이학석사학위논문

A comparison study of statistical methods for the analysis of metagenome data

메타게놈 데이터 분석을 위한 통계적 방법론 비교

2018 년 02 월

서울대학교 대학원

협동과정 생물정보학과

이찬영

A comparison study of statistical methods for the analysis of metagenome data

by

Chanyoung Lee

A thesis
submitted in fulfillment of the requirement
for the degree of Master
in
Bioinformatics

Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University
Feb, 2018

A comparison study of statistical methods for the analysis of metagenome data

지도교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함

2018 년 2 월

서울대학교 대학원
생물정보협동과정 생물정보학 전공
이 찬 영

이찬영의 이학석사 학위논문을 인준함

2018 년 2 월

위 원 장	원 성 호 **(인)**
부위원장	박 태 성 **(인)**
위 원	이 승 연 **(인)**

Abstract

A comparison study of statistical methods for the analysis of metagenome data

Chanyoung Lee
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

With the advent of next-generation sequencing (NGS) technology, sequencing microorganisms from varied samples facilitates association analysis between feature and environment. Several statistical methods have been proposed for analyzing metagenome data such as Metastats, metagenomeSeq, ZIBSeq, ANCOM, edgeR, and DESeq2. Each method has assumed its own specific distribution and model assumptions. While there have been some comparative studies on these methods, the comparison is rather limited and the results have been varied depending on how to generate simulation datasets. In this study, we systematically investigate the properties of these statistical methods for finding differentially abundant features (DAF). In addition, centered log-ratio transformation and permutation logistic regression model (CLR Perm) were applied to metagenome data. We compare their performances using simulation data generated from the Human Microbiome Project (HMP). We first assessed the type I error rate of each method over different levels of sparsity. CLR Perm, metagenomeSeq and ANCOM methods yielded well preserved type I error rates regardless of sparsity. In the power comparison study, CLR Perm showed the highest power among the methods preserving type I error. Furthermore, we applied the methods to real data on colorectal cancer (CRC) to compare our results with existing taxonomic markers of CRC. In conclusion, we recommend using a combination of CLR Perm and metagenomeSeq for the analysis of metagenome data because there are differences in the list of significant taxa discovered by CLR Perm and metagenomeSeq.

Keywords: Differentially abundant feature, Metagenome, 16S rRNA, Association test.

Student number: 2015-20510

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Next Generation Sequencing(NGS) technique contributes DNA-based analysis for various biological studies including metagenomics where researchers extract the genome of microorganisms from sample [1]. There are two approaches in metagenomics. The one is based on 16s rRNA and the other is based on whole genome shotgun sequencing. The former approach has been used mainly to identify the taxa of microorganisms present in a sample, whereas the later has been used to perform functional analyses [2]. Although the both approaches can be identifying taxa, the whole genome shotgun sequencing is not cost effective only to find a taxonomic biomarker. In this study, we focused on the 16S rRNA approach to find taxonomic biomarkers related to treatment (group).

Until now, most of the metagenome (or microbiome) studies have aimed to identify taxa and compare microbial communities between different environments. For this reason, the metagenome studies have been performed mainly by using α and β diversities which are measures for summarizing within-sample variation and between-sample variation, respectively [3]. Additionally, there have been many studies demostrating an association between the occurrence of human diseases and the presence of a specific microorganisms [4]. These are such as psoriasis [5], reflux oesophagitis [6], obesity [7], childhood-onset asthma [8], inflammatory bowel disease [9] ,functional bowel diseases [10], colorectal cancer (CRC) [11], cardiovascular disease [12] etc. For this reason, Microbiome-wide association studies (MWAS), which investigate not only α and β diversities, but also ascertain the relationship between the microorganisms and phenotypes, have become more popular [13].

MWAS are performed mainly by whole genome shotgun sequencing to identify functional markers through the feature table consisting of genes or protein families. However, when analyzing the Operational Taxonomic Unit (OTU) table, the same statistical methods used in 16S rRNA analysis can also be applied. Several statistical methods have been developed for analyzing OTU table, as summarized in Table 1.

Table 1. Summary of methods

Methods		Characteristics			
		Distribution of taxon(taxa)	***Normalization***	***Covariate adjustment***	***Advantage***
Univariate	Metastats (Permutation t-test)	Non-parametric	TSS	x	-
	Metastats (fisher exact test)	Hypergeometric	No	x	-
	ANCOM	Non-parametric	Log-transformed ratios	o	reflecting relation among taxon possible to analysis about repeated sample
	metagenome Seq1	Zero-inflated Gaussian	CSS	o	reflecting sparsity
	metagenome Seq2	Zero-inflated log-normal	CSS	o	reflecting sparsity
	ZIBSeq	Zero-inflated beta	TSS	o	reflecting dispersion and sparsity
	DESeq2	Negative Binomial	RLE	o	reflecting similarity of RNA-Seq over-dispersion
	edgeR	Negative Binomial	TMM	o	reflecting similarity of RNA-Seq over-dispersion
	CLR Perm	Not assume	Centered-log ratio transformation	o	reflecting relation among taxon. not need to consider sparsity on distribution.
Multivariate (Multiple)	PerMANOVA	Non-parametric	Distance	o	reflecting phylogenetic distance
	Dirichlet-multinomial regression	Dirichlet-multinomial	Distance	o	reflecting phylogenetic distance variable selection based on penalized model.
	aMiSPU	not assume (independent variable)	Distance × TSS	o	reflecting phylogenetic distance
	MiRKAT	not assume (independent variable)	Distance	o	reflecting multiple phylogenetic distance

Statistically, they can be classified into univariate models and multivariate models. The hypothesis of univariate model is testing a single taxon abundance differences between the treatment groups. While that of multivariate is to test whether or not microbial communities differs between treatment groups. The univariate models are sometimes called as differentially abundant features (DAFs) finding methods. In this study, we focused on univariate models.

Many univariate models have been proposed for analyzing OTU table. Among them, Metastats [14] was proposed using the t-test and Fisher's exact test. However, Metastats does not consider many zeros in metagenome data and relative abundances of compositional data. To consider these zeros, metagenomeSeq [15] and ZIBSeq [16] were proposed using zero-inflated mixture models. Analysis of composition of microbiome (ANCOM) [17] was also proposed to solve the problem of using relative abundances of compositional data. DESeq2 [18] and edgeR [19], originally proposed for RNA sequencing data, can also be used for finding DAFs. Table 1 summarizes the distributions, normalization methods, and the ability of handling covariates for each method. Although this study focused on DAFs tests (univariate), multivariate (multiple) methods also introduced in Table 1 [20-23].

Most OTU tables contain many zero counts at a particular taxon. This phenomenon is called a sparsity problem which may occur when the sequencing depth is not deep enough to detect a rare taxa [24]. The sparsity is defined as a ratio of zero for each taxon. Thus, when sparsity is high, that is when there are many zeros in a OTU table, it is difficult to know whether zero abundance is due to the absence of taxon in environment or to the lack of sequencing depth.

While many methods for finding DAFs are available, each method has its own specific distributional and model assumptions (Table 1). They were shown to have different performances depending on simulation datasets. In this study, we systematically investigate the properties of univariate models listed in Table 1 using simulation data generated from the Human Microbiome Project (HMP) data [25, 26].

Through the HMP data, we generated a binary phenotype and kept to the natural characteristics of the real-world data such as sparsity and their relative composition. We paid special attention to the effect of sparsity on the performance of these statistical methods. Furthermore, we performed colorectal cancer (CRC) data analysis (Baxter et al. [27]), and compared the results with previous studies related to CRC.

Chapter 2

Materials and Methods

2.1 Simulation materials (HMP)

The OTU table generated from 16s rRNA variable region v3-5 and metadata used in this study were downloaded from HMP QIIME Community Profiling website (http://hmpdacc.org/HMQCP/). The v3-5 data is composed of 4743 specimens of 235 screened healthy adults from 124 males and 111 females. Among the 18 body site data of HMP, stool data was selected for our comparison study because of its larger sample size than other sites [25]. Characteristics of the HMP stool data at genus level are summarized in Figure 1. Bar-plot in Figure 1 shows variation

