Multi-group Analysis of Compositions of Microbiomes with Covariate Adjustments and Repeated Measures

Shyamal Peddada (✉ Shyamal.peddada@nih.gov)

  National Institute of Environmental Health Sciences   https://orcid.org/0000-0003-1276-7252

Huang Lin

  National Institute of Environmental Health Sciences

Article

Keywords:

Additional Declarations: There is **NO** Competing Interest.

Multi-group Analysis of Compositions of Microbiomes with Covariate Adjustments and Repeated Measures

Huang Lin[1] **and Shyamal Das Peddada**[1,*]

[1]Biostatistics and Computational Biology Branch, NIEHS, NIH, Research Triangle Park, NC, USA
[*]shyamal.peddada@nih.gov

ABSTRACT

Microbiome differential abundance analysis methods for a pair of groups are well established in the literature. However, many microbiome studies involve multiple groups, sometimes even ordered groups, such as stages of a disease, and require different types of comparisons. Standard pairwise comparisons are not only inefficient in terms of power and false discovery rates, but they may not address the scientific question of interest. In this paper, we propose a general framework for performing a wide range of multi-group analyses with covariate adjustments and repeated measures. We demonstrate the effectiveness of our methodology through two real data sets. The first example explores the effects of aridity on the soil microbiome, and the second example investigates the effects of surgical interventions on the microbiome of IBD patients.

Introduction

Increasingly researchers are recognizing the role of the microbiome in human health and diseases[1,2]. Accordingly, it has become a common practice for researchers to investigate differences in microbial compositions and to identify differentially abundant taxa among two or more study or experimental groups. Most existing methods for differential abundance (DA) analysis, such as ANCOM[3], ANCOM-BC[4], LOCOM[5], etc., are designed for discovering differentially abundant taxa between two groups. A variety of strategies are used in the literature for performing multi-group differential abundance analysis. For example, some researchers use the above methods to analyze two groups at a time using a false discovery rate (FDR) threshold within each pairwise comparison and pool the results from all such pairwise comparisons to interpret the data. Such a strategy does not account for the fact that multiple tests and multiple pairwise comparisons are being performed when controlling for false discovery rates. Some researchers perform global beta diversity analyses for multiple groups and follow up with pairwise differential abundance analyses for groups that appear to be different. While all such approaches are intuitive and reasonable, they may not be designed to answer the specific scientific question of interest. Secondly, the statistical properties of such procedures are not well understood. For example, they may not control the overall FDR for multiple tests and multiple pairwise comparisons.

It is well recognized in gene expression studies that when there are more than two groups, the definitions of FDR depend upon the study design and the hypotheses of interest. Standard procedures, such as the BH procedure[6], are designed for testing multiple hypotheses between two groups. When there are multiple groups, the standard concept of FDR, and methods controlling the corresponding error rates, need to be modified according to the study design and type of analyses to be performed[7–9]. A variety of multi-group studies are performed by researchers, such as: (1) *Multiple pairwise comparisons*: A dietitian may be interested in making all pairwise comparisons of the gut microbial compositions among subjects receiving diets D_1, D_2 or D_3. Furthermore, for each pairwise comparison, D_1 vs. D_2, D_1 vs. D_3, and D_2 vs. D_3, the goal is often to identify taxa whose abundance increased (or decreased). (2) *Multiple pairwise comparisons against a specific reference group*: Same as in scenario (1), but the investigator is only interested in a subset of pairwise comparisons. For example, compare each group with a reference group. Suppose D_1 is the standard diet, and the researcher may be interested in identifying taxa whose abundance may have increased (or decreased) for subjects receiving a new diet D_2, compared to D_1, and similarly new diet D_3, compared to D_1. (3) *Pattern analysis over ordered study groups*: In some instances, an investigator may be interested in discovering trends or patterns in abundances of taxa over ordered groups, such as the health of subjects, changes in climate, doses of a drug, and others. For instance, during normal pregnancy, women not only undergo major physiological and hormonal changes, but they also experience significant changes in their gut and vaginal microbiome[10]. These changes are necessary for maternal metabolism, immune response, and hormonal changes to support pregnancy and to provide healthy flora for babies at birth[11,12]. In fact, the alpha diversity of the microbiome dramatically decreases temporally during pregnancy[13]. During the

1

first trimester of a healthy pregnancy, the gut microbiota resembles that of a healthy non-pregnant woman[13]. However, as the pregnancy progresses from the first to the third trimester, there is a reduction in the relative abundance of anti-inflammatory butyrate-producing bacteria and an increase in the pro-inflammatory phylum Proteobacteria, *Bifidobacterium spp.* in the phylum Actinobacteria, and lactic acid-producing bacteria in preparation for energy demands of lactation[12]. Thus, in many scientific investigations, researchers may be interested in studying changes in the microbiome over an ordered set of conditions. The patterns of microbial abundance may not always be monotonic. They may display other shapes, such as an umbrella or an inverted umbrella with the location of the peak or trough unknown a priori. Additionally, depending upon the scientific question of interest, such as microbial changes during pregnancy, repeated measures are taken on the same subject. Although the pattern analyses mentioned here could be accomplished by conducting a sequence of pairwise tests over adjacent ordered groups, such a strategy may have lower power than a test designed for pattern analysis, as will be demonstrated in the analysis of soil aridity data described later in this paper.

The focus of this paper is to develop methodologies for performing multi-group differential abundance analyses for studies such as the above-noted ones. To the best of our knowledge, there does not exist a formal methodology for performing such analyses, with the exception of ANCOM-II[14]. While ANCOM-II considered some of the above testing problems, it does not develop a formal framework for bias correction and is heuristic. The more recent methodology LinDA[15], which uses a model similar to the one developed in ANCOM-II, does not address the above multi-group testing problems. Thus, there is a major gap in the literature for analyzing multi-group microbiome studies, which will be filled by the methodology developed in this paper called ANCOM-BC2.

Before developing ANCOM-BC2, we first make some modifications to ANCOM-BC[4] for testing multiple hypotheses regarding two groups which would result in a better control of FDR. This framework will then be extended to the multi-group testing problem considered in this paper. Although the ANCOM-BC methodology accounted for sample-specific bias, we now also account for taxon-specific bias. This is important because sequencing efficiencies can vary across taxa, leading to a taxon-specific bias when some taxa are preferentially measured over others during the sequencing experimental workflow. For example, gram-positive bacteria have stronger cell walls than gram-negative bacteria, making them harder to extract during the data preparation step. As a result, gram-positive bacteria may be underrepresented in the observed abundances, leading to biased results if not properly accounted for in the analysis[16]. Also, it is well-known in the analysis of other omics data that small signals are often associated with small variances. Consequently, in such cases, the value of the test statistics is inflated which potentially results in a highly significant p-value, even though the effect sizes are trivially small. Inspired by Significance Analysis of Microarrays (SAM)[17] methodology, we regularize the variance to avoid inflated values for the test statistics and hence moderate the p-values to avoid false positives and false discoveries. Lastly, zeros are a common problem for many DA methods, including ANCOM-BC, that work with log-abundance data. Often such methods use pseudo-counts to deal with zero before taking logarithms. However, the choice of pseudo-count can impact the results of rare taxa containing excess zeros, potentially leading to an increase in false discoveries[14, 18, 19] and inflate FDR. To mitigate this issue, we conduct a sensitivity analysis for pseudo-count addition and assign a sensitivity score to each taxon, indicating the likelihood of a false positive result for a particular taxon that is declared significant. A larger sensitivity score indicates a higher risk of a false positive, thus providing a tool for a researcher when interpreting the results.

Using constrained statistical inference-based methods[7] and mixed directional false discovery rate (mdFDR) methods for multiple pairwise comparisons[8, 9], along with the above-noted modifications to ANCOM-BC, we develop ANCOM-BC2 for multi-group microbiome studies in this paper. ANCOM-BC2 allows modeling covariates as well as repeated measures. Details of the method are described in the Methods Section. The performance of ANCOM-BC2 is evaluated using extensive simulation studies under a variety of settings. The results are described in the Results Section. The study designs for the simulation experiments are provided in the Supplementary Methods. ANCOM-BC2 is illustrated using two publicly available data, namely, soil microbiome data and irritable bowel disease data.

Results

Simulations: Continuous and Binary Exposures

In light of practical applications, we conducted simulation studies under two common scenarios, when the exposure variable is continuous or it is binary while adjusting for covariates in the model. To benchmark the performance of ANCOM-BC2, we compared it to its predecessor, ANCOM-BC[4], which was originally developed for binary covariates, as well as two other state-of-the-art DA methods: 1) LinDA[15], which is based on linear regression and employs centered log-ratio transformed abundance data, and 2) LOCOM[5], a logistic regression-based approach that obviates the need for pseudo-counts and utilizes permutation methods to address overdispersion and small sample sizes.

We utilized a subset of data from the Quantitative Microbiome Project (QMP)[20] consisting of 106 samples and 91 operational taxonomic units (OTUs) to generate simulated data. By using real data as a template, we ensured that the data-generating process did not favor our methods, enabling a fair comparison across all methods. We applied the effects of the exposure and

adjusting covariates to the log abundance data based on the QMP data template. The log fold-changes (based on the natural log) of these variables ranged from -2 to 2, which corresponds to a fold-change between 0.14 to 7.4 in the original scale. To evaluate the robustness of differential abundance (DA) methods, we conducted a comprehensive simulation study that incorporated two sources of bias into the synthetic data. To account for sequencing efficiency differences, we applied a feature-specific bias that was sampled from a uniform distribution ($U[0.1, 1]$) and applied to each taxon. Additionally, we included a sample-specific bias that ensured the presence of rare taxa (with more than 50% zeros across samples) to test the robustness of the DA methods to pseudo-count addition. The sample-specific bias was also highly correlated with the exposure of interest to assess the methods' performance against batch effects, which are a concern in large-scale omics studies. To test the robustness of the DA methods regarding the violation of the assumption that most taxa are not differentially abundant in most DA methods, we considered four different cases of the underlying true proportion of DA taxa, ranging from 10% to 90%. For ANCOM-BC2, we used a sensitivity analysis to pseudo-count addition, setting the cutoff value at one, which means any taxon with a sensitivity score greater than one would be automatically declared not significant. To control the false discovery rate due to multiple testing, we utilized the Holm-Bonferroni method[21] instead of the Benjamini-Hochberg (BH) procedure[6], as it does not rely on assumptions about the dependence structure among the underlying p-values and is known to be more robust when dealing with inaccurate p-values[22]. Supplementary Figures 1 and 2 provide additional details on the simulation study design and the results obtained. More information is available in the Supplementary Methods.

Figure 1a displays the simulation results for DA analysis with respect to a continuous exposure. ANCOM-BC2 outperformed all other methods by maintaining a FDR control under the nominal level of 0.05, even with highly dynamic microbial environments, where up to 90% of taxa were differentially abundant. Also, as expected, ANCOM-BC2 demonstrated increased power with larger sample sizes, achieving comparable power to the other methods when the sample size was relatively large (> 50). In contrast, all competing methods had substantially higher FDR than ANCOM-BC2. LOCOM had severely inflated FDR and reduced power with limited sample sizes, reaching up to 50% FDR and as low as 20% power in simulations with 10 samples and 10% of taxa being differentially abundant. ANCOM-BC and LinDA had the highest power among all methods, but both methods exhibited FDR exceeding the nominal level in most simulation scenarios. This increase in FDR correlated with the increase in sample size, suggesting a systematic bias in their test statistics due to the addition of pseudo-counts. Proper sensitivity analysis regarding pseudo-count addition is necessary to prevent random false positives and FDR inflation for rare taxa in both methods.

Simulation results for differential abundance analysis (DA) with respect to a binary exposure are shown in Figure 1b. The results are consistent with those in Figure 1a. The competing methods had substantially inflated FDR as compared to ANCOM-BC2, and their FDRs increased monotonically with the increase in sample size. ANCOM-BC2 outperformed all other methods in terms of maintaining a uniformly small FDR and reasonable power. Note that given the striking differences in FDR, it may not be reasonable to compare powers.

Simulations: Multiple Groups
The simulation settings for multi-group comparisons were consistent with those outlined in the previous section. Similarly, we conducted a sensitivity analysis for ANCOM-BC2 with a cutoff value of one for pseudo-count addition and controlled the FDR due to multiple testing using the Holm-Bonferroni method[21].

Multiple pairwise comparisons against a reference group.   In this simulation study, we evaluated the performance of ANCOM-BC2 in multiple comparisons against a reference group, where the exposure variable consisted of three groups and adjusting covariates were present. We compared ANCOM-BC2 against ANCOM-BC and LinDA. We did not include LOCOM because, currently, it is not designed for multiple groups. Figure 1c illustrates that ANCOM-BC2 effectively controlled the mdFDR[8,9], the combination of both Type I and the directional errors in the FDR framework (see the Methods section for details), below the nominal level of 0.05 and maintained a power greater than 0.8 when the sample size per group exceeded 30. On the other hand, neither ANCOM-BC nor LinDA controlled the FDR, highlighting the advantage of ANCOM-BC2 over existing methods in controlling mdFDR in multi-group comparisons.

