sample size (N) is equal to 50 and the ratio of cases and controls is set to be 1:3 at missing rate 10%. Identity working correlation matrix and robust score statistic are used for mTMAT.

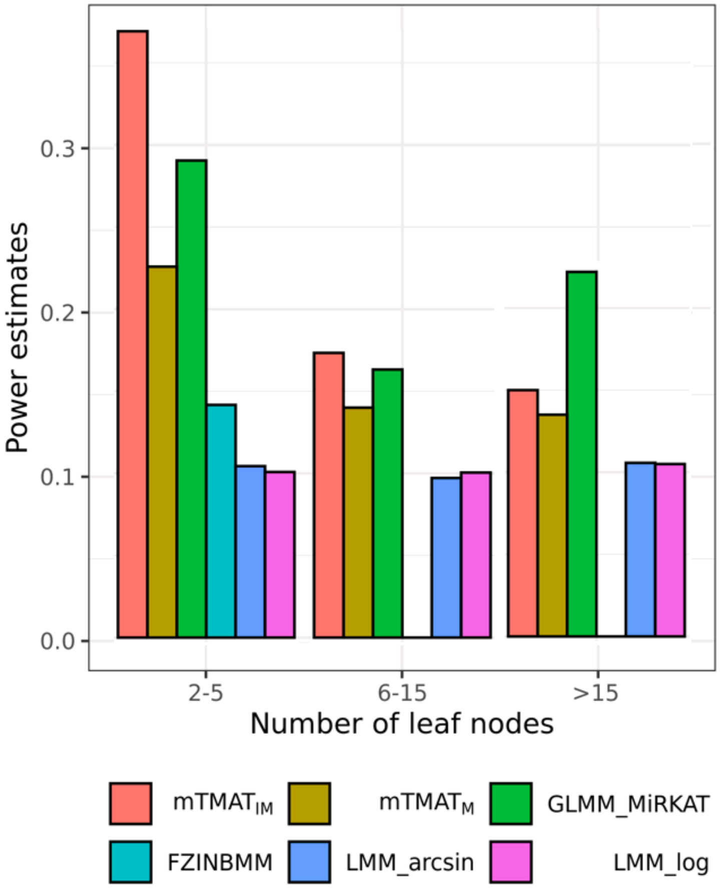


Figure 3.6. Effect of numbers of leaf nodes on power estimates. Families were categorized into four different groups according to the number of leaf nodes, and for each taxon, power estimates at the 0.05 significance level were calculated with 500 replicates. I generated simulation data based on read counts from datasets, and the results were combined. I considered families with more than one OTU. I assumed the total sample size (*N*) = 50 at missing rate 10%, p = 50%, $\boldsymbol{\beta}$ = 0.02 and the ratio of cases and controls is set to be 1:3 at missing rate 10%.

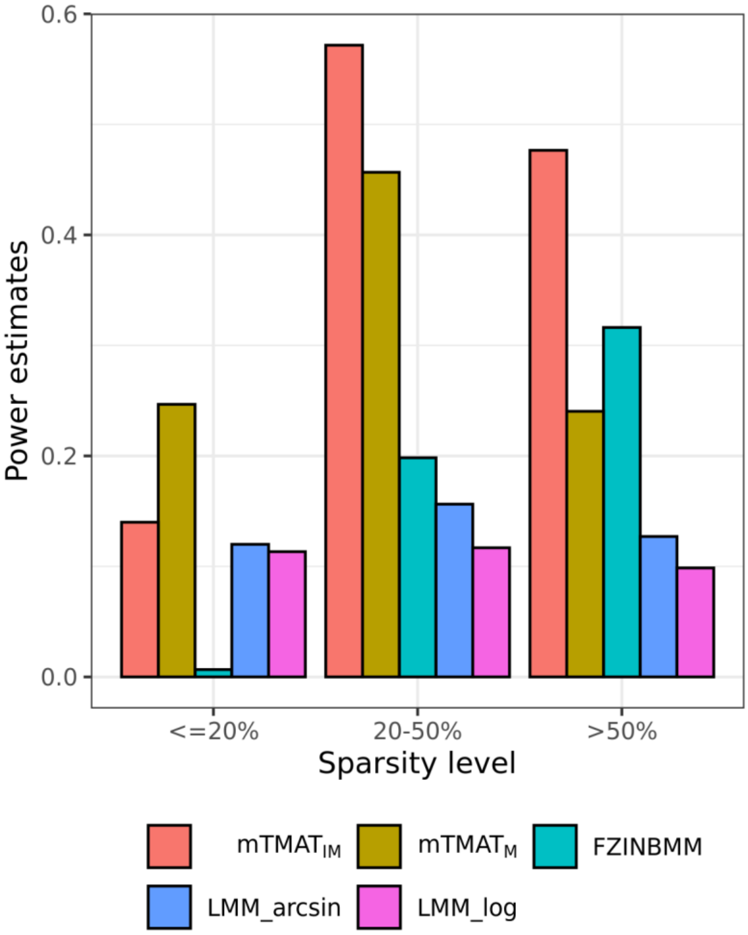


Figure 3.7. Effect of sparsity on power estimates. For each genus, I calculated its sparsity as the proportion of subjects with no reads (abundance of 0). Genera were sorted by their sparsity and categorized into three different groups, and for each taxon, power estimates at the 0.05 significance level were calculated with 500 replicates. I generated simulation data based on read counts from the dataset and considered genera with more than one OTU. I assumed the total sample size (N) = 50 at missing rate 10%, p = 50%, $\boldsymbol{\beta}$ = 0.02 and the ratio of cases and controls is set to be 1:3 at missing rate 10%.

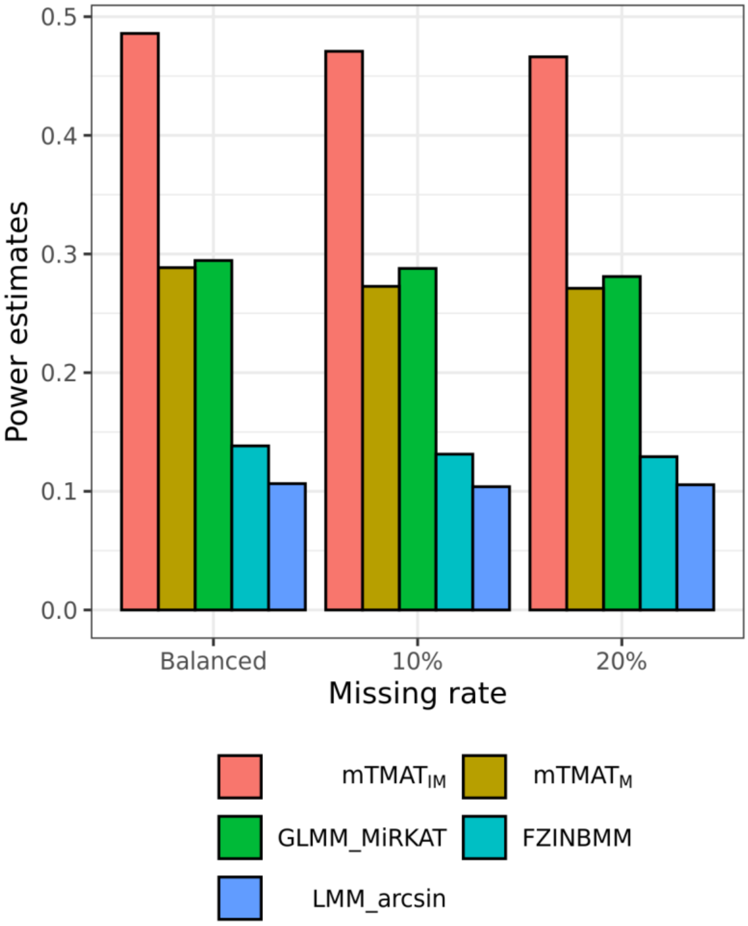


Figure 3.8. Effect of missing rate on power estimates. For each genus, power estimates at the 0.05 significance level were calculated with 500 replicates and compared in the different level of missing rates. I generated simulation data based on read counts from the dataset and considered genera with more than one OTU. I assumed the total sample size (*N*) = 50, p = 50%, $\boldsymbol{\beta}$ = 0.02 and the ratio of cases and controls is set to be 1:3.

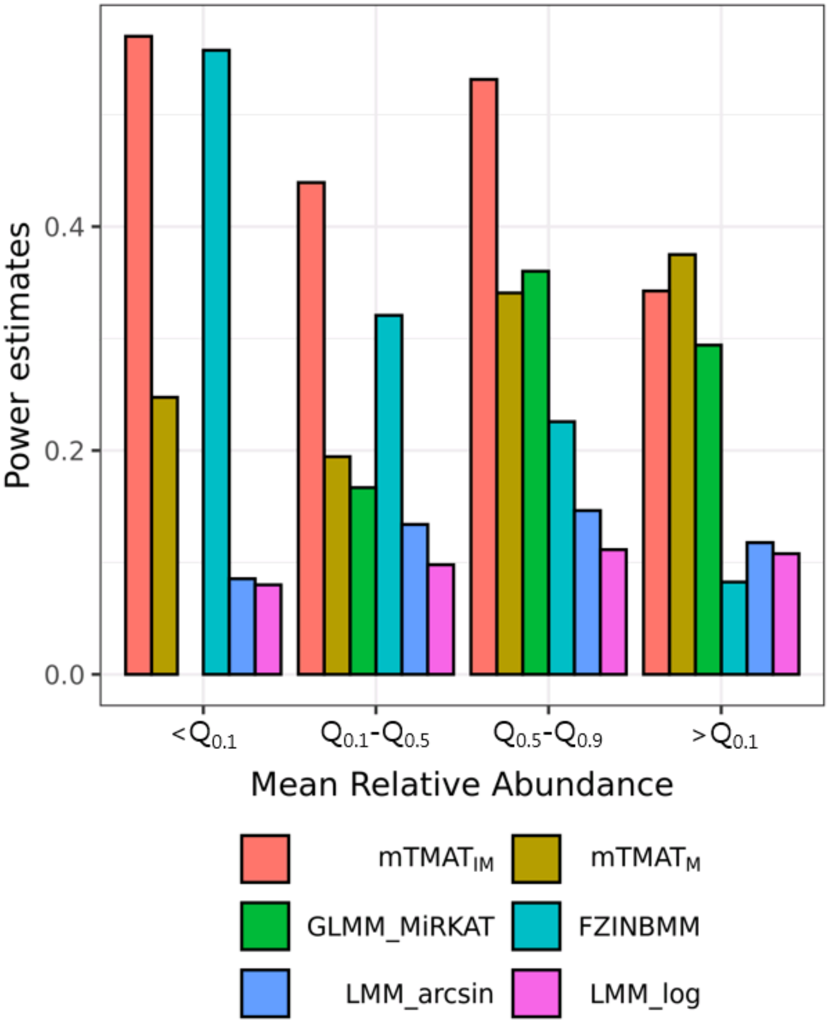


Figure 3.9. Effect of mean relative abundance on power estimates. For each genus, power estimates at the 0.05 significance level were calculated with 500 replicates and compared in the different level of mean relative abundance. I generated simulation data based on read counts from the dataset and considered genera with more than one OTU. I assumed the total sample size (N) = 50, p = 50%, $\boldsymbol{\beta}$ = 0.02 and the ratio of cases and controls is set to be 1:3.

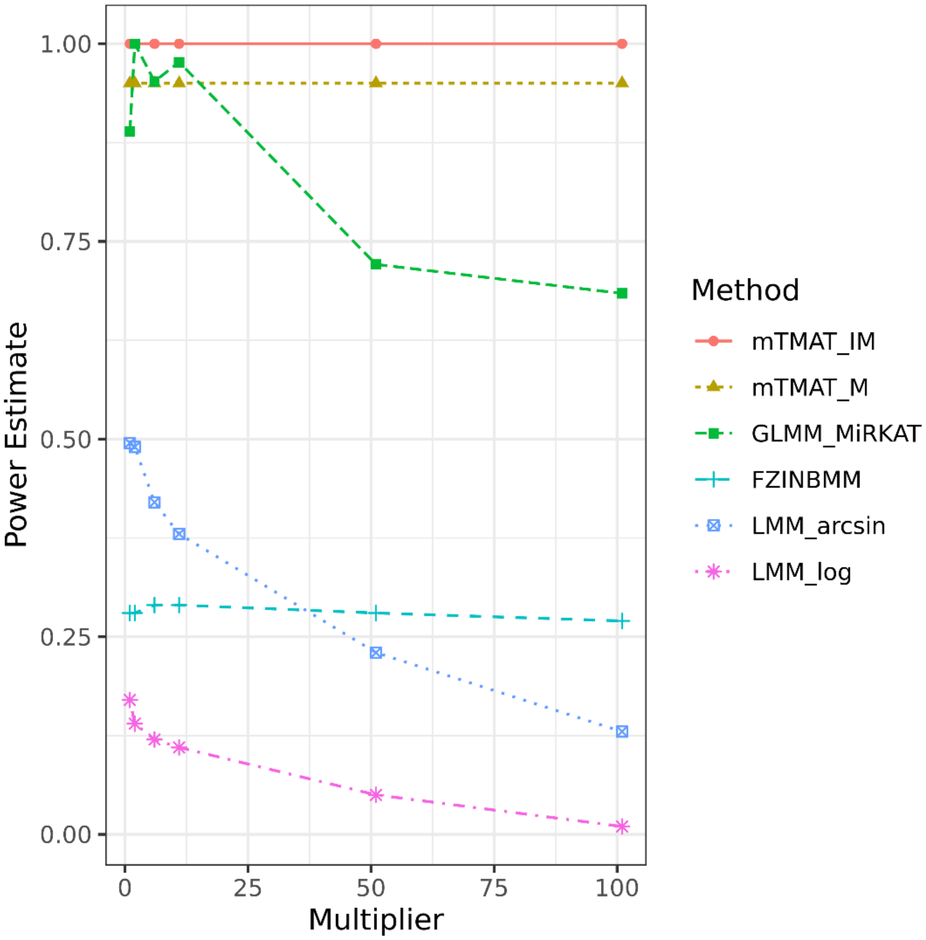


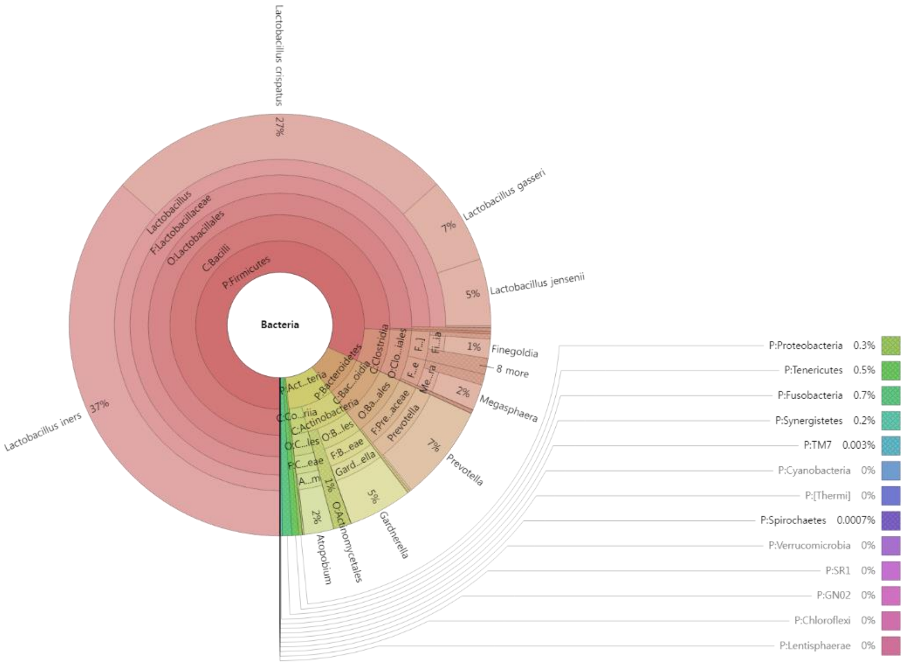
Figure 3.10. Effect of compositional bias on power estimates. Power estimates at the 0.05 significance level were calculated with 500 replicates with different level of multiplier. I assumed the total sample size (*N*) = 50, p = 50%, $\boldsymbol{\beta}$ = 0.15 and the ratio of cases and controls is set to be 1:3 at missing rate 10%.

Real data analysis

The pregnant datasets were analyzed with mTMAT, GLMM-MiRKAT, FZINBMM, LMM with the arcsine square root transformation and LMM with log transformation. The pregnant dataset includes the race, days after the first visit (GDColl), house hold income, maternal education, gender of baby. Overall composition is described in Figure 3.11 and the overall composition change was clear when 300 and more days has passed. Figure 3.12 shows that the change can be related with the pregnancy state. PERMANOVA analysis result shows the associated phenotype that explained microbiome variability. Race was the most significant covariates with p-value = 0.06 (Figure 3.13). Table 3.8 show that mTMAT$_{IM}$ found 11 significant genera. FZINBMM, LMM-arcsine and LMM-log found 16, 14 and 14 significant genera respectively. As shown in simulation study, most of detected genera as significant only by FZINBMM can be false positives. mTMAT$_{IM}$ shared most of significant genera with other methods. The most significant genera was *Lactobacillus*, which was consistent with the original paper [72]. Figure 3.14 shows a Venn diagram comparing the numbers of significant genera implicated by the various applied methods. As LMM-arcsine and LMM-log differ only in their transformation, those two methods shared all the 16 detected genera. FZINBMM detected two more genera that was no detected by any other methods. mTMAT$_{IM}$ shared all the 12 detected genera with other methods.

Figure 3.15 shows the distribution of OTUs under *Lactobacillus*. Lactobacillus has five leaf nodes and the relative abundances of all the leaf node m = 1, 2, 3, 4 and 5 were higher in pregnant group. *Lactobacillus* has been found to be more abundant in pregnant group than in healthy group and the absence of vaginal *Lactobacillus* species can increases the risks of preterm delivery [74]. Figure 3.16-22 showed the OTU distributions of other associated genera. These results confirm that the genera identified using mTMAT may be associated with delivery. Thus, it can be concluded that mTMAT successfully detected associated genera.

