Identifying unmeasured heterogeneity in microbiome data via quantile thresholding (QuanT)

Jiuyao Lu[1], Glen A. Satten[2], Katie A. Meyer[3], Lenore J. Launer[4], Wodan Ling[5], Ni Zhao[6,*]

Corresponding Authors: Ni Zhao, nzhao10@jhu.edu

[1]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, 265 South 37th Street, 19104 PA, USA. jiuyaolu@wharton.upenn.edu

[2]Department of Gynecology and Obstetrics, Emory University School of Medicine, 101 Woodruff Circle, 30322 Atlanta, USA. gsatten@emory.edu

[3]Nutrition Research Institute and Department of Nutrition, University of North Carolina, 500 Laureate Way, 28081 Kannapolis, USA. ktmeyer@email.unc.edu

[4]Laboratory of Epidemiology and Population Science, NIA, NIH, 7201 Wisconsin Ave, 20814 Bethesda, USA. lenore_launer@nih.gov

[5]Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, 402 East 67th Street, 10065 New York, USA. wol4002@med.cornell.edu

[6]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, 21205 Baltimore, USA. nzhao10@jhu.edu

Abstract

Background: Microbiome data, like other high-throughput data, suffer from technical heterogeneity stemming from differential experimental designs and processing. In addition to measured artifacts such as batch effects, there is heterogeneity due to unknown or unmeasured

factors, which lead to spurious conclusions if unaccounted for. With the advent of large-scale multi-center microbiome studies and the increasing availability of public datasets, this issue becomes more pronounced. Current approaches for addressing unmeasured heterogeneity in high-throughput data were developed for microarray and/or RNA sequencing data. They cannot accommodate the unique characteristics of microbiome data such as sparsity and over-dispersion.

Results: Here, we introduce Quantile Thresholding (QuanT), a novel non-parametric approach for identifying unmeasured heterogeneity tailored to microbiome data. QuanT applies quantile regression across multiple quantile levels to threshold the microbiome abundance data and uncovers latent heterogeneity using thresholded binary residual matrices. We validated QuanT using both synthetic and real microbiome datasets, demonstrating its superiority in capturing and mitigating heterogeneity and improving the accuracy of downstream analyses, such as prediction analysis, differential abundance tests, and community-level diversity evaluations.

Conclusions: We present QuanT, a novel tool for comprehensive identification of unmeasured heterogeneity in microbiome data. QuanT's distinct non-parametric method markedly enhances downstream analyses, serving as a valuable tool for data integration and comprehensive analysis in microbiome research.

Keywords: microbiome data, unmeasured heterogeneity, batch effects, zero inflation, conditional quantile regression

Background

Next-generation sequencing technology has facilitated large-scale microbiome profiling to explore associations with various disease and health outcomes. Given the large scale, samples must be collected and profiled in different runs or even from various studies. Furthermore,

there has been a significant increase in the availability of publicly accessible microbiome data[1-3], and integrating these diverse datasets can be invaluable in advancing scientific discoveries[4, 5]. When integrating these diverse datasets, it is crucial for researchers to recognize potential variations within these datasets, which might not be immediately apparent to external researchers. Much of this unwanted variation is attributable to hidden technical or biomedical heterogeneity, such as instrumental drift, undocumented differences in study design, or even lab labeling errors. Failing to account for these hidden heterogeneities can obscure the true biological signals and lead to spurious findings. The situation can be even worse when including publicly available data[1-3], because sources of variability such as batch labels may not be included in the distribution data.

In genomic studies, numerous statistical methods have been developed to correct for batch effects when batch indicators are available, with ComBat being among the most widely adopted. In the microbiome field, several methods have also been proposed to address batch effects when indicators are known, including approaches such as MMUPHin[6] and DEBIAS-M[7] and ConQuR[8], among others. These approaches rely on the availability of batch indicators. In contrast, no methods are designed to address *unmeasured/unknown* heterogeneity in microbiome studies, although techniques developed for genomic data, such as surrogate variable analysis (SVA)[9] and remove unwanted variation (RUV)[10], may be applicable. SVA is a two-step procedure for identifying surrogate variables (detection of hidden factors and then construction of surrogate variables), a general framework that was adapted by subsequent methods, including ours here. RUV employs factor analysis on control genes (via spike-in experiment or housekeeping gene estimation) to identify hidden heterogeneity. However, SVA and RUV are not ideally suited for microbiome data. First, they do not address the large number of zeros and the over-dispersion in microbiome data. Second, they only detect the underlying

structure in the (possibly log-transformed) mean expression while ignoring other possible changes in distribution. Third, RUV requires users to specify the number of hidden directions: incorrect specification of these parameters may result in false positive or negative discoveries.

Here, we present the Quantile Thresholding (QuanT) approach, the first tool for comprehensive identification of unmeasured heterogeneity in microbiome data. QuanT applies quantile regression[11] at multiple quantile levels to taxa relative abundances, and then forms dichotomized residuals. QuanT distills information from these dichotomized residuals to output a set of unit vectors that represent the hidden heterogeneity. As a non-parametric approach, QuanT is robust to the complex distributional attributes of microbiome data, handles sparseness in a natural way, and can detect heterogeneity beyond simple mean shifts. The quantile surrogate variables (QSVs) output by QuanT can be easily incorporated into downstream analysis. QSVs can be treated as known variables and embedded into any existing analytic workflows (e.g., regression models) that allow for covariate adjustment. Alternatively, QuanT can be combined with the recently developed microbiome batch effect removal method, ConQuR[8], to produce batch-free microbial profiles for downstream analysis.

Methods

Overview of QuanT

We first give an overview of the method before delving into a detailed explanation of each step. We propose a two-stage algorithm, depicted in Fig. 1, to identify unmeasured heterogeneity in microbiome relative abundance. Similar to existing methods such as SVA, our algorithm, QuanT, first employs matrix decomposition to detect hidden directions pertaining to unmeasured factors in the data, followed by a second stage to refine these detected directions. In the first stage, we fit quantile regression models at multiple quantile levels to the relative

abundance of each taxon with the primary variable of interest and important covariates, and then generate dichotomized, centered quantile residuals. These columns of residuals are concatenated into a single matrix of discretized residuals R. We then conduct a singular value decomposition (SVD) of R to uncover the systematic structure of the residuals and select singular vectors based on their directional information and variability. These selected vectors, termed stage-1 singular vectors, capture systematic variation independent of the known variables.

The stage-1 singular vectors are unrelated to the known variables as these known variables were included as covariates in the quantile regression models. As a result, adjusting for stage-1 singular vectors in downstream analysis would not correct for the confounding effects of the unmeasured heterogeneity (although efficiency may be improved). Therefore, in the second stage, we perform a "correlation calibration" step, in which we weight the "features" (i.e., combinations of quantile level and taxon). Specifically, we upweight features that are strongly marginally associated with the stage-1 singular vectors but weakly associated with the covariates, and downweight those with the opposite characteristics. Our second stage is an iterative procedure such that in each iteration, we fit quantile regression models and decompose a matrix of discretized residuals, similar to the first stage, but using the weighted features for model fitting. Each iteration will yield new "iterative" singular vectors (iSVs) by decomposing the residual matrix. The iSVs at convergence, i.e., when the resulting iSVs are close enough to the iSVs in the previous iteration, are called Quantile Surrogate Variables (QSVs), which are the output of our algorithm. This kind of correlation calibration step has been shown to be necessary in similar tasks when analyzing gene expression data[9, 12-15].

