Supplementary file 1

$YuanjingMa^1$, $YuanLuo^2$, $HongmeiJiang^1$

1. Department of Statistics, Northwestern University

2. Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University

In this supplementary file, we present the parameter estimation for zero-inflated Poisson mixture model and zero-inflated Negative Binomial model in Section 1 and 2 respectively. We discuss the multiple correction in Section 3. The detailed simulation studies and real data analysis are presented in Sections 4 and 5, respectively. The Supplementary Figures are attached at the end of this file.

1 Parameter estimation for zero-inflated Poisson mixture model

To estimate the parameters in zero-inflated Poisson regression model, we make the following notations:

$$\mathbf{B} = \begin{bmatrix} 1\ \kappa(1)\ log(S_1) \\ 1\ \kappa(2)\ log(S_2) \\ \ddots \\ 1\ \kappa(n)\ log(S_n) \end{bmatrix},$$

$$\boldsymbol{\beta_i} = [\beta_{0i}, \beta_{1i}, \beta_{2i}]^T,$$

$$\boldsymbol{\mu_i} = [\mu_{i1}, \mu_{i2}, ..., \mu_{in}]^T,$$

$$\mathbf{A} = \begin{bmatrix} 1\ log(S_1) \\ 1\ log(S_2) \\ \ddots \\ 1\ log(S_n) \end{bmatrix},$$

$$\boldsymbol{\alpha_i} = [\alpha_{0i}, \alpha_{1i}]^T,$$

$$\boldsymbol{\Pi_i} = [\pi_{i1}, \pi_{i2}, ..., \pi_{in}]^T,$$

$$\boldsymbol{c_i} = [c_{i1}, c_{i2}, ..., c_{in}]^T$$

and we use $\boldsymbol{B_j}$ and $\boldsymbol{A_j}$ to denote the j-th row of matrix $\mathbf{B}$ and $\mathbf{A}$.

The parameters can be estimated using EM algorithm combined with maximum likelihood estimation. The likelihood for ZIP mixture model is

$$L(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}; \boldsymbol{c_i}) = \prod_{c_{ij}=0} (\pi_{ij} + (1-\pi_{ij})e^{-\mu_{ij}}) \prod_{c_{ij}>0} ((1-\pi_{ij})e^{-\mu_{ij}} \mu_{ij}^{c_{ij}}/c_{ij}!),$$

which can be re-parameterized using Equation (4) and (5) in the paper:

$$\begin{aligned} L(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}; \boldsymbol{c_i}) &= \prod_{c_{ij}=0} (\frac{e^{\boldsymbol{A_j\alpha_i}}}{1+e^{\boldsymbol{A_j\alpha_i}}} + \frac{e^{-exp(\boldsymbol{B_j\beta_i})}}{1+e^{\boldsymbol{A_j\alpha_i}}}) \prod_{c_{ij}>0} \frac{e^{-exp(\boldsymbol{B_j\beta_i})+\mathbf{B}c_{ij}\beta_{ij}}}{(1+e^{\boldsymbol{A_j\alpha_i}})c_{ij}!} \\ &= \prod_{c_{ij}=0} \frac{e^{\boldsymbol{A_j\alpha_i}} + e^{-exp(\boldsymbol{B_j\beta_i})}}{1+e^{\boldsymbol{A_j\alpha_i}}} \prod_{c_{ij}>0} \frac{e^{-exp(\boldsymbol{B_j\beta_i})+\boldsymbol{B_j\beta_i}c_{ij}}}{(1+e^{\boldsymbol{A_j\alpha_i}})c_{ij}!} \end{aligned}.$$

The log-likelihood $l(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}; \boldsymbol{c_i})$ is:

$$\sum_{c_{ij}=0} log(e^{\boldsymbol{A_j\alpha_i}} + e^{-exp(\boldsymbol{B_j\beta_i})}) + \sum_{c_{ij}>0} (\boldsymbol{B_j\beta_i}c_{ij} - e^{\boldsymbol{B_j\beta_i}}) - \sum_{j=1}^{n} log(1+e^{\boldsymbol{A_j\alpha_i}}) - \sum_{c_{ij}>0} log(c_{ij}!).$$

It is complicated to use above log-likelihood directly for computing maximum a posteriori. Adding a latent variable and using EM algorithm will largely ease the computation. Suppose we have a latent variable

$$\Delta_{ij} = \begin{cases} 1 & c_{ij} \in \{0\} \\ 0 & c_{ij} \in Poisson(\mu_{ij}) \end{cases},$$

and $\boldsymbol{\Delta_i}$ represents $[\Delta_{i1}, \Delta_{i2}, ..., \Delta_{in}]^T$, then the likelihood with complete data ($\boldsymbol{c_i}$, $\boldsymbol{\Delta_i}$) is:

$$\begin{aligned} L_c(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}; \boldsymbol{c_i}, \boldsymbol{\Delta_i}) &= \prod_{j=1}^{n} \pi_{ij}^{\Delta_{ij}} (1-\pi_{ij})^{(1-\Delta_{ij})} (\frac{e^{-\mu_{ij}}\mu_{ij}^{c_{ij}}}{c_{ij}!})^{1-\Delta_{ij}} \\ &= \prod_{j=1}^{n} \frac{e^{\boldsymbol{A_j\alpha_i}\Delta_{ij}}}{(1+e^{\boldsymbol{A_j\alpha_i}})^{\Delta_{ij}+(1-\Delta_{ij})}} \frac{e^{-exp(\boldsymbol{B_j\beta_i})(1-\Delta_{ij})} e^{\boldsymbol{B_j\beta_i}c_{ij}(1-\Delta_{ij})}}{c_{ij}!^{(1-\Delta_{ij})}}. \end{aligned} \tag{1}$$

The corresponding log-likelihood of complete data ($\boldsymbol{c_i}$, $\boldsymbol{\Delta_i}$) is:

$$\begin{aligned} l_c(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}; \boldsymbol{c_i}, \boldsymbol{\Delta_i}) = &\sum_{j=1}^{n} [\boldsymbol{A_j\alpha_i}\Delta_{ij} - log(1+e^{\boldsymbol{A_j\alpha_i}})] + \\ &\sum_{j=1}^{n} (1-\Delta_{ij})(\boldsymbol{B_j\beta_i}c_{ij} - e^{\boldsymbol{B_j\beta_i}}) - \\ &\sum_{j=1}^{n} (1-\Delta_{ij}) log(c_{ij}!). \end{aligned} \tag{2}$$

The EM algorithm is applied iteratively to update Δ_{ij} and $\{\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}\}$, for the k-th iteration:

– E-step: estimate the expected value of Δ_{ij} given $\{\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}\}$.

$$
\begin{aligned}
z_{ij}^{(k)} &= E[\Delta_{ij}|\boldsymbol{\alpha_i}^{(k-1)}, \boldsymbol{\beta_i}^{(k-1)}] \\
&= Pr(\Delta_{ij} = 1|\alpha_{i0}^{(k-1)}, \alpha_{i1}^{(k-1)}, \beta_{i0}^{(k-1)}, \beta_{i1}^{(k-1)}, \beta_{i2}^{(k-1)}) \\
&= \frac{\pi_{ij}^{(k-1)} I_{\{0\}}(c_{ij})}{\pi_{ij}^{(k-1)} I_{\{0\}}(c_{ij}) + (1-\pi_{ij}^{(k-1)}) e^{-\mu_{ij}^{(k-1)}}} \\
&= \frac{e^{\boldsymbol{A_j}\boldsymbol{\alpha_i}^{(k-1)}}}{e^{\boldsymbol{A_j}\boldsymbol{\alpha_i}^{(k-1)}} + e^{-exp(\boldsymbol{B_j}\boldsymbol{\beta_i}^{(k-1)})}} \\
&= \begin{cases} (1 + e^{-\boldsymbol{A_j}\boldsymbol{\alpha_i}^{(k-1)} - exp(\boldsymbol{B_j}\boldsymbol{\beta_i}^{(k-1)})})^{-1} & if c_{ij} = 0 \\ 0 & if c_{ij} > 0 \end{cases}
\end{aligned},
$$

and $\boldsymbol{z_i}$ represents $[z_{i1}, z_{i2}, ..., z_{in}]^T$.

– M-step: maximize log-likelihood (2) with regards to $(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i})$ given z_{ij}.

The log-likelihood (2) can be partitioned as:

$$
\begin{aligned}
l_c(\boldsymbol{\alpha_i}^{(k)}, \boldsymbol{\beta_i}^{(k)}; \boldsymbol{c_i}, \boldsymbol{z_i}^{(k)}) = l_c(\boldsymbol{\alpha_i}^{(k)}; \boldsymbol{c_i}, \boldsymbol{z_i}^{(k)}) + l_c(\boldsymbol{\beta_i}^{(k)}; \boldsymbol{c_i}, \boldsymbol{z_i}^{(k)}) \\
- \sum_{j=1}^{n} (1 - z_{ij}^{(k)}) log(c_{ij}!),
\end{aligned} \tag{3}
$$

where

$$
l_c(\boldsymbol{\alpha_i}^{(k)}; \boldsymbol{c_i}, \boldsymbol{z_i}^{(k)}) = \sum_{j=1}^{n} [\boldsymbol{A_j}\boldsymbol{\alpha_i}^{(k)} z_{ij}^{(k)} - log(1 + e^{\boldsymbol{A_j}\boldsymbol{\alpha_i}^{(k)}})], \tag{4}
$$

$$
l_c(\boldsymbol{\beta_i}^{(k)}; \boldsymbol{c_i}, \boldsymbol{z_i}^{(k)}) = \sum_{j=1}^{n} (1 - z_{ij}^{(k)})(\boldsymbol{B_j}\boldsymbol{\beta_i}^{(k)} c_{ij} - e^{\boldsymbol{B_j}\boldsymbol{\beta_i}^{(k)}}). \tag{5}
$$

The partition shows an advantage that: $l_c(\boldsymbol{\alpha_i}; \boldsymbol{c_i}, \boldsymbol{\Delta_i})$ and $l_c(\boldsymbol{\beta_i}; \boldsymbol{c_i}, \boldsymbol{\Delta_i})$ can be maximized with regards to $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$ separately. Detailedly, $\boldsymbol{\alpha_i}^{(k)}$ can be found using weighted logistic regression. Suppose the number of zero counts among $\{c_{ij}\}$ is n_0, and we denote these zero counts as $\{c'_{i1}, c'_{i2}, ..., c'_{in_0}\}$. The parameters $\boldsymbol{\alpha_i}$ of (4) are the same as parameters in the weighted logistic regression with response variables $\boldsymbol{c_i^*} = (c_{i1}, c_{i2}, ..., c_{in}, c'_{i1}, c'_{i2}, ..., c'_{in_0})$, covariates matrix $\boldsymbol{A}_*^T = (\boldsymbol{A}_1^T, \boldsymbol{A}_2^T, ..., \boldsymbol{A}_n^T, \boldsymbol{A}_1'^T, \boldsymbol{A}_2'^T, ..., \boldsymbol{A}_{n_0}'^T)$, $\boldsymbol{\pi}_i^* = (\pi_{i1}, \pi_{i2}, ..., \pi_{in}, \pi'_{i1}, \pi'_{i2}, ..., \pi'_{in_0})$ and corresponding weights $\boldsymbol{w}^{(k)} = (1\text{-}z_{i1}^{(k)}, ..., 1\text{-}z_{in}^{(k)}, z_{i_1}^{'(k)}, ..., z_{in_0}^{'(k)})$. The computation of $\boldsymbol{\beta}_i^{(k)}$ can be derived using a weighted log-linear Poisson regression with weight $1\text{-}\boldsymbol{z_i}^{(k)}$. For more details, readers can refer to Lambert (1992).

2 Parameter estimation for zero-inflated Negative Binomial model

The likelihood function of ZINB is

$$L(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}, \phi_i; \boldsymbol{c_i}) = \prod_{c_{ij}=0} (\pi_{ij} + (1-\pi_{ij})(1+\phi_i\mu_{ij})^{-1/\phi_i}) \prod_{c_{ij}>0} ((1-\pi_{ij})\frac{\Gamma(c_{ij}+1/\phi_i)}{\Gamma(1/\phi_i)c_{ij}!}(1+\phi_i\mu_{ij})^{-1/\phi_i}(1+\frac{1}{\phi_i\mu_{ij}})^{-c_{ij}}).$$

Re-parameterized using Equation (4) and (5) in the paper, we get

$$\begin{aligned} L(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}, \phi_i; \boldsymbol{c_i}) = & \prod_{c_{ij}=0} (\frac{e^{\boldsymbol{A_j\alpha_i}}}{1+e^{\boldsymbol{A_j\alpha_i}}} + \frac{(1+\phi_i e^{\boldsymbol{B_j\beta_i}})^{-1/\phi_i}}{1+e^{\boldsymbol{A_j\alpha_i}}}) \\ & \prod_{c_{ij}>0} (\frac{1}{1+e^{\boldsymbol{A_j\alpha_i}}}\frac{\Gamma(c_{ij}+1/\phi_i)}{\Gamma(1/\phi_i)c_{ij}!}(1+\phi_i e^{\boldsymbol{B_j\beta_i}})^{(-1/\phi_i)}(1+\frac{1}{\phi_i e^{\boldsymbol{B_j\beta_i}}})^{-c_{ij}}). \end{aligned} \tag{6}$$

The log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}, \phi_i; \boldsymbol{c_i}) = & \sum_{j=1}^{n}(-log(1+e^{\boldsymbol{A_j\alpha_i}}))+ \\ & \sum_{c_{ij}=0} log(e^{\boldsymbol{A_j\alpha_i}} + (1+\phi_i e^{\boldsymbol{B_j\beta_i}})^{-1/\phi_i})+ \\ & \sum_{c_{ij}>0} (log\Gamma(c_{ij}+1/\phi_i) - log\Gamma(1/\phi_i) - log(c_{ij}!))+ \\ & \sum_{c_{ij}>0} (-\frac{1}{\phi_i}log(1+\phi_i e^{\boldsymbol{B_j\beta_i}}) - c_{ij}log(1+\frac{1}{\phi_i e^{\boldsymbol{B_j\beta_i}}})). \end{aligned} \tag{7}$$

Parameters $\{\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}, \phi_i\}$ can be estimated using BFGS algorithm. It is a quasi-Newton optimization method which only requires the first derivatives of log-likelihood function. For more details, readers can refer to Nocedal and Wright (pp. 136-142) (second edition).