A. Baseline



B. 0-200 days from baseline

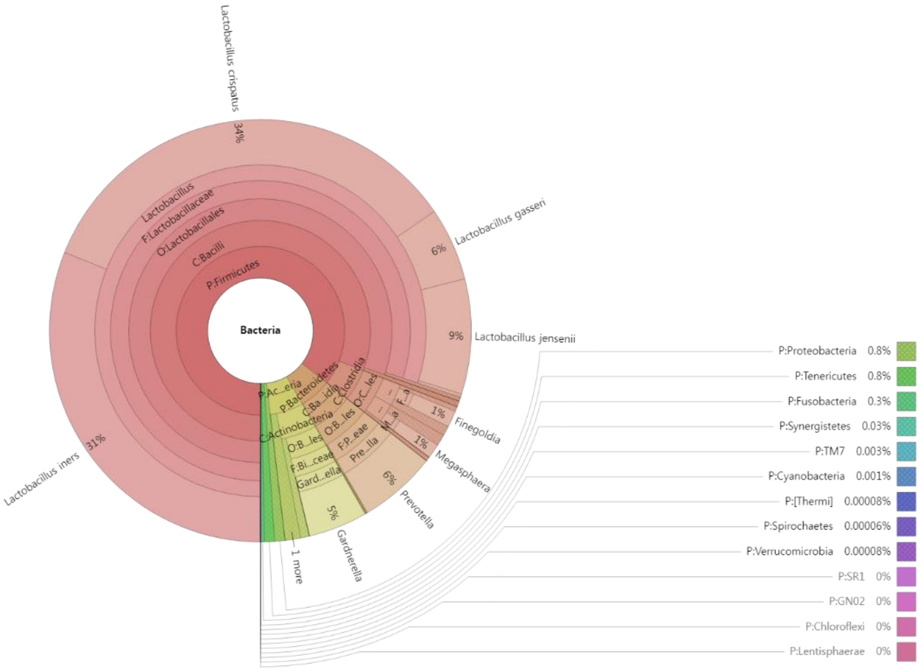
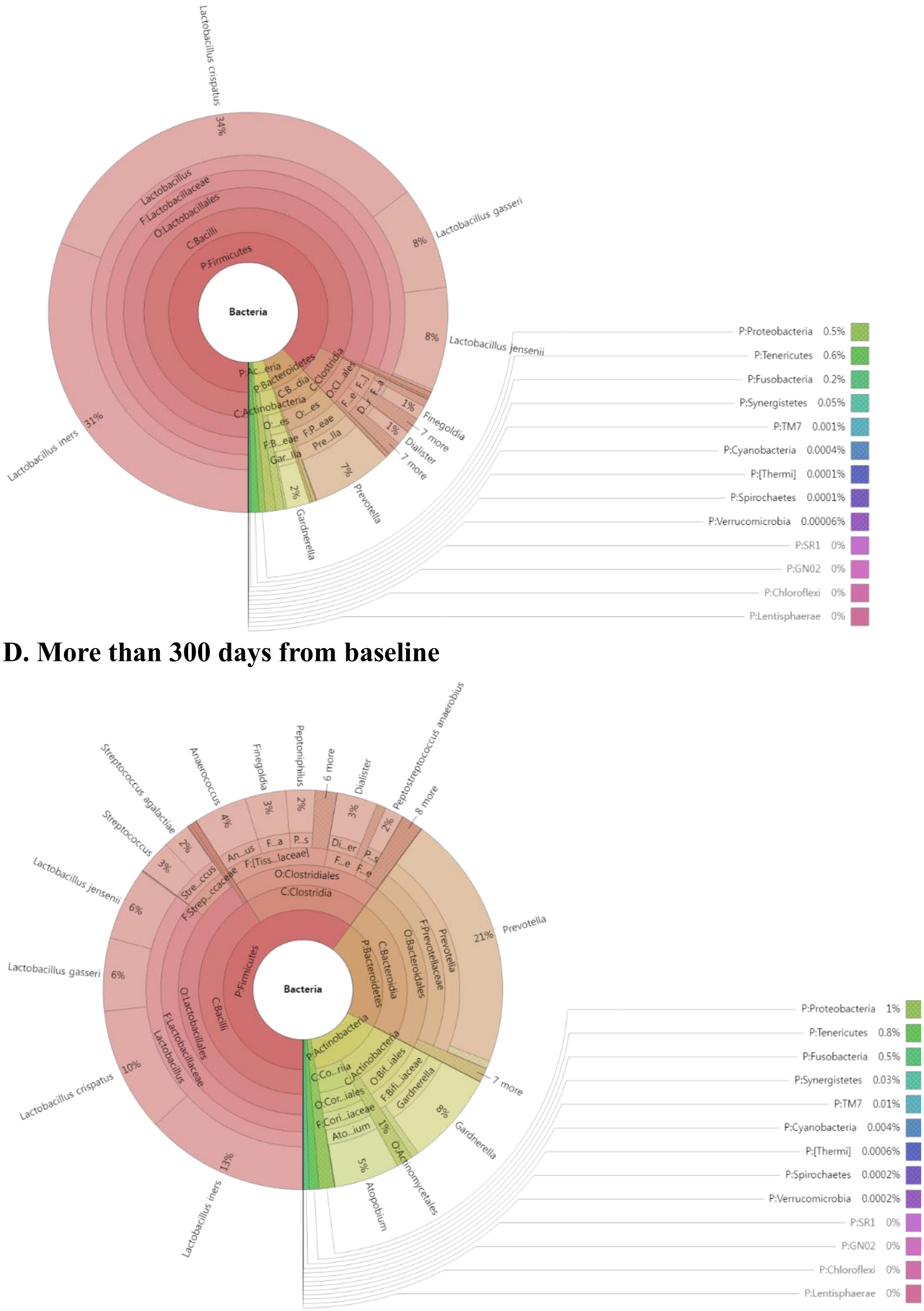


Figure 3.11. Change of microbial composition. Krona plots showing the mean relative abundance of bacterial taxa with different time range at the species level.

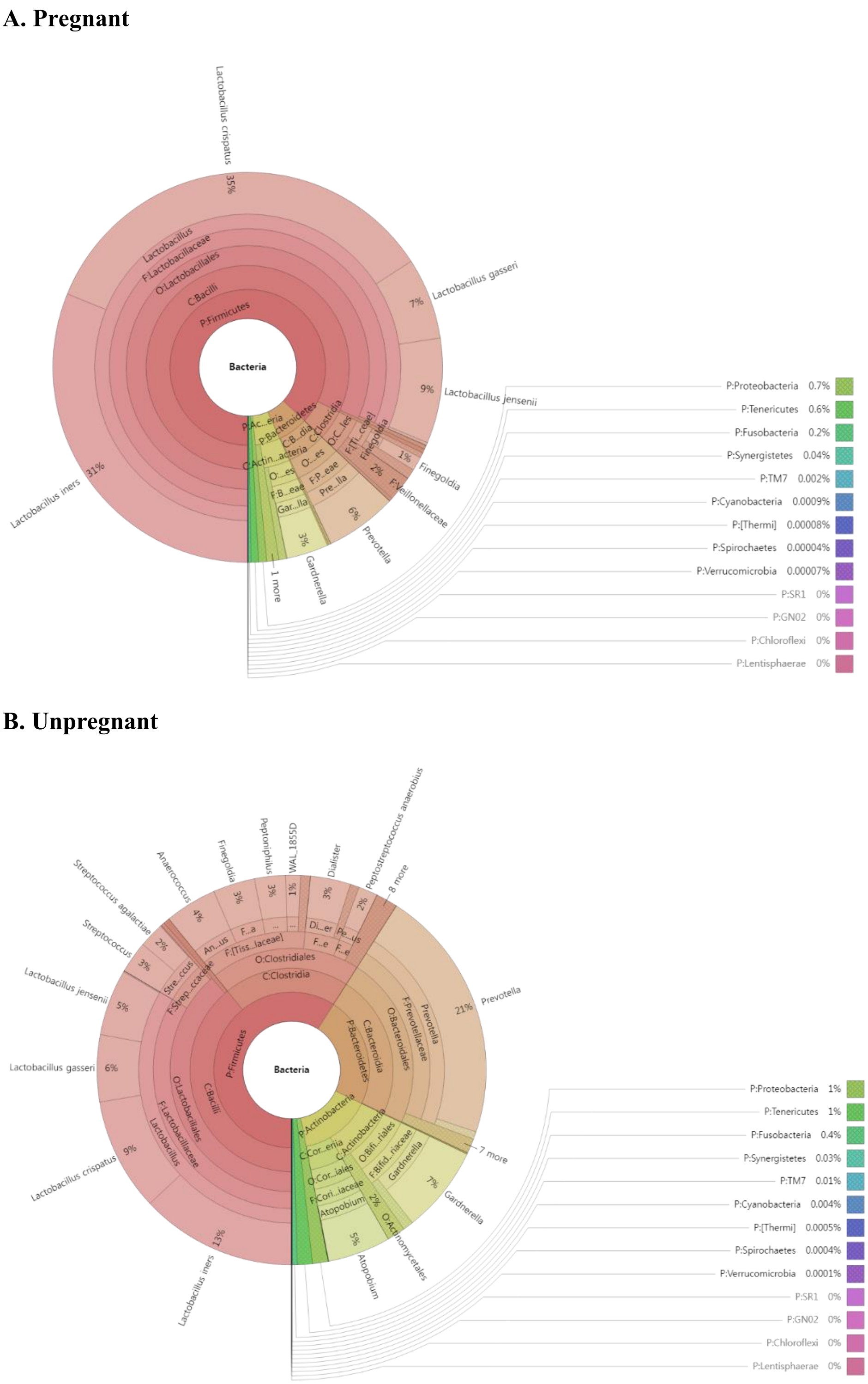
(Continued)



C. 200-300 days from baseline

D. More than 300 days from baseline

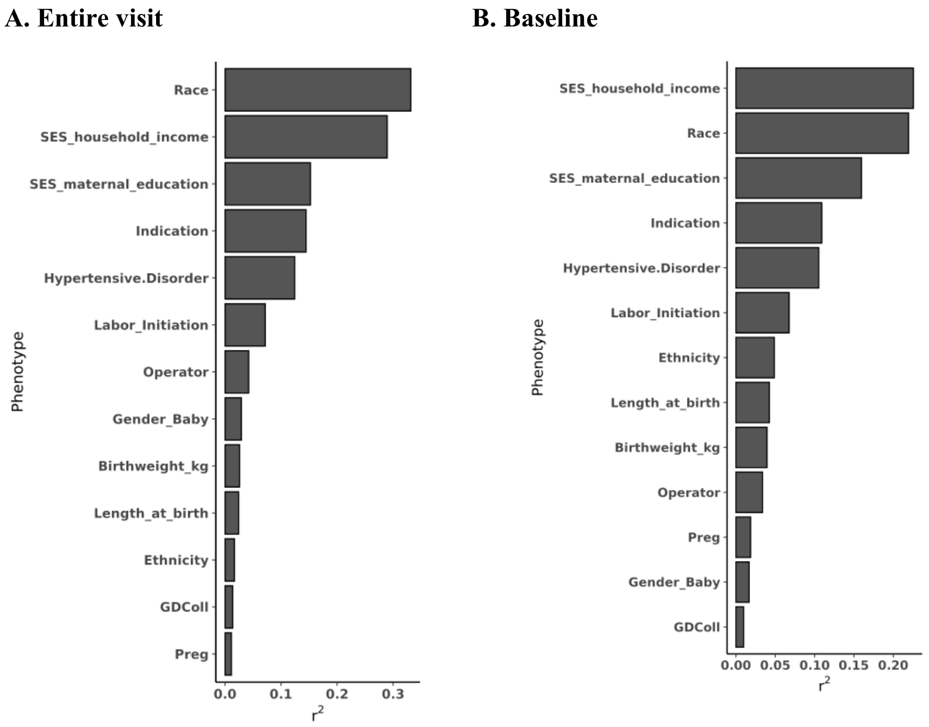
Figure 3.11. Continued



A. Pregnant

B. Unpregnant

Figure 3.12. Microbial composition by pregnant groups. Krona plots showing the mean relative abundance of bacterial taxa with different pregnant group at the species level.



A. Entire visit

B. Baseline

Figure 3.13. The relative importance of variables. Relative proportions of variance attributable to each variable were calculated with PERMANOVA. Pldist and bray-curtis distance is used for the calculation of beta-diversity.

Table 3.8. Association analysis results of Pregnant dataset. Results for genera significantly associated with at least one method at the FDR-adjusted 0.05 significance level were summarized

Family	Genus	mTMAT$_{IM}$	mTMAT$_{M}$	GLMM-MiRKAT	FZINBMM	LMM arcsine	LMM log
F:Campylobacteraceae	*Campylobacter*	0.01984	0.02275	NA	1.90E-26	5.45E-22	5.40E-11
F:Veillonellaceae	*Dialister*	0.01984	0.02275	0.06983	4.21E-32	2.51E-21	2.02E-13
F:[Tissierellaceae]	*Finegoldia*	0.01984	0.02110	NA	4.13E-29	2.55E-15	6.81E-11
F:Lactobacillaceae	*Lactobacillus*	0.01984	0.01549	0.00416	2.58E-44	9.01E-76	1.65E-73
O:Clostridiales	*O:Clostridiales*	0.01984	0.02110	NA	1.88E-40	1.77E-23	8.57E-08
F:[Tissierellaceae]	*Peptoniphilus*	0.01984	0.02110	0.00416	1.40E-22	7.75E-30	5.09E-24
F:Porphyromonadaceae	*Porphyromonas*	0.01984	0.02522	NA	3.56E-36	1.18E-12	2.20E-06
F:Streptococcaceae	*Streptococcus*	0.01984	0.02110	NA	1.30E-61	2.37E-23	8.87E-13
F:Actinomycetaceae	*Varibaculum*	0.01984	0.02110	NA	1.89E-18	2.63E-09	0.00612
F:[Tissierellaceae]	*Anaerococcus*	0.02110	0.02275	0.01247	3.87E-44	5.24E-42	3.44E-30
F:Prevotellaceae	*Prevotella*	0.02599	0.02110	0.00416	1.88E-40	1.75E-31	6.82E-29
F:[Tissierellaceae]	*1-68*	0.04797	0.04647	NA	2.23E-20	8.12E-09	1.43E-05
F:[Tissierellaceae]	*WAL_1855D*	0.05706	0.03594	NA	1.18E-18	2.08E-11	3.68E-09
F:Actinomycetaceae	*Mobiluncus*	0.09401	0.05409	NA	4.13E-17	3.71E-12	6.16E-07
F:Coriobacteriaceae	*Atopobium*	0.13031	0.10538	NA	5.08E-28	1.07E-10	4.75E-09
F:Mycoplasmataceae	*Ureaplasma*	0.53138	0.85988	NA	0.38339	0.40960	0.09996
F:Corynebacteriaceae	*Corynebacterium*	0.55456	0.34057	NA	6.83E-06	0.28831	0.93412
F:Bifidobacteriaceae	*Gardnerella*	0.84255	0.67380	NA	5.59E-05	0.00165	0.00171
F:Actinomycetaceae	*Actinomyces*	0.96391	0.79316	NA	0.11827	0.82055	0.70798
F:Staphylococcaceae	*Staphylococcus*	0.99925	0.79316	NA	0.00014	0.53031	0.59088

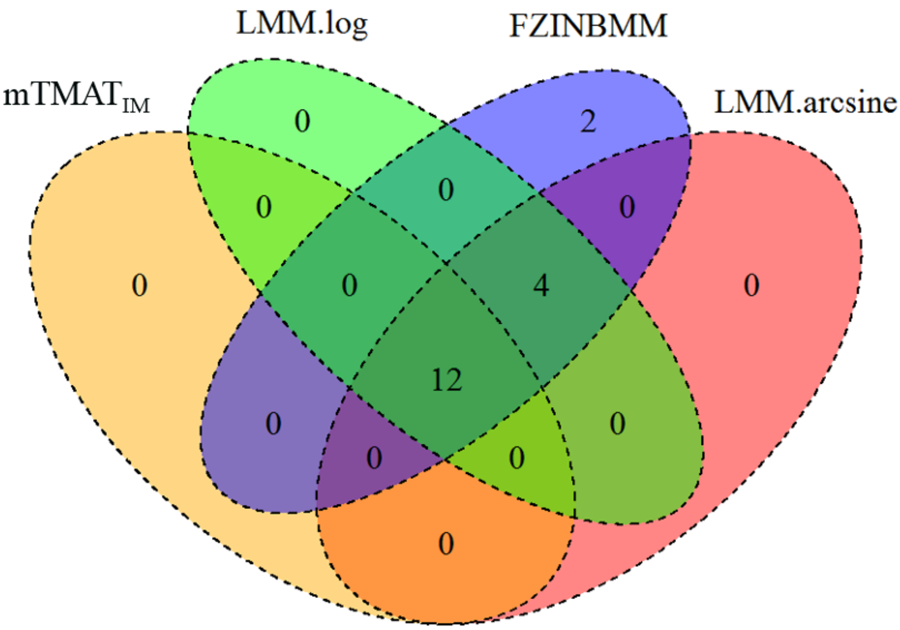


Figure 3.14. Comparison of significantly associated genera among different statistical methods. The number of significantly associated genera at the FDR-adjusted 0.05 significance level are compared among different methods.

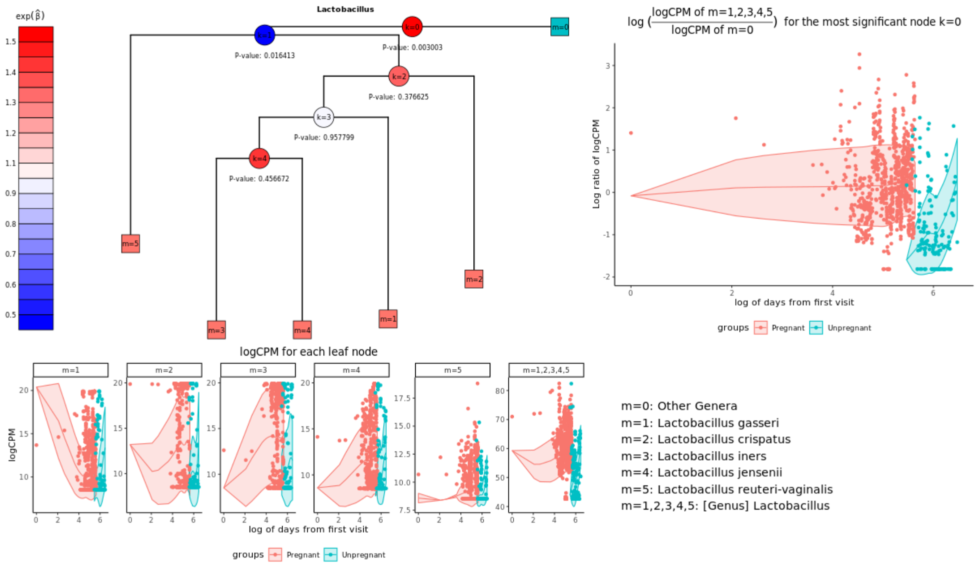


Figure 3.15. OTU distributions of significantly associated genus *Lactobacillus*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of $C^{k}{}_{ij}/D^{k}{}_{ij}$ between cases and controls after adjusting for covariates.

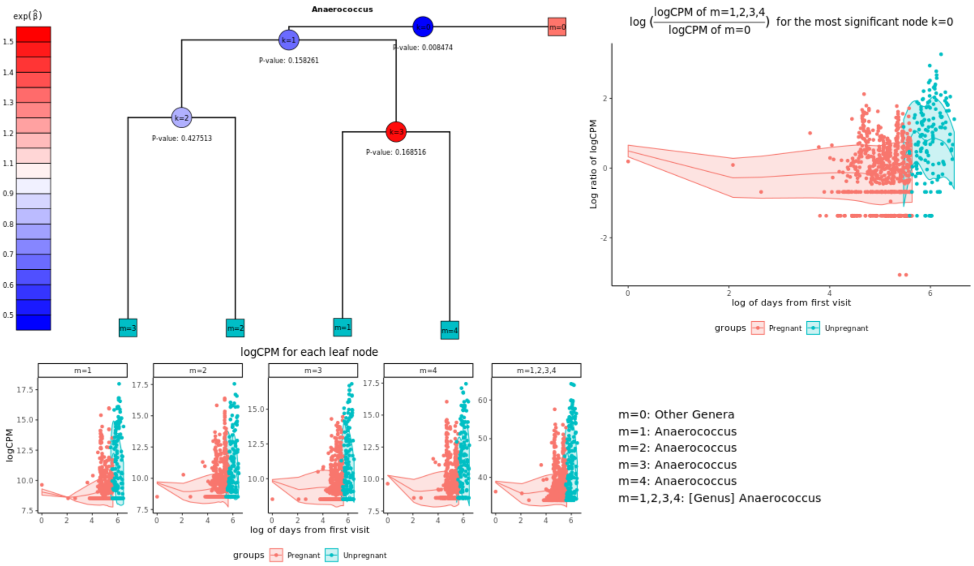


Figure 3.16. OTU distributions of significantly associated genus *Anaerococcus*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^k_{ij}/D^k_{ij} between cases and controls after adjusting for covariates.

