Zhang and Dao *BMC Bioinformatics* 2020, **21**(Suppl 9):205
https://doi.org/10.1186/s12859-020-3530-x

BMC Bioinformatics

METHODOLOGY Open Access

A distance based multisample test for high-dimensional compositional data with applications to the human microbiome

Qingyang Zhang* and Thy Dao

From The 20th International Conference on Bioinformatics & Computational Biology (BIOCOMP 2019)
Las Vegas, NV, USA. 29 July–01 August 2019

*Correspondence: qz008@uark.edu
Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA

Abstract

Background: Compositional data refer to the data that lie on a simplex, which are common in many scientific domains such as genomics, geology and economics. As the components in a composition must sum to one, traditional tests based on unconstrained data become inappropriate, and new statistical methods are needed to analyze this special type of data.

Results: In this paper, we consider a general problem of testing for the compositional difference between K populations. Motivated by microbiome and metagenomics studies, where the data are often over-dispersed and high-dimensional, we formulate a well-posed hypothesis from a Bayesian point of view and suggest a nonparametric test based on inter-point distance to evaluate statistical significance. Unlike most existing tests for compositional data, our method does not rely on any data transformation, sparsity assumption or regularity conditions on the covariance matrix, but directly analyzes the compositions. Simulated data and two real data sets on the human microbiome are used to illustrate the promise of our method.

Conclusions: Our simulation studies and real data applications demonstrate that the proposed test is more sensitive to the compositional difference than the mean-based method, especially when the data are over-dispersed or zero-inflated. The proposed test is easy to implement and computationally efficient, facilitating its application to large-scale datasets.

Keywords: Microbiome, Compositional data, High dimensionality, Centered log-ratio transformation, Multisample test, Distance correlation

Background

Data that lie on the simplex $\mathcal{S}^{d-1} = \left\{(x_1, x_2, ..., x_d), s.t. \min_j x_j \geq 0, \sum_{j=1}^{d} x_j = 1\right\}$ are often called $(d-1)$-dimensional compositional data, and they arise in many scientific disciplines such as genomics, geology and economics [1–3]. As the components in a composition must sum to one, classic statistical tests including two-sample t-test and

Wilcoxon rank-sum test become inappropriate as they require unconstrained data, and directly applying these standard methods to compositional data could result in misleading inference [4]. To overcome this difficulty, Aitchison (1982) proposed to use a log-ratio transformation to relax the unit-sum constraint, so that some classic tests can be applied to the transformed data. For instance, the generalized likelihood ratio test based on log-ratios [4] has been widely used to test compositional difference between groups due to its simplicity and good empirical performance.

It is noteworthy that Aitchison's test only applies to low-dimensional settings where the dimension is less than sample size. In recent microbiome and metagenomic studies, however, the compositional data are often high-dimensional. For instance, in the Human Microbiome Project, it is common to have hundreds to thousands of bacterial taxa while only tens of samples are available for analysis. To this end, Cao et al. (2017) developed a powerful two-sample test for high-dimensional means using a centered log-ratio transformation [3]. Cao et al.'s test achieves satisfactory statistical power under high-dimensional sparse settings, and the consistency of the test has been well established under some regularity conditions. Nevertheless, this test has several shortcomings which has limited its application in practice. For instance, Cao et al.'s test can only deal with two-sample comparison, and its validity depends on a list of regularity conditions on the underlying covariance matrices. In addition, this test is a maximum-type test, and its performance relies on the sparsity assumption, i.e., only a small proportion of components in the composition are different across groups.

In this paper, we formulated a new hypothesis from a Bayesian point of view, to handle high-dimensionality and over-dispersion that are commonly seen in recent microbiome data. Different from those mean-based hypotheses, we assumed that the abundances follow a multinomial model with random composition parameters, and redefined the compositional equivalency using the distribution of random compositions. To directly target the distributional difference in composition, we suggested a distance based non-parametric test for detecting the difference between multiple groups. Unlike most existing tests for compositional data, our method does not rely on any data transformation, sparsity assumption or regularity conditions on the covariance matrix, but directly analyzes the compositions. Simulation studies demonstrated that our test is more sensitive to the compositional difference than the mean-based method, especially when the data are over-dispersed or zero-inflated. The proposed test is easy to implement and computationally efficient, facilitating its application to large-scale datasets.

The remainder of the paper is structured as follows: "Methods" section formulates the hypothesis testing and introduces our distance based method. "Results" section compares the proposed method with a log-ratio based test using simulated data from negative binomial models. In addition, we apply the new method to two real datasets on human throat microbiome and intestinal microbiome. "Discussion" section discusses our method with some future perspectives. "Conclusions" section concludes the paper.

Methods

Problem formulation

In this part, we briefly reviewed the test by Cao et al. (2017), and then formulated our new hypothesis. We begin with the notations. Let $k \in \{1, 2, ..., K\}$ be the group index and $j \in \{1, 2, ..., p\}$ be the index of components in the composition. Denote

by $\boldsymbol{X}^{(k)} = \left(\boldsymbol{X}_1^{(k)}, ..., \boldsymbol{X}_{n_k}^{(k)}\right)^T$ the observed $n_k \times p$ data matrix for group k, where $\boldsymbol{X}_i^{(k)} = \left(X_{i1}^{(k)}, ..., X_{ip}^{(k)}\right)^T$ represents the composition for subject i that lie on the $(p-1)$-dimensional simplex. We assume that the observed compositional data $\boldsymbol{X}^{(k)}$ arise from a latent matrix $\boldsymbol{W}^{(k)} = \left(\boldsymbol{W}_1^{(k)}, ..., \boldsymbol{W}_{n_k}^{(k)}\right)^T$ $\left(\boldsymbol{W}_1^{(k)}, ..., \boldsymbol{W}_{n_k}^{(k)}\right.$ are iid samples$\left.\right)$ by normalization

$$X_{ij}^{(k)} = \frac{W_{ij}^{(k)}}{\sum_{h=1}^{p} W_{ih}^{(k)}},$$

where the unobserved $\boldsymbol{W}^{(k)}$ may refer to the true abundance of bacterial taxa for microbiome data.

As the true abundances $\boldsymbol{W}^{(k)}$ are generally unknown, Cao et al. (2017) formulated a testable hypothesis for two groups:

$$H_0 : E\left(\log\left(\boldsymbol{W}_1^{(1)}\right)\right) = E\left(\log\left(\boldsymbol{W}_1^{(2)}\right)\right) + c\mathbf{1}_p, \text{for some } c \in \mathbb{R},$$
$$H_\alpha : E\left(\log\left(\boldsymbol{W}_1^{(1)}\right)\right) \neq E\left(\log\left(\boldsymbol{W}_1^{(2)}\right)\right) + c\mathbf{1}_p, \text{for any } c \in \mathbb{R},$$

where $\mathbf{1}_p$ stands for the vector of p ones. The hypothesis above is mean-based and it can be tested through a centered log-ratio transformation

$$Y_{ij}^{(k)} = \log \frac{X_{ij}^{(k)}}{\left(\Pi_{h=1}^{p} X_{ih}^{(k)}\right)^{1/p}}, k = 1, 2; i = 1, ..., n_k.$$

It can be shown that the centered log-ratio variables $Y_{ij}^{(k)}$'s are only weakly dependent and satisfy certain desired properties, and Cao et al. (2017) suggested the following test statistics based on these log-ratio variables

$$\mathcal{M}_n = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le j \le p} \frac{\left(\bar{Y}_j^{(1)} - \bar{Y}_j^{(2)}\right)^2}{\hat{\gamma}_{jj}},$$

where $\bar{Y}_j^{(k)} = \sum_{i=1}^{n_k} Y_{ij}^{(k)}/n_k$, $\hat{\gamma}_{jj} = \sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(Y_{ij}^{(k)} - \bar{Y}_j^{(k)}\right)^2/(n_1+n_2)$. The p-value can be then obtained through Gumbel distribution

$$p = 1 - \left\{\exp\exp\left(-\frac{1}{2}\mathcal{M}_n - 2\log p + \log\log p + \log\pi\right)\right\}^{-1}.$$