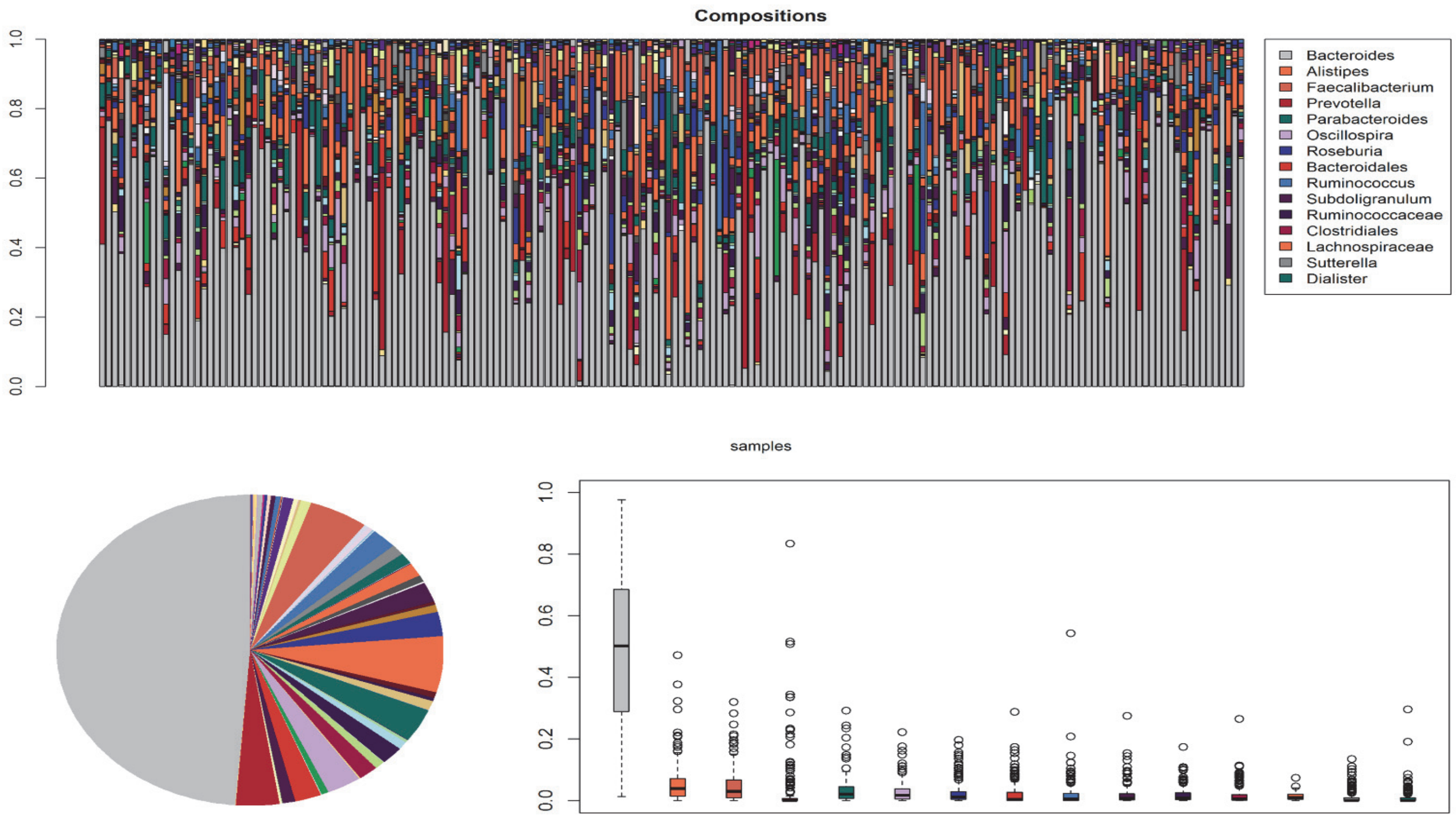


Fig. 1. Basic characteristic of HMP stool dataset using relative abundance.

between samples. Each stacked bar represents an individual and each color represents relative abundance of taxon. Pie-chart in Figure 1 shows the mean values of relative abundance of taxa across the sample. Box-plot in Figure 1 shows the variation of some taxon with a large mean of relative abundance. Individual variability seems pretty large. Especially, grey colored taxon shows dominantly higher proportion across all samples and larger variability than other taxa.

2.2 Colorectal cancer data (CRC)

Until now, it has been difficult to obtain metagenome data the true significance of whose taxonomic markers for a disease was known. For this reason, the performance of statistical methods had not been evaluated on real data. In this study, we use CRC data published by Baxter at al. [27]. The data relate to the development of a model for improving the accuracy of the fecal immunochemical test (FIT) [28]. In their study, the prediction model for diagnosing of colonic lesions was developed using the random forest method. Although, the significant markers of association test and the markers in prediction model have slightly different meanings due to different evaluation criteria, we compared significant taxa of association tests with the taxa used in prediction model.

We downloaded 16s rRNA variable region v4 dataset from https://github.com/SchlossLab/Baxter_glne007Modeling_GenomeMed_2015. The data are composed of 292 samples: 172 normal samples and 120 CRC samples. We used 335 OTUs in our study that account for at least 5 percent of abundance across the taxa. The characteristics of the CRC data at genus level are summarized in Figure 2. Bar-plot in Figure 2 shows variation between case and control. Unlike the HMP data, Pie-chart in Figure 2 does not show dominantly higher proportion. Box-plot in Figure2 shows that the top 15 taxa in order of highest proportion show similar abundances and variations.

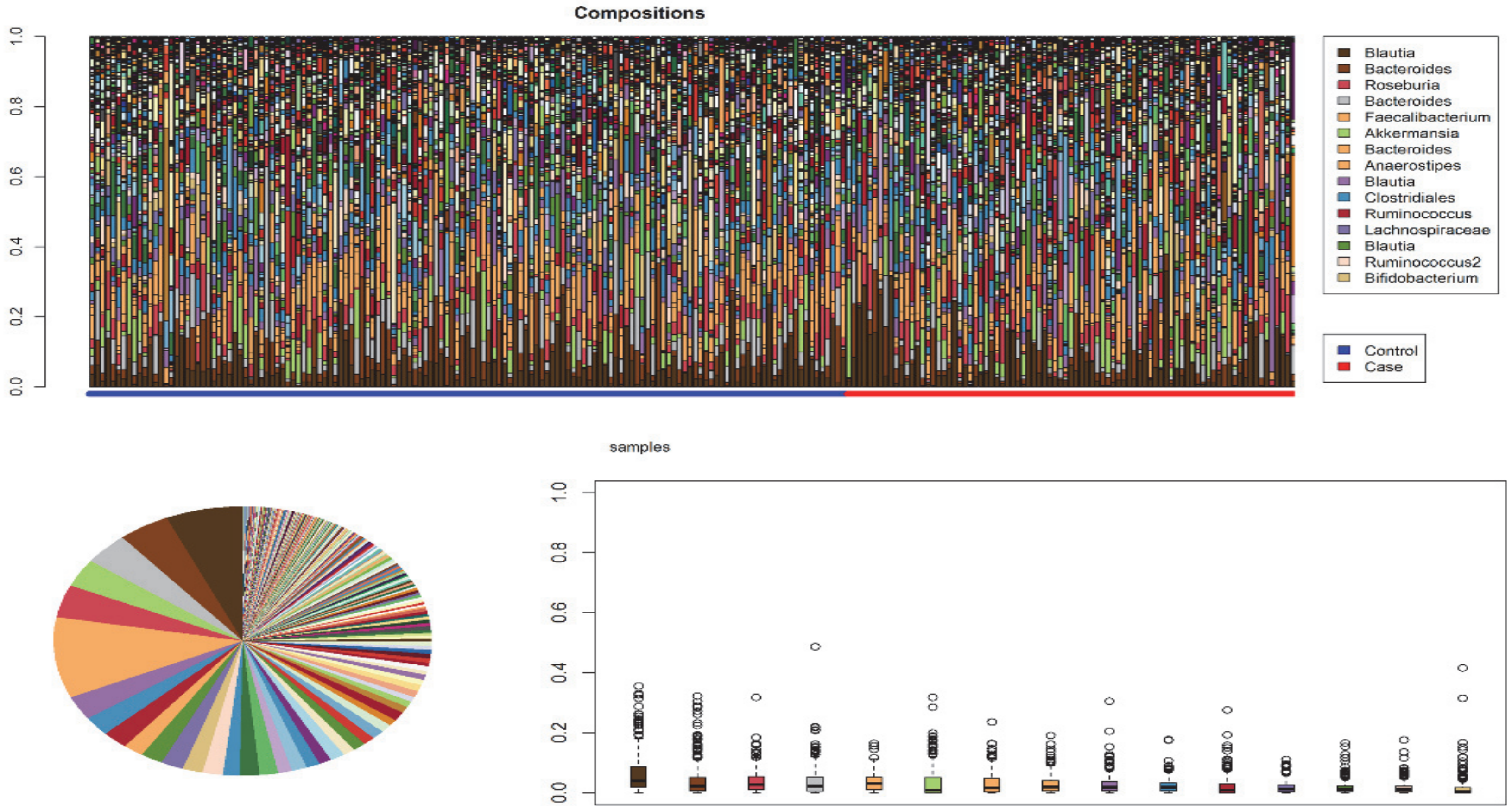


Fig. 2. Basic characteristic of colorectar cancer dataset using relative abundance.

2.4 Existing methods

We briefly review existing methods for finding DAF. Their characteristics are well summarized in Table 1.

- Metastats

In Metastats, taxa counts are normalized by library size (sample total). This normalization process is called total sum scaling (TSS) and the normalized abundance is called relative abundance (RA). In Metastats, either a permutation t-test based on Welch's t-test statistics [29] or Fisher's exact method is used [30]. Fisher's exact method is applied only if the total feature count is less than the number of subject in each treatment. Otherwise, the permutation t-test is used.

- metagenomeSeq

metagenomeSeq was developed to take sparsity in OTU table into account. At first, a binary indicator is introduced to represent whether the feature count 0 is actually zero or is caused by the lack of sequencing depth (library size)[15]. Since there is no way of knowing the true value of this indicator, it is treated as a latent variable. The log2-transformed feature count is assumed to follow a Gaussian distribution. To account for extra zero counts, the zero inflated Gaussian mixture model (metagenomeSeq1) is defined. The Expectation-maximization (EM) algorithm [31] is used for fold-change estimation. In the E-step, the mean model is expressed by a regression model which adopts Cumulative Sum Scaling (CSS) normalization factor and a parameter of fold-change. Finally assessing significance is done through a moderate t-test [32] using estimated parameter of fold change in the mean model. Recently, the zero inflated log normal models (metagenomeSeq2) were also developed [33]. metagenomeSeq1 and metagenomeSeq2 are different in their distributional assumptions and parameter estimation steps.

