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.

Currently, no methods are designed to address unmeasured heterogeneity in microbiome studies, although techniques developed for genomic data, such as surrogate variable analysis (SVA)[6] and remove unwanted variation (RUV)[7], 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[8] 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[9], 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 select "features" (i.e., combinations of quantile level and taxon) strongly marginally associated with the stage-1 singular vectors but weakly associated with the covariates. 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 only using the selected 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[6, 10-13].

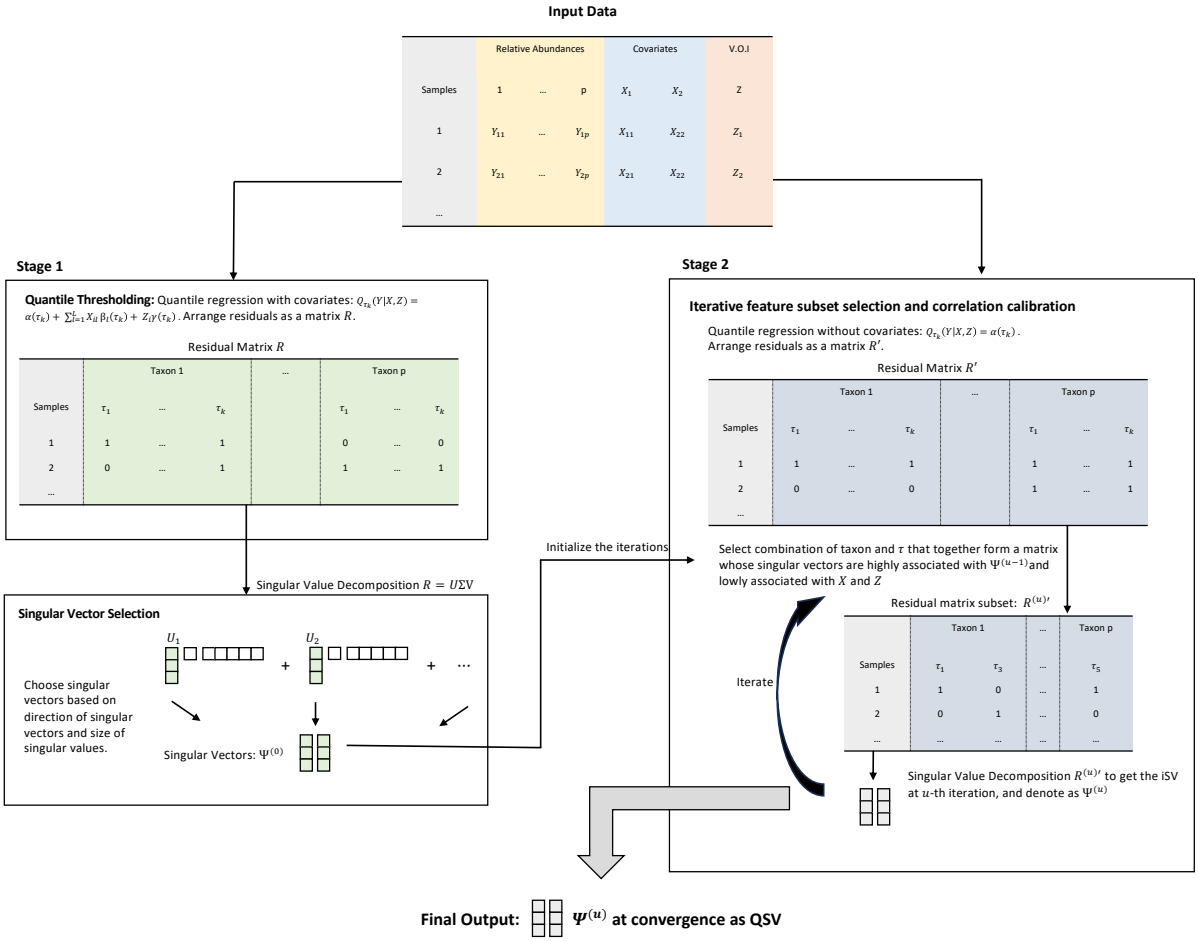

Fig. 1 | Flowchart of QuanT. V.O.I: variable of interest; QSV: quantile surrogate variable; $\tau_1 \ldots \tau_k$: quantile levels. Stage 1 conducts quantile regression for each taxon at multiple quantile levels using V.O.I (Z) and covariates (X). Residuals are thresholded to create a binary matrix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

125 and over-dispersed. First, by using quantile regressions instead of least squares or generalized
126 linear models, QuanT can gather information from the whole distribution of the taxa rather
127 than just the mean, which is critical for microbiome data. Secondly, SVA and RUV search for
128 significant directions by identifying singular vectors that represent a disproportionately high
129 amount of variation. QuanT, on the other hand, relies on both the direction of a singular vector
130 and the variance it explains, and thus allows us to detect significant directions that may harbor
131 important underlying structure but explain only a modest proportion of variation. In the second
132 stage of correlation calibration, QuanT uses a data-adaptive approach to select the p-value
133 cutoffs for the two pairs of comparisons (highly associated with the prior stage singular vectors
134 while minimally associated with variable of interest and important covariates), which contrasts
135 with SVA where a fixed cutoff is used. Our investigation shows that there are no universally
136 best cutoffs, and our adaptive approach allows for the optimal calibration (within our choice of
137 cutoffs) and thus allows for better FDR control in downstream analyses.
138
139 We now describe our notation and proceed to a more in-depth explanation of the method's
140 components.
141 **Notation**
142 Assume the relative abundance data form a $n \times p$ matrix Y, in which Y_{ij} is the relative
143 abundance of the j-th taxon in the i-th sample. We denote by $Y_{\cdot j}$ the j-th column of Y
144 comprising data on the j-th taxon, and treat it as the outcome in the proposed regression
145 framework. Suppose that in addition each sample i has a primary variable of interest (denoted
146 by Z_i) and a set of L covariates (denoted by X_i) to adjust for. We use quantile regression at
147 each of K quantile levels in our regression framework; these levels are represented by the
148 vector, $\vec{\tau}$. By default, $K = 5$, and $\vec{\tau}$ is the 10th, 25th, 50th, 75th, and 90th quantiles of the

distribution formed by the proportions of zeros across all taxa in the data. Specifically, if these quantiles correspond to proportions of 0.3, 0.5, 0.56, 0.72, and 0.9 respectively in the dataset, then τ is defined as (0.3, 0.5, 0.56, 0.72, 0.9).