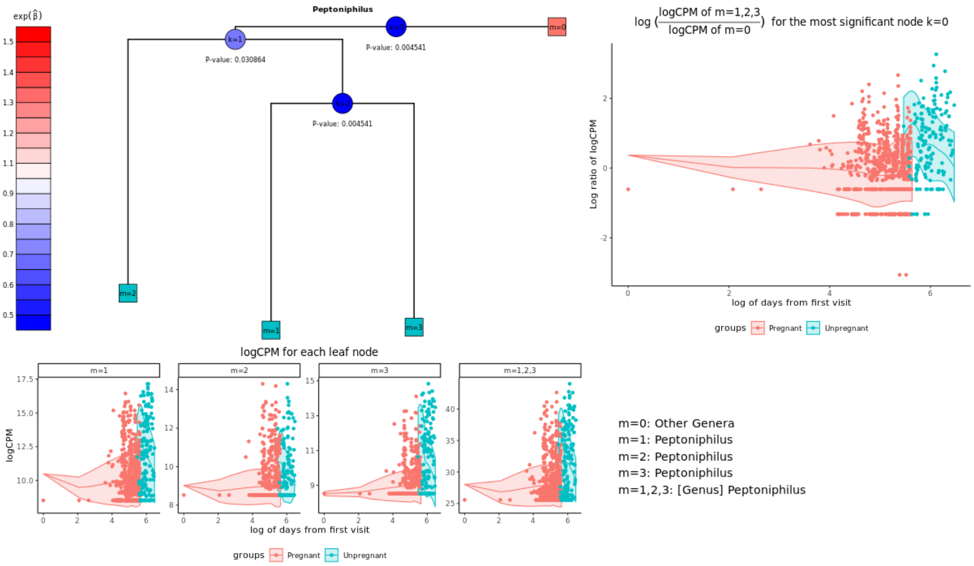


Figure 3.17. OTU distributions of significantly associated genus *Peptoniphilus*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^k_{ij}/D^k_{ij} between cases and controls after adjusting for covariates.

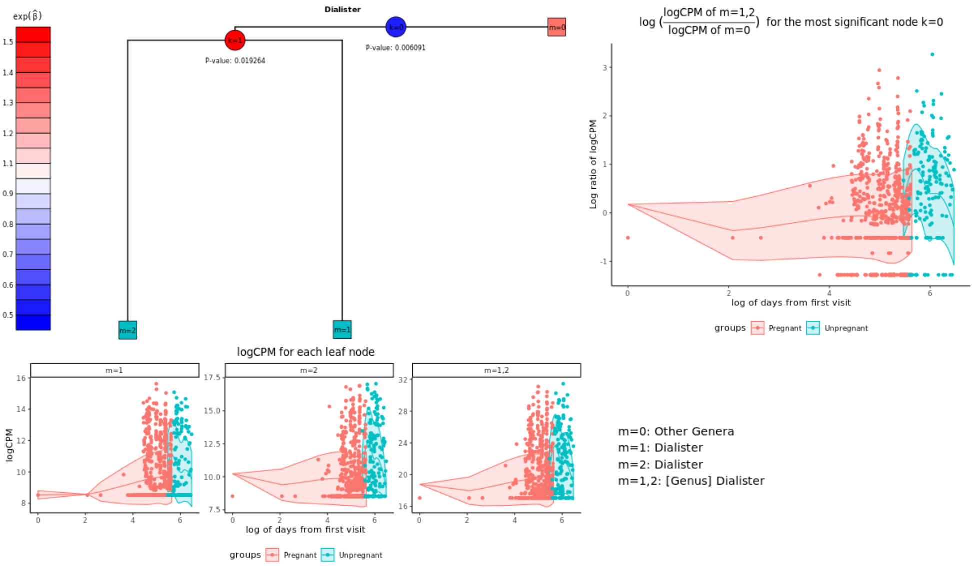


Figure 3.18. OTU distributions of significantly associated genus *Dialister*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^{k}_{ij}/D^{k}_{ij} between cases and controls after adjusting for covariates.

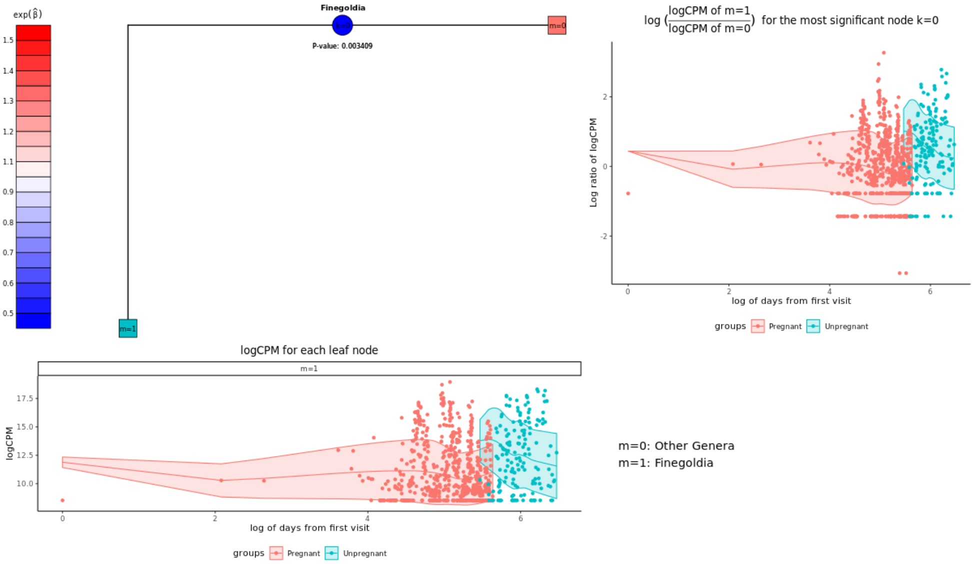


Figure 3.19. OTU distributions of significantly associated genus *Finegoldia*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{\text{IM}}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^{k}_{ij}/D^{k}_{ij} between cases and controls after adjusting for covariates**.**

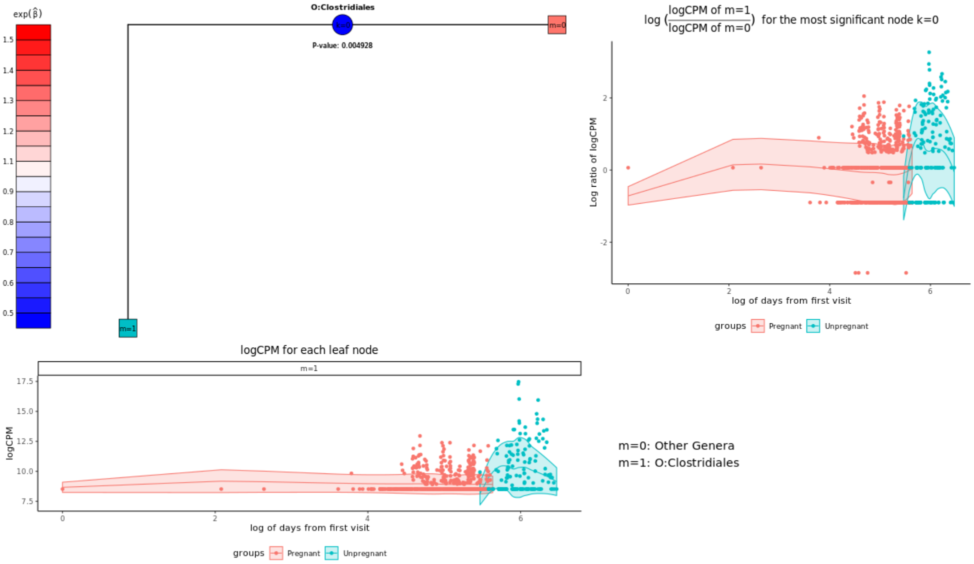


Figure 3.20. OTU distributions of significantly associated unclassified *Clostridiales*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\hat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\hat{\boldsymbol{\beta}}$) indicates the mean difference of C^{k}_{ij}/D^{k}_{ij} between cases and controls after adjusting for covariates.

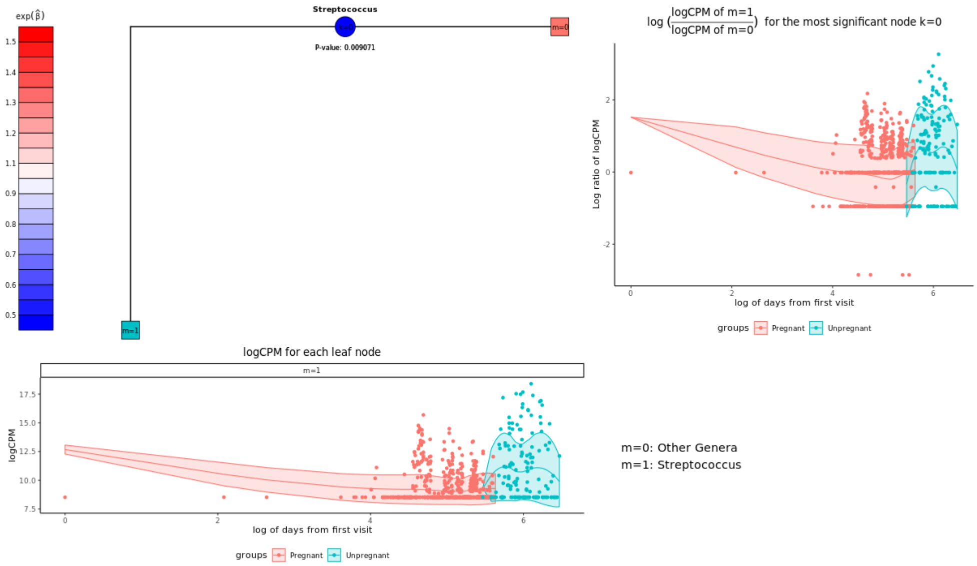


Figure 3.21. OTU distributions of significantly associated genus *Streptococcus*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^{k}_{ij}/D^{k}_{ij} between cases and controls after adjusting for covariates.

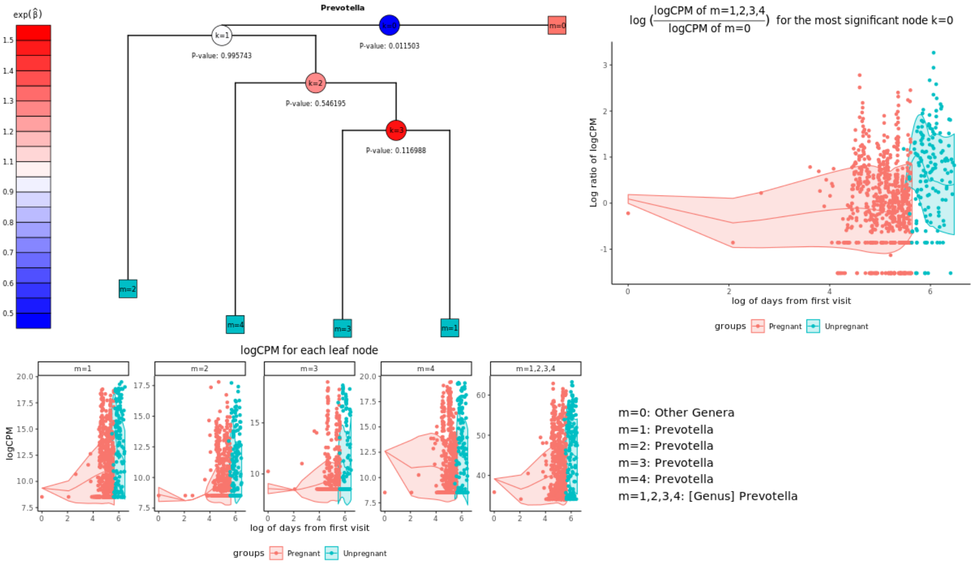


Figure 3.22. OTU distributions of significantly associated genus *Prevotella*. Relative proportions of OTUs belonging to significantly associated genera according to mTMAT$_{IM}$ were calculated. The blue internal node indicates that OTUs in left test leaf nodes are more abundant in category in blue than the category in red. Each OTU has its corresponding leaf node, and leaf nodes in blue and red indicate that they are more frequently observed in the category in blue and red, respectively. For exp($\widehat{\boldsymbol{\beta}}$), indicates the maximum likelihood estimate for the quasi-likelihood, and exp($\widehat{\boldsymbol{\beta}}$) indicates the mean difference of C^k_{ij}/D^k_{ij} between cases and controls after adjusting for covariates.

3.4. Discussion

The importance of microbiome-host interactions has been known for more than a century [53], and it has been shown that the occurrence of many human diseases is related to bacterial communities.

Microbiome data is phylogenetic number structured and has some unique properties, including high dimensionality, rarity and heterogeneity beyond composition. These unique properties cause multiple statistical problems when analyzing data across microbial composition and integrating multi-omics data such as large p and small n, dependencies, over-dispersion and zero inflation.

On the other hand, the classical correlation and related methods in the whole microbial association study are still applied in the actual study and used in the development of new methods. But by the problems of metagenomic analysis , the performance traditional approaches is normally low especially for more complex models such as longitudinal models including linear mixed models and generalized linear mixed model.

Here, I propose a new method for detecting OTUs associated with host diseases. mTMAT statistics are based on quasi-scores for internal nodes in a phylogenetic tree. In addition to this, mTMAT includes non-independent correlation structure and various correlation structure and the robust estimation is considered. The use of log CPM transformation and integrating abundances using phylogenetic tree helps the convergence of GEE approaches of mTMAT.

Then those statistics are combined into a single statistic with a minimum p-value. By using such quasi-score statistics, mTMAT can identify differences among OTUs significantly associated with host diseases, while existing statistical method, such as GLMM-MiRKAT cannot. Furthermore, by the nature of the proposed statistics, the statistical scores for internal nodes are independent, and the minimum p-value can be directly calculated. I compared the performance of mTMAT with those of GLMM-MiRKAT, LMM-arcsine, LMM-log and FZINBMM under various simulation scenarios. According to my results, mTMAT correctly controlled the nominal type-1 error rate and was statistically the most powerful method for detecting associations with host diseases in my simulation studies. Also, methodologies using permutation-based p-values, such as GLMM-MiRKAT, are computationally very slow compared to mTMAT.

However, despite the flexibility of mTMAT, the proposed method has several limitations. First, mTMAT combines the statistics for each test internal nodes and multiple comparison occurs when the number of leaf nodes is large. Also, the performances of the methods described in this paper can be dependent on the simulation setting even though the simulation tried to reflect the characteristics of metagenome dataset and used a the real microbial count data. The choice of database and OTU clustering methods can affect the statistical properties of mTMAT. mTMAT can help researchers easily perform fast and effective microbiome-wide association analysis, which provides a comprehensive understanding of how the entire microbiota interacts with the human body.

Chapter 4. Longitudinal Measurement of Urine Microbiome Reveals the Role of uncultured *Lachnospiraceae* on Type-2 Diabetes Pathogenesis

4.1. Introduction

Recent studies have revealed that the intestinal microbiota plays essential roles in host energy homeostasis, body adiposity, blood sugar control, insulin sensitivity, hormone secretion, and metabolic diseases, such as type 2 diabetes (T2D) and obesity [54-56]. However, most of these studies utilized stool samples and therefore provide limited information when compared to those from the direct sampling of the intestinal mucosa which is not possible in most cases. In addition, the composition of microbial communities in stool samples is affected by the compartment they reside in, such as the mucous membrane [75]. Moreover, microbial communities differ based on their source, ranging from the intestines, skin, and airways, which are frequently studied, to urine and blood, which are normally sterile environments [76]. Therefore, it is important to understand the function of not only the intestinal microbiota, but also the entire whole-body microbiota.

Extracellular vesicles (EVs) have been recently suggested to be the main messengers between intestinal microorganisms and the host. EVs have been shown to travel long distances within the body [77], and have been used as biomarkers of atopic dermatitis, alcoholic hepatitis, and asthma [78-81]. Microbiota-derived EVs can enter the circulatory system through the intestinal barrier and are expected to play a key role in the development of insulin resistance, and thus may provide important clues into T2D pathogenesis. For example, EVs derived from

Pseudomonas panacis present in the stool samples of high-fat-diet-fed mice infiltrated the gut barrier and blocked the insulin pathway in skeletal muscle and adipose tissue, and induced the development of insulin resistance and glucose intolerance [82]. However, microbiota-derived EVs are highly variable because they are affected by various factors, such as age and sex. Therefore, caution should be exercised when inferring causal relationships through statistical analysis of microbiota data. Longitudinal microbiota studies can allow stronger inferences than cross-sectional studies [57] and the detection of microorganisms related to the progression of T2D among healthy subjects. However, the existing studies are predominantly cross-sectional in nature and are based on correlation analyses, and therefore are limited with respect to providing an understanding of the exact roles of the intestinal microbiota and EVs in the development of metabolic diseases.

Therefore, in this study, I investigated the prospective Korean Association REsource project (KARE) cohort [83]. I used a cohort of Korean adults of 40 years of age or older for tracking the changes during different stages of T2D progression. By tracking changes in microbiota-derived EVs in urine samples collected three times over four years, I investigated the microorganisms potentially associated with T2D. Furthermore, using genomic and metabolite data from the KARE cohort, I conducted a multi-omics analysis to investigate the role of microorganisms potentially involved in the pathogenesis of T2D. I expected my findings to provide clues as to how microbes, the substances they produce, and their by-products interact with the human body and affect the development of metabolic diseases. In addition, using genomic data, I evaluated the causal relationships among microbial organisms, and clinical measures, with the aim of clarifying the relationship between T2D and microorganisms.

4.2. Materials and Methods

Ethics statement

The protocol used in this study was approved by the Institutional Review Board (IRB No. E1801/001-004) in Seoul National University.

Cohort and study design

The KARE cohort is a prospective study cohort involving subjects from the rural community of Ansung and the urban community of Ansan in South Korea. The KARE project began in 2001 as part of the Korean Genome Epidemiology study [67]. I used data from urine samples taken, and stored at −80°C from 2013, 2015, and 2017, which I refer to as phases 1, 2, and 3 in this study. For 1,891 subjects for whom urine samples were available, age, sex and BMI were matched by 2:1:1 propensity score matching. As a result, healthy group (healthy in all phases, $N = 328$), a T2D-at-risk group (T2D-at-risk in all phases, $N = 164$), and a T2D group (T2D in any of the three phases, $N = 164$) were selected. From the remaining unmatched subjects, 35 T2D subjects were also included. Consequently, 691 subjects were finally included, and their 2,072 urine samples were utilized for microbiota analysis. Metagenomic, metabolite, clinical, and genomic data were used for analyses (Figure 4.1). This study was approved by the Institutional Review Board of the Korea University Ansan Hospital, and written informed consent was given by all study subjects.

