Microbiome Data Integration via Shared Dictionary Learning

Bo Yuan* and Shulei Wang*
University of Illinois at Urbana-Champaign

(October 4, 2024)

Abstract

Data integration is a powerful tool for facilitating a comprehensive and generalizable understanding of microbial communities and their association with outcomes of interest. However, integrating data sets from different studies remains a challenging problem because of severe batch effects, unobserved confounding variables, and high heterogeneity across data sets. We propose a new data integration method called MetaDICT, which initially estimates the batch effects by weighting methods in causal inference literature and then refines the estimation via a novel shared dictionary learning. Compared with existing methods, MetaDICT can better avoid the overcorrection of batch effects and preserve biological variation when there exist unobserved confounding variables or data sets are highly heterogeneous across studies. Furthermore, MetaDICT can generate comparable embedding at both taxa and sample levels that can be used to unravel the hidden structure of the integrated data and improve the integrative analysis. Applications to synthetic and real microbiome data sets demonstrate the robustness and effectiveness of MetaDICT in integrative analysis. Using MetaDICT, we characterize microbial interaction, identify generalizable microbial signatures, and enhance the accuracy of disease prediction in an integrative analysis of colorectal cancer metagenomics studies.

1 Introduction

Recent advances in metagenomic sequencing technologies make it possible to profile the microbiome communities from hundreds to thousands of samples in scientific studies (Turnbaugh et al., 2009; Yatsunenko et al., 2012; Franzosa et al., 2019). These microbiome studies provide glimpses into the complex microbial ecosystem and improve our understanding of the interactions between the microbes and their host. Although each study has already

*Address for Correspondence: Department of Statistics, University of Illinois at Urbana-Champaign, 605 E. Springfield Ave., Champaign, IL 61820 (Email: boyuan5@illinois.edu, shuleiw@illinois.edu).

yielded interesting results, the findings from different studies are not always consistent and the power could be limited due to the small sample size of each study (Langdon et al., 2016; Duvallet et al., 2017). One promising strategy to obtain generalizable discoveries is the integrative analysis of the data sets from multiple studies (Wirbel et al., 2019; Ma et al., 2022). However, integrative microbiome data analysis presents unique quantitative challenges as the data from different studies are collected across times, locations, or sequencing protocols and thus suffer severe batch effects and are highly heterogeneous. When handled inappropriately, the batch effects and high heterogeneity could lead to increased false discoveries and reduced accuracy in the downstream integrative analysis (Ling et al., 2022).

In order to correct batch effect and facilitate a valid integrative analysis, one of the most popular strategies in microbiome studies is to apply the regression models to correct the batch effects, where the sequencing count of each taxon is the outcome and covariates include batch and each sample's observed covariates (Ritchie et al., 2015; Johnson et al., 2007; Gibbons et al., 2018; Zhang et al., 2020; Ma et al., 2022; Ling et al., 2022; Ye et al., 2023). The primary assumption behind such a strategy is that the conditional distribution/expectation of sequencing count remains the same after successfully adjusting the effects of batch and observed covariates. The covariate adjustment methods can correct the batch effects efficiently when all confounding covariates are observed and adjusted appropriately. However, this assumption could be invalid, and overcorrection happens when there are some important unmeasured confounding covariates, such as lifestyle. Besides the covariate adjustment strategy, another popular strategy is to utilize the intrinsic structure of data to correct batch effects (Haghverdi et al., 2018; Butler et al., 2018; Hie et al., 2019; Korsunsky et al., 2019; Welch et al., 2019; Amodio et al., 2019; Barkas et al., 2019). The main advantage of such a strategy is that exploring intrinsic structure does not rely on extrinsic covariates and is thus robust to unmeasured confounding covariates. Most methods in this strategy are designed for single-cell RNA sequencing data since the covariate adjustment is not applicable in single-cell RNA sequencing data due to no access to cell-level covariates. However, this strategy does not utilize the valuable information in observed covariates when available, and the assumptions adopted in existing methods are not general enough to work for the microbiome data. For example, microbiome data may be separated poorly into several groups, while the commonly used anchor for the single-cell RNA-seq data integration is a multi-cluster structure due to different cell types. The challenges above raise questions about whether it is possible to develop a new data integration method for microbiome data that combines the advantages of these two popular strategies. This paper shows that this is feasible.

