Biometrika (2023), **110**, 1, *pp.* 169–185 https://doi.org/10.1093/biomet/asac029
Advance Access publication 23 May 2022

Robust differential abundance test in compositional data

By SHULEI WANG

Department of Statistics, University of Illinois at Urbana-Champaign, 605 E. Springfield Avenue, Champaign, Illinois 61820, U.S.A.
shuleiw@illinois.edu

Summary

Differential abundance tests for compositional data are essential and fundamental in various biomedical applications, such as single-cell, bulk RNA-seq and microbiome data analysis. However, because of the compositional constraint and the prevalence of zero counts in the data, differential abundance analysis on compositional data remains a complicated and unsolved statistical problem. This article proposes a new differential abundance test, the robust differential abundance test, to address these challenges. Compared with existing methods, the robust differential abundance test is simple and computationally efficient, is robust to prevalent zero counts in compositional datasets, can take the data's compositional nature into account, and has a theoretical guarantee of controlling false discoveries in a general setting. Furthermore, in the presence of observed covariates, the robust differential abundance test can work with covariate-balancing techniques to remove potential confounding effects and draw reliable conclusions. The proposed test is applied to several numerical examples, and its merits are demonstrated using both simulated and real datasets.

Some key words: Compositional data; Covariate balancing; Differential abundance test; Multiple testing.

1. Introduction

Compositional data, in which all components are nonnegative and their sum is equal to 1, naturally arise in a wide range of modern scientific applications, including human microbiome studies, nutritional science, genomics studies and geochemistry. An essential and fundamental task in these scientific applications is differential abundance testing, which aims to identify a set of differential components across experimental conditions (Robinson et al., 2010; Fernandes et al., 2014; Kharchenko et al., 2014; Law et al., 2014; Love et al., 2014; Mandal et al., 2015; Risso et al., 2018; Morton et al., 2019). However, differential abundance analysis in compositional data is a complicated and challenging statistical problem because of the constant-sum constraint. Applying standard statistical analysis directly to compositional data ignores the data's compositional nature, and hence can result in an ill-defined statistical hypothesis and false-positive scientific discovery (Vandeputte et al., 2017; Weiss et al., 2017; Morton et al., 2019; Brill et al., 2020; Lin & Peddada, 2020).

To account for the constant-sum constraint and gain reliable scientific insights from compositional data, many differential abundance testing methods have been proposed for different types of biomedical compositional datasets, including single-cell, bulk RNA-seq and microbiome

data (Robinson et al., 2010; Paulson et al., 2013; Fernandes et al., 2014; Kharchenko et al., 2014; Law et al., 2014; Love et al., 2014; Mandal et al., 2015; Lê Cao et al., 2016; Butler et al., 2018; Risso et al., 2018; Morton et al., 2019; Lin & Peddada, 2020; Martin et al., 2020). These methods have been very successful in many applications and have helped researchers make significant progress in various scientific fields. However, most of the existing methods require taking a ratio, or log-ratio, of two proportions and thus implicitly assume that each component's proportions are strictly greater than zero. Unfortunately, zero counts are prevalent in some biomedical compositional datasets, such as microbiome data (Cao et al., 2020), because of insufficient sequence depth, also called technical zero, or biological variation, also called structural zero. Using pseudo-counts, i.e., replacing zero counts with a small positive number, could alleviate this zero-count problem. However, one usually does not know how to choose a pseudo-count for a given dataset, and this strategy could potentially lead to inflated false-positive discoveries, as noted by Brill et al. (2020).

These challenges raise the question of whether there exists a differential abundance test for compositional data that has the following properties: (i) it is simple, easy to implement and computationally efficient; (ii) it is robust to prevalent zero counts in compositional datasets, eliminating the need for an extra step to handle the zero-count problem; (iii) it can take the compositional nature of the data into account; and (iv) it has a theoretical guarantee of controlling false discoveries in a general setting. This article develops a new robust differential abundance test that meets all these requirements.

Specifically, motivated by the studies of Morton et al. (2019) and Brill et al. (2020), the differential abundance analysis is formulated as a reference-based hypothesis testing problem. That is, the hypothesis is defined with respect to a set of unknown reference components, and its identification conditions are studied. Our investigation shows that under such a reference-based hypothesis, the differential components can be recovered completely by a series of directional comparisons on renormalized proportions in the noiseless case. Based on this observation, we introduce a new iterative method for identifying the differential components in compositional data. Unlike existing methods, the new method only needs to evaluate standard two-sample t-test statistics, also known as Welch's t test, on renormalized proportions in each iteration, so that one does not need to worry about the zero-count problem any more. Owing to the reference-based hypothesis, the new method follows a two-stage procedure to identify the differential components in each iteration: (i) the t-test statistics of all components are integrated to determine the testing direction; (ii) the differential components are identified by a componentwise directional two-sample test. The idea of such a two-stage strategy is inspired by the empirical Bayesian analysis perspective (Robbins, 1951; Efron, 2012) and has been used in other large-scale simultaneous hypothesis testing problems. For example, Efron (2004) proposed empirically estimating the mean and variance of the null distribution before conducting large-scale hypothesis testing. Unlike these existing methods, we iteratively apply the two-stage strategy. To demonstrate the merit of the proposed method, we study its theoretical properties. More concretely, we show that the familywise error rate can be controlled at the α level asymptotically in a general setting, and the differential components can be identified completely with high probability when the signal-to-noise ratio is large enough. The theoretical analysis may be of interest in itself because of the challenges that arise in handling the dependency among test statistics caused by renormalization and the iterative procedure.

As more observational studies collect compositional data, another difficulty in differential abundance analysis is reducing the bias introduced by the potential confounding covariates. As the proposed robust differential abundance test comprises a series of standard two-sample t tests, it can work with most covariate-balancing techniques designed for the potential outcome

framework (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015). In particular, we combine the proposed test with weighting methods (Rosenbaum, 1987; Robins et al., 2000; Imai & Ratkovic, 2014; Chan et al., 2016) to remove the potential confounding effect of observed covariates. Furthermore, the idea of employing this iterative method in the robust differential abundance test can also be extended to control the false discovery rate and test the linear association with a continuous outcome. An `R` package implementing the robust differential abundance test is available at `https://github.com/lakerwsl/RDB`.

2. Model and reference-based hypothesis

2.1. *Model and notation*

Suppose we wish to compare the abundance of d components, such as d different taxa. The absolute abundance of these d components in a sample can be represented by a nonnegative vector $A = (A_1, \ldots, A_d)$, where $A_i \geqslant 0$ is the absolute abundance of the ith component. Instead of directly observing the absolute abundance A, in many real applications only count data $N = (N_1, \ldots, N_d)$ are collected, where N_i is the number observed of the ith component. Here, given the absolute abundance A, we assume that the count data are drawn from a multinomial model

$$N \mid P, N^* \sim \mathrm{Mu}(N^*, P), \tag{1}$$

where $P = (P_1, \ldots, P_d)$ is the relative abundance of the d components and $N^* = \sum_i N_i$ is the total count observed in a sample. The relative abundance P is defined as $P_i = A_i/A^*$, where $A^* = \sum_{i=1}^d A_i$. As the total count N^* is usually not proportional to the absolute abundance, we normalize the count data as a compositional vector, i.e., $\hat{P} = (\hat{P}_1, \ldots, \hat{P}_d)$ where $\hat{P}_i = N_i/N^*$. Clearly, the compositional data $\hat{P}$ that we observe constitute an empirical version of the relative abundance P, and thus reflect only the information on relative abundance rather than absolute abundance.