Cao et al.'s test targets the difference in high-dimensional means, and its validity relies on several assumptions on the underlying covariance matrices, which are impractical to check in reality. Here, we considered a different hypothesis based on the distribution of composition instead of means. Under multinomial model, we have

$$\boldsymbol{W}_i^{(k)} \sim \text{Multinomial}\left(N_i^{(k)}, \boldsymbol{\pi}_i^{(k)}\right),$$

where $N_i^{(k)}$ stands for the total abundance of bacterial taxa for sample i from group k, and $\boldsymbol{\pi}_i^{(k)}$ represents the true composition. In order to model over-dispersion, we assumed random parameters, $N_i^{(k)} \sim f_N(\alpha)$ and $\boldsymbol{\pi}_i^{(k)} = \left(\pi_{i1}^{(k)}, ..., \pi_{ip}^{(k)}\right) \sim f_\pi\left(\boldsymbol{\Theta}^{(k)}\right)$, where α and $\boldsymbol{\Theta}^{(k)}$ are hyper-parameters. We then define the compositional equivalence between two groups based on the distribution of parameter $\boldsymbol{\pi}$:

Definition 1 *Two groups k and k′ are said to be compositionally equivalent if* $f_\pi\left(\boldsymbol{\Theta}^{(k)}\right) = f_\pi\left(\boldsymbol{\Theta}^{(k')}\right)$.

By Definition 1, we formulate the null and alternative hypotheses for K groups

$$H_0 : f_\pi\left(\Theta^{(1)}\right) = ... = f_\pi\left(\Theta^{(K)}\right),$$
$$H_\alpha : f_\pi\left(\Theta^{(k)}\right) \neq f_\pi\left(\Theta^{(k')}\right) \text{ for some } k \text{ and } k'.$$

Throughout this paper, we assume that the total abundance or sequencing depth $N_i^{(k)}$ is independent of $\boldsymbol{\pi}_i^{(k)}$, and $N_i^{(k)} \sim f_N(\alpha)$ for $i \in \{1, ..., n_k\}$ and $k \in \{1, 2, ..., K\}$, therefore testing H_0 amounts to testing the distributional equality of the compositions between K groups. Let $f_{\boldsymbol{X}}^{(k)}(\boldsymbol{x})$ be the density function of $\boldsymbol{X}_i^{(k)}$, one can test the following equivalent hypothesis

$$H_0^* : f_{\boldsymbol{X}}^{(1)}(\boldsymbol{x}) = ... = f_{\boldsymbol{X}}^{(K)}(\boldsymbol{x}) \text{ for all } \boldsymbol{x},,$$
$$H_\alpha^* : f_{\boldsymbol{X}}^{(k)}(\boldsymbol{x}) \neq f_{\boldsymbol{X}}^{(k')}(\boldsymbol{x}) \text{ for some } \boldsymbol{x},\ k \text{ and } k',$$

Here, it is noteworthy that H_0^* is equivalent to the independence between the composition $\boldsymbol{X}$ and the grouping variable $k \in \{1, 2, ..., K\}$ (i.e., phenotype), which converts the problem to testing the independence between the continuous random vector and a categorical variable.

Distance based test

In this part, we proposed a distance based method to test H_0^*, i.e., to detect the association between composition and phenotype. To begin with, we briefly introduce the notion of distance covariance. The distance covariance between two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ (can be of different sizes and different types) is defined as the square root of

$$\mathrm{dCov}^2(\boldsymbol{X}, \boldsymbol{Y}) = \int_{R^{d_x+d_y}} \frac{\|\phi_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{t}, \boldsymbol{s}) - \phi_{\boldsymbol{x}}(\boldsymbol{t})\phi_{\boldsymbol{y}}(\boldsymbol{s})\|^2}{c_{d_x} c_{d_y} \|\boldsymbol{t}\|_{d_x}^{1+d_x} \|\boldsymbol{s}\|_{d_y}^{1+d_y}} d\boldsymbol{t} d\boldsymbol{s}, \tag{1}$$

where $\phi(\cdot)$ represents a characteristic function, d_x and d_y are the dimensions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, $c_{d_x} = \frac{\pi^{(1+d_x)/2}}{\Gamma\{(1+d_x)/2\}}$ and $c_{d_y} = \frac{\pi^{(1+d_y)/2}}{\Gamma\{(1+d_y)/2\}}$. Unless otherwise specified, $\|\boldsymbol{z}\|_{d_z}$ denotes the Euclidean norm of $\boldsymbol{z} \in \mathbb{R}^{d_z}$, and $\|\phi\|^2 = \phi\bar{\phi}$ for a complex-valued function ϕ and its conjugate $\bar{\phi}$.

One remarkable property of distance covariance is that $\mathrm{dCov}(\mathbf{X}, \mathbf{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically independent, indicating that the distance covariance can also capture nonlinear associations. In their seminal work, Szekely et al. (2007) also provided the following alternative definition of distance covariance based on Euclidean distance and established its equivalency to the original definition in Eq. (1) (see Theorem 1, [5]):

$$\mathrm{dCov}^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{Cov}(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|, \|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|) - 2\mathrm{Cov}(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|, \|\boldsymbol{Y}_1 - \boldsymbol{Y}_3\|),$$

where $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$, $(\boldsymbol{X}_2, \boldsymbol{Y}_2)$ and $(\boldsymbol{X}_3, \boldsymbol{Y}_3)$ be three independent copies of $(\boldsymbol{X}, \boldsymbol{Y})$. Here, we choose to use this alternative definition to derive the explicit formula of distance covariance between composition and phenotype. For ease of notations, let Y be the phenotype, taking values from a discrete set $\{1, 2, ..., K\}$ with probabilities $\{p_1, ..., p_K\}$, and $\boldsymbol{X} = \{X_1, ..., X_p\}$ be the composition. For illustration purpose, here we assume Y is nominal (without ordering between categories), however, our test can be easily extended to ordinal Y and the formula is given in the "Discussion" section. Let $(\boldsymbol{X}_1, Y_1)$, $(\boldsymbol{X}_2, Y_2)$ and $(\boldsymbol{X}_3, Y_3)$ be three independent copies of $(\boldsymbol{X}, Y)$, we define $\|Y_1 - Y_2\| = 1$, if $Y_1 \neq Y_2$ and

0 otherwise. In addition, we define expected inter-point distance as

$$\mathcal{D}_{ij} = E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| | Y_1 = i, Y_2 = j), i, j = 1, ..., K.$$

The distance covariance between Y and $\boldsymbol{X}$ can then be derived from the second definition

$$\begin{aligned}
E(\|Y_1 - Y_2\|) &= 1 - \sum_{i=1}^{K} p_i^2, \\
E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|) &= \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j \mathcal{D}_{ij}, \\
E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \|Y_1 - Y_2\|) &= \sum_{i\neq j} p_i p_j \mathcal{D}_{ij} = \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j \mathcal{D}_{ij} - \sum_{i=1}^{K} p_i^2 \mathcal{D}_{ii}, \\
E(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \|Y_1 - Y_3\|) &= \sum_{j=1}^{K}\sum_{i\neq l} p_i p_j p_l \mathcal{D}_{ij} = \sum_{i=1}^{K}\sum_{j=1}^{K} p_i(1-p_i) p_j \mathcal{D}_{ij}.
\end{aligned}$$

Summarizing the results above, we have

$$\mathrm{dCov}(\boldsymbol{X}, Y) = 2\sum_{i=1}^{K}\sum_{j=1}^{K} p_i^2 p_j \mathcal{D}_{ij} - \sum_{i=1}^{K} p_i^2 \mathcal{D}_{ii} - \left(\sum_{i=1}^{K} p_i^2\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j \mathcal{D}_{ij}\right).$$