This paper introduces MetaDICT, a new two-stage data integration method for microbiome data. With a similar spirit to existing covariate adjustment methods, the first stage of MetaDICT obtains an initial estimation of batch effects via adjusting commonly observed covariates. However, MetaDICT adopts the weighting method in casual inference literature to adjust covariates instead of the commonly used regression-based methods because several studies confirm that the batch effects affect the sequencing counts multiplicatively rather than additively (Harismendy et al., 2009; McLaren et al., 2019). In the second stage of MetaDICT, the estimation of batch effects is further refined via a novel shared dictionary learning. Through shared dictionary learning, MetaDICT explores the intrinsic structures

that are sufficiently flexible to capture the characteristics of the microbiome data and can thus disentangle the batch effect from the heterogeneous data sets robustly. Thanks to the shared dictionary learning, MetaDICT can better address the overcorrection of batch effects and preserve biological variation than existing methods in diverse settings, including ones where there are unmeasured confounding covariates or data sets are highly heterogeneous across studies. Beyond batch effect correction, shared dictionary learning in MetaDICT also generates the embedding at both taxa and sample levels that can be used to unravel the hidden structure of the integrated data and improve the integrative analysis. Comprehensive numerical experiments presented in this paper demonstrate the efficacy of MetaDICT in correcting batch effect, reducing false discoveries, and enhancing the power of integrative analysis. In particular, we apply MetaDICT to an integrative analysis of five colorectal cancer metagenomics studies where each study is conducted in a different country. In the integrative analysis, MetaDICT can successfully separate the batch effect and effect of country, reveal the microbial functional similarity, detect population structure, identify previously documented and novel microbial signatures of colorectal cancer, and improve the accuracy and generalizability of disease diagnosis.

2 Results

2.1 Overview of MetaDICT

This section presents an overview of MetaDICT, while the Methods section comprehensively explains the proposed method. As summarized in Figure 1, the newly proposed MetaDICT consists of two stages: the first stage provides an initial estimation of batch effect via covariate balancing and the second stage refines the estimation by shared dictionary learning. MetaDICT defines the batch effect as heterogeneous capturing efficiency in sequencing measurement, which is highly influenced by the variations in technical factors and external conditions of sequencing procedures (Lander, 1999; Morgan et al., 2010). Because measurement efficiency usually affects observed abundances in a multiplicative way (Harismendy et al., 2009; McLaren et al., 2019), the first stage of MetaDICT estimates the measurement efficiency by the weighting method in causal inference (Imbens and Rubin, 2015), one of the most popular covariate balancing methods. The initial estimation from the weighting method is accurate when all confounding variables are observed in different studies. However, this strategy might result in overcorrection when we only observe a few common variables across studies or have difficulty in measuring some important confounding variables in practice, like lifestyle. MetaDICT further refines the estimation in the presence of unobserved confounding variables to increase the robustness.

Unlike covariate adjustment, the second stage of MetaDICT aims to improve the estimation of measurement efficiency via exploring two types of intrinsic structures in microbiome data: shared dictionary and the measurement efficiency's smoothness. Despite the variation in sequencing procedures, the microbes interact and coexist as an ecosystem similarly in different studies (Woyke et al., 2006; Chaffron et al., 2010). Motivated by this observation,