- ZIBSeq

The key idea of ZIBSeq lies in that relative abundance after TSS normalization is composed of large number of zeros and results in the skewed distribution [16]. Thus, the zero inflated beta distribution is considered with the beta distribution reparametrized by mean and dispersion parameters. With the assumption of relative abundance to follow zero inflated beta distribution, regression model is used with the logit link. For parameter estimation, R package GAMLSS can be used to find MLEs of model parameters numerically [34]. For testing hypothesis, Wald statistics is used with its p-value obtained from the chi-squared distribution.

- ANCOM

ANCOM was developed to compensate for a weakness in the methods that calculate the relative abundance (RA) of features: the RA value can vary greatly due to a small change in the abundance of one of the features [35]. The key idea behind ANCOM is to use a hypothesis test which is based on computing the pairwise log ratio. The preceding methods used the following hypothesis test : For the $i(= 1, \dots, q)$th taxon and the $k(= 1, \dots, K)$ group, the hypothesis,

$$H_{0i} : \mathrm{E}\left(\log\left(\mu_i^{(1)}\right)\right) = \mathrm{E}\left(\log\left(\mu_i^{(2)}\right)\right),$$

has been commonly used, where $\mu_i^{(k)}$ represents the mean abundance of the *k*th group. Instead of this hypothesis, ANCOM considered the following $q - 1$ sub-hypotheses:

$$H_{0ri} : \mathrm{E}\left(\log\left(\mu_i^{(1)}/\mu_r^{(1)}\right)\right) = \mathrm{E}\left(\log\left(\mu_i^{(2)}/\mu_r^{(2)}\right)\right),\ \forall \mathrm{r}(\neq \mathrm{i}).$$

Then, all pairwise $q(q - 1)/2$ p-values are calculated by a non-parametric procedure such as Wilcoxon rank sum test $(\mathrm{K} = 2)$, the Kruskal-Wallis test $(\mathrm{K} \geq 3)$, and Freidman test for repeated sample.From the $q(q - 1)/2$ p-values for each taxon i, W_i, the number of rejected sub-hypotheses H_{0ri} , $r \neq i$, can be computed. Finally, through empirical cumulative density function of W_i, decision rule of significance taxa is defined.

- DESeq2 and edgeR

Both methods use the negative binomial model. The main differences lie in normalization and parameter estimation procedures. edgeR uses the trimmed mean of M-value (TMM) normalization and DESeq2 uses relative log expression (RLE) normalization [36] Additionally, DESeq2 and edgeR use slightly different estimation procedures for the dispersion parameters. The detailed step of estimating dispersion parameters was introduced by Love et al. [18] and Robinson et al. [19]. In this study, the likelihood ratio test (LRT) statistics and Wald statistics are used for DESeq2, while only LRT statistics are used for edgeR.

2.4 Permutation logistic regression with centered log-ratio transformation (CLR Perm)

In previous existing methods, taxon counts or abundances are treated as random variable. Therefore, the counts and abundances have distributional assumption. For this reason, complex distributional assumptions such as the zero-inflated mixture models are required by the sparsity of metagenome data. Moreover, when computing the RA, a small change in the abundance of one taxon affects the whole microbiota composition because the overall abundance of compositional data is constrained to sum to one.

In this study, we simply treated taxon counts as constant and binary trait as random variable. From this point of view, we used logistic regression and centered log-ratio transformation developed by John Aitchison [35] to overcome the constant sum constrain. In addition, permutation method [29] is applied to Wald statistics of logistic regression to reflect exact null distribution based on observation.

- Centered log-ratio transformation (CLR)

  Let, y_{ij} is raw taxon count. For the $i(=1,\ldots,Q)$th taxon and the $j(=1,\ldots,N)$ sample. To avoid the geometric mean becoming zero, $y_{ij}^{*}\left(=y_{ij}+1\right)$ is considered by adding a pseudo count 1. After adding pseudo count for each count, a vector, $\boldsymbol{y}_j^{*}=\left(y_{1j}^{*},y_{2j}^{*},\ldots,y_{Qj}^{*}\right)$, is defined for sample j, (where, $y_{ij}^{*}\geq 0$ and $\sum_{i=1}^{Q}y_{ij}^{*}=l_j$). l_j is library size of sample j. Then CLR transformed values x_{ij} is defined as below :

$$x_{ij}=ln\frac{y_{ij}^{*}}{g\left(\boldsymbol{y}_j^{*}\right)}\ where,g\left(\boldsymbol{y}_j^{*}\right)=\sqrt[Q]{y_{1j}^{*}\cdot y_{2j}^{*}\cdot\ldots\cdot y_{Qj}^{*}}$$

- Permutation logistic regression

  Let $Y_j\sim\ iid\ bernouli(\pi)$ and $\boldsymbol{x_j}=(x_{0j},x_{1j},x_{2j},\ldots,x_{Qj})$ is above transformed value. the success probability is $\Pr\left[Y_j=1|\boldsymbol{x}_j\right]=\pi\left(\boldsymbol{x}_j\right)=$

$\frac{\exp\left(\sum_{i=0}^{Q}\beta_i x_{ij}\right)}{1+\exp\left(\sum_{i=0}^{Q}\beta_i x_{ij}\right)}$, where $x_{0j} = 1$. Then we can define likelihood $L(\beta)$ as below :

$$L(\beta) = \sum_i (\sum_j y_j x_{ij})\beta_i - N \cdot \log\left[1 + \exp\left(\sum_i \beta_i x_{ij}\right)\right]$$

Estimate $\widehat{\beta_\iota}$ is obtained from $\frac{\partial L(\beta\,)}{\partial(\beta_i)} = 0$. standard error of $\widehat{\beta_\iota}$ is square-root of $\left[-\frac{\partial^2 L(\beta\,)}{\partial(\beta_i)^2}\ \right]^{-1}$. Statistics Z_i is defined as below :

$$Z_i \ = \frac{\widehat{\beta_\iota}}{s.e(\widehat{\beta_\iota})}$$

Finally, Storney and Tibshirani permutation method [29] is used to assess significance under the null ($H_0 :\ \beta_i$=0). we randomly shuffle Y_j labels over *K* times and compute statistics $Z_{i1}, Z_{i2}, \ldots, Z_{iK}$ for *i* th taxon. After calculating a set of statistics, p-value , p_i, of *i* th taxon is computed as below :

$$p_i \ = \sum_{k=1}^{K} \frac{I[|Z_{ik}| \geq |Z_i|]}{K}$$

Chapter 3

Simulation

3.1 Simulation model

To compare the performance of each method for finding DAFs, we used HMP stool data (genus level). For simplicity, we fixed the HMP stool data and generated binary responses from the binomial distribution. Let Y_{ij}^{k} be a random variable generated from $Bernouli\left(\pi(x_{ij})\right)$ and x_{ij} be scaled relative abundance, where $i(=1,\ldots,165)$ represents causal taxon, $j(=1,\ldots,180)$ samples, and $k(=1,\ldots,100)$ replicate numbers. The success probability $\pi(x_{ij})$ is defined as a function of x_{ij} given as follows:

$$\pi(x_{ij}) = \frac{\exp(\beta_0+\beta_1 x_{1j}+\cdots+\beta_i x_{ij}+\cdots+\beta_{165}x_{165j})}{1+\exp(\beta_0+\beta_1 x_{1j}+\cdots+\beta_i x_{ij}+\cdots+\beta_{165}x_{165j})} \quad (1)$$

The parameter β_0 in model (1) is an intercept and is given by $-\sum_{j=1}^{n}\frac{\beta_i x_{ij}}{n}$ to let $\pi(x_{ij})$ have values near 0.5 and β_i vary from 0.5 to 1.5 by increments of 0.5. Using the given model, the power and type I error are computed by assuming that only one taxon is truly causal and the other 164 taxa are non-causal. The power for detecting a causal taxon i can be computed when $\beta_i \neq 0$ and $\beta_{i'} = 0$ for all $i' \neq i$. The type I error rate of detecting non-causal taxon i' can similarly be evaluated for all $i' \neq i$.

3.2 Power and type I error rate

Assuming that only one taxon is truly causal and the other 164 taxa are non-causal, we compared the performance of each method at different levels of sparsity based on their empirical power and type I error rates. A detailed procedure for computing power and the type I error rates is given in Figure 3. Figure 3 (a) shows 165 tables of binary traits generated by model (1). The first table in Figure 3 (a) shows the case when the first taxon is causal, the second table shows the case when the second taxon is