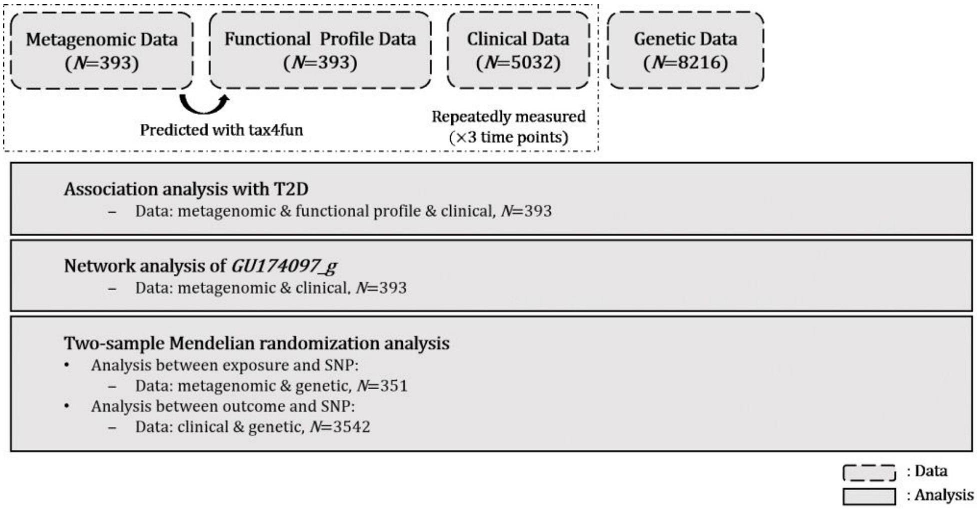

Metagenomic Data
(N=393)
Functional Profile Data
(N=393)
Clinical Data
(N=5032)
Genetic Data
(N=8216)
Predicted with tax4fun
Repeatedly measured
(×3 time points)
Association analysis with T2D
– Data: metagenomic & functional profile & clinical, N=393
Network analysis of GU174097_g
– Data: metagenomic & clinical, N=393
Two-sample Mendelian randomization analysis
• Analysis between exposure and SNP:
– Data: metagenomic & genetic, N=351
• Analysis between outcome and SNP:
– Data: clinical & genetic, N=3542
: Data
: Analysis

Figure 4.1. Summary chart of data analysis.

Operational definition of T2D and related phenotypes

Participants were categorized into controls, T2D-at-risk patients, and T2D patients. T2D and T2D-at-risk patients were diagnosed on the basis of criteria provided by the American Diabetes Association. The detailed criteria are provided in Table 4.1. T2D status was then stratified into *T2D-at-risk/T2D* (0 for healthy; 1 for T2D-at-risk and T2D) and *binary_T2D* (0 for healthy and T2D-at-risk; 1 for T2D). In addition, I considered other T2D-related phenotypes, such as body mass index (BMI), HbA1c, fasting glucose and insulin, 60- and 120-minute plasma glucose, and insulin levels after a 75 g oral glucose tolerance test for the analysis. Age, total cholesterol, high-density lipoprotein (HDL) cholesterol, triglyceride, kidney- and liver-related disease indicators—such as levels of blood urea nitrogen (BUN), creatinine, aspartate aminotransferase (AST), and alanine aminotransferase (ALT)—C-reactive protein (CRP), white blood cell (WBC) count, red blood cell (RBC) count, hemoglobin, hematocrit, and platelet count were also collected. The homeostatic model assessment for insulin resistance (HOMA-IR) was calculated using fasting glucose and fasting insulin levels [84]. Descriptive statistics for all variables were generated using the Rex software (RexSoft Inc., Seoul, Korea) (Table 4.2) [85].

Table 4.1. Criteria used in this study for the diagnosis of T2D and T2D-at-risk patients

Diagnosis Condition	Fasting glucose (mg/dL)	120 min glucose (mg/dL)	HbA1c (%)	Diabetic medicine taken
Without Diabetes	<100	<140	<5.7	No
T2D-at-risk (Prediabetes)	100-125	140-199	5.7-6.4	No
T2D (Diabetes)	≥126	≥200	≥6.5	Yes

Table 4.2. Descriptive statistics for KARE cohort subjects

Variable	Category	Statistics	Phase 1 (N=393)	Phase 2 (N=393)	Phase 3 (N=393)	Total (N=1179)
Sex						
	Male	n(%)	192(48.85%)	192(48.85%)	192(48.85%)	576(48.85%)
	Female	n(%)	201(51.15%)	201(51.15%)	201(51.15%)	603(51.15%)
	Total	n(%)	393(100.00%)	393(100.00%)	393(100.00%)	1179(100.00%)
Age						
		n	393	393	393	1179
		mean ± Std	57.23 ± 5.88	59.20 ± 5.91	61.29 ± 5.93	59.24 ± 6.13
		Q1, Q3	53.00, 60.00	55.00, 62.00	57.00, 64.00	55.00, 63.00
		min ~ max	49.00 – 77.00	51.00 – 79.00	53.00 – 81.00	49.00 – 81.00
Hemoglobin						
		n	393	393	393	1179
		mean ± Std	5.55 ± 0.42	5.55 ± 0.45	5.67 ± 0.60	5.59 ± 0.50
		Q1, Q3	5.30, 5.80	5.30, 5.80	5.30, 5.90	5.30, 5.80
		min ~ max	4.10 – 7.40	4.20 – 7.90	4.30 – 11.10	4.10 – 11.10
Glu0						
		n	393	393	393	1179
		mean ± Std	92.52 ± 9.47	92.06 ± 11.58	94.80 ± 14.44	93.12 ± 12.05
		Q1, Q3	86.00, 97.00	85.00, 96.00	87.00, 99.00	86.00, 98.00
		min ~ max	71.00 – 141.00	62.00 – 159.00	71.00 – 224.00	62.00 – 224.00
Glu60						
		n	372	364	344	1080
		mean ± Std	161.60 ± 46.01	164.50 ± 49.51	172.60 ± 48.88	166.08 ± 48.30
		Q1, Q3	127.00, 194.00	127.00, 199.00	140.00, 204.00	130.00, 200.00
		min ~ max	50.00 – 304.00	66.00 – 373.00	54.00 – 423.00	50.00 – 423.00
Glu120						
		n	372	364	344	1080
		mean ± Std	136.10 ± 40.59	141.66 ± 43.85	144.25 ± 53.05	140.57 ± 46.04
		Q1, Q3	105.75, 158.25	112.00, 166.00	106.75, 173.25	107.00, 166.00
		min ~ max	47.00 – 287.00	58.00 – 331.00	61.00 – 447.00	47.00 – 447.00

(Continued)

Table 4.2. Continued

Variable	Category	Statistics	Phase 1 (N=393)	Phase 2 (N=393)	Phase 3 (N=393)	Total (N=1179)
BUN						
		n	393	393	393	1179
		mean ± Std	15.37 ± 3.90	15.71 ± 3.93	15.90 ± 3.83	15.66 ± 3.89
		Q1, Q3	12.60, 17.60	12.90, 18.20	12.90, 18.10	12.90, 18.00
		min ~ max	6.70 – 32.00	7.40 – 32.20	7.80 – 32.10	6.70 – 32.20
Creatinine						
		n	393	393	393	1179
		mean ± Std	1.00 ± 0.18	1.00 ± 0.17	0.96 ± 0.18	0.99 ± 0.18
		Q1, Q3	0.85, 1.12	0.87, 1.11	0.82, 1.08	0.85, 1.11
		min ~ max	0.59 – 1.72	0.67 – 1.97	0.50 – 1.79	0.50 – 1.97
AST						
		n	393	393	393	1179
		mean ± Std	25.60 ± 7.18	24.41 ± 8.03	25.78 ± 11.99	25.26 ± 9.32
		Q1, Q3	21.00, 28.00	20.00, 27.00	21.00, 27.00	21.00, 27.00
		min ~ max	14.00 – 72.00	13.00 – 121.00	14.00 – 181.00	13.00 – 181.00
ALT						
		n	393	393	393	1179
		mean ± Std	24.48 ± 11.96	22.61 ± 10.27	24.08 ± 13.46	23.72 ± 11.98
		Q1, Q3	17.00, 28.00	16.00, 26.00	16.00, 27.00	16.00, 27.00
		min ~ max	6.00 – 118.00	8.00 – 74.00	10.00 – 158.00	6.00 – 158.00
Total_cholesterol						
		n	393	393	393	1179
		mean ± Std	200.81 ± 37.31	196.22 ± 33.81	193.37 ± 36.05	196.80 ± 35.85
		Q1, Q3	175.00, 226.00	173.00, 218.00	169.00, 215.00	172.00, 219.00
		min ~ max	116.00 – 330.00	109.00 – 313.00	95.00 – 341.00	95.00 – 341.00
HDL_cholesterol						
		n	393	393	393	1179
		mean ± Std	49.38 ± 12.86	47.58 ± 12.34	46.89 ± 12.53	47.95 ± 12.61
		Q1, Q3	40.00, 58.00	39.00, 55.00	38.00, 54.00	39.00, 55.00
		min ~ max	25.00 ~ 102.00	19.00 ~ 100.00	24.00 ~ 123.00	19.00 ~ 123.00

(Continued)

Table 4.2. Continued

Variable	Category	Statistics	Phase 1 (N=393)	Phase 2 (N=393)	Phase 3 (N=393)	Total (N=1179)
Triglyceride						
		n	393	393	393	1179
		mean ± Std	136.87 ± 81.85	132.22 ± 81.01	130.86 ± 79.97	133.32 ± 80.92
		Q1, Q3	85.00, 162.00	84.00, 156.00	83.00, 151.00	84.00, 157.00
		min ~ max	34.00 – 878.00	37.00 – 901.00	35.00 – 714.00	34.00 – 901.00
CRP						
		n	393	393	393	1179
		mean ± Std	1.40 ± 4.46	1.30 ± 2.47	1.25 ± 2.52	1.32 ± 3.28
		Q1, Q3	0.39, 1.23	0.40, 1.12	0.37, 1.14	0.39, 1.15
		min ~ max	0.01 – 77.37	0.04 – 23.36	0.08 – 33.92	0.01 – 77.37
WBC						
		n	393	393	393	1179
		mean ± Std	5.15 ± 1.64	4.98 ± 1.43	5.30 ± 1.53	5.14 ± 1.54
		Q1, Q3	4.20, 5.80	4.00, 5.60	4.30, 5.90	4.20, 5.80
		min ~ max	2.10 – 22.80	2.30 – 12.50	2.00 – 17.90	2.00 – 22.80
RBC						
		n	393	393	393	1179
		mean ± Std	4.60 ± 0.44	4.60 ± 0.42	4.45 ± 0.40	4.55 ± 0.43
		Q1, Q3	4.29, 4.86	4.31, 4.91	4.17, 4.73	4.26, 4.85
		min ~ max	3.55 – 7.04	3.40 – 6.24	3.41 – 5.70	3.40 – 7.04
Hematocrit						
		n	393	393	393	1179
		mean ± Std	42.88 ± 4.25	43.67 ± 3.89	42.35 ± 3.86	42.97 ± 4.04
		Q1, Q3	40.10, 45.40	41.10, 46.40	39.50, 44.70	40.20, 45.60
		min ~ max	32.80 – 64.90	33.30 – 56.10	33.50 – 52.50	32.80 – 64.90
Platlet						
		n	393	393	393	1179
		mean ± Std	240.83 ± 55.77	237.89 ± 53.00	247.41 ± 57.25	242.04 ± 55.46
		Q1, Q3	201.00, 273.00	201.00, 273.00	212.00, 281.00	203.00, 275.00
		min ~ max	104.00 – 492.00	86.00 – 453.00	88.00 – 471.00	86.00 – 492.00

(Continued)

Table 4.2. Continued

Variable	Category	Statistics	Phase 1 (N=393)	Phase 2 (N=393)	Phase 3 (N=393)	Total (N=1179)
Ins0						
		n	393	393	393	1179
		mean ± Std	8.33 ± 3.85	8.88 ± 4.05	9.14 ± 3.70	8.78 ± 3.88
		Q1, Q3	6.20, 9.30	6.40, 10.50	6.50, 11.20	6.35, 10.45
		min ~ max	1.60 – 33.00	1.60 – 29.10	2.10 – 25.90	1.60 – 33.00
Ins60						
		n	372	364	344	1080
		mean ± Std	53.37 ± 43.47	50.95 ± 37.41	34.65 ± 24.76	46.59 ± 37.19
		Q1, Q3	26.88, 63.90	26.18, 62.00	18.68, 42.05	24.00, 57.30
		min ~ max	2.40 – 345.90	2.50 – 248.40	2.10 – 185.40	2.10 – 345.90
Ins120						
		n	372	364	343	1079
		mean ± Std	50.94 ± 44.81	49.63 ± 42.07	41.41 ± 47.13	47.47 ± 44.82
		Q1, Q3	23.15, 61.10	19.70, 65.80	17.60, 49.75	20.20, 60.35
		min ~ max	1.60 – 343.00	2.40 – 227.00	1.00 – 670.00	1.00 – 670.00
BMI						
		n	393	393	393	1179
		mean ± Std	25.01 ± 2.93	24.87 ± 2.93	24.84 ± 3.01	24.90 ± 2.96
		Q1, Q3	23.25, 26.75	23.11, 26.72	23.03, 26.64	23.10, 26.72
		min ~ max	17.05 – 34.60	16.95 – 35.15	15.33 – 34.99	15.33 – 35.15

EV isolation and DNA extraction

Urine samples were subjected to differential centrifugation at 10,000 × *g*, 4 °C for 10 min using a microcentrifuge (Labogene 1730R; Bio-Medical Science, Seoul, Korea) to isolate EVs [86]. To remove bacteria, foreign particles, and waste, the supernatant was filtered through a 0.22-micrometer filter (Inchpor2 Syringe Filter; Inchemtec, Seoul, Korea). The isolated EVs were boiled at 100 °C for 40 min and centrifuged at 18,214 x g, 4 °C for 30 min to eliminate the floating particles and impurities. The supernatant was collected and subjected to DNA extraction using a PowerSoil® DNA Isolation Kit (MO BIO Laboratories, Carlsbad, CA, USA) according to the manufacturer’s protocol. DNA was quantified using the QIAxpert system (Qiagen, Hilden, Germany).

16S rRNA sequence data processing

Paired-end sequencing of the V3-V4 region of the bacterial 16S rRNA gene were conducted at MD Healthcare (Seoul, Korea) with the MiSeq Reagent Kits v3 (600 cycles, Illumina, San Diego, CA, USA) using the widely used primers 16S_V3_F (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGG-GNGGCWGCAG -3') and 16S_V4_R (5'- GTCTCGTGGGCTCGGAGATGTGTA-TAAGAGACAGGACTACHVGGGTATCTAATCC-3'). Adaptor sequences were detected and removed using the CUTADAPT software (https://cutadapt.readthedocs.io) with a minimum overlap of 11, maximum error rate of 10%, and a minimum length of 10 [41]. Sequences were merged using CASPER (http://best.snu.ac.kr/casper) with a mismatch ratio of 0.27 and filtered by the Phred (Q) score, resulting in sequences 350–550 bp in length [42, 68]. After the merged sequences were dereplicated, chimeric sequences were detected and removed using VSEARCH (https://github.com/torognes/vsearch) and the Silva Gold reference database for chimeras [43]. The open-reference Operational Taxonomic Units (OTU) picking was conducted based on the EzTaxon database using UCLUST (http://www.drive5.com/usearch) [46, 69]. For each OTU, I calculated its proportion among all OTUs and determined the mean value across all subjects. If the resulting value was <0.001, the OTU was excluded [70]. Among the 691 subjects, those with a read count <3,000 or for whom genomic data were not available in any phase were excluded. As a result, 1179 samples from 393 subjects, including 70 genera, were used for subsequent analyses.

Prediction of functional profiles from 16S rRNA metagenomic data

The functional potential of microbial communities can be predicted from their phylogeny. Tax4fun uses evolutionary modeling to predict metagenomes based on 16S data using the SILVA reference genome database. The SILVA-based 16S rRNA profile was used to estimate a taxonomic profile of prokaryotic Kyoto Encyclopedia of Genes and Genomes (KEGG) organisms. The estimated abundances of KEGG organisms were normalized using the 16S rRNA copy number obtained from National Center for Biotechnology Information (NCBI) genome annotations. Finally, the normalized taxonomic abundances were used to linearly combine the precomputed functional profiles of the KEGG organisms to predict the functional profile of the microbial community [87]. Similar to the analysis of OTUs, I calculated the mean of the relative proportions across all subjects for each functional profile. If the resulting value was <0.001, the functional profile was excluded from the analysis. As a result, 238 functional profiles were retained for analysis.

Analysis of bacterial composition and microbial variance

I calculated alpha- and beta-diversity indices using R (v3.6.2) after read number normalization with the Rarefy function in the R package GUniFrac (v1.1). The R package Fossil (v0.4.0) was used to obtain Chao1 and ACE diversity indexes. Shannon index and Simpson's index of diversity were calculated using the Vegan package in R (v2.5.6). Taxonomy-based ring-charts were created using the Krona tool [88]. PERMANOVA is a non-parametric multivariate analysis of variance method based on pairwise distances [89]. The R package pldist was used to obtain the microbial variance for individuals in repeated measurements of microbial profiles. pldist summarizes within-individual shifts in the microbiome composition and compares these compositional shifts across individuals. It calculates dissimilarities depending on a novel transformation of relative abundances, which then are extended to more than two time points and are incorporated into a chosen beta-diversity, in this case, Bray–Curtis dissimilarity. PERMANOVA was performed for biochemistry-related KARE phenotypes using the adonis function in R. PERMANOVA can be applied to the cross-sectional data, and thus the phenotypes were averaged for phase 1,2 and 3.

Statistical analysis of the effect of the microbiome on T2D and diabetes-risk indicators

For each taxon and functional profile, a generalized linear mixed model (LMM) with logit link function was used to find associations with *binary_T2D* and *T2D-at-risk/T2D*, whereas a LMM was used for log-transformed diabetes-risk indicators. Random effect with compound symmetry structure for each time point was incorporated to adjust the similarity of the T2D status of the same subject at different time points, and the sandwich estimator was used to find a robust estimate against the misspecified covariance matrix. To accommodate the multiple testing problem, *p*-values were adjusted for the false discovery rate (FDR) using the Benjamini–Hochberg method [27].

Network analysis of a T2D-related taxon based on multi-omics data

To assess overall associations using repeatedly measured multi-omics data, I first modeled a LMM using the log-transformed diabetes-risk indicators as response variables and age in phase 1 and sex as explanatory variables with a compound symmetry structure for its covariance structure. Then, I modeled a LMM with a T2D-related taxon as response variables with the same covariates and covariance structure. For each combination of diabetes-risk indicators and a T2D-related taxon, two different set of residuals were obtained, and Spearman correlations between the residuals were calculated.

To calculate simple correlations among diabetes-risk indicators and a chosen taxon, network analysis was conducted. Edge width was calculated as $-\log_{10}$ of the *p*-value. The network was visualized using the R package visNetwork (v2.0.8).

Genotyping, imputation, and quality control

Quality control and genotype imputation were performed according to the standard quality control and imputation protocols for the genotypes of 8,842 KARE cohort participants [90]. After quality control, 8,216 subjects with 17,716,215 SNPs were included in the analysis. In total, 351 subjects with a read count <3,000 or non-missing T2D status for all phases were used for a genome-wide association study (GWAS) of metagenomic data. Among the subjects not included in metagenome GWAS, 3,542 subjects had KARE phenotypes for the three phases and they were used for a GWAS of KARE phenotypes. The subjects in metagenome GWAS were excluded for the purposes of a two-sample MR study. Details are provided in Figure 4.1.

