Statistics in Biosciences (2021) 13:291–312
https://doi.org/10.1007/s12561-021-09306-6

Distance-Based Analysis with Quantile Regression Models

Shaoyu Li[1] · Yanqing Sun[1] · Liyang Diao[2] · Xue Wang[3]

Received: 30 January 2020 / Revised: 6 January 2021 / Accepted: 2 March 2021 /
Published online: 27 March 2021

Abstract

Non-standard structured, multivariate data are emerging in many research areas, including genetics and genomics, ecology, and social science. Suitably defined pairwise distance measures are commonly used in distance-based analysis to study the association between the variables. In this work, we consider a linear quantile regression model for pairwise distances. We investigate the large sample properties of an estimator of the unknown coefficients and propose statistical inference procedures correspondingly. Extensive simulations provide evidence of satisfactory finite sample properties of the proposed method. Finally, we applied the method to a microbiome association study to illustrate its utility.

Keywords Pairwise distance · Quantile regression · Asymptotic property · Microbiome association study · Ecology

1 Introduction

It has long been known that some microbes play critical roles in human health. For example, *Clostridium difficile* infections have been reported for more than 30 years [8], with the Centers for Disease Control reporting nearly half a million Americans infected in 2015 and a mortality rate of 1.3% within the first 30 days of diagnosis [23]. Not all microbiome health associations are due to infection by a single pathogenic bacteria, however. Even *C. difficile* infections often occur opportunistically, after a subject's microbiome becomes significantly altered, such as through the use of broad-spectrum antibiotics. Usage of such antibiotics significantly depletes the normal microbial diversity in the gut, thereby allowing pathogenic strains to proliferate. A dysbiosis, or imbalance, of the gut

✉ Shaoyu Li
sli23@uncc.edu

[1] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA

[2] Seres Therapeutics, Cambridge, MA, USA

[3] Department of Health Sciences Research, Mayo Clinic, Jacksonville, FL, USA

microbiome has been associated with conditions as varied as inflammatory bowel disease [14], graft-versus-host disease [34], and response to checkpoint inhibitor theory [7]. In the last decade, the role that the gut microbiome as a whole system plays in human disease has become widely appreciated. The emergence of high-throughput sequencing technologies, particularly 16S rRNA sequencing and whole metagenomics shotgun sequencing, has allowed the generation of microbiome data at unprecedented quantities and speeds. It is often now the analysis of these data and the extraction of meaningful biological signals that has become the bottleneck.

There are several essential features of microbiome data that challenge existing statistical methods. Microbiome data are typically high dimensional, with hundreds of species observed in a single subject's gut microbiome. Additionally, microbiome data are often compositional, given as abundance profiles, or they could be represented as the number of reads assigned to a species or other taxonomic level, and are, therefore, non-normally distributed. Another consideration is that microbiome data can be considered as phylogenetically structured, so that two samples that appear on the surface compositionally distinct may be phylogenetically or functionally similar.

Classical statistical methods for vectorially structured multivariate data, such as multivariate analysis of variance (MANOVA) and the Kruskal-Wallis test, become unsuitable. It is instead common to describe variation in multivariate outcomes by analyzing distance among all pairs of sample units. This distance measure could be a classical metric such as the Manhattan and Euclidean distance, or a study and data-type-specific measure, for example, the widely used identity-by-state (IBS) genetic distance in genetic association studies [24, 36], and the $\beta-$diversity metric in ecological studies. $\beta-$diversity is one type of biodiversity measurement for ecological data and is traditionally used to measure the number of species as well as the distribution of their abundances between two ecological communities [22]. Commonly used $\beta-$diversity measures, including the Bray-Curtis dissimilarity measure [3] and Jaccard distance [15], quantify the compositional dissimilarity between samples based on abundance distributions. More recently, UniFrac and generalized UniFrac distances were developed specifically for microbiome data to allow the incorporation of phylogenetic relatedness of species between samples [4, 27]. Once a distance measure is selected, pairwise distance between all samples is calculated and aggregated in a distance matrix. Statistical methods based on such distance matrices are termed distance-based methods [5, 22].

Distance-based analysis tools have been widely used in ecological research for decades [22] and are gaining attention across multiple fields, including genomics [35], social science [30], and microbiome studies [26]. Permutational multivariate analysis of variance (PERMANOVA) [1, 2, 29] and the distance-based F-test (DBF) [31, 32] are extensions of MANOVA to distance matrices, which examine the within and/or between-group variations of the pairwise distances. PERMANOVA is commonly used in microbiome studies to determine the significance of segregation of samples by a distance matrix, as it is nonparametric unlike the MANOVA, and unaffected by data sparsity [1]. The Mantel test [28] and the least-squares linear regression model for distance matrices [5, 6, 22, 25], on the other hand, are regression models of pairwise distances. Regression models are more flexible in terms of incorporating multiple covariates and handling different experimental designs.

Although existing methods for analyzing pairwise distances, current approaches have limitations. First, pairwise distances are likely to be positively skewed due to their non-negativity. For example, pairwise-weighted UniFrac distances in our application example are positively skewed (left panel of Fig. 1). Quantiles at tails could be significantly different, even when the median/mean of pairwise distances is the same (e.g., right panel, Fig. 1). As such, a quantile regression (QR) model is more suitable for the analysis of pairwise distances. Quantile regression models τth quantile of a response variable Y condition on a $(p+1) \times 1$ vector of covariates $x = (1, x_1, \dots, x_p)^T$ as $Q_\tau(Y|x) = x^T\beta(\tau)$, $\tau \in (0, 1)$. It requires minimal distributional assumptions and, therefore, is more robust. Also, by allowing the entire spectrum of the conditional distribution of the response variable to be related to a group of covariates, it provides much richer information on the distributional changes of the response variable than least-squares regression. To the best of our knowledge, however, there is no existing literature on distance-based quantile regression. Second, existing distance-based analysis tools rely on either a permutation procedure or a distribution approximation approach [31, 32] for statistical inference. Permutation testing is known to be computationally expensive, especially when there are tens of thousands of tests, as required in genome-wide association studies. In addition, permutation testing gives severely inflated type I error rate when heteroscedasticity presents. The distribution approximation-based approach is much faster; however, it is not generally applicable as it requires known closed form for the moments of the test statistic. Therefore, we propose a quantile regression (QR) model of pairwise distances and investigate the asymptotic properties of an estimator of the model parameter for an efficient statistical inference procedure.

The rest of the article is organized as follows. The QR model for pairwise distances, model parameter estimation, and asymptotic results are presented in Sect. 2. Hypothesis testing is discussed in Sect. 3. Section 4 demonstrates the finite sample performance of our proposed method by numerical simulations. We apply the

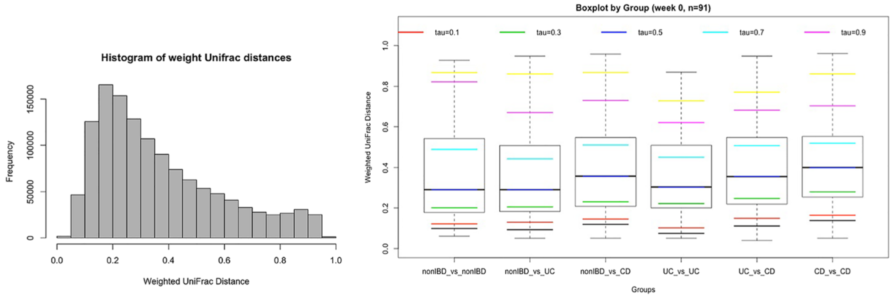


Fig. 1 (Left panel) A histogram of pairwise-weighted UniFrac distances for the iHMP dataset. (Right panel) Boxplot of pairwise-weighted UniFrac distances by group: pairwise distance between healthy controls (non-IBD vs non-IBD), healthy controls and UC patients (non-IBD vs UC), healthy controls and CD patients (non-IBD vs CD), UC patients (UC vs UC), UC patients and CD patients (UC vs CD), and CD patients (CD vs CD)

proposed method to a motivating microbiome association study and report the results in Sect. 5. Concluding remarks are given in Sect. 6.