Multiple pairwise comparisons.   In this simulation study, we evaluated the performance of ANCOM-BC2 for all possible pairwise comparisons, rather than focusing on comparisons against a specific group. As previously mentioned, the competing methods considered in this paper are not equipped to handle pairwise comparisons in their existing forms, which led to their exclusion from this simulation study. Figure 2a demonstrates that, on average, ANCOM-BC2 effectively controlled the mdFDR well below the nominal level of 0.05 while maintaining substantial power (> 0.8) when the sample size per group was not exceedingly small (n > 30).

Pattern analysis.   Similar to pairwise comparisons, pattern analysis stands out as an important feature of ANCOM-BC2. In this simulation study, we benchmarked the scenario of a monotonically increasing pattern, where the log fold-change between the second group and the reference group, denoted as δ, ranged from 0.5 to 2, and the third group had the log fold-change equal

to $\delta + 1$ with respect to the reference group. The "discovery" in the pattern analysis was defined as the identification of a taxon that was monotonically increasing across the three groups. As shown in Figure 2b, ANCOM-BC2 was successful in controlling the false discovery rate (FDR) in this simulation setting and maintained high power (> 0.8) in most cases. However, under the most extreme scenario where 90% of taxa were truly differentially abundant, ANCOM-BC2 experienced a loss of power. It is important to note that in this simulation study, ANCOM-BC2 pattern analysis controlled the FDR in a very conservative manner, with an FDR equal to 0. Increasing the cutoff value for the sensitivity analysis of pseudo-count addition or using a less stringent FDR controlling method instead of the Holm-Bonferroni method could improve the power and the FDR/power tradeoff (data not shown).

Simulations: Correlated Samples

We also performed benchmarks comparing ANCOM-BC2 to LinDA with respect to correlated microbiome data. Note that ANCOM-BC and LOCOM were not included since neither of them is designed for correlated experimental groups, such as repeated measures. We considered two scenarios in this simulation study. The first scenario contained only a random intercept effect in the data, while the second had both random intercept and random slope effects. Both random effects had variances equal to 1, and a correlation coefficient of 0.5 was set if both were present. The exposure variable had three levels, i.e., there were three experimental groups. The simulation study included a continuous covariate. The remaining simulation settings were the same as indicated in the previous section, and we provide further details in the Supplementary Methods. The results were similar in both scenarios (Fig. 3). ANCOM-BC2 controlled the FDR well below the nominal level of 0.05 in all simulation settings while maintaining adequate power as the sample size increased. Meanwhile, although LinDA had uniformly larger power than ANCOM-BC2, it suffered inflated FDR, particularly when the sample size was large (e.g., the FDR could reach 70% when the sample size per group was 100).

Illustration: Soil Microbiome and Aridity

The threat of climate change not only threatens the dryland ecosystems with increasing temperatures and aridity but there is a concern that some of the microbes native to such conditions may be trans-locating to other environments with the expansion of deserts and desert ecologies. There is growing literature on desert microbiome[23]. The translocation of such bacteria may affect vegetation, plant life, and human health. Most importantly, there are concerns regarding the spread of antimicrobial or antibacterial resistance-causing microbes, which may make infections harder to treat and increase the risk of spreading diseases. Accordingly, there is interest in identifying microbial species that are native to extremely hot conditions, such as hyper-arid soils.

Recently, Neilson et al.[2] investigated the differences in soil microbiomes according to soil aridity in the Atacama Desert in Chile. They classified soil into three ordered categories based on aridity, namely, arid, margin, and hyper-arid, and sequenced data from 63 sample pits from 18 sites in the desert. Since they did not conduct formal differential abundance analyses of those data, we reanalyzed those data using the ANCOM-BC2 methodology. To begin with, we conducted a trend analysis of species richness with respect to the ordered aridity categories (arid to hyper-arid) (Fig. 4a). The trend test using the ORIOGEN[24] software yielded a p-value < 0.0001, suggesting a significant loss of species richness with the increase in aridity. This finding regarding the species richness is consistent with the original publication of Neilson et al.[2], except that we are able to provide a p-value for trend using ORIOGEN.

Next, we performed a pattern analysis to discover patterns in microbial taxa abundances over the ordered soil categories using the arid soil as the reference category. We discovered *Thermobaculum*, *Geodermatophilus*, and *Rubrobacter* to increase in mean abundance with the aridity of the soil (p < 0.05), and among these, *Thermobaculum* had a significant trend after correcting for multiple testing with a multiple testing adjusted p-value < 0.0001 (Fig. 4b). *Thermobaculum spp.* is well-known to exist in temperatures as high as 90 degrees Celsius[25], and it is documented to contain antibacterial or antimicrobial-resistant genes[26,27]. Similarly, the two Actinobacteria genera, *Geodermatophilus* and *Rubrobacter*, are also known to be antibacterial-resistant genera. For a review of these taxa, one may refer to Montero-Calasanz et al.[28] and Li et al.[29], respectively. Thus, using our proposed methodology, we discover genera that have an increase in abundance with respect to aridity and which may have antibacterial resistance properties.

Not surprisingly, nitrogen-cycling microbes are significantly diminished by increasing aridity in desert soils. For example, the presence/absence results (Supplementary Table 1) show that *Nitrobacter*, a common contributor to nitrification, along with putative broadly distributed N2 fixers *Sinorhizobium*, *Rhizobium*, and *Azospirillum*, were not detected in hyper-arid samples. The results of the ANCOM-BC2 pattern analysis also show that increasing aridity correlates with significant reductions in the abundance of taxa typically associated with fertile soils (Fig. 4b). Genera such as *Ca. Nitrososphaera*, which are chemolithoautotrophs and have essential biogeochemical roles as nitrifying organisms[30]; *Paenibacillus*, which contains many species that promote crop growth through nitrogen fixation, phosphate solubilization, production of the phytohormone indole-3-acetic acid (IAA), and release of siderophores that enable iron acquisition[31]; and *Pseudonocardia*, which has been reported to

achieve associative nitrogen fixation and protect their hosts against soil-borne pathogenic infection[32], monotonically decreased with aridity.

ANCOM-BC2 sensitivity analysis revealed that none of the significant taxa exceeded the threshold of 1, indicating that these discoveries are likely true positives (Supplementary Fig. 3).

Illustration: Gut Microbial Composition of IBD Patients after Surgery

Numerous studies have delved into the role of the microbiome in inflammatory bowel disease (IBD); however, only a few have specifically addressed the impact of surgery on the gut microbiome and its subsequent consequences. We employ the ANCOM-BC2 to analyze a longitudinal dataset obtained from Fang et al.[33] to investigate the changes in the gut microbiome following gastrointestinal surgery in IBD patients. The dataset utilized in this study consists of 322 stool samples collected from 125 patients. Of these, 46 patients were diagnosed with ulcerative colitis (UC) and 79 with Crohn's disease (CD). Stool samples were obtained from each subject at approximately 6-month intervals, beginning at the baseline time point. Specifically, 21 patients provided one sample, 38 patients provided two samples, 41 patients provided three samples, 23 patients provided four samples, and 2 patients provided five samples. Of the total patient population, 87 (70.0%) had no history of intestinal surgery, while 22 CD patients had undergone ileocolonic resection, and 13 CD patients and 3 UC patients had undergone different types of colectomies. These surgeries had occurred prior to the collection of the baseline stool sample. For the purposes of this study, we focused on comparing the microbial compositions between patients who had not undergone gastrointestinal surgery, those who had undergone ileocolonic resection, and those who had undergone colectomies. We adjusted the ANCOM-BC2 model for IBD disease type (UC vs. CD) and two potential confounders, namely disease state (inactive vs. active) and antibiotic use (absent vs. present). The results are depicted in Fig. 5.

We performed multiple pairwise comparisons among the three groups controlling the overall mdFDR at 0.05 using ANCOM-BC2. Ileocoloic section is the surgical removal of the diseased section of the ileum, which is the junction area between the small and last intestines. In contrast, colectomy is the surgical removal of most or all of the large intestine. Interestingly, our analysis revealed that almost no microbial species were differentially abundant between the two surgical groups of patients, except for *H. parainfluenzae* and *C. perfringens*, which are more abundant in the colectomy group, and conversely, *C. aldense*, and *B. producta* are more abundant only in the ileocolonic section group.

We saw reduced abundances of several native gut bacterial species in patients who underwent either ileocolonic resection or colectomy compared to patients with no intestinal surgery. Specifically, These species include *Bacteroides spp.* (*ovatus* and *uniformis*), *Butyricicoccus pullicaecorum*, *Dorea* (*formicigenerans* and *longicatena*), *Faecalibacterium prausnitzii*, *Roseburia spp.* (*faecis* and *inulinivorans*), and *Ruminococcus torques*. Most of these bacterial species are involved in the production of short-chain fatty acids (SCFAs), such as butyrate, propionate, and acetate[34–40]. SCFAs play essential roles in maintaining gut health, supporting gut barrier function, possessing anti-inflammatory properties, and providing energy sources for colonocytes. Thus, although the surgical intervention was necessary for these patients, the surgery may have impacted the host immune response and health of these patients. The IBD patients who underwent any of these surgeries may require probiotics or other forms of microbial supplementation.

We also conducted an ANCOM-BC2 sensitivity analysis, revealing that none of the significant taxa exceeded a threshold of 1. This suggests that the observed findings are likely true positives (Supplementary Fig. 4).

Discussion

In this article, we introduced a general framework called ANCOM-BC2 for performing differential abundance analysis when the exposure variable is continuous, binary, or (ordered) categorical. The proposed methodology allows for adjusting for covariates and repeated measures (longitudinal measures) while controlling for FDR or mdFDR when the exposure variable has more than two groups, and the researcher is interested in inferring if the abundance of a taxon increased or decreased for each pairwise comparison. Furthermore, using the theory of constrained statistical inference, ANCOM-BC2 allows researchers to infer patterns in microbial abundance over ordered categories of exposure variables. For example, it allows a researcher to test if a particular microbe increased (or decreased) in abundance over ordered disease categories (very healthy to least healthy). This is a unique feature of ANCOM-BC2. Lastly, ANCOM-BC2 overcomes some of the limitations of ANCOM-BC for controlling the FDR while maintaining high power, especially with the presence of rare taxa with excess zeros.

The power of ANCOM-BC2's pattern analysis was demonstrated in the soil microbiome data analyzed in this paper. When standard pairwise analyses were performed, we only discovered *Paenibacillus* that was significantly differentially abundant across different groups (data not shown). However, using the pattern analysis, we discovered several taxa display increasing or decreasing trends over the ordered soil aridity groups. This is because, unlike pairwise comparisons, pattern analysis uses constrained inference methods, which "borrow" information from neighboring ordered groups. Thus, increasing the effective sample size and the power[7,41,42].

Intrinsically, the ileocolonic section and colectomy are surgically removing different regions of the intestines, and yet based on our analysis of the IBD data, there were no significant differences in the abundance of the majority of the gut bacteria. Furthermore, the two groups of patients have similarly reduced abundances of these bacteria relative to those who did not undergo either of the two surgeries. Based on these findings, it may be reasonable to hypothesize that most species of gut microbiota are spatially uniformly distributed in the ileum and large intestines.

Methods

Notation

The notations described in ANCOM-BC2 methodology are summarized in Table 1.

Notation	Description
i	Sample index, $i = 1, 2, \ldots, n$.
j	Taxon index, $j = 1, 2, \ldots, d$.
k	Index of fixed effects, $k = 1, 2, \ldots, p$.
l	Index of Random effects, $l = 1, 2, \ldots, q$.
x_{ik}	The kth fixed effect of interest for the ith sample.
z_{il}	The lth random effect of interest for the ith sample.
$A_{ij}\ddagger$	Unobserved abundance of jth taxon in a unit volume of ecosystem of ith sample.
$O_{ij}\ddagger$	Observed abundance of jth taxon in a random specimen taken from a unit volume of ecosystem of ith sample.
$E_{ij}\ddagger$	Random error for taxon j in sample i.
$S_i\dagger$	Sample-specific sampling fraction.
$C_j\dagger$	Taxon-specific sequencing efficiency.
$a_{ij}\ddagger$	$\log A_{ij}$.
$o_{ij}\ddagger$	$\log O_{ij}$.
$e_{ij}\ddagger$	Random error for taxon j in sample i in log scale.
$s_i\dagger$	Sample-specific sampling fraction in log scale.
$c_j\dagger$	Taxon-specific sequencing efficiency in log scale.

† Parameter;
‡ Random variable.