3 Multiple testing correction

Here, we intuitively justify the separate application of BH procedure. By applying BH procedure at level q to two different test models separately, we still can roughly control the overall false positive rate at level q. Let TP and FP be the total number of true positives and false positives among the statistically significant OTUs. Suppose the number of true positives and false positives in ZIP and ZINB regression models are TP_1, FP_1, and TP_2, FP_2, which are unobservable random variables, respectively. When the number of tests is large, according to Benjamini (2008), we have $E(\frac{FP}{FP+TP}) \approx \frac{E(FP)}{E(FP+TP)}$. With separate correction, we have $\frac{E(FP_1)}{E(FP_1+TP_1)} \leq q$ and $\frac{E(FP_2)}{E(FP_2+TP_2)} \leq q$.

The overall FDR is

$$\begin{aligned} E\left(\frac{FP_1 + FP_2}{FP_1 + TP_1 + FP_2 + TP_2}\right) &\approx \frac{E(FP_1 + FP_2)}{E(FP_1 + TP_1 + FP_2 + TP_2)} \\ &\leq \frac{q \cdot E(FP_1 + TP_1) + q \cdot E(FP_2 + TP_2)}{E(FP_1 + TP_1 + FP_2 + TP_2)} \\ &\leq q. \end{aligned}$$

4 Simulation

To compare the performance of RioNorm2 to that of DESeq, DESeq2, metagenomeSeq, RAIDA, and Omnibus, we run simulation studies where we can control the settings and the true differential abundance of each OTU. Specifically, we will focus on evaluating the impact of sample size, library size, and effect size on the performance of different methods. We explore two different simulation settings which are based on different distributions. The first simulation setting is adopted from McMurdie and Holmes (2014); the second one is based on the Dirichlet-Multinomial distribution.

4.1 Simulation setting 1 - McMurdie and Holmes (2014)

We adopt similar simulation setup as simulation setting B in McMurdie and Holmes (2014). To mimic OTU counts observed in the real world, we choose the real dataset "GlobalPattern" in R package "phyloseq" (McMurdie and Holmes, 2013) as our simulation template. "GlobalPattern" dataset contains 26 samples collected from 9 different environments (feces, freshwater, freshwater(creek), mock, ocean, sediment (estuary), skin, soil, and tongue) with more than two samples in each environment. We have 3 parameters: sample size per condition (which takes value 25, 35, and 50), median library size (which takes value 5000, 10000, and 50000), and effect size (which takes value 2, 3, 4, and 5). The number of truly differentially abundant OTUs is fixed to be 30. For each combination of parameters and environments, 10 OTU tables are generated. Readers can refer to Figure S1 for a detailed diagram illustration.

Firstly, OTU counts are summed across all samples in one environment of the "GlobalPattern" dataset to derive a single "pseudo-population" OTU table. Then, we randomly pick a library size for each sample from 26 library sizes in the "GlobalPattern" dataset with replacement. These library sizes are scaled to have the pre-determined median library size. Secondly, the OTU counts of each simulated sample are generated by sampling from the "pseudo-population" OTU table using multinomial distribution. Thirdly, to add artificial effect, the simulated samples of an environment are divided into two equally-sized conditions, control and test, and the effect size is multiplied to the count values of a random subset of OTUs in the test condition only. Each of these randomly

perturbed OTUs is differentially abundant between two conditions. The above process for generating OTU tables will be repeated 10 times for each combination of the 3 parameters (sample size, library size, and effect size) and the 9 environments. Therefore, for each combination of sample size per condition, median library size, and effect size, we will have 90 simulated OTU tables (10 times for each of the 9 environments) which can be used to calculate the mean and standard deviation of evaluation metrics. For implementation details, we have done two trimming steps in the above process. We want to avoid performing the differential abundance test on OTUs with very low prevalence level, since there is not enough power to make conclusions on these OTUs. First, we trim the OTU table to only keep those relatively prevalent ones. We rank all OTUs inversely based on their prevalence (defined as how many times an OTU is observed in samples), if tie, based on their count sum across samples. Only top abundant 1000 OTUs of each environment will be kept for deriving the "pseudo-population" OTU table. Second, before multiplying any OTU with the effect size, we further trim OTUs such that the prevalence of each remaining OTU is bigger than 0.05 and the sum of each OTU counts across samples is bigger than 0.5 times the number of samples. In order to reduce computational time, instead of using the iterative way to search for riOTUs (Algorithm 1), we fix h value to be 0.03th quantile of dissimilarity distribution.

4.2 Simulation setting 2 - Dirichlet-Multinomial distribution

To incorporate more variation, we conduct the second simulation study based on Dirichlet-Multinomial distribution. We modify the simulation setting 1 by adding one extra step: using the Dirichlet distribution to generate the probability parameters for the multinomial distribution for each sample. We have 4 parameters: sample size per condition (which takes value 25, 35, and 50), median library size (which takes value 5000, 10000, and 50000), effect size (which takes value 2, 3, 4, and 5), and the proportion of DA-OTUs (which varies from 10% to 15%). For each combination of parameters, 10 OTU tables are generated. To mimic the real-world dataset and reduce the complexity of the simulation, we derive the parameters from real samples in the mock environment of "GlobalPattern" dataset instead of using samples from all the nine environments.

Here are the details. Firstly, OTU counts are summed across all samples in the mock environment to derive a single "pseudo-population". Then, we randomly pick a library size for each sample from 26 library sizes in the "GlobalPattern" dataset with replacement. These library sizes are scaled to have the pre-determined median library size. Secondly, for each sample, we generate the probabilities of OTUs using Dirichlet distribution with "pseudo-population" as its parameters. Then the OTU counts of each simulated sample are generated by the multinomial distribution.

Thirdly, to add artificial effect, the simulated samples of an OTU-table are divided into two equally-sized conditions, control and test, and the effect size is multiplied to the count values of a randomly selected subset of OTUs in the test condition only. Each of these randomly perturbed OTUs is differentially abundant between two conditions. We do the same trimming process as mentioned in the simulation setting 1. In order to reduce the computational time, we fix the h value as 0.025th quantile of dissimilarity distribution.

False discovery rate (FDR) and power are used to evaluate the performance of different methods. Figures S15 and S16 show the FDR and power respectively for the Dirichlet-Multinomial based simulation across all settings with 10% of DA-OTUs. Figurea S17 and S18 show the FDR and power when the proportion of DA-OTUs is 15%. In all these four plots, the horizontal axis indicates various effect sizes; row panels represent the median library size and column panels represent the number of samples per condition. Each curve traces the mean FDR and power across all replicates. RioNorm2 and RAIDA can control FDR around 5% across all simulation settings; other four approaches have severe FDR inflation that are near or much higher than 20%. This is within our expectation since these four approaches capture the proportion change of non DA-OTUs due to the abundance change of true DA-OTUs and tend to detect more false positives when total abundances of DA-OTUs across different conditions are large. When the effect size is small, compared to all 5 other approaches, RioNorm2 constantly yields high detection power. When the effect size is large, RioNorm2 has comparable results as RAIDA. In summary, RioNorm2 and RAIDA perform the best when the abundance change of DA-OTUs does not impact or suppress other OTUs abundance. They achieve high power at the same time controlling for FDR. RioNorm2 surpasses RAIDA under small effect sizes.

Besides the above mentioned five approaches, we also compare RioNorm2 to ANCOM (Mandal15) which is becoming a standard method due to its implementation in Qiime2. Since the computation time of ANCOM is long (usually taking 20 minutes for an OTU table with around 1000 features and 50 samples using MacBook Pro with 2.8GHz intel Core i7), we only use a subset of data generated using the simulation setting 2. We tune the ANCOM hyper-parameter and choose the one that gives the best results. Figures S19, S20, S21 and S22 show that ANCOM has very low FDR value (close to 0), which trades off its ability to detect DA-OTUs effectively. Compared to ANCOM, RioNorm2 has FDR rate lower but close to 5% and gives high detection power across various effect sizes.

5 Real data analysis

5.1 Metastatic melanoma data

To test the performance of our framework, we consider a newly published differential abundance study which is related to the cancer treatment efficiency by Matson et al.(2018). They found that anti-PD-1 "based immunotherapy has a major beneficial impact on only a subset of cancer patients. Among the variables that could contribute to the different treatment effects is the different composition of patients' microbiome. Matson et al. (2018) analyzed baseline stool samples from metastatic melanoma patients before immunotherapy treatment through integration of 16S ribosomal RNA gene sequencing, metagenomic shotgun sequencing, and quantitative polymerase chain reaction for selected bacteria. A significant association was observed between commensal microbial composition and clinical response.

Here, we will focus on the microbiome data collected through 16S rRNA targeted sequencing. The output OTU table consists of 10385 OTUs and 42 samples. Among 42 patients, 16 of them have responded to the treatment while the other 26 are non-responders. Matson et al.(2018) applied the nonparametric permutation test for differential abundance analysis. The 16S sequencing revealed 62 OTUs of differential abundance in responders versus non-responders using p-values without multiple testing correction. Since the authors did not apply the multiple testing correction, a large amount of differentially abundant OTUs will be false positive. If we apply Benjamini-Hochberg procedure to the raw p-values from permutation test, no OTUs are statistically differential between responders and non-responders. If we compare the differentially abundant OTUs detected from RioNorm2 to those selected based on unadjusted p-values from permutation test, 4 OTUs are found in common. Therefore, our framework can increase the power of differential abundance analysis.

We also compare our results with those derived from RAIDA, Omnibus and DESeq2. RAIDA detects 12 DA-OTUs; 2 are shared with the permutation test, but there is no overlap with RioNorm2. Omnibus detects 53 DA-OTUs; 6 are shared with RioNorm2, 3 are shared with RAIDA and 20 are shared with the permutation test. DESeq2 detects 3 DA-OTUs; among them, 1 is shared with RioNorm2 and the permutation test, 2 are shared with Omnibus, and no common DA-OTUs are shared with RAIDA.

5.2 Inflammatory bowl disease (IBD) data

The IBD dataset can be downloaded from Qiita with study ID 11336 (https://qiita.ucsd.edu). There are 4323 OTUs and 67 samples, of which 18 samples are from the control group, the other

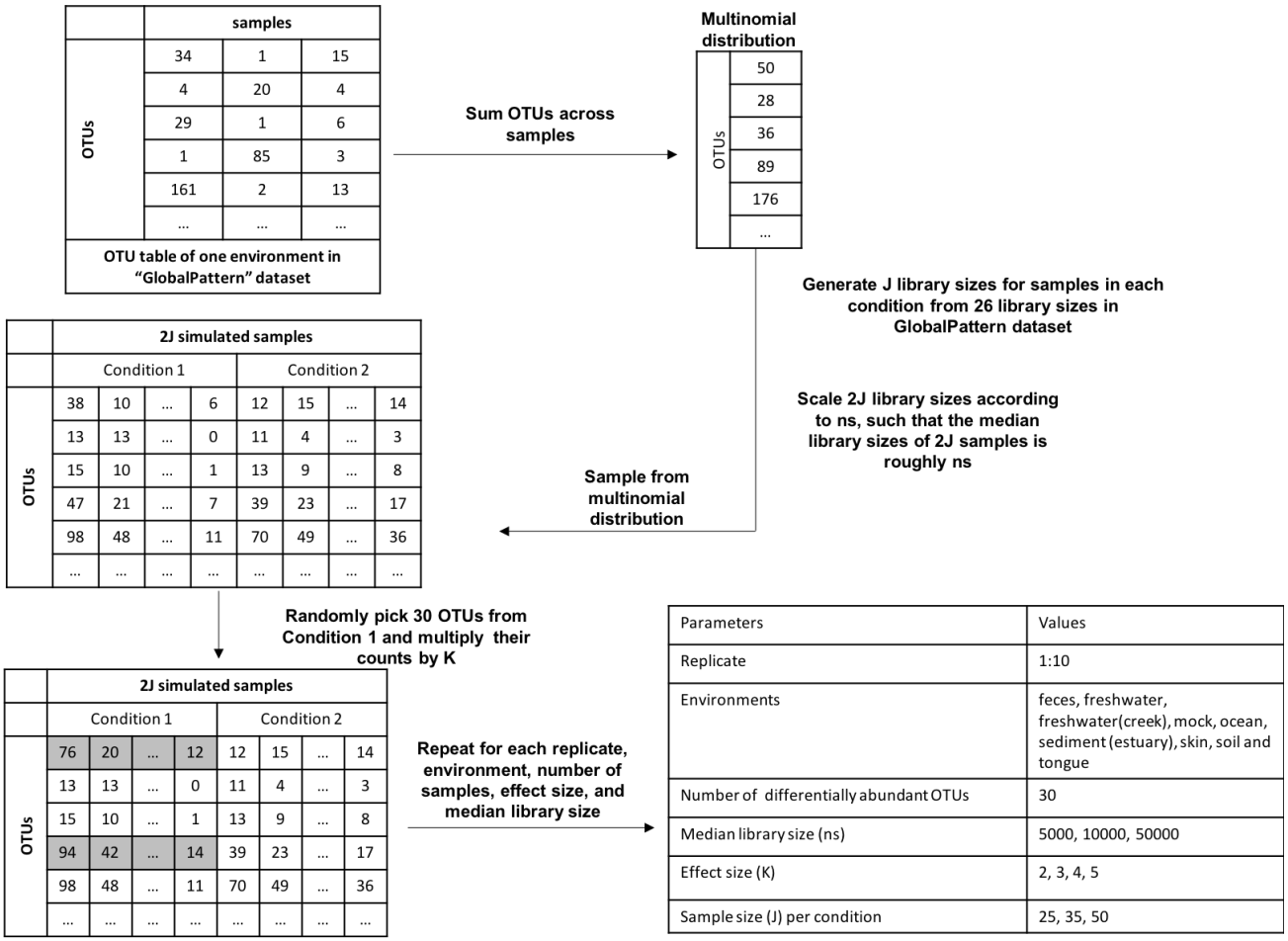
49 samples are from the disease group. OTUs that are observed in at least 80% of samples with average count greater than 5 are selected to build microbiome network. Applying algorithm 1 in the paper, RioNorm2 detects 6 riOTUs. The taxonomy information shows that all riOTUs share the same evolution path from kingdom to order (see Figure S27). RioNorm2 has advantages to detect biologically meaningful riOTUs for normalization. Size factors are calculated using Equation (2) in the paper. We found that 3 samples from the disease group have zero size factors; therefore, we exclude them from the later differential abundance test. We keep OTUs shown up in at least 20% of samples for differential abundance test. 953 OTUs are remained after filtering; among them, 897 OTUs are over-dispersed and 56 OTUs are non-overdispersed.