In the differential abundance test we assume two populations of interest, such as the treated group and the control group, which can be written as $\pi_1(A)$ and $\pi_2(A)$, respectively. As we wish to compare the abundance of the two populations, we draw m_1 and m_2 samples from each population:

$$A_{k,1}, A_{k,2}, \ldots, A_{k,m_k} \sim \pi_k(A) \quad (k = 1, 2).$$

For the jth sample in the kth population, we observe count data $N_{k,j} = (N_{k,j,1}, \ldots, N_{k,j,d})$ drawn from the model (1) given the absolute abundance $A_{k,j}$. For a fair comparison, these count data can be normalized as the empirical relative abundance $\hat{P}_{k,j} = (\hat{P}_{k,j,1}, \ldots, \hat{P}_{k,j,d})$. Therefore, the goal of the differential abundance test is to compare the two populations π_1 and π_2 based on the observed compositional data $\hat{P}_{k,1}, \hat{P}_{k,2}, \ldots, \hat{P}_{k,m_k}$ $(k = 1, 2)$. Hereafter, we always use i as the index of the component, j as the index of the sample and k as the index of the population. After introducing the data-generating model, we also need to define a hypothesis for the differential abundance test. It seems straightforward to define a rigorous hypothesis in such a two-sample case. However, we will see that the data's compositional nature makes it difficult to define a statistically identifiable and scientifically meaningful hypothesis.

The general goal of the differential abundance test is to identify a set of components with different abundances across two populations. This task can be naturally formulated as a hypothesis

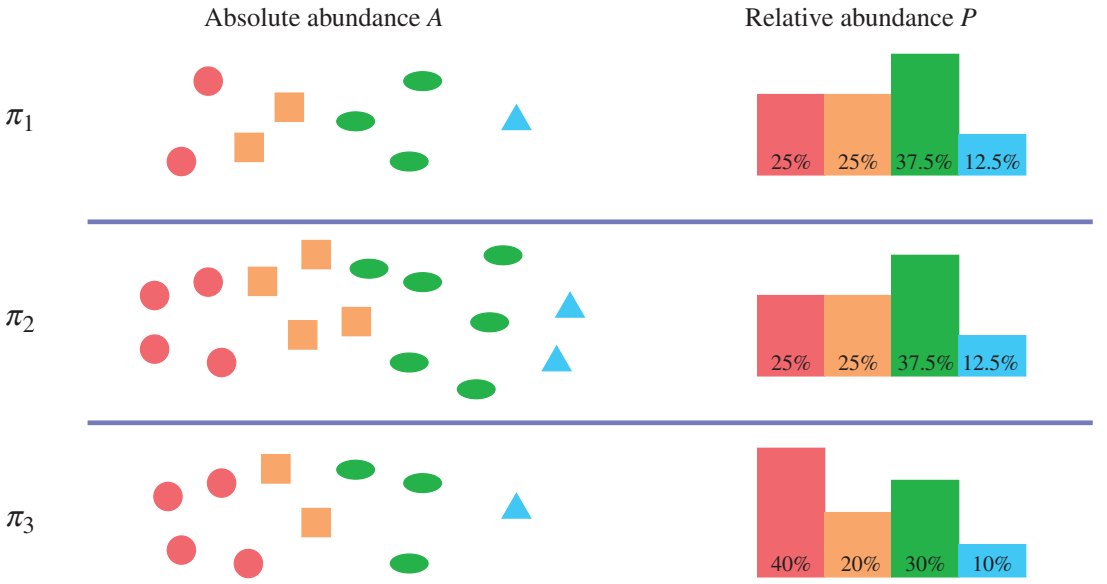

Fig. 1. Comparison of absolute abundances and relative abundances.

testing problem on the absolute abundance,

$$H_{i,0} : E_{\pi_1}(A_i) = E_{\pi_2}(A_i) \quad \text{versus} \quad H_{i,1} : E_{\pi_1}(A_i) \neq E_{\pi_2}(A_i), \tag{2}$$

where $E_{\pi_k}(A_i)$ represents the expected absolute abundance of the ith component in the kth population. However, the following example suggests that no method can consistently test the hypothesis in (2) based on the compositional data unless further assumptions are made.

Example 1. Consider the distributions π_1 and π_2 in Fig. 1. Examining the absolute abundance of π_1 and of π_2, we see that the absolute abundances of all components are doubled in π_2 compared with π_1. However, observing only relative abundance, we would conclude that there is no difference between π_1 and π_2. In other words, we cannot distinguish the two populations by only observing the relative abundance.

Example 1 suggests that the information provided by the compositional data is insufficient to answer a question such as (2). To make the hypothesis identifiable, one possible solution is to ignore the compositional nature of the data and directly test the relative abundance,

$$H_{i,0} : Q_{1,i} = Q_{2,i} \quad \text{versus} \quad H_{i,1} : Q_{1,i} \neq Q_{2,i}, \tag{3}$$

where $Q_{1,i} = E_{\pi_1}(P_i)$ and $Q_{2,i} = E_{\pi_2}(P_i)$ are the expected relative abundances of the ith component in the two populations. Although a hypothesis such as (3) is testable based on the compositional data, it may lead to a false discovery as finding a meaningful connection to absolute abundance is difficult. To illustrate this, consider the following example.

Example 2. Consider the distributions π_1 and π_3 in Fig. 1. In terms of absolute abundance, only the red component's absolute abundance is greater in π_3 than in π_1, and the absolute abundances of the remaining components stay the same across the populations. However, examining relative abundance, we observe that the red component's relative abundance increases from π_1 to π_3

while the relative abundances of the other components decrease owing to the compositional constraint. Simply put, the changes in relative abundance are not equivalent to the changes in absolute abundance.

In the above example, only the red component is the driver component; however, all the components belong to the alternative hypothesis in (3). These two examples indicate that it is important to define an identifiable and interpretable statistical hypothesis carefully for the compositional data.

2.2. *Reference-based hypothesis*

As every piece of information introduced by compositional data is relative, it is necessary to define reference frames to analyse the compositional data, as pointed out by Pawlowsky-Glahn & Buccianti (2011) and Morton et al. (2019). The idea of reference frames has also been employed in standard compositional data analysis. For example, the geometric mean of all components is viewed as a reference component in the centre log-ratio transformation. A change in compositional data can be interpreted as a change with respect to the reference components through the reference frames.

Specifically, motivated by the work of Brill et al. (2020), a subset of components I_0 is called a reference set if there exists a positive constant $b > 0$ such that the relative abundance of components in I_0 changes by the same amplitude:

$$Q_{1,i} = bQ_{2,i}, \quad i \in I_0. \tag{4}$$

The definition of the reference set suggests that the relative relationships within the reference set I_0 do not change across the two populations, i.e., $Q_{1,i}/Q_{1,i'} = Q_{2,i}/Q_{2,i'}$ for any $i, i' \in I_0$, so that I_0 can be seen as a benchmark. For example, the orange, green and blue components in Example 2 can be seen as a reference set, although their relative abundances differ between π_1 and π_3. Based on the reference set I_0, we compare everything with the components in I_0, and thus the problem of testing for differential components can be cast as the following reference-based hypothesis testing problem:

$$H_{i,0} : \frac{Q_{1,i}}{\sum_{i\in I_0} Q_{1,i}} = \frac{Q_{2,i}}{\sum_{i\in I_0} Q_{2,i}} \quad \text{versus} \quad H_{i,1} : \frac{Q_{1,i}}{\sum_{i\in I_0} Q_{1,i}} \neq \frac{Q_{2,i}}{\sum_{i\in I_0} Q_{2,i}}. \tag{5}$$

Under the above hypothesis, a change in some components is interpreted as a change with respect to the reference set I_0. Unlike in the conventional statistical hypothesis testing problem, the reference-based hypothesis at each component is defined by all components. If a component belongs to the reference set, then the null hypothesis is naturally true. The definition in (4) also defines a set of nondifferential components. If we further assume that the absolute abundance within the reference set is unchanged across two populations, i.e.,

$$E_{\pi_1}(A_i) = E_{\pi_2}(A_i), \quad i \in I_0, \tag{6}$$

then hypothesis (2) is roughly equivalent to hypothesis (5). In other words, under the assumption (6), the compositional data can be used to infer the differential components defined based on absolute abundance. The assumption (6) cannot be diagnosed and verified based on the compositional data alone.