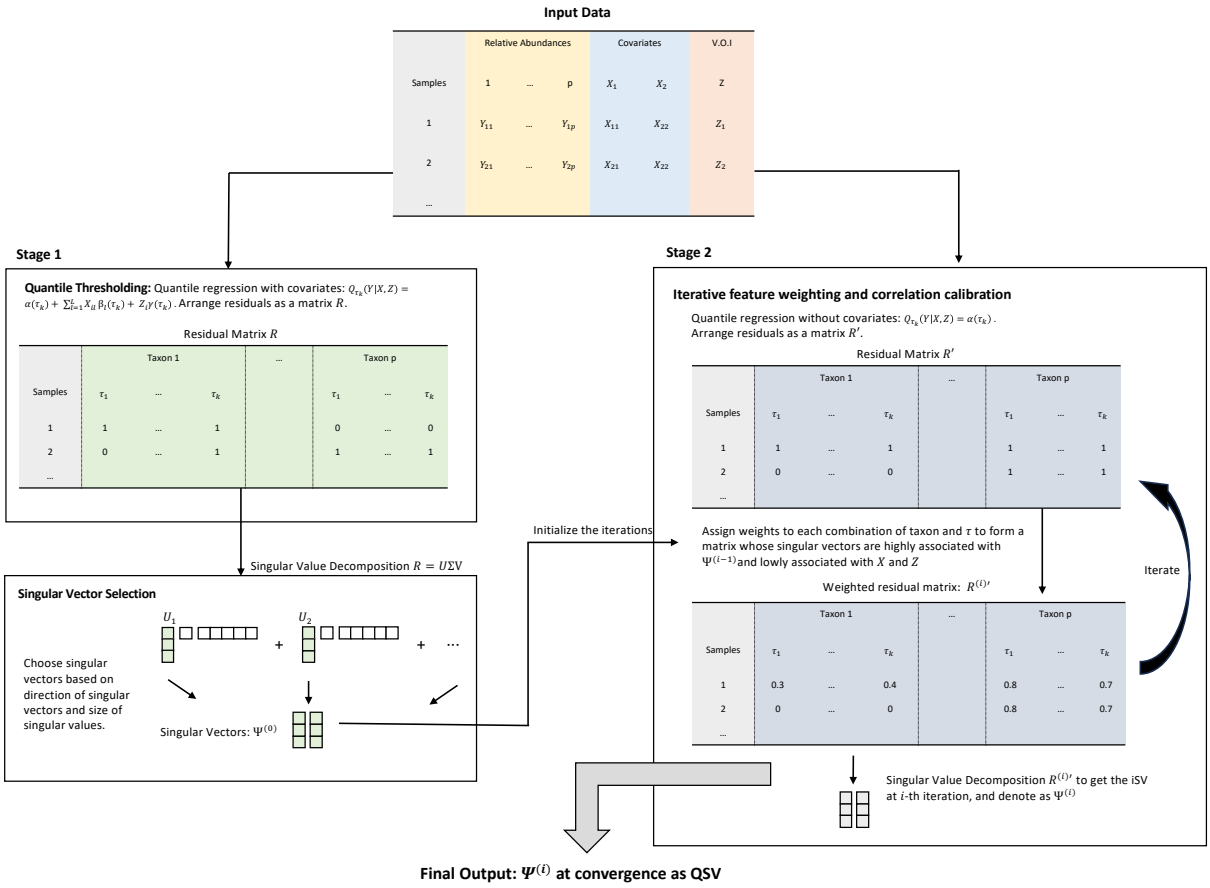


Fig. 1 | Flowchart of QuanT. V.O.I: variable of interest; QSV: quantile surrogate variable; $\tau_1 \ldots \tau_k$: quantile levels. Stage 1 conducts quantile regression for each taxon at multiple quantile levels using V.O.I (Z) and covariates (X). Residuals are thresholded to create a binary matrix R, which is analyzed through singular value decomposition to identify key directions. Stage 2 implements iterative feature weighting and correlation calibration. Stage 2 starts by fitting quantile regression and thresholding without covariates at multiple quantile levels, followed by the weighting of taxon and quantile levels so that the singular vectors of the weighted residual matrix $R^{(u)\prime}$ are highly associated with the prior stage singular vectors but minimally associated with X and Z. The singular vectors at convergence of Stage 2 are QSVs, the output of QuanT.

Comparing to the existing approaches such as SVA and RUV, QuanT offers several novel modifications/improvements to make it suitable for microbiome data, which is typically sparser and over-dispersed. First, by using quantile regressions instead of least squares or generalized

linear models, QuanT can gather information from the whole distribution of the taxa rather than just the mean, which is critical for microbiome data. Secondly, SVA and RUV search for significant directions by identifying singular vectors that represent a disproportionately high amount of variation. QuanT, on the other hand, relies on both the direction of a singular vector and the variance it explains, and thus allows us to detect significant directions that may harbor important underlying structure but explain only a modest proportion of variation.

We now describe our notation and proceed to a more in-depth explanation of the method's components.

Notation

Assume the relative abundance data form a $n \times p$ matrix Y, in which Y_{ij} is the relative abundance of the j-th taxon in the i-th sample. We denote by $Y_{\cdot j}$ the j-th column of Y comprising data on the j-th taxon, and treat it as the outcome in the proposed regression framework. Suppose that in addition each sample i has a primary variable of interest (denoted by Z_i) and a set of L covariates (denoted by X_i) to adjust for. We use quantile regression at each of K quantile levels in our regression framework; these levels are represented by the vector, $\vec{\tau}$. By default, we set the quantile levels as an equally spaced sequence from 0.1 to 0.9 with a step size of 0.005 corresponding to K=161.

Note that some taxon-quantile combinations may not convey useful information. For example, if 70% of the observations for a specific taxon j are zeros, any quantile levels $\vec{\tau}$ less than 70% will be irrelevant and are thus disregarded for this taxon. Thus, the effective length of $\vec{\tau}$ varies by taxon. This approach makes $\vec{\tau}$ data adaptive, utilizing more quantile levels for taxa with more informative distributions.

Stage 1: Obtaining singular vector directions that capture the underlying heterogeneity structure.

Quantile thresholding

For each feasible pair of taxon j and quantile level τ_k, we fit the quantile regression model $Q_{\tau_k}(Y_{\cdot j}|X,Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) + Z_i\gamma(\tau_k)$ to the data $Y_{\cdot j}$ in which $Q_{\tau_k}(Y|X,Z) = \inf\{y: F_{Y|X,Z}(y) > \tau_k)\}$ is the τ_k -th quantile function. Parameters $(\hat{\alpha}_j(\tau_k), \hat{\beta}_{lj}(\tau_k), \hat{\gamma}_j(\tau_k))$ are obtained by solving the minimization problem

$$(\hat{\alpha}_j(\tau_k), \hat{\beta}_{lj}(\tau_k), \hat{\gamma}_j(\tau_k)) = \arg\min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \rho_{\tau_k}\left(Y_{ij} - \alpha(\tau_k) - \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) - Z_i\gamma(\tau_k)\right),$$

where ρ_{τ_k} is the check function[16] $\rho_{\tau_k(x)} = \tau_k x - I\{x < 0\}x$.

After fitting the quantile regression model, we threshold the microbial abundances against the estimated quantiles and center them as $I\{Y_{ij} > \hat{\alpha}_j(\tau_k) + \sum_{l=1}^{L} X_{il}\,\hat{\beta}_{lj}(\tau_k) + Z_i\hat{\gamma}_j(\tau_k)\} - (1-\tau_k)$. For each pair of taxon j and quantile level τ_k, this yields a n-dimensional vector of discretized (binary) centered residuals. We gather the results from different pairs into a matrix R with n rows (samples) and no more than pK columns (processed features) as we omit the uninformative pairs of taxon and quantile level. R captures the heterogeneity structure such as the residual matrix constructed by SVA but is more robust and distribution-free and handles zero cells in a natural way.

Singular vector selection

We conduct SVD of the binarized and centered residual matrix R to obtain

$$R = U\Sigma V.$$

We aim at selecting a subset of the singular vectors, i.e., columns of U, to represent the important structure in the data. We use an integrated approach to select the singular vectors. Specifically, we select singular vectors 1) whose distributions are statistically different from the marginal distribution of the singular vectors if there is no underlying structure in the binary centered residual matrix (a.k.a., complete null, details are discussed in the next paragraph), and 2) which explain a large proportion of variability of the binary centered residual matrix. We use the SelectiveSeqStep procedure in the knockoff paper[17] to achieve this goal based on the two criteria simultaneously. This dual goal selection is a novel feature of QuanT. The detailed rationale and procedure are as follows.

Under the complete null that there is no underlying structure in the binary centered residual matrix, each singular vector direction should point to a uniformly distributed point on the n-dimensional sphere; this implies that the components of each singular vector should follow a Gaussian distribution[18]. Thus, we can test whether a singular vector is an important direction by testing the normality of each singular vector direction using the Lilliefors test[19] or the Anderson-Darling test[20] or the Shapiro-Francia test[21] and obtain its corresponding p-value. We also order the singular vectors based on the size of their corresponding singular values. We aim to select from the top m singular vectors which have p-values smaller than a cutoff c (e.g., 0.05). The choice of m is determined by controlling the false discovery rate (denoted by q) of the selected singular vectors. Under the null hypothesis, p-value is uniformly distributed; thus the ratio between the number of singular vectors with p-values greater than c and those with p-values less than c should be $\frac{1-c}{c}$. Roughly speaking, the singular vectors with p-values > c can be considered generally null, and therefore, we can estimate the number of falsely

discovered singular vectors based on them, which is $\#\{j \leq m: p_j > c\}/\frac{1-c}{c}$, where # represents the size of a set. This is similar to the idea of q-value[22-24], a commonly used approach for multiple testing problems. The proportion of falsely discovered singular vectors, if we select the top m singular vectors, are estimated by $\frac{\#\{j \leq m: p_j > c\}}{\#\{j \leq m: p_j \leq c\} \vee 1} / \frac{1-c}{c}$. We pick the m such that this proportion is controlled below a false discovery rate threshold q if the top m singular vectors are selected and would exceed q if the top $m+1$ singular vectors were selected:

$$\hat{m} = \min\left\{m: \frac{\#\{j \leq m: p_j > c\}}{\#\{j \leq m: p_j \leq c\} \vee 1} / \frac{1-c}{c} > q\right\} - 1.$$

After determining $\hat{m}$, we select the singular vectors from the top $\hat{m}$ singular vectors with p-values smaller than c. Note that our approach allows for a non-monotonic selection of singular vectors, yet primarily focus on those that account for a substantial portion of variability in the data.

We define Φ to be the matrix with the s-th column, denoted as Φ_s, equal to the s-th stage-1 singular vector.

Stage 2: Iterative feature weighting and correlation calibration

Similar to SVA[9, 12-15], we conduct a second stage to refine the stage-1 singular vectors that we call correlation calibration. The reason is that, by regressing out X and Z to construct the binarized centered residuals in the quantile regression stage, the selected singular vectors are, loosely speaking, uncorrelated with the primary variable Z and covariates X (if we had used linear regression, this statement would be exact). Thus, they are not able to capture unmeasured confounding effects, as confounders must be correlated with Z as well. Nonetheless, it is the unmeasured confounders that pose the greatest threat to the integrity of a study.

Because the stage-1 singular vectors represent the important features of the full set of residuals R, we seek to reweight the columns of R that will allow us to identify directions that not only reflect the underlying data structure but also maintain potential correlations with X and Z, thereby enhancing the robustness of downstream analysis against hidden confounding factors.