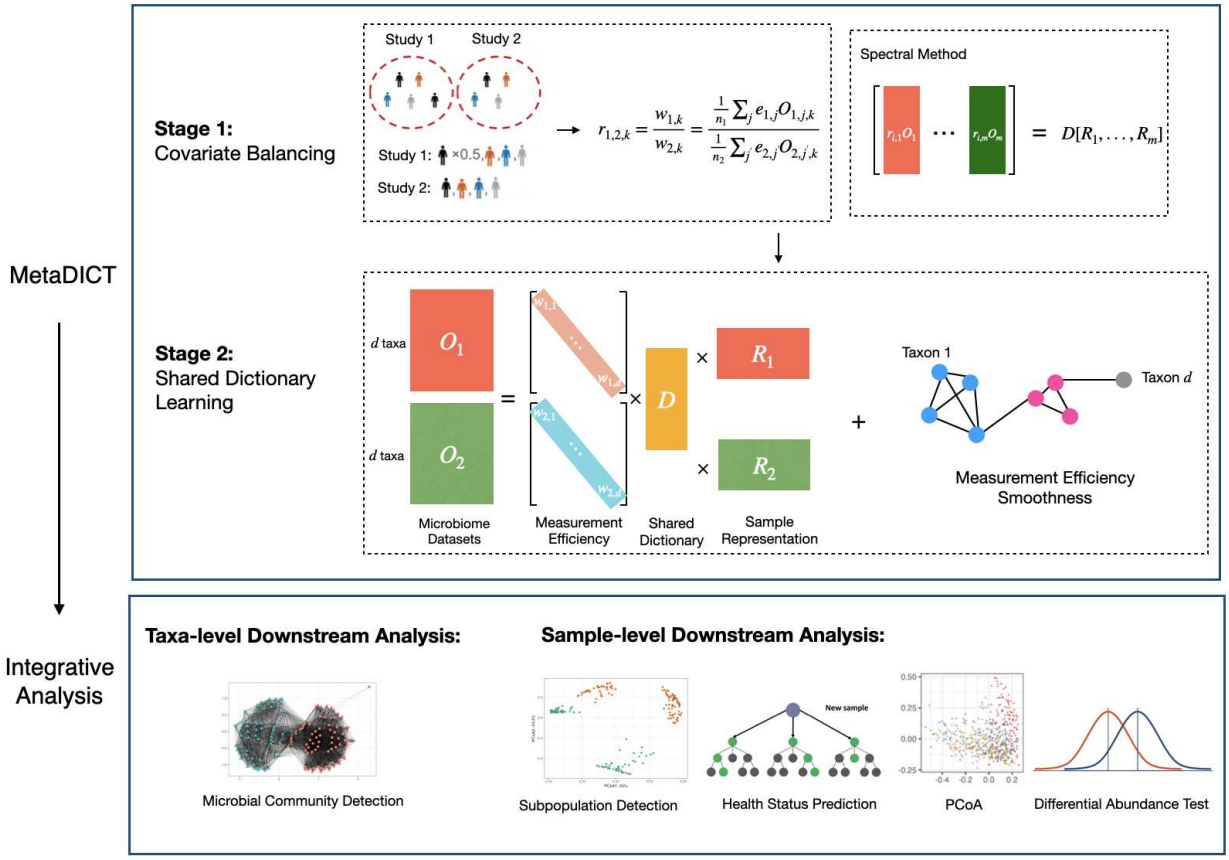

Figure 1: **A summary of MetaDICT for integrative analysis.** Stage 1: the weighting method in the causal inference literature adjusts the commonly observed covariates to estimate the batch effect initially. Stage 2: the batch effect estimation is refined by exploring the intrinsic structures of microbiome data: the shared dictionary of microbial profile and measurement efficiency's smoothness. In addition to batch effect correction, MetaDICT also generates embedding of taxa and samples for efficient integrative analysis.

MetaDICT introduces a shared dictionary of microbial absolute abundance to capture such a universal structure across studies, making it possible to separate batch effects from biological variation in absolute abundance. Besides the shared dictionary, another important observation is that microbes with close taxonomic sequences tend to have similar capturing efficiency in each study (Krsek and Wellington, 1999; Polz and Cavanaugh, 1998; Carrigg et al., 2007; Benjamini and Speed, 2012). This observation indicates that measurement efficiency is smooth with respect to the similarity among taxa, allowing borrowing strength from similar taxa. The second stage of MetaDICT solves a nonconvex optimization problem to explore the above two intrinsic structures of microbiome data, which is initialized by a spectral method and the estimation in the first stage. By utilizing intrinsic structures, MetaDICT can adjust the batch effect robustly, avoid overcorrection efficiently, and maintain the biological variation in the downstream integrative analysis. In addition to batch effect correction, the estimated shared dictionary in MetaDICT can naturally offer embeddings of taxa and samples, revealing the microbial communities and improving the performance of

downstream analysis. In the following sections, we demonstrate the robustness and effectiveness of MetaDICT in correcting batch effect and improving the downstream integrative analysis.