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 is 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[14] $\rho_{\tau_k(x)} = \tau_k x - I\{x < 0\}x$.

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[15] 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[16]. 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[17] or the Anderson-Darling test[18] 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 $\#\{j \leq m: p_j > c\}/\frac{1-c}{c}$. This is similar to the idea of q-value[19-21], 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 largest m such that this proportion is smaller than a false discovery rate threshold q,

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

After determining $\widehat{m}$, we select the singular vectors from the top $\widehat{m}$ 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 subset selection and correlation calibration

Similar to SVA[6, 10-13], 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 a subset of 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 select columns (combinations of taxon and quantile levels) of R' that represent underlying data structure and less 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 X

and Z solely $Q_{\tau_k}(Y|X,Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) + Z_i\gamma(\tau_k)$ 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[22] 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 conveys 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 seek to select combinations of j and τ_k such that $p_{jk}^{(1)} < c_1$ and $p_{jk}^{(2)} > c_2$.

Choices of $c_1$and c_2 are crucial for QuanT. Our empirical results indicate that there are no universally optimal cutoffs and they have to be determined adaptively. Therefore, we consider a two-dimensional grid search using the candidate cutoffs 0.001, 0.01, and 0.1 (a total of nine combinations). For each combination of c_1 and c_2, we select those columns from R' that satisfy that $p_{jk}^{(1)} < c_1$ and $p_{jk}^{(2)} > c_2$ and 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 $d_s = \arg\max_{d} |\mathrm{corr}\left(\Phi_s, \Psi_d^{(1)}\right)|$. Across all choices of c_1 and c_2, we choose the one pair that maximizes an aggregated correlation score s, defined as $\underset{s}{\mathrm{median}}\, |\mathrm{corr}\left(\Phi_s, \Psi_{d_s}^{(1)}\right)|$, across all selected stage-1 singular vectors. 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 (e.g., 20) of iterations. In the second case, we calculate the final singular vector as a weighted average of the last 10 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

1. For each feasible pair of taxon j and quantile level τ_k, fit the quantile regression models $Q_{\tau_k}\left(Y_{\cdot j} \middle| X, Z\right) = \alpha_j(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_{lj}(\tau_k) + Z_i \gamma_j(\tau_k)$ to each column of the relative abundance matrix Y.

2. Threshold the microbial abundances against the estimated quantiles and center it as $I\{Y_{ij} > \hat{\alpha}_j(\tau_k) + \sum_{l=1}^{L} X_{il}\,\hat{\beta}_{lj}(\tau_{\mathrm{k}}) + Z_i\hat{\gamma}_j(\tau_k)\} - (1-\tau_k)$ to form columns of a binarized and centered residual matrix R.
3. Conduct SVD on $R = U\Sigma V$.
4. 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.
5. For each feasible pair of taxon j and quantile level τ_k, fit the reduced quantile regression model $Q_{\tau_k}(Y_{\cdot j}) = \alpha_j(\tau_k)$ to the relative abundance data Y and obtain the binarized and centered residual matrix correlated with X and Z, denoted as R'.
6. Iterative procedure for correlation calibration:

   For the first iteration, let $\Psi^{(1)} = \Phi$.

   For iteration u, denote the iterative singular vector as $\Psi^{(\mathrm{u})}$ with the s-th column as $\Psi_{\mathrm{s}}^{(\mathrm{u})}$:

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

Pair 2: $Q_{\tau_k}(Y|X,Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) + Z_i\gamma(\tau_k)$,

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

b. Select combinations of j and τ_k that $p_{jk}^{(1)} < c_1$ and $p_{jk}^{(2)} > c_2$. $c_1$and c_2 are selected based on the criteria depicted in the section above. Compose a reduced matrix by selecting columns of R' that satisfy the above criteria, denote it as $R^{(u)\prime}$.

c. Conduct SVD on $R^{(u)\prime}$.

d. 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 the iSV at iteration u as $\Psi^{(u)}$.

7. Repeat Step 6 iteratively until convergence or until a large number of iterations are achieved.

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

---

1-3. Step 1 – 3 is the same as the Algorithm 1.

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

5. 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.

6. Fit the models $Q_{\tau_k}\left(r_{\cdot j}^0 \middle| X, Z\right) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_l(\tau_k) + Z_i \gamma(\tau_k)$, and obtain the binarized and centered residual matrix, denote by R^{0*}

7. Conduct SVD of R^{0*} and calculate statistic $T_k^0 = \frac{d_{0k}^2}{\sum d_{0l}^2}$

8. 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 20.