The definition of hypothesis (5) depends on the choice of the reference set. If domain knowledge or extra information on the reference set is available in advance, (5) is well-defined based on

the known I_0 and ready to be tested by the compositional data. However, in practice a more realistic situation is that the reference set is unknown in advance. In this case, different choices of reference set can lead to inconsistent conclusions of the null hypothesis, and hence issues of model identifiability. The following proposition shows that assumptions are required to make the hypothesis testing problem identifiable.

PROPOSITION 1. *If the reference set I_0 in (5) is assumed to satisfy $|I_0| > d/2$, then the null and alternative hypotheses in (5) are well-defined. In contrast, there exists an instance of two reference sets $I_{0,1}$ and $I_{0,2}$ defined in (4) with $|I_{0,1}|, |I_{0,2}| \leqslant d/2$ such that the null hypotheses defined by $I_{0,1}$ and $I_{0,2}$ contradict each other.*

This proposition suggests that when the reference set is unknown, it is necessary to assume the existence of a large enough reference set to ensure that the hypothesis for the differential abundance test in compositional data is well-defined. Hereafter, we always assume that the reference set I_0 in (5) satisfies $|I_0| > d/2$.

2.3. *Existing methods for reference-based hypothesis testing*

To test the hypothesis in (5), several different methods have been proposed that use standard compositional analysis techniques. Mandal et al. (2015) proposed a method called analysis of composition of microbiome. To test whether the ith component is a differential one, Mandal et al.'s method compares the ith component with all other components by testing

$$\begin{aligned} & H_{i,i',0} : E_{\pi_1}\{\log(P_i/P_{i'})\} = E_{\pi_2}\{\log(P_i/P_{i'})\} \\ \text{versus} \quad & H_{i,i',1} : E_{\pi_1}\{\log(P_i/P_{i'})\} \neq E_{\pi_2}\{\log(P_i/P_{i'})\} \end{aligned}$$

for all $i' \neq i$. After $d(d-1)/2$ hypothesis tests, their method summarizes all the decisions by the number of null hypothesis rejections,

$$W_i = \sum_{i' \neq i} I\big[E_{\pi_1}\{\log(P_i/P_{i'})\} \neq E_{\pi_2}\{\log(P_i/P_{i'})\}\big],$$

where $I(\cdot)$ denotes the indicator function. The ith component is a differential one if $W_i > d/2$. The main disadvantage of this method is that comparing $O(d^2)$ hypotheses is time-consuming in practice. Using a similar idea, Brill et al. (2020) proposed first identifying the reference set by evaluating

$$S_i = \text{Median}\big[\big|E_{\pi_1}\{\log(P_i/P_{i'})\} - E_{\pi_2}\{\log(P_i/P_{i'})\}\big| : i' \neq i\big].$$

The ith component belongs to the reference set if S_i is small. After the reference set is estimated, the differential components can be identified by comparing each component with the estimated reference set. However, it is unclear how to estimate the reference set effectively, and the effect of a misspecified reference set on the results remains unknown. Unlike the previous two methods, Morton et al. (2019) proposed ranking the components by the ratio between the two populations. Specifically, the ratio for the ith component is defined as

$$T_i = E_{\pi_1}\{\log(P_i)\} - E_{\pi_2}\{\log(P_i)\}.$$

The mode of T_i is defined as T^* such that $|\{i : T_i = T^*\}|$ is the largest. Then the differential components are defined as those with $T_i \neq T^*$. However, with this method it can be difficult to estimate the mode of T_i when there is noise.

In all the above methods, the important tool used to account for the compositional nature of the data is the ratio or log-ratio of two proportions. This idea also commonly appears in the existing literature on compositional data analysis (Aitchison, 1983; Pawlowsky-Glahn & Buccianti, 2011). Its main advantages are that it can represent the relative information in a clean form and can be easily incorporated into various classical multivariate analyses when the proportions are strictly greater than zero. However, the ratio between the two proportions can be unstable and even ill-defined in the presence of zero counts and measurement error. Unfortunately, zero counts are prevalent in many compositional datasets, such as microbiome datasets. To overcome this problem, a popular practice is to replace zero counts with a small number, also called a pseudo-count. However, choosing this small number for a given dataset remains a challenge. In addition, as shown in Brill et al. (2020), mishandling zero counts can lead to inflated false discoveries.

3. Robust differential abundance test

3.1. *Infinite sample size*

The discussion in the previous section raises the question of whether there is a method of testing for differential components that is robust to prevalent zero counts in compositional data. To answer this question, we introduce a new robust framework for the differential abundance test in compositional data, which examines only the absolute difference between two proportions. Without loss of generality, we always write the largest reference set as I_0; this means that $H_{i,1}$ in (5), which is based on I_0, is true as long as $i \notin I_0$, and the set of differential components is $I_1 = [d] \setminus I_0$ where $[d] = \{1, \ldots, d\}$. Based on this definition, I_0 is essentially the set of all nondifferential components, so we can use these two concepts interchangeably.

Compared with conventional hypothesis testing, the main difficulty in testing a reference-based hypothesis is that the null hypothesis of the ith component cannot be identified based only on data at the ith component, as (5) is defined based on the relative relationship with respect to components in I_0. In this section we show that testing such hypotheses can be achieved by a series of directional comparisons on renormalized proportions. To illustrate the idea, we first consider the ideal case where the number of observed samples is infinite, that is, $Q_{1,i}$ and $Q_{2,i}$ are known, and the sum of the components between the two conditions is strictly greater than zero, $Q_{1,i} + Q_{2,i} > 0$ for $i \in [d]$.

According to (4), a common feature of all nondifferential components, i.e., those in I_0, is that $R_i = Q_{1,i} - Q_{2,i}$ has the same sign for $i \in I_0$. More concretely, for all $i \in I_0$ we have $R_i > 0$ if $b > 1$, $R_i < 0$ if $b < 1$ and $R_i = 0$ if $b = 1$. This observation motivates us to use all components to infer the sign of $b - 1$, and then identify the differential components by comparing the signs of R_i and $b - 1$. In particular, the assumption $|I_0| > d/2$ suggests that the sign of the median of all the R_i, denoted by $M\{R_i : 1, \ldots, d\}$, always equals the sign of R_i for $i \in I_0$. Therefore, $b - 1$ has the same sign as $M\{R_i : 1, \ldots, d\}$. As all components with a different sign of $b - 1$ are differential components, we can identify differential components by comparing the signs of R_i and $M\{R_i : 1, \ldots, d\}$. Briefly, all components with $\text{sign}(R_i) \neq \text{sign}(M\{R_i : 1, \ldots, d\})$ are differential components. The differential components identified by the above strategy are all correct ones, but many other differential components may be ignored. For example, if $b > 1$, the above strategy cannot identify components with $Q_{1,i} > bQ_{2,i}$ even if they are also differential components based on the definition in (5).