To accomplish this, we first conduct a similar thresholding procedure as in stage 1, but on a reduced quantile regression model $Q_{\tau_k}(Y_{\cdot j}) = \alpha_j(\tau_k)$, i.e., without regressing X and Z out, and construct the binarized and centered residuals (denoted by R'). We then weight columns (combinations of taxon and quantile levels) of R' to prioritize those that represent underlying data structure while being minimally associated with X and Z. Specifically, we use an iterative procedure that balances the two goals—capturing the underlying heterogeneity and maintaining appropriate correlation with X and Z.

In the first iteration, for each feasible pair of taxon j and quantile level τ_k, we fit two pairs of quantile regression models to the columns of R'. The first pair compares the reduced quantile regression model $Q_{\tau_k}(Y) = \alpha(\tau_k)$ to the model that includes the stage-1 singular vectors $Q_{\tau_k}(Y|\Phi) = \alpha(\tau_k) + \sum_{s=1}^{\hat{m}} \Phi_s \theta_s(\tau_k)$. The second pair compares the model with the stage-1 singular vectors $Q_{\tau_k}(Y|\Phi) = \alpha(\tau_k) + \sum_{s=1}^{\hat{m}} \Phi_s \theta_s(\tau_k)$ to the full model $Q_{\tau_k}(Y|X, Z, \Phi) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_l(\tau_k) + Z_i \gamma(\tau_k) + \sum_{s=1}^{\hat{m}} \Phi_s \theta_s(\tau_k)$. Note that the two models in the first pair can assess whether the taxon-quantile combination can capture information in the stage-1 singular vectors, and the second pair of models allows us to assess whether the taxon-quantile combination convey information about the known covariates, conditional on the stage-1 singular vectors. Thus, these two pairs of models facilitate the

identification of directions that relate to the stage-1 singular vectors without being overly influenced by X and Z.

For each pair of hypotheses at each taxon j and quantile level τ_k, we calculate a rank-based pseudo-F statistic[25] to generate a p-value, comparing whether the larger model fits the data better than the smaller model. With this, we obtain two sets of p-values, with $p_{jk}^{(1)}$ ($p_{jk}^{(2)}$) corresponding to the p-value comparing the first (second) pair of models for taxon j and quantile level τ_k. Intuitively, a small $p_{jk}^{(1)}$ implies that the stage-1 singular vectors corresponding to taxon j and quantile level τ_k convey important information on the data structure; yet a large $p_{jk}^{(2)}$ suggests that taxon j and quantile level τ_k is less impacted by the known covariates given the corresponding stage-1 singular vectors. Thus, we will weight each column of R' so that columns with small $p_{jk}^{(1)}$ values and large $p_{jk}^{(2)}$ values receive a higher weight.

To be specific, we apply the local FDR method[26] to the two sets of p-values separately to obtain two sets of local FDR estimates. This transforms each set of p-values to a set of local FDR estimates. Each taxon-quantile pair corresponds to two p-values before the transformation, and hence corresponds to two local FDR estimates after the transformation. These two local FDR estimates answer the following questions: given the observed data for this specific taxon-quantile pair, what is the posterior probability that it is impacted by the unmeasured heterogeneity and what is the posterior probability that it is not impacted by the known covariates. We then use the product of the two local FDR estimates as the weight for the taxon-quantile pair. We multiply each column of R' by its weight and obtain a reweighted matrix $R^{(u)\prime}$. We then conduct SVD to obtain a new set of singular vectors, which we denote as $\Psi^{(1)}$, in which the superscripts indicate the iteration cycle, and we are now at iteration 1.

Note that R' is obtained without regressing out X and Z, and thus, $\Psi^{(1)}$ can be correlated with X and Z. Note also that the number of the new singular vectors $\Psi^{(1)}$ can be larger than the stage-1 singular vectors Φ. We seek to only maintain the directions that are most correlated with Φ. To do this, for each column, Φ_s, we search for the columns of $\Psi^{(1)}$ that have the highest absolute correlation with Φ_s. In other words, for each column Φ_s, we choose the d_s-th column of $\Psi^{(1)}$ in which $d_s = \arg\max_{d} |\mathrm{corr}\left(\Phi_s, \Psi_d^{(1)}\right)|$. The final selected singular vectors are called the iSVs at the first iteration.

For later iterations, we replace the stage-1 singular vectors in the two pairs of regression models with the iSV of the previous iteration and follow the same procedure. The iteration stops when iSVs from two adjacent iterations are the close enough, or when we reach a sufficient number (T_{max}, with default = 10) of iterations. In the second case, we calculate the final singular vector as a weighted average of the last $T_{max}/2$ iterations. The final singular vectors are the QSV output of our algorithm. Please refer to the Algorithm 1 for individual steps.

Algorithm 1: QuanT algorithm

Input: Y, X, Z

Output: Ψ (quantile surrogate variables)

---

1. Stage 1 residualization:

   For each feasible pair of taxon j and quantile level τ_k, fit quantile regression

   $Q_{\tau_k}\left(Y_{\cdot j} \middle| X, Z\right) = \alpha_j(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_{lj}(\tau_k) + Z_i \gamma_j(\tau_k);$

   Compute $r_{jk} = I\left\{Y_{ij} > \hat{\alpha}_j(\tau_k) + \sum_{l=1}^{L} X_{il}\, \hat{\beta}_{lj}(\tau_k) + Z_i \hat{\gamma}_j(\tau_k)\right\} - (1 - \tau_k);$

Stack r_{jk} across j and k to form matrix R.

2. Conduct SVD on $R = U\Sigma V$.

3. Apply normality test to each singular vector to get p-values. From the top m singular vectors, select singular vectors whose p-value is less than a cutoff c using the SelectiveSeqStep procedure. Here, m is selected so that the FDR is smaller than a prespecified value of q. These form the matrix consisting of stage-1 singular vectors Φ.

4. For each feasible pair of taxon j and quantile level τ_k, fit the reduced quantile regression model $Q_{\tau_k}(Y_{\cdot j}) = \alpha_j(\tau_k)$;

   Compute $r'_{jk} = I\{Y_{ij} > \tilde{\alpha}_j(\tau_k)\} - (1 - \tau_k)$, where $\tilde{\alpha}_j(\tau_k)$ is the estimate under this reduced model.

   Stack r'_{jk} across j and k to form matrix R'.

5. Iterative procedure for correlation calibration:

   a. Initialize $\Psi^{(0)} \leftarrow \Phi$, set $u \leftarrow 0$

   b. Repeat until convergence or until $u = T_{max}$ (maximum number of iteration)

      i. For each taxon j and quantile τ_k, compare pairs of models to obtain $p_{jk}^{(1)}$ and $p_{jk}^{(2)}$:

      Pair 1: $Q_{\tau_k}(Y) = \alpha(\tau_k)$,

      vs. $Q_{\tau_k}(Y|\Psi^{(u)}) = \alpha(\tau_k) + \sum_{s=1}^{\hat{m}} \Psi_s^{(u)} \theta_s(\tau_k)$;

Pair 2: $Q_{\tau_k}\left(Y|\Psi^{(\mathrm{u})}\right) = \alpha(\tau_k) + \sum_{s=1}^{\hat{m}} \Psi_s^{(u)}\theta_s(\tau_k)$,

vs. $Q_{\tau_k}\left(Y|X, Z, \Psi^{(\mathrm{u})}\right) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) + Z_i\gamma(\tau_k) +$

$\sum_{s=1}^{\hat{m}} \Psi_s^{(u)}\theta_s(\tau_k)$

ii. Compute Weights:

Apply the local FDR method to $p_{jk}^{(1)}$ to obtain $q_{jk}^{(1)}$;

Apply the local FDR method to $p_{jk}^{(2)}$ to obtain $q_{jk}^{(2)}$;

Calculate the weight as $w_{jk} = \left(1 - q_{jk}^{(1)}\right)q_{jk}^{(2)}$;

Form $R^{(u)\prime}$ by multiplying the corresponding column of $R'\ by\ w_{jk}$ for each combination of j and τ_k

iii. Conduct SVD on $R^{(u)'}$.

iv. Select the set of singular vectors of $R^{(u)'}$ that shares the strongest absolute correlation with the stage-1 singular vectors Φ. Denote these as $\Psi^{(u+1)}$.

v. Convergence check: If $\left|\left|\Psi^{(u+1)} - \Psi^{(u)}\right|\right|_F \leq \epsilon$, then break and set $\Psi = \Psi^{(u+1)}$; else $u \leftarrow u + 1$.

c. If convergence is not reached after T_{max}iterations:

i. Calculate the weight $\overline{w}_{jk}$ as the average of w_{jk} in the last $\frac{T_{max}}{2}$ iterations;

ii. Form $R^{(u)'}$ by multiplying the corresponding column of $R'by\ \overline{w}_{jk}$ for each combination of j and τ_k.

iii. Conduct SVD on $R^{(u)'}$.

iv. Select the set of singular vectors of $R^{(u)'}$ that shares the strongest absolute correlation with the stage-1 singular vectors Φ. Denote these as Ψ.

6. Return Ψ.

Singular vectors selection when no stage-1 singular vectors are found

In rare cases, we may encounter the situation where no singular vector is selected in the first stage. This usually happens when the hidden heterogeneity variable is normally distributed or when the sample size is not large enough. In this case, we provide an alternative approach that selects important singular vectors solely based on the proportion of variabilities explained. This approach is very similar to the principal component/eigengene selection methods in SVA, except that the least squares regression models are replaced by appropriate quantile regressions. The detailed algorithm can be found in the Algorithm 2. Note that this variant in our algorithm only applies to calculating the stage-1 singular vectors; we still need the same iterative correlation calibration procedure as in the standard QuanT algorithm (Algorithm 1).

Algorithm 2: alternative stage-1 algorithm when no singular vector is selected through the normality test

1-2 Step 1 – 2 is the same as the Algorithm 1.