Table 1. Summary of notations

ANCOM-BC2 for fixed effects models

Model assumptions

Assumption 1 (Multiplicative model for observed abundances)*:*

$$O_{ij} = S_i C_j A_{ij} E_{ij}.$$

Assumption 1 indicates that, in expectation, the observed abundance of a taxon in a random sample is in constant proportion to the abundance in a unit volume of the ecosystem of the sample. This proportion can be decomposed into two parts: (1) sample-specific sampling fraction, and (2) taxon-specific sequencing efficiency.

According to Assumption 1, for non-zero observed abundance, the above multiplicative model can be transformed into an additive model by log transformation

$$o_{ij} = s_i + c_j + a_{ij} + e_{ij}.$$

Assumption 2 (Linear model for log abundances)*:* For each taxon j, $a_{ij}, i = 1, \ldots, n$ are independently distributed, and

$$a_{ij} = \mathbf{b_j}^T \mathbf{x_i} + e_{ij}^{(a)},$$

where

(1) $\mathbf{x_i} = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ are the covariates of interest (including the intercept) for the ith sample,

(2) $\mathbf{b_j} = (b_{j0}, b_{j1}, b_{j2}, \ldots, b_{jp})^T$ are the corresponding coefficients for $\mathbf{x_i}$.

(3) $e_{ij}^{(a)}, i = 1, \ldots, n$ are independently distributed random errors for log abundances with $E(e_{ij}^{(a)}) = 0, Var(e_{ij}^{(a)}) = \sigma_{jj}^{(a)}$.

Assumption 3 (Independent random error for log observed abundances)*:* Assume there are random errors, $e_{ij}^{(o)}, i = 1, \ldots, n, j = 1, \ldots, d$, for log observed abundances o_{ij}, which are independently distributed with heteroskedasticity:

$$E(e_{ij}^{(o)}) = 0, \quad Var(e_{ij}^{(o)}) = \sigma_{ij}^{(o)}, \quad e_{ij}^{(o)} \perp\!\!\!\perp e_{ij}^{(a)}.$$

Regression framework

Based on the Assumptions 2 and 3, o_{ij} can be modeled as:

$$o_{ij} = s_i + c_j + \mathbf{b_j}^T \mathbf{x_i} + e_{ij}^{(a)} + e_{ij}^{(o)} := s_i + c_j + \mathbf{b_j}^T \mathbf{x_i} + e_{ij}, \tag{1}$$

with

$$E(o_{ij}) = s_i + c_j + \mathbf{b_j}^T \mathbf{x_i}, \quad Var(o_{ij}) = Var(e_{ij}) = \sigma_{jj}^{(a)} + \sigma_{ij}^{(o)} := \sigma_{ij}^{(t)}.$$

where $\sigma_{ij}^{(t)}$ denotes the total variance.

Model (1) can also be written in a vector notation as follows:

$$\mathbf{o_j} = \mathbf{s} + c_j \mathbf{1} + \mathbf{X} \mathbf{b_j} + \mathbf{e_j}, \tag{2}$$

with

$$E(\mathbf{e_j}) = (0, \ldots, 0)^T,$$
$$E(\mathbf{o_j}) = \mathbf{s} + c_j \mathbf{1} + \mathbf{X} \mathbf{b_j},$$
$$Cov(\mathbf{o_j}) = \begin{bmatrix} \sigma_{1j}^{(t)} & 0 & \ldots & 0 \\ 0 & \sigma_{2j}^{(t)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{nj}^{(t)} \end{bmatrix}.$$

where

(1) $\mathbf{1} = (1, 1, \ldots, 1)^T,$

(2) $\mathbf{o_j} = (o_{1j}, o_{2j}, \ldots, o_{nj})^T,$

(3) $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T,$

(4) $\mathbf{b_j} = (b_{j0}, b_{j1}, b_{j2}, \ldots, b_{jp})^T,$

(5) $\mathbf{e_j} = (e_{1j}, e_{2j}, \ldots, e_{nj})^T,$

(6) $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}.$

It is important to note that within each sample i, for taxa $l \neq m$, o_{il} and o_{im} are not necessarily independent due to correlations between a_{il} and a_{im}. Thus vectors $\mathbf{o_l}$ and $\mathbf{o_m}$ are not independent random vectors.

Remove the effect of taxon-specific sequencing efficiency To eliminate the effect of c_j, we first center the log observed abundances across samples, i.e.

$$y_{ij} := o_{ij} - \bar{o}_{\cdot j} = (s_i - \bar{s}) + \mathbf{b_j}^T (\mathbf{x_i} - \bar{\mathbf{x}}) + (e_{ij} - \bar{e}_{\cdot j}),$$
$$:= \theta_i + \beta_j^T \mathbf{x_i} + \varepsilon_{ij}, \tag{3}$$

where

(1) $\beta_{jk} = b_{jk}$ for $k = 1, \ldots, p$, and $\beta_{j0} = \sum_{k=1}^p b_{jk} \bar{x}_{\cdot k},$

(2) $Var(\varepsilon_{ij}) = \frac{(n-1)^2}{n^2} \sigma_{ij}^{(t)} + \frac{1}{n^2} \sum_{i' \neq i} \sigma_{i'j}^{(t)} := \sigma_{ij}.$

Algorithm 1 Iterative MLE

Initialize:

    For $j = 1, \ldots, d$

    $\theta \leftarrow \mathbf{0}$

    $\mathbf{y_j^{(crt)}} \leftarrow \mathbf{y_j} - \theta = \mathbf{y_j}$

    $\beta_j \leftarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y_j^{(crt)}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y_j}$

while not converge **do**

    $\theta \leftarrow \frac{1}{d}\sum_{j=1}^{d}(\mathbf{y_j} - \mathbf{X}\beta_j)$

    $\mathbf{y_j^{(crt)}} \leftarrow \mathbf{y_j} - \theta$

    $\beta_j \leftarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y_j^{(crt)}}$

end while

Estimation of sample-specific bias As can be seen from (3), β_j are not identifiable without determining the nuisance parameter θ_i. We define bias-corrected log abundance $y_{ij}^{(crt)} = y_{ij} - \theta_i$, then the ordinary least squares (OLS) estimators of θ_i and β_j can be obtained by iteratively solving the following equations. For ease of exposition, the algorithm is described in the vector form, i.e. $\mathbf{y_j} = (y_{1j}, y_{2j}, \ldots, y_{nj})^T$, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$, etc.

Upon convergence,

$$\theta^* = \frac{1}{d}\sum_{j=1}^{d}(\mathbf{y_j} - \mathbf{X}\beta_j^*), \quad \mathbf{y_j^{(crt)*}} = \mathbf{y_j} - \theta^*, \quad \beta_j^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y_j^{(crt)*}}. \tag{4}$$

Therefore

$$
\begin{aligned}
\theta^* &= \frac{1}{d}\sum_{j=1}^{d}(\mathbf{y_j} - \mathbf{X}\beta_j^*) = \frac{1}{d}\sum_{j=1}^{d}(\mathbf{y_j} - \mathbf{P}\mathbf{y_j^{(crt)*}}) \\
&= \frac{1}{d}\sum_{j=1}^{d}(\mathbf{y_j} - \mathbf{P}\mathbf{y_j} + \mathbf{P}\theta^*) = \frac{1}{d}\sum_{j=1}^{d}\left[\mathbf{y_j^{(crt)}} + \theta - \mathbf{P}(\mathbf{y_j^{(crt)}} + \theta) + \mathbf{P}\theta^*\right] \\
&= (\mathbf{I} - \mathbf{P})\theta + \mathbf{P}\theta^* + \frac{1}{d}\sum_{j=1}^{d}(\mathbf{I} - \mathbf{P})\mathbf{y_j^{(crt)}} \\
&= (\mathbf{I} - \mathbf{P})\theta + \mathbf{P}\theta^* + \frac{1}{d}\sum_{j=1}^{d}\varepsilon_j,
\end{aligned}
\tag{5}
$$

where

(1) $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix onto $\mathscr{C}(\mathbf{X})$, the column space of $\mathbf{X}$,

(2) $\varepsilon_j = (\mathbf{I} - \mathbf{P})\mathbf{y_j^{(crt)}}$ with $E(\varepsilon_j) = \mathbf{0}$.

Rearranging (5), we see that

$$(\mathbf{I} - \mathbf{P})\theta^* = (\mathbf{I} - \mathbf{P})\theta + \frac{1}{d}\sum_{j=1}^{d}\varepsilon_j.$$

Taking expectations on both sides leads to

$$(\mathbf{I} - \mathbf{P})[E(\theta^*) - \theta] = \mathbf{0}.$$

As $\mathbf{I} - \mathbf{P}$ is an orthogonal projector onto $\mathscr{C}(\mathbf{X})$, the above equation holds as long as either of the following is valid:

(1) $E(\theta^*) - \theta = \mathbf{0}$,

(2) $E(\theta^*) - \theta \in \mathscr{C}(\mathbf{X})$.

It is sufficient to consider (2) because (1) is the trivial case. If (1) were true then from (4) we deduce that there is no sample-specific effect and that $E(\beta_j^*) = \beta_j$. Suppose (2) is true, then there exists a vector $\delta \neq \mathbf{0} \in \mathbb{R}^p$, such that

$$E(\theta^*) = \theta - \mathbf{X}\delta. \tag{6}$$

Then by combining with equation (4), we have

$$E(\beta_j^*) = \delta + \beta_j. \tag{7}$$

We shall denote θ^* and β_j^* obtained from the above iterative algorithm as preliminary estimators of θ and β_j, respectively. Without loss of generality, throughout this paper we assume $\mathbf{X}^T\mathbf{X}$ is a full rank matrix. If it is not a full rank matrix, then we may use any generalized inverse of $\mathbf{X}^T\mathbf{X}$ because $\mathbf{X}\beta_j^*$ in (5) is invariant of the choice of generalized inverse $(\mathbf{X}^T\mathbf{X})^g$ used in $\beta_j^* = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{y_j^{(crt)}}$. Thus the preliminary estimator θ^* provided above is invariant of the choice of generalized inverse used in deriving β_j^*. Furthermore, throughout this paper we are interested in testing hypothesis regarding linearly estimable parameters $\mathbf{A}\beta_j$, i.e. $\mathscr{C}(\mathbf{A}^T) \subset \mathscr{C}(\mathbf{X}^T)$[43]. Consequently, the estimator $\mathbf{A}\beta_j^*$ is invariant of the generalized inverse used in the estimation of β_j^*. Hence throughout this text, for simplicity of exposition, we shall assume $\mathbf{X}^T\mathbf{X}$ is of full rank.

For each taxon $j = 1, \ldots, d$, by (7), β_j^* is a biased estimator if $\delta \neq \mathbf{0}$. Suppose we wish test the following hypothesis

$$H_0 : \mathbf{A}\beta_j = \mathbf{A}\beta_j^0,$$
$$H_1 : \mathbf{A}\beta_j \neq \mathbf{A}\beta_j^0.$$

Under the null hypothesis, $E(\mathbf{A}\beta_j^*) - \mathbf{A}\beta_j^0 = \mathbf{A}\delta \neq \mathbf{0}$ and hence biased. The next step is to estimate this bias δ and accordingly modify the estimator $\mathbf{A}\beta_j^*$ so that the resulting estimator is asymptotically centered at $\mathbf{A}\beta_j^0$ under the null hypothesis and hence the test statistic is asymptotically centered at zero.