2 Statistical Model

2.1 Model Specification and Parameter Estimation

Suppose a set of i.i.d. observations $(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, 2, \dots, n$, with $\boldsymbol{x}_i \in \mathcal{R}^p$ and $\boldsymbol{y}_i \in \mathcal{R}^q$, p, q are fixed numbers. Calculate the pairwise distance for the response variable, $y_{ij} \equiv s(\boldsymbol{y}_i, \boldsymbol{y}_j), i = 1, 2, \dots, n, j > i$, and the pairwise distance for the covariates between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $\boldsymbol{x}_{ij} \equiv (1, s_1(x_{i1}, x_{j1}), \dots, s_p(x_{ip}, x_{jp}))$, where $s, s_1, \dots, s_p$ are pre-selected, known functions that quantify the pairwise distances. Depending on the subject of an application, various distance metrics may be used. In microbiome studies, for example, the Bray-Curtis measure of β-diversity is commonly used for measuring dissimilarity between sample profiles. Once a distance metric is decided, dissimilarity between microbiome profiles are calculated and aggregate in a distance matrix, which is square and symmetric. Vectorize the upper triangle of each distance matrix and denote them by $\mathbf{y} = (y_{12}, y_{13}, \dots, y_{1n}, y_{23}, \dots, y_{2n}, \dots, y_{n-1,n})^T$. Similarly, distance matrices on covariates are calculated and vectorized: $\boldsymbol{X} = (\boldsymbol{x}_{12}^T, \boldsymbol{x}_{13}^T, \dots, \boldsymbol{x}_{n-1,n}^T)^T$. Although pairwise distances on each individual covariate are used here for model demonstration, distances can also be defined using sub-groups of covariates or all covariates. For example, $\boldsymbol{x}_{ij} \equiv (1, s_1(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_j^{(1)}), s_2(\boldsymbol{x}_i^{(2)}, \boldsymbol{x}_j^{(2)}))$, where $[(\boldsymbol{x}_i^{(1)})^T, (\boldsymbol{x}_i^{(2)})^T] = \boldsymbol{x}_i^T$. Let $F(y|\boldsymbol{x}_{ij}) = P(y_{ij} \le y|\boldsymbol{x}_{ij}), i = 1, 2, \dots, n, j > i$, we model τ quantile of the conditional distribution of y_{ij} given $\boldsymbol{x}_{ij}$, $Q_\tau(y_{ij}|\boldsymbol{x}_{ij}) = F^{-1}(\tau)$, through the following regression:

$$Q_\tau(y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta}(\tau) \quad for \quad \tau \in (0, 1), \tag{1}$$

where $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \dots, \beta_p(\tau))^T$ is a $(p+1) \times 1$ vector of unknown coefficients representing the effect of $\boldsymbol{x}_{ij}$ on the τth quantile of y_{ij}, and $\boldsymbol{\beta}(\tau)$ can be different at different τ.

We estimate the unknown coefficients $\boldsymbol{\beta}(\tau)$ by minimizing the following objective function [16–19]:

$$S_n(\boldsymbol{\beta}(\tau)) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \rho_\tau(y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}(\tau)), \tag{2}$$

where function $\rho_\tau(u) = u(\tau - \mathbf{I}(u < 0))$ is the so-called "check" function. An estimate of $\boldsymbol{\beta}(\tau)$ is obtained by minimizing $S_n(\boldsymbol{\beta})$ using linear programming (LP) via the *rq()* function in the contributed R package *quantreg*.

The asymptotic results for the i.i.d case are not applicable because the pairwise distances are dependent: all $y_{ij}, j > i$ involve the ith observation and are correlated. Therefore, we investigate the large sample properties of $\hat{\boldsymbol{\beta}}(\tau) = \text{argmin}_{\beta} S_n(\boldsymbol{\beta}(\tau))$ by taking into account the pairwise correlation structure.

2.2 Large Sample Properties

For simplicity, we use notation $\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}(\tau)$ in the following demonstration. Since the "check" function $\rho_\tau(u)$ is not differentiable at $u = 0$, a normalized subgradiant of $S_n(\boldsymbol{\beta})$ is used and denoted as $U_n(\boldsymbol{\beta})$ [11]:

$$U_n(\boldsymbol{\beta}) \equiv \sqrt{n}\binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} q(y_{ij}, \boldsymbol{x}_{ij}; \boldsymbol{\beta}),$$

where each component of the vector $q(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta})$ is a convex combination of the left and right partial derivatives of $\rho_\tau(\cdot;\boldsymbol{\beta})$ with respect to the corresponding component of $\boldsymbol{\beta}$, that is,

$$q(\cdot;\boldsymbol{\beta}) = \alpha\frac{\partial^- \rho_\tau(\cdot;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} + (1-\alpha)\frac{\partial^+ \rho_\tau(\cdot;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \quad \alpha \in [0,1]. \tag{3}$$

So, when function $\rho_\tau(\cdot;\boldsymbol{\beta})$ is differentiable, $q(\cdot;\boldsymbol{\beta}) = \frac{\partial\rho_\tau(\cdot;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$, and when $\rho_\tau(\cdot;\boldsymbol{\beta})$ is not differentiable, $q(\cdot;\boldsymbol{\beta}) = \tau\frac{\partial^-\rho_\tau(\cdot;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} + (1-\tau)\frac{\partial^+\rho_\tau(\cdot;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = (\tau(\tau-1) + (1-\tau)\tau)\boldsymbol{x}_{ij}^T = 0$, with $\alpha = \tau$. Noticing that both the objective function $S_n(\boldsymbol{\beta})$ and the normalized subgradiant $U_n(\boldsymbol{\beta})$ have the form of a second-order U-statistic, we were inspired to incorporate the large sample theories of U-statistics into the framework of regular quantile regression model and derive the asymptotic consistency and normality of $\hat{\boldsymbol{\beta}}$. We state the major results here, assumptions (A1)–(A8) and detailed proofs are referred to the web-based supporting material.

Theorem 1 *Under Assumptions (A1) through (A3),* $S_n(\boldsymbol{\beta}) - E[S_n(\boldsymbol{\beta})]$ *converges to zero almost surely uniformly over* $\boldsymbol{\beta} \in \Theta$.

The strong law of large numbers (SLLN) for U-statistics [9] implies that $S_n(\boldsymbol{\beta}) \to E(S_n(\boldsymbol{\beta}))$ a.s. We further show that $S_n(\boldsymbol{\beta})$ is almost surely *Lipschitz* continuous and, therefore, stochastically equicontinuous. Then $S_n(\boldsymbol{\beta}) \to E[S_n(\boldsymbol{\beta})]$ almost surely over Θ according to the Stochastic Ascoli Lemma.

Theorem 2 (**Asymptotic Consistency**) *Under Assumptions (A1) through (A5), the estimator defined by minimizing* $S_n(\boldsymbol{\beta})$ *is strongly consistent for* $\boldsymbol{\beta}_0 = \text{argmin}_{\boldsymbol{\beta}\in\Theta} E[S_n(\boldsymbol{\beta})]$.

We show that $\boldsymbol{\beta}_0$ is an unique minimizer of $E(S_n(\boldsymbol{\beta}))$. Combining the almost surely uniform convergence of $S_n(\boldsymbol{\beta})$ to $E(S_n(\boldsymbol{\beta}))$ by Theorem 1, we have that $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} S_n(\boldsymbol{\beta})$ is strongly consistent of $\boldsymbol{\beta}_0$ following the consistency theorem for M-estimator [12, 13].

Theorem 3 (**Asymptotic Normality**) *Under Assumptions (A1) through (A8), if* $\hat{\boldsymbol{\beta}}$ *is consistent for* $\boldsymbol{\beta}_0$*, then* $\hat{\boldsymbol{\beta}}$ *satisfies the following asymptotic linearity relation*

$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=-A_0^{-1}\frac{2}{\sqrt{n}}\sum_{i=1}^n r(y_i,\boldsymbol{x}_i,\boldsymbol{\beta}_0)+o_p(1)$, *and has asymptotic normal distribution,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{\mathrm{d}} N\big(\boldsymbol{0},A_0^{-1}V_0A_0^{-1}\big), \tag{4}$$

with $A_0=\frac{\partial\lambda(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}'}$ *and* $V_0=4E[r(y_i,\boldsymbol{x}_i,\boldsymbol{\beta}_0)r(y_i,\boldsymbol{x}_i,\boldsymbol{\beta}_0)']$, *where* $r(y_i,\boldsymbol{x}_i,\boldsymbol{\beta}_0)\equiv E(q(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta}_0)|y_i,\boldsymbol{x}_i)$ *and* $\lambda(\boldsymbol{\beta}_0)\equiv E[q(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta}_0)]=E[r(y_i,\boldsymbol{x}_i,\boldsymbol{\beta}_0)]$.

We first prove that $U_n(\boldsymbol{\beta}_0)+\sqrt{n}\lambda(\hat{\boldsymbol{\beta}})\xrightarrow{p}0$. Even though $U_n(\boldsymbol{\beta})$ is not differentiable, its expected value $E(U_n(\boldsymbol{\beta}))$ is. The Taylor series expansion of $\lambda(\hat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$ yields $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=A_0^{-1}\sqrt{n}\lambda(\hat{\boldsymbol{\beta}})+o_p(1)$, where A_0 is the Hessian matrix. Replace $\sqrt{n}\lambda(\hat{\boldsymbol{\beta}})$ by $-U_n(\boldsymbol{\beta}_0)$ and apply the asymptotic normality theorem for U-statistics, we prove the asymptotic linearity and the normality.

2.3 Estimating the Covariance Matrix

Statistical inference about the unknown coefficients is always important in real applications. Although we have derived the asymptotic results, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ involves unknown matrices A_0 and V_0 and needs to be estimated.