Mendelian randomization (MR) analysis

MR uses genetic variants, which are not associated with conventional confounders of observational studies, and therefore is considered analogous to randomized controlled trials [91]. There are two types of MR: two-sample MR and one-sample MR. Two-sample MR uses two independent datasets with non-overlapping samples for the associations of SNP-exposure and SNP-outcome (as opposed to one-sample MR) and is less likely to lead to inflated type 1 error rates and false-positive findings when compared to one-sample MR. Two-sample MR was conducted to identify effect of a microbial taxon on KARE phenotypes by using no overlapping samples.

For two-sample MR, the average *F*-statistic was used to avoid weak instrument bias. The inverse-variance-weighted (IVW) method, Cochran's *Q* test, and MR-PRESSO global test were used to confirm the heterogeneity assumption, and I^2 was used for the no measurement error (NOME) assumption. To enhance the validity of the MR analysis, I considered the extensive range of existing MR methods, including IVW, MR-egger, MR-egger with SIMEX correction, median-weighted method, and MR-PRESSO, and I selected the recommended MR method based on the violations of MR assumptions [92].

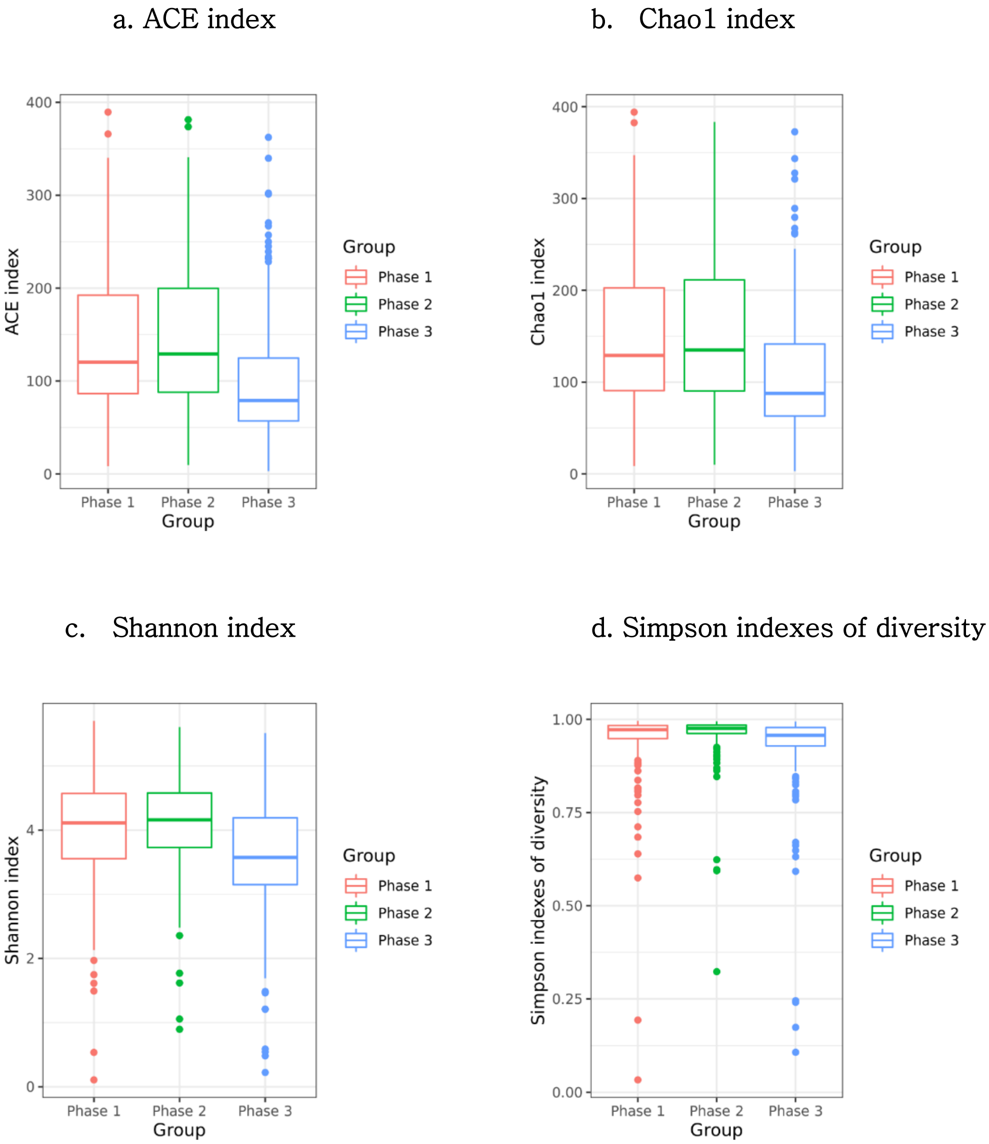
4.3. Results

Longitudinal changes in the urine microbial composition over 4 years

Alpha diversity of urine microbiome decreased during the follow-up period, which may have been an effect of aging (Figure 4.2). An Nonmetric Multi-Dimensional Scaling plot based on beta diversity also revealed a gradual change in the composition of microbiota according to age (Figure 4.3). The overall microbiome composition is presented in Figure 4.4 *Akkermansia muciniphila* was the most abundant taxon in all phases. Figure 4.5 shows that the abundance of two major genera, *Bacteriodes* and *Akkermansia,* decreased during the follow-up period.

T2D and other clinical traits explained by microbial variance

I investigated the associations between various clinical phenotypes and microbial composition using PERMANOVA (Figure 4.6). HbA1c, WBC, hematocrit, *binary_T2D*, and age in phase 1 significantly explained the changes in the microbial composition during the follow-up period (*p* = 0.0061, 0.0107, 0.0110, 0.0409, and 0.0290, respectively; FDR-adjusted *p* = 0.1027, 0.1027, 0.1027, 0.2290, and 0.2030, respectively). HbA1c and *binary_T2D* explained a certain amount of variance in the microbial changes in the study cohort during the 4 years, indicating that the longitudinal change in microbiome composition may be more closely associated with T2D-related phenotypes than with other clinical traits.



a. ACE index

b. Chao1 index

c. Shannon index

d. Simpson indexes of diversity

Figure 4.2. Box plots of alpha diversity indices for phases 1, 2, and 3.

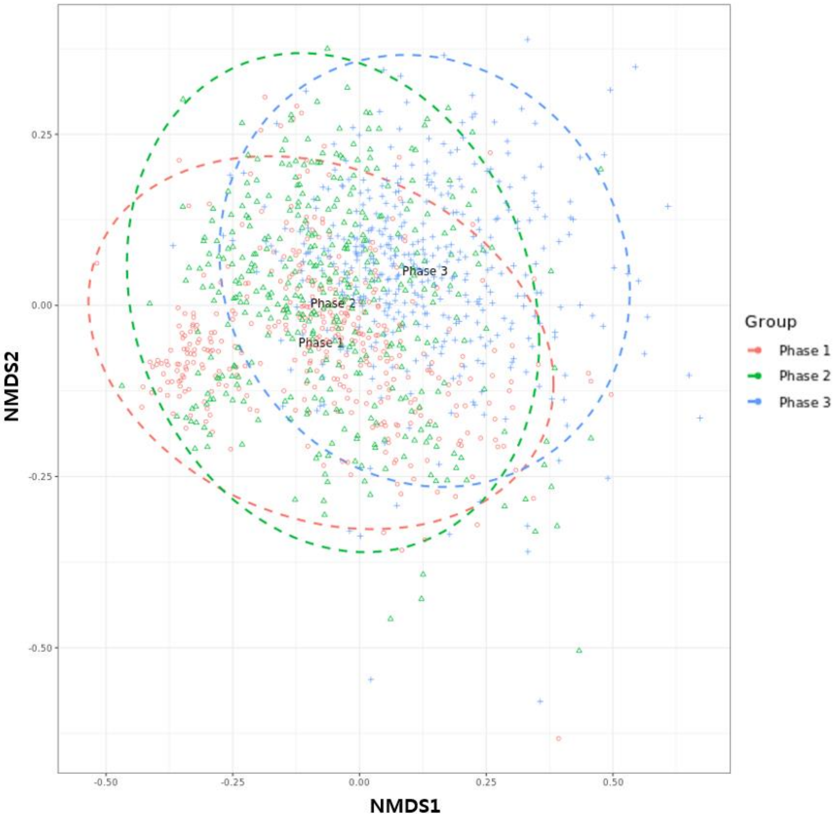
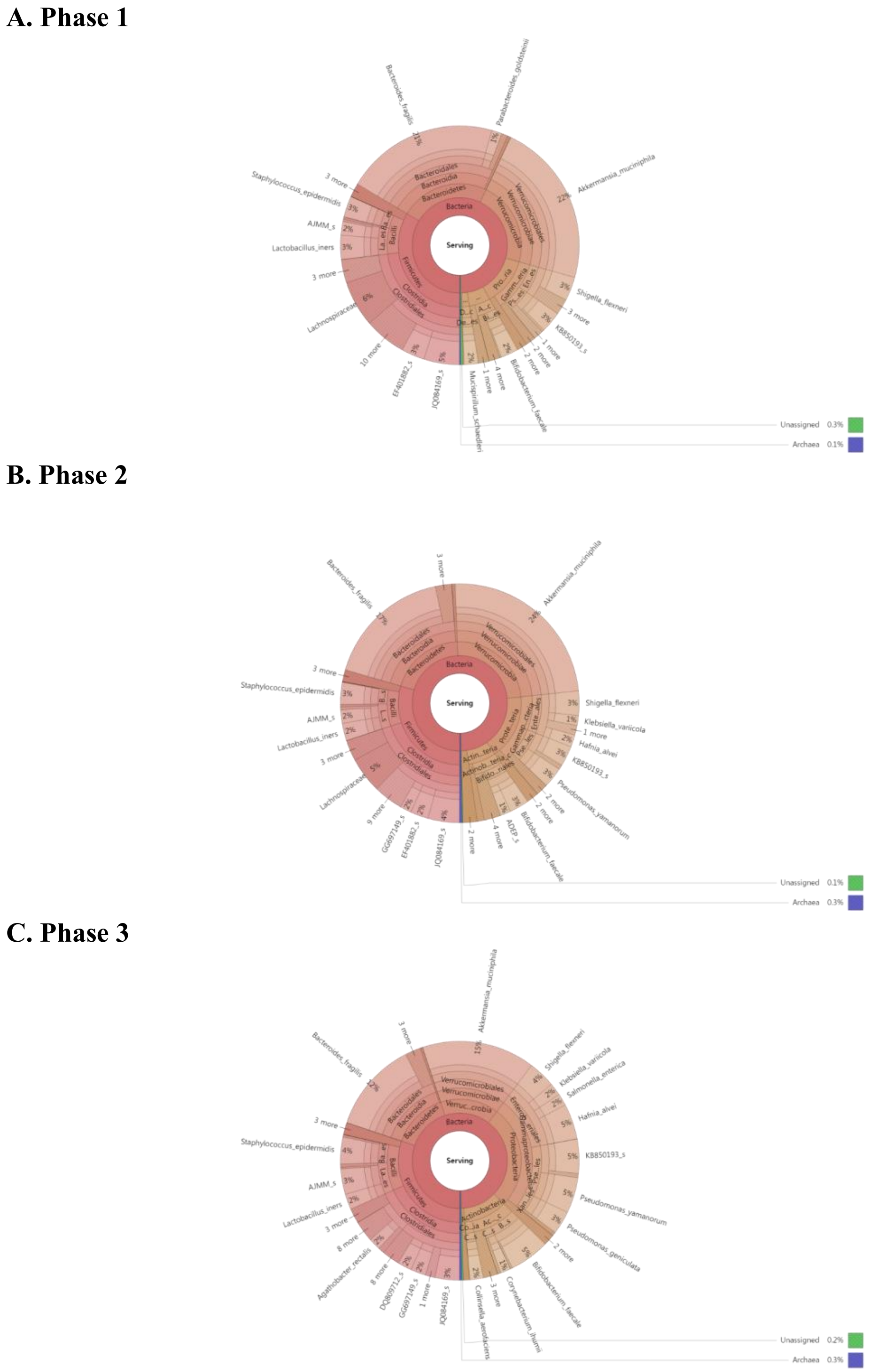


Figure 4.3. NMDS plot of Bray–Curtis beta diversities for phases 1, 2, and 3.

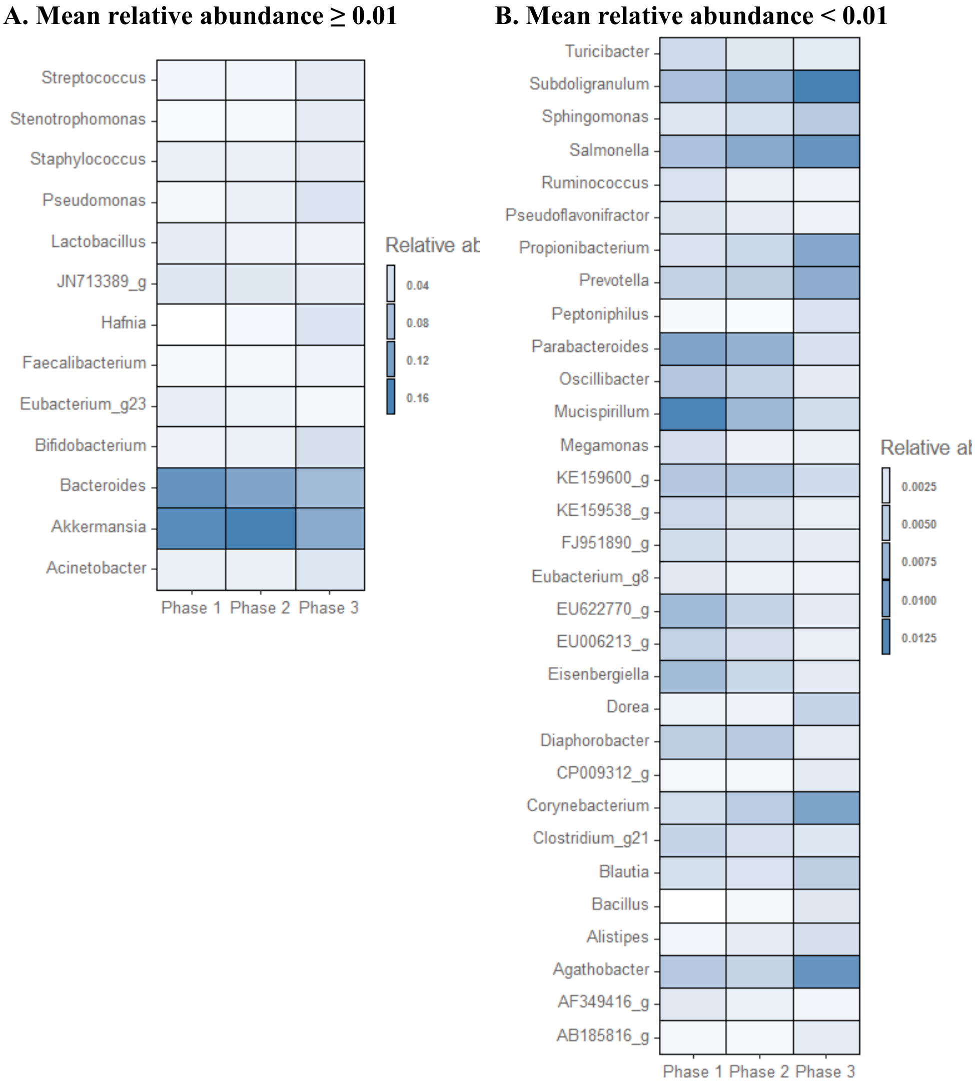


A. Phase 1

B. Phase 2

C. Phase 3

Figure 4.4. Taxonomic composition. The mean relative abundances of bacterial taxa at different taxonomic levels are shown with Krona plot for phase 1, 2, and 3.



A. Mean relative abundance ≥ 0.01

B. Mean relative abundance < 0.01

Figure 4.5. Mean relative abundances of genera in urine samples from healthy and T2D-at-risk subjects and T2D patients during a four-year follow-up (2013-2017). Genera whose relative abundance significantly decreased or increased ($p < 0.01$) are shown. p-values were calculated and are presented in Table 4.2. under the repeated measurement ANOVA model with compound symmetry correlation structure among the same subjects.

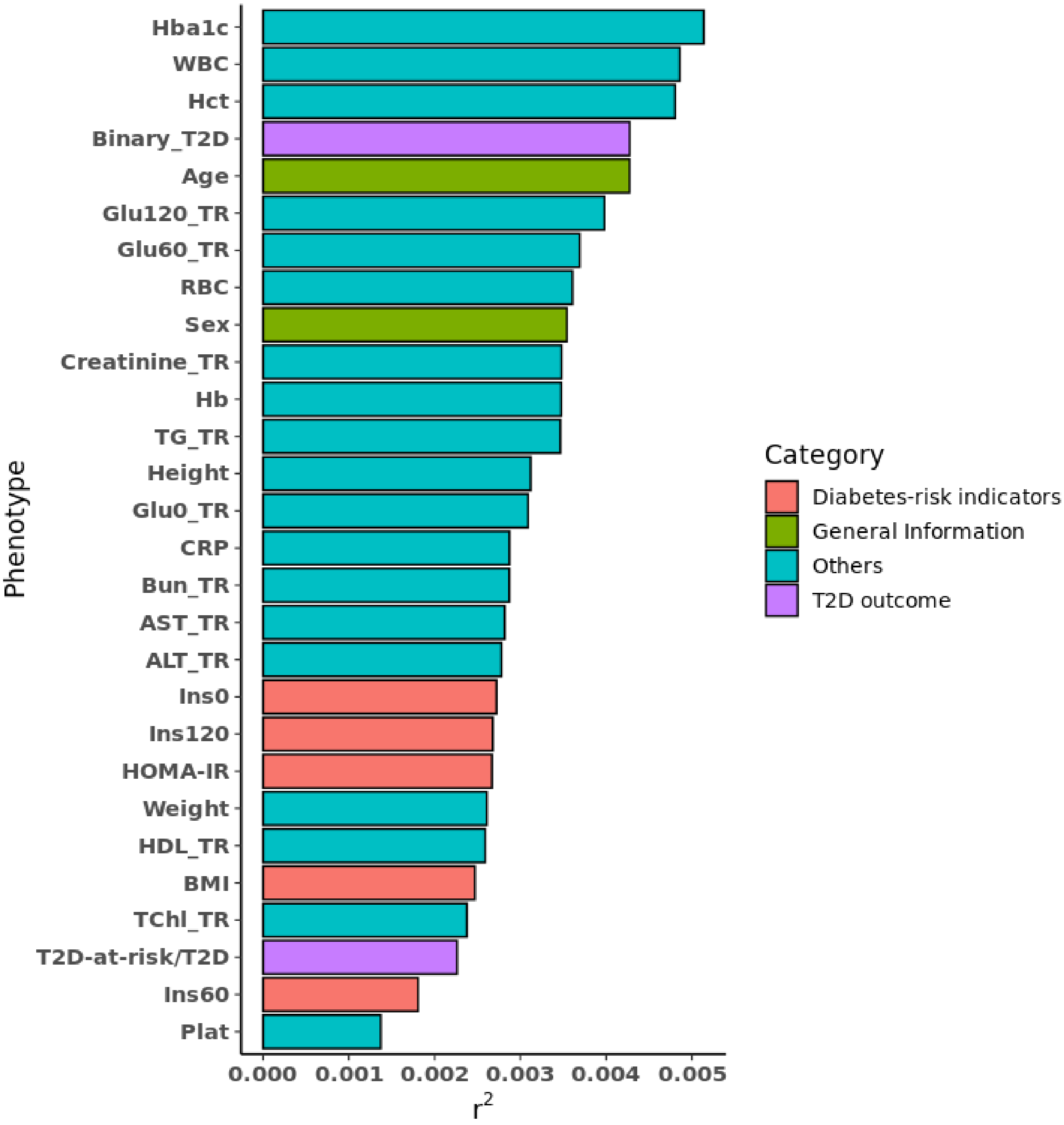


Figure 4.6. Relative importance of variables. Relative proportions of variance attributable to each variable were calculated with PERMANOVA using pldist based on Bray–Curtis beta-diversity. Every trait was categorized into four groups, including pre-defined diabetes-risk indicators, general information (age and sex), T2D outcomes (Binary_T2D and T2D-at-risk/T2D), and others.