By Cauchy-Schwarz inequality, it can be shown that $\mathrm{dCov}(\boldsymbol{X}, Y) \geq 0$ and the equality holds if and only if $\mathcal{D}_{ii} = \mathcal{D}_{jj} = \mathcal{D}_{ij}$ for all (i, j)'s. When $K = 2$, we have the following special case

$$\mathrm{dCov}(\boldsymbol{X}, Y) = 2p^2(1-p)^2(2\mathcal{D}_{12} - \mathcal{D}_{11} - \mathcal{D}_{22}).$$

The sample version of $\mathrm{dCov}(\boldsymbol{X}, Y)$ can be expressed as

$$\widehat{\mathrm{dCov}}(\boldsymbol{X}, Y) = 2\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{p}_i^2 \hat{p}_j \hat{\mathcal{D}}_{ij} - \sum_{i=1}^{K} \hat{p}_i^2 \hat{\mathcal{D}}_{ii} - \left(\sum_{i=1}^{K} \hat{p}_i^2\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{p}_i \hat{p}_j \hat{\mathcal{D}}_{ij}\right).$$

Let n_i be the sample size in group i, the maximum likelihood estimate of p_i is $\hat{p}_i = n_i/n$, and the sample inter-point distance can be computed as follows:

$$\hat{\mathcal{D}}_{ij} = \frac{1}{n_i n_j}\sum_{m=1}^{n_i}\sum_{l=1}^{n_j} \|\boldsymbol{X}_m^{(i)} - \boldsymbol{X}_l^{(j)}\|, \tag{2}$$

$$\hat{\mathcal{D}}_{ii} = \frac{2}{n_i(n_i-1)}\sum_{m=1}^{n_i}\sum_{l=1}^{n_i} \|\boldsymbol{X}_m^{(i)} - \boldsymbol{X}_l^{(i)}\|, \tag{3}$$

where $\left\{\boldsymbol{X}_1^{(i)}, ..., \boldsymbol{X}_{n_i}^{(i)}\right\}$ and $\left\{\boldsymbol{X}_1^{(j)}, ..., \boldsymbol{X}_{n_j}^{(j)}\right\}$ stand for samples of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, respectively.

As the distribution of sample distance covariance is impractical to evaluate [5], we suggest a simple permutation procedure to obtain p-values. In practice, one can randomly shuffle the vector of Y for M times, and calculate sample distance covariance between composition and the permuted Y, then the permutation p-value can be computed as the proportion of distance covariance from permuted data that exceed the observed one.

It is noteworthy that in addition to distance correlation, there are many other dependence measures that could be used in our framework, including the energy-divergence metric [6], multiscale graph correlation [7] and projection correlation [8], among others.

One may refer to Josse and Holmes (2014) [9] for a general review of existing dependence measures between random vectors, and Szekely and Rizzo (2013) [10] for a review of energy- and distance-based measures.

Results

Numerical study

We conduct three simulation studies to compare the distance based test and the log-ratio based test [3] in detecting the compositional differences between groups. In the first study, we focus on two-group comparison under various high-dimensional and over-dispersed models. The dimension is fixed at $p = 200$ and two different sample sizes $n_1 = n_2 = 50$ and $n_1 = n_2 = 100$ are used. The abundance $W_{ij}^{(k)}$ are generated from three different settings

- **Setting 1**: $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$, $i = 1, ..., n_i, j = 1, ..., p$, $r_j^{(1)} \sim \text{Unif}(0.1, 1)$, $r_j^{(2)} = r_j^{(1)}$, $\mu_j^{(1)} \sim \text{Unif}(10, 15)$. Let $\boldsymbol{I} = \{\boldsymbol{I}_+, \boldsymbol{I}_-\}$ be the set of taxa with different abundances in two conditions, $\mu_j^{(2)} = \mu_j^{(1)} + \Delta$ for $j \in \boldsymbol{I}_+$ and $\mu_j^{(2)} = \mu_j^{(1)} - \Delta$ for $j \in \boldsymbol{I}_-$, $\mu_j^{(2)} = \mu_j^{(1)}$ for $j \notin \boldsymbol{I}$, $|\boldsymbol{I}_+| = |\boldsymbol{I}_-| = dp$, where $|\cdot|$ represents set cardinality, and d is the proportion of differential means. We chose $d = 5\%, 20\%$, representing relatively sparse and dense signals in mean difference, and used $\Delta = \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$.
- **Setting 2**: Same as Setting 1, but $\mu_j^{(1)} \sim \text{Unif}(5, 10)$.
- **Setting 3** (Negative binomial model with excess zeros): $W_{ij}^{(k)} = 0$ with probability π ($\pi = 10\%, 20\%$), and $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$ with probability $1 - \pi$. Other settings are same as in Setting 1, and we used $d = 10\%$, $\Delta = \{0.5, 1.0, 1.5, 2.0, 2.5\}$ in the simulation.

We compute the composition $X_{ij}^{(k)}$ by normalizing the abundance $W_{ij}^{(k)}$, and test the null hypothesis using compositional data at the level of 0.05. For Cao et al.'s test, we calculate the test statistics $\mathcal{M}_n$ and directly compute the p-value using Gumbel distribution. For our distance correlation test, p-value is computed based on 5,000 permutations.

For each setting, we simulate 1,000 datasets, and compare the true positive rates (TPRs) by the two tests. Figures 1, 2 and 3 summarize the TPRs under three settings. It can be seen that our distance based test consistently outperforms the log-ratio based method in all settings. Particularly, in the dense setting ($d = 20\%$), our test achieves substantially higher TPR than the log-ratio test. For instance, in Setting 1, when $\Delta = 2.0$, $n_1 = n_2 = 50$, our test achieves a high TPR of 0.97 while the TPR by log-ratio test is only 0.41. However, when Δ is subtle, e.g., $\Delta = 0.50$, both tests fail to detect the difference, even for relatively large sample size, e.g., $n_1 = n_2 = 100$.

In the second simulation study, we investigate the effect of dimension on the true positive rate. The sample size is fixed at $n_1 = n_2 = 100$, and the dimension p is varied from 100 to 500. The abundance $W_{ij}^{(k)}$ are generated from two different settings (similar to settings 1 and 3)

- **Setting 4**: $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$, $i = 1, ..., n_i, j = 1, ..., p$, $r_j^{(1)} \sim \text{Unif}(0.1, 1)$, $r_j^{(2)} = r_j^{(1)}$, $\mu_j^{(1)} \sim \text{Unif}(10, 15)$. Let $\boldsymbol{I} = \{\boldsymbol{I}_+, \boldsymbol{I}_-\}$ be the set of taxa with different