First we make the following observations. Since $E(\beta_j^*) = \delta + \beta_j$, we note that as $n \to \infty$, for finite dimension d,

$$\Sigma_j^{-\frac{1}{2}}(\beta_j^* - (\delta + \beta_j)) \to_d N_p(\mathbf{0}, \mathbf{I}), \tag{8}$$

where

$$\Sigma_j = \lim_{n \to \infty}(\mathbf{X}^T\mathbf{X})^{-1}(\sum_{i=1}^{n} \sigma_{ij}^2 \mathbf{x_i x_i}^T)(\mathbf{X}^T\mathbf{X})^{-1}. \tag{9}$$

Since

$$E(\theta^* + \mathbf{X}\beta_j^*) = \theta - \mathbf{X}\delta + \mathbf{X}(\delta + \beta_j) = \theta + \mathbf{X}\beta_j,$$

i.e. $\theta^* + \mathbf{X}\beta_j^*$ is an unbiased estimator of $\theta + \mathbf{X}\beta_j$, hence a possible estimator of Σ_j is given by

$$\hat{\Sigma}_j = (\mathbf{X}^T\mathbf{X})^{-1}(\sum_{i=1}^{n} (y_{ij} - \theta_i^* - \beta_j^{*T}\mathbf{x_i})^2 \mathbf{x_i x_i}^T)(\mathbf{X}^T\mathbf{X})^{-1}. \tag{10}$$

Under some mild regularity conditions[44], with finite d, we have the following consistency result

$$n(\hat{\Sigma}_j - \Sigma_j) \to_p \mathbf{0}, \quad \text{as } n \to \infty. \tag{11}$$

Therefore, replacing Σ_j with $\hat{\Sigma}_j$ in (8) and appealing to Slutsky's theorem, we have

$$\hat{\Sigma}_j^{-\frac{1}{2}}(\beta_j^* - (\delta + \beta_j)) \to_d N_p(\mathbf{0}, \mathbf{I}), \quad \text{as } n \to \infty.$$

By (9) and (11), under some mild regularity conditions[44], for finite d, we obtain

$$\hat{\Sigma}_j \to_p \mathbf{0}, \quad \text{as } n \to \infty.$$

Consequently,

$$\beta_j^* \to_p \delta + \beta_j, \text{ as } n \to \infty. \tag{12}$$

The above observation regarding the convergence of β_j^* plays a critical role in the following. Since the sampling fraction is constant for all taxa within a sample, we pool information across taxa within each sample when estimating δ. We model each taxon abundance using the following Gaussian mixture model. For the jth taxon and the kth covariate, let C_0 denote the set of taxa that are not differentially abundant with respect to x_{ik}, i.e. $C_0 = \{j \in (1,2,\ldots,d) : \beta_{jk} = 0\}$, C_1 denote the set of taxa whose abundance decreases with x_{ik}, i.e. $C_1 = \{j \in (1,2,\ldots,d) : \beta_{jk} < 0\}$, and let C_2 denote the set of taxa whose abundance increases with x_{ik}, i.e. $C_2 = \{j \in (1,2,\ldots,d) : \beta_{jk} > 0\}$. Let π_r denote the probability that a taxon belongs to set C_r, $r = 0,1,2$. For simplicity of estimation of parameters, similar to GEE, we shall assume that $\beta_{jk}^*, j = 1,2,\ldots,d$, are independently distributed. As commonly done in the analyses of various 'omics data, we ignore the underlying correlation structure when estimating δ. Thus, we model the distribution of β_{jk}^* by Gaussian mixture model as follows:

$$f(\beta_{jk}^*) = \pi_0 \phi\left(\frac{\beta_{jk}^* - \delta_k}{v_{j0}}\right) + \pi_1 \phi\left(\frac{\beta_{jk}^* - (\delta_k + l_1)}{v_{j1}}\right) + \pi_2 \phi\left(\frac{\beta_{jk}^* - (\delta_k + l_2)}{v_{j2}}\right), \tag{13}$$

where

(1) ϕ is the standard normal density function,

(2) $\delta_k, \delta_k + l_1$, and $\delta_k + l_2$ are means for $\beta_{jk}^*|C_0, \beta_{jk}^*|C_1$, and $\beta_{jk}^*|C_2$, respectively. $l_1 < 0$, $l_2 > 0$,

(3) v_{j0}, v_{j1}, and v_{j2} are variances of $\beta_{jk}^*|C_0, \beta_{jk}^*|C_1$, and $\beta_{jk}^*|C_2$, respectively.

Note that instead of fitting a multivariate Gaussian mixture model for all covariates together, we choose to fit a univariate Gaussian mixture model repeatedly for every single covariate. This repetition is simply because the sets of taxa $\{C_0, C_1, C_2\}$ are not necessarily the same for different covariates. Also, note that for a categorical covariate of s levels, it contains s coefficients, e.g. $\beta_{j1}, \ldots, \beta_{js}$, and we shall fit the Gaussian mixture model for these s coefficients separately.

For computational simplicity, we assume that $v_{j1} > v_{j0}, v_{j2} > v_{j0}$. Thus, Without loss of generality for $\kappa_1, \kappa_2 > 0$, let $v_{j1} = v_{j0} + \kappa_1$ and $v_{j2} = v_{j0} + \kappa_2$. While this assumption is not a requirement for our method, it is reasonable to assume that variability among differentially abundant taxa is larger than that among the null taxa. By making this assumption, we simplify the computation.

Assuming samples are independent, we begin by first estimating $v_{j0}^2 = Var(\beta_{jk}^*)$. Note that v_{j0}^2 is the function of heteroscedastic variances, a consistent estimator of v_{j0}^2, which we refer to as $\hat{v}_{j0}^2$, is the kth diagonal element of $\hat{\Sigma}_j$ stated in (10). In all future calculations, we plug in $\hat{v}_{j0}^2$ for v_{j0}^2. This is similar in spirit to many statistical procedures involving nuisance parameters. The following lemma is useful in the sequel.

Lemma 1 (Introducing the latent variable in calculating log-likelihood[45])*:*

$$\log f(x|\theta) = E_{f(z|x,\theta)}[\log f(z|\theta) + \log f(x|z,\theta)].$$

Let $\Theta = (\delta_k, \pi_1, \pi_2, \pi_3, l_1, l_2, \kappa_1, \kappa_2)^T$ denote the set of unknown parameters, then for each taxon the log-likelihood can be reformulated using Lemma 1, as follows:

$$\Theta \leftarrow \arg\max_{\Theta} \sum_{j=1}^{d} \sum_{r=0}^{2} p_{r,j}[\log Pr(j \in C_r) + \log f(\beta_{jk}|j \in C_r)]. \tag{14}$$

Then the E-M algorithm is described as follows:

- E-step: Compute conditional probabilities of latent variables. Define $p_{r,j} = Pr(j \in C_r|\beta_{jk}, \Theta) = \frac{\pi_r \phi\left(\frac{\beta_{jk} - (\delta_k + l_r)}{v_{jr}}\right)}{\sum_r \pi_r \phi\left(\frac{\beta_{jk} - (\delta_k + l_r)}{v_{jr}}\right)}, r = 0,1,2; j = 1,\ldots,d$, which are conditional probabilities representing the probability that an observed value follows each distribution. Note that $l_0 = 0$.

- M-step: Maximize the likelihood function with respect to the parameters, given the conditional probabilities.

We shall denote the resulting estimator of δ_k upon convergence of the algorithm by $\hat{\delta}_k^{EM}$.

As stated in Lin and Peddada[4], compared to $\hat{v}_{j0}^2$, the variance and covariance contributed by $\hat{\delta}_k^{EM}$ is negligible when the number of non-differentially abundant taxa is large, such as when analyzing the microbiome data at the OTU/ASV or species level of the phylogenetic tree.

The above procedure is applied to every $\beta_{jk}, k = 1, \ldots, p$, eventually, we obtain the estimator of δ as

$$\hat{\delta}^{EM} = (\hat{\delta}_1^{EM}, \hat{\delta}_2^{EM}, \ldots, \hat{\delta}_p^{EM})^T. \tag{15}$$

Therefore, the final estimator of β_j is defined as

$$\hat{\beta}_j = \beta_j^* - \hat{\delta}^{EM}, \tag{16}$$

with

$$\hat{\beta}_j \to_p \beta_j, \quad \text{as } n \to \infty, \tag{17}$$

given that $\hat{\delta}^{EM}$ is a good approximation of δ.

The estimation procedure is summarized in Algorithm 2.

Algorithm 2 E-M algorithm

1: **Input:**
      $\beta_j^*, \Sigma_j, j = 1, \ldots, d$
2: **procedure** E-M(β_j^*, Σ_j)
3:      **return** $\hat{\delta}_k^{EM}, k = 1, \ldots, p$
4: **end procedure**
5: **for** $k = 1, \ldots, p$ **do**
6:      $\hat{\beta}_{jk} \leftarrow \beta_{jk}^* - \hat{\delta}_k^{EM}$
7: **end for**

For taxon j, we now describe our methodology for testing the following hypotheses

$$H_0 : \mathbf{A}\beta_j = \mathbf{A}\beta_j^0,$$
$$H_1 : \mathbf{A}\beta_j \neq \mathbf{A}\beta_j^0.$$

From Slutsky's theorem, as $n \to \infty$, the following test statistic is approximately central chi-square distributed under the null hypothesis

$$\begin{aligned}
W_j &= (\mathbf{A}\hat{\beta}_j - \mathbf{A}\beta_j^0)^T (\mathbf{A}\hat{\Sigma}_j \mathbf{A}^T)^{-1} (\mathbf{A}\hat{\beta}_j - \mathbf{A}\beta_j^0) \\
&= (\mathbf{A}\beta_j^* - \mathbf{A}\hat{\delta}^{EM} - \mathbf{A}\beta_j^0)^T (\mathbf{A}\hat{\Sigma}_j \mathbf{A}^T)^{-1} (\mathbf{A}\beta_j^* - \mathbf{A}\hat{\delta}^{EM} - \mathbf{A}\beta_j^0) \\
&\to_d \chi_q^2,
\end{aligned}$$

where $q = rank(A)$.

To control the FDR due to multiple testing, we recommend applying Holm-Bonferroni method[21] instead of Benjamini-Hochberg (BH) procedure[6] because the Holm-Bonferroni method does not require any assumptions regarding the dependence structure in the underlying p-values, and is also known to be a better method to control FDR when p-values are not accurate[22].

Sampling-specific biases estimation   After obtaining $\hat{\delta}^{EM}$, the estimator of sample-specific biases θ is defined as follows:

$$\hat{\theta} = \frac{1}{d} \sum_{j=1}^d (\mathbf{y_j} - \mathbf{X}\hat{\beta}_j). \tag{18}$$

Let $\Sigma^{(i)} = [\sigma_{lm}^{(i)}]_{l,m=1,\ldots,d}$ denote the $d \times d$ covariance matrix of $\varepsilon^{(i)} = (\varepsilon_1^{(i)}, \varepsilon_2^{(i)}, \ldots, \varepsilon_d^{(i)})^T$, where $\sigma_{lm}^{(i)}$ is the $(l,m)^{th}$ element of $\Sigma^{(i)}$ and $\sigma_{jj}^{(i)}$ is the jth diagonal element of $\Sigma^{(i)}$. Furthermore, suppose

Assumption 4 (Sparse correlations among taxa)*:*

$$\sigma_{jj}^{(i)} < K < \infty,$$
$$\frac{\sum_{l \neq m}^d \sigma_{lm}^{(i)}}{d^2} = o(1).$$

From Assumption [4], we have

$$0 \leq \mathbf{1}^T \Sigma^{(i)} \mathbf{1} = \sum_{l=1}^{d} \sum_{m=1}^{d} \sigma_{lm}^{(i)} = \sum_{j=1}^{d} \sigma_{jj}^{(i)} + \sum_{l \neq m}^{d} \sigma_{lm}^{(i)} \leq dK + \sum_{l \neq m}^{d} \sigma_{lm}^{(i)}.$$

Hence

$$0 \leq \frac{\mathbf{1}^T \Sigma^{(i)} \mathbf{1}}{d^2} \leq \frac{K}{d} + \frac{\sum_{l \neq m}^{d} \sigma_{lm}^{(i)}}{d^2} = o(1).$$

Thus, for each taxon $j = 1, 2, \ldots, d$, we have

$$\frac{1}{d} \sum_{j=1}^{d} (\mathbf{y_j} - (\theta + \mathbf{X}\beta_j)) \to_p 0, \quad \text{as } d \to \infty. \tag{19}$$

Therefore, according to (17) and (19), as both $n, d \to \infty$,

$$\hat{\theta} \to \theta. \tag{20}$$

Remark 1. Regularization of variance: To avoid the spurious detection of significance due to extremely small standard errors, particularly for rare taxa, we incorporated a small positive constant in the denominator of the ANCOM-BC2 test statistic for each taxon. This approach was inspired by the Significance Analysis of Microarrays (SAM) methodology[17]. Specifically, the regularization factor was set as the 5th percentile of the distribution of standard errors for each fixed effect, unless otherwise specified.

Remark 2. Sensitivity analysis for the pseudo-count addition: To mitigate the risk of inflated false positive rates resulting from the choice of pseudo-count in ANCOM-BC2, we conducted a sensitivity analysis to assess the impact of varying pseudo-count values on differential abundance results. This is particularly important, as several studies have shown that the choice of pseudo-count can significantly influence the results of differential abundance analysis methods[18, 19]. Specifically, we surveyed a range of pseudo-count values from 0 to 1 with 0.01 increments, and a sensitivity score was assigned to each taxon to evaluate its sensitivity to the choice of pseudo-count. The sensitivity score was defined as the standard deviation of the resulting negative log p-values, with a larger score indicating greater sensitivity to the pseudo-count choice. Consequently, a significant result for a taxon with a high sensitivity score is likely to be a false positive.

Multi-group comparison

In some applications, for a given taxon, researchers are interested in drawing inferences regarding differential abundance among different pairs of experimental groups. We refer to this kind of problem as a multi-group comparison problem, and extra caution needs to be exercised to correct p-values due to multiple comparisons. For simplicity, we drop the subscript j (taxon index) in the following discussions.