3. Let d_k be the *k*-th singular value. For the *k*-th singular vector, construct the test statistic as $T_k = \frac{d_k^2}{\sum d_l^2}$

4. Form a matrix R^0 by permuting each row of R independently to remove any structure in the matrix. Denote the (i, j) entry of R^0 by r_{ij}^0.
5. Fit the models $Q_{\tau_k}(r_{\cdot j}^0 | X, Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_l(\tau_k) + Z_i \gamma(\tau_k)$, and obtain the binarized and centered residual matrix, denote by R^{0*}
6. Conduct SVD of R^{0*} and calculate statistic $T_k^0 = \frac{d_{0k}^2}{\sum d_{0l}^2}$
7. Repeat steps 5-7 for a total of B times to obtain null statistics T_K^{0b}, $b = 1, \ldots B$. We set B as 10.
8. Compute the p-value for singular vector k as

$$p_k = \frac{\#\{T_K^{0b} \geq T_k;\ b = 1, \ldots B\}}{B}.$$

We conservatively force monotonicity among the p-values. Thus, set $p_k = max\{p_{k-1}, p_k\}$.

9. Select the singular vectors with $p_k \leq 0.1$.
10. Carry out correlation calibration (steps 6-7 in Algorithm 1)

Simulation settings

We conducted simulations to evaluate QuanT's performance when the true signal is known. We simulated data that mimicked two datasets from the integrative Human Microbiome Project[27]: a gut microbiome dataset from the Inflammatory Bowel Disease Multi-omics Database (IBDMDB) project[28], and a vaginal microbiome dataset from the Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI) project[29]. These two datasets represent microbiome data from two body sites that are of great research interest. They also have contrasting attributes in terms of sparsity and over-dispersion level, which makes them useful

304 for evaluating our proposed method QuanT. We used two methods to simulate data that
305 mimicked the template data: the parametric Dirichlet-Multinomial (DM) model and the non-
306 parametric approach in MIDASim[30]. We additionally simulated a third scenario where
307 microbiome absolute abundances were generated using a log-linear model, in order to evaluate
308 QuanT's performance in compositional settings.
309
310 For each sample i, we simulated a clinical variable of interest (denoted by Z_i) and a hidden
311 batch variable (denoted as B_i) through a bivariate Bernoulli distribution, with both variables
312 having a marginal mean of 0.5. A correlation of 0.3 was set between them. Note that because
313 the "batch" variable is simulated to be correlated with both the microbiome data and Z, thus, it
314 is a confounder. Here, we use the "batch" variable to represent all the hidden heterogeneities.
315 For the parametric simulations, we adopted the simulation setup from a previous study[31]
316 to simulate taxa abundances using the Dirichlet-Multinomial (DM) distribution. In our
317 simulations, we designated the batch variable to impact most taxa (referred to as Batch Taxa),
318 while the variable of interest, Z, influenced only a small subset (referred to as Clinical Taxa).
319 Specifically, for the IBDMDB dataset, we randomly selected 250 of the 301 taxa to be Batch
320 Taxa, among which 40 were also Clinical Taxa. Similarly, for the MOMS-PI dataset, we
321 selected 200 of the 233 taxa as Batch Taxa, with 50 of these being Clinical Taxa. We first fitted
322 the DM model to the template datasets (IBDMDB and MOMS-PI) and obtained the estimated
323 mean vectors π_0. Subsequently, we formed two additional vectors π_B and π_Z to incorporate
324 batch and clinical effects. Both vectors were initially set to be the same as π_0. We then
325 randomly permuted the frequencies within π_B and π_Z corresponding to the Batch Taxa and
326 Clinical Taxa. Finally, a sample-specific frequency was defined as $\pi_i = (1 - w_B B_i -$
327 $w_z Z_i)\pi_0 + w_B B_i \pi_B + w_z Z_i \pi_z$, where (w_B, w_z) determine the underlying batch and clinical

effects. We considered three combinations for these parameters: $(w_B, w_z) \in \{(0,0.5), (0.5,0.5), (0.8,0.2)\}$. When $w_B = 0$, there is no underlying hidden heterogeneity/batch effect. The count data for each sample were generated using the DM distribution with π_i and an overdispersion parameter of 0.02.

For the non-parametric simulations, we used the Microbiome Data Simulator (MIDASim)[30], which is designed to generate microbiome data that capture the underlying distributional and correlation structure of a template dataset. It offers a nonparametric quantile matching approach, and a parametric approach based on log-linear models. Here, we utilized the non-parametric version because it captures the underlying data structure more accurately. Via MIDASim, for each template data set, we first generated vectors of relative abundances and zero-proportions. After choosing the Batch Taxa and the Clinical Taxa, we generated sample-specific relative abundance and zero-proportion vectors in the same way as in the DM simulation and also used the same (w_B, w_z) combinations. Count data for each sample were generated by inputting the sample-specific vectors into MIDASim and choosing to use its non-parametric approach.

In the third simulation scenario, we again simulated microbiome read counts using the Dirichlet–Multinomial model, but introduced batch and clinical effects differently. Instead of modifying the mean frequency vectors π_B and π_Z as in the first simulation, we first generated read counts from the fitted DM model using π_0, and then directly adjusted the counts. Specifically, for samples with $B_i = 1$, read counts of Batch Taxa were multiplied by w_B; for samples with $Z_i = 1$, read counts of Clinical Taxa were multiplied by w_z. We considered three combinations for these parameters: $(w_B, w_z) \in \{(1,4), (4,4), (5,3)\}$.

351 ## Results

352 To rigorously assess QuanT, we conducted extensive simulation studies based on real-world
353 gut and virginal microbiome datasets from the HMP2 projects[27]. We also applied QuanT and
354 competing methods to three curated microbiome datasets, each of which has distinct sources
355 of technical heterogeneity: a gut microbiome study on cardiovascular diseases, which displays
356 a moderate technical variation arising from sequencing in multiple runs; an integrated dataset
357 compiled from various HIV gut microbiome studies; and a consolidated dataset encompassing
358 multiple colon cancer studies. Through both visual and quantitative comparisons, we
359 demonstrate that QuanT outperforms existing methods by robustly and comprehensively
360 uncovering underlying microbiome data heterogeneity, which is beyond mere shifts in means.
361 When incorporated into the analytical workflow, QuanT effectively reduces the influence of
362 unwanted heterogeneity, enhancing false discovery rate control, maintaining sensitivity of
363 discovery, as well as improving prediction accuracy.
364

365 **Application to the CARDIA study, a single large-scale epidemiological investigation**

366 We assessed QuanT using multiple datasets. We first applied it to a study containing
367 traditional batch variation: samples are collected under one protocol but handled in different
368 batches. The Coronary Artery Risk Development in Young Adults (CARDIA) Study[32],
369 initiated in 1985-1986, recruited young adults aged between 18 and 30 with the goal of
370 uncovering determinants of cardiovascular disease (CVD) throughout adult life. Subject
371 enrollment was balanced according to race, gender, education, and age. Extensive data on
372 risk factors related to CVD have been collected, including health behaviors, medical history
373 (including medication use), and clinical risk factors, such as blood pressure. Stool samples
374 were also collected for some visits. Here, we used the microbiome data generated from the

375 stool samples collected during the Year 30 follow-up (2015-2016). The samples were
376 processed through four DNA extraction/library preparation batches, and subsequently
377 sequenced through seven runs using the Illumina MiSeq 2x300 platform. Specifically, each
378 DNA extraction batch was allocated to two sequencing runs, except for the final batch, which
379 was sequenced by itself during the seventh run. Consequently, the data were processed in
380 seven distinct batches. Following sequencing, forward reads were processed through the
381 DADA2[33] pipeline for quality control and derivation of amplicon sequence variants (ASVs),
382 and taxonomy was assigned using the Silva reference database[34]. The data were agglomerated
383 to the genus level, and taxa that did not yield any reads in all samples were omitted from the
384 dataset. After preprocessing, the data contains 633 samples and 145 taxa. Analyses were
385 restricted to participants for whom all covariate data were available.
386 In this study, we considered high blood pressure (HBP) (non-HBP = 0, HBP = 1) as
387 the variable of interest, and gender (Male = 0, Female = 1) and race (White = 0, Black = 1) as
388 adjustment covariates. To assess QuanT and the competing methods, we treated the batch
389 indicators as unknown and used these methods to identify the now-hidden
390 heterogeneities/batches. The performance of these methods was assessed based on three
391 criteria: 1) consistency and sensitivity in downstream differential abundance analysis for
392 individual taxa; 2) enhancement in prediction accuracy for the variable of interest; and 3) the
393 efficacy of batch effect removal while preserving the effect from the variable of interest at the
394 entire community level. For SVA, we used the svaseq function, which added a pseudo count
395 of 1 to all counts and then applied log transformation as described in Leek (2014)[15], and used
396 the iterative reweighted algorithm[12]. The default parameter setting was used. For RUV, we
397 used the RUVg approach in Risso *et al.* (2014)[10] with an estimated group of null taxa as
398 controls, rather than RUVr and RUVs as these approaches can only identify hidden directions

399 that are uncorrelated with the variables of interest. For hyperparameters, we selected control
400 taxa using the Benjamini–Hochberg procedure on edgeR p-values at a 0.05 threshold, and set
401 the number of latent factors following the guidance in Risso *et al.* (2014)[10].

