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[1]

Abstract

Microbiome data exhibit technical and biomedical heterogeneity due to varied processing and experimental designs, which may lead to spurious results if uncorrected. Here, we introduce the Quantile Thresholding (QuanT) method, a comprehensive non-parametric hidden variable inference method that accommodates the complex distributions of microbial read counts and relative abundances. We apply QuanT to synthetic and real data sets and demonstrate its ability to identify unmeasured heterogeneity and improve downstream analysis.

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

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, 21205 Baltimore, USA. [2]Department of Gynecology and Obstetrics, Emory University School of Medicine, 101 Woodruff Circle, 30322 Atlanta, USA. [3]Nutrition Research Institute and Department of Nutrition, University of North Carolina, 500 Laureate Way, 28081 Kannapolis, USA. [4]Laboratory of Epidemiology and Population Science, NIA, NIH, 7201 Wisconsin Ave, 20814 Bethesda, USA. [5]Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, 402 East 67th Street, 10065 New York, USA.

Next-generation sequencing technology has facilitated large-scale microbiome profiling to explore associations with various diseases and health outcomes. Given the large scale, samples must be collected and profiled in different runs or even from various studies. Inevitably, unwanted variation arises from the differences in experimental/bioinformatics procedures. Much of this unwanted variation stems from hidden technical or biomedical heterogeneity, such as lost lab labels, instrumental drifts, and undocumented differences in study designs. Failing to account for these hidden heterogeneities can obscure true signals.

Currently, no methods are designed to address unmeasured heterogeneity in microbiome studies, although techniques developed for genomic data, such as surrogate variable analysis (SVA)[1] and remove unwanted variation (RUV)[2], may be applicable. SVA is a two-step procedure for identifying 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 heterogeneities. However, SVA and RUV may not be ideally suited for microbiome data. First, they do not address the inflated zeros and over-dispersion typical in microbiome data. Second, they detect the underlying structure in the (possibly log-transformed) mean expression while ignoring other critical aspects of the complex data. Third, RUV requires users to specify the number of hidden directions: incorrect specification of these parameters may result in false positives or negatives.

Here, we present the Quantile Thresholding (QuanT), the first comprehensive tool for unmeasured heterogeneity identification in microbiome data. QuanT uses quantile regression[3] at multiple quantile levels, and thresholds the abundances against the estimated quantiles to obtain dichotomized residuals. QuanT distills information about the underlying data structure and outputs a set of unit vectors that represents the hidden heterogeneity. As a non-parametric approach, QuanT is robust to the complex distributional attributes of microbiome data, handles zero-inflated data in a natural way, and can detect higher-order heterogeneity beyond simple mean shifts. Fig. 1 illustrates the workflow of QuanT.