2.2 MetaDICT Corrects Batch Effect Robustly

This section designs a series of numerical experiments to evaluate the MetaDICT's performance of correcting batch effect. As summarized in Figure 1, one unique feature of the MetaDICT is to explore the intrinsic structure of microbiome data via shared dictionary learning. The first set of numerical experiments aims to assess whether the shared dictionary learning in the MetaDICT can improve the initial estimation in the first stage. To mimic the real data, we generate the synthetic data from a microbiome data set collected by He et al. (2018). We consider two criteria to assess the performance: the mean absolute error of estimated measurement efficiency at each taxon (Figure 2(a)) and the Pearson correlation coefficient between the estimated and true measurement efficiency (Figure 2(b)). These comparisons suggest that utilizing the intrinsic structure improves the accuracy of estimated measurement efficiency at both taxa and data set levels. Exploiting the smoothness of measurement efficiency is one of the major reasons why shared dictionary learning can increase accuracy. As illustrated in Figure S1(a), the penalty for measurement efficiency's smoothness enables borrowing strength from similar taxa, thus improving the performance (Figure S1(b)). Furthermore, we also validate the robustness of MetaDICT in a wide range of settings, including experiments when the number of data sets, sample sizes per data set, the balance of data set sizes, and the measurement efficiency smoothness level are different (Figure S2). The results presented in Figure 2, S1, and S2 indicate that MetaDICT can recover measurement efficiencies accurately and robustly.

The next set of numerical experiments investigates whether the accurate measurement efficiency estimation in MetaDICT can help correct batch effects and maintain the biological variation in the presence of unobserved confounding variables. In this set of experiments, we compare MetaDICT with three state-of-the-art microbiome data integration methods: ConQuR (Ling et al., 2022), ComBat-Seq (Zhang et al., 2020), and MMUPHin (Ma et al., 2022). The synthetic data include data sets from two studies, and the microbial absolute abundance relies on a binary biological variable, such as the indicator of health status. Due to the batch effect, PCoA plots of the unprocessed data show a distribution change across studies and less separated clusters between groups defined by the biological variable (Figure 2(c) and (d)). We consider two possible scenarios to correct the batch effect: the confounding biological variable is 1) not observed or 2) observed. When the biological variable is observed, only ConQuR and MetaDICT can adjust the effect of the biological variable well, thus successfully correcting the batch effect and maintaining a decent amount of biological variation (Figure 2(c)). On the other hand, if the biological variable is not observed, MetaDICT can still separate clusters defined by the biological variable, while ConQuR wrongly reduces the biological variation (Figure 2(d)). Besides the binary biological variable, we consider a similar experiment when the biological variable is continuous (Figure S3). The results suggest that exploring the intrinsic structure of microbiome data in MetaDICT can significantly