402 Fig. 2a and 2b compare the consistency of differential abundance analysis results
403 across various models. Here, differential abundance analysis was conducted using linear
404 regression on relative abundances of each taxon, controlling for age, gender, and identified
405 hidden variables. For benchmarking purposes, analyses were conducted adjusting for true
406 batch/group indicators (TG), which was regarded as the gold standard, and with no correction
407 (NC), which served as the baseline. Fig. 2a shows the number of differentially abundant taxa
408 identified from the CARDIA dataset by Benjamini-Hochberg adjusted p-value < 0.05. TG
409 leads to 4 discoveries. Considering these as true signals, NC identifies the 4 true discoveries
410 and has 1 additional false discovery; QuanT identifies 3 true discoveries and has 1 additional
411 false discovery. SVA and RUV lead to no discoveries. Therefore, QuanT demonstrates
412 superior performance compared to SVA and RUV when applied to the CARDIA dataset,
413 identifying more potential true discoveries while maintaining a reasonable false discovery
414 proportion. Fig. 2b shows the top K detected differentially abundant taxa. Treating TG as the
415 gold standard and calculating the proportion of overlapping discoveries, QuanT has
416 comparable performance with NC and SVA, and is substantially better than RUV.

417 Fig. 2c shows the prediction accuracy using the original and the ConQuR corrected
418 microbiome data, where the corrected data were adjusted for TG, or surrogate variables
419 estimated by SVA, RUV or QuanT. Random forest models were fit with five-fold cross-
420 validation to predict HBP status, and the area under the receiver operating characteristic curve
421 (ROC-AUC) was used to evaluate the accuracy. We note that this is not a genuine predictive
422 analysis, but rather a metric to quantify how well the corrected microbiome data can explain

423 the variable of interest. As shown in Fig. 2c, RUV and QuanT corrected data predicts HBP
424 more accurately than using TG or SVA, and prediction using the unadjusted data (NC) leads
425 to the worst accuracy. Fig. 2d shows the principal coordinate analysis (PCoA) plots using the
426 Bray-Curtis distance, comparing with the uncorrected and ConQuR-corrected datasets.
427 Assessed by the PERMANOVA R^2 (which evaluates the multivariate association between the
428 microbial community and a certain variable, here quantifies the proportion of microbiome data
429 variability explained by the variable), correction with TG leads to the most significant
430 reduction in batch effects, which was expected. Visually, it not only equalizes the mean and
431 dispersion (centroid and size of the ellipses) across groups but also removes higher-order
432 batch/group effects (shown by angles of the ellipses). Correction for QuanT, SVA, and RUV's
433 estimated surrogate variables leads to comparable efficacy in reducing batch effects and
434 maintaining the effects of high blood pressure (HBP).
435 It is worth noting that the CARDIA study has well-balanced batch designs; thus the
436 analysis with NC leads to reasonable and comparable results as the analyses with adjustments.
437
438

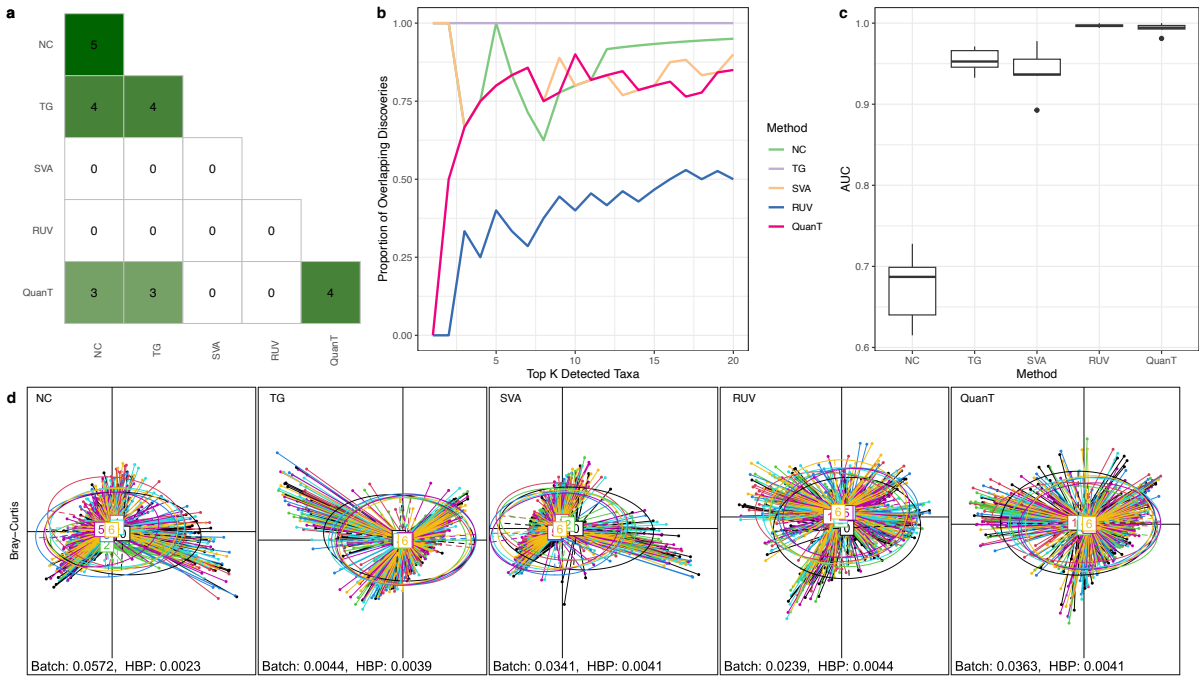


Fig. 2 | Evaluation on the CARDIA data. NC: no batch correction; TG: correction using true group indicators. **a** Heatmap representing the number of differentially abundant taxa identified in the CARDIA dataset by Benjamini-Hochberg adjusted p-value < 0.05; diagonal elements are numbers detected, and off-diagonal elements are numbers detected in common. **b** Proportions of overlapping taxa in the top K detected differentially abundant taxa between each method and TG in the CARDIA dataset. **c** Average cross-validated area under the receiver operating characteristic curves (ROC-AUC) of predicting HBP from the ConQuR-corrected microbiome data via random forest. Larger values indicate stronger predictive capabilities. **d** PCoA plots of the ConQuR-corrected data based on Bray-Curtis distance, with PERMANOVA R^2 of batch ID and HBP.

Application to HIVRC, an integrated dataset with multiple HIV studies

Our second example used data from the HIV re-analysis consortium (HIVRC)[5]. In this consortium study, raw 16s rRNA gene sequencing data were solicited from 22 studies that investigated the relationship between the gut microbiome and HIV infection. Data from 17 of these studies were subsequently processed through a common pipeline with taxa assigned using

455 Resphera Insight[35]. Note that the studies in the final-processed data have distinct experimental
456 designs and sequencing protocols (Supp. Tab. 2 of Tuddenham, et al. 2020[17]). Thus, the data
457 have more pronounced variability than that observed within a single study, such as the
458 CARDIA study.

459 We selected two studies with similar library sizes (5000-5500) but very different study
460 designs (an observational study in the US[36] and a clinical trial in Europe[37]), which include 37
461 and 24 samples respectively. After preprocessing, the data contains 126 taxa. We applied
462 QuanT, SVA, and RUV to the data aggregated at the genus level. HIV status (Negative = 0,
463 Positive = 1) was regarded as the variable of interest, while age, gender and BMI were
464 considered as covariates. We retained complete cases only and conducted the same set of
465 analyses as on the CARDIA data.

466 Fig. 3a and 3b compare the consistency of differential abundance analysis results
467 obtained analyzing these data using various methods. Fig. 3a shows the number of
468 differentially abundant taxa identified in the HIVRC dataset by Benjamini-Hochberg adjusted
469 p-value < 0.05. We identified 19 discoveries adjusting for true group indicators (TG), which
470 were considered as the gold standard. QuanT leads to 17 true discoveries and 1 false discovery.
471 NC leads to 11 true discoveries and 16 false discoveries. SVA leads to 11 true discoveries and
472 21 false discoveries. RUV leads to 8 true discoveries and 5 false discoveries. Therefore, QuanT
473 again performs better than the other methods when applied to the HIVRC dataset and has the
474 lowest FDR compared to competitors. In Fig. 3b, we consider the top K detected differentially
475 abundant taxa. QuanT is the most consistent with the gold standard, TG, followed by NC, SVA,
476 and RUV.

477 Fig. 3c shows the average cross-validated ROC-AUC of predicting HIV from the
478 ConQuR-corrected data via random forest. TG, SVA, RUV, and QuanT all perform perfectly

and are better than no correction. Fig. 3d shows the PCoA plots using Bray-Curtis distance, comparing with the uncorrected and ConQuR-corrected datasets. Visually, TG, RUV and QuanT lead to better harmonization of the microbiome data from the two distinct HIV studies. As quantified by PERMANOVA R^2, QuanT leads to the greatest reduction of the effects of the unwanted heterogeneity, followed by RUV, SVA and TG. On the other hand, SVA maintains the HIV effects the most, followed by TG, QuanT and RUV (figure 3d).