Global test   For a given taxon and a total of $g + 1$ experimental groups (including the reference group), researchers may want to test whether there exists at least one group that is significantly different from others. For ease of exposition, we split the covariates $\mathbf{X}$ into two parts, where $\mathbf{X_1}$ stands for the group assignment, and $\mathbf{X_2}$ denotes the remaining covariates. Note that the difference of group effects against the reference group is estimable, while the individual group effect is not. For simplicity, in the discussions of multi-group comparisons among group 0 to group g, we assume group 0 is the reference group. We use $\beta_k, k = 1, \ldots, g$ to denote the group effect, but notice that it actually estimates $\beta_k - \beta_0$. We rewrite the model stated in (3) as

$$\mathbf{y} = \theta + \mathbf{X_1}\beta + \mathbf{X_2}\gamma + \varepsilon, \tag{21}$$

where

(1) θ is the sample-specific bias,

(2) β is the vector of group effects (as compared to group 0) of the order $g \times 1$,

(3) $\mathbf{X_1}$ is the design matrix of the order $n \times g$ consisting of 0s and 1s,

(4) $\mathbf{X_2}$ is the known matrix of other covariates (including the intercept) of the order $n \times (p - g + 1)$ with the corresponding regression parameter vector γ of the order $(p - g + 1) \times 1$.

The global test intends to test

$$H_0 : \cap_{k \in \{1,\ldots,g\}} \beta_k = 0,$$
$$H_1 : \cup_{k \in \{1,\ldots,g\}} \beta_k \neq 0,$$

which can be reformulated as

$$H_0 : \mathbf{A}\beta = \mathbf{0},$$
$$H_1 : \mathbf{A}\beta \neq \mathbf{0},$$

where

$$\mathbf{A} = \mathbf{I}_g = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix}$$

with the test statistic

$$W = (\mathbf{A}\hat{\beta})^T (\mathbf{A}\hat{\Sigma}^{(g)}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\beta}) \to_d \chi_g^2, \quad \text{as } n \to \infty,$$

where $\hat{\Sigma}^{(g)}$ is the corresponding sub-matrix of $\hat{\Sigma}$ defined in equation (10).

Similarly, to control the FDR due to multiple testing, we recommend applying Holm-Bonferroni method[21] instead of Benjamini-Hochberg (BH) procedure[6] due to the underlying complex dependence structure between taxa.

Example 0.1. Suppose there are 3 groups, namely, groups 0 (reference), 1, and 2, and no other covariates. For each sample $i, i = 1, \ldots, n$, we have:

$$y_i = \theta_i + \mu + \beta_1 I\{group = 1\} + \beta_2 I\{group = 2\} + \varepsilon_i.$$

To test whether there is at least one group among 0, 1, and 2, that is significantly different from others, we would like to test:

$$H_0 : \beta_1 = \beta_2 = 0,$$
$$H_1 : \beta_1 \neq 0 \cup \beta_2 \neq 0,$$

which is the same as testing:

$$H_0 : \mathbf{A}\beta = 0,$$
$$H_1 : \mathbf{A}\beta \neq \mathbf{0},$$

where $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\beta = (\beta_1, \beta_2)^T$.

Multiple pairwise comparisons If we are interested in knowing whether the abundance increased or decreased between various pairs of groups, then it amounts to testing the following hypotheses:

$$H_{0,k,k'} : \beta_k = \beta_{k'}$$
$$H_{1,k,k'} : \{\beta_k < \beta_{k'}\} \cup \{\beta_k > \beta_{k'}\},$$

where $k \neq k' \in \{1, \ldots, g\}$. Denote the test statistic for a given pairwise comparison as

$$W_{kk'} = \frac{\hat{\beta}_k - \hat{\beta}_{k'}}{\sqrt{\widehat{Var}(\hat{\beta}_k) + \widehat{Var}(\hat{\beta}_{k'})}} \to_d N(0,1), \quad \text{as } n \to \infty,$$

where $\widehat{Var}(\hat{\beta}_k), \widehat{Var}(\hat{\beta}_{k'})$ are the kth and k'^{th} diagonal elements of $\hat{\Sigma}^{(g)}$, respectively. Thus, the raw p-value for comparing group k and group k' is defined as:

$$p_{kk'} = 2[1 - \phi(|W_{kk'}|)].$$

For comparing with the reference group (group 0), the hypotheses become:

$$H_{0,k} : \beta_k = 0$$
$$H_{1,k} : \{\beta_k < 0\} \cup \{\beta_k > 0\}.$$

We also replace $\hat{\beta}_{k'}$ and $\widehat{Var}(\hat{\beta}_{k'})$ with 0s in the test statistic.

Note that the null and alternative hypotheses for the global test are denoted as H_0 and H_1, a Type I error might occur due to wrongly rejecting H_0 or correctly rejecting H_0 but wrongly rejecting $H_{0,k,k'}$. A directional error might occur due to correctly rejecting H_0 but wrong assignment of the direction between β_k and $\beta_{k'}$ while correctly rejecting $H_{0,k,k'}$. In this case, we need to control the error rate combining both Type I and the directional errors in the FDR framework, which is referred to as mixed directional FDR (mdFDR)[8, 9].

Definition 1 (mdFDR)*:* Let $V(j)$ denote the indicator function of at least one Type I error or directional error committed, i.e.

$$V(j) = \begin{cases} 1 & \text{if Type I or directional error occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, mdFDR is defined as the expected proportion of Type I and directional errors among all discovered taxa.

$$mdFDR = E\Big(\frac{\sum_{j=1}^{d} V(j)}{\max(R, 1)}\Big),$$

where R denote the number of taxa discovered.

To control the mdFDR for all pairwise tests, we adopt the general mdFDR controlling procedure[9], and do the following:

(1) Apply global test method stated above to obtain the p-value for each taxon. We denote these p-values as screening p-values. Apply BH procedure to identify taxa that are differentially abundant in at least one pairwise comparison. Let R denote the number of taxa discovered.

(2) For each taxon discovered in step 1, apply any mixed directional family wise error (mdFWER) controlling procedure, such as Holm-Bonferroni (default), Hochberg, etc., to the pairwise p-values ($p_{kk'}$) at level $R\alpha/d$.

(3) For a given taxon discovered in step 1, if a pairwise hypothesis is rejected in step 2, then we declare $\beta_k < \beta_{k'}$ or $\beta_k > \beta_{k'}$ according to $W_{kk'} < 0$ or > 0.

It has been proved that under the assumption of independence of p-values obtained from the global test, the mdFDR of the above procedure is strongly controlled at level α[9].

Example 0.2. Suppose there are 3 groups, namely, groups 0 (reference), 1, and 2, and no other covariates. For each sample $i, i = 1, \ldots, n$, we have:

$$y_i = \theta_i + \mu + \beta_1 I\{group = 1\} + \beta_2 I\{group = 2\} + \varepsilon_i.$$

To test whether the taxon is differentially abundant between group 1 and 0 (reference), we would like to test:

$$H_0 : \beta_1 = 0,$$
$$H_1 : \{\beta_1 < 0\} \cup \{\beta_1 > 0\},$$

with the test statistic:

$$W_{10} = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}.$$

Additionally, if we want to test whether the taxon is differentially abundant between group 1 and 2:

$$H_0 : \beta_1 = \beta_2,$$
$$H_1 : \{\beta_1 < \beta_2\} \cup \{\beta_1 > \beta_2\}.$$

The test statistic is:

$$W_{12} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_2)}}.$$

Test against a specific group   Often researchers are interested in knowing whether the abundance increased or decreased in an ecosystem relative a pre-specified group, say the control group. Again, assume group 0 is the reference group and $\beta_0 = 0$, then one may be interested in testing the following hypotheses:

$$H_{0,k} : \beta_k = 0,$$
$$H_{1,k} : \{\beta_k < 0\} \cup \{\beta_k > 0\},$$

where $k \in \{1, \ldots, g\}$.

As before, the pairwise test statistic is defined as follows:

$$W_k = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}} \to_d N(0,1), \quad \text{as } n \to \infty,$$

where $\widehat{Var}(\hat{\beta}_k)$ is the kth diagonal elements of $\hat{\Sigma}^{(g)}$. Thus, the raw p-value for comparing group k and group 1 is defined as

$$p_k = 2[1 - \phi(|W_k|)].$$

Likewise, we apply the mdFDR controlling procedure for all pairwise tests. To improve power, we modify the global test mentioned earlier to a Dunnet-based[46–48] test as described below:

(1) The test statistic $W = \max\limits_{k \in \{1, \ldots, g\}} |W_k|$,

(2) Generate $W_k^{(b)} \sim N(0,1), k = 1, \ldots, g$.

(3) Compute $W^{(b)} = \max\limits_{k \in \{1, \ldots, g\}} |W_k^{(b)}|$.

(4) Repeat the above steps B times, we get the null distribution of W.

The screening p-value is calculated as:

$$p = \frac{1}{B} \sum_{b=1}^{B} I(W^{(b)} > W).$$

Pattern analysis

When the experimental groups are ordered naturally, such as doses of exposure or duration of exposure or stages of a disease, etc., for a given taxon, researchers may be interested in testing whether the abundance of the taxon is changing with the ordered experimental groups according to some specific pattern. Thus, the null and alternative hypotheses one wants to test become (assume group 0 is the reference group):

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_g = 0,$$
$$H_1 : \beta = (\beta_1, \ldots, \beta_g)^T \in \mathbb{C},$$

where $\mathbb{C}$ is one or a collection of patterns. Examples of patterns are given below.

Example 0.3 (Simple order).

$$\mathbb{C}_1 = \{0 \le \beta_1 \le \beta_2 \le \ldots \le \beta_g\} \quad \text{with at least one strict inequality.} \tag{22}$$

Example 0.4 (Tree order).

$$\mathbb{C}_2 = \{\beta_k \ge 0, k = 1, \ldots, g\} \quad \text{with at least one strict inequality.} \tag{23}$$

Example 0.5 (Umbrella order).

$$\mathbb{C}_4 = \{0 \le \beta_1 \le \ldots \le \beta_{k-1} \le \beta_k \ge \beta_{k+1} \ldots \ge \beta_g\} \quad \text{with at least one strict inequality.} \tag{24}$$

Estimation of β under a certain pattern (constraint) can be obtained by solving the following convex optimization problem[49]:

$$\hat{\beta}^{opt} = \underset{\beta \in \mathbb{C}}{\arg\min}(\hat{\beta} - \beta)^T \hat{\Sigma}^{(g)-1}(\hat{\beta} - \beta), \tag{25}$$

where $\hat{\Sigma}^{(g)}$ is the corresponding sub-matrix of $\hat{\Sigma}$ defined in (10). The solution to (25) can be numerically obtained by using a suitable convex optimization algorithm, such as CVRX[50].

Example 0.6. Suppose there are 3 groups, namely, groups 0 (reference), 1, and 2, and no other covariates. For each sample $i, i = 1, \ldots, n$, we have:

$$y_i = \theta_i + \mu + \beta_1 I\{group = 1\} + \beta_2 I\{group = 2\} + \varepsilon_i.$$

To test whether the group effect is monotonically increasing, we would like to test:

$H_0 : \beta_1 = \beta_2 = 0,$
$H_1 : \beta \in \mathbb{C} = \{0 \leq \beta_1 \leq \beta_2\}, \quad$ with at least one strict inequality.

The estimation of β under $\mathbb{C}$ can be obtained by solving:

$$\hat{\beta}^{opt} = \underset{\beta \in \mathbb{R}^2}{\arg\min}(\hat{\beta} - \beta)^T \hat{\Sigma}^{(g)-1}(\hat{\beta} - \beta),$$

s.t. $\mathbf{A}\beta \geq 0,$

where $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$, and $\beta = (\beta_1, \beta_2)^T$.

Once the constrained estimator is obtained, there exist a variety of options to test the above hypotheses. For example, one may consider William's type of statistic[51]. We adopt the following definitions from Peddada et al. (2002)[7] to facilitate the construction of the test statistic.

Definition 2 (Linked parameters): Two parameters in a given pattern are said to be linked if the inequality between them is specified *a priori*.

Definition 3 (Nodal parameter): For a given pattern, a parameter is said to be nodal if it is linked with every other parameter in the profile.

For example, every parameter is a nodal parameter in $\mathbb{C}_1$; no nodal parameter in $\mathbb{C}_2$; and β_k is the only nodal parameter in $\mathbb{C}_3$.

Definition 4 (Norm of maximum difference): Define the norm $l_\infty(\mathbb{C})$ of pattern $\mathbb{C}$ as the maximum difference between the estimates of two linked parameters.

For example, $l_\infty(\mathbb{C}_3) = \max\{\hat{\beta}_k, \hat{\beta}_k - \hat{\beta}_g\}$.

Given a collection of potential patterns, $\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_T$, the William's type of test statistic is defined as:

$W = \max\{l_\infty(\mathbb{C}_t), t = 1, \ldots, T\},$
with $\quad t^{opt} = \arg\max\{l_\infty(\mathbb{C}_t), t = 1, \ldots, T\},$

where t^{opt} is regarded as the optimal pattern for the microbial abundance of a specific taxon.