Estimating V_0 is relatively straightforward. The conditional expectation $r(y_i,\boldsymbol{x}_i,\boldsymbol{\beta})$ can be estimated by its sample mean $\hat{r}(y_i,\boldsymbol{x}_i,\boldsymbol{\beta})=\frac{1}{n-1}\sum_{j\neq i}q(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta})$. By construction $\frac{1}{n}\sum_{i=1}^n\hat{r}(y_i,\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})=\binom{n}{2}^{-1}\sum_{1\le i<j\le n}q(y_{ij},\boldsymbol{x}_{ij},\hat{\boldsymbol{\beta}})=U_n(\hat{\boldsymbol{\beta}})/\sqrt{n}=o_p(1/\sqrt{n})$. An estimator of V_0 using $\hat{r}$ is given by

$$\hat{V}_0=\frac{4}{n}\sum_{i=1}^n\hat{r}(y_i,x_i,\hat{\boldsymbol{\beta}})\hat{r}(y_i,x_i,\hat{\boldsymbol{\beta}})^T. \tag{5}$$

The estimation of A_0 is challenging. We adopt a simpler bootstrap approach by Honoré and Hu [10]. The simpler bootstrap approach utilizes the relationship between the proposed estimator $\hat{\boldsymbol{\beta}}$ and a "split-sample"-based estimator $\tilde{\boldsymbol{\beta}}$ defined as follows:

$$\tilde{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\,S_n^*(\boldsymbol{\beta})=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\,\frac{1}{n^*}\sum_{i=1}^{n*}\rho_\tau(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta}),\quad j=i+n^*. \tag{6}$$

$\tilde{\boldsymbol{\beta}}$ is the minimizer based on $n^*=int(n/2)$ non-overlapping, and therefore, independent, pairwise distances. That means the general large sample properties for M-estimators can be applied; hence, $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\rightarrow N(\boldsymbol{0},A_1^{-1}V_1A_1^{-1})$, where $A_1=\partial E[\partial S_n^*(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}]/\partial\boldsymbol{\beta}'$ and $V_1=2Var[q_{ij}(\boldsymbol{\beta}_0)]$. Notice that, $A_1=A_0$ under random sampling, an estimate of A_0 can be obtained by estimating A_1 using the "split-samples."

Specifically, we obtain a series of estimates $\tilde{\boldsymbol{\beta}}^b$, for $b=1,2,\dots,B$, using bootstrap. Then we calculate the sample covariance matrix of $\tilde{\boldsymbol{\beta}}$ and denote it by $\widehat{Var}(\tilde{\boldsymbol{\beta}})$. A_1^{-1} can then be solved from the following equation, with the constraint that A_1^{-1} is symmetric and positive definite: $\widehat{Var}(\tilde{\boldsymbol{\beta}}) = A_1^{-1}\hat{V}_1A_1^{-1}$, where $\hat{V}_1 = 2B^{-1}\sum_{b=1}^{B}((n^*)^{-1}\sum_{i=1}^{n^*} q_{i,i+n^*}^{\otimes 2}(\tilde{\boldsymbol{\beta}}^b))$. These equations are special cases of the continuous time albegraic Riccati equations [20], and they are known to have unique and non-negative definite solutions. We use the Schur method by Laub [21] to solve the equation for A_1^{-1}.

The simpler bootstrap approach is computationally much more efficient than ordinary bootstrap procedure, because we only need to refit models using "split-samples," which is much smaller in size. For example, when sample size $n = 100$, the simpler bootstrap procedure refits QR models with 50 independent pairwise distances, while, ordinary bootstrap approach refits QR model with $\begin{pmatrix} 100 \\ 2 \end{pmatrix} = 4950$ data points.

3 Hypothesis Testing

Hypothesis testing is an essential component of statistical inference because it is often of practical interest to test if a certain covariate is significantly associated with the response variable. This testing problem can be accommodated by considering the null hypothesis: $H_0 : \beta_k(\tau) = 0$ that coefficient of the kth covariate in the τth quantile model is zero. We may use a Wald-type t-statistic $T = \frac{\hat{\beta}_k(\tau)}{se(\hat{\beta}_k(\tau))}$ and calculate p values using our derived asymptotic normality result.

As one would expect, the normal approximation can be unsatisfactory when the sample size is small. In such cases, empirical p values may be calculated using the following permutation procedure: (1) simultaneously permute the rows and columns of the pairwise distance matrix of the kth covariate ($\boldsymbol{x}_{ij}^k = s_k(x_{ik}, x_{jk})$), while keeping distance matrices of the response variable and other covariates unchanged; (2) fit the QR model using the vectorized permuted distance matrix and calculate the value of test statistic; (3) Repeat (1)–(2) a large number of times B, say $B = 1000$, and an empirical p value can be obtained by comparing the observed test statistic value in the original model with the permuted ones. Specifically, $pvalue = \frac{\#(|\hat{\beta}_k^b| > |\hat{\beta}_k|)+1}{B}$. Additionally, we can test if the covariate is associated with any considered quantiles using $\hat{\beta}_{k,\max} = \max_\tau(|\hat{\beta}_k(\tau)|)$. We denote the p value for the supremum test by $p_{\max} = \frac{\#(\hat{\beta}_{k,\max}^b > \hat{\beta}_{k,\max})+1}{B}$, where $\hat{\beta}_{k,\max}$ and $\hat{\beta}_{k,\max}^b$ are calculated in the original and permuted data, respectively.

We recommend conducting the permutation test only when the sample size is small for three reasons. First, fitting a QR model with $\begin{pmatrix} n \\ 2 \end{pmatrix}$ data points becomes exponentially slower as the sample size increases, and the permutation procedure becomes especially time consuming. Second, the performance of the proposed T test statistic using the asymptotic distribution is satisfactory when sample size is large. Based on our numerical simulations, a sample size of $n = 200$ may be considered large enough. Finally, permutation testing is not a universal solution. When heteroscedasticity presents, permutation testing has a severely inflated type I error rate.

4 Numerical Simulations

We conducted Monte-Carlo simulations to evaluate the finite sample performance of our proposed method, including estimation accuracy and empirical type I error rate and power for our proposed test.

4.1 Model Performance

We considered two scenarios by generating data from models with homoscedastic and heteroscedastic random errors, and called them scenarios I and II, respectively.

In scenario I, responses were generated through the following linear regression model: $y_{ij} = b_1 x_{ij} + b_2 z_{ij} + \varepsilon_{ij}$, $i = 1, 2, \dots, n, j = i+1, \dots, n$, with $x_i \sim Unif(0,1)$, $z_i \sim N(0,1)$, and the Euclidean distance was used for pairwise distance matrices. That is, $x_{ij} = |x_i - x_j|$, $z_{ij} = |z_i - z_j|$. The random error terms $\varepsilon_{ij} = \varepsilon_i + \varepsilon_j$, with $\varepsilon_i, i = 1, 2, \dots, n$ are i.i.d. random samples from a normal distribution $N(0, \sigma^2)$. The corresponding τth quantile regression model is $Q_\tau(y_{ij}|x_{ij}, z_{ij}) = \beta_0(\tau) + \beta_1(\tau)x_{ij} + \beta_2(\tau)z_{ij}$, $\beta_0(\tau)$ is the τth quantile of ε_{ij} and $(\beta_1(\tau), \beta_2(\tau)) = (b_1, b_2)$ are constants for all $\tau \in (0,1)$. We set $b_2 = 1$ and $b_1 = 0$ or 1, corresponding to the null and alternative hypothesis for testing $H_0 : \beta_1 = 0$. Two different values of $\sigma^2 (= 0.5, 1)$ were used.

In scenario II, we considered two different cases: case 1 has a discrete covariate and case 2 has only continuous covariates. In case 1, we generated data via a linear model $y_{ij} = b_1 x_{ij} + z_{ij} b_2 + (z_{ij}\eta_{ij} + \varepsilon_{ij})$, where x_{ij} and ε_{ij} were simulated as in scenario I. While, we simulated $z_i \sim Bernoulli(p = 0.5)$ and z_{ij} is the Manhattan distance between z_i and z_j. An additional random error, $\eta_{ij} = \eta_i + \eta_j$, $\eta_i \sim N(0,1), i = 1, 2, \dots, n$ was multiplied to z_{ij}. The corresponding true quantile regression model then is $Q_\tau(y_{ij}|x_{ij}, z_{ij}) = b_1 x_{ij} + b_2 z_{ij} + Q_\tau(z_{ij}\eta_{ij} + \varepsilon_{ij}) = F_1^{-1}(\tau) + b_1 x_{ij} + (b_2 + F_2^{-1}(\tau) - F_1^{-1}(\tau))z_{ij}$, where F_1, F_2 are cumulative distribution functions of ε_{ij} and $\varepsilon_{ij} + \eta_{ij}$, respectively. Clearly, $\beta_0(\tau) = F_1^{-1}(\tau)$ and $\beta_2(\tau) = b_2 + F_2^{-1}(\tau) - F_1^{-1}(\tau)$ are τ dependent, and $\beta_1(\tau) = b_1$ does not depend on τ. We set b_1 to be 0 or 1 corresponding to the null and alternative model for testing $H_0 : \beta_1(\tau) = 0$. We set $b_2 = F_1^{-1}(\tau) - F_2^{-1}(\tau)$ for $\tau = (0.1, 0.3, 0.5, 0.7, 0.9)$, respectively. For example, when we set $b_2 = F_1^{-1}(0.5) - F_2^{-1}(0.5)$, then, $\beta_2(0.5) = 0$ and $\beta_2(\tau) = F_1^{-1}(0.5) - F_2^{-1}(0.5) + F_2^{-1}(\tau) - F_1^{-1}(\tau) \neq 0$, when $\tau \neq 0.5$. Therefore, same data were used to exam the empirical type I error rate for testing $H_0 : \beta_2(0.5) = 0$ and study the empirical power of testing $H_0 : \beta_2(\tau) = 0, \tau \neq 0.5$.