9. Compute the p-value for singular vector *k* as

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

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

10. Select the singular vectors with $p_k \leq 0.1$.

11. Carry out correlation calibration (steps 6-7 in Algorithm 1)

---

290 ## Results

291 To rigorously assess QuanT, we conducted extensive simulation studies based on real-world
292 gut and virginal microbiome datasets from the HMP2 projects[23]. We also applied QuanT and
293 competing methods to three curated microbiome datasets, each of which has distinct sources
294 of technical heterogeneity: a gut microbiome study on cardiovascular diseases, which displays
295 a moderate technical variation arising from sequencing in multiple runs; an integrated dataset
296 compiled from various HIV gut microbiome studies; and a consolidated dataset encompassing
297 multiple colon cancer studies. Through both visual and quantitative comparisons, we
298 demonstrate that QuanT outperforms existing methods by robustly and comprehensively
299 uncovering underlying microbiome data heterogeneity, which is beyond mere shifts in means.
300 When incorporated into the analytical workflow, QuanT effectively reduces the influence of
301 unwanted heterogeneity, enhancing false discovery rate control, maintaining sensitivity of
302 discovery, as well as improving prediction accuracy.
303

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

305 We assessed QuanT using multiple datasets. We first applied it to a study containing
306 traditional batch variation: samples are collected under one protocol but handled in different
307 batches. The Coronary Artery Risk Development in Young Adults (CARDIA) Study[24],
308 initiated in 1985-1986, recruited young adults aged between 18 to 30 with the goal of
309 uncovering determinants of cardiovascular disease (CVD) throughout adult life. Subject
310 enrollment was balanced according to race, gender, education, and age. Extensive data on
311 risk factors related to CVD have been collected, including health behaviors, medical history
312 (including medication use), and clinical risk factors, such as blood pressure. Stool samples
313 were also collected for some visits. Here, we used the microbiome data generated from the

314 stool samples collected during the Year 30 follow-up (2015-2016). The samples were
315 processed through four DNA extraction/library preparation batches, and subsequently
316 sequenced through seven runs using the Illumina MiSeq 2x300 platform. Specifically, each
317 DNA extraction batch was allocated to two sequencing runs, except for the final batch, which
318 was sequenced by itself during the seventh run. Consequently, the data were processed in
319 seven distinct batches. The data were agglomerated to the genus level, and taxa that did not
320 yield any reads in all samples were omitted from the dataset. After preprocessing, the data
321 contains 633 samples and 145 taxa. Analyses were restricted to participants for whom all
322 covariate data were available.

323 In this study, we considered high blood pressure (HBP) (non-HBP = 0, HBP = 1) as
324 the variable of interest, and gender (Male = 0, Female = 1) and race (White = 0, Black = 1) as
325 adjustment covariates. To assess QuanT and the competing methods, we treated the batch
326 indicators as unknown and used these methods to identify the now-hidden
327 heterogeneities/batches. The performance of these methods was assessed based on three
328 criteria: 1) consistency and sensitivity in downstream differential abundance analysis for
329 individual taxa; 2) enhancement in prediction accuracy for the variable of interest; and 3) the
330 efficacy of batch effect removal while preserving the effect from the variable of interest at the
331 entire community level. For SVA, we used the svaseq function, which added a pseudo count
332 of 1 to all counts and then applied log transformation as described in Leek (2014)[13], and used
333 the iterative reweighted algorithm[10]. For RUV, we used the RUVg approach in Risso *et al.*
334 (2014)[7] with an estimated group of null taxa as controls, rather than RUVr and RUVs as these
335 approaches can only identify hidden directions that are uncorrelated with the variables of
336 interest.

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

Fig. 2c shows the prediction accuracy using the original and the ConQuR corrected microbiome data, where the corrected data were adjusted for TG, or surrogate variables estimated by SVA, RUV or QuanT. Random forest models were fit with five-fold cross-validation to predict HBP status, and the area under the receiver operating characteristic curve (ROC-AUC) was used to evaluate the accuracy. We note that this is not a genuine predictive analysis, but rather a metric to quantify how well the corrected microbiome data can explain the variable of interest. As shown in Fig. 2c, RUV and QuanT corrected data predicts HBP more accurately than using TG or SVA, and prediction using the unadjusted data (NC) leads to the worst accuracy. Fig. 2d shows the principal coordinate analysis (PCoA) plots using the

Bray-Curtis distance, comparing with the uncorrected and ConQuR-corrected datasets. Assessed by the PERMANOVA R^2 (which evaluates the multivariate association between the microbial community and a certain variable, here quantifies the proportion of microbiome data variability explained by the variable), correction with TG leads to the most significant reduction in batch effects, which was expected. Visually, it not only equalizes the mean and dispersion (centroid and size of the ellipses) across groups but also removes higher-order batch/group effects (shown by angles of the ellipses). Correction for QuanT, SVA, and RUV's estimated surrogate variables leads to comparable efficacy in reducing batch effects and maintaining the effects of high blood pressure (HBP).

It is worth noting that the CARDIA study has well-balanced batch designs, thus the analysis with NC leads to reasonable and comparable results as the analyses with adjustments.