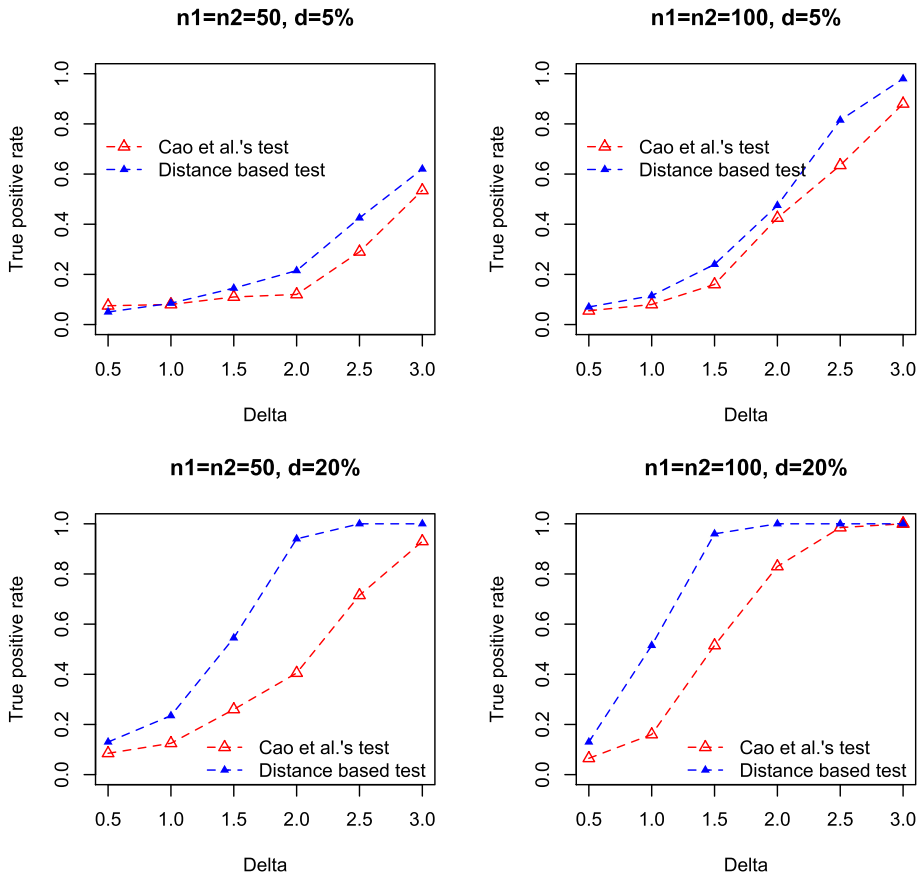

Fig. 1 True positive rate against mean difference Δ, by Cao et al.'s test (red) and our test (blue) in Setting 1. Results are based on 1,000 simulations

abundances in two conditions, $\mu_j^{(2)} = \mu_j^{(1)} + 1.5$ for $j \in \boldsymbol{I}_+$ and $\mu_j^{(2)} = \mu_j^{(1)} - 1.5$ for $j \in \boldsymbol{I}_-$, $\mu_j^{(2)} = \mu_j^{(1)}$ for $j \notin \boldsymbol{I}$, $|\boldsymbol{I}_+| = |\boldsymbol{I}_-| = 10$, where $|\cdot|$ represents set cardinality.

- **Setting 5** (Negative binomial model with excess zeros): $W_{ij}^{(k)} = 0$ with probability $\pi = 10\%$, and $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$ with probability $1 - \pi$. Other settings are same as in Setting 4.

Figure 4 summarizes the TPRs by the two tests based on 1,000 replicates and significance level $\alpha = 0.05$. It can be seen that the distance based test outperforms the log-ratio test especially when the dimension is relatively low. When the dimension is high, for instance $p = 500$, the two tests are comparable. More importantly, there is a substantial decrease of TPR as p increases, indicating that a feature screening could improve the test performance when p is large.

In the third study, we consider testing the compositional difference between multiple groups. We set $K = 4$ with sample sizes $n_1 = n_2 = n_3 = n_4 = 50$. The dimension p is fixed at 200. The abundance $W_{ij}^{(k)}$ are generated from the negative binomial model with excess zeros. Let $\pi = P\left(W_{ij}^{(k)} = 0\right)$, with probability $1 - \pi$, $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$, $i = 1, ..., n_i$, $j = 1, ..., p$, $r_j^{(1)} \sim \text{Unif}(0.1, 1)$, $r_j^{(3)} \sim \text{Unif}(0.1, 1)$, $r_j^{(2)} = r_j^{(1)}$, $r_j^{(4)} = r_j^{(3)}$, $\mu_j^{(1)} \sim \text{Unif}(10, 15)$, $\mu_j^{(3)} \sim \text{Unif}(10, 15)$. Let $\boldsymbol{I} = \{\boldsymbol{I}_+, \boldsymbol{I}_-\}$ be the set of taxa with different

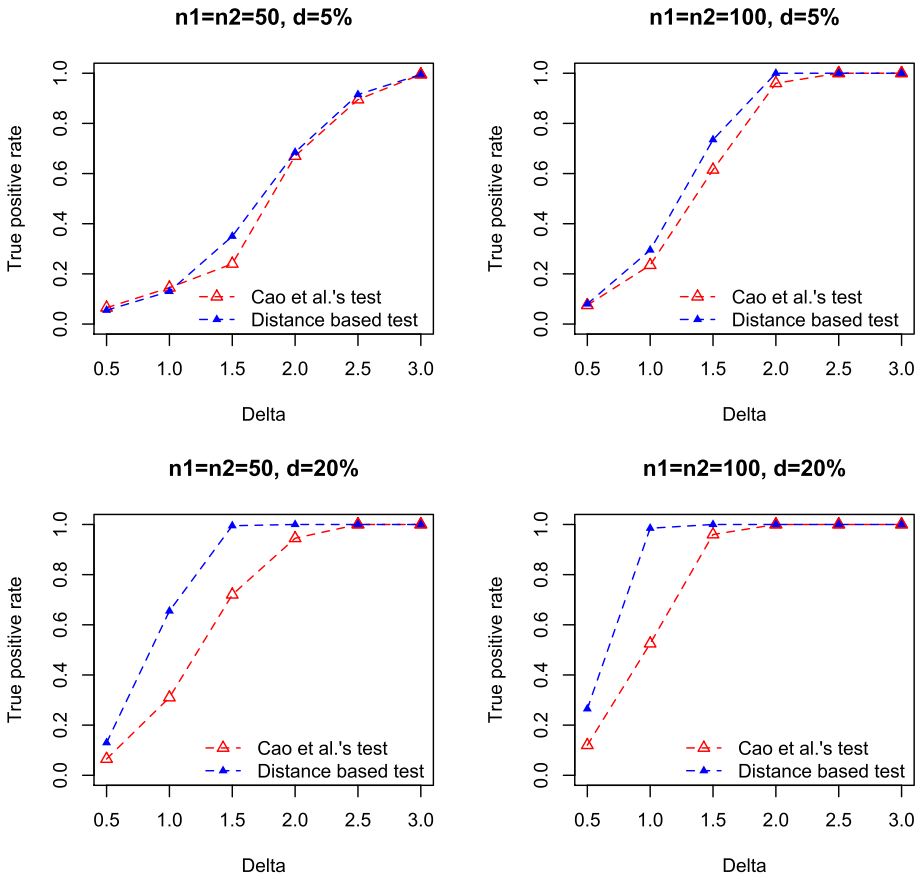

Fig. 2 True positive rate against mean difference Δ, by Cao et al.'s test (red) and our test (blue) in Setting 2. Results are based on 1,000 simulations

abundances in four conditions, $\mu_j^{(2)} = \mu_j^{(1)} + \Delta$ and $\mu_j^{(4)} = \mu_j^{(3)} + \Delta$ for $j \in \boldsymbol{I}_+$, $\mu_j^{(2)} = \mu_j^{(1)} - \Delta$ and $\mu_j^{(4)} = \mu_j^{(3)} - \Delta$ for $j \in \boldsymbol{I}_-$, $\mu_j^{(1)} = \mu_j^{(2)} = \mu_j^{(3)} = \mu_j^{(4)}$ for $j \notin \boldsymbol{I}$, $|\boldsymbol{I}_+| = |\boldsymbol{I}_-| = 20$, where $|\cdot|$ represents set cardinality. We use $\Delta = \{0.5, 1.0, 1.5, 2.0, 2.5\}$ and $\pi = \{10\%, 20\%\}$ in the simulation.

For the distance correlation test, p-value is computed based on 5,000 permutations. For Cao et al.'s test, we calculate the p-values by Gumbel distribution from six pairwise comparisons, and use the smallest p-value for decision-making. Figure 5 summarizes the TPRs by the two tests, where it can be seen that our proposed test performs consistently better than the log-ratio based test. Notably, in the setting $\pi = 20\%$ and $\Delta = 2.0$, the distance correlation test achieves a TPR of 0.83, compared to the TPR of 0.46 by the log-ratio test.