In case 2, we generated $y_{ij} = b_1 x_{ij} + b_2 z_{ij} + (1 + z_{ij})\varepsilon_{ij}$. x_{ij}, z_{ij}, and ε_{ij} were simulated similarly as in scenario I. The conditional τth quantile of y_{ij} is $Q_\tau(y_{ij}|x_{ij}, z_{ij}) = b_1 x_{ij} + b_2 z_{ij} + \sqrt{2}Z_\tau(1 + z_{ij})\sigma$, where Z_τ denotes τth quantile of a standard normal distribution. That means $\beta_0(\tau) = \sqrt{2}\sigma Z_\tau$, $\beta_1(\tau) = b_1$, and $\beta_2(\tau) = b_2 + Z_\tau\sqrt{2}\sigma$. We set $b_1 = 1$, $b_2 = -\sqrt{2}\sigma Z_\tau, \tau = (0.1, 0.3, 0.5, 0.7, 0.9)$, respectively. By setting $b_2 = -\sqrt{2}\sigma Z_\tau$ at different levels of τ, we generate data under the null hypothesis of $H_0 : \beta_2(\tau) = 0$, which also serves as an alternative case for testing $H_0 : \beta_2(\delta) = 0, \delta \neq \tau$. For example, when $b_2 = -\sqrt{2}\sigma Z_{0.1}$, $\beta_2(0.1) = 0$, but $\beta_2(\tau) = b_2 + \sqrt{2}\sigma Z_\tau = \sqrt{2}\sigma Z_\tau - \sqrt{2}\sigma Z_{0.1} \neq 0$, when $\tau \neq 0.1$.

Sample sizes of (30, 50, 100, 200) were considered, and 1,000 bootstrap/permuted samples were used in all procedures. Results reported here are based on 1,000 replicates. We also simulated data with $\varepsilon_i \sim exp(1)$ and $\varepsilon_i \sim Cauchy(1)$. Results of these additional simulations are referred to the online supplementary file.

4.2 Results

We summarized the empirical mean bias (EmpBias), the empirical standard deviation (EmpSD), and the average estimated standard deviation (EstSD) of $\hat{\boldsymbol{\beta}}$, and empirical coverage probability (EmpCP) of 95% confidence intervals. We calculated the empirical type I error rate and power for the proposed test statistic based on asymptotic normality and compare the results with permutation test. We denote the two methods as “A” and “P,” respectively, in the tables. We benchmark our method against permutation test because it is the most commonly used approach in many distance-based tools, including the Mantle test and PERMANOVA. Besides, the Mantle test and PERMANOVA are not directly comparable to DBQR. The Mantel test tests the null hypothesis $H_0 : r_{x_{ij},y_{ij}} = 0$ and PERMANOVA tests if location parameter in the original data is different between groups.

4.2.1 Parameter Estimation Results

Table 1 shows simulation results of parameter estimation. The estimated $\hat{\boldsymbol{\beta}}$ is virtually unbiased. Their mean bias converges to zero as sample size increases at all quantiles considered. For example, the mean bias for $\beta_2(\tau = 0.1)$ decreases from 0.0213 to −0.00003 as sample size n increases from 30 to 200 in scenario I. The simpler bootstrap approach tends to over-estimate the standard errors at lower quantiles and under-estimate them at upper quantiles, especially when sample size is small. As the sample size n increases, the mean estimated standard errors of $\hat{\boldsymbol{\beta}}$ agree better with the empirical sample standard error. Correspondingly, the EmpCP is bigger at lower quantiles and smaller at upper quantiles than the true confidence level, 95%, when n is small. And it approaches the true confidence level as sample size n reaches 200 from both ends. The proposed method is robust to heteroscedasticity and the results under scenario II (lower panel, Table 1) are comparable to those of the homoscedastic scenario. The mean bias of the estimator converges to zero as sample size increases. The standard error of $\hat{\boldsymbol{\beta}}$ was over-/under-estimated at the lower/upper quantiles, and therefore, larger/smaller empirical coverage probability when n is small. As n increases, the performance improves and the results become satisfactory when n reaches 200 for models with normal random errors. Results for models with $\varepsilon_i \sim exp(1)$ have similar patterns and are included in the supplementary file (Table C.1).

Table 1 Parameter estimation results under scenario I—homoscedastic errors (upper panel), and scenario II—heteroscedastic errors, case 1 (lower panel)

τ	n	EmpBias($\times 10^3$)			EmpSD($\times 10^3$)			EstSD($\times 10^3$)			EmpCP (%)		
		$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$
		Scenario I, $\varepsilon_i \sim N(0, \sigma^2 = 0.5)$											
0.1	50	3.64	26.18	13.78	315.03	378.77	144.37	347.90	458.40	150.26	95.0	98.4	95.5
	100	10.16	15.90	−1.47	219.28	255.93	97.39	231.16	280.39	101.47	96.1	97.8	96.3
	200	2.47	14.29	−0.03	148.51	174.00	69.17	158.26	181.77	70.48	96.1	95.8	94.8
	400	2.58	1.82	−0.08	105.35	112.89	46.46	109.45	120.71	48.43	96.0	96.4	95.3
0.3	50	2.14	−1.61	4.45	275.78	324.02	124.67	289.71	378.91	133.07	95.5	98.3	95.8
	100	5.69	7.72	−4.94	190.93	211.07	85.91	194.80	232.79	88.34	95.4	97.7	95.9
	200	−0.59	10.98	−0.69	131.33	148.91	59.45	135.23	153.06	60.68	95.1	95.9	95.4
	400	1.49	−0.23	−1.19	92.90	98.45	41.20	94.78	103.90	42.08	95.1	96.3	95.8
0.5	50	0.15	−10.09	−0.29	267.13	311.49	119.70	277.15	355.36	125.96	96.1	98.2	95.4
	100	4.36	1.61	−6.74	184.40	202.97	82.81	188.40	223.04	85.11	94.8	97.0	94.3
	200	−1.89	9.68	−1.40	128.87	143.00	58.03	131.25	147.80	58.68	95.3	95.5	95.2
	400	1.06	−2.18	−1.51	91.54	97.59	40.52	92.08	101.08	40.76	94.8	96.1	95.5
0.7	50	−2.52	−21.33	−4.46	270.62	319.69	122.83	277.58	354.63	123.27	95.9	97.0	94.2
	100	1.81	−2.17	−8.24	189.39	209.43	84.26	191.35	225.58	84.85	94.6	97.2	94.3
	200	−3.52	7.03	−1.78	133.69	146.11	59.05	133.66	150.19	59.30	94.8	95.5	94.7
	400	0.95	−3.47	−2.24	94.45	101.96	41.90	94.32	103.30	41.47	94.0	95.0	94.5
0.9	50	−4.33	−44.30	−15.75	317.21	376.42	142.82	282.00	357.10	115.42	89.2	93.8	88.3
	100	−4.01	−6.77	−11.21	217.59	240.29	97.26	206.75	241.65	86.79	92.4	94.6	90.6
	200	−7.56	5.32	−2.99	157.66	170.35	67.75	149.41	166.75	63.94	93.1	95.4	92.7
	400	0.70	−6.47	−3.18	107.92	117.45	48.02	107.24	117.44	46.23	93.3	95.1	94.7

Table 1 (continued)

τ	n	EmpBias(×10^3)			EmpSD(×10^3)			EstSD(×10^3)			EmpCP (%)		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
		Scenario II, $\varepsilon_i \sim N(0, \sigma^2 = 0.5)$											
0.1	50	25.69	13.20	10.33	287.21	425.95	291.16	356.24	586.33	340.50	97.6	99.7	97.2
	100	3.04	18.04	8.20	191.82	266.88	198.90	221.71	335.45	217.91	97.5	98.7	95.2
	200	3.05	7.72	0.32	133.36	188.84	136.79	144.43	203.38	143.27	96.2	96.4	96.2
	400	4.49	−0.85	−0.95	95.12	124.09	94.80	98.01	132.49	97.16	94.7	95.7	95.5
0.3	50	20.01	−22.41	0.02	248.87	362.07	235.87	270.44	435.37	259.36	96.8	98.7	96.9
	100	3.37	6.30	2.91	167.29	226.99	163.38	178.65	261.05	168.64	95.5	97.3	94.5
	200	2.34	−3.23	0.72	114.91	161.57	111.74	121.41	167.49	113.88	95.8	95.8	95.6
	400	3.13	−1.56	1.13	82.44	109.79	77.07	83.67	113.21	78.75	94.8	95.7	95.9
0.5	50	9.03	−15.70	−2.31	243.79	351.21	226.21	254.39	404.50	232.90	95.1	98.2	95.0
	100	4.77	−2.41	−3.13	163.46	216.58	156.52	170.89	248.10	155.36	96.4	97.6	94.1
	200	1.91	−9.37	1.05	110.24	156.23	106.82	117.31	161.03	106.47	96.3	96.1	95.0
	400	1.90	−0.99	2.63	79.97	107.07	73.92	81.40	109.83	74.60	95.7	95.2	94.4
0.7	50	1.45	−24.59	−1.95	252.34	365.75	229.43	254.25	398.42	244.74	94.2	97.5	96.0
	100	3.72	−6.91	−4.29	169.02	219.61	162.38	172.39	248.74	163.87	95.3	97.1	94.4
	200	2.70	−16.40	0.13	112.88	158.32	111.57	119.54	163.66	111.67	95.4	95.9	95.4
	400	1.61	−2.14	2.34	82.09	110.58	76.68	83.41	111.75	78.10	95.9	94.4	94.9
0.9	50	−8.01	−52.90	0.26	288.57	413.91	281.73	253.68	393.83	250.88	90.1	94.5	91.3
	100	−3.23	−13.42	−7.19	199.83	258.83	196.65	184.82	262.60	182.88	92.5	96.5	91.9
	200	0.08	−21.51	−3.10	131.73	182.06	136.00	132.57	179.87	130.36	95.1	94.9	94.1
	400	1.58	−4.23	0.84	92.99	124.68	94.42	94.11	124.46	93.32	94.9	94.3	94.5