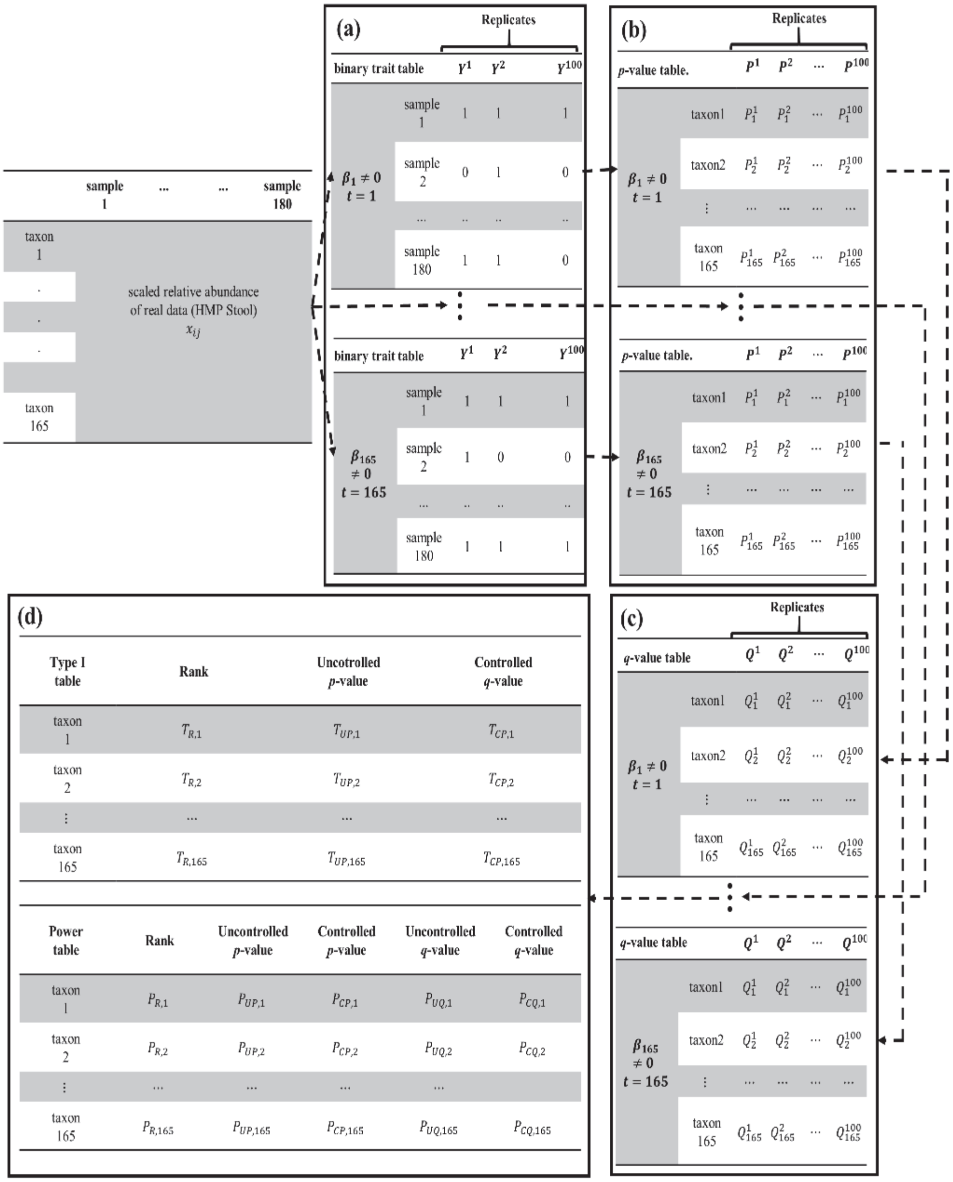

sample 1 ... ... sample 180
taxon 1
taxon 165
scaled relative abundance of real data (HMP Stool) x_{ij}
(a)
Replicates
binary trait table Y^1 Y^2 Y^{100}
$\beta_1 \neq 0$ $t = 1$
sample 1 1 1 1
sample 2 0 1 0
sample 180 1 1 0
$\beta_{165} \neq 0$ $t = 165$
sample 1 1 1 1
sample 2 1 0 0
sample 180 1 1 1
(b)
p-value table. P^1 P^2 ... P^{100}
taxon1 P_1^1 P_1^2 ... P_1^{100}
taxon2 P_2^1 P_2^2 ... P_2^{100}
taxon 165 P_{165}^1 P_{165}^2 ... P_{165}^{100}
(c)
q-value table Q^1 Q^2 ... Q^{100}
taxon1 Q_1^1 Q_1^2 ... Q_1^{100}
taxon2 Q_2^1 Q_2^2 ... Q_2^{100}
taxon 165 Q_{165}^1 Q_{165}^2 ... Q_{165}^{100}
(d)
Type I table | Rank | Uncotrolled p-value | Controlled q-value
taxon 1 $T_{R,1}$ $T_{UP,1}$ $T_{CP,1}$
taxon 2 $T_{R,2}$ $T_{UP,2}$ $T_{CP,2}$
taxon 165 $T_{R,165}$ $T_{UP,165}$ $T_{CP,165}$
Power table | Rank | Uncontrolled p-value | Controlled p-value | Uncontrolled q-value | Controlled q-value
taxon 1 $P_{R,1}$ $P_{UP,1}$ $P_{CP,1}$ $P_{UQ,1}$ $P_{CQ,1}$
taxon 2 $P_{R,2}$ $P_{UP,2}$ $P_{CP,2}$ $P_{UQ,2}$ $P_{CQ,2}$
taxon 165 $P_{R,165}$ $P_{UP,165}$ $P_{CP,165}$ $P_{UQ,165}$ $P_{CQ,165}$

Fig. 3. The Procedure of computing type1 error rates and powers.

causal, and so forth. Figure 3 (b) shows the tables of p-values for the test results. Figure 3 (c) shows the tables of q-values for multiple comparisons. Finally, Figure 3 (d) shows the tables of powers and type I error rates. The type I error rate (Uncontrolled p-value) was computed by counting the number of cases when the unadjusted p-value under the null hypothesis was smaller than 0.05, $T_{UP,i'} = \sum_{\{t:\beta_{i'}=0\}} \sum_{k=1}^{100} I\left[P_{i'}^{k} < 0.05\right] / (100 \cdot 164)$, where t is the index of a table as given in Figure 3. In order to restrict the type I error rate to 5% for methods with a high type I error rate, we numerically found the $C_{i'}$ thresholds for each taxon at which the Type I error rate (Controlled p-value) was limited to 0.05 with the following equation: $T_{CP,i'} = \sum_{\{t:\beta_{i'}=0\}} \sum_{k=1}^{100} I\left[P_{i'}^{k} < C_{i'}\right] / (100 \cdot 164)$. Additionally, we computed rank-based type I errors. the Type I error rate (Rank): $T_{Ri'} = \sum_{\{t:\beta_{i'}=0\}} \sum_{k=1}^{100} I[P_{i'}^{k} = P_{(1)}^{k}] / (100 \cdot 164)$, where $P_{(1)}^{k}$ is the minimum p-value in kth replicate. For power comparison, we considered five types of power. The first type of power P(Uncontrolled p-value), is based on p-values not adjusted by applying the multiple comparisons correction. Specifically, P(Uncontrolled p-value) was computed by counting the number of cases when the unadjusted p-value of the causal taxon was smaller than 0.05, $P_{UP,i} = \sum_{k=1}^{100} I\left[P_{i}^{k} < 0.05\right] / 100$. P(Controlled p-value) was computed in the same way as P(Uncontrolled p-value), but the threshold $C_{i'}$

obtained from $T_{Ci'}$ was used to compare the power of all the methods at the same type I error threshold. P(Controlled p-value) is defined as $P_{CP,i} = \sum_{k=1}^{100} I\left[P_i^k < C_{i'}\right]/100$, where $i = i'$. The second type of power, P(Uncontrolled q-value) uses the Benjamini-Hochberg correction-derived q-values for multiple comparison [37]. P(Uncontrolled q-value) were computed by counting the number of cases when the Benjamini-Hochberg q-value of causal taxon was smaller than 0.05, $P_{UQ,i} = \sum_{k=1}^{100} I[Q_i^k < 0.05]/100$. Controlled p-values $\widetilde{P_i^k} = P_i^k \cdot \frac{0.05}{C_{i'}}$, are computed to obtain P(Controlled q-value) using the threshold $C_{i'}$. After computing controlled p-values, controlled q-values, $\widetilde{Q_i^k}$, are obtained by applying the Benjamini-Hochberg correction to the adjusted p-values. Finally, P(Controlled q-value) was computed by counting the number of cases when the adjusted q-value of a causal taxon was smaller than 0.05: $P_{CQ,i} = \sum_{k=1}^{100} I[\widetilde{Q_i^k} < 0.05]/100$. P(Rank) was computed by ranking of p-values : $P_{R,i} = \sum_{k=1}^{100} I[P_i^k < P_{(1)}^k]/100$.

Chapter 4

Results

4.1 Simulation results

We performed simulation studies by varying β_i from 0.5 to 1.5 in increments of 0.5. Since the ANCOM method does not provide p-values, but does provide W_i statistics and a list of significant taxa, the power of ANCOM was computed on the basis of the significant taxa list, and the P(Rank) computed using W_i.