Two microbiome applications

Analysis of throat microbiome data

In this part, we use the proposed hypothesis and distance based test to reanalyze a throat microbiome dataset. Cigarette smokers have an increased risk of infectious diseases involving the respiratory tract, however, the consequences for global airway microbial community composition remains unclear. Charlson et al. (2010) used culture-independent high-density sequencing to analyze the microbiota from the right and left

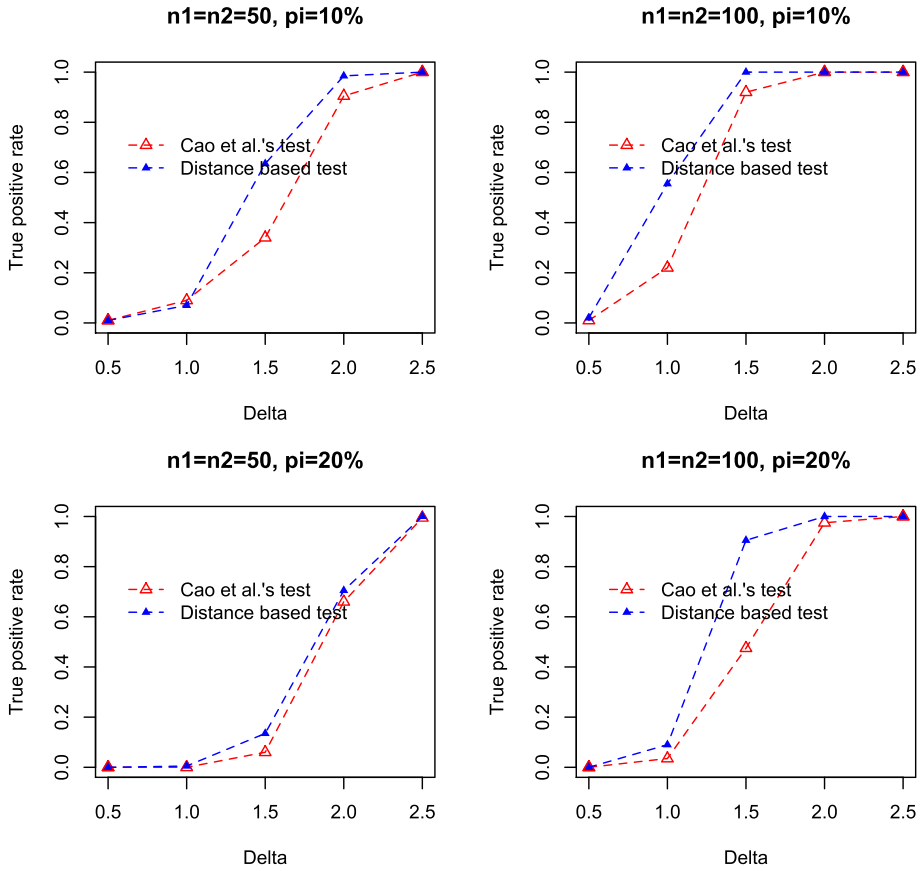

Fig. 3 True positive rate against mean difference Δ, by Cao et al.'s test (red) and our test (blue) in Setting 3. Results are based on 1,000 simulations

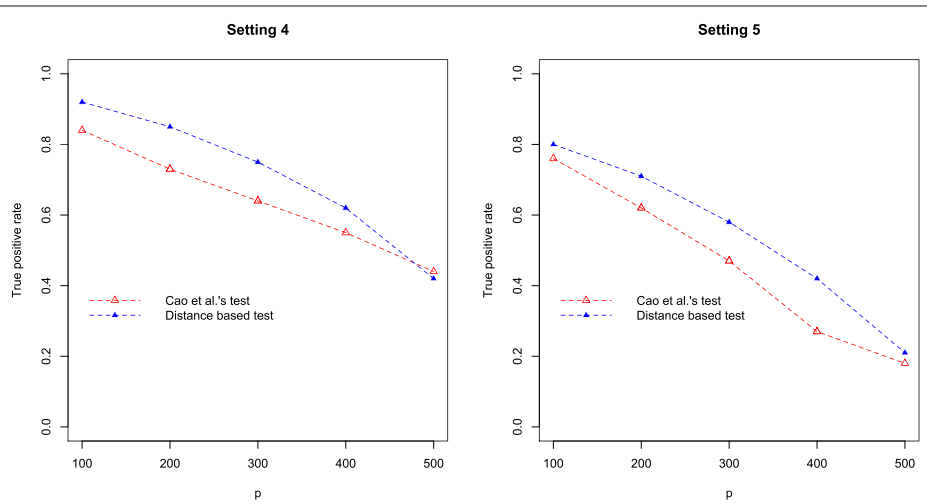

Fig. 4 True positive rate against dimension *p*, by Cao et al.'s test (red) and our test (blue) in the second simulation study (settings 4 and 5). Results are based on 1,000 simulations

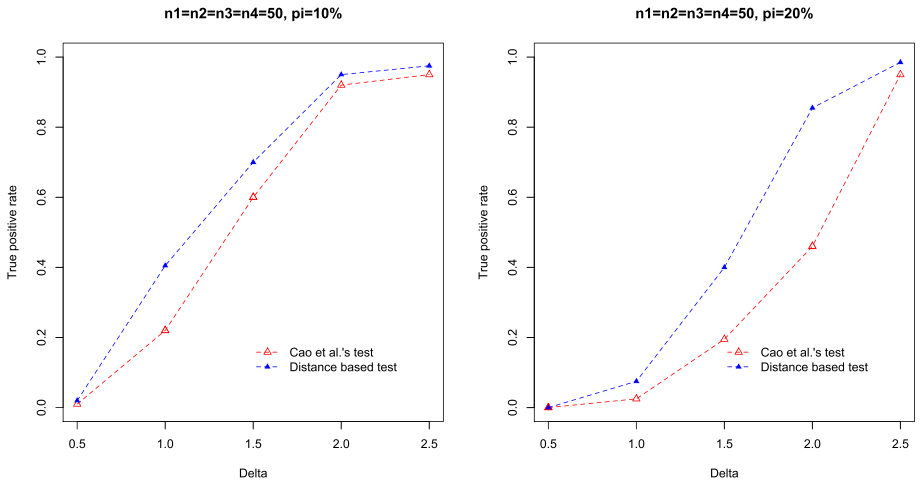

Fig. 5 True positive rate against mean difference Δ, by Cao et al.'s test (red) and our test (blue) in the third simulation study (zero-inflated negative binomial and multi-sample comparison). Results are based on 1,000 simulations

nasopharynx and oropharynx of 29 smoking and 33 nonsmoking healthy adults to assess microbial composition and effects of cigarette smoking [11]. Bacterial communities were profiled using 454 pyrosequencing of 16S sequence tags, aligned to 16S rRNA databases.

We are interested in whether there is any significant difference in microbial compositions between smokers and non-smokers. The processed data (observed abundance) were downloaded from R package *GUniFrac* [12], which included the read counts of 856 predefined operational taxonomic units (OTUs, also called phylotypes) on 62 samples. We first deleted OTUs with extremely small number of reads (less than 20 reads in total), resulting a final set of 190 OTUs.