Table 4.2. Microbial profiles and time effects in a repeated-measures ANOVA model

Genus	F-statistic	p-value	Phase 1	Phase 2	Phase 3
Hafnia	174.65565	0.00000	0.00001	0.01117	0.03280
Pseudomonas	162.59098	0.00000	0.00870	0.01858	0.03294
Bacteroides	132.10006	0.00000	0.13790	0.11525	0.08394
Mucispirillum	84.96168	0.00000	0.01282	0.00728	0.00345
Eubacterium_g23	84.65617	0.00000	0.02102	0.01629	0.00872
Parabacteroides	76.27374	0.00000	0.00930	0.00801	0.00291
EU622770_g	74.78402	0.00000	0.00701	0.00444	0.00214
Eisenbergiella	74.30869	0.00000	0.00695	0.00414	0.00213
Propionibacterium	72.51276	0.00000	0.00279	0.00393	0.00917
Bifidobacterium	72.35795	0.00000	0.01594	0.01785	0.03705
Stenotrophomonas	64.22521	0.00000	0.00608	0.00864	0.02140
Akkermansia	52.73753	0.00000	0.14830	0.16070	0.10442
Oscillibacter	51.21245	0.00000	0.00565	0.00441	0.00203
EU006213_g	49.79571	0.00000	0.00437	0.00301	0.00157
Streptococcus	45.45609	0.00000	0.01224	0.01224	0.02345
Bacillus	44.95760	0.00000	0.00005	0.00085	0.00231
Subdoligranulum	44.12874	0.00000	0.00630	0.00866	0.01310
Corynebacterium	40.86407	0.00000	0.00323	0.00510	0.00951
Faecalibacterium	39.78100	0.00000	0.00709	0.01036	0.01406
Dorea	36.51705	0.00000	0.00138	0.00126	0.00435
KE159538_g	36.11301	0.00000	0.00380	0.00277	0.00163
JN713389_g	29.40586	0.00000	0.03234	0.02888	0.02289
Salmonella	27.69458	0.00000	0.00615	0.00872	0.01110
Pseudoflavonifractor	26.63656	0.00000	0.00278	0.00180	0.00123
Lactobacillus	23.31283	0.00000	0.02182	0.01582	0.01435
Prevotella	21.74107	0.00000	0.00453	0.00504	0.00848
Diaphorobacter	21.42170	0.00000	0.00490	0.00536	0.00197
Acinetobacter	20.57114	0.00001	0.01888	0.01843	0.03170
Peptoniphilus	19.47782	0.00001	0.00068	0.00057	0.00283
Agathobacter	17.92011	0.00003	0.00547	0.00429	0.01108
AB185816_g	17.69532	0.00003	0.00080	0.00063	0.00196
AF349416_g	16.04020	0.00007	0.00219	0.00143	0.00101
Ruminococcus	15.89541	0.00007	0.00284	0.00151	0.00126
Alistipes	15.71696	0.00008	0.00104	0.00181	0.00300
Megamonas	14.32393	0.00017	0.00320	0.00141	0.00152
FJ951890_g	13.66239	0.00023	0.00345	0.00264	0.00191
Clostridium_g21	13.46607	0.00026	0.00436	0.00296	0.00255

Table 4.2. Continued

Genus	F-statistic	p-value	Phase 1	Phase 2	Phase 3
KE159600_g	12.42524	0.00045	0.00566	0.00582	0.00352
Turicibacter	12.12757	0.00052	0.00362	0.00235	0.00217
Blautia	10.26543	0.00141	0.00306	0.00272	0.00491
Sphingomonas	10.21799	0.00145	0.00259	0.00315	0.00540
CP009312_g	9.56734	0.00205	0.00074	0.00085	0.00201
Staphylococcus	8.43551	0.00378	0.01916	0.01962	0.02466
Eubacterium_g8	7.07873	0.00796	0.00222	0.00148	0.00135
Klebsiella	6.65326	0.01008	0.00819	0.00893	0.01147
Phascolarctobacterium	5.25604	0.02213	0.00156	0.00205	0.00080
GU174097_g	4.79592	0.02882	0.00295	0.00211	0.00210
Moraxella	4.45083	0.03520	0.00222	0.00429	0.00373
Collinsella	4.33469	0.03767	0.00963	0.00646	0.01270
Methanobrevibacter	3.53503	0.06046	0.00092	0.00205	0.00213
Actinomyces	3.26329	0.07123	0.00138	0.00286	0.00248
Cloacibacterium	3.02431	0.08242	0.00302	0.00300	0.00216
Micrococcus	2.13212	0.14464	0.00131	0.00118	0.00175
Enterococcus	2.08498	0.14915	0.00202	0.00224	0.00302
Fusobacterium	2.02367	0.15526	0.00100	0.00570	0.00206
Lactococcus	1.96258	0.16163	0.00203	0.00268	0.00147
Rhodococcus	1.93296	0.16483	0.00061	0.00124	0.00101
Anaerotruncus	1.79365	0.18087	0.00120	0.00134	0.00084
Romboutsia	1.77354	0.18333	0.00872	0.00711	0.00772
Haemophilus	1.59476	0.20702	0.00296	0.00223	0.00362
Paracoccus	1.56925	0.21069	0.00059	0.00154	0.00155
Megasphaera	1.46126	0.22709	0.00075	0.00171	0.00141
Escherichia	1.29997	0.25457	0.02143	0.02220	0.02491
Clostridium	1.10265	0.29401	0.00522	0.00738	0.00380
Unassigned	0.81100	0.36810	0.00186	0.00069	0.00131
Neisseria	0.41556	0.51935	0.00114	0.00090	0.00135
Ruminococcus_g2	0.34225	0.55870	0.00539	0.00434	0.00580
Gardnerella	0.11903	0.73018	0.00133	0.00950	0.00189
Veillonella	0.07176	0.78886	0.00105	0.00120	0.00100
EU842423_g	0.03733	0.84684	0.00122	0.00195	0.00130
Dialister	0.03148	0.85922	0.00153	0.00151	0.00147

A compound symmetry correlation structure among the same subjects was applied.

Taxa and functional profiles associated with T2D and diabetes-risk indicators

In an association analysis of 70 genera with *binary_T2D* and *T2D-at-risk/T2D* phenotypes, *GU174097_g*, an unclassified *Lachnospiraceae*, was found to exhibit a significant association with these phenotypes and was identified to be more abundant in healthy subjects than in diabetic and prediabetic patients (Table 4.3). Box plot and box plot of relative proportion in Figure 4.7 show that *GU174097_g* is slightly less observed in case group than control group in phase 1 and 3. In Figure 4.8, subjects were grouped according to their T2D status. *Healthy on Phase 1-3* group comprised subjects who were healthy in phases 1, 2, and 3; *T2D on Phase 1-3* group comprised subjects who had T2D in phases 1, 2, and 3; *Healthy to T2D-at-risk/T2D* group included the subjects who were healthy in phase 1 and became T2D patient or T2D-at-risk in phase 3. *T2D-at-risk/T2D to Healthy* group included subjects who were T2D-at-risk/T2D in phase 1 and healthy in phase 3. The relative abundance of GU174097_g of *Healthy to T2D-at-risk/T2D* subjects decreases as the development of T2D occurs. Conversely, Healthy to T2D-at-risk/T2D group had increased abundance when they become healthy. In summary, GU174097_g seems to move in clear association with the progression of diabetes over time, not simply whether or not diabetes.

To investigate T2D-associated microbial functional profiles, 238 functional profiles were evaluated, and significant associations at an FDR-adjusted significance of 0.1 are presented in Table 4.4. The *T2D-at-risk/T2D* phenotype was related to cationic antimicrobial peptide (CAMP), and the biosynthesis of fatty acids, coenzyme A (CoA), and secondary metabolites, and oxidative phosphorylation were significantly associated with the *Binary_T2D* phenotype at an FDR-adjusted significance of 0.1.

Next, I investigated the associations of log-transformed diabetes-risk indicators with genera, and significant associations at an FDR-adjusted significance of 0.1 were identified. Twelve, four, and twenty genera were significantly associated with HbA1c, glucose levels, and insulin levels, respectively. Particularly, *Hafnia* was associated with HbA1c and 60- and 120-minute insulin levels, and *AB185816_g* and *Akkermansia* were associated with HbA1c, fasting glucose, and 60-minute insulin levels (Table 4.5).

Table 4.3. Association analysis of T2D with bacterial genera

Phenotype	Genus	Estimate	StdErr	DF	p-value	FDR
T2D-at-risk/T2D	*GU174097_g*	−189.13	46.63	735	0.00006	0.00393
Binary_T2D	*JN713389_g*	−13.07	5.31	735	0.01411	0.38195
Binary_T2D	*Akkermansia*	−3.49	1.43	735	0.01489	0.38195
Binary_T2D	*Dialister*	−86.44	37.49	735	0.02140	0.38195
Binary_T2D	*Ruminococcus_g2*	−25.38	11.70	735	0.03039	0.38195
Binary_T2D	*KE159538_g*	−48.29	22.95	735	0.03568	0.38195
Binary_T2D	*Bifidobacterium*	6.71	3.21	735	0.03669	0.38195
Binary_T2D	*Eubacterium_g8*	−71.10	34.46	735	0.03944	0.38195
Binary_T2D	*Megamonas*	−65.20	32.53	735	0.04538	0.38195
Binary_T2D	*Pseudomonas*	7.74	3.91	735	0.04842	0.38195

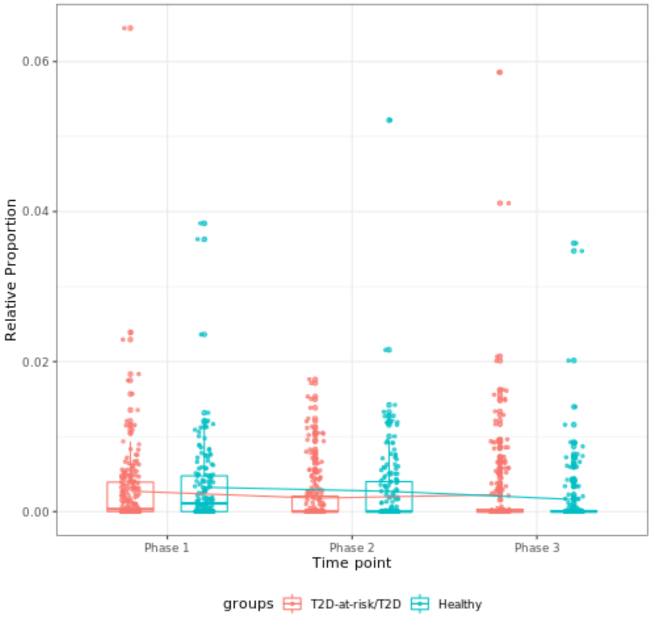


Figure 4.7. Relative proportions of GU714097_g in T2D-at-risk/T2D and Healthy groups.

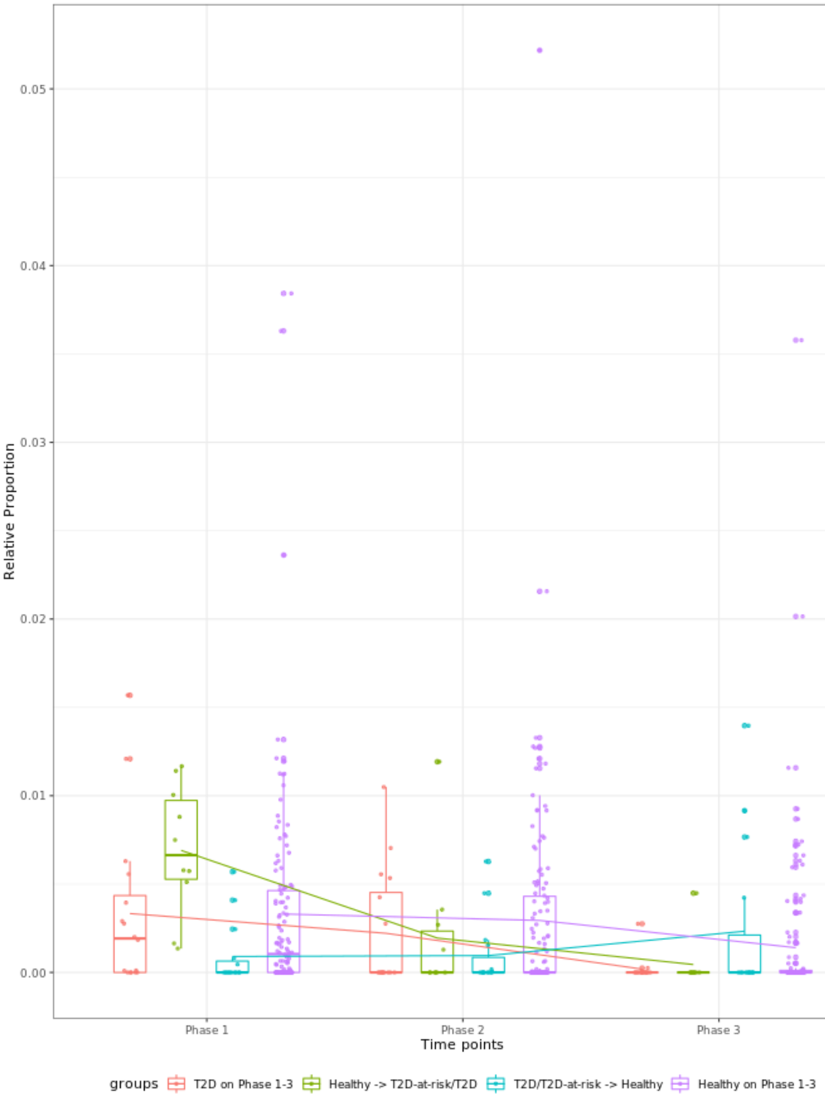


Figure 4.8. Change of relative proportions of *GU714097_g* in different T2D development groups. For each time point, mean relative proportions of GU174097_g are provided for each different T2D group.

Table 4.4. Association analysis of T2D with functional profiles predicted using tax4fun

Phenotype	KEGG Pathway	Estimate	StdErr	DF	p-value	FDR
T2D-at-risk/T2D	Cationic antimicrobial peptide (CAMP) resistance	−105.77	28.99	920	0.00028	0.03735
Binary_T2D	Bacterial secretion system	−104.43	34.02	920	0.00221	0.09027
Binary_T2D	Base excision repair	−236.06	79.69	920	0.00313	0.09027
Binary_T2D	Taurine and hypotaurine metabolism	378.99	129.15	920	0.00342	0.09027
Binary_T2D	Glycerophospholipid metabolism	−192.29	66.15	920	0.00374	0.09027
Binary_T2D	Pantothenate and CoA biosynthesis	−114.77	39.82	920	0.00404	0.09027
Binary_T2D	Fatty acid biosynthesis	75.86	27.60	920	0.00611	0.09528
Binary_T2D	beta-Lactam resistance	−65.95	24.46	920	0.00715	0.09528
Binary_T2D	Oxidative phosphorylation	−91.71	34.08	920	0.00726	0.09528
Binary_T2D	Biosynthesis of secondary metabolites	−22.08	8.32	920	0.00806	0.09528

Table 4.5. Association analysis of genera with diabetes-risk indicators

Phenotype	Genus	Estimate	p-value	FDR
Ins120	*Diaphorobacter*	8.7235	4.32E-06	0.000302
Hba1c	*Fusobacterium*	-0.6144	5.30E-06	0.000371
Hba1c	*Gardnerella*	-0.2588	3.31E-05	0.000772
Hba1c	*Hafnia*	0.1691	2.59E-05	0.000772
Hba1c	*Akkermansia*	-0.08445	4.69E-05	0.000822
Ins0	*Bacteroides*	-0.563	1.89E-05	0.001325
Glu60	*Faecalibacterium*	1.7566	2.37E-05	0.001351
Ins120	*Mucispirillum*	4.6197	4.77E-05	0.001668
Glu0	*AB185816_g*	1.9868	3.25E-05	0.002274
Ins60	*Paracoccus*	-3.1448	0.000117	0.008211
Hba1c	*AB185816_g*	1.184	0.000591	0.008271
Ins60	*Akkermansia*	0.8599	0.000644	0.014056
Ins60	*Bacillus*	-10.4746	0.000525	0.014056
Ins60	*Hafnia*	-2.1636	0.000803	0.014056
Ins60	*Eubacterium_g23*	3.3619	0.001052	0.014733
Glu0	*Gardnerella*	-0.2234	0.000428	0.014985
Hba1c	*KE159538_g*	-0.8552	0.001459	0.017027
Ins120	*Hafnia*	-2.996	0.000824	0.019217
Ins60	*Bifidobacterium*	-2.2198	0.001773	0.020682
Ins60	*AF349416_g*	11.2469	0.003463	0.030304
Ins60	*Clostridium*	2.7748	0.003281	0.030304
Ins60	*EU622770_g*	6.3967	0.005	0.03889
Ins60	*Subdoligranulum*	-4.5438	0.005617	0.039319
Glu0	*Akkermansia*	-0.08067	0.001739	0.040583
Ins60	*AB185816_g*	-13.129	0.008131	0.047429
Ins60	*Parabacteroides*	4.5801	0.008032	0.047429
Hba1c	*Collinsella*	0.182	0.005517	0.048343
Hba1c	*Pseudoflavonifractor*	-0.7954	0.005525	0.048343
Ins0	*Acinetobacter*	0.532	0.001467	0.050619
Ins0	*Fusobacterium*	2.5293	0.002169	0.050619
Hba1c	*Prevotella*	0.3263	0.008377	0.065156
Hba1c	*Agathobacter*	0.1971	0.014171	0.082664
Hba1c	*Bifidobacterium*	0.1111	0.012478	0.082664
Hba1c	*Faecalibacterium*	0.2225	0.013039	0.082664
Ins120	*Oscillibacter*	7.0022	0.004943	0.086495
Ins60	*Methanobrevibacter*	-3.4821	0.017655	0.095066

Associations of T2D-related unclassified *Lachnospiraceae* with diabetes-risk indicators

To confirm the association of *GU174097_g* with T2D, I performed an extensive validation analysis using clinical data. I first conducted an association analysis between *GU174097_g* and clinical variables to investigate the correlations among *GU174097_g* and diabetes-risk indicators (Table 4.6). *GU174097_g* was significantly and positively associated with the 60-minute insulin level among all glucose- and insulin-related variables.