Figure 4 (a) shows the type I errors (Rank), which have very low Type I error rate because of rank-based calculation. Figure 4 (c) shows Type I error (Controlled p-value) of each method at the same type I error rate (α=0.05). The results of type I errors (Uncontrolled p-value) showed that some methods did not preserve type I errors. Figure 4 (b) shows type I

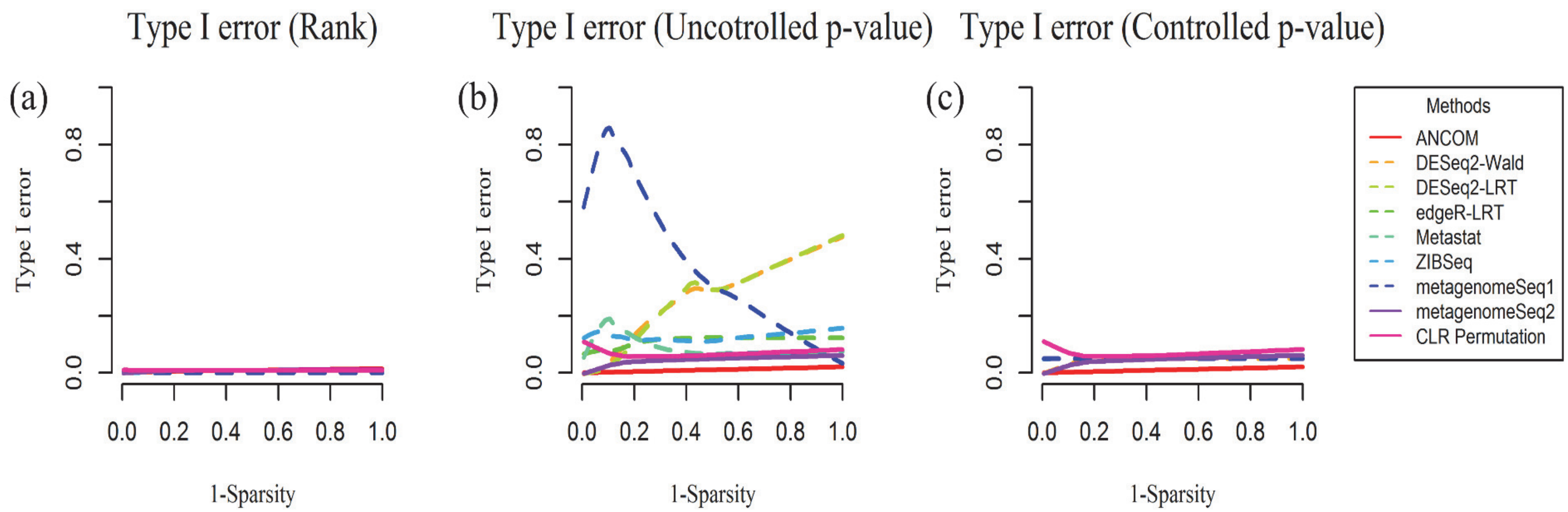


Fig. 4. Type I error of each method for finding DAFs Based on simulation.

error rates at different levels of sparsity. Three out of nine methods, represented by solid lines, were shown to have well preserved type I errors. However, the other methods represented by dashed lines, yielded inflated more type I errors. In particular, with increasing sparsity, the type I error of DESeq2 tended to decrease and that of metagenomeSeq1 tended to increase. This was due to the fact that the p-values of DESeq2 become close to one when sparsity is greater than 0.8, while those of metagenomeSeq1 become close to zero as sparsity increases. edgeR and ZIBSeq did not preserved type I errors, with similar patterns, as shown in Figure 4. Metastats showed a well-preserved type I errors when sparsity was low. However, when the sparsity is higher than 0.6, the type I error rate of Metastats tends to increase.

Figure 5 shows the results of power analysis. Figure 5 (a) and (b) show the plots of P(Rank) and P(Uncontrolled p-value) respectively over different levels of sparsity. High P(Rank) means that the causal taxon was frequently assigned the minimum p-value among all the other p-values. Among the methods well-preserving type I errors, CLR Perm showed the highest power; metagenomeSeq2 a high power, similar to CLR Perm. Figure 5 (d) shows the plot of P(Uncontrolled q-value) over different levels of sparsity. The ANCOM method showed the highest power among the methods well-preserving type I errors when applying

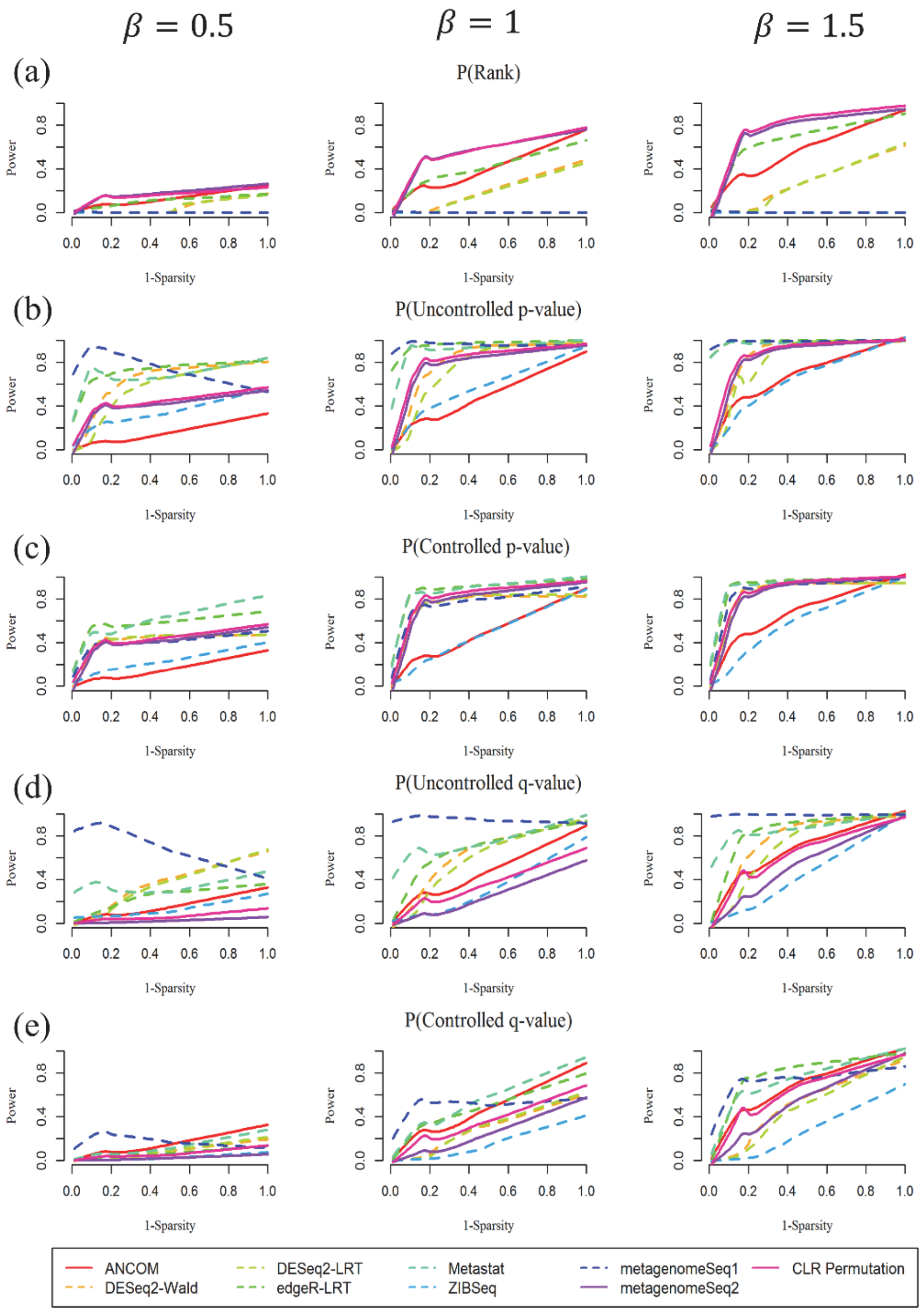


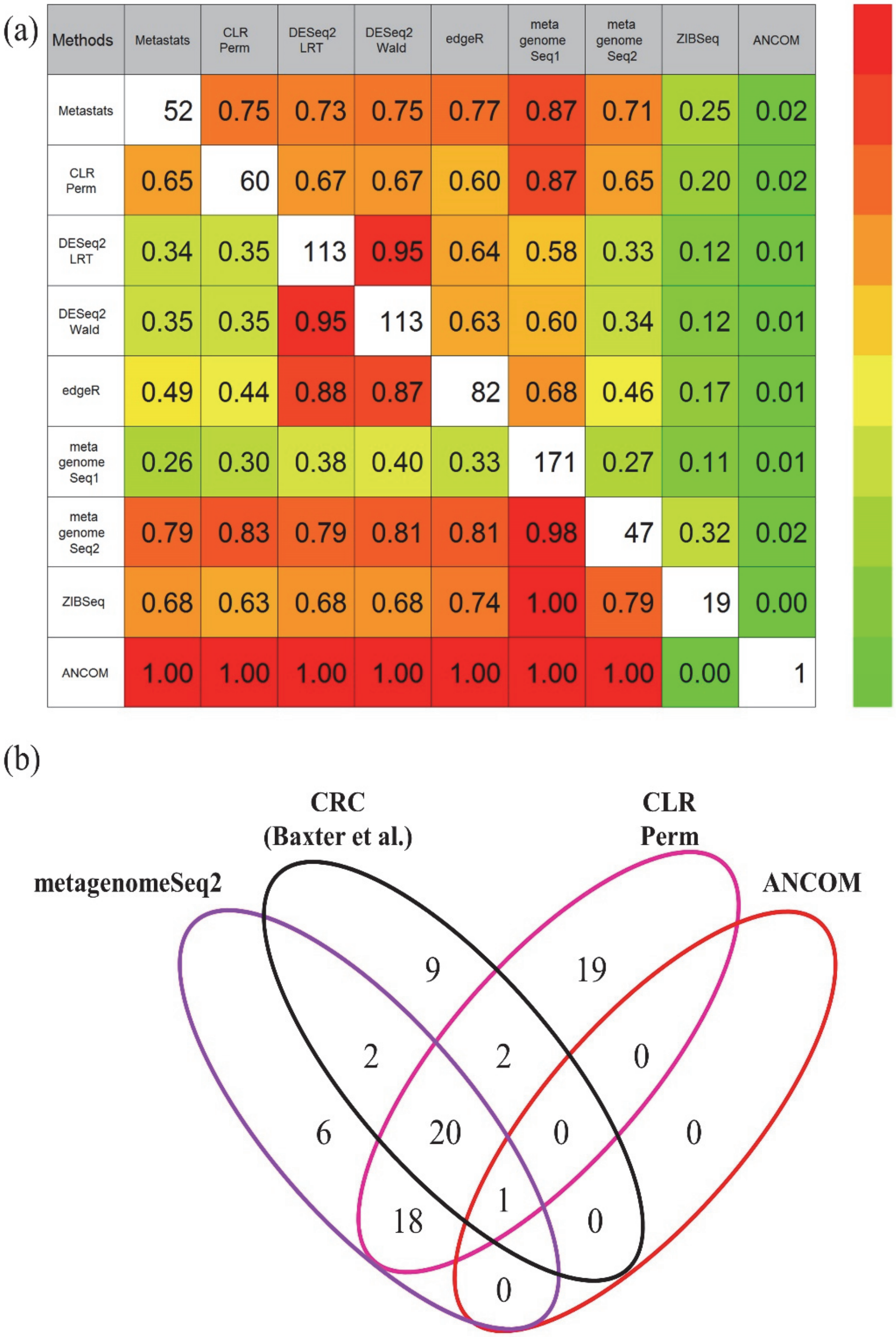
Fig. 5. Comparsion results of statistical powers of the methods. Different significant thresholds were used for each taxon to compare the P(Controlled *p*-value,*q*-value) of each method at similar type1 error rate(α=0.05).