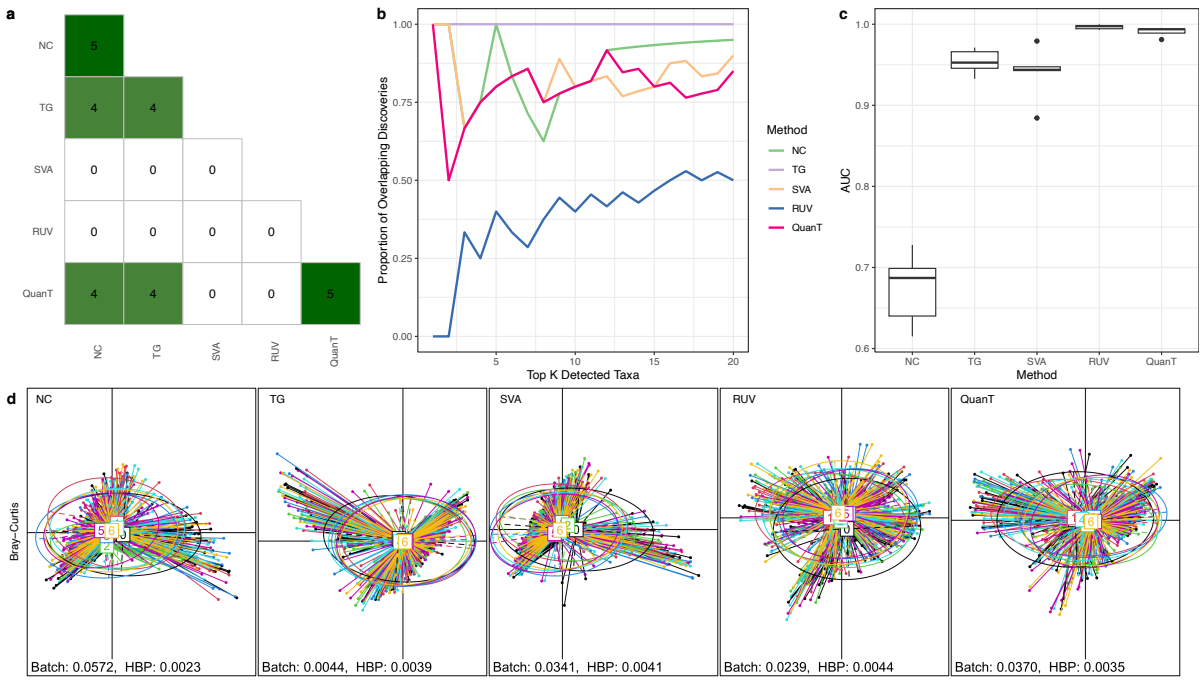

Fig. 2 | Evaluation on the CARDIA data. NC: no batch correction; TG: correction using true group indicators. **a** Heatmap representing the number of differentially abundant taxa identified in the CARDIA dataset by Benjamini-Hochberg adjusted p-value < 0.05; diagonal elements are numbers detected, off-diagonal 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 Resphera Insight[25]. Note that the studies in the final-processed data have distinct experimental designs and sequencing protocols (Supp. Tab. 2 of Tuddenham, et al. 2020[17]). Thus, the data have more pronounced variability than that observed within a single study, such as the CARDIA study.

We selected two studies with similar library sizes (5000-5500) but very different study designs (an observational study in the US[26] and a clinical trial in Europe[27]), which include 37 and 24 samples respectively. After preprocessing, the data contains 126 taxa. We applied QuanT, SVA, and RUV to the data aggregated at the genus level. HIV status (Negative = 0, Positive = 1) was regarded as the variable of interest, while age, gender and BMI were considered as covariates. We retained complete cases only and conducted the same set of analyses as on the CARDIA data.

401 Fig. 3a and 3b compare the consistency of differential abundance analysis results
402 obtained analyzing these data using various methods. Fig. 3a shows the number of
403 differentially abundant taxa identified in the HIVRC dataset by Benjamini-Hochberg adjusted
404 p-value < 0.05. We identified 17 discoveries adjusting for true group indicators (TG), which
405 were considered as the gold standard. QuanT leads to 15 true discoveries and 3 false discoveries.
406 NC leads to 9 true discoveries and 16 false discoveries. SVA leads to 9 true discoveries and 20
407 false discoveries. RUV leads to 2 true discoveries and 7 false discoveries. Therefore, QuanT
408 again performs better than the other methods when applied to the HIVRC dataset and has the
409 lowest FDR compared to competitors. In Fig. 3b, we consider the top K detected differentially
410 abundant taxa. QuanT is the most consistent with the gold standard, TG, followed by NC, SVA,
411 and RUV.
412 Fig. 3c shows the average cross-validated ROC-AUC of predicting HIV from the
413 ConQuR-corrected data via random forest. TG, SVA, RUV, and QuanT all perform perfectly
414 and are better than no correction. Fig. 3d shows the PCoA plots using Bray-Curtis distance,
415 comparing with the uncorrected and ConQuR-corrected datasets. Visually, TG, RUV and
416 QuanT lead to better harmonization of the microbiome data from the two distinct HIV studies.
417 As quantified by PERMANOVA R^2, QuanT leads to the greatest reduction of batch effects,
418 followed by RUV, SVA and TG. On the other hand, SVA maintains the HIV effects the most,
419 followed by TG, QuanT and RUV (figure 3d).
420