Under null hypothesis, the expectations for $\hat{\beta}_k, k = 1, \ldots, g$ are 0s; thus, we can construct the null distribution of W as follows:

(1) Generate $\hat{\beta}_k^{(b)} \sim \sqrt{\widehat{Var(\hat{\beta}_k)}}N(0, 1), l = 1, \ldots, g$.

(2) Obtain constrained regression estimators for $\hat{\beta}_k^{opt,(b)}$ using the convex optimization problem described above.

(3) Compute $W^{(b)} = \max\{l_\infty(\mathbb{C}_t), t = 1, \ldots, T\}$ using the simulated data under pre-specified patterns.

(4) Repeat the above steps B times, and we get the null distribution of W.

The raw p-value is calculated as

$$p = \frac{1}{B}\sum_{b=1}^{B}I(W^{(b)} > W).$$

We then apply the Holm-Bonferroni correction or BH procedure on raw p-values to control the FDR.

ANCOM-BC2 for mixed effects models

Similar to the fixed effects model stated in (3), for each taxon $j, j = 1, \ldots, d$, and each sample $i, i = 1, \ldots, n$, suppose each sample has n_i observations and $\sum_i n_i = n$. The offset-based mixed effects log-linear model is set up as

$$\mathbf{y_{ij}} = \theta_i \mathbf{1_{n_i}} + \mathbf{X_i}\beta_j + \mathbf{Z_i}\alpha_i + \varepsilon_{ij}, \tag{26}$$

where

(1) $\mathbf{y_{ij}}$ is the n_i-vector centered observed abundances,

(2) $\mathbf{1_{n_i}} = (1, \ldots, 1)^T \in \mathbb{R}^{n_i}$ is a vector of 1s,

(3) $\mathbf{X_i}$ is the $n_i \times p$ design matrix for fixed effects,

(4) β_j is the p-vector of fixed effects regression coefficients to be estimated,

(5) $\mathbf{Z_i}$ is the $n_i \times q$ design matrix for the random effects,

(6) α_i is the q-vector random effects,

(7) ε_{ij} is the n_i-vector residuals.

The following distributional assumptions are made

$$\alpha_i \sim N(\mathbf{0}, \mathbf{D_{q \times q}}),$$
$$\varepsilon_{ij} \sim N(\mathbf{0}, \sigma_j^2 \mathbf{I_{n_i}}),$$
$$\alpha_i \perp\!\!\!\perp \varepsilon_{ij} \quad \text{for } i = 1, \ldots, n.$$

Thus, for each taxon $j, j = 1, \ldots, d$, and each sample $i, i = 1, \ldots, n$, we have

$$\mathbf{y_{ij}} \sim N(\theta_i \mathbf{1_{n_i}} + \mathbf{X_i}\beta_j, \mathbf{H_{ij}}(\tau)),$$

where $\mathbf{H_{ij}}(\tau) = \mathbf{Z_i}\mathbf{D}\mathbf{Z_i}^T + \sigma_j^2 \mathbf{I_{n_i}}$ (or $\mathbf{H_{ij}}$ for short) denotes a general covariance matrix parametrized by τ.

Stack up observations across samples, we have:

$$\mathbf{y_j} = \theta + \mathbf{X}\beta_j + \mathbf{Z}\alpha + \varepsilon_j, \tag{27}$$

where

$$\mathbf{y_j} = \begin{bmatrix} \mathbf{y_{1j}} \\ \mathbf{y_{2j}} \\ \vdots \\ \mathbf{y_{nj}} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \mathbf{1_{n_1}} \\ \theta_2 \mathbf{1_{n_2}} \\ \vdots \\ \theta_n \mathbf{1_{n_n}} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \vdots \\ \mathbf{X_n} \end{bmatrix}, \quad \beta_j = \begin{bmatrix} \beta_{j1} \\ \beta_{j2} \\ \vdots \\ \beta_{jp} \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_1} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{Z_1} \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}, \quad \varepsilon_j = \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{nj} \end{bmatrix}.$$

That is,

$$\mathbf{y_j} \sim N\left(\theta + \mathbf{X}\beta_j, \mathbf{H_j}(\tau) = \begin{bmatrix} \mathbf{H_{1j}}(\tau) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{H_{2j}}(\tau) & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{H_{nj}}(\tau) \end{bmatrix}\right),$$

where $\mathbf{H_j}(\tau)$ (or $\mathbf{H_j}$ for short) is a block diagonal matrix.

Similarly, we estimate θ and β_j iteratively to obtain the corresponding preliminary estimators. As compared to Algorithm 1, the Maximum Likelihood (ML) is replaced with Restricted Maximum Likelihood (ReML)[52,53].

Algorithm 3 Iterative ReMLE

1: **Initialize:**
   For $j = 1, \ldots, d$
   $\theta \leftarrow \mathbf{0}$
   $\mathbf{y_j^{(crt)}} \leftarrow \mathbf{y_j} - \theta = \mathbf{y_j}$
   $\beta_j \leftarrow \text{ReML}(\mathbf{y_j^{(crt)}}) = \text{ReML}(\mathbf{y_j})$
2: **while** not converge **do**
3: $\quad \theta \leftarrow \frac{1}{d} \sum_{j=1}^{d} (\mathbf{y_j} - \mathbf{X}\beta_j)$
4: $\quad \mathbf{y_j^{(crt)}} \mathbf{y_j} - \theta$
5: $\quad \beta_j \leftarrow \text{ReML}(\mathbf{y_j^{(crt)}})$
6: **end while**

Note that the estimators for regression coefficients β_j and variance components τ are obtained iteratively by maximizing the following log-likelihood function:

$$\mathscr{L}(\tau|\mathbf{y_j}) = -\sum_{i=1}^{n} \log|\mathbf{H_{ij}}| - \sum_{i=1}^{n} \log|\mathbf{X_i}^T \mathbf{H_{ij}}^{-1} \mathbf{X_i}| - \sum_{i=1}^{n} (\mathbf{y_{ij}} - \mathbf{X_i}\beta_j)^T \mathbf{H_{ij}}^{-1} (\mathbf{y_{ij}} - \mathbf{X_i}\beta_j), \tag{28}$$

where $\beta_j \leftarrow (\mathbf{X}^T \mathbf{H_j}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{H_j}^{-1} \mathbf{y_j}$. As close-form solutions of (28) do not exist, Newton-Raphson method is usually employed[54].

Suppose on convergence, $\theta \leftarrow \theta^*, \mathbf{y_j^{(crt)}} \leftarrow \mathbf{y_j^{(crt)*}}, \mathbf{H} \leftarrow \mathbf{H}^*, \beta_j \leftarrow \beta_j^*$, we have

$$\theta^* = \frac{1}{d} \sum_{j=1}^{d} (\mathbf{y_j} - \mathbf{X}\beta_j^*),$$

$$\mathbf{y_j^{(crt)*}} = \mathbf{y_j} - \theta^*,$$

$$\beta_j^* = (\mathbf{X}^T \mathbf{H_j^{*-1}} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{H_j^{*-1}} \mathbf{y_j^{(crt)*}}.$$

It is easy to show that there exists a vector $\delta \in \mathbb{R}^p$, such that

$$E(\theta^*) = \theta - \mathbf{X}\delta,$$
$$E(\beta_j^*) = \delta + \beta_j.$$

i.e., β_j^* is a biased estimator for β_j.

Similar to the case of fixed effects model, we fit the Gaussian mixture model to each $\beta_{jk}, k = 1, \ldots, p$ separately, to correct the bias δ, and final estimators for β_j and θ are given by

$$\hat{\beta}_j = \beta_j^* - \hat{\delta}^{EM},$$

$$\hat{\theta} = \frac{1}{d} \sum_{j=1}^{d} (\mathbf{y_j} - \mathbf{X}\hat{\beta}_j).$$

Thus, for taxon j, the Wald statistic for the following hypotheses

$$H_0 : \mathbf{A}\beta_j = \mathbf{A}\beta_j^0,$$
$$H_1 : \mathbf{A}\beta_j \neq \mathbf{A}\beta_j^0,$$

is given by

$$W_j = (\mathbf{A}\hat{\beta}_j - \mathbf{A}\beta_j^0)^T (\mathbf{A}\hat{\Sigma}_j \mathbf{A}^T)^{-1} (\mathbf{A}\hat{\beta}_j - \mathbf{A}\beta_j^0)$$
$$= (\mathbf{A}\beta_j^* - \mathbf{A}\hat{\delta}^{EM} - \mathbf{A}\beta_j^0)^T (\mathbf{A}\hat{\Sigma}_j \mathbf{A}^T)^{-1} (\mathbf{A}\beta_j^* - \mathbf{A}\hat{\delta}^{EM} - \mathbf{A}\beta_j^0)$$
$$\rightarrow_d \chi_q^2,$$

where

(1) $\hat{\Sigma}_j = (\mathbf{X}^T \mathbf{H_j}^{*-1} \mathbf{X})^{-1},$

(2) $q = rank(\mathbf{A}).$

To draw statistical inference under inequality constraints in linear mixed effects model, we use William's type test statistic as specified in the last section, and adopt the Constrainted Linear Mixed Effects (CLME) framework[42] into ANCOM-BC. The procedure is summarized as follows:

(1) Obtain $\hat{\beta}_j$, the estimate of β_j under the null hypothesis,

(2) Compute the observed values of residuals and random effects: $\hat{\varepsilon}_j = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \hat{\mathbf{H}}_{\mathbf{j}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{H}}_{\mathbf{j}}^{-1}) \mathbf{y}_{\mathbf{j}}^{(\text{crt})*}$, and $\hat{\alpha} = \hat{\theta} \mathbf{Z}^T \hat{\mathbf{H}}_{\mathbf{j}}^{-1} \hat{\varepsilon}_j,$

(3) Standardize the observed values of the random effects and residuals. Define $\hat{\mathbf{a}} = SE(\hat{\alpha})^{-1} \hat{\alpha}$, and $\hat{\mathbf{e}} = SE(\hat{\varepsilon})^{-1} \hat{\varepsilon}$, where $SE(\cdot)$ denotes the standard error,

(4) Obtain bootstrap samples. Let $\mathbf{a}^{(b)}$ and $\mathbf{e}^{(b)}$ denote the bootstrap samples of $\mathbf{a}$ and $\mathbf{e}$, respectively. Then define $\hat{\alpha}^{(b)} = SE(\hat{\alpha})\mathbf{a}^{(b)}$ and $\hat{\varepsilon}^{(b)} = SE(\hat{\varepsilon})\mathbf{e}^{(b)}$. Finally construct the final bootstrap sample as: $\mathbf{y}_{\mathbf{j}}^{(b)} = \hat{\theta} + \mathbf{X}\hat{\beta}_j + \mathbf{Z}\hat{\alpha}^{(b)} + \hat{\varepsilon}^{(b)},$

(5) Repeat B times ($b = 1, \ldots, B$), we construct the null distribution for the test statistic.

(6) Compute raw p-values,

(7) Apply the Holm-Bonferroni correction (default) or BH procedure on raw p-values to control the FDR.

Data availability

Simulation data can be found on the corresponding GitHub repository[55]. The soil microbiome data for aridity can be found in Qiita: `https://qiita.ucsd.edu/study/description/10360`. The gut microbiome data for IBD can be found in Qiita: `https://qiita.ucsd.edu/study/description/11546`.

Code availability

ANCOM-BC2 has been implemented in the R package ANCOMBC, which is available on Bioconductor at `https://www.bioconductor.org/packages/release/bioc/html/ANCOMBC.html`. All analyses shown in the paper can be found on the corresponding GitHub repository[55].

References

1. Tierney, B. T. *et al.* Multidrug-resistant acinetobacter pittii is adapting to and exhibiting potential succession aboard the international space station. *Microbiome* **10**, 1–14 (2022).

2. Neilson, J. W. *et al.* Significant impacts of increasing aridity on the arid soil microbiome. *MSystems* **2**, e00195–16 (2017).

3. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. ecology health disease* **26**, 27663 (2015).

4. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. communications* **11**, 1–11 (2020).

5. Hu, Y., Satten, G. A. & Hu, Y.-J. Locom: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. *Proc. Natl. Acad. Sci.* **119**, e2122788119 (2022).

6. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

7. Peddada, S. D. *et al.* Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference. *Bioinformatics* **19**, 834–841 (2003).

8. Guo, W., Sarkar, S. K. & Peddada, S. D. Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. *Biometrics* **66**, 485–492 (2010).

9. Grandhi, A., Guo, W. & Peddada, S. D. A multiple testing procedure for multi-dimensional pairwise comparisons with application to gene expression studies. *BMC bioinformatics* **17**, 104 (2016).

10. Gohir, W. *et al.* Pregnancy-related changes in the maternal gut microbiota are dependent upon the mother's periconceptional diet. *Gut microbes* **6**, 310–320 (2015).

11. Wu, H.-J. & Wu, E. The role of gut microbiota in immune homeostasis and autoimmunity. *Gut microbes* **3**, 4–14 (2012).