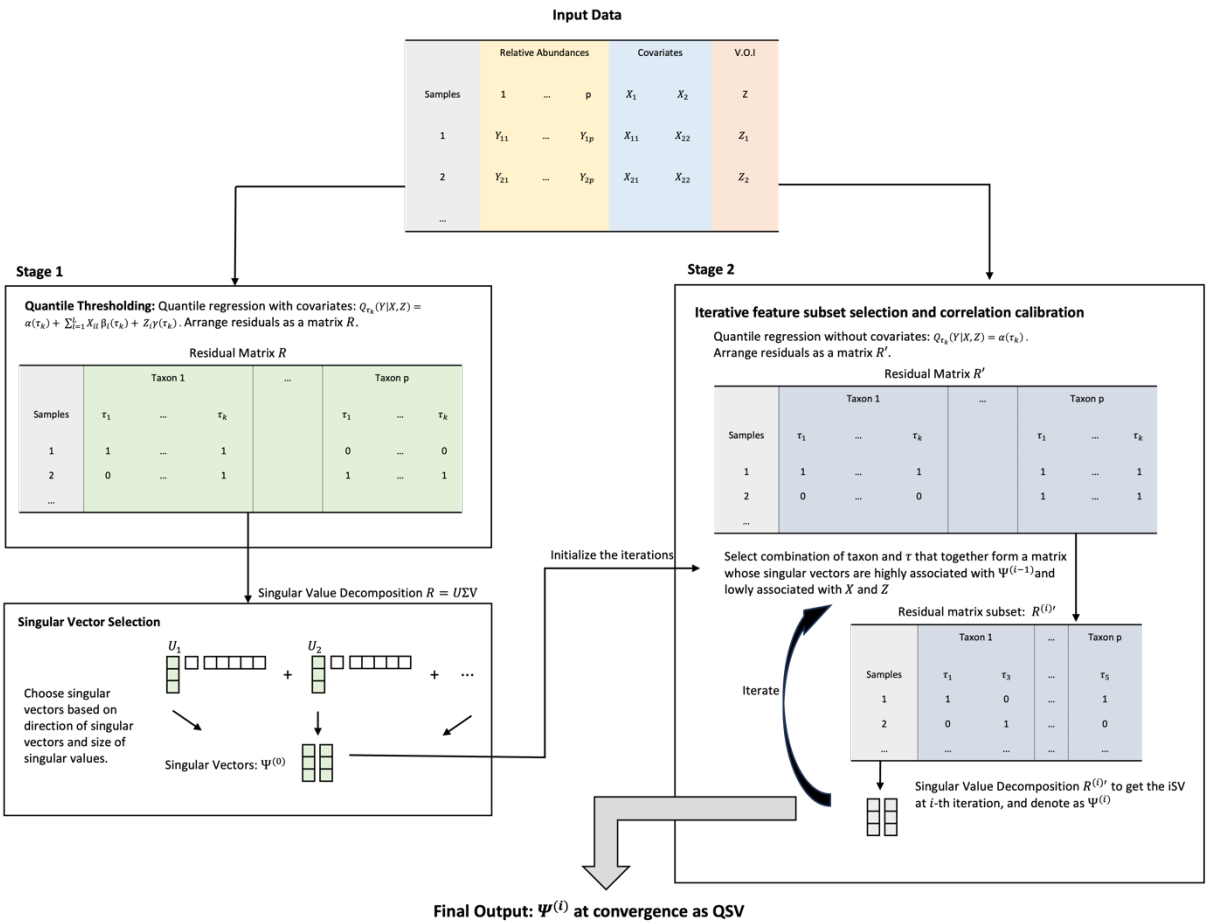

Figure 1: Flowchart of QuanT. V.O.I: variable of interest; QSV: quantile surrogate variable. Stage 1 conducts quantile regression for each taxon at multiple quantile levels using V.O.I and covariates. Residuals are thresholded to create a binary matrix R, which is analyzed through singular value decomposition 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 across multiple levels, followed by the selection of taxon and quantile levels so that the singular vectors of the reduced residual matrix subset $R^{(i)\prime}$ are highly associated with the prior stage singular vectors but minimally associated with X and Z. Stage 2 is an iterative procedure and the singular vectors at convergence of Stage 2 are QSVs, the output of QuanT.

The QuanT output, namely quantile surrogate variables (QSVs), 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 covariate adjustment. Alternatively, QuanT can be combined with the recently-developed microbiome batch effect removal method, ConQuR[4], to produce batch-free microbial profiles for downstream analysis.

We applied QuanT to three datasets involving multiple known groupings (batches or studies) to demonstrate its substantial improvement of downstream analyses. In each dataset, we focused on one variable of interest and aimed to identify taxa associated with this variable while controlling other covariates. Treating the group indices as unknown, we utilized QuanT to identify QSVs, SVA for Surrogate Variables (SVs), and RUV for RUV factors. Subsequently, we conducted linear regression on relative abundances, adjusting for the hidden variables identified by each method. For benchmarking, we performed linear regression adjusted for the true group indicators (as the gold standard) and without any adjustment (as a baseline result).

Data from the Coronary Artery Risk Development in Young Adults (CARDIA) Study[5] showcased classic batch variation with 633 samples processed in the same study in 7 different batches. We considered high blood pressure (HBP) as the variable of interest and gender and race as adjustment covariates. Our second example used data from the HIV re-analysis consortium (HIVRC)[6], which consolidated multiple 16s rRNA sequencing studies through the same bioinformatics pipeline (Resphera Insight). Unlike the CARDIA data, the HIVRC dataset featured greater diversity due to variations in study designs/protocols. We analyzed two studies from HIVRC due to their significant differences (an observational study in the US[7] and a clinical trial in Europe[8]), with HIV status as the variable of interest, and age, gender, and BMI as covariates. Our third example came from two colorectal cancer (CRC) studies[9, 10] from the curatedMetagenomicData package[11], focusing on different patient phenotypes (adenocarcinoma, adenoma, advanced adenoma, carcinoma, and control) as the variable of interest, adjusted by age and gender.

Figures 2a and 2b compare the consistency of differential abundance analysis results across various models: the model with true group adjustment (TG), the baseline model with no correction (NC), and models adjusting for hidden variables identified by each method. QuanT consistently outperforms SVA and RUV, aligning more closely with the TG model, showing fewer false discoveries and more true positives (Fig 2a). In the CARDIA study, QuanT's performance is comparable to the NC model, likely due to its balanced design. Additionally,

in the HIVRC and CRC datasets, QuanT surpasses the NC approach, which tends to yield false discoveries.