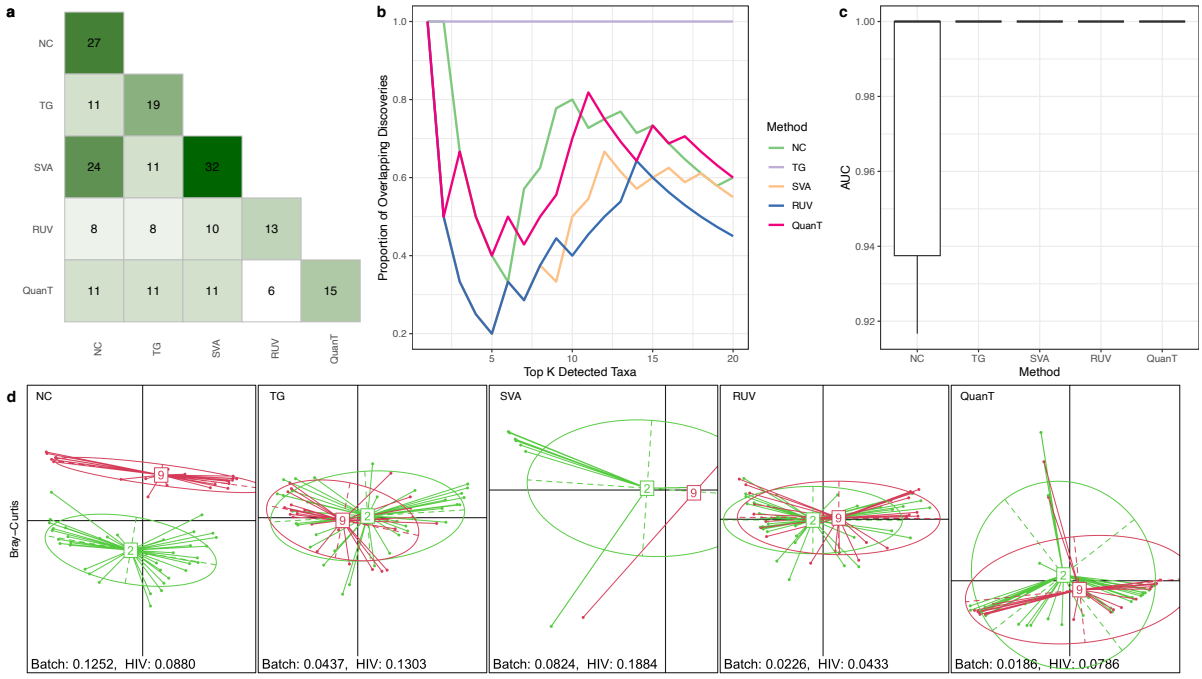

Fig. 3 | Evaluation on the HIVRC data. NC: no batch correction; TG: correction using true group indicators. **a** Heatmap representing the number of differentially abundant taxa identified in the HIVRC dataset by Benjamini-Hochberg adjusted p-value < 0.05; diagonal elements are numbers detected, off-diagonal 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.

Application to an integrated microbiome dataset in colon cancers

Our third example came from two colorectal cancer (CRC) studies[28, 29] from the curatedMetagenomicData package[30]. The CuratedMetagenomicData data package offers standardized, curated, and analysis-ready metagenomic datasets derived from human microbiome studies, facilitating comprehensive meta-analyses and cross-study comparisons by ensuring consistent data processing and annotation. We selected two CRC studies[28, 29] from

437 this package, which include 154 and 80 stool samples, respectively. After filtering, 166 genera
438 were included in the data. These studies collectively cover samples from three distinct health
439 conditions: colorectal cancer (CRC), adenoma, and disease-free controls, with CRC subdivided
440 into adenocarcinoma and carcinoma, and adenoma into simple adenoma and advanced
441 adenoma.

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

447 Fig. 4a and 4b compare the consistency of differential abundance analysis results across
448 various methods. Fig. 4a shows the number of differentially abundant taxa identified in the
449 CRC dataset by Benjamini-Hochberg adjusted p-value < 0.05. TG (gold standard) leads to 8
450 discoveries. QuanT leads to 4 true discoveries with no false discovery. NC leads to 6 true
451 discoveries and 24 false discoveries. SVA leads to 6 true discoveries and 23 false discoveries.
452 RUV leads to 1 true discovery with no false discovery. Therefore, NC and SVA lead to inflated
453 false discovery proportion when applied to the CRC dataset. RUV and QuanT have reasonable
454 false discovery proportion control, while QuanT is notably better at identifying true signals
455 than RUV. In Fig. 3b, we consider the top K detected differentially abundant taxa. QuanT is
456 the most consistent with the gold standard, TG. NC, SVA, and RUV perform similarly are all
457 outperformed by QuanT.

458 Fig. 4c presents the prediction accuracy, measured by negative cross entropy, for the
459 five phenotype groups using the original and ConQuR-corrected CRC microbiome data.
460 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.