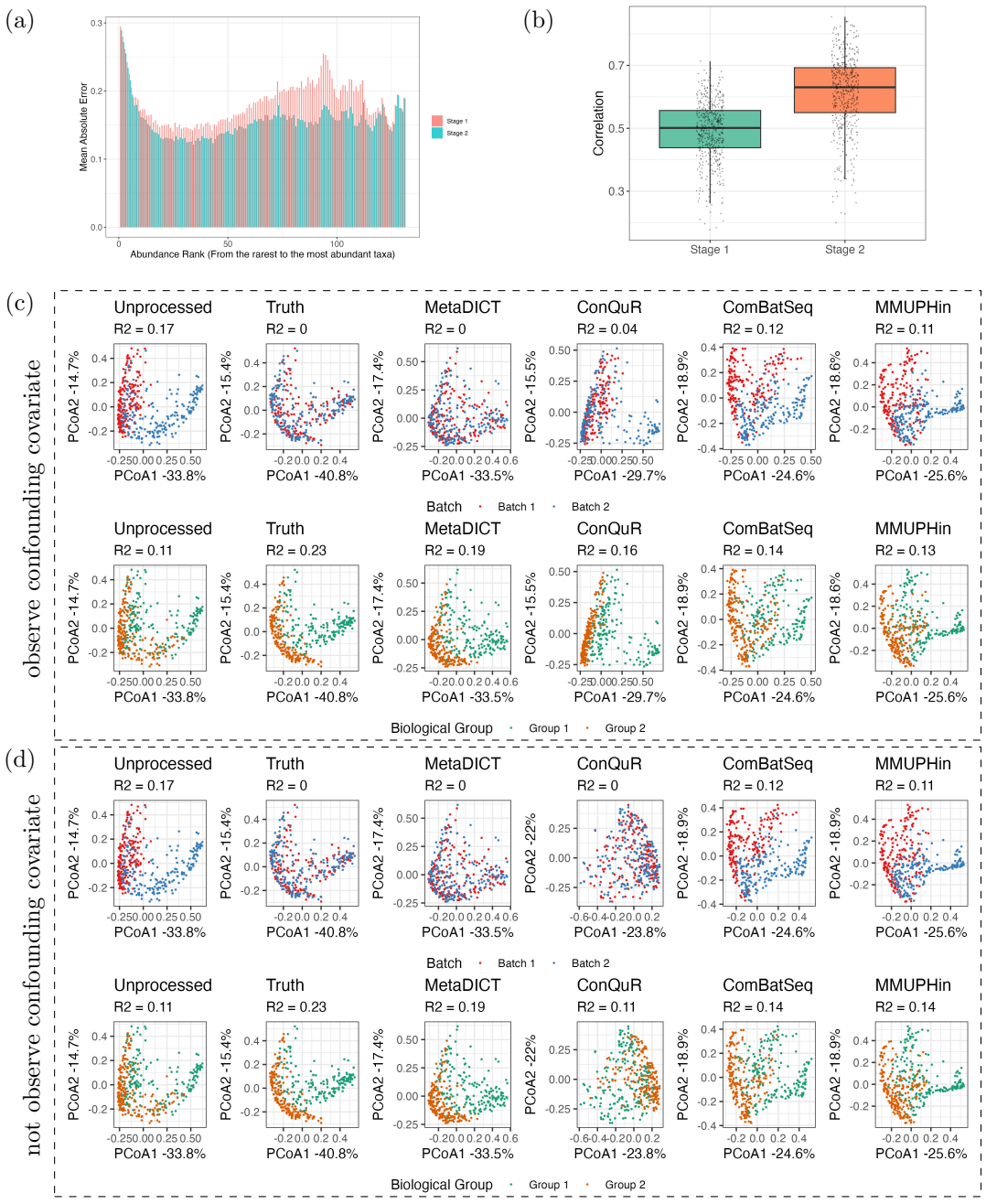

Figure 2: **MetaDICT estimates measurement efficiency accurately and corrects batch effects robustly.** Figures (a) and (b) compare the estimation accuracy of measurement efficiency between stages 1 and 2, showing that exploring intrinsic structure can significantly improve estimation accuracy. In (a), the mean absolute error of each taxon is compared, with taxa ordered from least to most abundant. In (b), the y-axis represents the Pearson correlation coefficients between the estimated measurement efficiency and the ground truth. Figures (c) and (d) compare the performance of four data integration methods, demonstrating that MetaDICT corrects batch effects robustly and maintains biological variation effectively. Figure (d) shows the PCoA plots and R^2 in PERMANOVA when the biological variable is not observed in advance, while Figure (c) presents the results when the biological variable is used as input for all methods.

increase the robustness of batch effect correction.