Second, I established an association network of the diabetes-risk indicators. An association network of diabetes-risk indicators is important because the same observed correlations can imply completely different biological processes. For example, if high levels of glucose or HbA1c tend to appear with high levels of insulin, insulin resistance may be present. However, if high levels of glucose or HbA1c tend to appear with low levels of insulin, insulin secretion may have reduced the glucose or HbA1c levels. Network analysis indicated strong associations among the diabetes-risk indicators (Figure 4.9). Particularly, the 60-minute insulin level exhibited a strong negative correlation with the HbA1c level, suggesting that the former can decrease the latter.

Lastly, I investigated the association of diabetes-risk indicators and the progression of T2D. *Healthy to T2D-at-risk/T2D group* showed increase of 60 min glucose and insulin level as T2D progress. Fasting glucose and insulin were more associated with the cross-sectional status of T2D not the progression of T2D (Figure 4.10).

Table 4.6. Association analysis of T2D-related phenotypes with *GU174097_g*

Phenotype	Rho	p-value	FDR
Ins60	0.0950	0.0018	0.0026
HbA1c	−0.0434	0.1548	0.1872
Glu0	0.0306	0.3165	0.3647
Ins0	−0.0301	0.3236	0.3721
HOMA-IR	−0.0270	0.3756	0.4255
Glu60	−0.0229	0.4536	0.5045
Ins120	0.0177	0.5622	0.6096
Glu120	0.0072	0.8138	0.8424
BMI	−0.0070	0.8186	0.8469

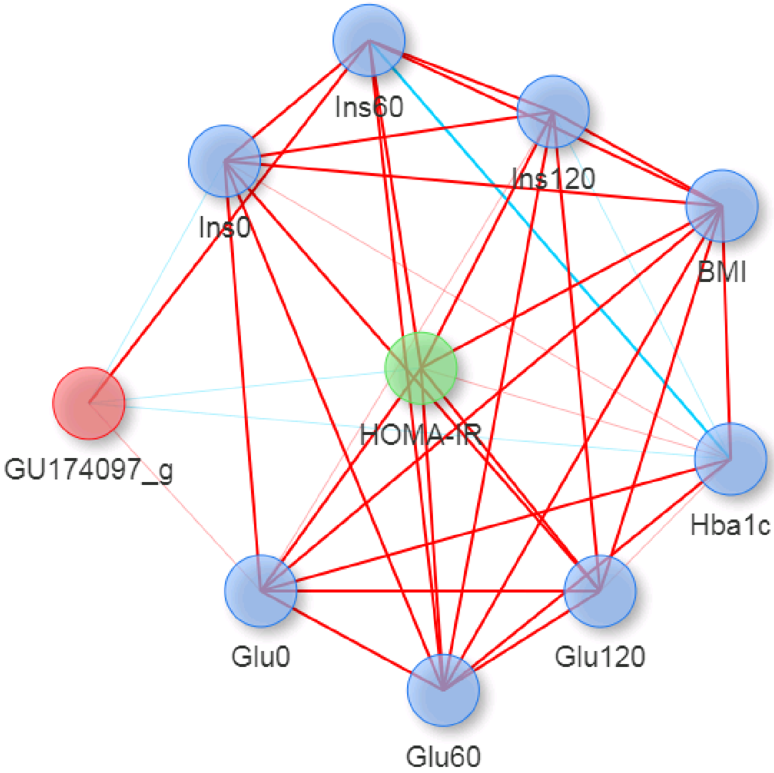


Figure 4.9. Network of GU174097_g and KARE phenotypes. When FDR < 0.05, edge width is in bold; otherwise, edge width is not included. When rho is positive, edges are colored red; otherwise, they are colored blue. Blue, green, and red nodes represent KARE phenotypes, Homeostatic Model Assessment for Insulin Resistance and *GU174097_g*, respectively.

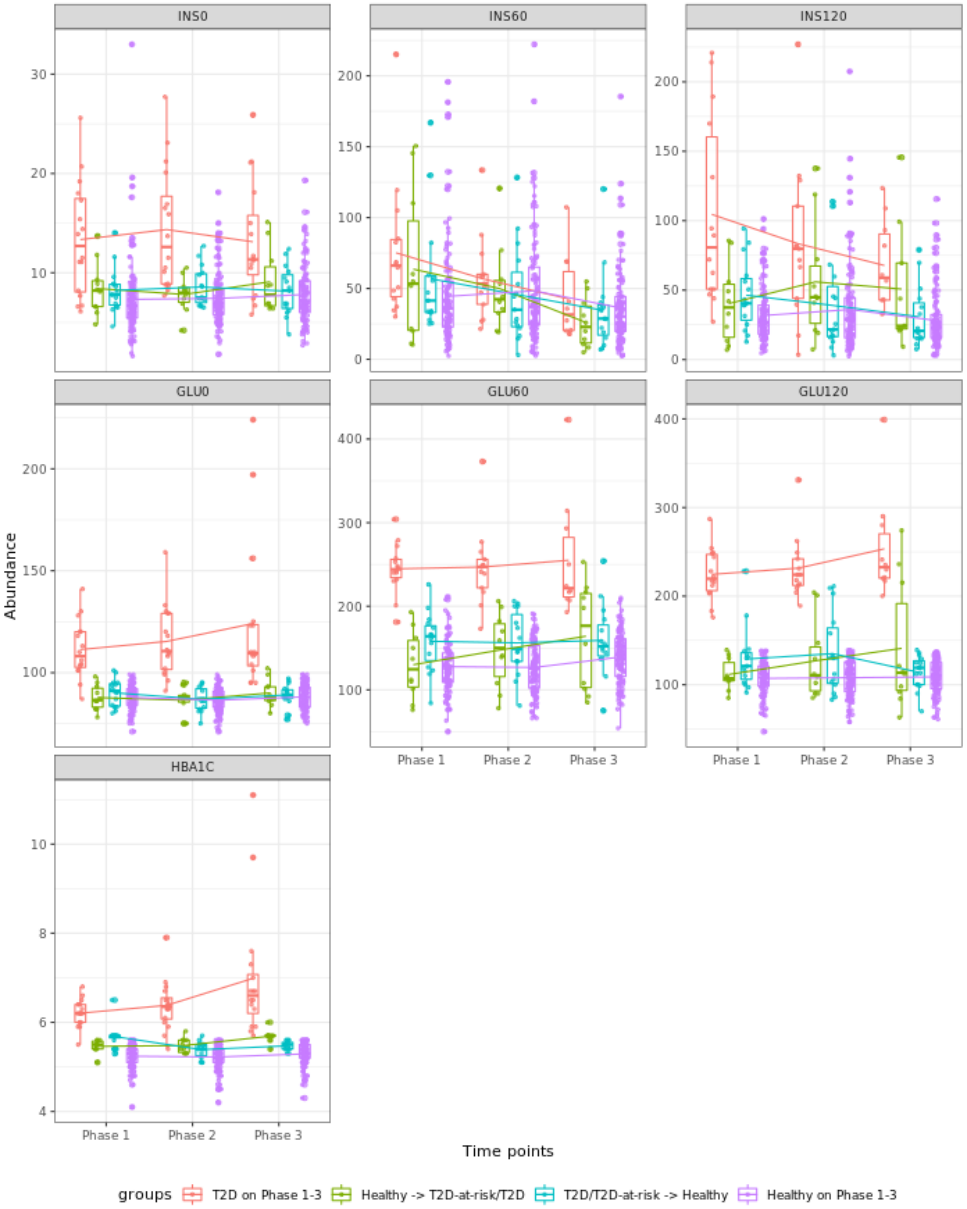


Figure 4.10. Abundance of T2D risk indicators in different T2D development groups. For each time point, boxplot of the abundance are provided for each different T2D group. The line indicates the mean abundance values.

Causal relationship between the T2D-related taxon and diabetes-risk indicators

To verify whether a causal relationship existed between the abundance of *GU174097_g* on diabetes-risk indicators, a two-sample MR analysis was performed. Extensive assumption checks were conducted to enhance the validity of two-sample MR analysis (Table 4.7). No weak instrument bias was observed (*F*-statistic > 10). However, NOME assumptions were violated for all tests because *GU174097_g* had 7 SNPs as their instrument variables, and this value was not sufficiently large for $I^2 > 90$. In this case, if heterogeneity exists, MR-Egger (SIMEX) is recommended; otherwise, IVW is recommended. Because the InSIDE assumption cannot be statistically tested [93], the weighted median method—a robust method used in case of violation of InSIDE assumption—has to be considered with each recommended method [92]. Therefore, the IVW method was used to estimate all causal effects. There was no significant causal association at an FDR-adjusted significance of 0.05.

Table 4.7. Statistical analysis to check the assumption required for two-sample Mendelian randomization

Outcome	*GU174097_g* N=7 F=10.45				
	I^2	Q	Q´	RSS	Suggested MR method
HbA1c	0*	0.533	0.644	0.651	IVW
Glu0	0*	0.372	0.181	0.183	IVW
Glu60	0*	0.660	0.686	0.694	IVW
Glu120	0*	0.355	0.457	0.473	IVW
Ins0	0*	0.542	0.668	0.666	IVW
Ins60	0*	0.117	0.174	0.184	IVW
Ins120	0*	0.463	0.257	0.271	IVW
BMI	0*	0.758	0.840	0.846	IVW

Q, heterogeneity test from IVW; *Q*´, *P*-value for Cochran's Q test; *RSS*, *P* for MR-PRESSO global test, *F*, mean F statistic; I^2, I^2 value from MR-Egger. *F*-test checks weak instrument bias, I^2 checks the NOME assumption and *Q* and *Q*´, and *RSS* checks the heterogeneity assumption.

Table 4.8. Two-sample Mendelian randomization causal effect.

Outcome	MR methods	*GU174097_g*	
		Estimate (95% CI)	FDR
HbA1c	Weighted Median	0.632 (−1.7011, 2.965)	0.595
	MR-Egger	−0.997 (−7.212, 5.2179)	0.753
	MR-Egger (SIMEX)	0.7775 (−1.0014, 2.5564)	0.431
	IVW	**0.0695 (−1.7213, 1.8604)**	**0.939**
	MR-PRESSO	0.0695 (−1.436, 1.5751)	0.931
Glu0	Weighted Median	−0.0985 (−3.5584, 3.3613)	0.955
	MR−Egger	−7.8089 (−16.9558, 1.338)	0.141
	MR−Egger (SIMEX)	2.4721 (−0.1467, 5.0909)	0.185
	IVW	**0.2604 (−2.8322, 3.353)**	**0.869**
	MR-PRESSO	0.2604 (−2.8322, 3.353)	0.874
Glu60	Weighted Median	−2.7662 (−8.5932, 3.0608)	0.919
	MR-Egger	−8.4801 (−24.0515, 7.0914)	0.658
	MR-Egger (SIMEX)	−1.5038 (−6.3334, 3.3259)	0.708
	IVW	**−2.2526 (−6.749, 2.2438)**	**0.530**
	MR−PRESSO	−2.2526 (−5.8922, 1.387)	0.471
Glu120	Weighted Median	−3.1981 (−9.4609, 3.0647)	0.490
	MR-Egger	−7.5848 (−24.8455, 9.6759)	0.455
	MR-Egger (SIMEX)	−3.7682 (−10.031, 2.4945)	0.437
	IVW	**−4.2316 (−8.9665, 0.5034)**	**0.240**
	MR-PRESSO	−4.2316 (−8.8492, 0.3861)	0.368
Ins0	Weighted Median	0.0809 (−6.646, 6.8077)	0.981
	MR-Egger	0.1588 (−17.656, 17.9735)	0.986
	MR-Egger (SIMEX)	3.0289 (−2.0456, 8.1034)	0.442
	IVW	**1.1125 (−4.0195, 6.2444)**	**0.829**
	MR-PRESSO	1.1125 (−3.1119, 5.3368)	0.780
Ins60	Weighted Median	7.518 (−6.0625, 21.0984)	0.468
	MR-Egger	10.0106 (−33.5852, 53.6064)	0.787
	MR-Egger (SIMEX)	8.2196 (−5.8634, 22.3025)	0.457
	IVW	**3.2629 (−8.3365, 14.8623)**	**0.859**
	MR-PRESSO	3.2629 (−8.3365, 14.8623)	0.807
Ins120	Weighted Median	−11.4001 (−26.8887, 4.0885)	0.447
	MR-Egger	24.9815 (−13.0067, 62.9696)	0.399
	MR-Egger (SIMEX)	−11.541 (−27.7749, 4.6929)	0.324
	IVW	**−7.7912 (−20.2658, 4.6835)**	**0.663**
	MR-PRESSO	−7.7912 (−20.2658, 4.6835)	0.551
BMI	Weighted Median	0.5582 (−1.772, 2.8884)	0.639
	MR-Egger	−1.1856 (−7.5103, 5.139)	0.890
	MR-Egger (SIMEX)	0.4651 (−0.9768, 1.907)	0.833
	IVW	**−0.1097 (−1.9344, 1.715)**	**0.906**
	MR-PRESSO	−0.1097 (−1.3438, 1.1245)	0.867

The recommended MR method is highlighted in bold letters.

4.4. Discussion

Recent microbiome studies have shown that T2D is associated with gut microbial dysbiosis [94-96], which can result in altered intestinal barrier functions and remodeled host metabolic and signaling pathways [97]. Intestinal bacteria can affect insulin resistance by triggering inflammation via lipid polysaccharides, which are a component of gram-negative bacterial cell walls [98]. In addition, microbiota-derived EVs are expected to affect insulin resistance and may help understand the pathogenesis of T2D [82]. Various bacterial metabolites, such as short-chain fatty acids (SCFAs), can modulate the functioning of multiple signaling pathways involved in the maintenance of human health and can protect against insulin resistance [98, 99].

The human microbiota is highly variable, and this variability can be explained by the effect of various external factors, such as diet, exercise, mobility, medication, and cohabitation patterns. Many of these external factors affect the risk of developing metabolic diseases and are age-related [100]. In other words, the intestinal microbiota and the host phenotype alter substantially with aging, and the effect of the intestinal microbiota on the host phenotype is dependent on the age of the host. The estimation of within-subject covariate effects is robust against between-subject confounders, and longitudinally measured microbiome data enable the identification of microbiota effects on the risk of diseases in the host. As most existing studies are cross-sectional in nature, the validity and interpretation of their results are limited. Therefore, longitudinal studies are needed to investigate the association between the human microbiome and diseases.

My longitudinal study revealed that a low abundance of *GU174097_g* can be a risk factor for T2D development. To date, *GU174097_g* has not been cultured. Multi-omics data, including host genomic data, T2D-related metabolites, clinical information, and predicted functional meta-genomic profiles, were utilized to extensively validate my results. *GU174097_*g is a member of the family *Lachnospiraceae*, and the association between *Lachnospiraceae* and T2D risk has been reported in several studies [101, 102]. SCFA pathways, including the propanediol and acrylate pathways in *Lachnospiraceae* play important roles in mediating the effects of *Lachnospiraceae* on T2D, and microbial organisms producing SCFAs affect epigenetic regulation in T2D patients and reduce the risk of developing T2D [99, 103]. I found that *GU174097_g* is positively correlated with the 60-minute insulin level, which in turn is negatively correlated with the HbA1c level. This indicates that *GU174097_g* reduces the HbA1c level and thus, the risk of developing T2D, by stimulating insulin secretion.

Next, I aimed to elucidate how *GU174097_g* affects T2D through 60-minute insulin and HbA1c. Multiple mechanisms for these associations, including various metabolites produced by the microbiome, such as SCFAs, have been previously suggested. [103, 104].

Interestingly, *Coprococcus*, a member of the *Lachnospiraceae* family, is one of the major butyrate-producing bacteria. It uses metabolic intermediates essential to produce butyrate, a type of SCFA. SCFAs are considered to be beneficial for health and to protect against T2D [105]. Thus, I hypothesize that *GU174097_g* produces SCFAs, which can increase insulin secretion and decrease the HbA1c level, ultimately reducing the risk of developing T2D.

This study had several limitations. First, as it was based on the metagenomic profiles of EVs, the microbial compositions observed can differ from—and need to be further compared with—those of the intestinal microbiota. Second, as the genus-level taxonomy of *GU174097_g* is unknown, ecological and biological information is limited. Third, the causal relationship found in the discovery MR analysis failed to replicate. The published summary statistics of microbial GWAS are limited especially for EV, and the sample size in the microbial GWAS in this study was small. Therefore, the number of SNPs used as instrument variables in the MR analysis was small. Future studies should include a large sample size to identify more associated SNPs and increase the power of MR analysis. In this way, more mechanisms underlying T2D pathogenesis would be identified. Fourth, although extensive methods were used to validate the assumptions in the MR analysis and enhance the validity of the causal analysis, the MR results were not easy to interpret because diabetes-risk indicators are highly correlated and interact with each other. Additional *in-vivo* and *in-vitro* experiments may clarify the associations identified in this study.

Conclusions

This study revealed that *GU174097_g*, an unclassified *Lachnospiraceae*, is associated with T2D. This findings indicate that *GU174097_g* is associated with the progression of T2D and may lower the risk of developing T2D. Although the mechanism by which *GU174097_g* affect T2D development has not been elucidated, the results suggest that large-scale longitudinal studies and *in-vivo* and *in-vitro* experiments should be employed to unravel the underlying biological mechanisms.

Chapter 5. Conclusions

The statistical method TMAT integrates and normalized abundances of microorganisms based on phylogenetic tree information and hence corrected zero-inflated problems. Compositional bias is also handled because the abundance is based on a proportion between two OTUs. I developed AMAA package for microbiome association analysis that include the procedures of making microbial count table with different clustering methods and databases, unifying the preprocessing steps for various microbiome association test statistics, conducting metagenome-wide association analysis and comparison of the results. However, the comparison of the results are limited because it is hard to identify OTUs or ASVs based on different dataset, clustering method and databases. In further research, more sophisticated strategy to combine consensus sequences from different microbial clusters need to be applied.

The statistical method mTMAT implements TMAT and it also corrected zero-inflated problems and compositional bias with integration based on phylogenetic tree and taking a proportion between two OTUs. mTMAT also considers correlatation between repeatedly measured samples and developed for longitudinal microbiome analysis. The performance of the statistical method can be influenced by the simulation setup, and further validation of mTMAT can be improved in simulation setups with different properties.

Longitudinal analysis to find type-2 diabetes-associated microorganisms was conducted and provided evidences including estimated functional genes, network analysis, analysis of metabolomics, GWAS, mendelian randomization analysis and meta-analysis. These findings indicate that *GU174097_g* may lower the risk of developing T2D. Large-scale longitudinal studies and *in*-*vivo* and *in*-*vitro* experiments should be employed to unravel the underlying biological mechanisms.