multiple comparisons correction to the *p*-values. In order to include other methods that did not preserve the type I errors, we adjusted their threshold values so that their type I errors were controlled at the 5% significance level. Figure 4 (c) showed that type I error rates (Controlled pvalue) are well-preserved for all methods when using the adjusted thresholds. Using these thresholds, we re-computed the power, P(Controlled *p*-value), as shown in Figure 5 (c). Though, all the methods showed very similar power, they yielded two clusters: one cluster having higher power includes DESeq2, edgeR, Metastat, metagenomeSeq, CLR Perm and the other cluster with lower power includes ZIBSeq, and ANCOM. However, it is not easy to estimate these threshold values when analyzing real data. Figure 5 (e) shows the P(Controlled *q*-value) plots. metagenomeSeq1, edgeR and Metastats have high power when applying Benjamini-Hochberg correction to controlled *p*-value. As a result, all methods except CLR Perm, metagenomeSeq2, and ANCOM suffered from inflation of false positive errors.

4.2 Colorectal cancer data results

Until now, we compared the performances of these methods through simulation results. We next confirmed the consistency of detecting significant OTUs by each method through CRC data as shown in the heatmap in Figure

6 (a). Using the Venn diagram in figure 6 (b), we compared the significant OTUs detected by the three methods with low type I error rate with 34 OTUs detected by Baxter et al. in detail. The CRC data consists of 353 OTUs. Among those OTUs, 34 OTUs were used to construct a prediction model to diagnose colorectal cancer by Baxter et al. In their study, the prediction model showed an AUC (area under the curve) value of 0.84. Figure 6 (a) shows the consistency of significant OTU detection by each method. The diagonal in figure 6(a) shows the number of significant OTUs detected at a level of significance 0.05 by each method. The rest of the cells in figure 6(a) display the ratio of the number of significant OTUs found by both the row name method and column name methods (numerator) to the number of significant OTUs found only by the row name method (denominator). Thus, the halves on either side of the diagonal are not symmetric. As the ratio increases from zero to one, the color changes from green to red. From Figure 6 (a), we can see how consistently two different methods are in finding significant OTUs. The DESeq2 and metagenomeSeq1 methods found a large number of significant OTUs. These results seem to be due to their high type I error rate as shown in the simulation. Except ZIBSeq, all the methods found the OTU also found by ANCOM. Among the



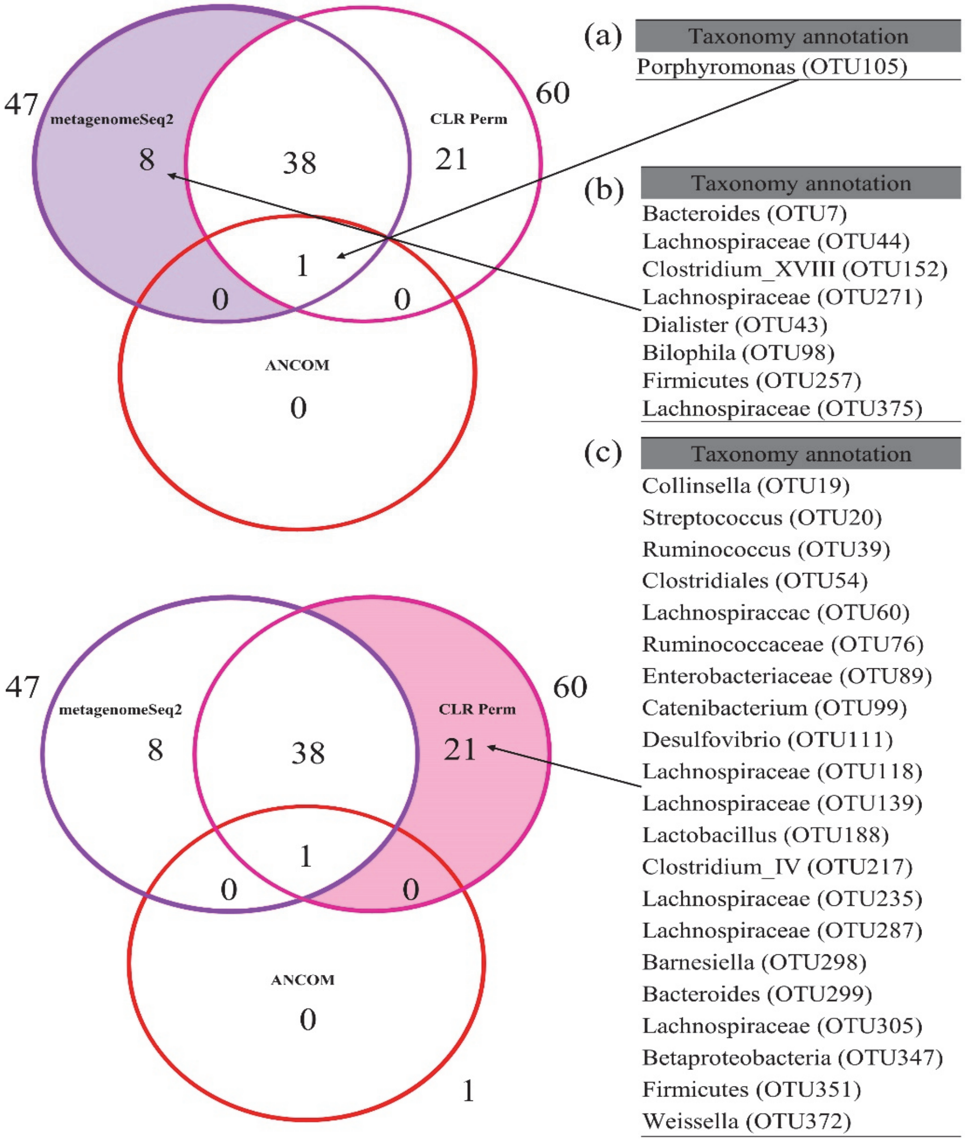
(a)

Methods	Metastats	CLR Perm	DESeq2 LRT	DESeq2 Wald	edgeR	meta genome Seq1	meta genome Seq2	ZIBSeq	ANCOM
Metastats	52	0.75	0.73	0.75	0.77	0.87	0.71	0.25	0.02
CLR Perm	0.65	60	0.67	0.67	0.60	0.87	0.65	0.20	0.02
DESeq2 LRT	0.34	0.35	113	0.95	0.64	0.58	0.33	0.12	0.01
DESeq2 Wald	0.35	0.35	0.95	113	0.63	0.60	0.34	0.12	0.01
edgeR	0.49	0.44	0.88	0.87	82	0.68	0.46	0.17	0.01
meta genome Seq1	0.26	0.30	0.38	0.40	0.33	171	0.27	0.11	0.01
meta genome Seq2	0.79	0.83	0.79	0.81	0.81	0.98	47	0.32	0.02
ZIBSeq	0.68	0.63	0.68	0.68	0.74	1.00	0.79	19	0.00
ANCOM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1

Fig. 6. (a) The heatmap and (b) Venn diagram of Colorectal cancer data analysis.

methods well-preserved type I error rates, CLR Perm identified 83 percent (39/47) of the discoveries made by metagenomeSeq2. For a more detailed comparison, we used the results of ANCOM, metagenomeSeq2, and CLR Perm with the result encompassing 34 OTUs obtained by Baxter et al. The results are shown in Figure 6 (b). Through the three methods, we found 25 OTUs among the 34 OTUs. The list of significant OTUs found through the methods preserving type I errors is shown in Figure 7. Figure 7 (a) shows a significant taxon found in ANCOM—Porphyromonas—a well-known OTU related to colorectal cancer [38-42]. Figure 7 (b) shows the list of significant OTUs found only by metagenomeSeq2 and Figure 7 (c) shows the list of significant OTUs found only by CLR Perm. Although a detailed analysis based sequencing is required to acquire complete taxonomical information, most of the genera listed in figure 7 (b) and (c) are frequently mentioned as being associated with colorectal cancer [38-42]. In summary, through metagenomeSeq2 and CLR Perm, we could find OTUs known to be related to CRC from various previous studies.