To overcome this problem, we generalize the above idea by repeatedly applying the strategy to renormalized proportions. To be specific, let $U_{(t)}$ be the set of identified differential component

candidates and let $V_{(t)}$ be the rest of the components at each iteration t. We set $U_{(0)} = \emptyset$ and $V_{(0)} = [d]$ to initialize the iterative procedure. Instead of assessing R_i directly, we opt to examine the absolute difference after normalization with respect to a set $I \subset [d]$:

$$R_i(I) = \frac{Q_{1,i}}{\sum_{i\in I} Q_{1,i}} - \frac{Q_{2,i}}{\sum_{i\in I} Q_{2,i}}, \quad i \in I.$$

Defining $b(I) = b \times (\sum_{i\in I} Q_{2,i}/\sum_{i\in I} Q_{1,i})$, we know that $b(I)$ is the ratio of $Q_{1,i}/\sum_{i\in I} Q_{1,i}$ and $Q_{2,i}/\sum_{i\in I} Q_{2,i}$ for all $i \in I_0$, and the median of $R_i(I)$ has the same sign as $b(I) - 1$ when $I_0 \subset I$. Here, we define the median of $R_i(I)$ as

$$M(I) = \text{Median}\{R_i(I) : i \in I\}.$$

After determining the sign of $M(I)$, we can define the sets of differential components $W^+(I) = \{i \in I : R_i(I) > 0\}$, $W^-(I) = \{i \in I : R_i(I) < 0\}$ and $W^0(I) = \{i \in I : R_i(I) = 0\}$.

Using this notation, we now define an iterative procedure for identifying the differential components. For any $t = 0, 1, \ldots$ we repeat the following steps.

(i) Find the median $M(V_{(t)})$ based on $R_i(V_{(t)})$ for $i \in V_{(t)}$.
(ii) Find all components with a different sign from $M(V_{(t)})$:

$$W_{(t)} = \begin{cases} W^+(V_{(t)}) \cup W^-(V_{(t)}), & M(V_{(t)}) = 0, \\ W^-(V_{(t)}) \cup W^0(V_{(t)}), & M(V_{(t)}) > 0, \\ W^+(V_{(t)}) \cup W^0(V_{(t)}), & M(V_{(t)}) < 0. \end{cases}$$

(iii) Let $U_{(t+1)} = U_{(t)} \cup W_{(t)}$ and $V_{(t+1)} = V_{(t)} \setminus W_{(t)}$. If $W_{(t)} = \emptyset$, stop the loop.

After the loop stops at $t = T$, we set $\hat{I}_0 = V_{(T)}$ and $\hat{I}_1 = U_{(T)}$; $\hat{I}_0$ is an estimation of the largest reference set I_0, and $\hat{I}_1$ estimates the set of differential components I_1. The following theorem shows that I_0 and I_1 can be recovered by $\hat{I}_0$ and $\hat{I}_1$ completely in at most $|I_1| + 1$ iterations.

THEOREM 1. *Consider estimating I_0 and I_1 by $\hat{I}_0$ and $\hat{I}_1$ defined above. If $|I_0| > d/2$, then*

$$T \leqslant |I_1| + 1, \quad \hat{I}_0 = I_0, \quad \hat{I}_1 = I_1.$$

This theorem says that one can identify the differential components accurately by a series of directional comparisons on renormalized proportions when the sample size is infinite. In contrast to existing methods, the proposed iterative procedure does not evaluate the ratio between two proportions; therefore we need not worry about the zero-count problem.

3.2. *Finite sample size*

The iterative procedure described in the previous subsection cannot be applied directly to compositional data, because in practice only finite samples are observed. As there is uncertainty in data, we redefine the testing procedure in the previous subsection to account for the randomness in data. Given a subset $I \subset [d]$, we consider a standard two-sample t-test statistic after normalization with respect to I,

$$\hat{R}_i(I) = \frac{\bar{P}_{1,i}(I) - \bar{P}_{2,i}(I)}{\{\hat{\sigma}^2_{1,i}(I)/m_1 + \hat{\sigma}^2_{2,i}(I)/m_2\}^{1/2}},$$

where $\bar{P}_{k,i}(I) = \bar{P}_{k,i} / \sum_{i \in I} \bar{P}_{k,i}$ with $\bar{P}_{k,i} = \sum_{j=1}^{m_k} \hat{P}_{k,j,i}/m_k$ and

$$\hat{\sigma}^2_{k,i}(I) = \frac{1}{m_k - 1} \sum_{j=1}^{m_k} \left(\frac{\hat{P}_{k,j,i}}{\sum_{i \in I} \bar{P}_{k,i}} - \bar{P}_{k,i}(I) \right)^2.$$

Similar to $M(I)$, we also define the median of $\hat{R}_i(I)$,

$$\hat{M}(I) = \text{Median}\{\hat{R}_i(I) : i \in I\}.$$

Suppose $\hat{M}(I)$ is close to zero in the sense that $-M < \hat{M}(I) < M$; then we can expect that $b(I)$ is close to 1 and that all components with $\hat{R}_i(I)$ far away from zero are differential components. This is equivalent to conducting two-sided hypothesis testing, and we define the set $W^{\pm}(I, D) = \{i \in I : |\hat{R}_i(I)| > D\}$ where D is a critical value that corrects for the multiple testing. In contrast, if $\hat{M}(I)$ is larger than M, we can expect $b(I)$ to be greater than 1, and the differential components can be identified by one-sided hypothesis testing, i.e., $W^{-}(I, D) = \{i \in I : \hat{R}_i(I) < -D\}$. Similarly, when $\hat{M}(I) < -M$, we collect the differential components in the set $W^{+}(I, D) = \{i \in I : \hat{R}_i(I) > D\}$.

As in the previous subsection, we define $U_{(t)}$ to be the set of selected differential components and $V_{(t)}$ to be the rest of the components at each iteration t, which are initially chosen as $U_{(0)} = \emptyset$ and $V_{(0)} = [d]$. Given a threshold M and a series of critical values $(D^{\pm}_{(t)}, D^{+}_{(t)}, D^{-}_{(t)})$ for $t = 0, 1, \ldots,$ we update $U_{(t)}$ and $V_{(t)}$ by the following procedure.

(i) Find the median $\hat{M}(V_{(t)})$ based on $\hat{R}_i(V_{(t)})$ for $i \in V_{(t)}$.
(ii) Set

$$W_{(t)} = \begin{cases} W^{\pm}(V_{(t)}, D^{\pm}_{(t)}), & -M < \hat{M}(V_{(t)}) < M, \\ W^{-}(V_{(t)}, D^{-}_{(t)}), & \hat{M}(V_{(t)}) \geqslant M, \\ W^{+}(V_{(t)}, D^{+}_{(t)}), & \hat{M}(V_{(t)}) \leqslant -M. \end{cases}$$

(iii) Let $U_{(t+1)} = U_{(t)} \cup W_{(t)}$ and $V_{(t+1)} = V_{(t)} \setminus W_{(t)}$. If $W_{(t)} = \emptyset$, stop the loop.