We apply the BH control of FDR separately on ZIP and ZINB tests at level 0.05. After correction, 25 OTUs are differentially abundant; a1l of them are overdispersed OTUs. Among the 25 DA-OTUs detected by the RioNorm2 test, 18 are taxa in the order Clostridiales (Figure S28). The consistency of large number of Clostridiales associations is remarkable. Besides, among these 25 OTUs, 9 OTUs are observed in at least 30 samples and 14 OTUs are observed in at least 20 samples. Box plots of top observed 9 DA-OTUs are plotted (Figure S29). We also compare our results with those derived from the RAIDA, Omnibus and DESeq2. RAIDA detects 7 DA-OTUs and none of them are shared with the RioNorm2 test. Omnibus detects 65 DA-OTUs; 15 are shared with RioNorm2 and 7 are shared with RAIDA. DESeq2 detects 66 DA-OTUs; among them, 2 are shared with RioNorm2, 1 is shared with RAIDA and 5 are shared with Omnibus.

References

1. Lambert, Diane. "Zero-inflated Poisson regression, with an application to defects in manufacturing." Technometrics 34, no. 1 (1992): 1-14.
2. Wright, S., and Nocedal, J. (1999). Numerical optimization. Springer Science, 35(67-68), 7.
3. Benjamini, Y. (2008). Comment: Microarrays, empirical Bayes and the two-groups model. Statistical Science, 23(1), 23-28.

List of Figures



Parameters	Values
Replicate	1:10
Environments	feces, freshwater, freshwater(creek), mock, ocean, sediment (estuary), skin, soil and tongue
Number of differentially abundant OTUs	30
Median library size (ns)	5000, 10000, 50000
Effect size (K)	2, 3, 4, 5
Sample size (J) per condition	25, 35, 50

Fig. S1. Detailed graphical summary of simulation procedure

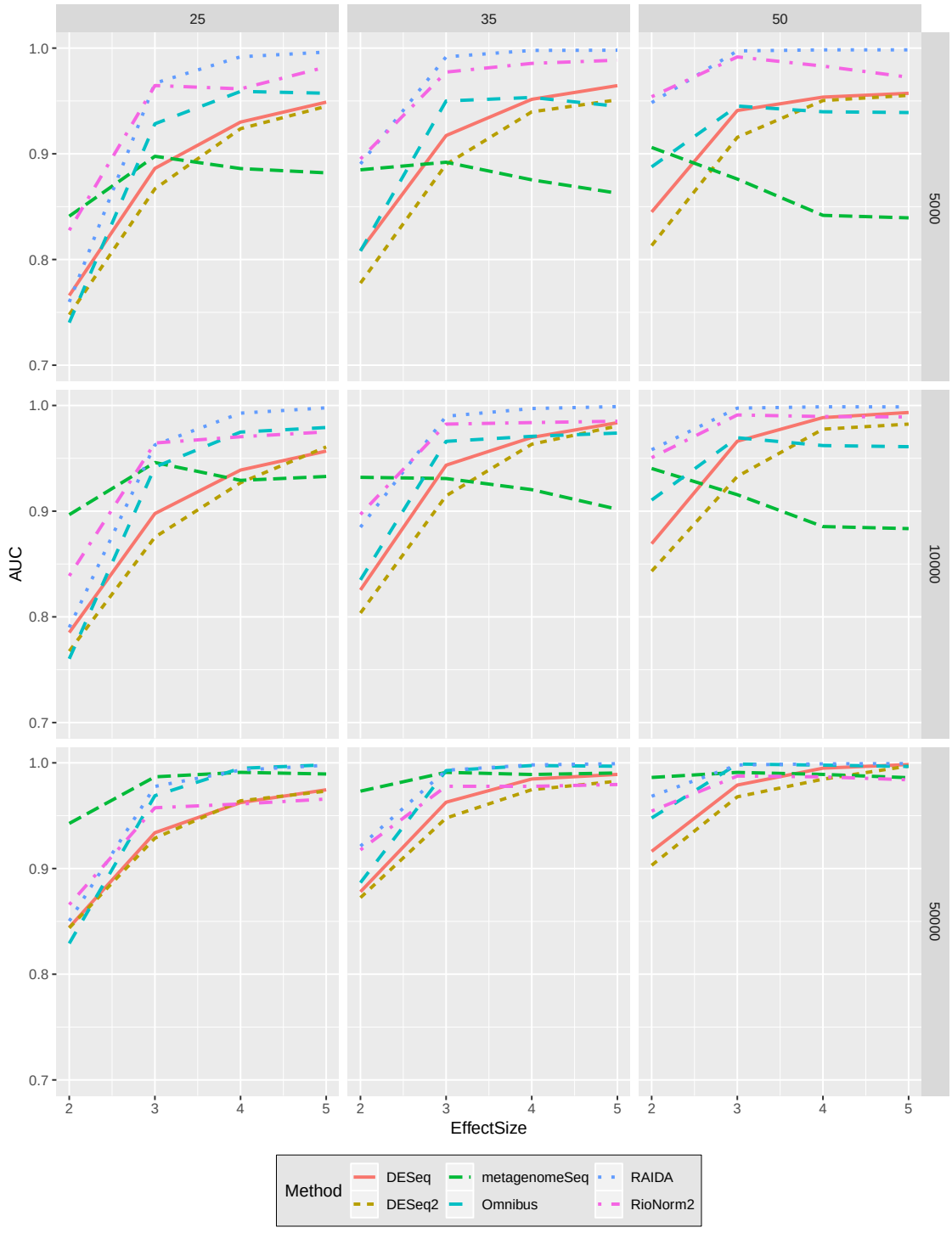


Fig. S2. Comparison of different methods in terms of AUC for various effect sizes in the simulation setting 1. Panel rows represent the median library size, and panel columns represent the sample size per condition.

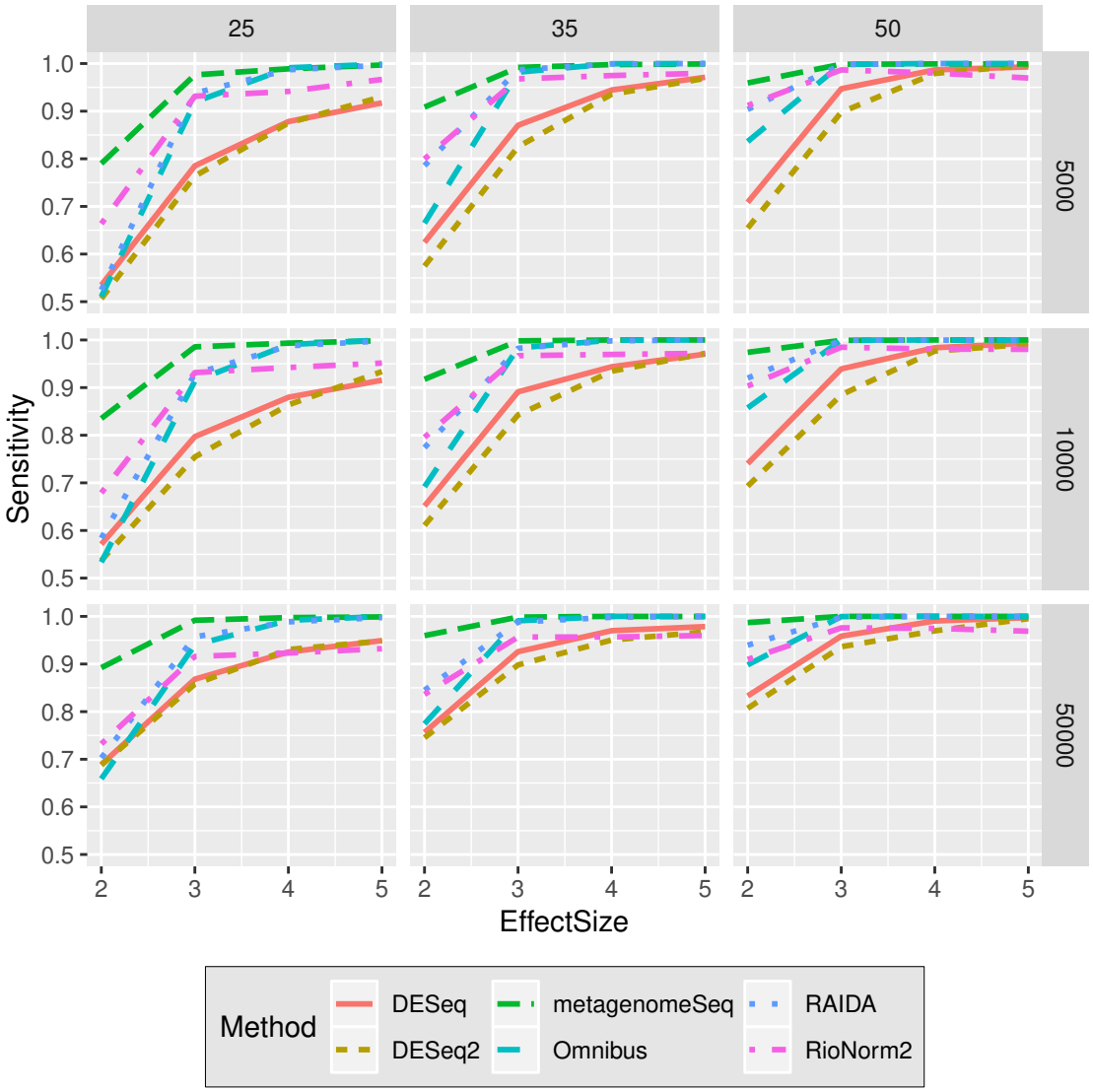


Fig. S3. Comparison of different methods in terms of sensitivity for various effect sizes in the simulation setting 1. Panel rows represent the median library size, and panel columns represent the sample size per condition.

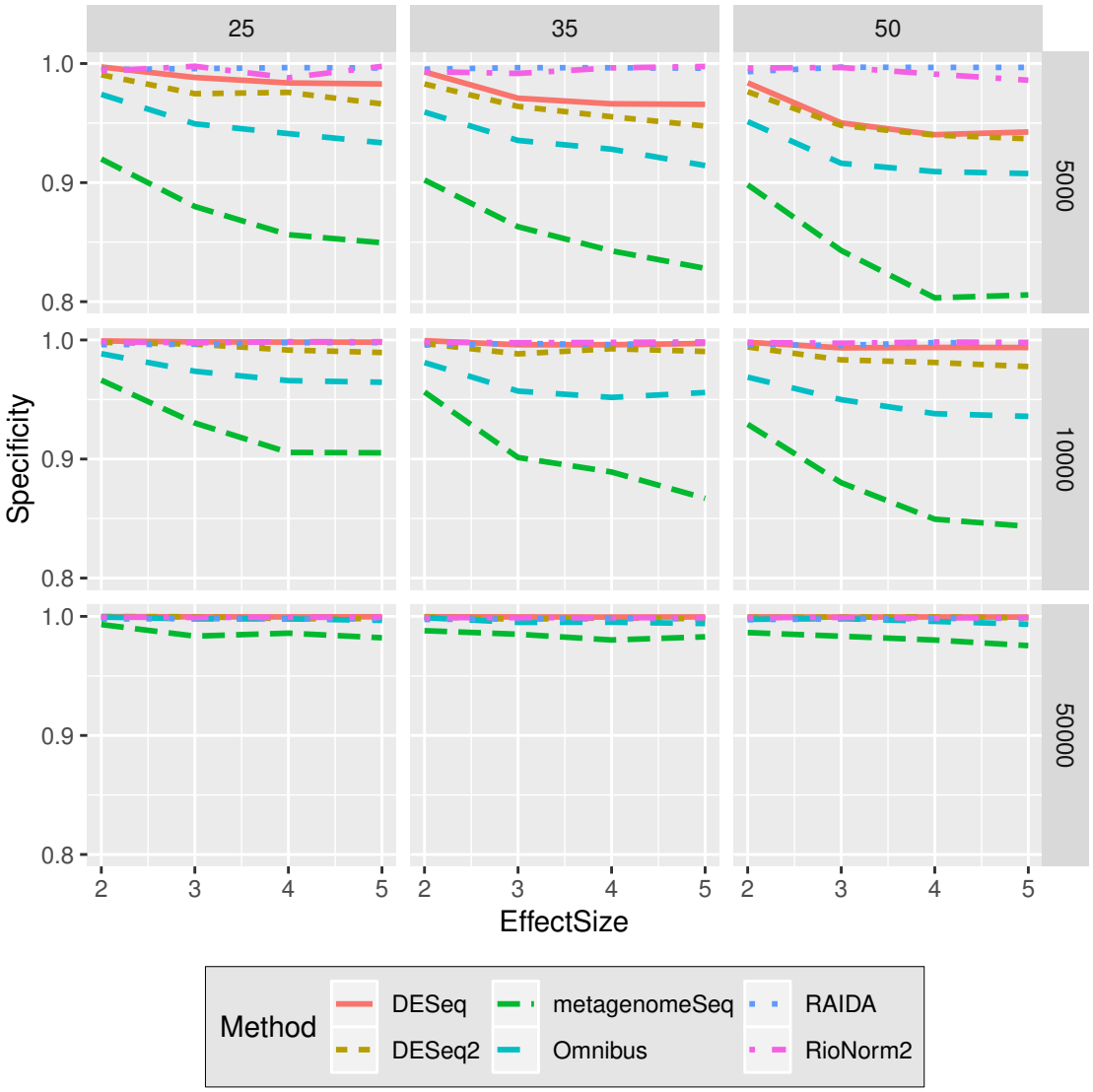


Fig. S4. Comparison of different methods in terms of specificity for various effect sizes in the simulation setting 1. Panel rows represent the median library size, and panel columns represent the sample size per condition.