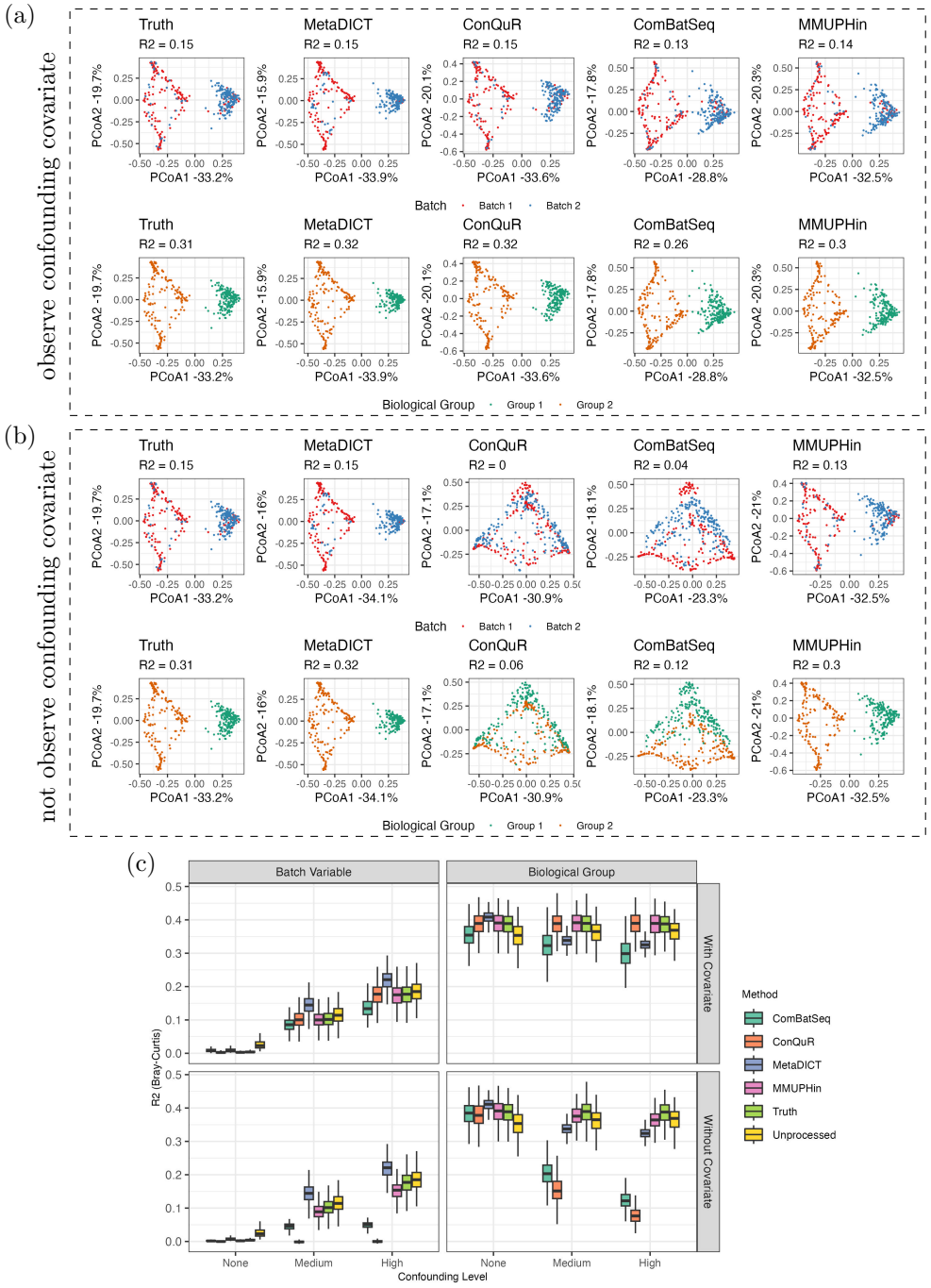

Figure 3: **MetaDICT avoids overcorrection of batch effects in the presence of a distribution shift in absolute abundance across data sets.** All experiments include a distribution shift in absolute abundance but no batch effect. Figures (a) and (b) compare the performance of four data integration methods via PCoA plots and R^2 in PERMANOVA. Figure (c) shows the box plot of R^2 in PERMANOVA when the confounding level of biological variables varies. These results show that MetaDICT can perform robustly and avoid overcorrection when there is a distribution shift in absolute abundance.