After the loop stops at $t = T$, we set $\hat{I}_0 = V_{(T)}$ and $\hat{I}_1 = U_{(T)}$. To implement this procedure, we still need to choose a threshold M and a sequence of positive critical values $(D^{\pm}_{(t)}, D^{+}_{(t)}, D^{-}_{(t)})$ for $t = 0, 1, \ldots$ to control false discovery.

3.3. *Familywise error rate control*

Unlike the conventional false discovery setting, there are several unique challenges here. First, the test statistics $\hat{R}_i(I)$ are generally dependent for different $i \in I$ because of the compositional constraint after renormalization. Moreover, in certain applications there may be some dependency structure between components, such as in microbiome data (Hawinkel et al., 2019). Second, the two-step procedure suggests that we need to consider the potential error in the first step when choosing the critical values for the second step. Third, as an iterative procedure, the possible dependence after renormalization plays an important role in false discovery control.

To choose the threshold M for the median, we need to consider the potential dependency structure between the $\hat{R}_i(I)$. In particular, let $P_i = \{\hat{P}_{k,j,i} : k = 1, 2; j = 1, \ldots, m_k\}$ and let S_l be

a subset of $[0,1]^{m_1+m_2}$ for all $1 \leqslant l \leqslant |I_0|$. For any subset S_l we assume that the P_i with $i \in I_0$ can be ordered as $P_{(1)}, \ldots, P_{(|I_0|)}$ to satisfy

$$\sum_{l=1}^{|I_0|} \left\| E\{(B_l - \mu_l)^2 \mid B_1^{l-1}\} \right\|_\infty + 2 \sum_{l=1}^{|I_0|} \sum_{l'=1}^{l-1} \left\| E(B_l - \mu_l \mid B_1^{l'}) \right\|_\infty \leqslant C_p \sum_{l=1}^{|I_0|} \mu_l, \tag{7}$$

where $C_p > 0$ is some constant, $B_l = I(P_{(l)} \in S_l)$, $\mu_l = E(B_l)$ and $B_1^l = \{B_1, \ldots, B_l\}$. We can set $C_p = 1$ when the P_i are independent of each other. Given condition (7), we choose the median's threshold to be $M = (C_p \log d/d)^{1/2}$. In doing so, we can ensure that $\mathrm{pr}\{\hat{M}(I) < -M, b(I) \geqslant 1\} \to 0$ and $\mathrm{pr}\{\hat{M}(I) > M, b(I) \leqslant 1\} \to 0$.

When choosing the critical values $(D^{\pm}_{(t)}, D^{+}_{(t)}, D^{-}_{(t)})$, it is necessary to consider renormalization at each iteration. Specifically, let q_α be the upper α-quantile of the random variable $\max_{i \in I_0}\{\sup_{r_i} r_i z_{1,i} + (1 - r_i^2)^{1/2} z_{2,i}\} = \max_{i \in I_0}(z_{1,i}^2 + z_{2,i}^2)^{1/2}$, where $z_{1,i}$ and $z_{2,i}$ are independent standard Gaussian random variables. For simplicity, we can choose an upper bound for q_α in practice, $\tilde{q}_\alpha = (2 \log d - 2 \log \alpha)^{1/2}$. In one-sided testing, we choose $D^{+}_{(t)} = D^{-}_{(t)} = q_\alpha$. When choosing a critical value in two-sided testing, it is also necessary to consider the effect of potential error in the first step. In particular, we choose the critical value $D^{\pm}_{(t)}$ as

$$D^{\pm}_{(t)} = q_\alpha + r_Q M,$$

where $r_Q = 8 C_r U_f / L_{1/4}$ with U_f, $L_{1/4}$ and C_r defined in the Supplementary Material. When $r_Q {C_p}^{1/2} \ll d^{1/2}$, we have $D^{\pm}_{(t)} = q_\alpha\{1 + o(1)\}$. Based on the above thresholds and critical values, it can be shown that the proposed procedure can asymptotically control the familywise error rate at the α level.

THEOREM 2. *Consider estimating I_0 and I_1 by $\hat{I}_0$ and $\hat{I}_1$ defined in § 3.2. If (7) is satisfied and Assumptions 1–6 in the Supplementary Material hold, then*

$$\mathrm{pr}(\hat{I}_1 \cap I_0 \neq \emptyset) \leqslant \alpha + o(1).$$

Theorem 2 says that the familywise error rate can be controlled at the α level asymptotically by the proposed method. In § 4.2 it is shown that the critical values $(D^{\pm}_{(t)}, D^{+}_{(t)}, D^{-}_{(t)})$ can also be chosen to control the false discovery rate.

In the proposed test we need to specify three parameters: the threshold M for the median, the critical value T for each component, and the effect size of the first step, r_Q, if we set $D^{\pm}_{(t)} = T + r_Q M$ and $D^{+}_{(t)} = D^{-}_{(t)} = T$. The parameters M and r_Q mainly control the decision of the testing direction in each iteration and its effect. Although larger M and r_Q will increase the robustness of the test to testing the wrong direction in the correlated case, they lead to lower power in the second step. The choices of $M = (2 \log d/d)^{1/2}$ and $r_Q = 0.2$ are supported by our experience and used in all numerical experiments in this paper. The simulation results suggest that the performance of the robust differential abundance test is not very sensitive to the choices of M and r_Q, even in the correlated case. However, we need to carefully choose the critical value T, as it plays an important role in the test. The simulation experiments suggest that the theoretical choice $T = (2 \log d - 2 \log \alpha)^{1/2}$ works well; however, as illustrated in § 4.2, T can also be chosen automatically when we aim to control the false discovery rate.

3.4. *Power analysis*

To conduct a power analysis, we consider a three-component mixture model. Specifically, we assume that the absolute abundances of each component A_i are independent of each other and we split $\{1, \ldots, d\}$ into three sets, I_0, I_1^+ and I_1^-. If the component is nondifferential, i.e., $i \in I_0$, we assume that A_i has the same distribution under π_1 and π_2. In contrast, if $i \in I_1^+$, we assume that A_i under π_2 has the same distribution as $A_i + \delta_+$ under π_1. Similarly, if $i \in I_1^-$, then A_i under π_2 has the same distribution as $A_i - \delta_-$ under π_1. Here, δ_+ and δ_- are two constants. Under this model, $I_1 = I_1^- \cup I_1^+$ is the set of differential components. We assume that the sample size in the two groups is the same, $m_1 = m_2 = m$. We also assume the variance of $\hat{P}_{k,j,i}$ to be the same for $k = 1, 2$, denoted by σ_i^2. We now characterize the sufficient conditions for reliably identifying differential components.

THEOREM 3. *Suppose that δ_+ and δ_- satisfy*

$$E_{\pi_1}\left(\frac{\delta_+ - A_i/|I_1^+|\delta_+/\sum_{i\in[d]}A_i}{\sum_{i\in[d]}A_i + \delta^*}\right) \geqslant (2+\epsilon)\sigma_i\left(\frac{2\log d}{m}\right)^{1/2}, \quad i \in I_1^+$$

$$E_{\pi_1}\left(\frac{A_i/|I_1^-|\delta_-/\sum_{i\in[d]}A_i - \delta_-}{\sum_{i\in[d]}A_i + \delta^*}\right) \geqslant (2+\epsilon)\sigma_i\left(\frac{2\log d}{m}\right)^{1/2}, \quad i \in I_1^-,$$

where $\delta^* = |I_1^+|\delta_+ - |I_1^-|\delta_-$ and ϵ is some small constant. If Assumptions 5 and 6 in the Supplementary Material hold and $r_Q C_p{}^{1/2} \ll d^{1/2}$, then

$$\mathrm{pr}(I_1 \subset \hat{I}_1) \to 1.$$

Theorem 3 suggests that if the difference between two groups is sufficiently large, i.e., if the signal-to-noise ratio is of rate $(\log d/m)^{1/2}$, the proposed method can identify differential components consistently.