We used -log10(*p*-value) of CRC data to obtain the dendrogram in Figure 8; except for ANCOM which does not provide *p*-values. The remaining eight methods group well according to their underlying distributions and characteristics (zero inflated mixture model,



(a)

Taxonomy annotation
Porphyromonas (OTU105)

(b)

Taxonomy annotation
Bacteroides (OTU7)
Lachnospiraceae (OTU44)
Clostridium_XVIII (OTU152)
Lachnospiraceae (OTU271)
Dialister (OTU43)
Bilophila (OTU98)
Firmicutes (OTU257)
Lachnospiraceae (OTU375)

(c)

Taxonomy annotation
Collinsella (OTU19)
Streptococcus (OTU20)
Ruminococcus (OTU39)
Clostridiales (OTU54)
Lachnospiraceac (OTU60)
Ruminococcaceae (OTU76)
Enterobacteriaceae (OTU89)
Catenibacterium (OTU99)
Desulfovibrio (OTU111)
Lachnospiraceae (OTU118)
Lachnospiraceae (OTU139)
Lactobacillus (OTU188)
Clostridium_IV (OTU217)
Lachnospiraceae (OTU235)
Lachnospiraceae (OTU287)
Barnesiella (OTU298)
Bacteroides (OTU299)
Lachnospiraceae (OTU305)
Betaproteobacteria (OTU347)
Firmicutes (OTU351)
Weissella (OTU372)

Fig. 7. (a) A significant taxon found in three methods simultaneously (b) The list of taxa only found in metagenomeSeq2 . (c) The list of taxa nly found in CLR Perm.

permutation based method, negative binomial model). The -log10(*p*-value) of metagenomeSeq1 seems to be far from the -log(*p*-value) of the other methods due to its high type I error rate. Based on the simulation results, we can identify the highest-power method with the preserving type I error rate for each group. First, among the zero-inflated models, we recommend metagenomeSeq2. Second, of the two permutation-based models, we recommend CLR Perm. Third, although negative-binomial models (edgeR, DESeq2) show high type I error rates, we recommend edgeR because edgeR shows a better performance than DESeq2 with respect to P(rank), P(Controlled *p*-value) and P(Controlled *q*-value). Finally, for selecting the fewest number of important taxa, ANCOM is recommended due to its high P(Uncontrolled *q*-value).

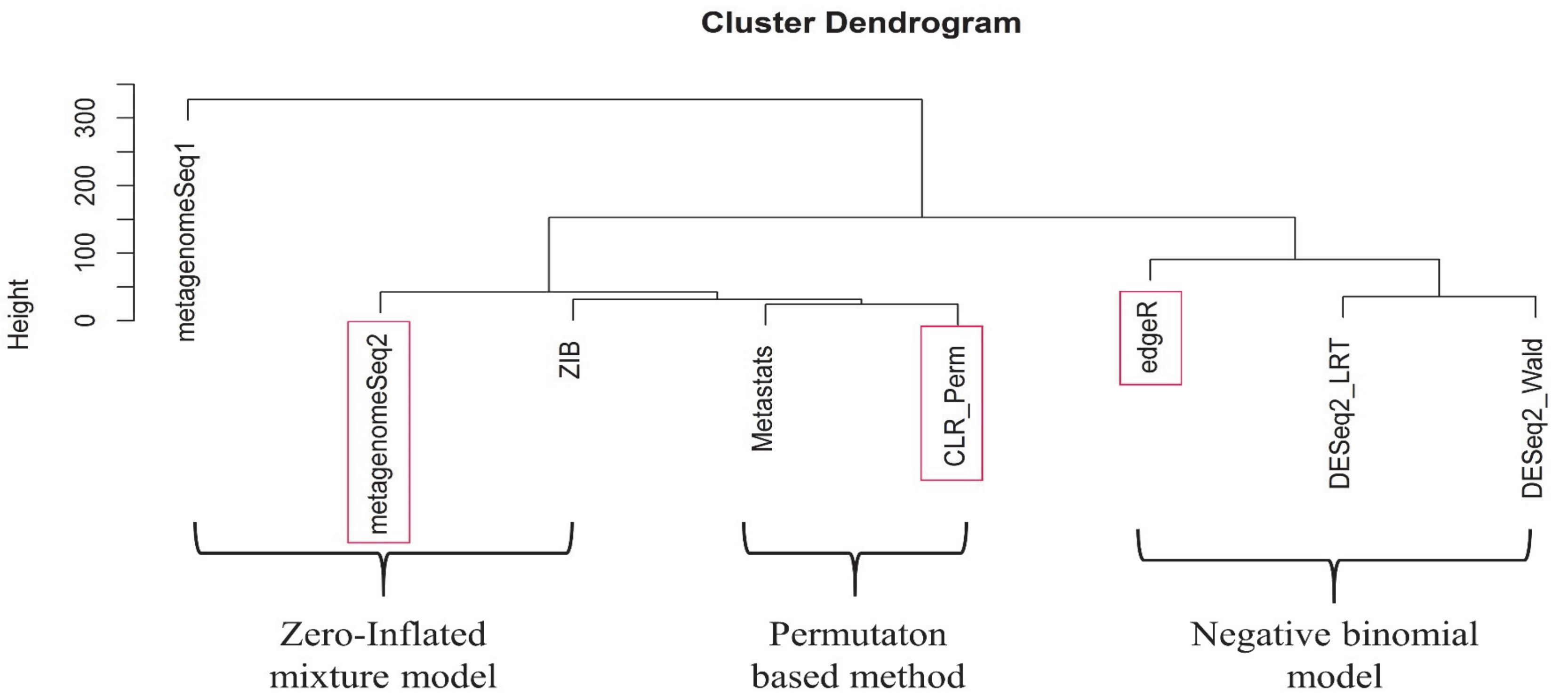


Fig. 8. The above dendrogram was obtained by using the –log10(*p*-value) of each method. The methods group well according to their characteristics (Zero-inflated model, Permutation based model, Negative binomial model).

Chapter 5

Discussion

Since 16S rRNA has become suitable for taxonomic identification based on NGS [43], microbiome-wide association studies have been performed at an accelerated pace for finding the linkage between microbiota/microbiome and various diseases. While many DAF finding methods have been developed in recent years, it is not easy to choose the most appropriate method for real-world data analysis due to the differing assumptions of the methods and differing characteristics of metagenome data. In our study, we reviewed DAF finding methods and compared their performance with respect to type I error rates and their powers.

In order to compare these methods more systematically, we generated simulation data based on HMP data. Especially, we investigated the effect of sparsity on type I error and empirical power. Our comparison results showed that metagenomeSeq2, CLR Perm and ANCOM preserved the type I errors. The power of CLR Perm is highest among the methods preserved type I error; metagenomeSeq2 showed a similar power. However, there was a difference in the list of significant OTUs discovered by CLR Perm and metagenomeSeq2. In conclusion, we recommend using a combination of metagenomeSeq2 and CLR Perm for the analysis of metagenome data.

However, there are some limitations in our comparison study. First, HMP data may not reflect the characteristics of the original DAF finding studies because they were conducted on only healthy individuals. Second, we only investigated the effect of sparsity on the performances of each method. In addition to sparsity, the effects of some additional key parameters need to be investigated. Third, in our comparison only one taxon was assumed to be a causal taxon; it is necessary to consider methods with multiple causal taxa. Some methods such as ANCOM can take into account multiple taxa simultaneously, while others cannot.

Finally, Although the prediction model using 34 OTUs showed high AUC 0.84 in Baxter et al., the high AUC does not mean that 34 OTUs is truly associated with colorectal cancer. However, the analysis was performed with the expectation that OTUs to predict colorectal cancer with high accuracy would be highly associated with colorectal cancer.

Bibliography

[1] J. R. Marchesi, and J. Ravel, “The vocabulary of microbiome research: a proposal,” *Microbiome,* vol. 3, Jul 30, 2015.

[2] N. Shah, H. Tang, T. G. Doak, and Y. Ye, "Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics," *Biocomputing 2011*, pp. 165-176, 2011.

[3] L. Jost, “Entropy and diversity,” *Oikos,* vol. 113, no. 2, pp. 363-375, May, 2006.

[4] I. Cho, and M. J. Blaser, “APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease,” *Nature Reviews Genetics,* vol. 13, no. 4, pp. 260-270, Apr, 2012.

[5] Z. Gao, C. H. Tseng, B. E. Strober, Z. H. Pei, and M. J. Blaser, “Substantial Alterations of the Cutaneous Bacterial Biota in Psoriatic Lesions,” *Plos One,* vol. 3, no. 7, Jul 23, 2008.

[6] R. M. Peek, and M. J. Blaser, “Helicobacter pylori and gastrointestinal tract adenocarcinomas,” *Nature Reviews Cancer,* vol. 2, no. 1, pp. 28-37, Jan, 2002.

[7] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, “An obesity-associated gut microbiome with increased capacity for energy harvest,” *Nature,* vol. 444, no. 7122, pp. 1027-1031, Dec 21, 2006.

[8] Y. Chen, and M. Blaser, “Inverse associations of Helicobacter pylori with asthma and allergy,” *American Journal of Epidemiology,* vol. 165, no. 11, pp. S100-S100, Jun 1, 2007.

[9] W. S. Garrett, C. A. Gallini, T. Yatsunenko, M. Michaud, A. DuBois, M. L. Delaney, S. Punit, M. Karlsson, L. Bry, J. N. Glickman, J. I. Gordon, A. B. Onderdonk, and L. H. Glimcher, “Enterobacteriaceae Act in Concert with the Gut Microbiota to Induce Spontaneous and Maternally Transmitted Colitis,” *Cell Host & Microbe,* vol. 8, no. 3, pp. 292-300, Sep 16, 2010.