This section's last set of numerical experiments studies if MetaDICT can avoid overcorrection of batch effects when the data sets are heterogeneous across studies. When there is a distribution shift in the observed sequencing counts across studies, it is challenging to distinguish the variation results from the batch effects or heterogeneity in the absolute abundance. We first consider an ideal setting with no batch effects but distribution shifts in the absolute abundance due to a confounding biological variable. Similar to the previous experiments, we apply the four different data integration methods and consider the same two possible scenarios when the confounding biological variable is not observed and observed. As expected, the unevenly distributed biological variable naturally leads to a distribution shift in absolute abundance across studies (Figure 3(a) and (b)). Applying data integration methods preserves such heterogeneity if the biological variable is observed (Figure 3(a)). However, ConQuR and ComBat-Seq overcorrect the batch effects when the biological variable is not observed, while MetaDICT and MMUPHin are robust in such a setting (Figure 3(b)). This observation indicates that some existing methods might wrongly consider the effect of confounding variables as batch effects and remove it when the corresponding confounding variable is not observed. Similar phenomena of overcorrection are also observed when we vary the confounding level of biological variables (Figure 3(c)). Furthermore, we also design two numerical experiments when both batch effect and heterogeneity in absolute abundance are present (Figure S4), and the choice of study completely confounds the observed covariate (Figure S5). All these numerical experiments show that MetaDICT can better address the issue of overcorrection and preserve the biological variation than existing data integration methods when there is a distribution shift in absolute abundance across studies.

2.3 MetaDICT Reveals Hidden Structure via Embedding

Another unique feature of MetaDICT is the embedding derived from the estimated shared dictionary, and this section presents a series of numerical experiments to demonstrate its merit. We first study if the taxa embedding in MetaDICT can recover the universal structure of microbial communities. Similar to the previous section, we generate synthetic data sets with five microbial communities by modifying the real microbiome data set collected by He et al. (2018) (Figure S6(a)). In this experiment, we consider six community detection methods: clustering on the taxa embedding from MetaDICT, clustering on a single data set, clustering on the integrated data sets corrected by ConQuR, ComBat-Seq, and MMUPHin, and clustering on the combined unprocessed data sets. We use the adjusted Rand index to evaluate the performance of community detection. As illustrated in Figure 4(a), the taxa embedding from MetaDICT can separate different microbial communities, while other methods tend to underestimate the number of communities because of overlapping across communities. We further compare the performance of these clustering methods on a wide range of experiments. Specifically, we consider various settings: 1) the confounding biological variable is observed or not (Figure 4(c)); 2) the number of data sets varies (Figure S7(a)); 3) the signal of community is different (Figure S7(b)); 4) the sample size of each data set is different (Figure S7(c)). These results suggest that the batch effect can greatly impact the microbial communities' structure, and community detection performance relies highly on

the quality of data integration. On the one hand, microbial community detection could be inaccurate if we cannot adjust the batch effect appropriately, like overcorrecting the batch effect when the confounding covariate is not observed. On the other hand, data integration can achieve a better performance than other methods when MetaDICT corrects batch effects. These comparisons indicate the embedding of MetaDICT can effectively integrate the strength of multiple data sets to uncover the universal structure of microbial communities.