4. EXTENSIONS OF THE ROBUST DIFFERENTIAL ABUNDANCE TEST

4.1. *Covariate balancing*

Besides compositional data from the two populations, in observational studies one often observes several covariates for each sample. Specifically, the observed data are $(\hat{P}_{k,1}, X_{k,1}), \ldots, (\hat{P}_{k,m_k}, X_{k,m_k})$ for $k = 1, 2$, where $X_{k,j} \in \mathbb{R}^d$ are the observed covariates. Imbalance between the distributions of $X_{1,j}$ and $X_{2,j}$ can give rise to bias when covariates are related to both the treatment assignment and the outcome of interest (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015). To remove the potential confounding effect, we use covariate-balancing techniques. As the proposed robust differential abundance test is composed of a series of two-sample t tests, we adopt one of the most widely used covariate-balancing methods, namely a weighting method, to reduce the bias introduced by the confounding effects (Rosenbaum, 1987; Robins et al., 2000; Chan et al., 2016; Yu & Wang, 2021). Specifically, we use a weighting method to estimate weights $w_{k,j}$ for each sample from the observed covariates $X_{k,j}$ for $k = 1, 2$ and $j = 1, \ldots, m_k$. Here, we assume $\sum_{j=1}^{m_k} w_{k,j} = 1$ for $k = 1, 2$. Instead of a standard two-sample t test, we employ a weighted version of the two-sample t test for a subset $I \subset [d]$,

$$\hat{R}_{i,w}(I) = \frac{\bar{P}_{1,i,w}(I) - \bar{P}_{2,i,w}(I)}{\{\hat{V}_{i,w}(I)\}^{1/2}}$$

where $\bar{P}_{k,i,w}(I) = \bar{P}_{k,i,w}/(\sum_{i\in I}\bar{P}_{k,i,w})$ with $\bar{P}_{k,i,w} = \sum_{j=1}^{m_k} w_{k,j}\hat{P}_{k,j,i}$ and $\hat{V}_{i,w}(I)$ is an estimator for the variance of $\bar{P}_{1,i,w}(I) - \bar{P}_{2,i,w}(I)$. The statistic $\hat{R}_{i,w}(I)$ can then replace $\hat{R}_i(I)$ in the method described in § 3.2. We can use any weighting method as long as we can estimate the variance. In the numerical experiments in the next subsection, we consider the empirical balancing calibration weighting method proposed by Chan et al. (2016) and implemented in the R package ATE (R Development Core Team, 2023), as it provides an estimator of the variance.

4.2. *False discovery rate control*

The choices of $D^{+}_{(t)}$, $D^{-}_{(t)}$ and $D^{\pm}_{(t)}$ in § 3.3 are designed for controlling the familywise error rate in a correlated case. However, one may also want to control the false discovery rate when aiming for high power in practice. We now provide, without proof, a Benjamini–Hochberg-like procedure for our iterative method to control the false discovery rate (Benjamini & Hochberg, 1995). Specifically, if we choose $D^{+}_{(t)} = D^{-}_{(t)} = T$ and $D^{\pm}_{(t)} = T + r_Q M$ for some constant T, a plug-in estimator for the upper bound on the false discovery proportion is

$$\widehat{\mathrm{FDP}}(T) \leqslant \frac{|I_0|\,\mathrm{pr}(Z > T)}{\max\{|\hat{I}_1(T)|, 1\}},$$

where $|\hat{I}_1(T)|$ is the total number of null hypotheses rejected by the iterative method and $Z = (z_1^2 + z_2^2)^{1/2}$, with z_1 and z_2 being independent standard Gaussian random variables. This estimator naturally leads to the following procedure for choosing the threshold:

$$\hat{T} = \inf\left[0 \leqslant T \leqslant q_\alpha : \frac{d\,\mathrm{pr}(Z > T)}{\max\{|\hat{I}_1(T)|, 1\}} \leqslant \alpha\right].$$

After choosing $\hat{T}$, we can set $D^{+}_{(t)} = D^{-}_{(t)} = \hat{T}$ and $D^{\pm}_{(t)} = \hat{T} + r_Q M$ in the iterative method. Similar to the Benjamini–Hochberg procedure, the guarantee of this procedure may need different components to have independence or some special correlation structure, as a plug-in estimator is used to estimate the false discovery proportion. In practice, when the number of iterations is small, Z can also be chosen as $|z|$ where z is a standard Gaussian random variable.

4.3. *Continuous outcome*

Although the discussions in § 2 and § 3 focused on two-sample comparison, the idea of this iterative method can also be applied to testing the association between compositional data and a continuous outcome. More concretely, suppose we observe $(\hat{P}_1, Y_1), \ldots, (\hat{P}_m, Y_m)$, where $\hat{P}$ denotes the compositional data and Y is a continuous variable of interest. If we are interested in testing the linear association between the compositional data and outcome, we can replace the two-sample t test with a correlation test

$$\hat{R}_i(I) = r_i(I)\left\{\frac{m-2}{1-r_i^2(I)}\right\}^{1/2},$$

where $r_i(I)$ is the Pearson correlation coefficient between the renormalized compositional data $\hat{P}_{j,i}/(\sum_{i\in I}\hat{P}_{j,i})$ and the outcome Y_j. The theoretical analysis for testing linear association is similar to that in the two-sample comparison case. However, it remains unclear whether this framework can be extended to testing nonlinear association between compositional data and a continuous outcome.

Table 1. *Comparison of differential abundance tests when components are correlated*

		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
		FWER	Power	FWER	Power	FWER	Power
S1	RDB	0.02 (0.01)	0.57 (0.01)	0.01 (0.01)	0.59 (0.02)	0.01 (0.01)	0.58 (0.01)
	ANCOM.BC	0.63 (0.05)	0.72 (0.01)	0.60 (0.05)	0.73 (0.01)	0.63 (0.05)	0.71 (0.01)
	ANCOM	0.17 (0.04)	0.61 (0.01)	0.19 (0.04)	0.63 (0.01)	0.22 (0.04)	0.60 (0.01)
	DESeq2	0.98 (0.01)	0.73 (0.01)	0.95 (0.02)	0.74 (0.01)	0.92 (0.03)	0.72 (0.01)
	Wilcoxon	0.65 (0.05)	0.61 (0.01)	0.58 (0.05)	0.63 (0.01)	0.61 (0.05)	0.60 (0.01)
	Wilcoxon.TSS	0.70 (0.05)	0.63 (0.01)	0.70 (0.05)	0.65 (0.01)	0.66 (0.05)	0.63 (0.01)
	DACOMP	0.41 (0.05)	0.67 (0.01)	0.45 (0.05)	0.69 (0.01)	0.38 (0.05)	0.66 (0.01)
S2	RDB	0.01 (0.01)	0.29 (0.01)	0.01 (0.01)	0.30 (0.01)	0.00 (0.00)	0.32 (0.01)
	ANCOM.BC	0.37 (0.05)	0.43 (0.01)	0.43 (0.05)	0.44 (0.01)	0.46 (0.05)	0.44 (0.01)
	ANCOM	0.07 (0.03)	0.33 (0.01)	0.08 (0.03)	0.32 (0.01)	0.07 (0.03)	0.34 (0.01)
	DESeq2	0.75 (0.04)	0.46 (0.01)	0.73 (0.04)	0.46 (0.01)	0.74 (0.04)	0.46 (0.01)
	Wilcoxon	0.29 (0.05)	0.41 (0.01)	0.39 (0.05)	0.42 (0.01)	0.39 (0.05)	0.44 (0.01)
	Wilcoxon.TSS	0.43 (0.05)	0.43 (0.01)	0.44 (0.05)	0.44 (0.01)	0.47 (0.05)	0.47 (0.01)
	DACOMP	0.21 (0.04)	0.37 (0.01)	0.29 (0.05)	0.38 (0.01)	0.28 (0.05)	0.39 (0.01)