12. Koren, O. *et al.* Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* **150**, 470–480 (2012).

13. Mesa, M. D. *et al.* The evolving microbiome from pregnancy to early infancy: a comprehensive review. *Nutrients* **12**, 133 (2020).

14. Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of microbiome data in the presence of excess zeros. *Front. microbiology* **8**, 2114 (2017).

15. Zhou, H., He, K., Chen, J. & Zhang, X. Linda: linear models for differential abundance analysis of microbiome compositional data. *Genome biology* **23**, 1–23 (2022).

16. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **8**, e46923 (2019).

17. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**, 5116–5121 (2001).

18. Costea, P. I., Zeller, G., Sunagawa, S. & Bork, P. A fair comparison. *Nat. methods* **11**, 359–359 (2014).

19. Paulson, J. N., Bravo, H. C. & Pop, M. Reply to:" a fair comparison". *Nat. methods* **11**, 359–360 (2014).

20. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).

21. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. journal statistics* 65–70 (1979).

22. Lim, C., Sen, P. K. & Peddada, S. D. Robust analysis of high throughput screening (hts) assay data. *Technometrics* **55**, 150–160 (2013).

23. Belov, A. A., Cheptsov, V. S. & Vorobyova, E. A. Soil bacterial communities of sahara and gibson deserts: Physiological and taxonomical characteristics. *AIMS microbiology* **4**, 685 (2018).

24. Peddada, S., Harris, S., Zajd, J. & Harvey, E. Oriogen: order restricted inference for ordered gene expression data. *Bioinformatics* **21**, 3933–3934 (2005).

25. Botero, L. M. *et al.* Thermobaculum terrenum gen. nov., sp. nov.: a non-phototrophic gram-positive thermophile representing an environmental clone group related to the chloroflexi (green non-sulfur bacteria) and thermomicrobia. *Arch. microbiology* **181**, 269–277 (2004).

26. Lau, C. H.-F., van Engelen, K., Gordon, S., Renaud, J. & Topp, E. Novel antibiotic resistance determinants from agricultural soil exposed to antibiotics widely used in human medicine and animal farming. *Appl. Environ. Microbiol.* **83**, e00989–17 (2017).

27. Oyejobi, G. K., Sule, W. F., Akinde, S. B., Khan, F. M. & Ogolla, F. Multidrug-resistant enteric bacteria in nigeria and potential use of bacteriophages as biocontrol. *Sci. The Total. Environ.* 153842 (2022).

28. Montero-Calasanz, M. d. C. *et al.* Geodermatophilus poikilotrophi sp. nov.: a multitolerant actinomycete isolated from dolomitic marble. *BioMed research international* **2014** (2014).

29. Li, J. L. *et al.* Antichlamydial dimeric indole derivatives from marine actinomycete rubrobacter radiotolerans. *Planta Medica* **83**, 805–811 (2017).

30. Walker, C. B. *et al.* Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc. Natl. Acad. Sci.* **107**, 8818–8823 (2010).

31. Grady, E. N., MacDonald, J., Liu, L., Richman, A. & Yuan, Z.-C. Current knowledge and perspectives of paenibacillus: a review. *Microb. cell factories* **15**, 1–18 (2016).

32. Chen, H. *et al.* One-time nitrogen fertilization shifts switchgrass soil microbiomes within a context of larger spatial and temporal variation. *Plos one* **14**, e0211310 (2019).

33. Fang, X. *et al.* Gastrointestinal surgery for inflammatory bowel disease persistently lowers microbiome and metabolome diversity. *Inflamm. bowel diseases* **27**, 603–616 (2021).

34. López-Almela, I. *et al.* Bacteroides uniformis combined with fiber amplifies metabolic and immune benefits in obese mice. *Gut Microbes* **13**, 1–20 (2021).

35. Horvath, T. D. *et al.* Bacteroides ovatus colonization influences the abundance of intestinal short chain fatty acids and neurotransmitters. *Iscience* **25**, 104158 (2022).

36. Chang, S.-C. *et al.* A gut butyrate-producing bacterium butyricicoccus pullicaecorum regulates short-chain fatty acid transporter and receptor to reduce the progression of 1, 2-dimethylhydrazine-associated colorectal cancer. *Oncol. Lett.* **20**, 1–1 (2020).

37. Peterson, C. T. *et al.* Short-chain fatty acids modulate healthy gut microbiota composition and functional potential. *Curr. Microbiol.* **79**, 128 (2022).

38. Zhou, Y. *et al.* F. prausnitzii and its supernatant increase scfas-producing bacteria to restore gut dysbiosis in tnbs-induced colitis. *AMB express* **11**, 1–10 (2021).

39. Nie, K. *et al.* Roseburia intestinalis: A beneficial gut organism from the discoveries in genus and species. *Front. Cell. Infect. Microbiol.* 1147 (2021).

40. Rau, M. *et al.* Fecal scfas and scfa-producing bacteria in gut microbiome of human nafld as a putative link to systemic t-cell activation and advanced disease. *United Eur. gastroenterology journal* **6**, 1496–1507 (2018).

41. Farnan, L., Ivanova, A. & Peddada, S. Constrained inference in biological sciences: linear mixed effects models under constraints. *PLoS ONE* **9**, e84778 (2014).

42. Jelsema, C. M. & Peddada, S. D. Clme: an r package for linear mixed effects models under inequality constraints. *J. statistical software* **75** (2016).

43. Rao, C. R. *Linear statistical inference and its applications*, vol. 2 (Wiley New York, 1973).

44. Peddada, S. D. & Smith, T. Consistency of a class of variance estimators in linear models under heteroscedasticity. *Sankhyā: The Indian J. Stat. Ser. B* 1–10 (1997).

45. McLachlan, G. & Krishnan, T. *The EM algorithm and extensions*, vol. 382 (John Wiley & Sons, 2007).

46. Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* **50**, 1096–1121 (1955).

47. Dunnett, C. W. & Tamhane, A. C. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Stat. medicine* **10**, 939–947 (1991).

48. Dunnett, C. W. & Tamhane, A. C. A step-up multiple test procedure. *J. Am. Stat. Assoc.* **87**, 162–170 (1992).

49. Silvapulle, M. J. & Sen, P. K. *Constrained statistical inference: Order, inequality, and shape constraints* (John Wiley & Sons, 2011).

50. Fu, A., Narasimhan, B. & Boyd, S. Cvxr: An r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582* (2017).

51. Williams, D. A. Some inference procedures for monotonically ordered normal means. *Biometrika* **64**, 9–14 (1977).

52. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).

53. Harville, D. A. Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385 (1974).

54. Lindstrom, M. J. & Bates, D. M. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Stat. Assoc.* **83**, 1014–1022 (1988).

55. Lin, H. Multi-group analysis of compositions of microbiomes with covariate adjustments and repeated measures. https://github.com/FrederickHuangLin/ANCOM-BC2-Code-Archive (2023).

Acknowledgements

Author contributions

Both authors contributed equally to the theory and methodology described in this paper. All numerical works and computations were conducted by H.L. who developed ANCOM-BC2 pipeline in R that is freely and publicly available. Please contact H.L. for software requests.

Competing interests

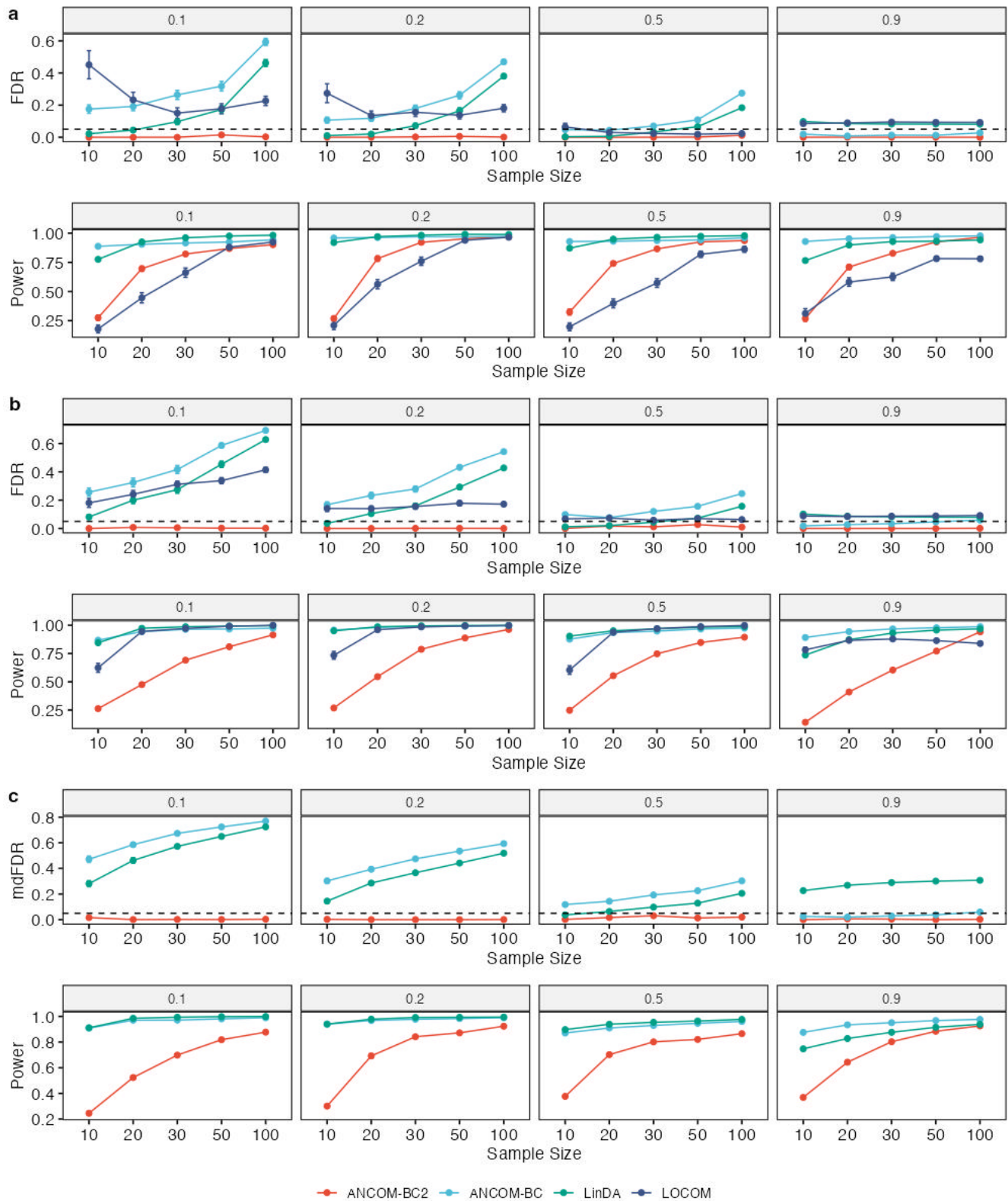Authors declare no conflicts of interest.