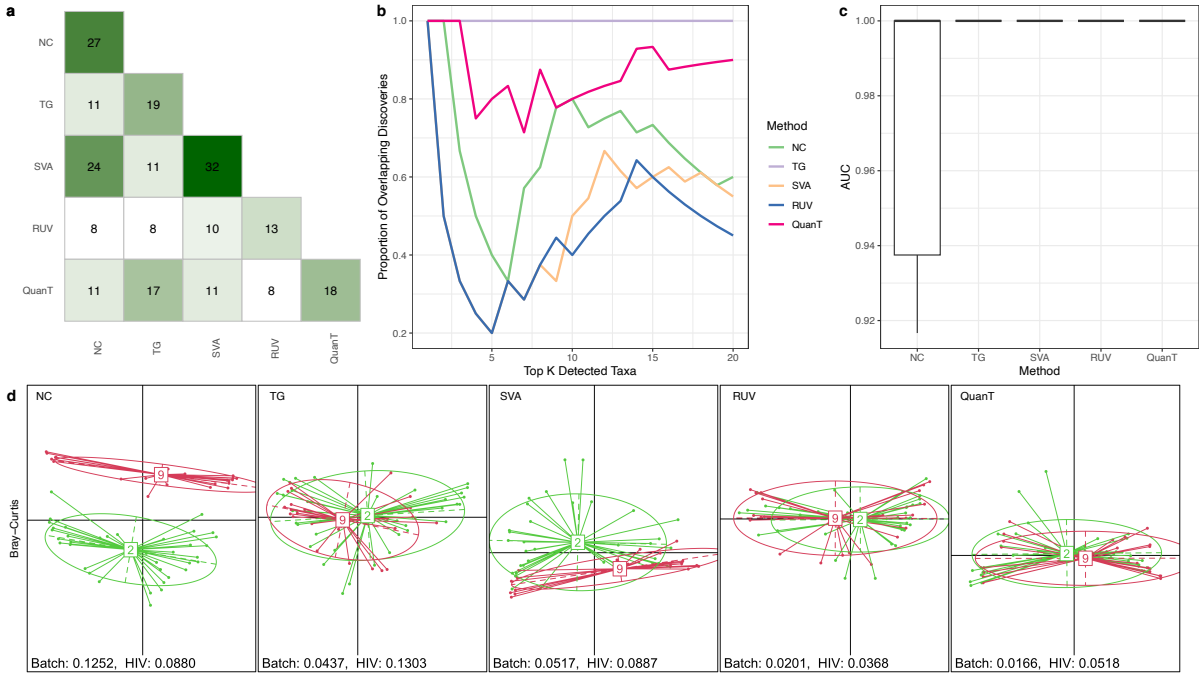


Fig. 3 | Evaluation on the HIVRC data. NC: no batch correction; TG: correction using true group indicators. **a** Heatmap representing the number of differentially abundant taxa identified in the HIVRC dataset by Benjamini-Hochberg adjusted p-value < 0.05; diagonal elements are numbers detected, off-diagonal elements are numbers detected in common. **b** Proportions of overlapping taxa in the top K detected differentially abundant taxa between each method and TG in the HIVRC dataset. **c** Average cross-validated area under the receiver operating characteristic curves (ROC-AUC) of predicting HIV from the ConQuR-corrected data via random forest. Larger values indicate stronger predictive capabilities. **d** PCoA plots of the ConQuR-corrected data based on the Bray-Curtis dissimilarity, with PERMANOVA R^2 of study ID and HIV.

495

496 **Application to an integrated microbiome dataset in colon cancers**

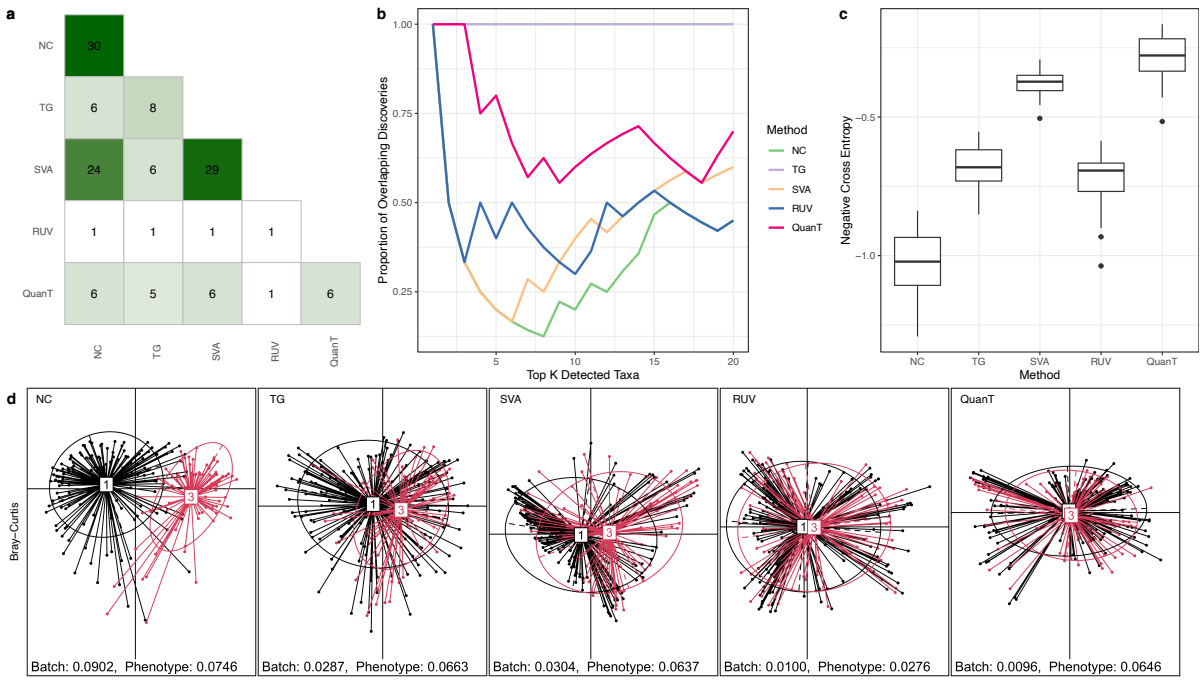
497 Our third example came from two colorectal cancer (CRC) studies[38, 39] from the
498 curatedMetagenomicData package[40]. The CuratedMetagenomicData data package offers
499 standardized, curated, and analysis-ready metagenomic datasets derived from human
500 microbiome studies, facilitating comprehensive meta-analyses and cross-study comparisons
501 by ensuring consistent data processing and annotation. The bacterial, fungal, and archaeal
502 taxonomic abundances for each sample were calculated with MetaPhlAn3[41], and metabolic
503 functional potential was calculated with HUMAnN3[41]. We selected two CRC studies[38, 39]
504 from this package, which include 154 and 80 stool samples, respectively. After filtering, 166
505 genera were included in the data. These studies collectively cover samples from three distinct
506 health conditions: colorectal cancer (CRC), adenoma, and disease-free controls, with CRC
507 subdivided into adenocarcinoma and carcinoma, and adenoma into simple adenoma and
508 advanced adenoma.

509 We applied QuanT, SVA, and RUV to this CRC dataset to find the hidden heterogeneity
510 directions, treating the true study indicators as unknown. We considered the five phenotype
511 categories (adenocarcinoma, carcinoma, adenoma, advanced adenoma, control) as the primary
512 variable of interest, age (Adult = 0, Senior = 1) and gender (Male = 0, Female = 1) as the
513 covariates to adjust for.

514 Fig. 4a and 4b compare the consistency of differential abundance analysis results across
515 various methods. Fig. 4a shows the number of differentially abundant taxa identified in the
516 CRC dataset by Benjamini-Hochberg adjusted p-value < 0.05. TG (gold standard) leads to 8
517 discoveries. QuanT leads to 5 true discoveries with 1 false discovery. NC leads to 6 true
518 discoveries and 24 false discoveries. SVA leads to 6 true discoveries and 23 false discoveries.

RUV leads to 1 true discovery with no false discovery. Therefore, NC and SVA lead to inflated false discovery proportion when applied to the CRC dataset. RUV and QuanT have reasonable false discovery proportion control, while QuanT is notably better at identifying true signals than RUV. In Fig. 3b, we consider the top K detected differentially abundant taxa. QuanT is the most consistent with the gold standard, TG. NC, SVA, and RUV perform similarly are all outperformed by QuanT.

Fig. 4c presents the prediction accuracy, measured by negative cross entropy, for the five phenotype groups using the original and ConQuR-corrected CRC microbiome data. Here, QuanT performs the best, followed by SVA, RUV, TG, and NC. Fig. 4d displays the PCoA plots using Bray-Curtis distance. Visually, QuanT leads to similar data harmonization as using the true study IDs, and outperforms SVA, RUV and NC. This is numerically confirmed by the PERMANOVA R^2 of the study ID and the phenotypes of interest.



534 **Fig. 4 | Evaluation on the CRC data.** NC: no batch correction; TG: correction using true group indicators. **a**
535 Heatmap representing the number of differentially abundant taxa identified in the CRC dataset by Benjamini-
536 Hochberg adjusted p-value < 0.05; diagonal elements are numbers detected, off-diagonal elements are numbers
537 detected in common. **b** Proportions of overlapping taxa in the top K detected differentially abundant taxa between
538 each method and TG in the CRC dataset. **c** Average cross-validated negative cross entropy of predicting HIV
539 from the ConQuR-corrected data via random forest. Larger values indicate stronger predictive capabilities. **d.**
540 PCoA plots of the ConQuR-corrected data based on the Bray-Curtis dissimilarity, with PERMANOVA R^2 of
541 study ID and Phenotype.
542

543 **Evaluation using simulated data**

544 We applied SVA, RUV and QuanT to the simulated datasets to identify the underlying hidden
545 heterogeneity, treating the batch variables as unknown. Like in the CARDIA, HIVRC, and
546 CRC analyses, we used linear regression models to conduct differential abundance analysis,
547 adjusting for hidden variables estimated by SVA, RUV, and QuanT. When data are simulated
548 from MIDASim, as shown in Fig. 5a, QuanT leads to better FDR control compared to SVA
549 and RUV. Fig. 5b shows that the discoveries by QuanT align more closely with the gold
550 standard than SVA, RUV, and the NC methods. Here, the gold standard varies: in the absence
551 of batch effects, NC is the gold standard, while when batch effects are present, the model
552 correcting for true batch ID is the gold standard. More simulation results based on MIDASim
553 (Supp. Fig. 2,4,6) and results based on parametric simulations (Supp. Fig. 1,3,5) lead to similar
554 conclusions.