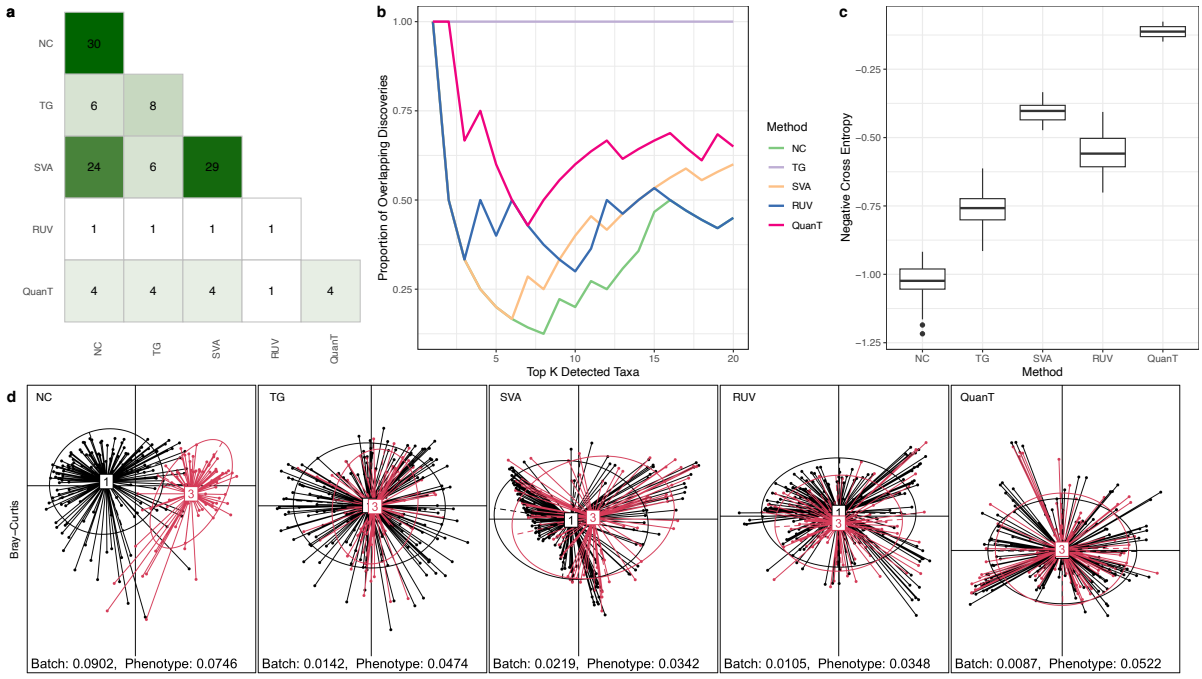

Fig. 4 | Evaluation on the CRC data. NC: no batch correction; TG: correction using true group indicators. **a** Heatmap representing the number of differentially abundant taxa identified in the CRC dataset by Benjamini-Hochberg adjusted p-value < 0.05; diagonal elements are numbers detected, off-diagonal 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 CRC dataset. **c** Average cross-validated negative cross entropy 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 Phenotype.

Evaluation using simulated data

477 We also conducted simulations to evaluate QuanT's performance when the true signal is known.
478 We simulated data that mimicked two datasets from the integrative Human Microbiome
479 Project[23], a gut microbiome dataset from the Inflammatory Bowel Disease Multi-omics
480 Database (IBDMDB) project[31], and a vaginal microbiome dataset from the Multi-Omic
481 Microbiome Study: Pregnancy Initiative (MOMS-PI) project[32]. These two datasets represent
482 microbiome data from two body sites that are of great research interest. They also have
483 contrasting attributes in terms of sparsity and over-dispersion level, which makes them ideal
484 for evaluating our proposed method QuanT. We used two methods to simulate data that
485 mimicked the template data: the parametric Dirichlet-Multinomial (DM) model and the non-
486 parametric approach in MIDASim[33].

487 For each sample i, we simulated a clinical variable of interest (denoted by Z_i) and a
488 hidden batch variable (denoted as B_i) through a bivariate Bernoulli distribution, with both
489 variables having a marginal mean of 0.5. A correlation of 0.3 was set between them. Note that
490 because the "batch" variable is simulated to be correlated with both the microbiome data and
491 Z, thus, it is a confounder. Here, we use the "batch" variable to represent all the hidden
492 heterogeneities.

493 For the parametric simulations, we adopted the simulation setup from a previous study[34]
494 to simulate taxa abundances using the Dirichlet-Multinomial (DM) distribution. In our
495 simulations, we designated the batch variable to impact most taxa (referred to as Batch Taxa),
496 while the variable of interest, Z, influenced only a small subset (referred to as Clinical Taxa).
497 Specifically, for the IBDMDB dataset, we randomly selected 250 of the 301 taxa to be Batch
498 Taxa, among which 40 were also Clinical Taxa. Similarly, for the MOMS-PI dataset, we
499 selected 200 of the 233 taxa as Batch Taxa, with 50 of these being Clinical Taxa. We first fitted
500 the DM model to the template datasets (IBDMDB and MOMS-PI) and obtained the estimated

mean vectors π_0. Subsequently, we formed two additional vectors π_B and π_Z to incorporate batch and clinical effects. Both vectors were initially set to be the same as π_0. We then randomly permuted the frequencies within π_B and π_Z corresponding to the Batch Taxa and Clinical Taxa. Finally, a sample-specific frequency was defined as $\pi_i = (1 - w_B B_i - w_z Z_i)\pi_0 + w_B B_i \pi_B + w_z Z_i \pi_z$, 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)[33], 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.