Two methods, including the log-ratio based test and distance based test, are applied to the compositional data. The proposed distance correlation test is implemented in the following steps

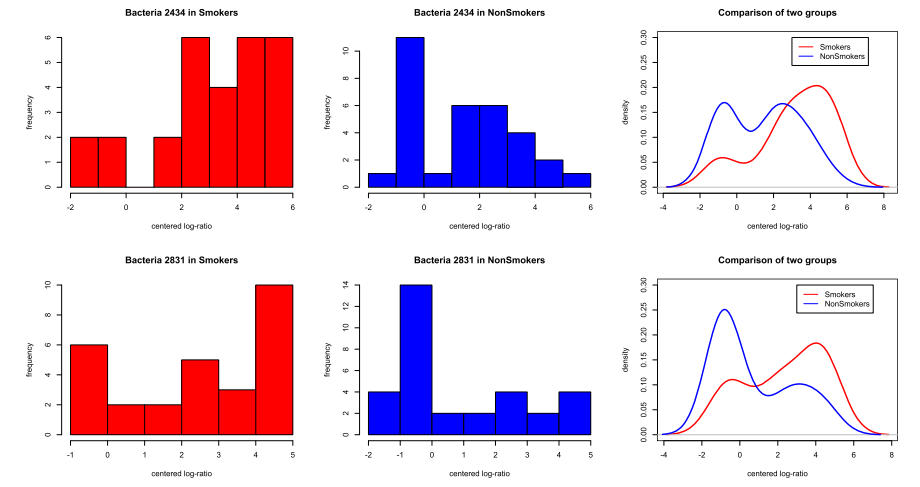

Fig. 6 Histograms and fitted density curves of centered log-ratios of two OTUs (bacteria 2434 and bacteria 2831), red for smokers and blue for non-smokers

- Step 1: Compute the composition X for each sample by normalizing the abundance W.
- Step 2: Calculate the sample proportions $\hat{p}_i$, and the inter-group distances $\hat{\mathcal{D}}_{ij}$ and $\hat{\mathcal{D}}_{ii}$, $i, j = 1, ..., K$ (e.g., Euclidean distance) using Eqs. (2) and (3).
- Step 3: Compute the permutation p-value based on $\widehat{\mathrm{dCov}}(\boldsymbol{X}, Y)$.

The proposed test yields a p-value of 0.0027, indicating a significant difference between smokers and non-smokers in microbial composition. In contrast, the test based on log-ratio transformation gives a p-value of 0.098, thus fails to reject the null hypothesis of equal means at the level of 0.05.

The disagreement between the two tests may indicate the existence of nonlinear effects and over-dispersion, because the log-ratio test only targets the mean difference while our test targets the distributional difference. We illustrate this point by carrying out additional analyses. Figure 6 gives two examples (bacteria 2434 and bacteria 2831), where the centered log-ratios exhibite substantially different distributions. However, the mean difference is not significant due to the nonlinear effect and heavy tails, which inflates the variance estimates in Cao et al.'s test.

We also compare the distributions of inter-point distance within smokers and non-smokers. Szekely et al. (2007) illustrated that if two multivariate distributions are

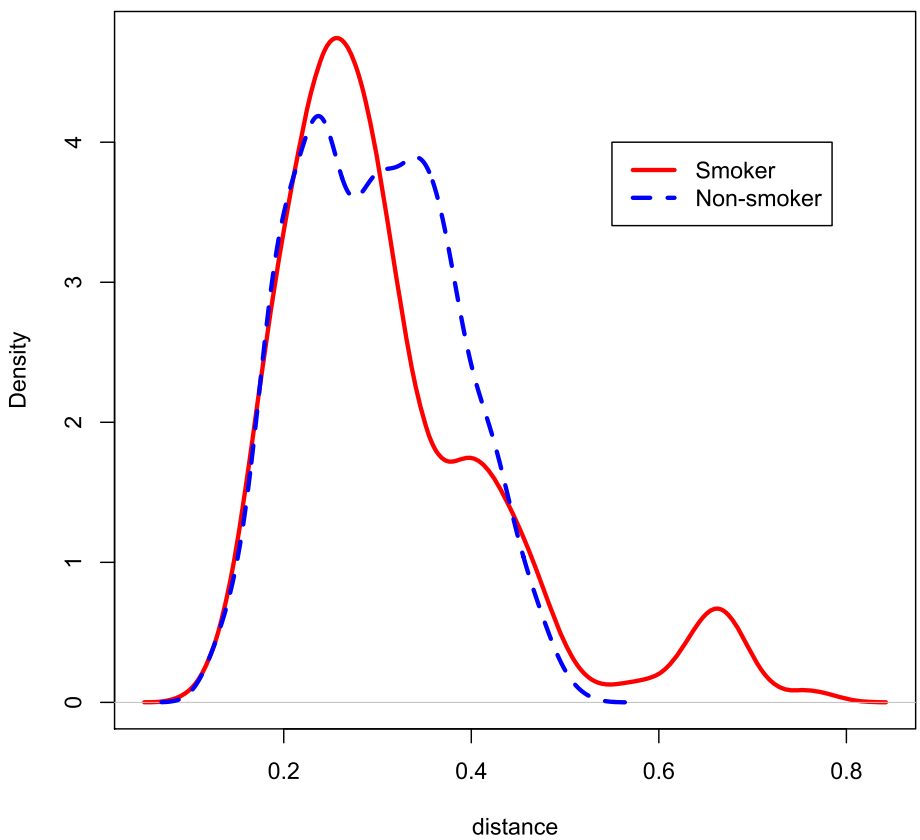

Fig. 7 Distributions of inter-point distance within the smoking group (red solid) and non-smoking group (blue dashed)

identical, the inter-point distances within each group have the same distribution. Figure 7 showed the inter-point distance distributions of two groups, where a substantial discrepancy was observed. Furthermore, we used the 3-minimum spanning tree (3-MST), a tree-based visualization method, to confirm our findings. Figure 8 shows the 3-MST based on the compositional data, where a connection in the network represents compositional similarity between two samples. In theory, if the two groups have the same distribution, then each sample has equal chance to connect with any other sample, regardless of which group it is from. However, it can be seen that certain samples from the same group formed clusters in the network. For instance, we identified a set of 12 smokers (circled) that are highly connected each other, but with very few connections with non-smokers, indicating a distributional difference in composition between the two groups.

Analysis of intestinal microbiome data

The microbial communities living in the human intestine have profound impact on our well-being and health. To understand the mechanisms that control this complex ecosystem, Lahti et al. (2014) conducted a deep phylogenetic analysis of the intestinal microbiota in 1,006 western adults from Europe and the United States [13]. The analysis

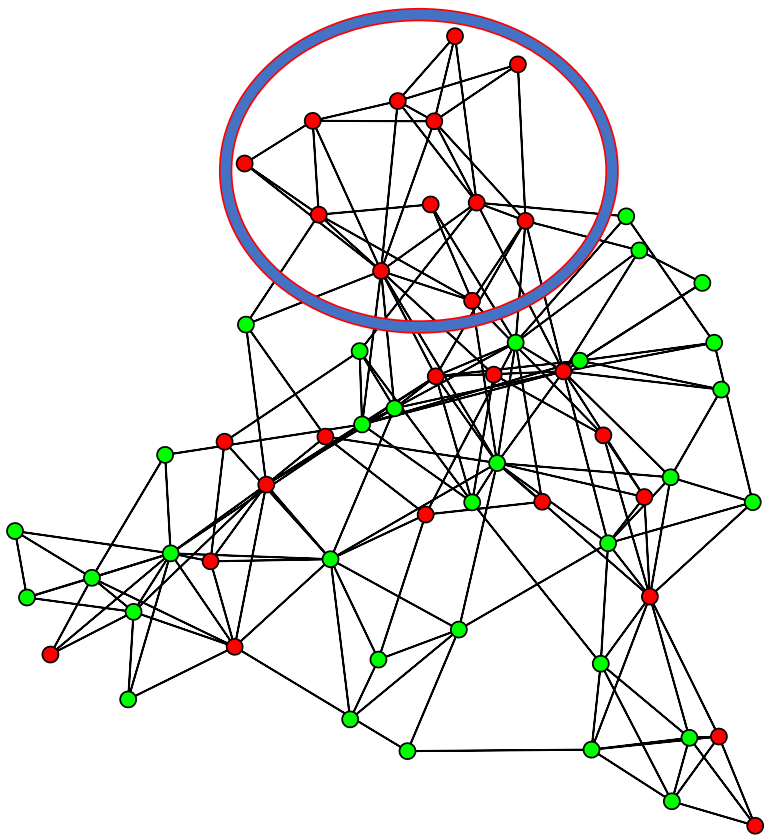

Fig. 8 The 3-MST for all samples based on Euclidean distance, red nodes are for smokers and green nodes are for non-smokers. A connection between two samples represents compositional similarity

is based on 130 genus-like phylogenetic groups that cover the majority of the known bacterial diversity of the human intestine. Clinical variables include age, nationality, BMI and DNA extraction method etc.