Table 2 Empirical type I error rate and power for testing the null hypothesis $H_0 : \beta_1(\tau) = 0$ under scenario I

n	τ										$p_{\max}$
	0.1		0.3		0.5		0.7		0.9		
	A	P	A	P	A	P	A	P	A	P	
	$\varepsilon \sim (N(0, 0.5))$, $\beta_1 = 0$										
50	0.016	0.048	0.017	0.048	0.018	0.049	0.030	0.058	0.062	0.055	0.052
100	0.022	0.052	0.023	0.047	0.030	0.045	0.028	0.049	0.054	0.049	0.048
200	0.042	0.061	0.041	0.055	0.045	0.054	0.045	0.045	0.046	0.046	0.052
400	0.044	0.055	0.041	0.051	0.039	0.047	0.040	0.048	0.044	0.051	0.051
	$\varepsilon \sim (N(0, 0.5))$, $\beta_1 = 1$										
50	0.689	0.799	0.792	0.869	0.811	0.887	0.794	0.870	0.739	0.740	0.839
100	0.950	0.982	0.986	0.994	0.990	0.995	0.988	0.995	0.957	0.972	0.994
200	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.997	1.000	1.000
	$\varepsilon \sim exp(1)$, $\beta_1 = 0$										
50	0.001	0.052	0.015	0.057	0.017	0.046	0.024	0.052	0.097	0.056	0.054
100	0.007	0.052	0.025	0.053	0.033	0.056	0.035	0.052	0.063	0.046	0.047
200	0.029	0.046	0.042	0.050	0.041	0.057	0.041	0.046	0.047	0.052	0.051
400	0.036	0.043	0.048	0.054	0.042	0.049	0.042	0.051	0.049	0.049	0.050
	$\varepsilon \sim exp(1)$, $\beta_1 = 1$										
50	0.917	0.992	0.820	0.905	0.628	0.726	0.447	0.512	0.327	0.267	0.351
100	0.977	0.995	0.954	0.966	0.872	0.901	0.697	0.744	0.450	0.430	0.574
200	1.000	1.000	1.000	1.000	0.998	0.998	0.969	0.970	0.706	0.703	0.911

A asymptotic; *P* permutation

4.2.2 Hypothesis Testing Results

Table 2 compares the empirical type I error rate and power for testing $H_0 : \beta_1(\tau) = 0$ at the nominal level 0.05 by using asymptotic normality (A) and the permutation test (P) for scenario I. As n increases, the empirical type I error rate of the asymptotic test approaches to the nominal level of 0.05 and gets around the nominal level when n reaches 200. However, when the sample size is small, p values calculated based on the asymptotic normality are conservative at lower quantiles and slightly liberal at upper quantiles. For example, when $n = 50$, the empirical type I error rates by asymptotic test are 0.016, 0.017, 0.018, and 0.030 for the (0.1, 0.3, 0.5, 0.7)th quantiles, and 0.062 for 0.9th quantile. The empirical type I error rate for the permutation test under scenario I is well controlled around the nominal level even when n is small. The asymptotic test has comparable power to the permutation test for large sample cases ($n >= 200$). But when n is small, the asymptotic test is less powerful. The empirical type I error rate of the permutation-based supremum test is also well controlled at around the nominal level regardless of sample size and quantile level. And its power increases as sample size increases. A similar pattern was observed in data with exponentially distributed random errors as shown in the lower panel of Table 2. However, an even larger sample size ($n = 400$) seems to be needed for the asymptotic test to have type I error around the nominal level and comparable power to the permutation test.

We examined the performance of testing $H_0 : \beta_2(\tau) = 0$ under scenario II, where $\beta_2(\tau)$ is τ dependent. The empirical type I error rate of the permutation test is greatly inflated in both cases as shown in Fig. 2 and Fig. I in Supplementary. While this may seem surprising at first, it should not. The permuting procedure not only breaks the association between the covariate of interest and the response variable, but also the underlying structure of heteroscedasticity. Therefore, the permuted test statistics are not samples from the true distribution under the null hypothesis. With normal random errors, the asymptotic test is still able to provide well-controlled type I error rate at different quantile levels, especially for the median (upper panel, Fig. 2). In case 2, where random errors are from a Cauchy distribution, our test is liberal at tail quantiles ($\tau = 0.1, 0.9$) but conservative at $\tau = 0.3, 0.5, 0.7$ compared to the normal and exponential errors with similar sample sizes (Table C.3-C.6 in supplementary). Although we do see a pattern of the empirical type I error rate approaches the nominal level as sample size increases, a sample size of 200 is still not large enough for satisfactory results. This indicates that a much larger sample size may be required to test the quantiles at two tails for heavy tail errors. The empirical type I error rates for case 1 in scenario II are summarized as in Fig. I (Supplementary). The permutation test has severely inflated type I error rate, and our method has well-controlled error rate with sufficiently large sample size. The empirical power of testing $H_0 : \beta_2(\tau) = 0$ with various effect sizes can be found in the supplementary (Tables C.3-C.6). Basically, as effect size and sample size increase, the empirical power of our test increases. Results for testing $H_0 : \beta_1(\tau) = 0$ under scenario II are very similar to that of Scenario I (supplementary, Table C.2).

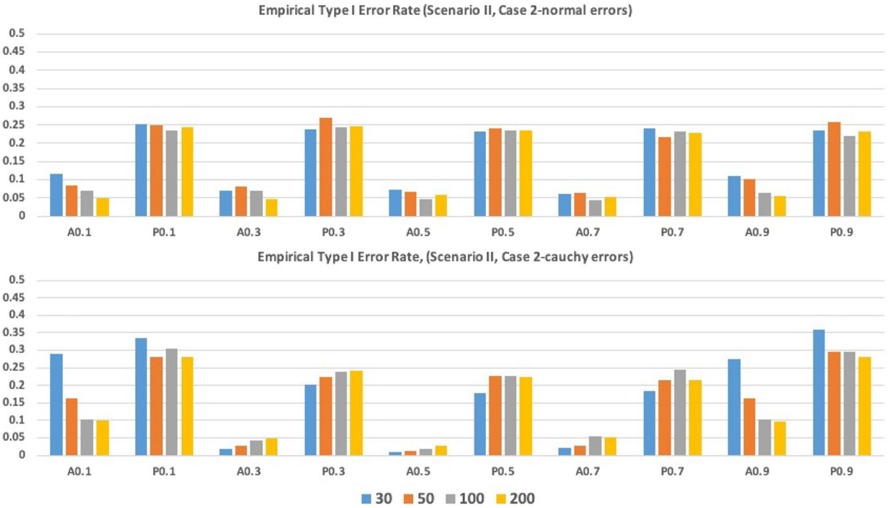


Fig. 2 Empirical type I error rate for testing the null hypothesis $H_0 : \beta_2(\tau) = 0$ under scenario II, case 2

Overall, a sufficiently large sample size is required to guarantee satisfactory performance of the asymptotic test. However, the permutation test fits in as a good back-up and provides controlled type I error rate and high power when sample size is small and there is no heteroscedasticity.

5 An Application Example

5.1 Association Between the Composition of Gut Microbiome and Inflammatory Bowel Diseases

Ulcerative colitis (UC) and Crohn's disease (CD), both of which fall under the umbrella of inflammatory bowel diseases (IBD), are chronic conditions of the gastrointestinal tract that affect several million individuals worldwide. While the exact causes of IBD are unknown, genetic, environmental, and more recently, microbial associations have been implicated over the years as potential contributors to the disease [14, 26, 33]. Various aspects of microbiome composition—here, the distribution of bacterial species in a subject's gut, have been found to be associated with disease status, including but not limited to an increased abundance of certain families of bacteria, such as the *Enterobacteriaceae*, a decreased abundance of other families of bacteria, such as the *Lachnospiraceae*, and in some cases an overall decreased diversity of bacteria in patients diagnosed with IBD [14]. It is arguable that disease status associates with a "re-set" of the whole microbial community, where the presence or abundance of many coordinating species varies and reaches a new equilibrium. Therefore, we look into pairwise distances between microbiome profiles and examine the association of the distribution of these distances and disease status, using our proposed method.