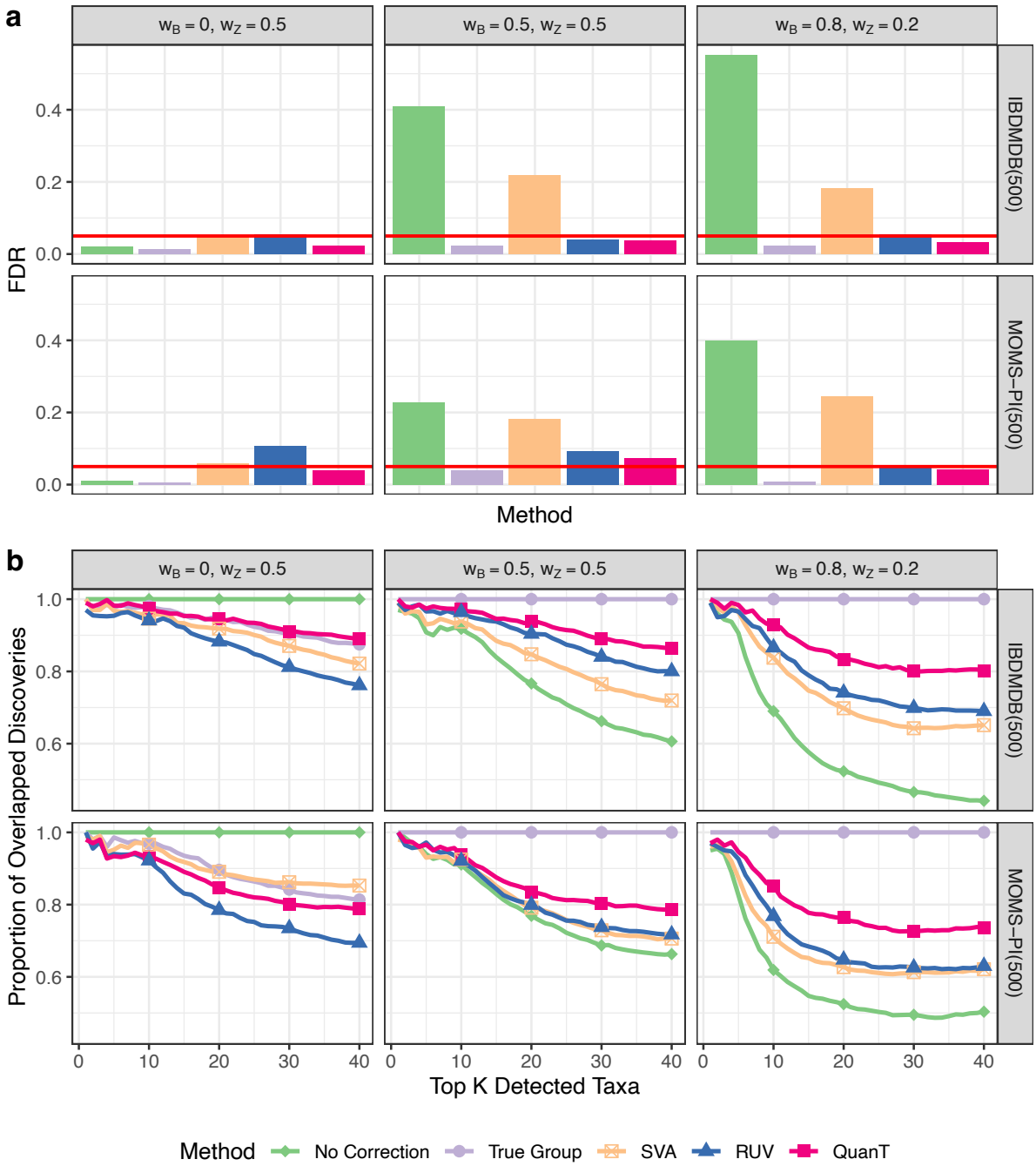

Fig. 5 | Evaluation on the simulated data. The data are simulated by MIDASim using the IBDMDB and MOMS-PI datasets as templates. **a** FDR of the differential abundance analysis via adjusted linear regression with BH-adjusted p-value < 0.05. **b** Proportions of overlapping taxa in the top K detected 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).

527

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

539 We also applied ConQuR to remove the true batch, SVs and QSV effects and generate
540 a "batch-free" dataset. Integrating QuanT with ConQuR in simulations (Supp. Fig. 7-10)
541 yielded microbial profiles where confounding batch effects were eliminated, while the effects
542 of the variable of interest were preserved. Finally, QuanT demonstrates scalable computational
543 performance for large-scale microbiome studies, as evidenced in Tab. 1 and Tab. 2.

544 In short, QuanT demonstrates improved performance compared to existing methods,
545 on simulated datasets and biological datasets, regardless of the effect size and the type of source
546 of unmeasured heterogeneity.

547

548 **Computation of QuanT**

549 We have streamlined the process of quantile regression and matrix decomposition within
550 QuanT by employing strategies for enhanced efficiency. The algorithm executes quantile

551 regression across various taxon and quantile levels independently, which lends itself well to
552 parallel processing. We leverage the doParallel package in R to expedite this computation. In
553 the second stage of QuanT, where SVD is performed, we focus on a few leading singular
554 vectors because it is generally improbable for a singular vector that accounts for minimal
555 variance to represent a critical direction established in the first stage. Consequently, it is
556 sufficient to compute only the leading singular vectors of the binarized residual matrix during
557 the correlation calibration phase. For this reason, QuanT only calculates the top 15 principal
558 components using the RSpectra R package.

559 Tables 1 and 2 present the computational time of QuanT for datasets with different
560 numbers of samples and taxa. Table 1 illustrates the scalability of QuanT's computational time
561 with respect to various numbers of samples (n) and taxa (p), benchmarked against SVA and
562 RUV. For each configuration of n and p, we generated ten datasets using the DM simulation
563 with parameters $(w_B, w_z) = (0.5, 0.25)$ and the IBDMDB dataset as the template. We also
564 compared the computational time of QuanT, SVA, and RUV for analyzing the three real-world
565 datasets (CARDIA, HIVRC, and CRC) (Tab. 2). Although QuanT exhibits slightly slower
566 computation than SVA and RUV, it operates at a comparable speed in general and offers
567 computational efficiency within acceptable limits.

568

569 **Table 1 | The computation time (in seconds) of QuanT with different numbers of samples**
570 **and taxa, compared with SVA and RUV. The data are simulated from the Dirichlet-**
571 **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	1.3	0.9	0.4	3.2	1.5	0.6	6.0
n=500	1.5	0.5	1.9	2.7	0.7	4.2	4.2	1.0	6.2
n=750	4.0	0.6	3.1	6.0	1.3	9.7	8.5	1.6	16.1
n=1000	8.5	0.8	3.6	11.7	1.4	13.4	15.0	1.9	23.7

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	12.9
HIVRC	61	126	0.1	0.6	1.4
CRC	234	166	0.9	5.9	1.1

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 specifically tailored to microbiome data, researchers often use methods developed for genomic studies. However, these methods do not account for the challenging characteristics of

584 microbiome data, such as sparsity and overdispersion. In addition, they focus solely on the
585 information underlying the mean expression and ignore other aspects of the data distribution.
586 As shown the "Results" section, these methods lead to increased false discoveries and/or
587 reduced power when analyzing microbiome studies.