We also conducted simulations to evaluate QuanT's performance. We simulated microbiome data that are impacted by both batch and covariate, in which the batch and the variable of interest can be correlated, so that batch is a confounder. Details of the simulation can be found in Online Methods (Simulation Studies Section). Like the CARDIA, HIVRC, and CRC analysis, we used linear regression models to conduct differential abundance analysis, adjusting for hidden variables detected by SVA, RUV, and QuanT. QuanT aligned more closely with the gold standard model than SVA, RUV, or the NC model (Fig 2c). Here, the gold standard varied: in the absence of batch effects, NC was the gold standard, while when batch effects were present, the model correcting for true batch ID was the gold standard. Supp. Fig. 1-5 confirm that QuanT effectively controls the false discovery rate and closely matches the gold standard. Furthermore, integrating QuanT with ConQuR in simulations (Supp. Fig. 6-9) yielded batch-free microbial profiles while preserving the impact of relevant variables and eliminating confounding effect. Finally, QuanT demonstrates scalable computational performance for large-scale microbiome studies, as evidenced in Supp. Tab. 1 and Supp. Tab. 2.

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.

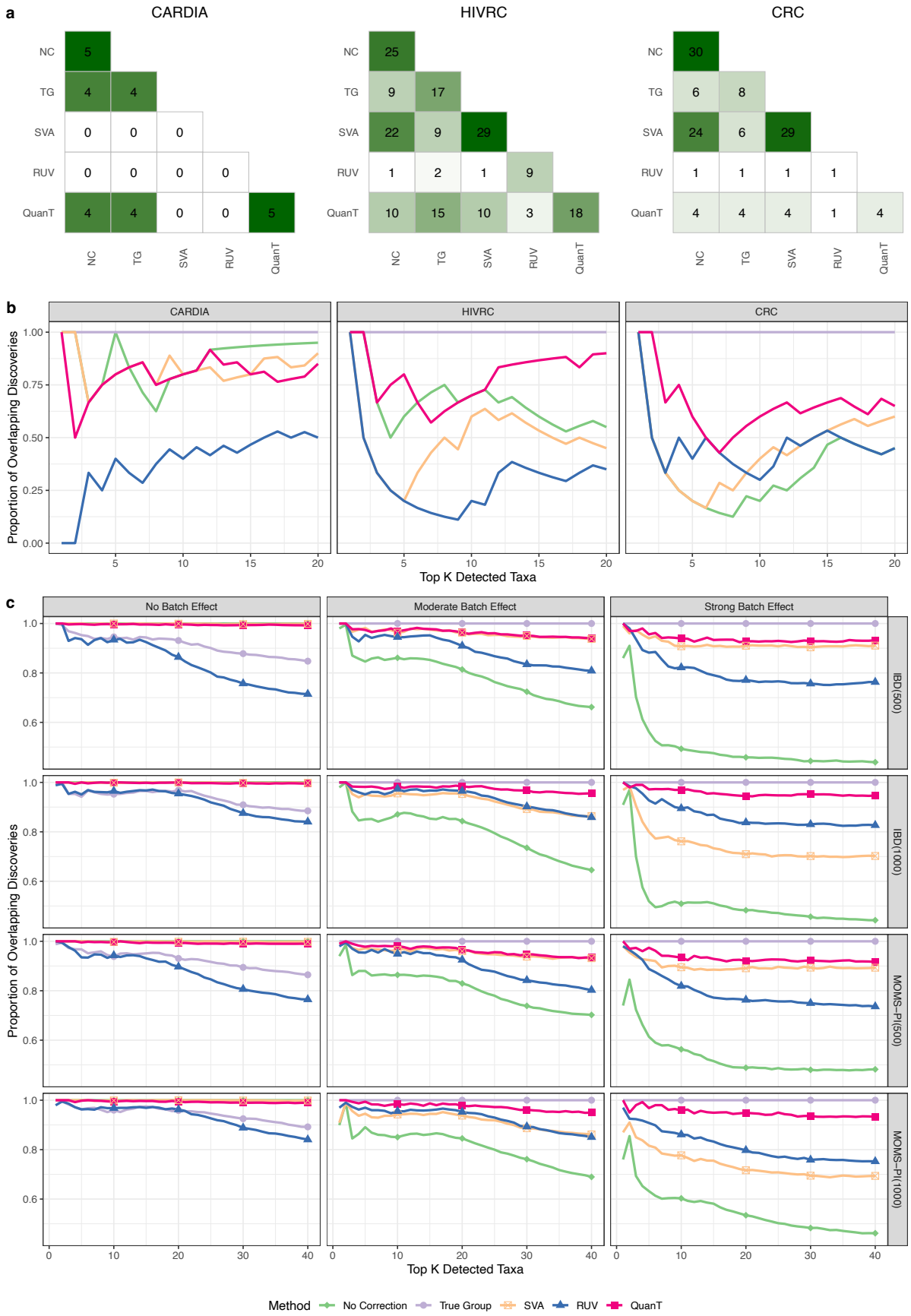
a
CARDIA
HIVRC
CRC
NC
TG
SVA
RUV
QuanT
b
Proportion of Overlapping Discoveries
Top K Detected Taxa
c
No Batch Effect
Moderate Batch Effect
Strong Batch Effect
IBD(500)
IBD(1000)
MOMS−PI(500)
MOMS−PI(1000)
Method
No Correction
True Group
SVA
RUV
QuanT