We downloaded species-level taxonomic profiles and metadata for UC, CD, and non-IBD subjects from the Integrative Human Microbiome Project website (https://ibdmdb.org/). This dataset consists of 1638 samples from 130 subjects. Multiple samples were collected from each subject and samples from each subject were assigned a time point relative to the first sample collected, on the interval of weeks over a period of about 57 weeks. Detailed information regarding study design, sample collection, and data preprocessing can be found in a recent publication [26]. Briefly, subjects were approached for enrollment into the study following routine colorectal cancer screening, suspected IBD, or other presentation of other gastrointestinal symptoms. Enrolled subjects were then subject to a colonoscopy, where IBD status was determined. We excluded samples that fail quality control and normalized the abundance of bacterial species in each sample to 100%. We conducted two independent analyses using the week 0 and week 8 samples, respectively. Both UC and CD are cyclical diseases where patients experience flares and periods of remission. Therefore, we do not expect over the course of the study for subjects to progressively develop more severe disease. Further, we do not expect that subjects' flares and remission periods are synced. However, clinical and biological markers of disease severity in this dataset were not well correlated (Lloyd-Price et al. 2019), and so determining which samples best represent those with severe disease or not is nontrivial, leading the authors to develop their own measure of dysbiosis (imbalance of the microbiome). The selection of time points reflect the presumption that at enrollment, subjects later diagnosed with IBD were in early stages of the disease, as they had never been diagnosed before. At week 8, these subjects would have been two months into a diagnosis, and while we would not expect all week 8 samples to represent a flare, we may expect at least more variability between IBD vs. non-IBD subjects. Our hypothesis is then that pairwise distances among week 8 samples are better at distinguishing disease groups than those of week 0 samples. An additional factor for time point selection was the relatively large sample sizes at these two time points: $n_0 = 91$ and $n_8 = 64$.

We considered six groups of pairwise distances: between non-IBD individuals (non-IBD *vs* non-IBD), non-IBD and UC patients (non-IBD *vs* UC), non-IBD and CD patients (non-IBD *vs* CD), UC patients (UC *vs* UC), UC and CD patients (UC *vs* CD), and CD patients (CD *vs* CD) (top panel, Fig. 3). We modeled the distribution of non-IBD *vs* non-IBD pairwise distances as the baseline, estimated, and tested how quantiles of pairwise distance in other five groups are different from the baseline by using our proposed quantile regression model. Specifically, by denoting "non-IBD" as group 0, "UC" as group 1, and "CD" as group 2, we define five dummy variables: $x_{01,ij}$, $x_{02,ij}$, $x_{11,ij}$, $x_{12,ij}$, $x_{22,ij}$ indicating membership of the five comparison groups. $x_{01,ij} = 1$ if ith sample is in group 0 and jth sample is in group 1; and $x_{01,ij} = 0$, otherwise. The other four dummy variables are similarly defined. y_{ij} is the pairwise distance between the ith and jth microbiome profiles. The conditional τth quantile of y_{ij} is modeled by $Q_\tau(y_{ij}|\boldsymbol{x}_{ij}) = \beta_0 + \beta_1(\tau)x_{01,ij} + \beta_2(\tau)x_{02,ij} + \beta_3(\tau)x_{11,ij} + \beta_4(\tau)x_{12,ij} + \beta_5(\tau)x_{22,ij}$. Here, $x^*_{ij} = \beta_1(\tau)x_{01,ij} + \beta_2(\tau)x_{02,ij} + \beta_3(\tau)x_{11,ij} + \beta_4(\tau)x_{12,ij} + \beta_5(\tau)x_{22,ij}$ can be viewed as a weighted pairwise distance between subjects i and j. $x^*_{ij} = \beta_k(\tau), k = 1, 2, \ldots, 5$ if two samples are in the kth comparison group that we considered. $\beta_k(\tau), k = 1, 2, \ldots, 5$ are the effect sizes of kth group on τth quantile of y_{ij}. We fitted quantile regression models at $\tau = (0.1, 0.3, 0.5, 0.7, 0.9)$ for the Bray-Curtis dissimilarity, UniFrac and weighted UniFrac distance. We tested the hypotheses $H_0 : \beta_k(\tau) = 0$ vs $H_a : \beta_k(\tau) \neq 0$ for $k = 1, \ldots, 5$, separately, and we compared our results with two other commonly used methods in microbiome association studies: the Mantel test [28] and PERMANOVA [1]. R functions: *partial.mantel.test* and *adonis* were used for these two tests.

5.2 Analysis Results

We observed significant difference in Bray-Curtis, weighted UniFrac, and unweighted UniFrac between the comparison groups at week 8. Pairwise Bray-Curtis dissimilarities observed between UC samples were found to be significantly greater than those between non-IBD samples ($\hat{\beta}_3(0.1) = 0.109, p = 0.003$; $\hat{\beta}_3(0.3) = 0.138, p = 0.001$; $\hat{\beta}_3(0.5) = 0.116, p = 0.018$; $\hat{\beta}_3(0.7) = 0.100, p = 0.035$; $\hat{\beta}_3(0.9) = 0.054, p = 0.091$). All effect sizes are estimated to be positive and $p_{3,\max} = 0.001$. Bray-Curtis dissimilarities between non-IBD and UC samples are also significantly greater than those

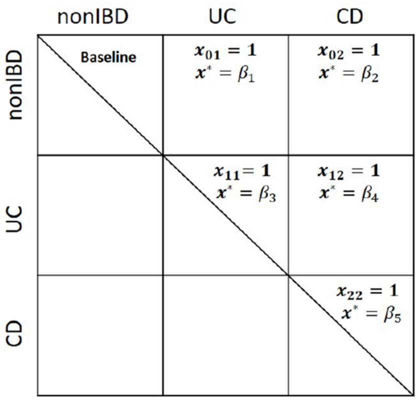
Fig. 3 Top: The defined six groups of pairwise distances: non-IBD vs non-IBD (baseline), non-IBD vs ▶ UC, non-IBD vs CD, UC vs UC, UC vs CD, and UC vs CD. Middle: Empirical quantiles of pairwise Bray-Curtis dissimilarity in the six groups. Within non-IBD individuals (black, solid dots), non-IBD vs UC (green, triangles), non-IBD vs CD (red, triangles), UC vs UC (blue, stars), UC vs CD (gray, circles), and CD vs CD (purple, stars) based on week 0 (left) and 8 (right) samples; Bottom: (left) Empirical quantiles of pairwise Bray-Curtis dissimilarity within non-IBD individuals (black, solid dots), non-IBD vs UC (green, triangles), and UC vs UC (blue, stars); (right) Empirical quantiles of pairwise Bray-Curtis dissimilarity between within-disease category (black, solid dots) and between-disease category (green triangles)

within non-IBD samples ($\hat{\beta}_1(0.1) = 0.066, p = 0.011$; $\hat{\beta}_1(0.3) = 0.070, p = 0.014$; $\hat{\beta}_1(0.5) = 0.059, p = 0.069$; $\hat{\beta}_1(0.7) = 0.069, p = 0.031$; $\hat{\beta}_1(0.9) = 0.034, p = 0.058$) and $p_{1,\max} = 0.010$. Bray-Curtis dissimilarities between UC samples, which are “within-group” measures, are also significantly greater than those between non-IBD and UC samples, which are “between-group” measures. This result is consistent with the biology of UC – given that week 8 does not represent any real biology in IBD subjects, there would be variability in disease severity at that time point among IBD subjects, but less so among non-IBD subjects.

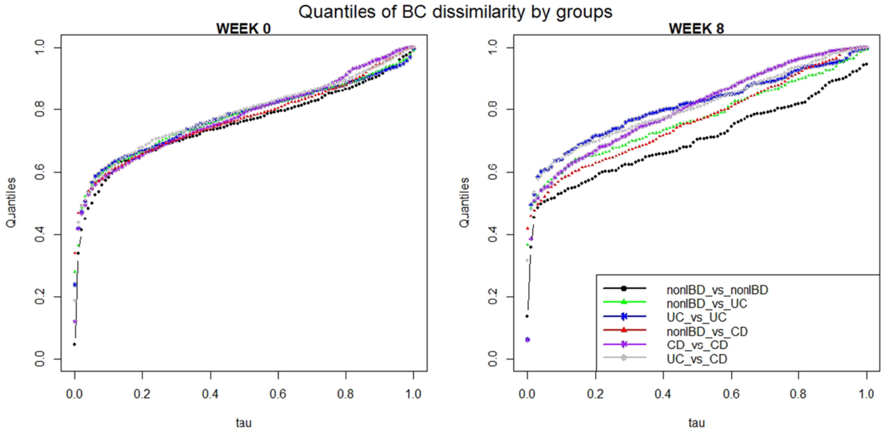
A similar pattern was observed when comparing Bray-Curtis dissimilarity in non-IBD *vs* CD and CD *vs* CD groups with the non-IBD *vs* non-IBD group. The Bray-Curtis dissimilarities between CD samples are the greatest, followed by the distances between non-IBD and CD samples, and the distances between non-IBD samples are the smallest (Table 3). The estimated effect sizes are about the same at all quantile levels. The distribution of UniFrac and weighted UniFrac distances tends to have heavier tails in the five groups compared to the baseline group. In particular, the upper tail quantiles of these distributions are significantly greater than in the baseline distribution. Besides, as the quantile level increases, the estimated effect sizes increase (for example, $\hat{\beta}_2(\tau) = (0.007, 0.021, 0.035, 0.056, 0.069)$; and $\hat{\beta}_5 = (0.031, 0.049, 0.070, 0.085, 0.109)$ for UniFrac distance, and $\hat{\beta}_1(\tau) = (0.004, 0.037, 0.066, 0.121, 0.285)$; and $\hat{\beta}_3 = (0.044, 0.079, 0.157, 0.283, 0.411)$ for weighted UniFrac distance). This finding agrees with empirical quantiles (Fig. 3 (middle-right panel) and Fig. II in the Supporting Material).