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

596 We applied QuanT to three microbiome datasets and a wide range of simulated datasets
597 to assess its performance, in comparison with competing methods. The microbiome datasets
598 include large-scale epidemiological studies with samples processed in different sequencing
599 runs, and curated microbiome datasets integrated from various publicly availabe studies. Our
600 simulations are comprehensive, generating microbiome profiles that mimic real-world
601 microbiome studies in gut and vaginal environments and using two distinct simulation
602 mechanisms. Visually and numerically, we showed that QuanT improved subsequent
603 differential abundance analysis, offering better FDR control and improved statistical power,
604 and providing results that align better with the gold standard in most setups. Frequently, the
605 results of differential abundance analysis following QuanT are comparable to those when the
606 true labels of batches or studies are known. It also offers comparable or improved performance

607 in predicting the variable of interest, as well as reducing the impact of hidden heterogeneity
608 while preserving key signals at the community-level.

609 Despite its advantages, QuanT has several limitations and potentials that could be
610 investigated in future research. Firstly, there might be certain underlying structure in the
611 samples, which should be well-specified to be addressed by QuanT. For instance, it is common
612 that microbiome studies are conducted in a longitudinal manner to investigate the dynamics of
613 the microbiome. Such studies typically produce abundance data collected across multiple time
614 points from the same individuals. QuanT makes no distributional assumptions, including
615 sample independence assumption. Thus, it is possible to incorporate time variables as
616 covariates in its regression models, hence it can be robust to longitudinal microbiome data.
617 However, it can be beneficial to consider the temporal nature of the data and use such
618 information explicitly, rather than treating the time variable in the same way as the other
619 covariates. Secondly, QuanT leads to well controlled false discovery rate in our real and
620 extensive simulation experiments, although it lacks theoretical guarantee. However, theoretical
621 guarantees might be derived given strict assumptions about the data generation mechanism,
622 which, in practice, is usually less important. Finally, a potential extension of QuanT is to
623 identify unmeasured heterogeneity in gene expression data, especially in single cell and spatial
624 transcriptomics data which have similar sparsity and over-dispersion as do microbiome data.
625 It would be worth investigating how QuanT performs when applied to these data types, which
626 might require some adaptations. However, these aspects are beyond the scope of this paper,
627 and we leave them to future studies.

628

629 **Conclusions**

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

List of abbreviations

QuanT: quantile thresholding

SVA: surrogate variable analysis

RUV: remove unwanted variation

SVD: singular value decomposition

iSV: iterative singular vectors

QSV: quantile surrogate variables

VOI: variable of interest

CARDIA: Coronary Artery Risk Development in Young Adults

CVD: cardiovascular disease

HBP: high blood pressure

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[31] and the vaginal microbiome dataset from Multi-Omic

676 Microbiome Study[32]: Pregnancy Initiative (MOMS-PI) project, which were used as templates
677 for simulation studies, are provided in the Bioconductor R package, HMP2Data[35].
678 The R package QuanT is available at https://github.com/jiuyaolu/QuanT in formats
679 appropriate for Macintosh, Windows, or Linux systems. A vignette demonstrating use of the
680 package is included (can be accessed at https://jiuyaolu.github.io/QuanT/QuanT.html).
681

682 **Competing interests**

683 The authors declare that they have no competing interests.
684

685 **Funding**

686 This work was supported in part by R01GM147162 (Lu, Satten, and Zhao), R01HL155417
687 (Ling), R01GM151301 (Ling) and R01GM155734 (Ling). The Coronary Artery Risk
688 Development in Young Adults Study (CARDIA) is supported by contracts 75N92023D00002,
689 75N92023D00003, 75N92023D00004, 75N92023D00005, and 75N92023D00006 from the
690 National Heart, Lung, and Blood Institute (NHLBI). The HIVRC data used in this study are
691 from work that was supported by the HIV Microbiome Re-analysis Consortium.
692

693 **Authors' contributions**

694 JL, GS, NZ and WL developed the method, analyzed the data, and wrote the manuscript. JL
695 prepared the software package and vignette. LL and KM contributed to the dataset generation
696 and manuscript development with technical input. All authors read and approved the final
697 manuscript.
698

699 **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. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**, e161 (2007).
7. 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).
8. Koenker, R. & Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50 (1978).
9. Ling, W. et al. Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications* **13**, 1-14 (2022).
10. Leek, J.T. & Storey, J.D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105**, 18718-18723 (2008).
11. 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).
12. 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).
13. Leek, J.T. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research* **42**, e161-e161 (2014).
14. Koenker, R. & Hallock, K.F. Quantile regression. *Journal of economic perspectives* **15**, 143-156 (2001).
15. Barber, R.F. & Candès, E.J. Controlling the false discovery rate via knockoffs. (2015).

16. Goldstein, S., Lebowitz, J.L., Tumulka, R. & Zanghi, N. Any orthonormal basis in high dimension is uniformly distributed over the sphere. (2017).
17. 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).
18. 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).
19. Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**, 479-498 (2002).
20. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics* **31**, 2013-2035 (2003).
21. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 (2003).
22. 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).
23. 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).
24. Friedman, G.D. et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology* **41**, 1105-1116 (1988).
25. 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).
26. Lozupone, C.A. et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* **14**, 329-339 (2013).
27. 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).
28. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature communications* **6**, 6528 (2015).
29. 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).
30. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023-1024 (2017).
31. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662 (2019).
32. Fettweis, J.M. et al. The vaginal microbiome and preterm birth. *Nature medicine* **25**, 1012-1021 (2019).
33. He, M., Zhao, N. & Satten, G.A. MIDASim: a fast and simple simulator for realistic microbiome data. *Microbiome* **12**, 135 (2024).

34. Hu, Y.-J. & Satten, G.A. Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics* **36**, 4106-4115 (2020).
35. 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).