Figure 2: QuanT removes the hidden heterogeneity more successfully than SVA and RUV in real and simulated datasets. NC: no batch correction, TG: correction using true group indicators. a. heatmap representing the number of differentially abundant taxa identified in CARDIA, HIVRC, and CRC by Benjamini-Hochberg adjusted p-value < 0.05; diagonal elements are number detected, off-diagonal elements are number detected in common. b & c Proportions of overlapping taxa in the top K detected differentially abundant taxa between each analysis in the CARDIA, HIVRC, and CRC datasets (b) and in simulated data (c). In panel c, rows (columns) represent the simulation setups with different datasets and sample sizes (batch effect).

Data availability

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[12] and the vaginal microbiome dataset from Multi-Omic Microbiome Study[13]: Pregnancy Initiative (MOMS-PI) project, which we use as templates for our simulation, are provided in the Bioconductor R package, HMP2Data.

Code availability

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).

Acknowledgements

This work was supported in part by R01GM147162 (Lu, Satten, and Zhao), R01HL155417 (Ling) and R01GM151301 (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. 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.

Author 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.

Competing interests

The authors declare no competing interests.

References

1. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**, e161 (2007).
2. 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).
3. Koenker, R. & Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50 (1978).
4. Ling, W. et al. Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications* **13**, 1-14 (2022).
5. Friedman, G.D. et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology* **41**, 1105-1116 (1988).
6. 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).
7. Lozupone, C.A. et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* **14**, 329-339 (2013).
8. 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).
9. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature communications* **6**, 6528 (2015).
10. 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).
11. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023-1024 (2017).
12. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662 (2019).
13. Fettweis, J.M. et al. The vaginal microbiome and preterm birth. *Nature medicine* **25**, 1012-1021 (2019).

Online Methods of "Identifying unmeasured heterogeneity in microbiome data via quantile thresholding (QuanT)"

Jiuyao Lu, Glen A. Satten, Katie A. Meyer, Lenore J. Launer, Wodan Ling, Ni Zhao

Methods

Overview of QuanT

We propose a two-stage algorithm to detect the unmeasured heterogeneity in microbiome relative abundance data (n samples, p taxa). In the first stage, we fit quantile regression models at multiple quantile levels to the relative abundance matrix with the primary variable of interest and important covariates and generate dichotomized, centered quantile residuals. These are concatenated into a single matrix of discretized residuals R. We then construct 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 known variables.

The stage-1-singular vectors are unrelated to the known variables as they were included in the quantile regression. As a result, adjusting for these singular vectors in downstream analysis could 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) which are strongly marginally associated with the stage-1 singular vectors but are weakly associated with the covariates. Our second stage is an iterative procedure that, in each iteration, we calculate new "iterative" singular vectors (iSVs) based on residuals from the selected features. The iSVs at convergence, i.e., when the resulting iSVs are the same as the previous iteration, are called Quantile Surrogate variables (QSVs), which are the output of our algorithm. This correlation calibration step has been shown to be necessary in similar tasks when analyzing gene expression data[1-5].