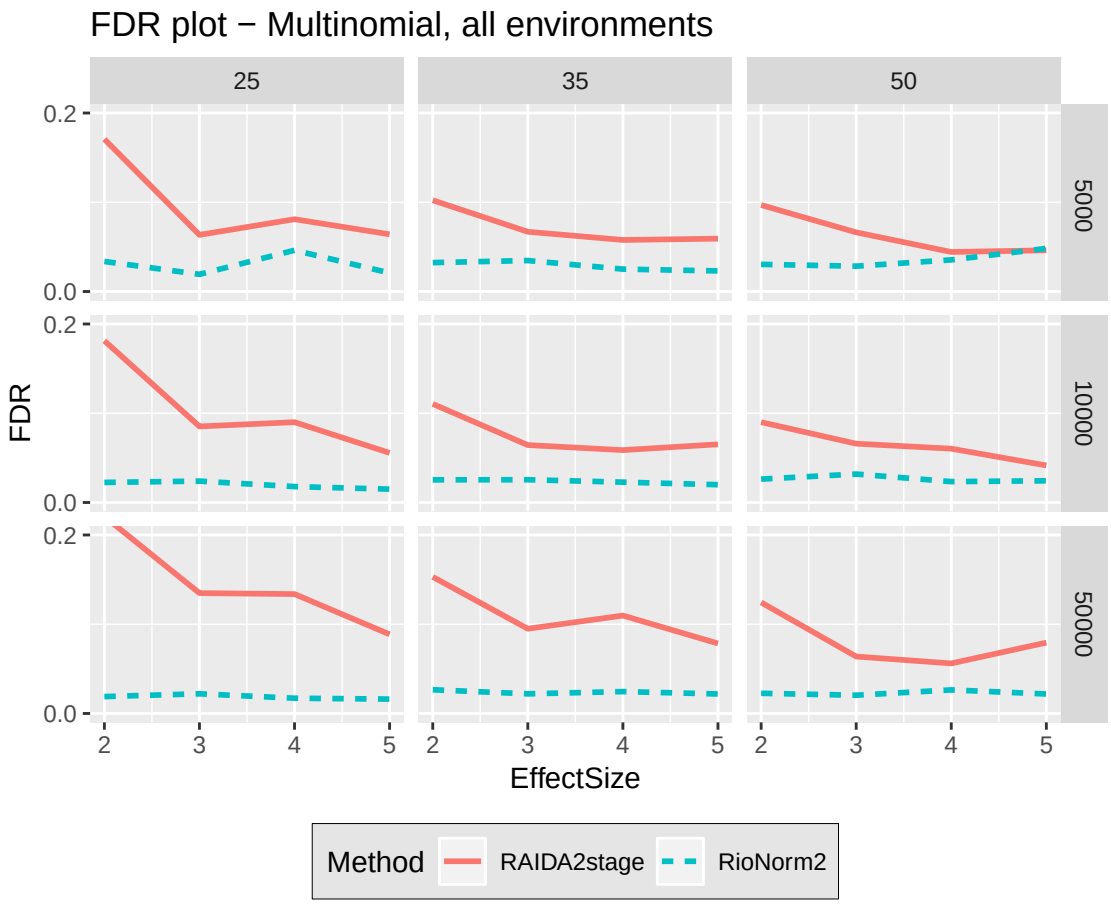


Fig. S5. Comparison of normalization step of RioNorm2 and RAIDA in terms of FDR for various effect sizes in the simulation setting 1. Panel rows represent the median library size, and panel columns represent the sample size per condition.

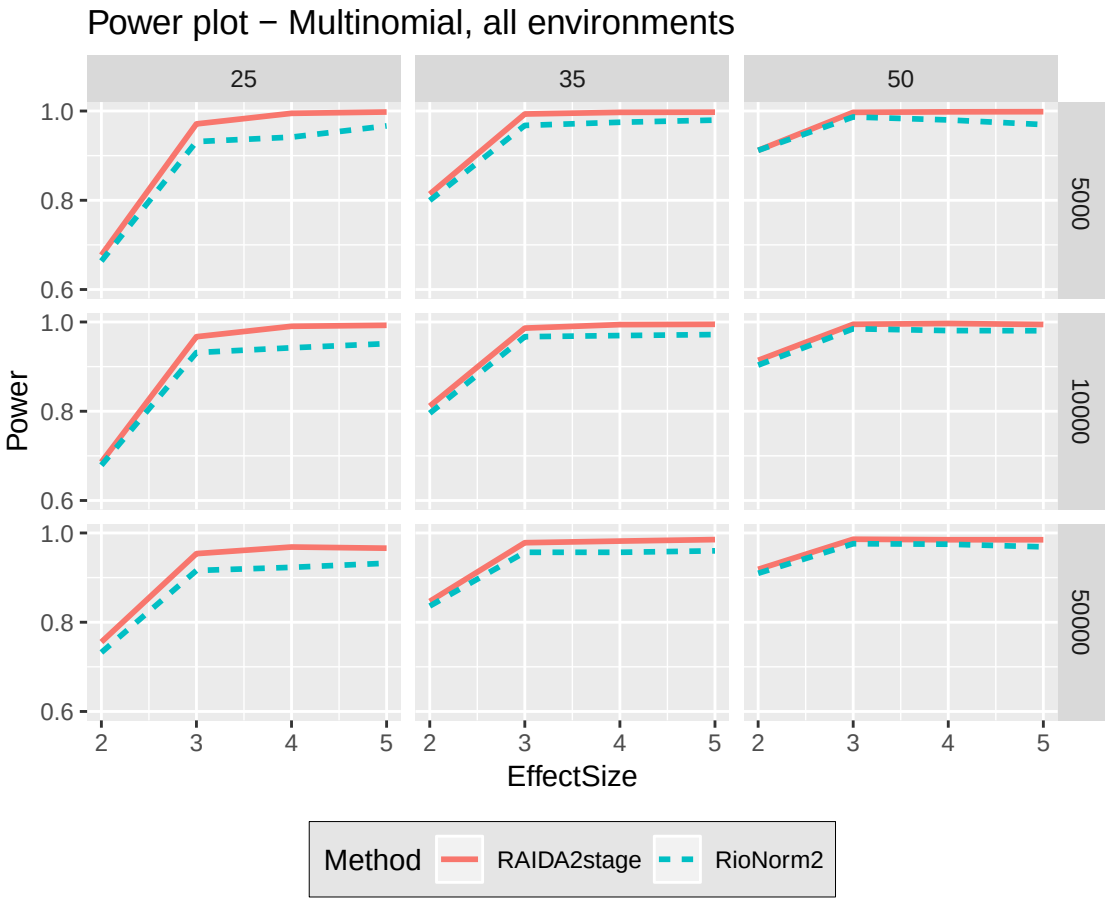


Fig. S6. Comparison of normalization step of RioNorm2 and RAIDA in terms of power for various effect sizes in the simulation setting 2. Panel rows represent the median library size, and panel columns represent the sample size per condition.

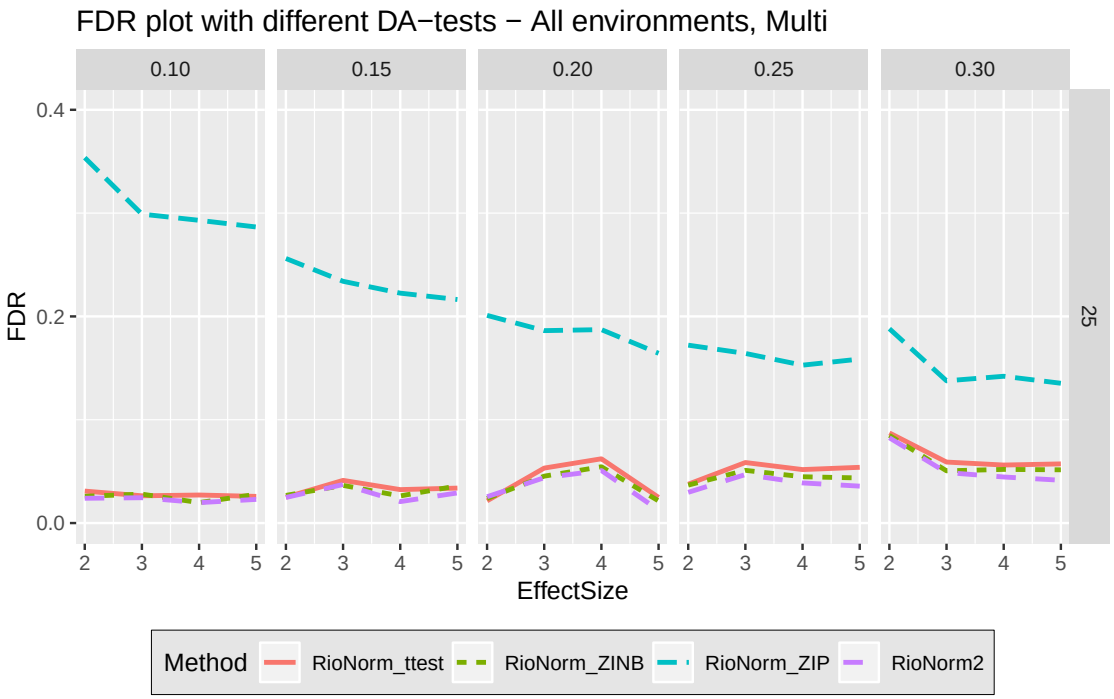


Fig. S7. Comparison of the 2-stage test in RioNorm2 to other differential abundance tests in terms of FDR for various effect sizes using a subset of data in the simulation setting 1. Panel columns represent different proportions of DA-OTUs. The number of samples per condition is 25 and the median library size is 5000.

Fig. S8. Comparison of the 2-stage test in RioNorm2 to other differential abundance tests in terms of power for various effect sizes using a subset of data in the simulation setting 1. Panel columns represent different proportions of DA-OTUs. The number of samples per condition is 25 and the median library size is 5000.

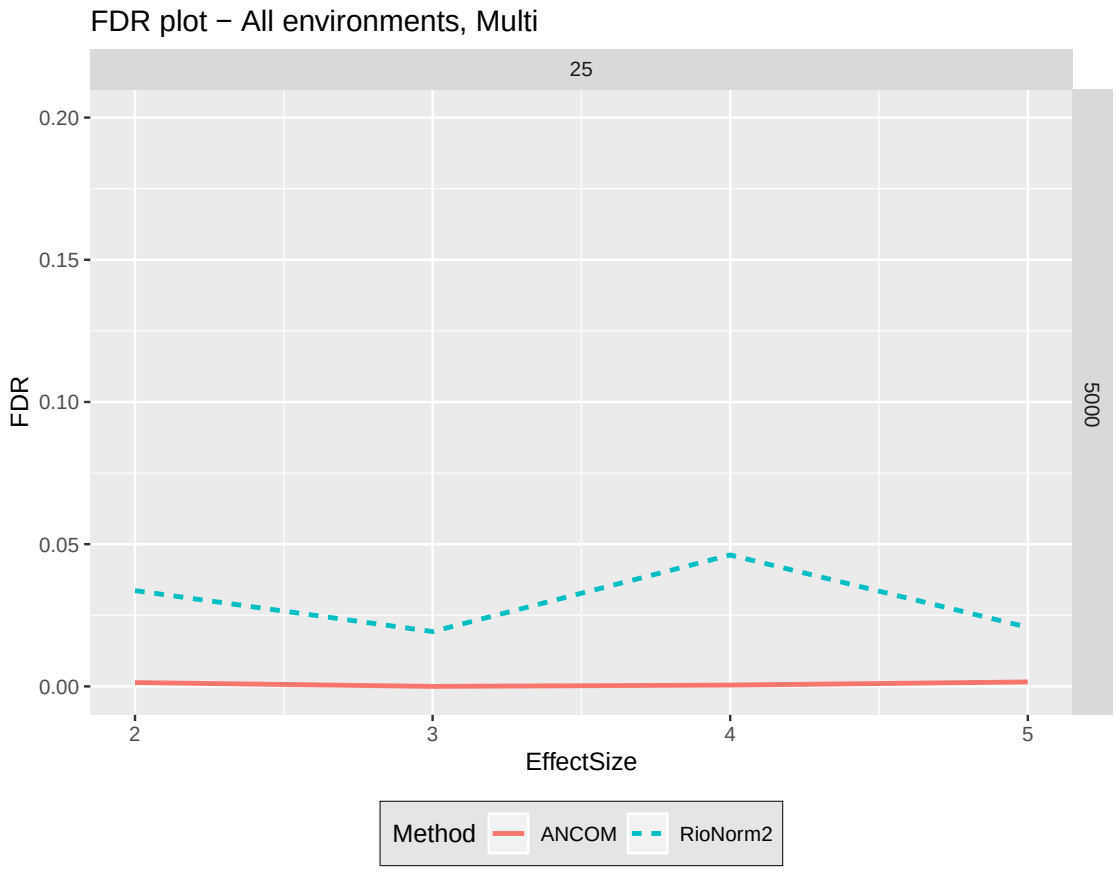


Fig. S9. Comparison of RioNorm2 and ANCOM in terms of FDR for various effect sizes using a subset of data in the simulation setting 1. The number of samples per condition is 25 and the median library size is 5000.