References

1. Schloss, P.D. and J. Handelsman, *Metagenomics for studying unculturable microorganisms: cutting the Gordian knot.* Genome Biology, 2005. **6**(8): p. 1-4.
2. Knight, R., et al., *Unlocking the potential of metagenomics through replicated experimental design.* Nature Biotechnology, 2012. **30**(6): p. 513-520.
3. Huttenhower, C., et al., *Structure, function and diversity of the healthy human microbiome.* Nature, 2012. **486**(7402): p. 207.
4. Friedman, J. and E.J. Alm, *Inferring correlation networks from genomic survey data.* PLoS Comput Biol, 2012. **8**(9): p. e1002687.
5. Aitchison, J. and J.J. Egozcue, *Compositional data analysis: where are we and where should we be heading?* Mathematical Geology, 2005. **37**(7): p. 829-850.
6. Videvall, E., et al., *Direct PCR offers a fast and reliable alternative to conventional DNA isolation methods for gut microbiomes.* Msystems, 2017. **2**(6).
7. Harrison, J., et al., *The quest for absolute abundance: the use of internal standards for DNA-barcoding in microbial ecology.* 2020.
8. Vandeputte, D., et al., *Practical considerations for large-scale gut microbiome studies.* FEMS Microbiology Reviews, 2017. **41**(Supplement_1): p. S154-S167.
9. Beiko, R.G., *Microbial malaise: how can we classify the microbiome?* Trends in Microbiology, 2015. **23**(11): p. 671-679.
10. Chen, E.Z. and H. Li, *A two-part mixed-effects model for analyzing longitudinal microbiome compositional data.* Bioinformatics, 2016. **32**(17): p. 2611-2617.
11. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nature Methods, 2010. **7**(5): p. 335-336.
12. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Applied and Environmental Microbiology, 2009. **75**(23): p. 7537-7541.
13. Chen, J., et al., *Associating microbiome composition with environmental covariates using generalized UniFrac distances.* Bioinformatics, 2012. **28**(16): p. 2106-2113.
14. Wu, C., et al., *An adaptive association test for microbiome data.* Genome Medicine, 2016. **8**(1): p. 56.
15. Koh, H., M.J. Blaser, and H. Li, *A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping.* Microbiome, 2017. **5**(1): p. 45.
16. Zhao, N., et al., *Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test.* The American Journal of Human Genetics, 2015. **96**(5): p. 797-807.

17. Xia, Y., J. Sun, and D.-G. Chen, *Introductory overview of statistical analysis of microbiome data*, in *Statistical Analysis of Microbiome Data with R*. 2018, Springer. p. 43-75.

18. Koh, H., et al., *A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies.* Frontiers in Genetics, 2019. **10**: p. 458.

19. Tsilimigras, M.C. and A.A. Fodor, *Compositional data analysis of the microbiome: fundamentals, tools, and challenges.* Annals of Epidemiology, 2016. **26**(5): p. 330-335.

20. Navas-Molina, J.A., et al., *Advancing our understanding of the human microbiome using QIIME*, in *Methods in enzymology*. 2013, Elsevier. p. 371-444.

21. Rideout, J.R., et al., *Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences.* PeerJ, 2014. **2**: p. e545.

22. Mahé, F., et al., *Swarm v2: highly-scalable and high-resolution amplicon clustering.* PeerJ, 2015. **3**: p. e1420.

23. Callahan, B.J., P.J. McMurdie, and S.P. Holmes, *Exact sequence variants should replace operational taxonomic units in marker-gene data analysis.* The ISME Journal, 2017. **11**(12): p. 2639-2643.

24. Callahan, B.J., et al., *DADA2: high-resolution sample inference from Illumina amplicon data.* Nature Methods, 2016. **13**(7): p. 581-583.

25. Balvočiūtė, M. and D.H. Huson, *SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?* BMC Genomics, 2017. **18**(2): p. 114.

26. Park, S.-C. and S. Won, *Evaluation of 16S rRNA databases for taxonomic assignments using mock community.* Genomics & Informatics, 2018. **16**(4): p. e24.

27. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society: Series B (Methodological), 1995. **57**(1): p. 289-300.

28. Consortium, H.M.J.R.S., *A catalog of reference genomes from the human microbiome.* Science, 2010. **328**(5981): p. 994-999.

29. Zhang, X., et al., *The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment.* Nature Medicine, 2015. **21**(8): p. 895-905.

30. Morgan, X.C., et al., *Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.* Genome Biology, 2012. **13**(9): p. R79.

31. Janda, J.M. and S.L. Abbott, *The genus Aeromonas: taxonomy, pathogenicity, and infection.* Clinical Microbiology Reviews, 2010. **23**(1): p. 35-73.

32. Law, C.W., et al., *Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biology, 2014. **15**(2): p. R29.

33. Beasley, T.M., S. Erickson, and D.B. Allison, *Rank-based inverse*

normal transformations are increasingly used, but are they merited? Behavior Genetics, 2009. **39**(5): p. 580-595.

34. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Research, 2012. **41**(D1): p. D590-D596.

35. Chun, J., et al., *EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences.* International Journal of Systematic and Evolutionary Microbiology, 2007. **57**(10): p. 2259-2261.

36. Pruesse, E., J. Peplies, and F.O. Glöckner, *SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes.* Bioinformatics, 2012. **28**(14): p. 1823-1829.

37. Li, K., M. Bihan, and B.A. Methé, *Analyses of the stability and core taxonomic memberships of the human microbiome.* PloS One, 2013. **8**(5): p. e63139.

38. Zeller, G., et al., *Potential of fecal microbiota for early-stage detection of colorectal cancer.* Molecular Systems Biology, 2014. **10**(11): p. 766.

39. Kwon, S., B. Lee, and S. Yoon. *CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing*. in *BMC Bioinformatics*. 2014. BioMed Central.

40. Giloteaux, L., et al., *Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome.* Microbiome, 2016. **4**(1): p. 30.

41. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. Journal, 2011. **17**(1): p. 10-12.

42. Kwon, S., B. Lee, and S. Yoon. *CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing*. in *BMC Bioinformatics*. 2014. BioMed Central.

43. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics.* PeerJ, 2016. **4**: p. e2584.

44. Yilmaz, P., et al., *The SILVA and "all-species living tree project (LTP)" taxonomic frameworks.* Nucleic Acids Research, 2014. **42**(D1): p. D643-D648.

45. McDonald, D., et al., *An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.* The ISME Journal, 2012. **6**(3): p. 610-618.

46. Yoon, S.-H., et al., *Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies.* International Journal of Systematic and Evolutionary Microbiology, 2017. **67**(5): p. 1613.

47. Herdin, M., et al. *Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels*. in *2005 IEEE 61st Vehicular Technology Conference*. 2005. IEEE.

48. Cleveland, W.S., *LOWESS: A program for smoothing scatterplots by robust locally weighted regression.* American Statistician, 1981. **35**(1): p. 54.

49. Pielou, E.C., *An introduction to mathematical ecology.* An Introduction

to Mathematical Ecology., 1969.

50. Guo, S., et al., *A Simple and Novel Fecal Biomarker for Colorectal Cancer: Ratio of Fusobacterium Nucleatum to Probiotics Populations, Based on Their Antagonistic Effect.* Clinical Chemistry, 2018: p. clinchem. 2018.289728.

51. Seelam, N.S., et al., *Production, characterization and optimization of fermented tomato and carrot juices by using Lysinibacillus sphaericus isolate.* Journal of Applied Biology & Biotechnology Vol, 2017. **5**(04): p. 066-075.

52. Kim, K.J., et al., *Phylogenetic tree-based microbiome association test.* Bioinformatics, 2020. **36**(4): p. 1000-1006.

53. Dethlefsen, L., M. McFall-Ngai, and D.A. Relman, *An ecological and evolutionary perspective on human–microbe mutualism and disease.* Nature, 2007. **449**(7164): p. 811.

54. Fan, Y. and O. Pedersen, *Gut microbiota in human metabolic health and disease.* Nature Reviews Microbiology, 2020: p. 1-17.

55. Johnson, E.L., et al., *Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes.* Journal of Molecular Medicine, 2017. **95**(1): p. 1-8.

56. Sanz, Y., et al., *Understanding the role of gut microbiome in metabolic disease risk.* Pediatric Research, 2015. **77**(1): p. 236-244.

57. VanderWeele, T.J., J.W. Jackson, and S. Li, *Causal inference and longitudinal data: a case study of religion and mental health.* Social Psychiatry and Psychiatric Epidemiology, 2016. **51**(11): p. 1457-1466.

58. Xia, Y., J. Sun, and D.-G. Chen, *Introductory overview of statistical analysis of microbiome data.* Statistical Analysis of Microbiome Data with R, 2018: p. 43-75.

59. Zhang, X., et al., *Negative binomial mixed models for analyzing longitudinal microbiome data.* Frontiers in Microbiology, 2018. **9**: p. 1683.

60. Zhang, X. and N. Yi, *Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data.* Bioinformatics, 2020. **36**(8): p. 2345-2351.

61. Zhan, X., et al., *A small-sample kernel association test for correlated data with application to microbiome association studies.* Genetic Epidemiology, 2018. **42**(8): p. 772-782.

62. Boos, D.D., *On generalized score tests.* The American Statistician, 1992. **46**(4): p. 327-333.

63. Mancl, L.A. and T.A. DeRouen, *A covariance estimator for GEE with improved small-sample properties.* Biometrics, 2001. **57**(1): p. 126-134.

64. Ristl, R., et al., *Simultaneous inference for multiple marginal generalized estimating equation models.* Statistical Methods in Medical Research, 2020. **29**(6): p. 1746-1762.

65. Wang, M., *Generalized estimating equations in longitudinal data analysis: a review and recent developments.* Advances in Statistics, 2014. **2014**.

66. McCullagh, P. and J.A. Nelder, *Generalized linear models*. 2019: Routledge.

67. Kim, Y., B.-G. Han, and K. Group, *Cohort profile: the Korean genome and epidemiology study (KoGES) consortium.* International Journal of Epidemiology, 2017. **46**(2): p. e20-e20.

68. Bokulich, N.A., et al., *Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.* Nature Methods, 2013. **10**(1): p. 57-59.

69. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-2461.

70. Li, K., M. Bihan, and B.A. Methé, *Analyses of the stability and core taxonomic memberships of the human microbiome.* PloS one, 2013. **8**(5).

71. Williams, J., et al., *microbiomeDASim: Simulating longitudinal differential abundance for microbiome data.* F1000Research, 2019. **8**.

72. Romero, R., et al., *The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women.* Microbiome, 2014. **2**(1): p. 1-19.

73. Matsen, F.A., R.B. Kodner, and E.V. Armbrust, *pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree.* BMC Bioinformatics, 2010. **11**(1): p. 1-16.

74. Farr, A., et al., *Role of Lactobacillus species in the intermediate vaginal flora in early pregnancy: a retrospective cohort study.* PLoS One, 2015. **10**(12): p. e0144181.

75. Zmora, N., et al., *Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features.* Cell, 2018. **174**(6): p. 1388-1405. e21.

76. Proal, A.D., P.J. Albert, and T.G. Marshall, *Inflammatory disease and the human microbiome.* Discovery Medicine, 2014(17): p. 257-265.

77. Ahmadi Badi, S., et al., *Microbiota-derived extracellular vesicles as new systemic regulators.* Frontiers in Microbiology, 2017. **8**: p. 1610.

78. Yang, J., et al., *Diagnostic models for atopic dermatitis based on serum microbial extracellular vesicle metagenomic analysis: a pilot study.* Allergy Asthma Immunol Res, 2020. **12**(5): p. 792-805.

79. An, J., et al., *Extracellular vesicle-derived microbiome obtained from exhaled breath condensate in patients with asthma.* Ann Allergy Asthma Immunol, 2021.

80. Lee, J.H., et al., *Metagenome analysis using serum extracellular vesicles identified distinct microbiota in asthmatics.* Scientific Reports, 2020. **10**(1): p. 15125.

81. Kim, S.S., et al., *Microbiome as a potential diagnostic and predictive biomarker in severe alcoholic hepatitis.* Aliment Pharmacol Ther, 2021. **53**(4): p. 540-551.

82. Choi, Y., et al., *Gut microbe-derived extracellular vesicles induce insulin resistance, thereby impairing glucose metabolism in skeletal muscle.* Scientific Reports, 2015. **5**(1): p. 1-11.

83. Cho, Y.S., et al., *A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits.* Nature Genetics, 2009. **41**(5): p. 527-534.

84. Pisprasert, V., et al., *Limitations in the use of indices using glucose and insulin levels to predict insulin sensitivity: impact of race and gender and superiority of the indices derived from oral glucose tolerance test in African Americans.* Diabetes Care, 2013. **36**(4): p. 845-853.

85. RexSoft. *Rex: Excel-based statistical software.* 2018; Available from: http://rexsoft.org/.

86. Lee, E.Y., et al., *Global proteomic profiling of native outer membrane vesicles derived from Escherichia coli.* Proteomics, 2007. **7**(17): p. 3143-3153.

87. Aßhauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data.* Bioinformatics, 2015. **31**(17): p. 2882-2884.

88. Ondov, B.D., N.H. Bergman, and A.M. Phillippy, *Interactive metagenomic visualization in a Web browser.* BMC Bioinformatics, 2011. **12**(1): p. 1-10.

89. Anderson, M.J., *A new method for non-parametric multivariate analysis of variance.* Austral Ecology, 2001. **26**(1): p. 32-46.

90. Gim, J., et al., *A between ethnicities comparison of chronic obstructive pulmonary disease genetic risk.* Frontiers in Genetics, 2020. **11**: p. 329.

91. Greco M, F.D., et al., *Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome.* Statistics in Medicine, 2015. **34**(21): p. 2926-2940.

92. Jin, H., S. Lee, and S. Won, *Causal evaluation of laboratory markers in type 2 diabetes on cancer and vascular diseases using various mendelian randomization tools.* medRxiv, 2020.

93. Bowden, J., *Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption.* International Journal of Epidemiology, 2017. **46**(6): p. 2097-2099.

94. Wen, L. and A. Duffy, *Factors influencing the gut microbiota, inflammation, and type 2 diabetes.* The Journal of Nutrition, 2017. **147**(7): p. 1468S-1475S.

95. Cani, P.D., et al., *Involvement of gut microbiota in the development of low-grade inflammation and type 2 diabetes associated with obesity.* Gut Microbes, 2012. **3**(4): p. 279-288.

96. Upadhyaya, S. and G. Banerjee, *Type 2 diabetes and gut microbiome: at the intersection of known and unknown.* Gut Microbes, 2015. **6**(2): p. 85-92.

97. Sharma, S. and P. Tripathi, *Gut microbiome and type 2 diabetes: where we are and where to go?* The Journal of Nutritional Biochemistry, 2019. **63**: p. 101-108.

98. Vallianou, N.G., T. Stratigou, and S. Tsagarakis, *Microbiome and diabetes: where are we now?* Diabetes Research and Clinical Practice,

2018. **146**: p. 111-118.

99. Puddu, A., et al., *Evidence for the gut microbiota short-chain fatty acids as key pathophysiological molecules improving diabetes.* Mediators of Inflammation, 2014. **2014**.

100. O'Toole, P.W. and I.B. Jeffery, *Gut microbiota and aging.* Science, 2015. **350**(6265): p. 1214-1215.

101. Vacca, M., et al., *The controversial role of human gut lachnospiraceae.* Microorganisms, 2020. **8**(4): p. 573.

102. Ibrahim, K.S., et al., *Characterisation of gut microbiota of obesity and type 2 diabetes in a rodent model.* Bioscience of Microbiota, Food and Health, 2020: p. 2019-031.

103. Sharon, G., et al., *Specialized metabolites from the microbiome in health and disease.* Cell Metabolism, 2014. **20**(5): p. 719-730.

104. Mahendran, Y., et al., *Association of ketone body levels with hyperglycemia and type 2 diabetes in 9,398 Finnish men.* Diabetes, 2013. **62**(10): p. 3618-3626.

105. Sasaki, K., et al., *In vitro human colonic microbiota utilises D-β-hydroxybutyrate to increase butyrogenesis.* Scientific Reports, 2020. **10**(1): p. 1-8.

Abstract

연구배경:

시퀀싱 기술의 발달과 시퀀싱 비용 감소는 미생물 군집에 대한 대규모 분석을 가능하게 하였고 메타 유전체학이 탄생하였으며 이 분야가 광범위하게 발전하였다. 구성 편향과 제로 팽창 문제는 메타 게놈 데이터의 연관 분석을 위한 통계적 방법도 수행하기 어렵게 만든다. 또한 이러한 문제는 반복 측정 내에서 복잡한 상관 관계를 고려해야하는 종단 분석의 모델링을 더 어렵게 만든다. 이러한 희박함과 다양한 데이터베이스 및 클러스터링 방법 선택은 미생물 군유 전체 데이터 세트의 이질성을 유도한다.

연구목적:

이 연구의 목적은 (1) 다양한 클러스터링 방법과 데이터베이스를 기반으로 결과를 비교할 수있는 구성 편향, 제로 인플레이션, 패키지 구현 등 문제를 수정하는 통계적 방법을 개발하는 것이다. (2) 구성 편향, 제로 인플레이션, 종단 데이터 세트 반복 측정 간의 상관 관계 등 문제를 수정하는 통계적 방법 개발, (3) 제 2형 당뇨병 위험 지표에 영향을 줄 수 있는 미생물을 식별하고 다중 오믹스 자료를 활용한 종단 연관분석을 통하여 이를 설명하는 생물학적 배경을 발견한다.

연구방법:

미생물 군유 전체 데이터의 특성을 수정하고 구성 편향 및 제로 팽창 문제를 수정하기 위해 풍부도를 정규화하고 트리 참조 트리 구조와 결합합니다. 전처리 절차와 다른 데이터베이스와의 결과 비교 및 클러스터링 방법을 포함하는 패키지가 개발되어 이질성 문제를 처리 할 수 있습니다. 반복 측정 값 간의 상관 관계는 각각 로버 스트 점수와 Wald 통계를 사용하여 일반화 된 추정 방정식을 반영한다. 제 2 형 당뇨병 위험 지표는 일반화 된 추정 방정식이있는 모델이며 생물학적 메커니즘은 추정 된 기능 게놈 및 SNP를 통해 탐색되었다. 목표 미생물과 제 2 형 당뇨병 위험 사이의 인과 관계를 조사하기 위해 Mendelian 무작위 분석도

수행되었다.

연구 결과 및 결론:

계통 발생 트리 기반 미생물 군집 연합 테스트 (TMAT)는 미생물 풍부도를 표준화하고 계통 발생 트리 구조와 결합하였다. 계통 발생 수를 기반으로 한 시퀀싱 판독의 통합은 제로 인플레이션을 줄이고 두 미생물 풍부 사이의 비율을 취하면 구성 편향을 수정하였다. 다양한 데이터베이스와 클러스터링 방법을 기반으로 한 파이프 라인 구축 미생물 수표를 포함하는 패키지 인 포괄적 인 미생물 군유 전체 연관 분석 (AMAA)과 메타 게놈 전체 연관 분석 방법을 개발하였으며 이를 통해 다양한 데이터베이스 또는 클러스터링 방식을 기반으로 한 통합 전처리 및 결과 비교를 통해 다양한 미생물 군유 전체 연관성 분석 방법을 편리하게 사용할 수 있을 것이다.

TMAT의 확장 버전 mTMAT는 강력한 분산 추정기를 사용하며 반복 측정 된 샘플에 적용 할 수 있다. mTMAT의 통계적 파워는 명목 유형 1 오류를 보존하는 대부분의 시나리오에서 다른 방법보다 우수하였다.

우리는 *Lachnospiraceae* 계통의 GU174097이 제 2 형 당뇨병 위험 지표와 상관 관계가 있음을 발견하였다. 또한 이 속은 단쇄 지방산 (SCFA)과 관련된 경로와 관련이있을 수 있음이 밝혀져 있으며 MR 분석 및 생물학적 배경 조사는 이 속이 당뇨의 위험을 증가시킬 수 있다는 가능성을 시사한다.