Notation

Assume we have an $n \times p$ matrix Y, in which the element Y_{ij} indicates the relative abundance of the j-th taxon in the i-th sample. Note that Y is a zero-inflated relative abundance variable, and we treat it as the outcome in the proposed regression framework. Suppose that each sample i has a primary variable of interest (denoted by Z_i) and a set of L covariates (denoted by X_i), that we want to adjust for. We use K quantile levels in our regression framework, represented by the vector, $\vec{\tau}$. By default, $K = 5$, and $\vec{\tau}$ is the 10th, 25th, 50th, 75th, and 90th quantiles of the proportion of zeroes across all taxa in the data. For instance, if the proportions of zeros across all taxa range from 0.2 to 0.95, the calculated quantiles may yield $\vec{\tau} = (0.3, 0.5, 0.56, 0.72, 0.9)$. Given the zero-inflated nature of microbiome data, 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 thus disregarded for this taxon. Thus, the effective length of $\vec{\tau}$ varies by taxon, allowing us to extract relevant information and elucidate the underlying structure of the microbiome data distribution. 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 $Q_{\tau_k}(Y|X,Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) + Z_i\gamma(\tau_k)$ to the data Y 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 $(\hat{\alpha}(\tau_k), \hat{\beta}(\tau_k), \hat{\gamma}(\tau_k))$ are obtained by solving the minimization problem

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

where ρ_{τ_k} is the check function[6] $\rho_{\tau_k(x)} = \tau_k x - I\{x < 0\}x$.

After fitting the quantile regression model, we threshold the microbial abundances against the estimated quantiles and center them as $I\{Y_{ij} > \hat{\alpha}(\tau_k) + \sum_{l=1}^{L} X_{il}\,\hat{\beta}_l(\tau_k) + Z_i\hat{\gamma}(\tau_k)\} - (1 - \tau_k)$. For one 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 like 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 that to obtain

$$R = U\Sigma V.$$

We aim at selecting a subset of the singular vectors, i.e., columns of U, that represent the important structure in the data. We use an integrated approach to select the singular vectors. 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 centered residual matrix (a.k.a, complete null) discussed in the next paragraph, and 2) which explain a large proportion of variability of the residual matrix. We use the SelectiveSeqStep procedure in the knockoff paper[7] to achieve this goal based on the two criteria at the same time. The detailed rationale and procedure are as follows.

Under the complete null that there is no underlying structure in the 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[8]. 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[9] or the Anderson-Darling test[10] 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[11-13], 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 the variability in the data.

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

Stage 2: Iterative feature subset selection and correlation calibration

Like SVA[1-5], we conduct a second stage to refine the stage-1-singular-vectors that we call correlation calibration. The reason lies in that, by regressing out X and Z to construct the binarized residuals, the selected singular vectors are, loosely speaking, uncorrelated with the primary variable Z and covariates X. Thus, they are not able to identify unmeasured confounders, as confounders must be correlated with Z. Nonetheless, it is these 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 simplistic quantile regression model $Q_{\tau_k}(Y)=\alpha(\tau_k)$, 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 using an iterative procedure that balances the two goals of 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 simplistic quantile model $Q_{\tau_k}(Y) = \alpha(\tau_k)$ with the model that includes the stage-1-singular-vectors: $Q_{\tau_k}(Y) = \alpha(\tau_k) + \sum_{s=1}^{\hat{m}} \Phi_s \theta_s(\tau_k)$. The second pair compares model with the stage-1-singular-vectors with a full model that $Q_{\tau_k}(Y) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_l(\tau_k) + Z_i \gamma(\tau_k) + \sum_{s=1}^{\hat{m}} \Phi_s \theta_s(\tau_k)$. 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 are valuable for identifying 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[14] 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 data from 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. Thus, we seek to select combinations of j and $\tau_k$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 of using the following 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 of 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 of Φ_s,we choose the d_s -th column of $\Psi^{(1)}$ in which $d_s = \arg\max_{\mathrm{d}} |\mathrm{corr}\left(\Phi_s, \Psi_d^{(1)}\right)|$.

Across all choices of c_1 and c_2, we choose the one pair that maximizes an aggregated correlation score across all selected stage-1-singular-vectors s $\underset{s}{\mathrm{median}}\, |\mathrm{corr}\left(\Phi_s, \Psi_{d_s}^{(1)}\right)|$. 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 earlier iteration and follow the same procedure. The iteration stops when iSVs from two adjacent iterations are the same, 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 QuanT algorithm for individual steps.

QuanT Algorithm

1. For each feasible pair of taxon j and quantile level τ_k, fit the quantile regression model $Q_{\tau_k}(Y|X,Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_l(\tau_k) + Z_i\gamma(\tau_k)$ to relative abundance data Y.
2. Threshold the microbial abundances against the estimated quantiles and center it as $I\{Y_{ij} > \hat{\alpha}(\tau_k) + \sum_{l=1}^{L} X_{il}\, \widehat{\beta}_l(\tau_\mathrm{k}) + Z_i\hat{\gamma}(\tau_k)\} - (1-\tau_k)$ to form columns of 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 stage-1-singular-vectors Φ.
5. For each feasible pair of taxon j and quantile level τ_k, fit the simplistic quantile regression model $Q_{\tau_k}(Y|X,Z) = \alpha(\tau_k)$ to relative abundance data Y and obtain the binarized and centered residuals, denoted as R'.