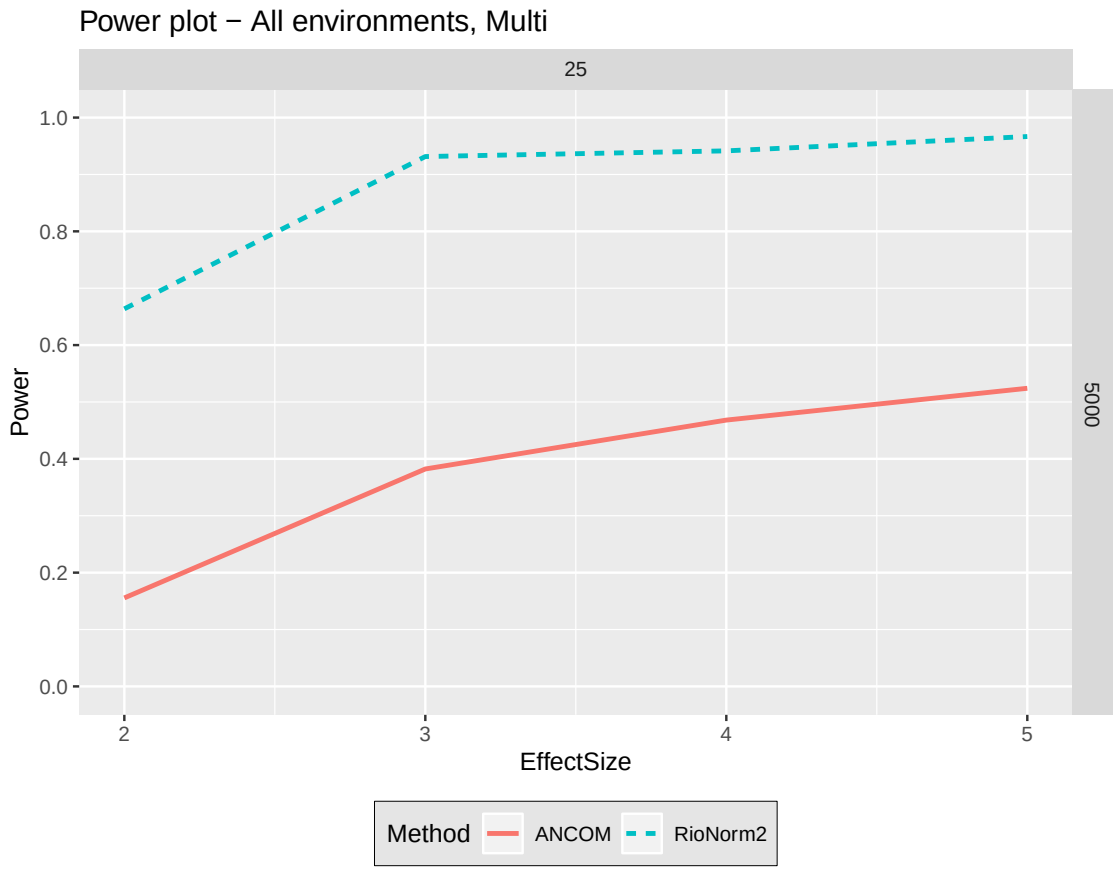


Fig. S10. Comparison of RioNorm2 and ANCOM in terms of power for various effect sizes using a subset of data in the simulation setting 1. The number of samples per condition is 25 and the median library size is 5000.

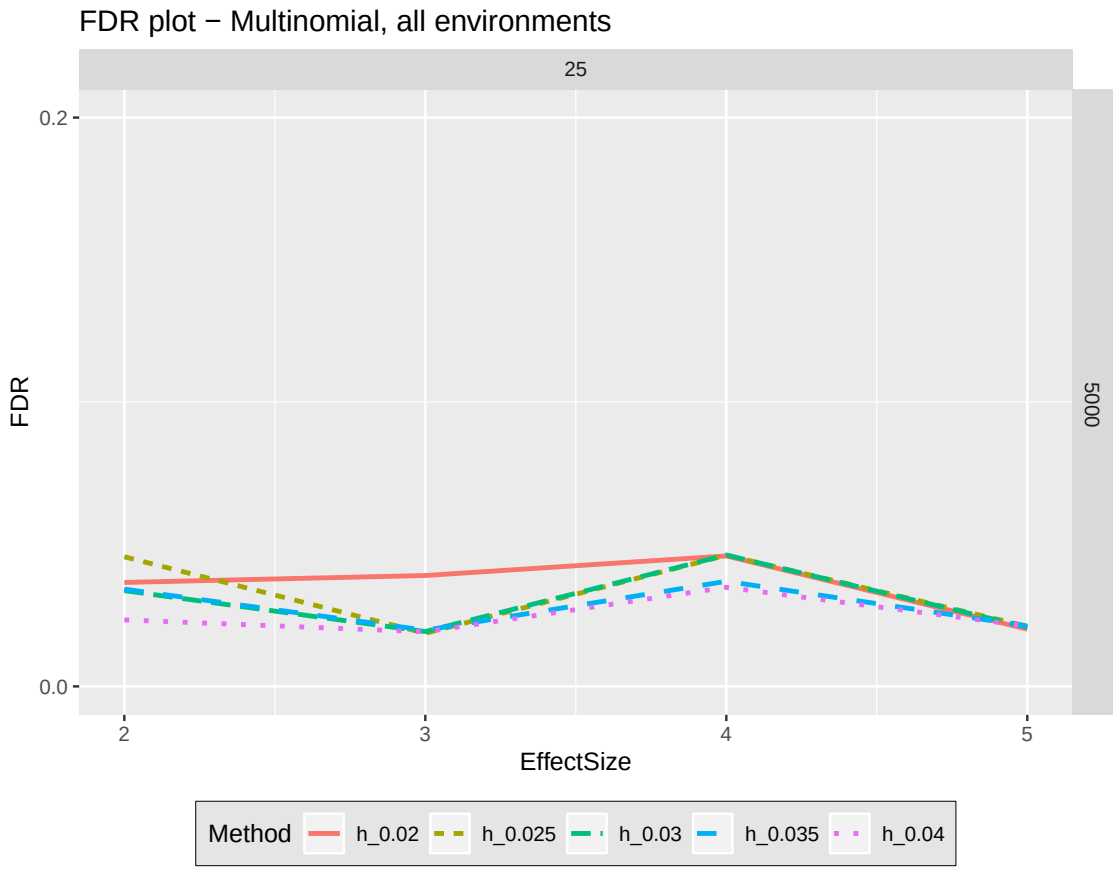


Fig. S11. FDR plot of robustness analysis of the h value in RioNorm2 with various effect sizes using a subset of data in the simulation setting 1. h value varies from 0.02th quantile to 0.04th quantile of dissimilarity distribution. The number of samples per condition is 25 and the median library size is 5000.

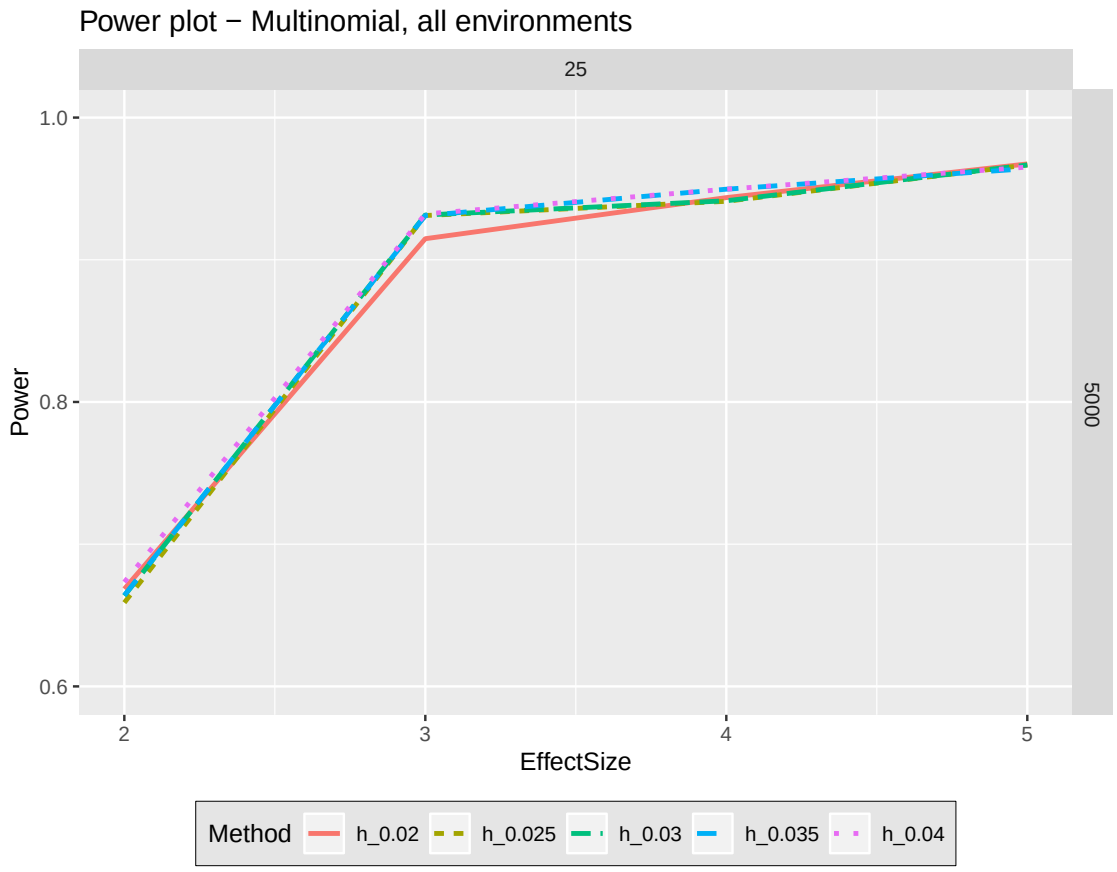


Fig. S12. Power plot of robustness analysis of the h value in RioNorm2 with various effect sizes using a subset of data in the simulation setting 1. h value varies from 0.02th quantile to 0.04th quantile of dissimilarity distribution. The number of samples per condition is 25 and the median library size is 5000.

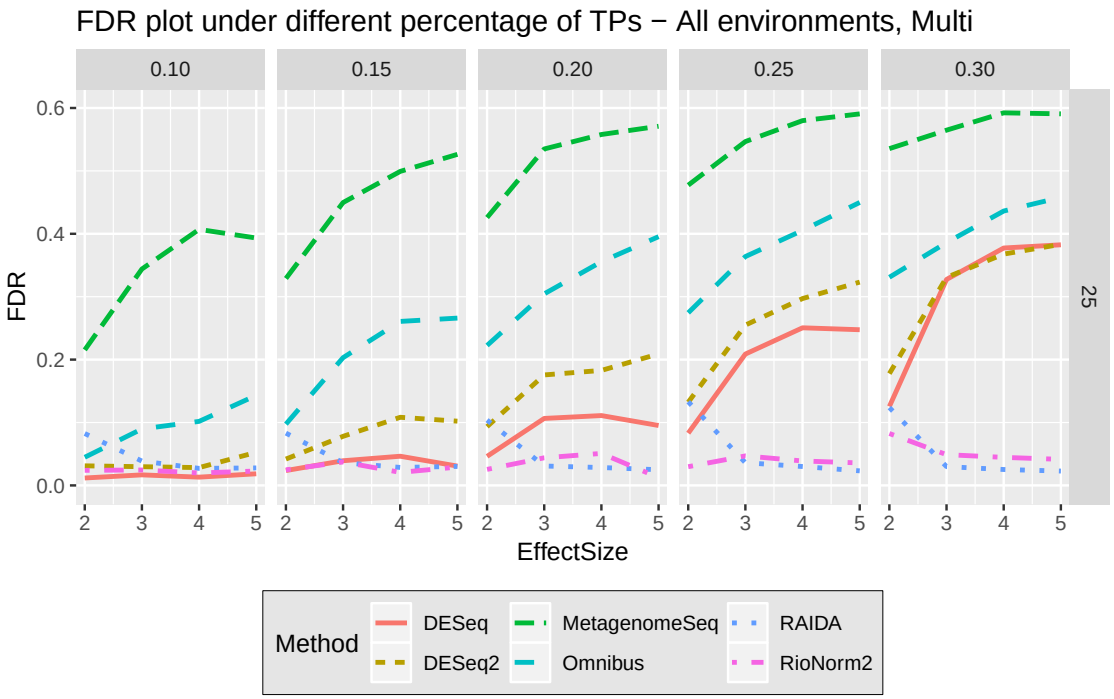


Fig. S13. Comparison of different methods in terms of FDR for various effect sizes using a subset of data in the simulation setting 1. Panel columns represent the proportion of DA-OTUs in each simulated OTU table. The number of samples per condition is 25 and the median library size is 5000.

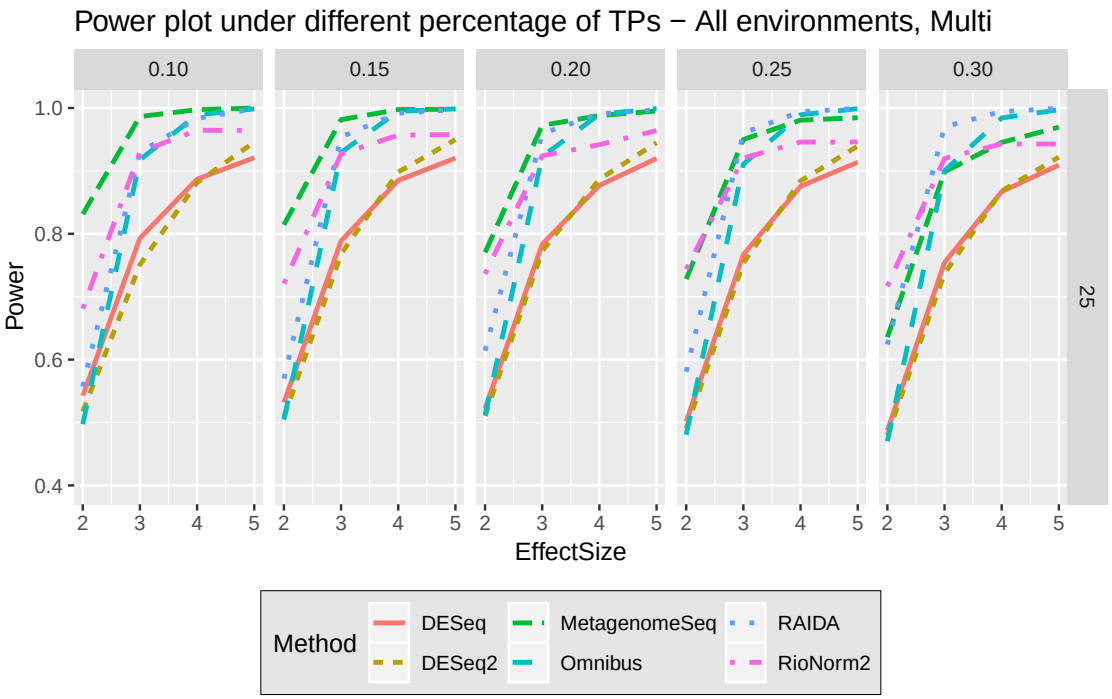


Fig. S14. Comparison of different methods in terms of power for various effect sizes using a subset of data in the simulation setting 1. Panel columns represent the proportion of DA-OTUs in each simulated OTU table. The number of samples per condition is 25 and the median library size is 5000.