555 We also applied ConQuR to remove the true batch, SVs and QSV effects and generate
556 a "batch-free" dataset. Integrating QuanT with ConQuR in simulations (Supp. Fig. 7-10)
557 yielded microbial profiles where confounding batch effects were eliminated, while the effects

558 of the variable of interest were preserved. Finally, QuanT demonstrates scalable computational
559 performance for large-scale microbiome studies, as evidenced in Tab. 1 and Tab. 2.
560 For our simulation scenario 3, we applied LinDA[42] and LDM-CLR[43], two modern tools
561 that are designed for compositional hypothesis in microbiome data. These results (Supp. Fig.
562 11-16) align with the conclusions from our other simulation studies.
563 In short, QuanT demonstrates improved performance compared to existing methods,
564 on simulated datasets and biological datasets, regardless of the effect size and the type of source
565 of unmeasured heterogeneity.
566

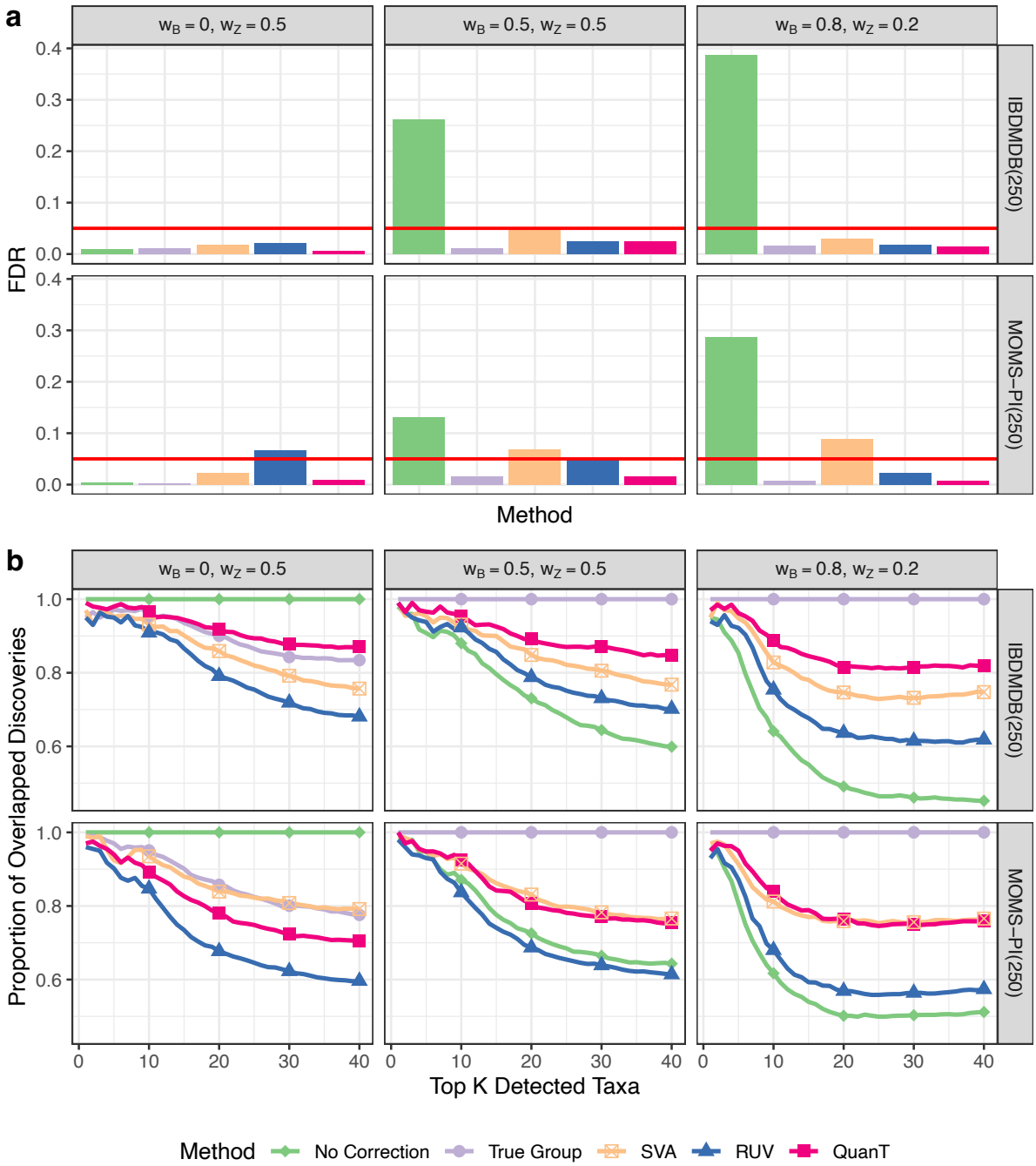


Fig. 5 | Evaluation on the simulated data. The data are simulated by MIDASim using the IBDMDB and MOMS-PI datasets as templates. **a** FDR of the differential abundance analysis via adjusted linear regression with BH-adjusted p-value < 0.05. **b** Proportions of overlapping taxa in the top K detected differentially abundant taxa between each method and the gold standard in simulated data. Simulation setups vary across different datasets (rows), and different combinations of batch and key variable effects (columns).

573 **Computation of QuanT**

574 We have streamlined the process of quantile regression and matrix decomposition within
575 QuanT by employing strategies for enhanced efficiency. The algorithm executes quantile
576 regression across various taxon and quantile levels independently, which lends itself well to
577 parallel processing. We leverage the doParallel package in R to expedite this computation. In
578 the second stage of QuanT, where SVD is performed, we focus on a few leading singular
579 vectors because it is generally improbable for a singular vector that accounts for minimal
580 variance to represent a critical direction established in the first stage. Consequently, it is
581 sufficient to compute only the leading singular vectors of the binarized residual matrix during
582 the correlation calibration phase. For this reason, QuanT only calculates the top 15 principal
583 components using the RSpectra R package.

584 Tables 1 and 2 present the computational time of QuanT for datasets with different
585 numbers of samples and taxa. Table 1 illustrates the scalability of QuanT's computational time
586 with respect to various numbers of samples (n) and taxa (p), benchmarked against SVA and
587 RUV. For each configuration of n and p, we generated ten datasets using the DM simulation
588 with parameters $(w_B, w_z) = (0.5, 0.25)$ and the IBDMDB dataset as the template. We also
589 compared the computational time of QuanT, SVA, and RUV for analyzing the three real-world
590 datasets (CARDIA, HIVRC, and CRC) (Tab. 2). Overall, QuanT exhibits slower computation
591 than SVA and RUV due to its multiple quantile regression procedures. Nevertheless, the run
592 time remains practical and the method is applicable to reasonably large microbiome datasets.

593

594 **Table 1 | The computation time (in seconds) of QuanT with different numbers of samples**
595 **and taxa, compared with SVA and RUV. The data are simulated from the Dirichlet-**
596 **Multinomial distribution using the IBDMDB dataset as the template.**

	p=100			p=200			p=300		
	SVA	RUV	QuanT	SVA	RUV	QuanT	SVA	RUV	QuanT
n=250	0.4	0.2	37.9	0.9	0.4	83.3	1.5	0.6	138.0
n=500	1.5	0.5	56.5	2.7	0.7	113.9	4.2	1.0	193.2
n=750	4.0	0.6	78.8	6.0	1.3	157.5	8.5	1.6	256.4
n=1000	8.5	0.8	103.3	11.7	1.4	208.0	15.0	1.9	334.7

Table 2 | The computation time (in seconds) of QuanT for the three datasets analyzed, compared with SVA and RUV.

	n	p	SVA	RUV	QuanT
CARDIA	633	145	4.1	2.0	107.0
HIVRC	61	126	0.1	0.6	52.8
CRC	234	166	0.9	5.9	79.1

Discussion

Identifying unmeasured heterogeneity is a challenging yet important task in microbiome data analysis. Failing to identify and adjust for the unmeasured heterogeneity can drastically increase the false discovery rate and hinder the detection of true signals. Due to lack of methods

607 specifically tailored to microbiome data, researchers often use methods developed for genomic
608 studies. However, these methods do not account for the challenging characteristics of
609 microbiome data, such as sparsity and overdispersion. In addition, they focus solely on the
610 information underlying the mean expression and ignore other aspects of the data distribution.
611 As shown the "Results" section, these methods lead to increased false discoveries and/or
612 reduced power when analyzing microbiome studies.

613 We present QuanT, the first comprehensive tool to identify unmeasured heterogeneity
614 in microbiome data. QuanT-uses quantile regression models and so can thoroughly investigate
615 the underlying data structure at multiple quantiles. QuanT outputs a set of vectors as surrogates
616 (QSV) for the unmeasured heterogeneity, which offers users the flexibility to incorporate them
617 in different types of downstream analysis. These vectors can be treated as known covariates
618 and adjusted for in regression models. They can also be used in conjunction with a batch effect
619 removal method, such as ConQuR, to generate microbial profiles without the effects of hidden
620 heterogeneity.