6. Iterative procedure for correlation calibration:

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

   For iteration i, denote the iterative singular vector as $\Psi^{(i)}$ with the s-th column as $\Psi_s^{(i)}$

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

      Pair 1: $Q_{\tau_k}(Y_j|X,Z) = \alpha(\tau_k)$,

      Vs $Q_{\tau_k}(Y_j) = \alpha(\tau_k) + \sum_{s=1}^{\hat{m}} \Psi_s^{(i)}\theta_s(\tau_k)$.

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

      Vs $Q_{\tau_k}(Y_j) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\,\beta_l(\tau_k) + Z_i\gamma(\tau_k) + \sum_{s=1}^{\hat{m}} \Psi_s^{(i)}\theta_s(\tau_k)$

   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. Conduct a reduced matrix by selecting columns of R' that satisfy the above criteria, denote it as $R^{(i)\prime}$

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

   d. Select from the singular vectors of $R^{(i)\prime}$the set of columns that shares the strongest absolute correlation with the stage-1-singular-vectors Φ. Denote these as the iSV at iteration i as $\Psi^{(i)}$

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 PC 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 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 QuanT Algorithm

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 model $Q_{\tau_k}(r_j^* | X, Z) = \alpha(\tau_k) + \sum_{l=1}^{L} X_{il}\, \beta_l(\tau_k) + Z_i \gamma(\tau_k)$, and obtain the binarized and centered residuals, denote by R^{0*}
7. Conduct singular value decomposition of R^{0*} and calculate statistic $T_k^0 = \frac{d_{0k}^2}{\sum d_{0l}^2}$
8. Repeat steps 5-7 a total of B times to obtain null statistics T_K^{0b} for $b = 1, \ldots B$. We set B as 20.
9. 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 $p_k = max\{p_{k-1}, p_k\}$.

10. Select the singular vectors with $p_k \leq 0.1$.
11. Carry out correlation calibration (steps 6-7 in QuanT algorithm)

Computation of QuanT

We have streamlined the process of quantile regression and matrix decomposition within QuanT by employing strategies for enhanced efficiency. The algorithm executes quantile regression across various taxon and quantile levels independently, which lends itself well to parallel processing. We leverage the doParallel package in R to expedite this computation. In the second stage of QuanT, where SVD is performed, we focus on a few leading singular

vectors because it is generally improbable for a singular vector that accounts for minimal variance to represent a critical direction established in the first stage. Consequently, it is sufficient to compute only the leading singular vectors of the binarized residual matrix during the correlation calibration phase. For this reason, QuanT only calculates the top principal components using the RSpectra R package. Supp. Tab. 1 and Supp. Tab. 2 present the computational time of QuanT for datasets with different number of samples and taxa.

Novelty of QuanT

Like existing approaches based on SVA, QuanT utilizes a two-stage procedure in which the matrix decomposition in the first stage detects the underlying hidden directions and a second stage refines this direction. However, QuanT offers many novel modifications/improvements to the existing approaches to make it suitable for the microbiome data that is typically sparser and has more outliers. First, by using quantile regressions instead of least squares, QuanT can gather information from the whole distribution of the data rather than just the means, which is critical for microbiome data. Secondly, SVA searches for significant directions by identifying singular vectors that represent more variations than expected by chance through a permutation. QuanT, on the other hand, relies on the 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 that explain only a modest proportion of variation. In the second stage of correlation calibration, QuanT uses a data-adaptive approach to select the p-value cutoffs (c_1 and c_2) for the two pairs of comparisons, which is in contrast to SVA where a fixed cutoff is used. Our investigation shows that there are no universally best cutoffs, and our adaptive approach allows for the optimal calibration (within our choice of cutoffs) and thus allows for better FDR control in downstream analyses.

Data sources and processing

CARDIA: a large-scale epidemiology study

The Coronary Artery Risk Development in Young Adults (CARDIA) Study[15], initiated in 1985-1986, recruited young adults aged between 18 to 30 with the goal of uncovering determinants of cardiovascular disease (CVD) throughout adult life. Subject enrollment was balanced according to race, gender, education (more than high school or not), and age.