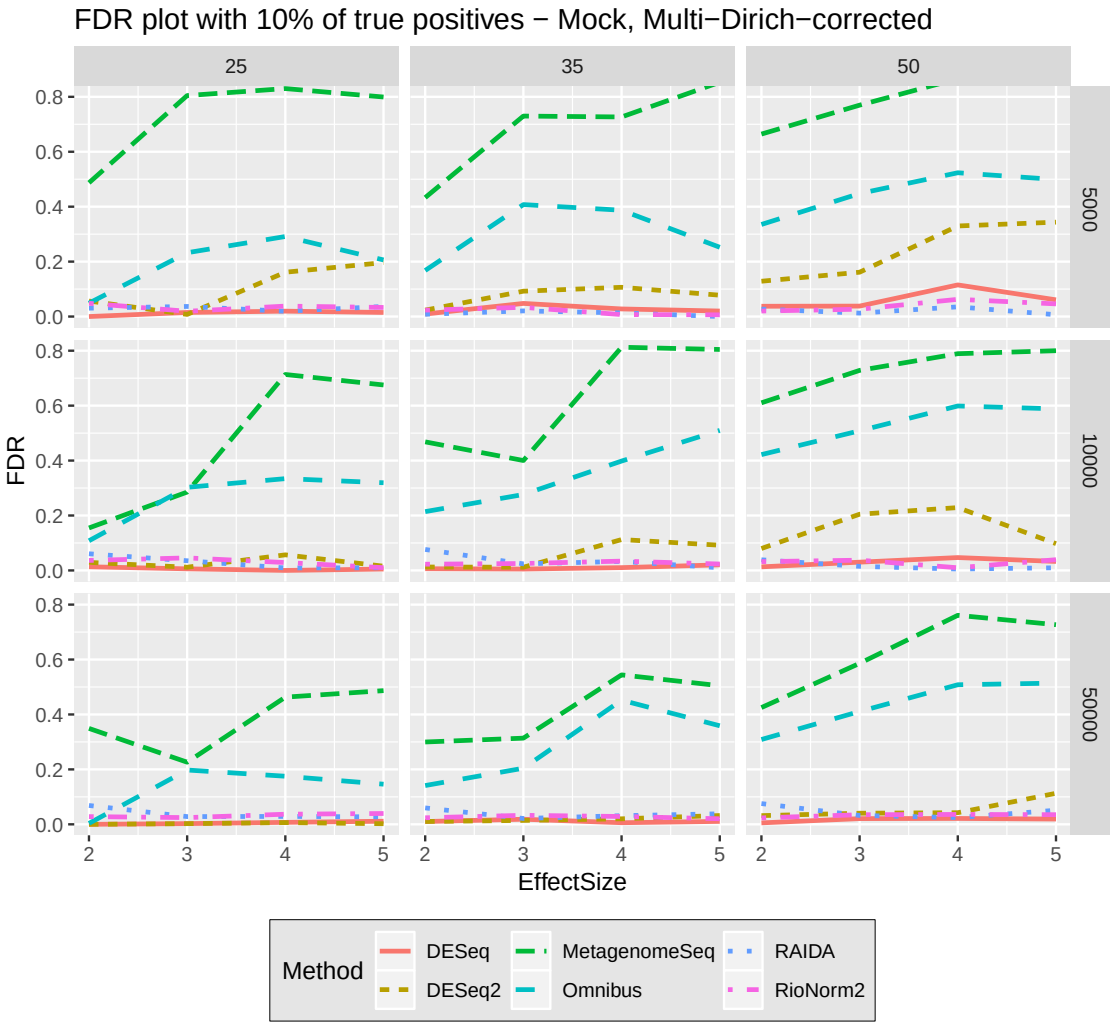


Fig. S15. Comparison of different methods in terms of FDR for various effect sizes in the simulation setting 2. The proportion of DA-OTUs is 10%. Panel rows represent the median library size, and panel columns represent the sample size per condition.

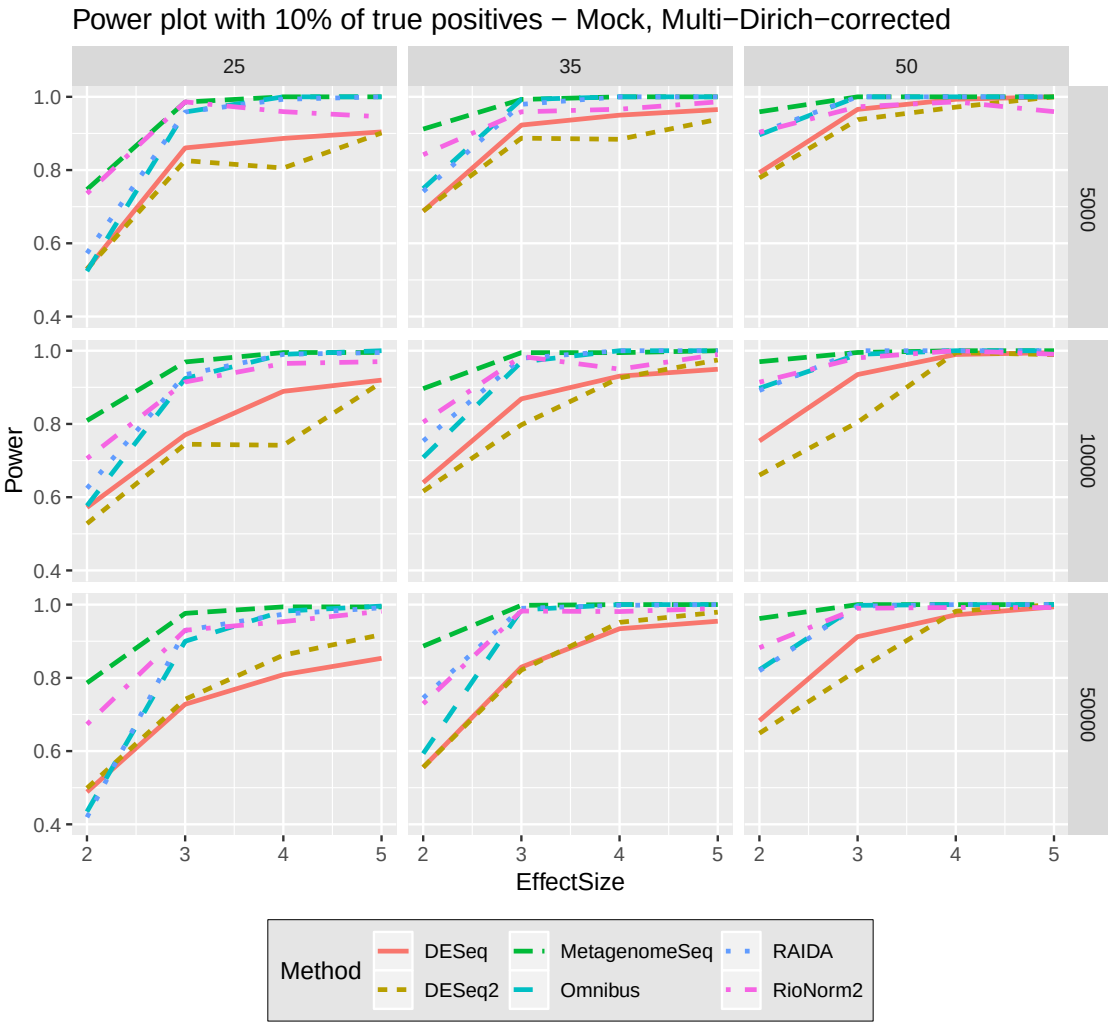


Fig. S16. Comparison of different methods in terms of power for various effect sizes in the simulation setting 2. The proportion of DA-OTUs is 10%. Panel rows represent the median library size, and panel columns represent the sample size per condition.

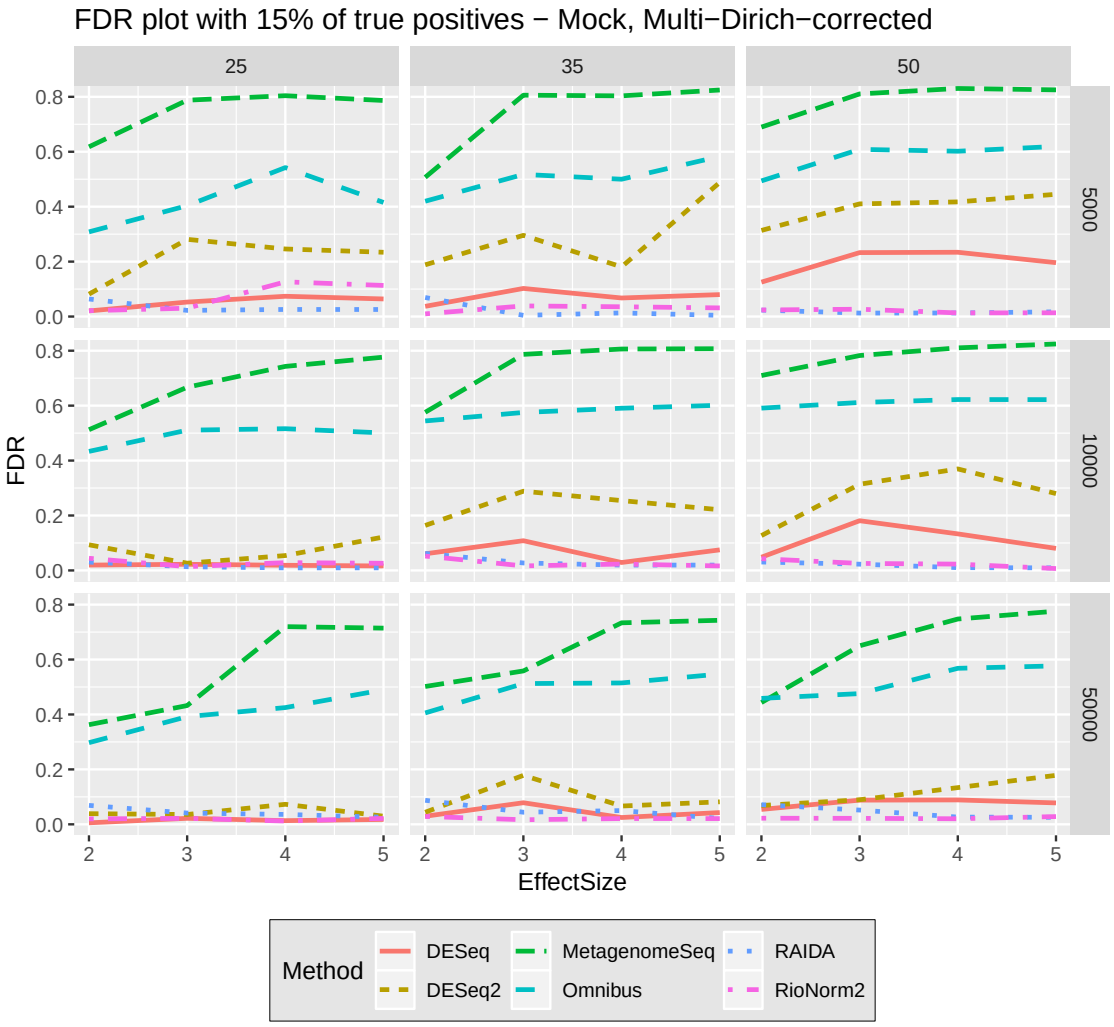


Fig. S17. Comparison of different methods in terms of FDR for various effect sizes in the simulation setting 2. The proportion of DA-OTUs is 15%. Panel rows represent the median library size, and panel columns represent the sample size per condition.