RDB, the proposed robust differential abundance test; ANCOM.BC, analysis of composition of microbiomes with bias correction (Lin & Peddada, 2020); ANCOM, analysis of composition of microbiomes (Mandal et al., 2015); DESeq2, differential expression analysis for sequence count data 2 (Love et al., 2014); Wilcoxon, the Wilcoxon rank-sum test without normalization; Wilcoxon.TSS, the Wilcoxon rank-sum test with total-sum scaling normalization; DACOMP, differential abundance testing with compositionality adjustment (Brill et al., 2020); FWER, familywise error rate.

5. Numerical experiments

5.1. *Simulation studies*

In simulation experiments we compare the proposed robust differential abundance test with six commonly used differential abundance tests. The tests we consider are the following: analysis of composition of microbiomes (Mandal et al., 2015), analysis of composition of microbiomes with bias correction (Lin & Peddada, 2020), differential abundance testing with compositionality adjustment (Brill et al., 2020), differential expression analysis for sequence count data 2 (Love et al., 2014), and the Wilcoxon rank-sum test without normalization and with total-sum scaling normalization. The simulation experiments consider four different ways to simulate observed data: Poisson–gamma distributions, log-normal distributions, log-normal distributions with covariates, and real-data shuffling. The Supplementary Material presents the detailed settings of the simulation experiments.

We first simulate the data from Poisson–gamma distributions and compare the seven methods as the number of components d, the number of differential components s and the sample size m vary. We aim to control the familywise error rate at the 10% level. The results are summarized in the Supplementary Material and suggest that the tests of Mandal et al. (2015), Lin & Peddada (2020) and Brill et al. (2020), and the proposed test can control the familywise error rate relatively well, while the other three tests suffer from inflated familywise error rate under our simulation settings. An interesting observation is that the familywise error rate in our test, and in the tests of Mandal et al. (2015) and Lin & Peddada (2020), is slightly inflated when the sample size is small; in constrast, the familywise error rate of the other methods is inflated as the sample size increases. A similar phenomenon is observed and explained in Lin & Peddada (2020). To compare performances on correlated data, we next simulate the data from log-normal distributions with a correlated covariance matrix and summarize the results in Table 1. From Table 1 we see that the proposed test can control the familywise error rate very well, while the other tests inflate

Table 2. *Comparison of computation time (in seconds) of differential abundance tests*

	$d = 100$	$d = 200$	$d = 500$
RDB	0.0041	0.0081	0.0225
ANCOM.BC	0.7156	1.2223	2.8303
ANCOM	15.6485	62.5443	400.1344
DESeq2	1.3138	1.3924	1.6444
Wilcoxon	0.0282	0.0533	0.1258
Wilcoxon.TSS	0.0280	0.0552	0.1386
DACOMP	0.4300	1.6985	8.3371

the familywise error rate. We also compare the computation time of these seven differential abundance tests when drawing data from Poisson–gamma distributions. The computation times summarized in Table 2 suggest that the proposed test is computationally very efficient. To better understand our test, we also study its performance when the sequencing depth is unbalanced. The results, reported in the Supplementary Material, suggest that the familywise error rate of our test is slightly inflated in the unbalanced sequencing depth.

Next we investigate the performance of our test extensions. As in the previous experiments, we simulate the data from both Poisson–gamma and log-normal distributions and compare the seven different methods; however, we now aim to control the false discovery rate at the 10% level. The results are summarized in the Supplementary Material. Compared with familywise error rate control, false discovery rate control can boost power for all methods. The proposed method and the methods of Mandal et al. (2015), Lin & Peddada (2020) and Brill et al. (2020) can control the false discovery rate well in the Poisson–gamma case; only the proposed method can control the false discovery rate in the correlated case, but we remain cautious as the method does not have a theoretical guarantee of controlling the false discovery rate under arbitrary correlated settings. We further compare the performance of different methods when some confounding variables are observed. We simulate the data from log-normal distributions with covariates. Results reported in the Supplementary Material show that our method can control both the familywise error rate and the false discovery rate after covariate-balancing adjustment. Finally, we study the performance of the different methods when the outcome is continuous. We draw the data from a Poisson–gamma distribution with a continuous outcome. The results suggest that the proposed method can control the familywise error rate well in such Poisson–gamma distributions.

5.2. *Analysis of gut microbiome data*

To further demonstrate the proposed method's practical merits, we apply it to a gut microbiota dataset of 476 samples (Yatsunenko et al., 2012). The samples are divided into three groups based on which countries they come from: 83 samples from Malawi, 83 samples from Venezuela and 310 samples from the U.S.A. The dataset consists of 11 905 operational taxonomic units, and we aggregate the data at the genus level, which yields 649 different genera. After aggregation, we have 82.5% zero entries in the dataset. Besides the countries, each sample in this dataset has two observed covariates, gender and age. The goal is to compare the microbiome communities of different countries and identify differential genera. Before conducting the differential abundance analysis, we first investigate the degree of covariance balancing in this dataset. A violin plot of age is shown in Fig. 2(a), and the proportions of females in Malawi, the U.S.A. and Venezuela are 63.9%, 58.1% and 56.6%, respectively, suggesting that neither age nor gender is well balanced across countries; therefore, covariate adjustment is needed in the differential abundance analysis.

We apply the robust differential abundance test to each pair of countries to compare their microbiome communities. The proposed test identifies 37 differential genera between Malawi