Extensive risk factors related to CVD have been collected, including health behaviors, medical history (including medication use), and clinical risk factors, such as blood pressure. Stool samples were also collected for some visits. For our purpose, we used the microbiome data generated from the stool samples collected during the Year 30 follow-up (2015-2016). The samples were processed through four DNA extraction/library preparation batches, and subsequently sequenced through seven runs using the Illumina MiSeq 2x300 platform. Each DNA extraction batch was allocated to two sequencing runs, except for the final batch, which was sequenced by itself during the seventh run. Consequently, the data were collected across seven distinct batches. The data were agglomerated to the genus level, and taxa that did not yield any reads in all samples used in our analyses were omitted from the dataset. After preprocessing, the data contains 633 samples and 145 taxa.

We applied QuanT, SVA, and RUV to the CARDIA dataset to find the hidden heterogeneity directions, pretending that the true batch indicators are unknown. For SVA, we used the svaseq function, which applied a moderated log transformation to count data as described in Leek[5], and used the iterative reweighted algorithm[2]. For RUV, we used the RUVg approach in Risso *et al.*[16] with an estimated group of null taxa as controls, rather than RUVr and RUVs as these approaches can only identify hidden directions that are uncorrelated with the variables of interest. We considered the high blood pressure (HBP) as the primary variable of interest, race (Black = 1, White = 0) and gender (Female = 1, Male = 0) as the covariates to adjust for. Only complete cases were retained. To investigate and compare the hidden directions identified by QuanT, SVA, and RUV, we incorporated the directions into downstream analysis by using ConQuR to remove the effects of the detected directions and generate a "batch-free" dataset.

HIVRC: the integration of multiple individual studies

We further assessed the performance of QuanT using data from the HIV re-analysis consortium (HIVRC)[17]. Briefly, raw 16s rRNA gene sequencing data were solicited from 22 studies that investigate the relationship between gut microbiome and HIV infection. Data from 17 of these studies were subsequently processed through a common pipeline with taxa assigned using Resphera Insight[18]. Details of data pre-processing and taxonomic assignment can be found in Tuddenham, et al. 2020[17].

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 in the CARDIA study. Measured by PERMANOVA R^2, the study ID explains 30.39% of the data variability, while the batch effects in CARDIA only contribute 5.66%. The library size also exhibits significant variation among studies, e.g., samples have 185-1000 reads in Study 6, whereas the library size was rarefied to 20000 reads in Studies 0 and 8.

We select two studies with similar library sizes (5000-5500) but very different study design (an observational study in the US[19] and a clinical trial in Europe[20]), which includes 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 (Positive = 1, Negative = 0) was regarded as the variable of interest, while age, gender (Female = 1, Male = 0) and BMI were considered as covariates. We retained complete cases only and conducted the same set of analyses as on the CARDIA data.

CRC: gut microbiome data in colon cancers from the CuratedMetagenomicData package

The CuratedMetagenomicData[21] data package offers standardized, curated, and analysis-ready metagenomic datasets derived from human microbiome studies, and thus facilitates comprehensive meta-analyses and comparisons across studies by ensuring consistent data processing and annotation. We selected two studies[22, 23] on colorectal cancer (CRC) from the curatedMetagenomicData package, which include 154 and 80 stool samples respectively. After filtering, there are 166 genera in the data. The two studies encompass samples from three distinct health conditions: colorectal cancer (CRC), adenoma, and control, with CRC further classified into two subtypes: adenocarcinoma and carcinoma and adenoma classified into adenoma and advanced adenoma.

We applied QuanT, SVA, and RUV to this CRC dataset to find the hidden heterogeneity directions, pretending that the true batch indicators are unknown. We considered the five phenotype categories (adenocarcinoma, carcinoma, adenoma, advanced adenoma, control) as the primary variable of interest, age (Adult vs. Senior) and gender (Female vs. Male) as the covariates to adjust for.

Simulation studies

We conducted extensive simulations to evaluate the performance of QuanT. We simulated data that mimicked two datasets from the integrative Human Microbiome Project[24]: 1) a gut microbiome dataset from the Inflammatory Bowel Disease Multi-omics Database (IBDMDB) project[25]; 2) a vaginal microbiome dataset from Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI) project. These two datasets represent microbiome data from two body sites that are of great research interest. They also have contrasting attributes in sparsity level and over-dispersion, and thus make them ideal for an evaluation of our proposed method. We used two methods to simulate data that mimicked the template data; the parametric Dirichlet-Multinomial distributions, used in many previous applications[26], and the non-parametric version of the newly proposed MIDASim[27] method.