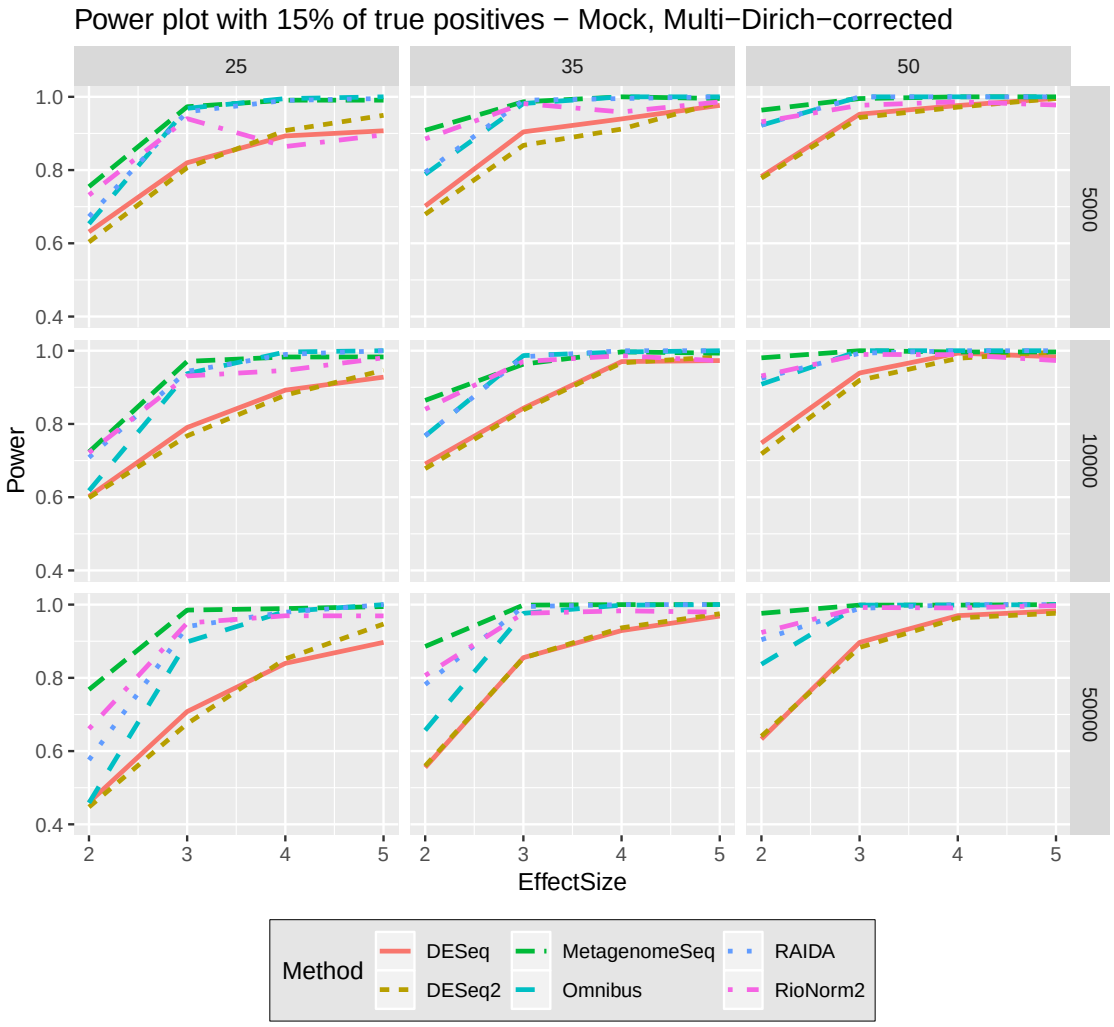


Fig. S18. Comparison of different methods in terms of power for various effect sizes in the simulation setting 2. The proportion of DA-OTUs is 15%. Panel rows represent the median library size, and panel columns represent the sample size per condition.

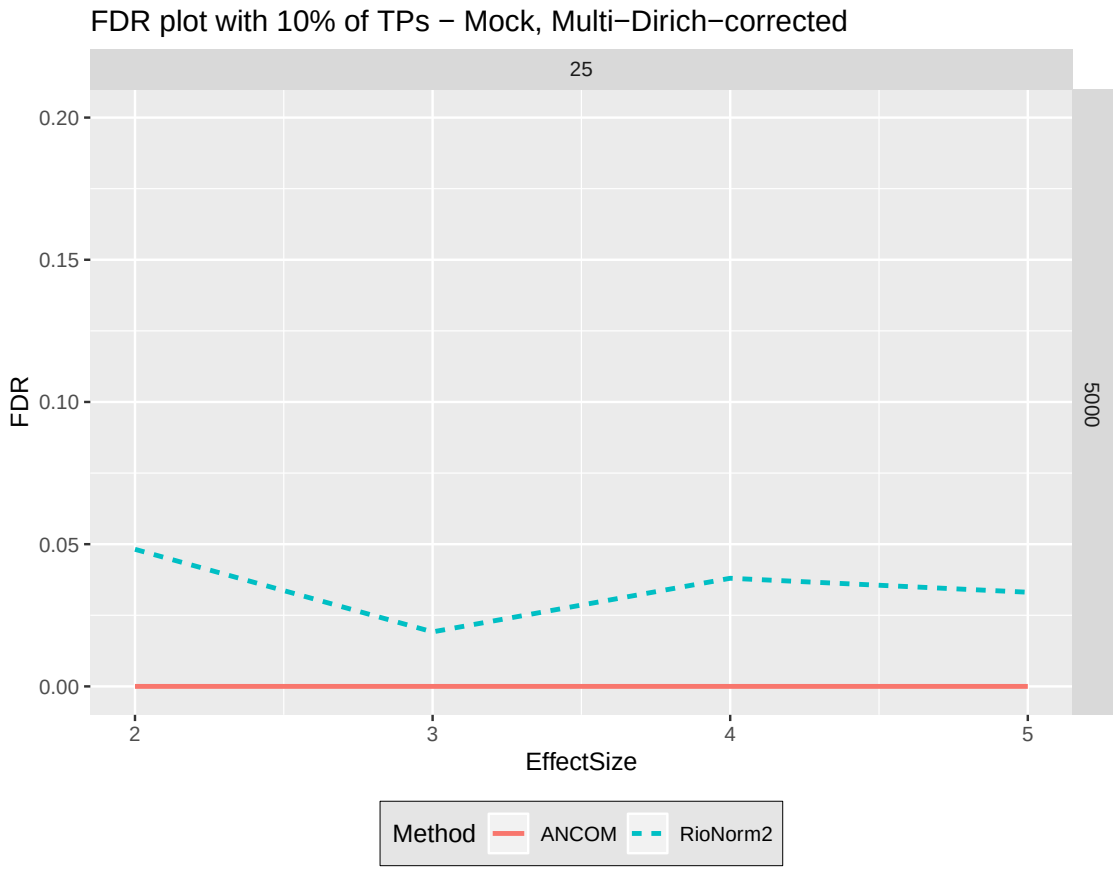


Fig. S19. Comparison of RioNorm2 and ANCOM in terms of FDR for various effect sizes using a subset of data in the simulation setting 2. The percentage of DA-OTUs is 10%. The number of samples per condition is 25 and the median library size is 5000.

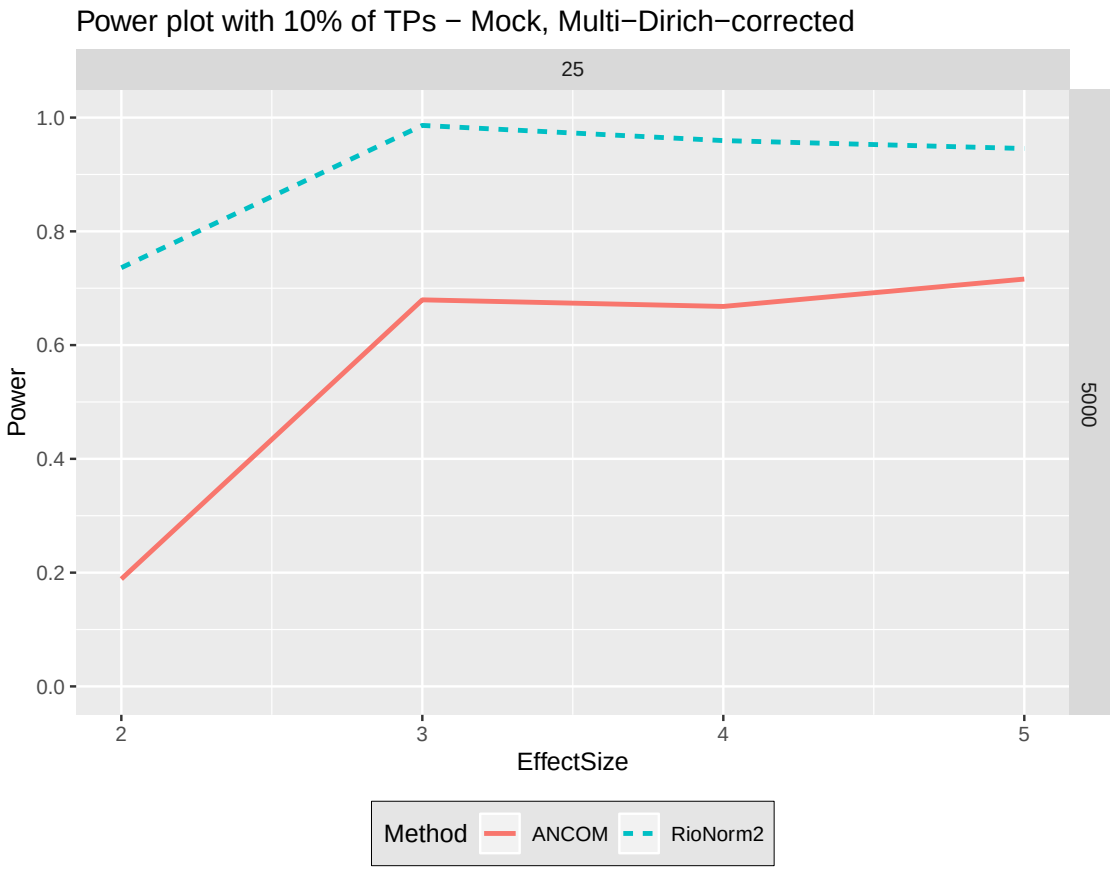


Fig. S20. Comparison of RioNorm2 and ANCOM in terms of power for various effect sizes using a subset of data in the simulation setting 2. The percentage of DA-OTUs is 10%. The number of samples per condition is 25 and the median library size is 5000.

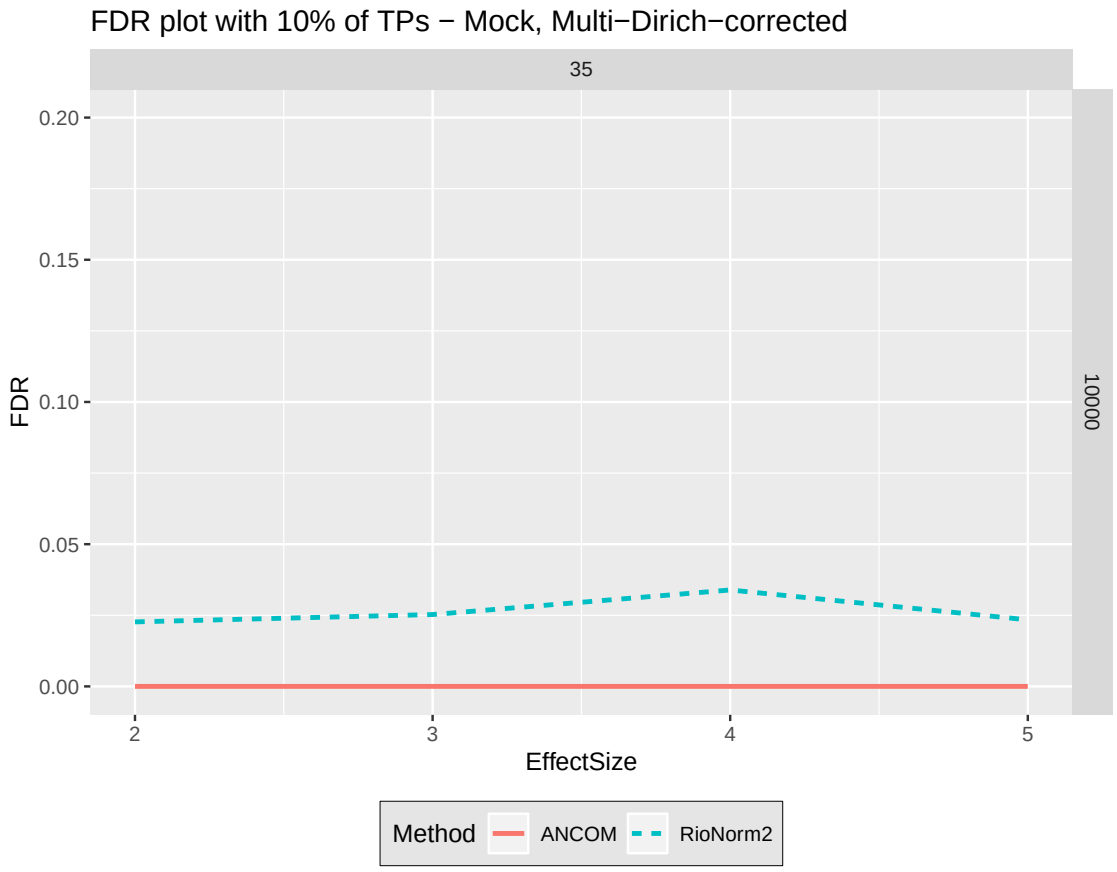


Fig. S21. Comparison of RioNorm2 and ANCOM in terms of FDR for various effect sizes using a subset of data in the simulation setting 2. The percentage of DA-OTUs is 10%. The number of samples per condition is 35 and the median library size is 10000.