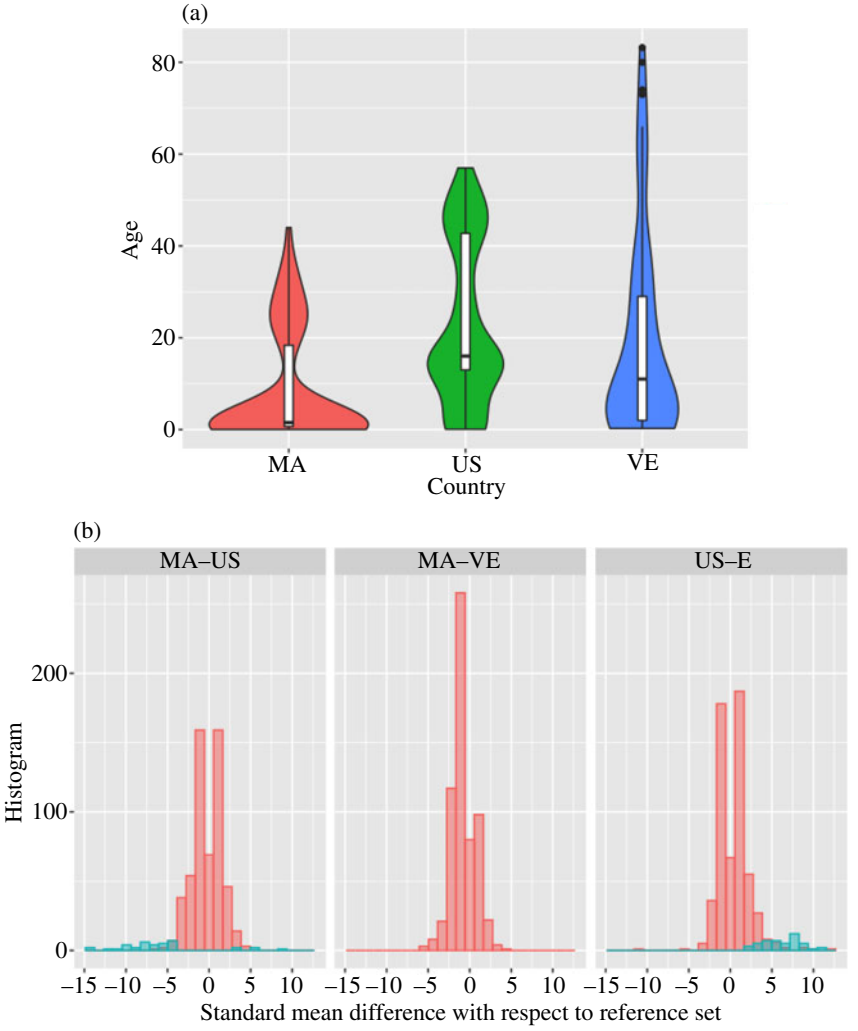

Fig. 2. (a) Violin plots of age for Malawi (MA), the U.S.A. (US) and Venezuela (VE); (b) histograms of the standard mean difference with respect to the reference set estimated by the proposed test for each pair of countries, where red represents false and blue represents true.

and the U.S.A., no differential genera between Malawi and Venezuela, and 47 between Venezuela and the U.S.A. The detailed genera are presented in the Supplementary Material. There are 28 common genera reported between Malawi versus the U.S.A. and Venezuela versus the U.S.A. In Fig. 2(b), all the estimated nondifferential genera are taken as the reference set in each pairwise comparison, and the standard mean differences with respect to this reference set are summarized. Figure 2(b) shows that differential and nondifferential genera are well separated by using this reference set as a benchmark. The results suggest that there is no significant difference between Malawi and Venezuela. All the differential genera between the U.S.A. and Venezuela have a greater abundance in the U.S.A. than in Venezuela. Most of the differential genera between the U.S.A. and Malawi have a greater abundance in the U.S.A. than in Malawi. Besides, there is large overlap between the differential genera in the U.S.A. versus Venezuela and those in the U.S.A. versus Malawi. Among the 28 common genera found by these two comparisons, 12 belong to order Clostridiales and phylum Firmicutes, which could be explained by a high-fat diet in the U.S.A. (Clarke et al., 2012). Therefore, the pairwise analysis results are consistent with each

other. We also apply the proposed test with false discovery rate control and the methods of Mandal et al. (2015) and Lin & Peddada (2020) to the same dataset. The results are summarized in the Supplementary Material and suggest some differences in the reported differential genera. From this comparison it can also be concluded that the robust differential abundance test is better able to identify the differential components from an angle of reference-based hypothesis.

6. CONCLUDING REMARKS

This work has focused on the robust differential abundance test with a standard two-sample t test for simplicity of presentation and theoretical analysis. However, as a general framework, the proposed test can also work with other popular two-sample location tests, such as the Mann–Whitney U test. Using more complicated tests in the robust differential abundance test could improve practical performance, but would also involve more theoretical analyses. In addition, we have mainly considered the use of weighting methods to adjust for the confounding effects of observed covariates. As mentioned in previous sections, it is possible to apply other covariate-balancing techniques. For example, the robust differential abundance test can work with another popular covariate-balancing technique, the matching method (Rosenbaum, 2002; Yu & Rosenbaum, 2019). Application of the test to microbiome data usually requires choosing a suitable taxonomic rank for microbes. Our test can also be performed with multi-scale adaptive techniques to detect differentially abundant microbes at a high resolution (Wang, 2021). Finally, it may be of interest to explore whether the proposed framework can test for other characteristics of abundances, such as dispersion and skewness (Martin et al., 2020).

ACKNOWLEDGEMENT

We thank the editor, associate editor and referees for valuable suggestions. This project was supported by the U.S. National Science Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material includes proofs of all the theorems, technical assumptions, and settings and additional results of the numerical experiments.

REFERENCES

AITCHISON, J. (1983). Principal component analysis of compositional data. *Biometrika* **70**, 57–65.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

BRILL, B., AMIR, A. & HELLER, R. (2020). Testing for differential abundance in compositional counts data, with application to microbiome studies. *arXiv:* 1904.08937v5.

BUTLER, A., HOFFMAN, P., SMIBERT, P., PAPALEXI, E. & SATIJA, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotech.* **36**, 411–20.

CAO, Y., ZHANG, A. & LI, H. (2020). Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika* **107**, 75–92.

CHAN, K. C. G., YAM, S. C. P. & ZHANG, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc.* B **78**, 673–700.

CLARKE, S. F., MURPHY, E. F., NILAWEERA, K., ROSS, P. R., SHANAHAN, F., O'TOOLE, P. W. & COTTER, P. D. (2012). The gut microbiota and its relationship to diet and obesity: New insights. *Gut Microbes* **3**, 186–202.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Am. Statist. Assoc.* **99**, 96–104.

EFRON, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R. & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, article no. 15.

Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. (2019). A broken promise: Microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinformatics* **20**, 210–21.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc.* B **76**, 243–63.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Kharchenko, P. V., Silberstein, L. & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Meth.* **11**, 740–2.

Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29.

Lê Cao, K., Costello, M., Lakis, V., Bartolo, F., Chua, X., Brazeilles, R. & Rondeau, P. (2016). MixMC: A multivariate statistical framework to gain insight into microbial communities. *PLoS One* **11**, e0160169.

Lin, H. & Peddada, S. (2020). Analysis of compositions of microbiomes with bias correction. *Nature Commun.* **11**, 1–11.

Love, M. I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, article no. 550.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. & Peddada, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial Ecol. Health Dis.* **26**, article no. 27663.

Martin, B. D., Witten, D. & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Statist.* **14**, 94–115.

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., Zengler, K. & Knight, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nature Commun.* **10**, 1–11.

Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Meth.* **10**, 1200–2.

Pawlowsky-Glahn, V. & Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. Chichester: John Wiley & Sons.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Commun.* **9**, 1–17.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. 2nd Berkeley Symp. Mathematical Statistics and Probability*. Berkeley, California: University of California Press, pp. 131–49.

Robins, J. M., Hernan, M. A. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–60.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *J. Am. Statist. Assoc.* **82**, 387–94.

Rosenbaum, P. R. (2002). *Design of Observational Studies*. New York: Springer.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., De Commer, L., Darzi, Y. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–11.

Wang, S. (2021). Multi-scale adaptive differential abundance analysis in microbial compositional data. *bioRxiv:* 10.1101/2021.11.02.466987.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A. et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, article no. 27.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P. et al. (2012). Human gut microbiome viewed across age and geography. *Nature* **486**, 222–7.

Yu, R. & Rosenbaum, P. R. (2019). Directional penalties for optimal matching in observational studies. *Biometrics* **75**, 1380–90.

Yu, R. & Wang, S. (2021). Treatment effects estimation by uniform transformer. *arXiv:* 2008.03738v2.

[*Received on* 10 *February* 2021. *Editorial decision on* 11 *April* 2022]