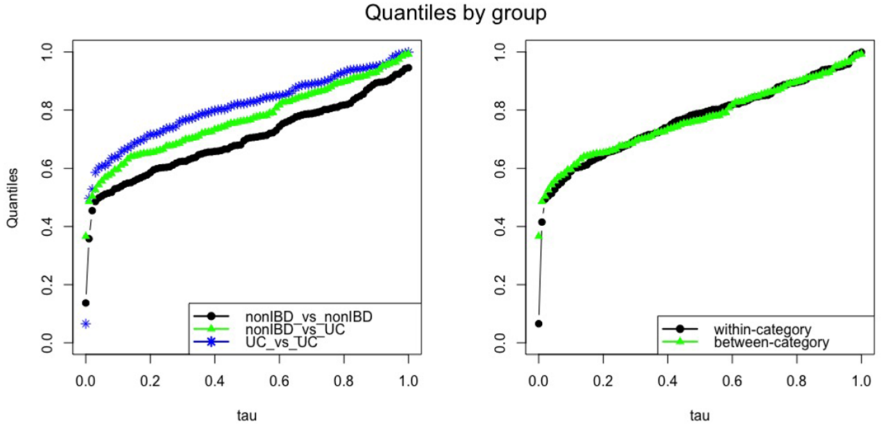
As we expected, quantiles of pairwise distances are not significantly different at week 0 (middle-left panel in Fig. 3 and Fig. II in the online supplementary file). Detailed model fitting and test results for week 0 samples are included in supplementary (Table C.7).

To lend additional support to our findings, we examined the dysbiotic status of week 0 and week 8 samples using the dysbiosis score defined by Lloyd-Price et al. [26].This score was developed by the authors as an independent measure of disease severity and is defined by a sample’s median Bray-Curtis dissimilarity to a set of non-IBD reference samples. This value was then compared to the respective non-IBD-to-reference distribution of Bray-Curtis dissimilarities. If the sample-to-reference value was in the 90th or greater percentile compared to the non-IBD-to-reference values, it was classified as a dysbiotic sample. Among all week 0 samples, 4 out of 24 non-IBD, 10 out 43 CD, and 6 out of 24 UC samples are determined to be dysbiotic, and there is no significant enrichment of dysbiotic samples in any disease



	nonIBD	UC	CD
nonIBD	Baseline	$x_{01} = 1$ $x^* = \beta_1$	$x_{02} = 1$ $x^* = \beta_2$
UC		$x_{11} = 1$ $x^* = \beta_3$	$x_{12} = 1$ $x^* = \beta_4$
CD			$x_{22} = 1$ $x^* = \beta_5$





category (Person’s Chi-square test, *pvalue* = 0.8067). There are 2 out of 13 non-IBD, 10 out 35 CD, and 6 out of 16 UC microbiome profiles are designated as dysbiotic at week 8. More CD samples are determined to be dysbiotic, though it is not statistically significant (Person’s Chi-square test, *pvalue* = 0.4218). We compared

Table 3 Estimate and testing p value of coefficients in $\tau = (0.1, 0.3, 0.5, 0.7, 0.9)$th quantile regression models for three pairwise distance measures: Bray-Curtis (Bray-Curtis), UniFrac distance (UniFrac), and weighted UniFrac distance (wUniFrac) based on 8-th week samples ($n = 64$). p values by the Mantle test and PERMANOVA are also include

Method	τ										$p_{\max}$
	0.1		0.3		0.5		0.7		0.9		
	$\hat{\beta}(\tau)$	p	$\hat{\beta}(\tau)$	p	$\hat{\beta}(\tau)$	p	$\hat{\beta}(\tau)$	p	$\hat{\beta}(\tau)$	p	
	Bray-Curtis (BC)										
x_{01}	0.066	0.011	0.070	0.014	0.059	0.069	0.069	0.031	0.034	0.058	0.010
x_{02}	0.047	0.030	0.045	0.057	0.061	0.031	0.074	0.005	0.066	0.001	0.009
x_{11}	0.109	0.003	0.138	0.001	0.116	0.018	0.100	0.035	0.054	0.091	0.001
x_{12}	0.113	0.001	0.114	0.001	0.103	0.001	0.107	0.001	0.085	0.001	0.001
x_{22}	0.071	0.001	0.098	0.001	0.119	0.001	0.135	0.001	0.094	0.001	0.055
Mantel	non-IBD vs UC, $p = 0.707$										
	non-IBD vs CD, $p = 0.193$										
PERMANOVA	non-IBD vs UC, $p = 0.108$										
	non-IBD vs CD, $p = 0.133$										
	UniFrac Distance										
x_{01}	−0.020	0.361	0.012	0.564	0.042	0.031	0.078	0.001	0.120	0.001	0.001
x_{02}	0.007	0.728	0.021	0.216	0.035	0.034	0.056	0.001	0.069	0.001	0.001
x_{11}	−0.011	0.671	0.049	0.088	0.085	0.003	0.143	0.001	0.157	0.001	0.001
x_{12}	0.003	0.852	0.044	0.005	0.076	0.001	0.098	0.001	0.128	0.001	0.001
x_{22}	0.031	0.158	0.049	0.010	0.070	0.001	0.085	0.001	0.109	0.001	0.001
Mantel	non-IBD vs UC, $p = 0.563$										
	non-IBD vs CD, $p = 0.089$										
PERMANOVA	non-IBD vs UC, $p = 0.225$										
	non-IBD vs CD, $p = 0.203$										
	wUniFrac Distance										
x_{01}	0.004	0.809	0.037	0.326	0.066	0.252	0.121	0.113	0.285	0.001	0.001
x_{02}	– 0.005	0.688	0.015	0.601	0.062	0.230	0.138	0.058	0.288	0.001	0.002
x_{11}	0.044	0.150	0.079	0.132	0.157	0.057	0.283	0.004	0.411	0.002	0.001
x_{12}	0.027	0.057	0.076	0.010	0.156	0.001	0.254	0.001	0.371	0.001	0.001
x_{22}	0.010	0.546	0.074	0.060	0.183	0.004	0.250	0.002	0.341	0.001	0.001
Mantel	non-IBD vs UC, $p = 0.299$										
	non-IBD vs CD, $p = 0.023$										
PERMANOVA	non-IBD vs UC, $p = 0.205$										
	non-IBD vs CD, $p = 0.196$										

the dysbiosis scores of all samples at the two time points using side-by-side box plots and observed that the dysbiosis score of non-IBD samples are much lower than that of UC and CD samples at week 8 than at week 0 (Supplementary, Fig. III).

However, both Mantel test and PERMANOVA fail to detect significant differences, possibly because these methods assume that the distances between individuals from a same disease category (for example, non-IBD *vs* non-IBD, UC *vs* UC, CD *vs* CD) have a similar underlying distribution, which is an assumption that obviously does not hold in these data. The bottom-left panel of Fig. 3 shows percentiles of Bray-Curtis dissimilarities in the UC *vs* UC group are the greatest (blue stars), followed by that in the non-IBD *vs* UC group (green triangles), percentiles of Bray-Curtis dissimilarities in the non-IBD *vs* non-IBD group are the smallest (black solid dots). If we ignore this underlying difference and exam the within- and between-disease category pairwise dissimilarities by merging Bray-Curtis dissimilarities in UC *vs* UC group to the non-IBD *vs* non-IBD group, the difference is completely masked (bottom-right panel, Fig. 3).

To sum up, our DBQR is more flexible in modeling and can provide detailed information about the association of a pairwise distance matrix with covariates of interest by providing estimates of effect sizes and comparisons across multiple quantile levels. In spite of the interesting findings, we acknowledge that only samples at two time points were used in our analysis to study the association and, therefore, illustrate the utility of our method in microbiome association studies. However, given that for the IBD cohort, the second time point likely reflected a mixture of more severe and less severe cases, the fact that our method could detect a significant difference among non-IBD and both UC and CD cohorts when a classic PERMANOVA could not is promising. For a complete understanding of the relationship between gut microbiome and IBD, a more in-depth analysis of the full dataset is warranted. Ideally, additional analyses would include relevant clinical metadata regarding clinically evaluated disease status at each time point, as both Crohn's disease and ulcerative colitis are chronic, and cyclical, conditions. A comprehensive listing of these types of clinical variables in this dataset is missing or highly sparse, and therefore, such an analysis was not possible. An additional algorithmic development could possibly to take into account the temporal structure of the samples in this study in addition to disease status.