One of the key research questions is whether different age groups have different microbiome compositions. We use the cutoffs suggested by Lahti et al. (2014) to define three age groups: young (18–40), middle-aged (41–60) and older (61–77). The distance test yields an overall p-value of 3.0×10^{-6}. In addition, we calculate the p-values for the three pairwise comparisons: 8.2×10^{-5} for young vs middle aged, 2.2×10^{-5} for young vs older, and 0.081 for middle-aged vs older, indicating a significant difference in microbiome compositions between young and middle-aged/older subjects, but a minor difference between middle-aged and older subjects. To confirm this finding, we identified a list of microbiome groups with different distributions between age groups. Figures 9 and 10 show two examples of these, including group 25 and group 60. The distribution of inter-point distance within each age group is given in Fig. 11, where a discrepancy can be observed between young and middle-aged/older subjects.

Discussion

Microbiome data are often compositional, high-dimensional and over-dispersed, which poses great challenges to the statistical analysis. To overcome these obstacles, in this work,

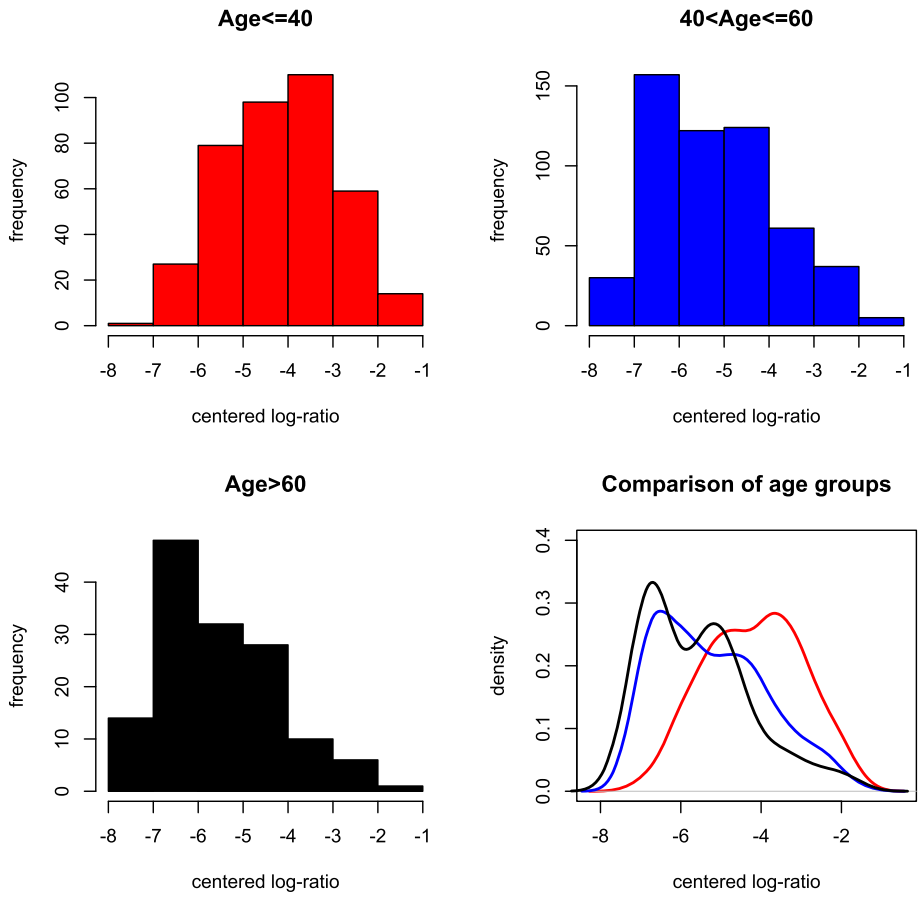

Fig. 9 Histograms and fitted density curves of the log-ratios of group 25, red for young, blue for middle-aged and black for older subjects

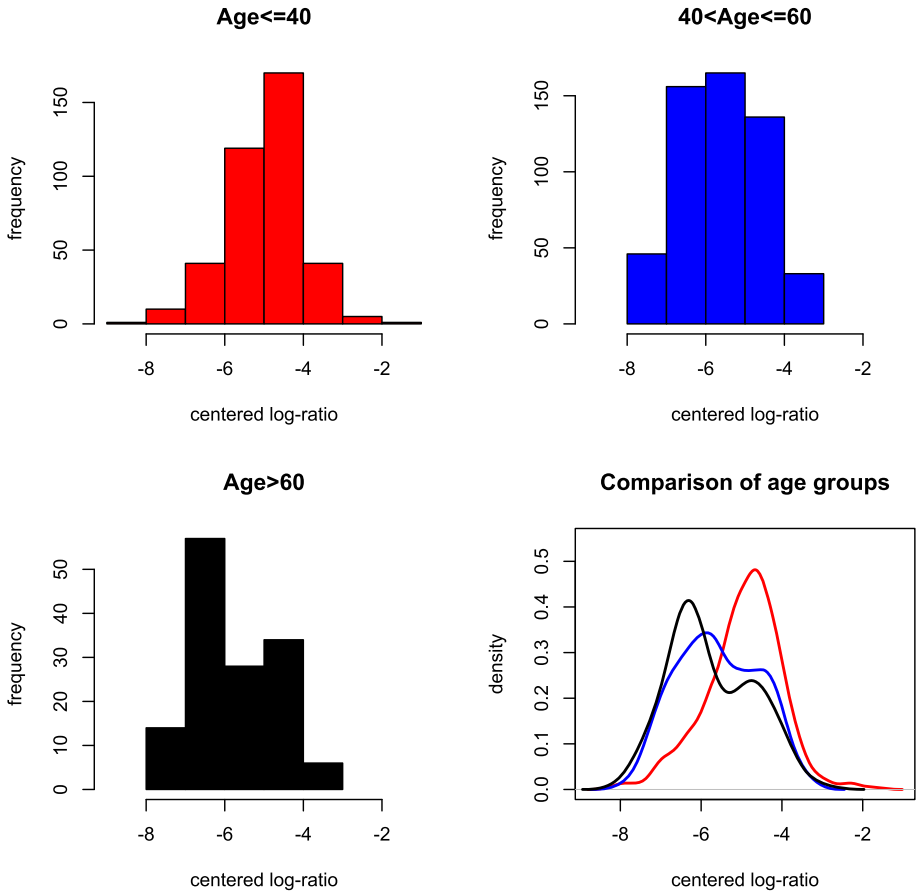

Fig. 10 Histograms and fitted density curves of the log-ratios of group 60, red for young, blue for middle-aged and black for older subjects

we formulated a new testable hypothesis from a Bayesian point of view, and suggested a nonparametric test to detect the compositional difference between multiple populations. Compared to the existing tests, our method has several advantages. First, the distance based test is free of parametric assumptions but directly targets the distributional difference, therefore it is capable of detecting nonlinear effects. The application in throat microbiome provided a good example, where the new test successfully captured the difference between two phenotypes, while the mean based test failed to do so. In addition, our method can deal with multiple groups, while most of existing methods are only for two-group comparison. Third, our test does not require sparsity assumption on the mean differences as in Cao et al.'s test, and in our simulation study, the new test worked quite well against both sparse and relatively dense alternatives.