Covariates and batches

For each sample i, we simulated a clinical variable of interest (denoted by Z_i) and a hidden batch variable (denoted as B_i) through a bivariate Bernoulli distribution, with both variables having a marginal mean of 0.5. A correlation of 0.3 was set between them so that the batch variable can be a potential confounder between microbiome and Z. Here, we use the "batch" variable to represent all the hidden heterogeneities. In the simulation, we set the batch variable to impact most of the taxa, but Z only impacts a small proportion of the taxa. For instance, using the IBDMDB as a template, we simulated the batch effect on 250 out of 301 taxa, with the variable of interest affecting 50 taxa. Similarly, in the MOMS-PI dataset, the batch variable impacted 200 of the 233 taxa, and the variable of interest affected 40 taxa. It is important to note that the batch variables were treated as unknown when using SVA, RUV and QuanT, and are treated as known for the gold standard analyses.

We defined two groups: "Batch Taxa" for those impacted by the batch effect and "Clinical Taxa" for those influenced by the variable of interest. The Batch Taxa were chosen at random, and the Clinical Taxa were a randomly selected subset within them. Our goal is to identify the batch variable as if it was unknown using the relative abundance data, and then test for the association between Z and taxa abundances adjusting for the hidden batch variable.

Simulation based on the Dirichlet-Multinomial distribution

We followed the simulation setup in a previous paper[26] to simulate taxa abundances using the Dirichlet-Multinomial (DM) distribution. We first fitted the DM model to the template datasets (IBDMDB and MOMS-PI) and obtained the estimated mean vectors π_0. We then formed two additional vectors π_B and π_Z to allow for a batch and clinical effect. Both vectors were initially set to be the same as π_0. We then randomly permuted the frequencies within π_B and $\pi_Z$that correspond to Batch Taxa and Clinical Taxa. We then defined a sample-specific frequency $\pi_i = (1 - w_B B_i - w_z Z_i)\pi_0 + w_B B_i \pi_B + w_z Z_i \pi_z$, where (w_B, w_z) determine the underlying batch effect and clinical effect. We considered three pairs of (w_B, w_z) . $(w_B, w_z) \in \{(0,0.5), (0.5,0.5), (0.8,0.2)\}$. When $w_B = 0$, there is no underlying hidden heterogeneity. We generated the count data for each sample using the DM distribution with π_i and an overdispersion parameter of 0.02.

Simulation using MIDASim

The Microbiome Data Simulator (MIDASim) 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 the log-linear models. Here, we utilized the non-parametric version because it captures the underlying data structure better. Via MIDASim, for each template data set, we can first generate 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 the same way as in the DM simulation using the same weights. Count data for each sample were generated by inputting the sample-specific vectors into MIDASim using the non-parametric approach.

Similar to our analysis of CARDIA and HIVRC data, we applied QuanT, SVA, and RUV to the simulated datasets, estimating the hidden heterogeneity controlling for Z. We applied simple linear regression models without adjustment, adjusting for true batch indicators, adjusting for SVs (from SVA), QSVs (from QuanT), and RUV factors. We also applied ConQuR to remove the true batch, SVs and QSV effects and generate a "batch-free" dataset.

Please refer to Supplement Section Simulation Studies and Supp. Fig. 1-9 for detailed description of the analysis and results.

References

1. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**, e161 (2007).
2. Leek, J.T. & Storey, J.D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105**, 18718-18723 (2008).
3. 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).
4. 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).
5. Leek, J.T. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research* **42**, e161-e161 (2014).
6. Koenker, R. & Hallock, K.F. Quantile regression. *Journal of economic perspectives* **15**, 143-156 (2001).
7. Barber, R.F. & Candès, E.J. Controlling the false discovery rate via knockoffs. (2015).
8. Goldstein, S., Lebowitz, J.L., Tumulka, R. & Zanghi, N. Any orthonormal basis in high dimension is uniformly distributed over the sphere. (2017).
9. 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).
10. 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).
11. Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**, 479-498 (2002).
12. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics* **31**, 2013-2035 (2003).
13. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 (2003).
14. 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).
15. Friedman, G.D. et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology* **41**, 1105-1116 (1988).
16. 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).

17. 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).
18. 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).
19. Lozupone, C.A. et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* **14**, 329-339 (2013).
20. 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).
21. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023-1024 (2017).
22. Feng, Q. et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature communications* **6**, 6528 (2015).
23. 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).
24. 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).
25. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662 (2019).
26. Hu, Y.-J. & Satten, G.A. Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics* **36**, 4106-4115 (2020).
27. He, M., Zhao, N. & Satten, G.A. MIDASim: a fast and simple simulator for realistic microbiome data. *Microbiome* **12**, 135 (2024).