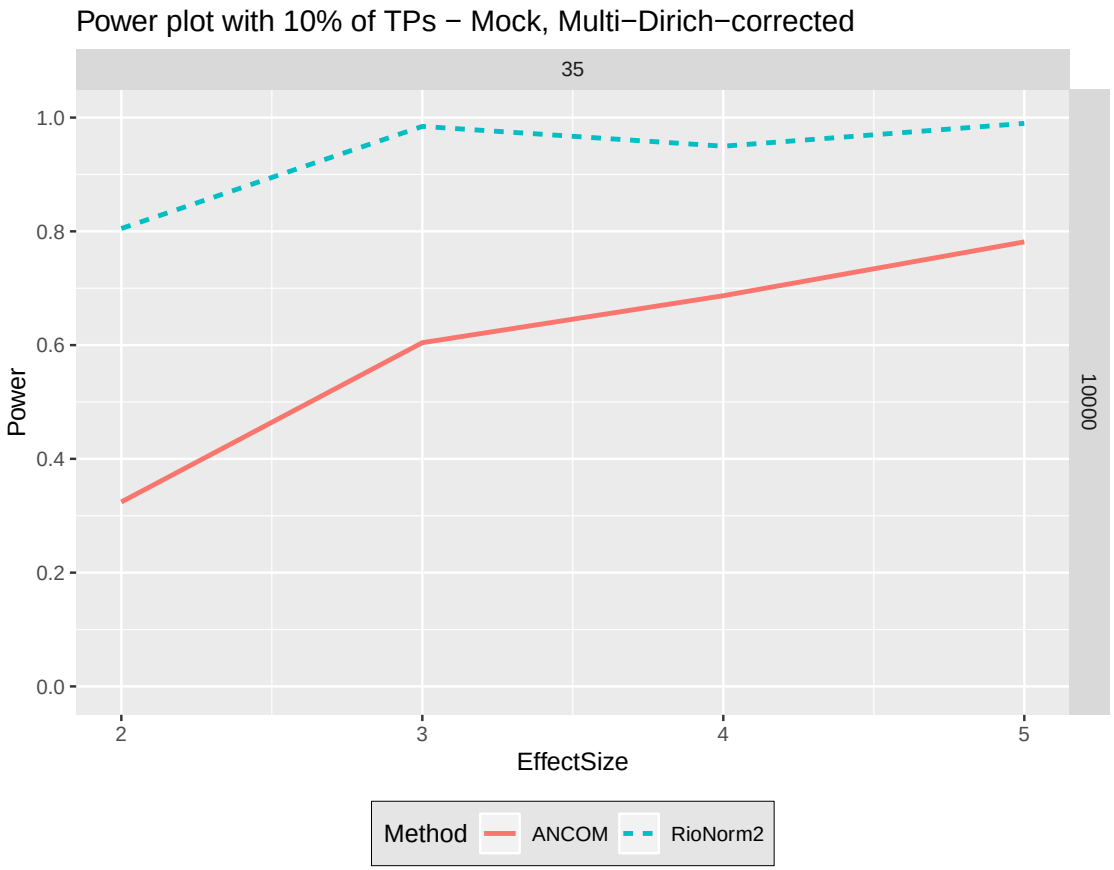


Fig. S22. Comparison of RioNorm2 and ANCOM in terms of power for various effect sizes using a subset of data in the simulation setting 2. The percentage of DA-OTUs is 10%. The number of samples per condition is 35 and the median library size is 10000.

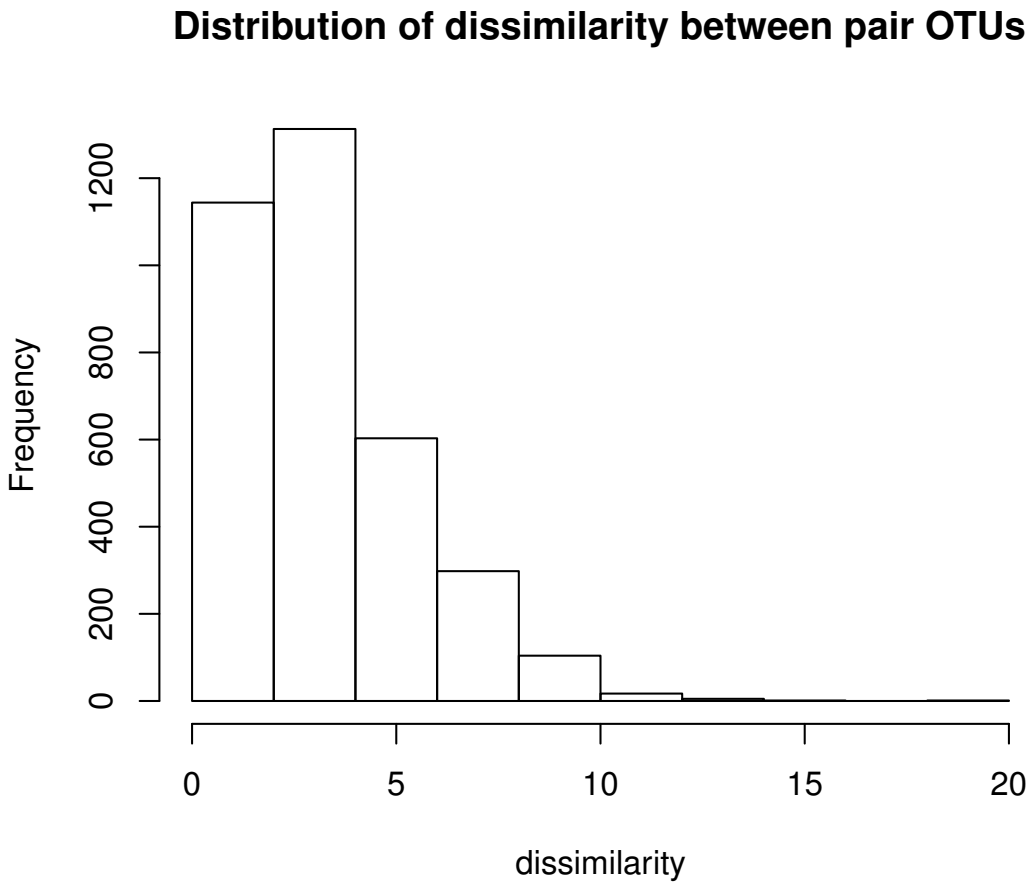


Fig. S23. Histogram of pairwise dissimilarity between OTUs that are observed in at least 80% of samples from metastatic melanoma data.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
189384	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
323231	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
301578	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
182854	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
513445	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
211706	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
184209	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
193591	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
1749079	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
589277	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
171559	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
3426658	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
1944498	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
577170	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__

Fig. S24. Taxonomy of riOTUs in metastatic melanoma data. The first column represents OTU ID. The second to the last column show the taxonomy from Kingdom to Species.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
364903	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
111364	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Christensenellaceae	g__	s__
107044	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Rikenellaceae	g__	s__
New.CleanUp.ReferenceOTU13318	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
New.CleanUp.ReferenceOTU4396	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
181155	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Veillonellaceae	g__Veillonella	s__dispar
187178	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
360268	k__Bacteria	p__Tenericutes	c__RF3	o__ML615J-28	f__	g__	s__
329703	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
275061	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
574689	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Dorea	s__
4395774	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
294856	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Eubacteriaceae	g__Pseudoramibacter_Eubacterium	s__
659361	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Dorea	s__
2696706	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
198151	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
New.CleanUp.ReferenceOTU9382	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__Oscillospira	s__
1111582	k__Bacteria	p__Firmicutes	c__Bacilli	o__Lactobacillales	f__Enterococcaceae	g__Enterococcus	s__
4418496	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Porphyromonadaceae	g__Parabacteroides	s__
New.CleanUp.ReferenceOTU2846	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
363389	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Veillonellaceae	g__Dialister	s__
191251	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Porphyromonadaceae	g__Parabacteroides	s__distasonis
182735	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
840914	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Prevotellaceae	g__Prevotella	s__copri
291090	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Porphyromonadaceae	g__Parabacteroides	s__distasonis
110317	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Peptococcaceae	g__	s__
352902	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__	g__	s__
806751	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
405780	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__Ruminococcus	s__
189936	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__[Mogibacteriaceae]	g__	s__
198830	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
215231	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
851323	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Porphyromonadaceae	g__Parabacteroides	s__
2876801	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__uniformis
336830	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
182052	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
174353	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales			
331571	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
583746	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Veillonellaceae	g__Dialister	s__
591671	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Dorea	s__
555945	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__	g__	s__
New.CleanUp.ReferenceOTU1467	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales			
523542	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae		
183579	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
4333897	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacteriales	f__Enterobacteriaceae	g__	s__

Fig. S25. Taxonomy of detected DA-OTUs in metastatic melanoma data. The first column represents OTU ID. The second to the last column show the taxonomy from Kingdom to Species.

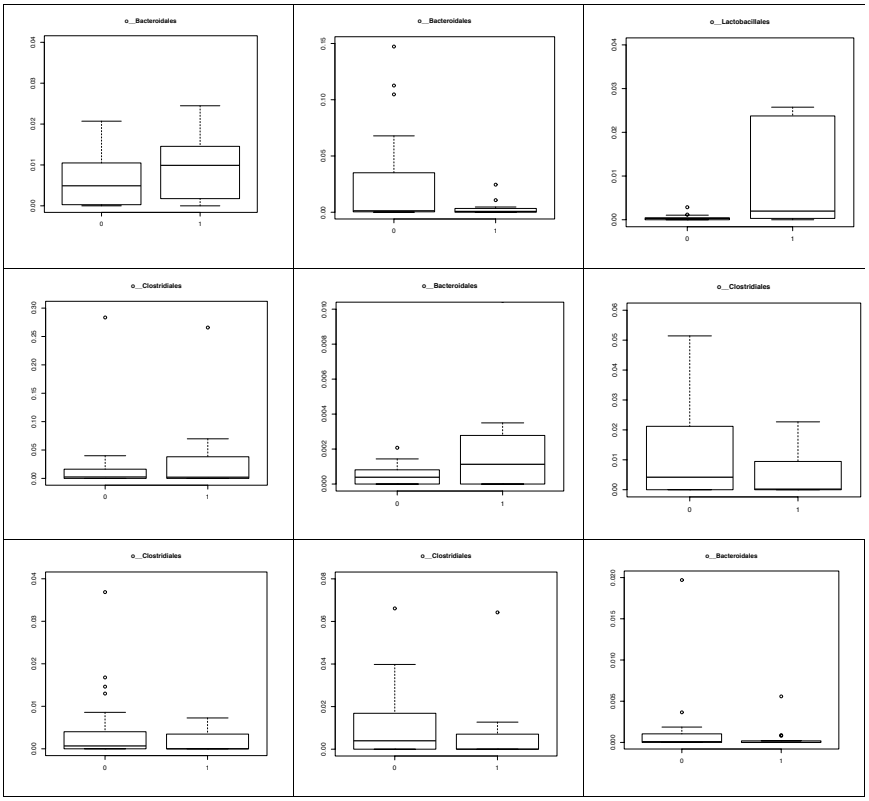


Fig. S26. Box plots of detected DA-OTUs in metastatic melanoma data. The y-axis represents the normalized counts using riOTUs.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
4465907	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Blautia	s__
2582660	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Blautia	s__
180155	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
4424063	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Dorea	s__formicigenerans
4478815	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Coprococcus	s__
4472399	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__

Fig. S27. Taxonomy of riOTUs in IBD data. The first column represents OTU ID. The second to the last column show the taxonomy from Kingdom to Species.

	Kingdom	Phylum	Class	Order	Family	Genus	Species
1504042	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__Oscillospira	s__
4403632	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Coprococcus	s__
4410166	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Prevotellaceae	g__Prevotella	s__copri
4318125	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
3856408	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
4442459	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Porphyromonadaceae	g__Parabacteroides	s__
592616	k__Bacteria	p__Firmicutes	c__Erysipelotrichi	o__Erysipelotrichales	f__Erysipelotrichaceae	g__	s__
814442	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Enterobacteriales	f__Enterobacteriaceae	g__	s__
2949328	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__Bacteroidaceae	g__Bacteroides	s__
4458306	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Veillonellaceae	g__Veillonella	s__dispar
2829179	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Veillonellaceae	g__Acidaminococcus	s__
4462599	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__	g__	s__
3422630	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
178242	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
4393466	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Clostridiaceae	g__SMB53	s__
190639	k__Bacteria	p__Bacteroidetes	c__Bacteroidia	o__Bacteroidales	f__[Odoribacteraceae]	g__Butyricimonas	s__
4436046	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__Dorea	s__
184678	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__Faecalibacterium	s__prausnitzii
4419459	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__	g__	s__
183650	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
305224	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__	s__
4404181	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__	s__
180136	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Ruminococcaceae	g__Oscillospira	s__
4344207	k__Bacteria	p__Firmicutes	c__Bacilli	o__Lactobacillales	f__Streptococcaceae	g__Streptococcus	s__anginosus
4390966	k__Bacteria	p__Firmicutes	c__Clostridia	o__Clostridiales	f__Lachnospiraceae	g__[Ruminococcus]	s__torques

Fig. S28. Taxonomy of detected DA-OTUs in IBD data. The first column represents OTU ID. The second to the last column show the taxonomy from Kingdom to Species.

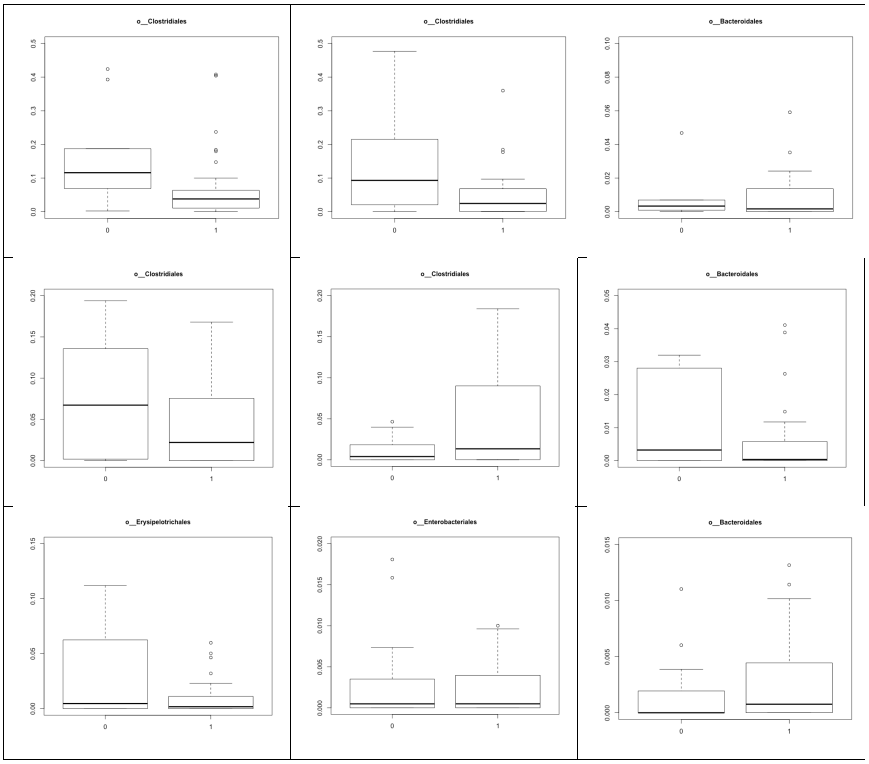


Fig. S29. Box plots of detected DA-OTUs in IBD data. The y-axis represents the normalized counts using riOTUs.