[10] C. Tana, Y. Umesaki, A. Imaoka, T. Handa, M. Kanazawa, and S. Fukudo, “Altered profiles of intestinal microbiota and organic acids may be the origin of symptoms in irritable bowel syndrome,” *Neurogastroenterology and Motility,* vol. 22, no. 5, pp. 512-+, May, 2010.

[11] M. Castellarin, R. L. Warren, J. D. Freeman, L. Dreolini, M. Krzywinski, J. Strauss, R. Barnes, P. Watson, E. Allen-Vercoe, R. A. Moore, and R. A. Holt, “Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma,” *Genome Research,* vol. 22, no. 2, pp. 299-306, Feb, 2012.

[12] Z. N. Wang, E. Klipfell, B. J. Bennett, R. Koeth, B. S. Levison, B. Dugar, A. E. Feldstein, E. B. Britt, X. M. Fu, Y. M. Chung, Y. P. Wu, P. Schauer, J. D. Smith, H. Allayee, W. H. W. Tang, J. A. DiDonato, A. J. Lusis, and S. L. Hazen, “Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease,” *Nature,* vol. 472, no. 7341, pp. 57-U82, Apr 7, 2011.

[13] J. A. Gilbert, R. A. Quinn, J. Debelius, Z. Z. Xu, J. Morton, N. Garg, J. K. Jansson, P. C. Dorrestein, and R. Knight, “Microbiome-wide association

studies link dynamic microbial consortia to disease," *Nature,* vol. 535, no. 7610, pp. 94-104, 2016.

[14] J. R. White, N. Nagarajan, and M. Pop, "Statistical methods for detecting differentially abundant features in clinical metagenomic samples," *PLoS computational biology,* vol. 5, no. 4, pp. e1000352, 2009.

[15] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop, "Differential abundance analysis for microbial marker-gene surveys," *Nature methods,* vol. 10, no. 12, pp. 1200-1202, 2013.

[16] X. Peng, G. Li, and Z. Liu, "Zero-inflated beta regression for differential abundance analysis with metagenomics data," *Journal of Computational Biology,* vol. 23, no. 2, pp. 102-110, 2016.

[17] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada, "Analysis of composition of microbiomes: a novel method for studying microbial composition," *Microbial ecology in health and disease,* vol. 26, no. 1, pp. 27663, 2015.

[18] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology,* vol. 15, no. 12, pp. 550, 2014.

[19] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, no. 1, pp. 139-140, 2010.

[20] C. Wu, J. Chen, J. H. Kim, and W. Pan, "An adaptive association test for microbiome data," *Genome Medicine,* vol. 8, May 19, 2016.

[21] J. Chen, and H. Z. Li, "Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis," *Annals of Applied Statistics,* vol. 7, no. 1, pp. 418-442, Mar, 2013.

[22] N. Zhao, J. Chen, I. M. Carroll, T. Ringel-Kulka, M. P. Epstein, H. Zhou, J. J. Zhou, Y. Ringel, H. Z. Li, and M. C. Wu, "Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test," *American Journal of Human Genetics,* vol. 96, no. 5, pp. 797-807, May 7, 2015.

[23] J. Chen, K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, R. G. Collman, F. D. Bushman, and H. Z. Li, "Associating microbiome composition with environmental covariates using generalized UniFrac distances," *Bioinformatics,* vol. 28, no. 16, pp. 2106-2113, Aug 15, 2012.

[24] M. C. Tsilimigras, and A. A. Fodor, "Compositional data analysis of the microbiome: fundamentals, tools, and challenges," *Annals of epidemiology,* vol. 26, no. 5, pp. 330-335, 2016.

[25] H. M. P. Consortium, "Structure, function and diversity of the healthy human microbiome," *nature,* vol. 486, no. 7402, pp. 207, 2012.

[26] B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, and J. H. Badger, "A framework for human microbiome research," *nature,* vol. 486, no. 7402, pp. 215, 2012.

[27] N. T. Baxter, M. T. Ruffin, M. A. M. Rogers, and P. D. Schloss, “Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions,” *Genome Medicine,* vol. 8, Apr 6, 2016.

[28] G. P. Young, "Fecal Immunochemical Test," *Encyclopedia of Cancer*, pp. 1103-1106: Springer, 2008.

[29] J. D. Storey, and R. Tibshirani, “Statistical significance for genomewide studies,” *Proceedings of the National Academy of Sciences,* vol. 100, no. 16, pp. 9440-9445, 2003.

[30] J. H. Zar, *Biostatistical analysis*: Pearson Education India, 1999.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1-38, 1977.

[32] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, *Bioinformatics and computational biology solutions using R and Bioconductor*: Springer Science & Business Media, 2006.

[33] J. N. Paulson, “Normalization and differential abundance analysis of metagenomic biomarker-gene surveys,” University of Maryland, College Park, 2015.

[34] D. M. Stasinopoulos, and R. A. Rigby, “Generalized additive models for location scale and shape (GAMLSS) in R,” *Journal of Statistical Software,* vol. 23, no. 7, Dec, 2007.

[35] J. Aitchison, “The Statistical-Analysis of Compositional Data,” *Journal of the Royal Statistical Society Series B-Methodological,* vol. 44, no. 2, pp. 139-177, 1982.

[36] E. Maza, “In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design,” *Frontiers in Genetics,* vol. 7, Sep 16, 2016.

[37] Y. Benjamini, 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)*, pp. 289-300, 1995.

[38] P. Louis, G. L. Hold, and H. J. Flint, “The gut microbiota, bacterial metabolites and colorectal cancer,” *Nature Reviews Microbiology,* vol. 12, no. 10, pp. 661-672, Oct, 2014.

[39] J. Ahn, R. Sinha, Z. H. Pei, C. Dominianni, J. Wu, J. X. Shi, J. J. Goedert, R. B. Hayes, and L. Y. Yang, “Human Gut Microbiome and Risk for Colorectal Cancer,” *Jnci-Journal of the National Cancer Institute,* vol. 105, no. 24, pp. 1907-1911, Dec, 2013.

[40] W. G. Chen, F. L. Liu, Z. X. Ling, X. J. Tong, and C. Xiang, “Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer,” *Plos One,* vol. 7, no. 6, Jun 28, 2012.

[41] J. R. Marchesi, B. E. Dutilh, N. Hall, W. H. M. Peters, R. Roelofs, A. Boleij, and H. Tjalsma, “Towards the Human Colorectal Cancer Microbiome,” *Plos One,* vol. 6, no. 5, May 24, 2011.

[42] T. T. Wang, G. X. Cai, Y. P. Qiu, N. Fei, M. H. Zhang, X. Y. Pang, W. Jia, S. J. Cai, and L. P. Zhao, “Structural segregation of gut microbiota between

colorectal cancer patients and healthy volunteers," *Isme Journal,* vol. 6, no. 2, pp. 320-329, Feb, 2012.

[43] B. J. Tindall, R. Rossello-Mora, H. J. Busse, W. Ludwig, and P. Kampfer, "Notes on the characterization of prokaryote strains for taxonomic purposes," *International Journal of Systematic and Evolutionary Microbiology,* vol. 60, pp. 249-266, Jan, 2010.

초 록

오늘 날 많은 연구자들이 인간 유전자 연구에 몰두하였지만 인간의 유전자만으로는 설명 할 수 없는 부분이 존재함을 인식하게 되었다. 이에 따라, 인체의 모든 유전자 보다, 수백 수천만 배 많은 제 2 의 유전자라고 불리는 인체 내 미생물 유전자(메타게놈)에 주목하기 시작하였으며 2008 년 미 국립 보건원 (National Institutes of Health)의 지원을 받아 휴먼 마이크로바이옴 프로젝트 (The Human Microbiome Project)가 시작되었다. 휴먼 마이크로바이옴 프로젝트를 통하여 건강한 사람의 미생물 군집 (microbiome)이 정의되었고. 이를 바탕으로 다양한 질병과 관련된 미생물군집이 규명되기 시작하였다. 여전히 밝혀지지 않은 많은 인간의 질병과 미생물간의 관계에 대한 연구가 지속됨에 따라 미생물과 인간의 다양한 표현형 (Phenotype)간의 연관성 (Association)을 연구하는 분야가 각광 받기 시작했다. 이러한 연구의 활성화로 인해 방대한 양의 메타게놈 자료를 다루기 위한 통계적 방법론에 대한 연구도 더불어 각광 받기 시작했다.

본 연구에서는 미생물과 인간의 다양한 표현형 간의 관계를 다루는 기존의 다양한 통계적 분석방법 (Statistical association test)들을 소개하고 중심화 로그 비 변환(Centered log-ratio transformation)과 로지스틱 회귀분석 (Logistic regression)을 활용한 통계적 분석 방법을 제시하였다. 시뮬레이션을 통하여 기존 방법과 새로 제시한 로지스틱 모형에 대한 통계적 검정력과 제 1종 오류를 비교하였으며. 대장암에 대한 실제 자료를 분석하였다. 그 결과, 새로 제시한 로지스틱 회귀분석 방법과 기존 방법 메타게놈식(metagenomeSeq)을 병행하여 사용시 대장암과 관련된 미생물들을 찾는데 성공적인 결과를 보였다.

주요어: 통계적 연관성 분석, 메타게놈, 표현형, 미생물

학 번: 2015-20510