621 We applied QuanT to three microbiome datasets and a wide range of simulated datasets
622 to assess its performance, in comparison with competing methods. The microbiome datasets
623 include large-scale epidemiological studies with samples processed in different sequencing
624 runs, and curated microbiome datasets integrated from various publicly available studies. Our
625 simulations are comprehensive, generating microbiome profiles that mimic real-world
626 microbiome studies in gut and vaginal environments and using two distinct simulation
627 mechanisms. Visually and numerically, we showed that QuanT improved subsequent
628 differential abundance analysis, offering better FDR control and improved statistical power,
629 and providing results that align better with the gold standard in most setups. Frequently, the
630 results of differential abundance analysis following QuanT are comparable to those when the

true labels of batches or studies are known. It also offers comparable or improved performance in predicting the variable of interest, as well as reducing the impact of hidden heterogeneity while preserving key signals at the community level.

Despite its advantages, QuanT has several limitations and potentials that could be investigated in future research. First, nowadays, it is common that microbiome studies are conducted in a longitudinal manner to investigate the dynamics of the microbiome. Such studies typically produce abundance data collected across multiple time points from the same individuals. For longitudinal data, one tentative approach is to incorporate time as an additional adjustment variable into the QuanT framework. However, this approach ignores the correlations between repeated samples from the same person and may lead to spurious conclusions. In contrast, in a well-designed study, multiple samples from the same individual are usually sequenced on the same plate. In such cases, baseline microbiome data can be used for the identification of unmeasured heterogeneity. Sensitivity analyses may then be conducted by leveraging microbiome samples that share similar batch structures—for example, by performing analyses separately on baseline samples and on samples collected at later time points, thereby assessing the consistency of results across these subsets. Secondly, QuanT leads to well-controlled false discovery rate in our real and extensive simulation experiments, although it lacks theoretical guarantee. However, theoretical guarantees might be derived given strict assumptions about the data generation mechanism, which, in practice, is usually less important. Finally, a potential extension of QuanT is to identify unmeasured heterogeneity in gene expression data, especially in single cell and spatial transcriptomics data which have similar sparsity and over-dispersion as do microbiome data. It would be worth investigating how QuanT performs when applied to these data types, which might

654 require some adaptations. However, these aspects are beyond the scope of this paper, and we
655 leave them to future studies.
656

657 Conclusions

658 In summary, we propose a versatile approach, QuanT, to uncover concealed heterogeneities
659 within microbiome data. QuanT stands out from existing approaches by detecting hidden data
660 structures beyond the mean and by easing the constraints of distributional assumptions. When
661 integrated into the analysis pipeline, QuanT effectively mitigates the confounding influences
662 of undesired latent variables, enhances false discovery rate control, and preserves statistical
663 power.
664

665 List of abbreviations

666 QuanT: quantile thresholding
667 SVA: surrogate variable analysis
668 RUV: remove unwanted variation
669 SVD: singular value decomposition
670 iSV: iterative singular vectors
671 QSV: quantile surrogate variables
672 VOI: variable of interest
673 CARDIA: Coronary Artery Risk Development in Young Adults
674 CVD: cardiovascular disease
675 HBP: high blood pressure
676 TG: true group

NC: no correction

ROC-AUC: area under the receiver operating characteristic curve

PCoA: principal coordinate analysis

HIVRC: HIV re-analysis consortium

CRC: colorectal cancer

IBDMDB: Inflammatory Bowel Disease Multi-omics Database

MOMS-PI: Multi-Omic Microbiome Study: Pregnancy Initiative

DM: Dirichlet-Multinomial

MIDASim: Microbiome Data Simulator

Declarations

Ethics approval and consent to participate

Not applicable. Because this study only involved secondary analysis of existing, de-identified datasets, it is not considered human subject research.

Consent for publication

Not applicable.

Availability of data and material

The CARDIA data can be shared upon request by submitting a CARDIA Data Set Request-Intent to Analyze Form to the CARDIA Coordinating Center, University of Alabama at Birmingham, at https://www.cardia.dopm.uab.edu/publications-2/publications-documents. The HIVRC data are available in Synapse under accession code syn18406854. The CRC data

are from the curatedMetagenomicData package at https://waldronlab.io/curatedMetagenomicData/. The Inflammatory Bowel Disease Multi-omics Database (IBDMDB) project[28] and the vaginal microbiome dataset from Multi-Omic Microbiome Study[29]: Pregnancy Initiative (MOMS-PI) project, which were used as templates for simulation studies, are provided in the Bioconductor R package, HMP2Data[44].

The R package QuanT is available at https://github.com/jiuyaolu/QuanT in formats appropriate for Macintosh, Windows, or Linux systems. A vignette demonstrating use of the package is included (can be accessed at https://jiuyaolu.github.io/QuanT/QuanT.html).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by R01GM147162 (Lu, Satten, and Zhao), R01HL155417 (Ling), R01GM151301 (Ling) and R01GM155734 (Ling). The Coronary Artery Risk Development in Young Adults Study (CARDIA) is supported by contracts 75N92023D00002, 75N92023D00003, 75N92023D00004, 75N92023D00005, and 75N92023D00006 from the National Heart, Lung, and Blood Institute (NHLBI). The HIVRC data used in this study are from work that was supported by the HIV Microbiome Re-analysis Consortium.

Authors' contributions

JL, GS, NZ and WL developed the method, analyzed the data, and wrote the manuscript. JL prepared the software package and vignette. LL and KM contributed to the dataset generation

and manuscript development with technical input. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Drs. Susan A. Tuddenham, James White, Wei Li A Koay, Cynthia L. Sears, and all members on the HIV Microbiome Re-analysis Consortium for collecting and processing the HIVRC dataset.

References

1. Oliveira, F.S. et al. MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic acids research* **46**, D684-D691 (2018).
2. Dhariwal, A., Chong, J., Habib, S., King, I.L., Agellon, L.B. & Xia, J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic acids research* **45**, W180-W188 (2017).
3. Stewart, C.J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583-588 (2018).
4. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A. & Alm, E.J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications* **8**, 1784 (2017).
5. Tuddenham, S.A., Koay, W.L.A., Zhao, N., White, J.R., Ghanem, K.G. & Sears, C.L. The impact of human immunodeficiency virus infection on gut microbiota α-diversity: an individual-level meta-analysis. *Clinical Infectious Diseases* **70**, 615-627 (2020).
6. Ma, S. et al. Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biology* **23**, 208 (2022).
7. Austin, G.I. et al. Processing-bias correction with DEBIAS-M improves cross-study generalization of microbiome-based prediction models. *Nature Microbiology*, 1-15 (2025).
8. Ling, W. et al. Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications* **13**, 1-14 (2022).
9. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**, e161 (2007).
10. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896-902 (2014).
11. Koenker, R. & Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50 (1978).

12. Leek, J.T. & Storey, J.D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105**, 18718-18723 (2008).
13. Leek, J.T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 733-739 (2010).
14. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).
15. Leek, J.T. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research* **42**, e161-e161 (2014).
16. Koenker, R. & Hallock, K.F. Quantile regression. *Journal of economic perspectives* **15**, 143-156 (2001).
17. Barber, R.F. & Candès, E.J. Controlling the false discovery rate via knockoffs. (2015).
18. Goldstein, S., Lebowitz, J.L., Tumulka, R. & Zanghi, N. Any orthonormal basis in high dimension is uniformly distributed over the sphere. (2017).
19. Lilliefors, H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association* **62**, 399-402 (1967).
20. Anderson, T.W. & Darling, D.A. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, 193-212 (1952).
21. Royston, P. A pocket-calculator algorithm for the shapiro-francia test for non-normality: An application to medicine. *Statistics in medicine* **12**, 181-184 (1993).
22. Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**, 479-498 (2002).
23. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics* **31**, 2013-2035 (2003).
24. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 (2003).
25. Gutenbrunner, C., Jurečková, J., Koenker, R. & Portnoy, S. Tests of linear hypotheses based on regression rank scores. *Journaltitle of Nonparametric Statistics* **2**, 307-331 (1993).
26. Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association* **96**, 1151-1160 (2001).
27. Integrative, H. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe* **16**, 276-289 (2014).
28. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662 (2019).
29. Fettweis, J.M. et al. The vaginal microbiome and preterm birth. *Nature medicine* **25**, 1012-1021 (2019).
30. He, M., Zhao, N. & Satten, G.A. MIDASim: a fast and simple simulator for realistic microbiome data. *Microbiome* **12**, 135 (2024).

31. Hu, Y.-J. & Satten, G.A. Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics* **36**, 4106-4115 (2020).
32. Friedman, G.D. et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology* **41**, 1105-1116 (1988).
33. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. & Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods* **13**, 581-583 (2016).
34. Callahan, B. Silva taxonomic training data formatted for DADA2 (Silva version 132). *Zenodo* (2018).
35. Daquigan, N., Seekatz, A.M., Greathouse, K.L., Young, V.B. & White, J.R. High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden. *NPJ biofilms and microbiomes* **3**, 35 (2017).
36. Lozupone, C.A. et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* **14**, 329-339 (2013).
37. Villar-García, J. et al. Impact of probiotic Saccharomyces boulardii on the gut microbiome composition in HIV-treated patients: a double-blind, randomised, placebo-controlled trial. *PloS one* **12**, e0173802 (2017).
38. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature communications* **6**, 6528 (2015).
39. Thomas, A.M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature medicine* **25**, 667-678 (2019).
40. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023-1024 (2017).
41. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *elife* **10**, e65088 (2021).
42. Zhou, H., He, K., Chen, J. & Zhang, X. LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome biology* **23**, 95 (2022).
43. Hu, Y.-J. & Satten, G.A. Compositional analysis of microbiome data using the linear decomposition model (LDM). *Bioinformatics* **39**, btad668 (2023).
44. Stansfield, J., Smirnova, E., Zhao, N., Fettweis, J., Waldron, L. & Dozmorov, M. HMP2Data: 16s rRNA sequencing data from the Human Microbiome Project 2. *R package version* **1** (2020).