6 Discussion

The association of environmental and disease covariates with pairwise distance matrices appears frequently in some fields of study—traditionally in ecology, where scientists have long used diversity metrics to summarize high-dimensional species abundance data, and more recently in studies of the microbiome, where scientists have adopted many of these ecological approaches to compare samples which contain information on hundreds and potentially thousands of species. While in some cases, this approach has been driven by necessity to reduce high-dimensional data, from a biological perspective the community structure of an ecosystem as a whole may hold answers that presence or abundance of individual species may not. The biological need for methods which can adequately model such associations is clear, and several methods have been adopted by the community, most of which are based on permutation testing. However, as the amount of data grows and the questions asked expand, the need for additional methods still stands.

In this work, we propose a quantile regression model for distance matrices, which uses pairwise distances between the original observations to study the association between these distances and other factors of interest, whether they be environmental variables, such as spatial distance between samples to each other, or clinical variables such as disease status. We derive asymptotic consistency and normality of the proposed estimator by incorporating the theories for U-statistics into the framework of a classical quantile regression model. We also propose a procedure to estimate the asymptotic covariance matrix for statistical inference. Results of numerical simulations under various settings suggest satisfactory performance of our proposed method, especially with a median to large sample size: empirical coverage probability is close to the true level, the empirical type I error rate is well controlled at around the nominal level, and empirical power is also comparable to that of the permutation-based approach. Importantly, in the case of heteroscedasticity where the empirical type I error rate of permutation test inflated severely, our approach still maintains the error rate around the nominal level.

The proposed method is computationally faster than a permutation-based approach. The simpler bootstrap used for the estimation of the covariance matrix in our method only requires $[n/2]$ data points to fit the quantile regression model while the permutation approach requires $\binom{n}{2}$ data points. Further, the linear programming used for fitting quantile regression can be very slow and memory intensive when the number of data points reaches the thousands. This shows the merits of our approach in large-scale studies with decent sample sizes.

Although the paper considers distances that measure dissimilarity, the concept of distance can be relaxed to a more general non-negative symmetric kernel function, $y_{ij} = s(y_i, y_j) = y_{ji}$. For example, in our application example, incorporating multiple dummy variables indexing group membership can be viewed as a "weighted" distance $x^*_{ij} = \beta_1 x_{ij,1} + \beta_2 x_{ij,2} + \beta_3 x_{ij,3} + \beta_4 x_{ij,4} + \beta_5 x_{ij,5}$ in a way that each individual effect β_k can be estimated and tested.

There are a few additional notes worth mentioning. Although the choice of distance measurement generally depends on the nature of the data as well as the objective of the study, it is possible that an optimal metric exists. We provide an overall conclusion that as long as the selected pairwise distance measure satisfies the fundamental assumptions for the large sample properties of U-statistics, the asymptotic results are valid. The selection of the optimal distance measure is beyond the scope of this work, but it certainly is an important concept we continue to pursue. Meanwhile, a drawback of distance-based analysis for microbiome analysis is that they cannot identify the individual contribution of species, which could be of great interest in real studies. While permutation tests seem straightforward and work well when sample sizes are small, they cannot account for heteroscedasticity, and indeed popular tests such as the PERMANOVA often suffer from the inability to distinguish significance due to differences in mean and significances due to differences in variance among samples being compared.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12561-021-09306-6.

Acknowledgements Li's research was partially supported by Junior Faculty Development Award from University of North Carolina at Charlotte. Sun's research was partially supported by NIH grant R37AI054165, NSF grant DMS1513072, and DMS1915829.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Anderson MJ (2008) A new method for non-parametric multivariate analysis of variance. Austral Ecol 26(1):32–46
2. Anderson MJ (November 2017) Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online, pp 32–46
3. Bray JR, Curtis JT (1957) An ordination of upland forest communities of Southern Wisconsin. Ecol Monogr 27:325–349
4. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28(16):2106–2113
5. Cuadras C, Arenas A (1990) A distance based regression model for prediction with mixed data. Commun Stat 19:2261–2279
6. Faraway JJ (2013) Regression for non-Euclidean data using distance matrices. J Appl Stat 41(11):2342–2357
7. Frankel AE, Deshmukh S, Reddy A, Lightcap J, Hayes M, McClellan S, Singh S, Rabideau B, Glover TG, Roberts B, Koh AY, (Jan-Dec, (2019) Cancer immune checkpoint inhibitor therapy and the gut microbiota. Integr Cancer Ther. https://doi.org/10.1177/1534735419846379
8. Ghose C (2013) Clostridium difficile infection in the twenty-first century. Emerg Microbios Infect 2:62
9. Hoeffding W (1961) The strong law of large numbers for u-statistics. Technical report
10. Honoré BE, Hu L (2017) Simpler bootstrap estimation of the asymptotic variance of U-statistic-based estimators. Econ J 21(1):1–10
11. Honoré BE, Powell JL (1994) Pairwise difference estimators of censored and truncated regression models. J Econ 64:241–278
12. Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101
13. Huber P (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, vol 1, pp 221–233
14. Huttenhower C, Kostic AD, Xavier RJ (2014) Inflammatory bowel disease as a model for translating the microbiome. Immunity 40(6):843–854
15. Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc vaudoise sci nat 44:223–270
16. Koenker R, Gilbert Bassett J (1978) Regression quantiles. Econometrica 46(1):33–50
17. Koenker R, Gilbert Bassett J (1982) Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50(1):43–61
18. Koenker R, Hallock KF (2001) Quantile regression. J Econ Perspect 15(4):143–156
19. Koenker R, Machado JAF (1999) Goodness of fit and related inference processes for quantile regression. J Am Stat Assoc 94(448):1296–1310
20. Lancaster P, Rodman L (1995) Algebraic riccati equations. Oxford University Press, Oxford
21. Laub A (1979) A schur method for solving algebraic riccati equations. IEEE Trans Autom Control 24:913–921. https://doi.org/10.1109/CDC.1978.267893
22. Legendre P, Legendre L (1998) Numerical Ecology. Developments in Environmental Modelling. Elsevier Science, ISBN 9780080537870. https://books.google.com/books?id=KBoHuoNRO5MC

23. Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, Dunn JR, Farley MM, Holzbauer SM, Meek JI, Phipps EC, Wilson LE, Winston LG, Cohen JA, Limbago BM, Fridkin SK, Gerding DN, McDonald LC (2015) Burden of clostridium difficile infection in the united states. N Engl J Med 372(9):825–834. https://doi.org/10.1056/NEJMoa1408913 (**PMID: 25714160**)
24. Li SY, Cui YH (2012) Gene-centric gene-gene interaction: A model-based kernel machine method. Ann. Appl. Stat. 6(3):1134–1161
25. Lichstein JW (2007) Multiple regression on distance matrices: a multivariate spatial analysis tool. Plant Ecol 188(2):117–131
26. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber TG, Hall AB, Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J, Shungin D, Vázquez-Baeza Y, White RA, Investigators I, Braun J, Denson LA, Jansson JK, Knight R, Kugathasan S, McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS, Clish CB, Franzosa EA, Vlamakis H, Xavier RJ, Huttenhower C (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569(7758):655–662
27. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appd Environ Microbiol 71(12):8228–8235
28. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27(2):209–220
29. McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82:290–297
30. McArtor DB, Lubke GH, Bergeman CS (2016) Extending multivariate distance matrix regression with an effect size measure and the asymptotic null distribution of the test statistic. Psychometrika 82(4):1052–1077
31. Minas C, Montana G (2014) Distance-based analysis of variance: approximate inference. Stat Anal Data Min 7(6):450–470
32. Minas C, Waddell SJ, Montana G (2011) Distance-based differential analysis of gene curves. Bioinformatics 27(22):3135–3141
33. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13:117–126. https://doi.org/10.1186/gb-2012-13-9-r79 (**ISSN 1465-6906**)
34. Shono Y, Docampo M, Peled J, Perobelli SM, Velardi E, Tsai J, Slingerland A, Smith OM, Young LF, Gupta J, Lieberman SR, Jay H, Ahr KF, Porosnicu Rodriguez KA, Xu K, Calarfiore M, Poeck H, Caballero S, Devlin SM, Rapaport F, Dudakov JA, Hanash AM, Gyurkocza B, Murphy GF, Gomes C, Liu C, Moss EL, Falconer SB, Bhatt AS, Taur Y, Pamer EG, van den Brink MRM, R JR (2016) Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. Sci Trans Med 8(339):339ra71. https://doi.org/10.1126/scitranslmed.aaf2311
35. Wessel J, Schork NJ (2006) Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis. Am J Hum Genet 79(5):792–806
36. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics 89(1):82–93