Figures

Figure 1. Comparisons of FDR (mdFDR) and power in identifying DA taxa in **(a)** continuous, **(b)** binary, or **(c)** categorical exposure. Synthetic data were generated based on the QMP data[20]. The X-axis shows the sample size (or sample size per group for the categorical covariate), and the Y-axis shows the FDR (mdFDR) or power. The proportion of true DA taxa is indicated in the panel title. Results are represented by the average of the corresponding measure (FDR, mdFDR, or power) $\pm$ standard errors (shown as error bars) across 100 simulation runs for each setting. The results demonstrate that ANCOM-BC2 outperformed all competing methods in terms of uniformly small FDR (mdFDR) and comparable power.
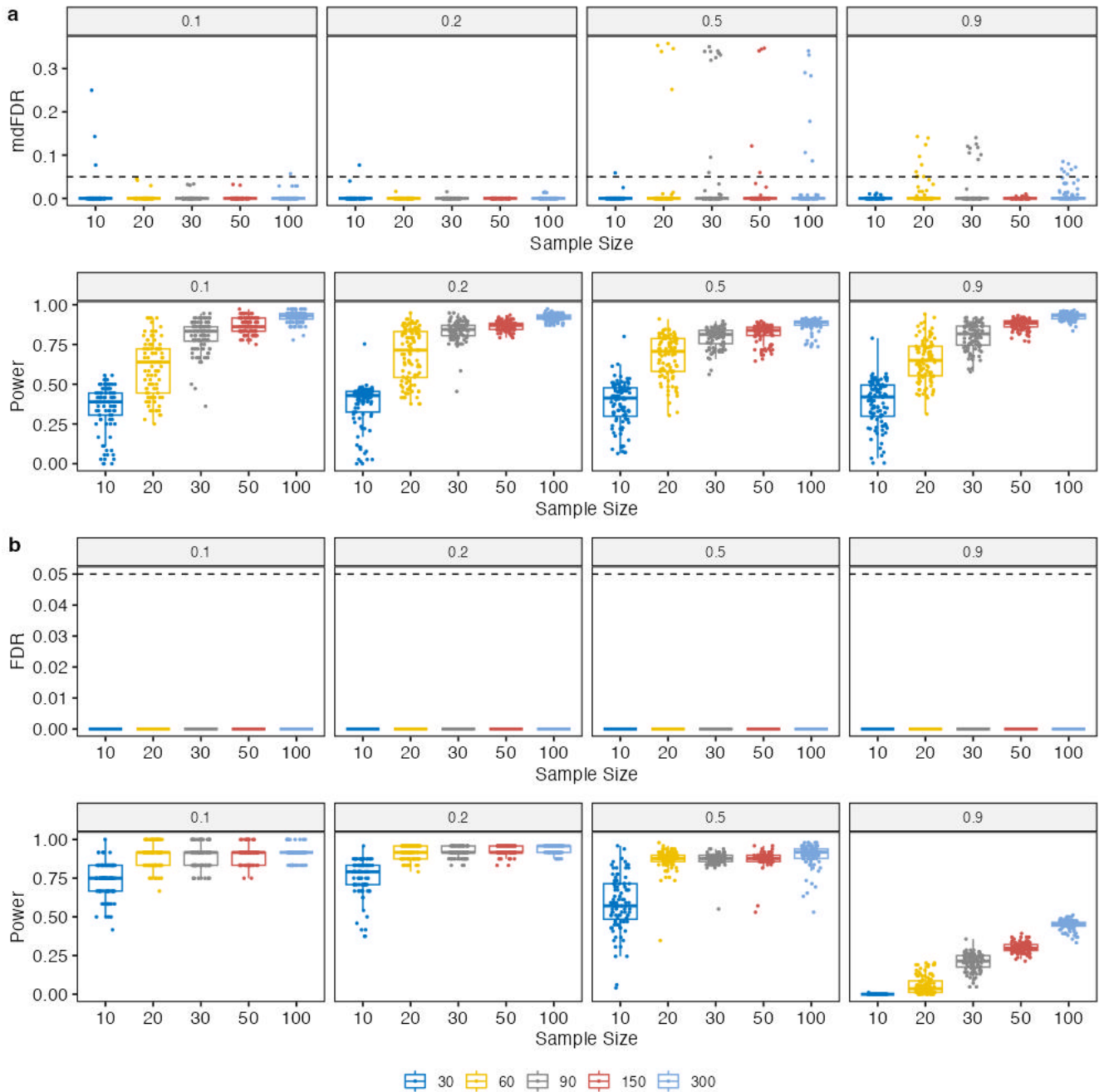
Figure 2. Comparisons of FDR (mdFDR) and power in identifying DA taxa in **(a)** multiple pairwise comparisons, and **(b)** pattern analysis. Synthetic data were generated based on the QMP data[20]. The X-axis shows the sample size per group, and the Y-axis shows the FDR (mdFDR) or power. The proportion of true DA taxa is indicated in the panel title. The box plot shows the distribution of the corresponding measure (FDR or power) across 100 simulation runs for each setting. Each box represents the interquartile range (IQR) of the data, with the horizontal line in the box representing the median. The whiskers extend to the furthest data point which is within 1.5 times the IQR. Any data points beyond this range are shown as individual points. Additionally, the data points are jittered to avoid overlap and provide a better visualization of the data. The results demonstrate that ANCOM-BC2 controlled FDR (mdFDR) while maintaining adequate power.
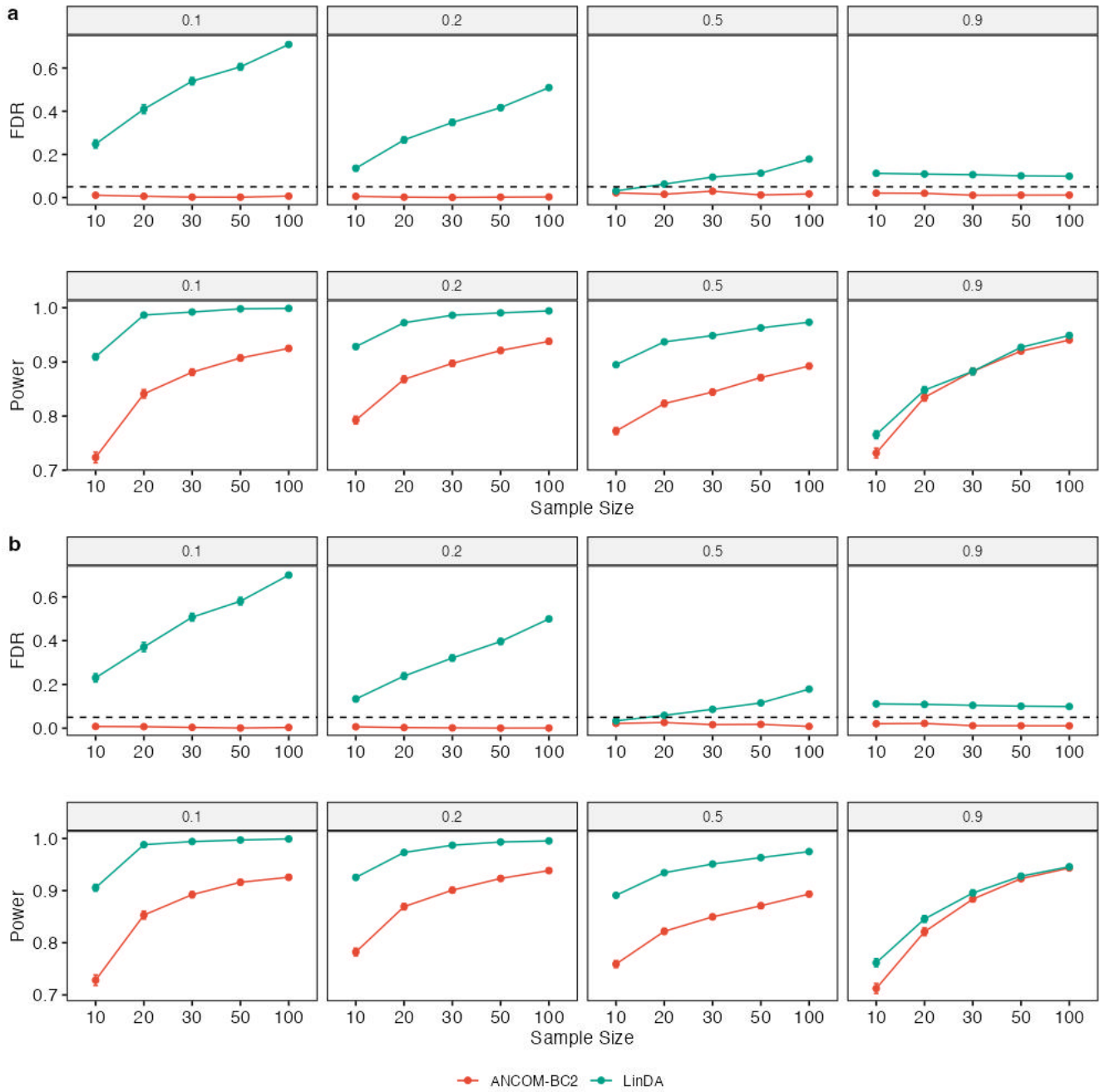
Figure 3. Comparisons of FDR and power in identifying DA taxa in **(a)** a random intercept model, and **(b)** a random coefficients model. Synthetic data were generated based on the QMP data[20]. The X-axis shows the sample size per group, and the Y-axis shows the FDR or power. The proportion of true DA taxa is indicated in the panel title. Results are represented by the average of the corresponding measure (FDR or power) ± standard errors (shown as error bars) across 100 simulation runs for each setting. The results demonstrate that ANCOM-BC2 outperformed LinDA in terms of uniformly small FDR and comparable power.
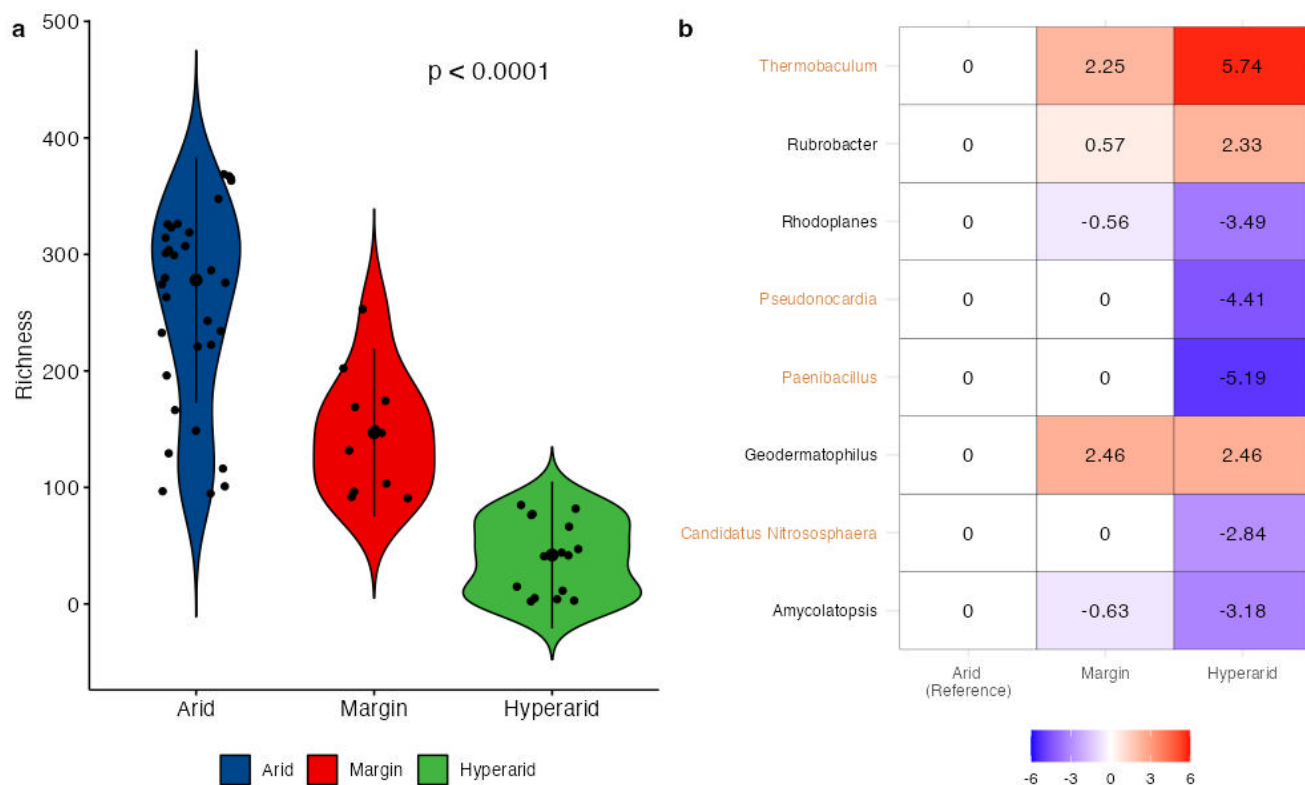
Figure 4. Differential abundance analysis for the impact of aridity in desert soil samples. **(a)** Violin plot of the relationship between aridity and microbial richness. The black dot in the middle of each violin represents the median value, while the black bar represents the interquartile range (IQR). The width of each violin reflects the density of data points at each value of microbial richness. Individual data points are also shown as jittered dots. The ORIOGEN software was utilized to conduct a trend test, which revealed a significant reduction in species richness with increasing aridity with a p-value < 0.0001. **(b)** Heatmap of ANCOM-BC2 pattern analysis with respect to aridity. Monotonic increasing and decreasing trends were evaluated across ordered soil categories, with arid soil as the reference category. The X-axis represents soil categories, while the Y-axis displays significant genera identified by ANCOM-BC2 pattern analysis. Each cell is color-coded with blue representing reduced abundance and red representing increased abundance, with the numbers on each cell indicating the log fold-change relative to the reference group (arid group). The Holm-Bonferroni method was utilized to correct for multiple testing, and genera colored in brown on the Y-axis were found to be significant after correcting for multiple testing.

Figure 5. Heatmap of ANCOM-BC2 pairwise analysis for the effect of surgical resection in a cohort of patients with IBD. The analysis included multiple pairwise comparisons among the three groups: ileocolonic resection, colectomy, and no intestinal surgery, while controlling the overall mdFDR at 0.05. The X-axis represents the specific comparisons made between ileocolonic resection vs. no intestinal surgery, colectomy vs. no intestinal surgery, and ileocolonic resection vs. colectomy. The Y-axis displays the significant species identified by ANCOM-BC2. Each cell in the heatmap is color-coded with blue representing reduced abundance and red representing increased abundance, and the numbers on each cell indicate the log fold-change. The Holm-Bonferroni method was utilized to correct for multiple testing.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.csv
- ANCOMBC2SupplementaryInformation.pdf