There are several possible extensions of the proposed test. First, the distance based method can be readily extended to ordinal phenotypes (or conditions), although we have been using nominal phenotypes for illustrative purpose. For ordinal phenotype, $Y \in \{1, 2, ..., K\}$, where there is a natural ordering $1 < 2... < K$, (e.g., {mild, moderate, severe} for severity of a disease, {I, II, III, IV} for cancer stage, or {non-smoking, light smoking, heavy smoking} for smoking status), we need predefine the distance matrix between categories i and j, for instance, $d_{ij} = |i - j|$, or $d_{ij} = |i - j|^2$. The distance covariance between

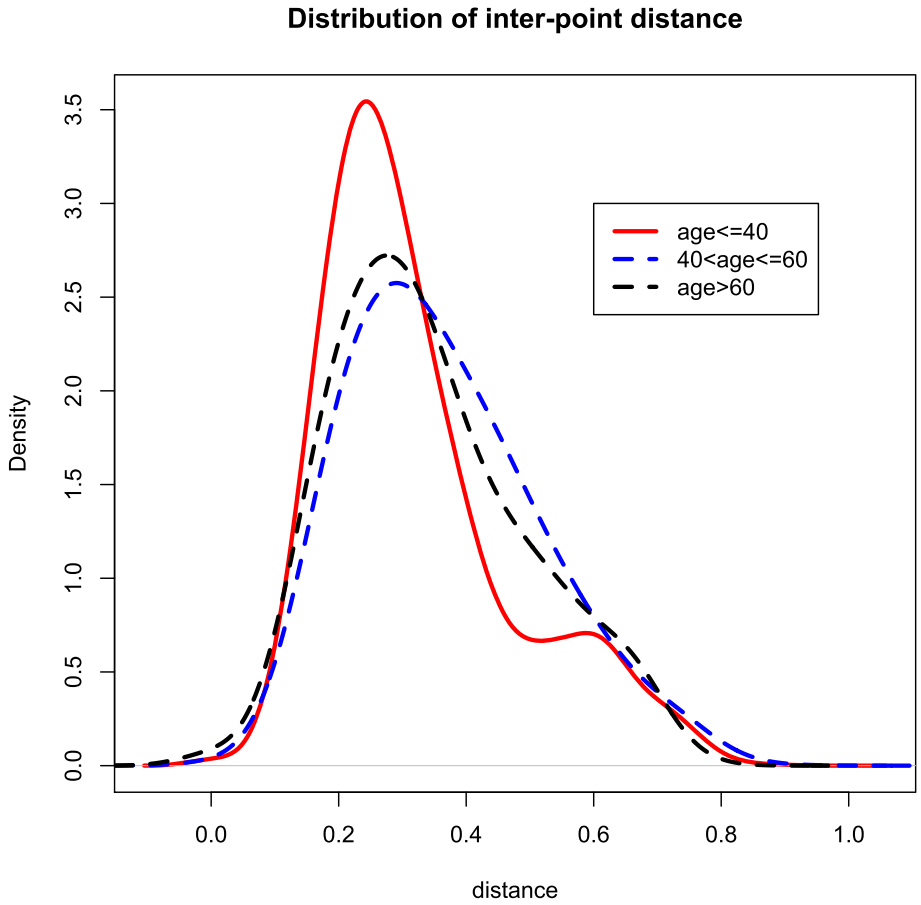

Fig. 11 Distributions of inter-point distance within each age group (red solid for young, blue for middle-aged and black for older subjects)

composition $\boldsymbol{X}$ and ordinal phenotype Y has the following expression

$$\mathrm{dCov}^2(\boldsymbol{X}, Y) = \left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j d_{ij}\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j D_{ij}\right) + \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j d_{ij} D_{ij} - 2\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{l=1}^{K} p_i p_j p_l d_{il} D_{ij},$$

and one may use the same permutation procedure to obtain p-values. In practice, the distance matrix d_{ij} should be carefully chosen to reflect the true spacings between categories. An inappropriate choice of d_{ij} may result in misleading conclusions. Second, our test might be improved by incorporating more information about bacteria taxa. For instance, one can assign different weights for different bacterial taxa based on their position in the polygenetic tree [14], and use weighted Euclidean distance to construct the test statistic.

In addition to the microbiome application that we illustrated in this paper, the proposed test can be readily applied to several other fields. For instance, the market share data in economics are compositional and often high-dimensional [15]. One may apply our test to detect the market share difference between multiple countries. In geology, it is often of

interest to study the compositions of species in sediment [16] and it is possible to apply our test to detect the difference in species compositions between multiple locations.

Conclusions

We formulate a Bayesian testing framework to identify the compositional differences between multiple populations. In addition, we propose to use the distance correlation measure to test the null hypothesis. Simulation studies and two real applications in the human microbiome demonstrate that our test is more sensitive to the compositional difference than the mean-based method, especially when the data are over-dispersed or zero-inflated. The proposed test is easy to implement and computationally efficient, facilitating its application to large-scale datasets.

Abbreviations

dCov: Distance covariance; OTU: Operational taxonomic unit; MST: Minimum spanning tree; TPR: True positive rate

Acknowledgements

QZ's research is supported in part by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 21 Supplement 9, 2020: Selected Articles from the 20th International Conference on Bioinformatics & Computational Biology (BIOCOMP 2019)*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-21-supplement-9.

Authors' contributions

QZ conceived of the presented idea and developed the theory. QZ and TD performed the computations and wrote the manuscript. Both authors read and approved of the final manuscript.

Funding

This work and publication costs are funded by the Arkansas Biosciences Institute (ABI-UAF17). The funder had no role in data analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The throat microbiome data can be downloaded from R package *GUniFrac* at https://cran.r-project.org/web/packages/GUniFrac/index.html. The intestinal microbiome data can be downloaded at https://datadryad.org/stash/dataset/doi:10.5061/dryad.pk75d.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have declared that no competing interests exist.

Received: 16 April 2020 Accepted: 30 April 2020 Published: 3 December 2020

References

1. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B. 1982;44(2):139–77.
2. Fry J, Fry T, McLaren K. Compositional data analysis and zeros in micro data. Appl Econ. 2010;32(8):953–9.
3. Cao Y, Lin W, Li H. Two-sample tests of high-dimensional means for compositional data. Biometrika. 2017;105(1):115–32.
4. Aitchison J. The statistical analysis of compositional data. Caldwell: Blackburn Press; 2003.
5. Szekely G, Rizzo M, Bakirov N. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35(6):2769–94.
6. Matteson D, James N. A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. J Am Stat Assoc. 2014;109(505):334–45.
7. Shen C, Priebe C, Vogelstein J. From Distance Correlation to Multiscale Graph Correlation; 2019. In Press. https://doi.org/10.1080/01621459.2018.1543125.
8. Zhu L, Xu K, Li R, Zhong W. Projection correlation between two random vectors. Biometrika. 2018;104(4):829–43.
9. Josse J, Holmes S. Measures of dependence between random vectors and tests of independence: a survey. 2014. arXiv:1307.7383.
10. Szekely G, Rizzo M. Energy statistics: A class of statistics based on distances. J Stat Plan Infer. 2013;143(8):1249–72.

11. Charlson E, Chen J, Custers-Allen R, Bittinger K, Li H, et al. Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS ONE. 2010;5(12):e15216.
12. Chen J, Bittinger K, Charlson E, Hoffmann C, Lewis J, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012;28(16):2106–13.
13. Lahti L, Salojarvi J, Salonen A, Scheffer M, de Vos W. Tipping elements in the human intestinal ecosystem. Nat Commun. 2014;5(4344):1–10.
14. Tang Y, Ma L, Nicolae D. A phylogenetic scan test on Dirichlet-tree multinomial model for microbiome data. Ann Appl Stat. 2018;12(1):1–26.
15. Morais J, Thomas-Agnan C, Simioni M. Using compositional and Dirichlet models for market share regression. J Appl Stat. 2018;45(9):1670–89.
16. Flood R, Bloemsma M, Weltje G, Barr I, O'Rourke S, et al. Compositional data analysis of Holocene sediments from the West Bengal Sundarbans, India: Geochemical proxies for grain-size variability in a delta environment. Appl Geochem. 2016;75:222–35.

Publisher's Note