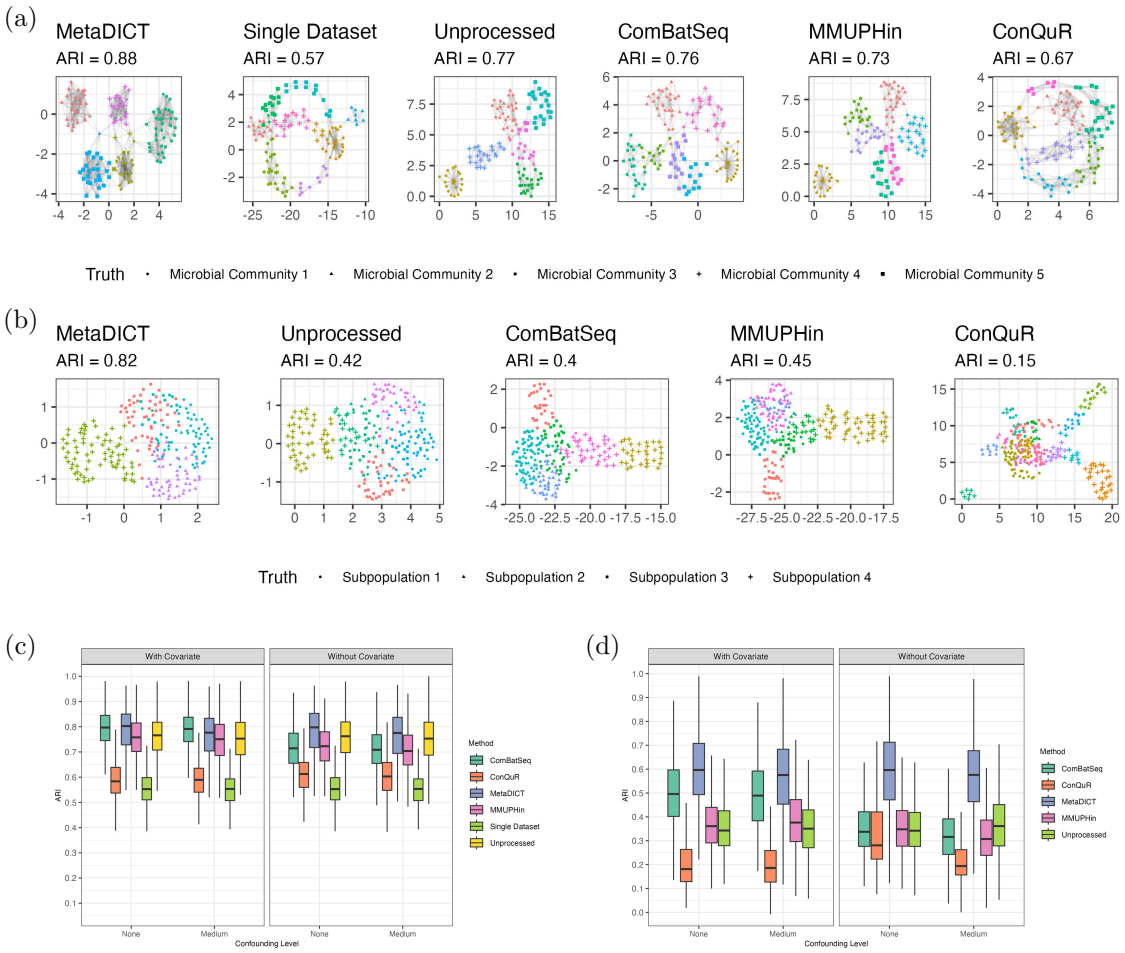

Figure 4: **MetaDICT reveals the hidden structure of microbiome data via embedding.** Figures (a) and (b) show examples of clustering results at taxa and sample levels, where colors represent the detected communities and shapes represent the true communities. Figures (c) and (d) compare the community detection accuracy of different methods and present adjusted Rand index in repeated experiments. All experiments suggest MetaDICT is a reliable approach to detecting communities at taxa and sample levels.

Besides taxa embedding, MetaDICT also generates embedding at the sample level, so the next set of experiments is designed to evaluate the efficiency of sample embedding. We consider the same synthetic data sets similar to the previous set of experiments and the adjusted Rand index as the performance measure. Since we are usually interested in sample subpopulations across studies in integrative analysis, we no longer apply clustering algorithms to each data set. To compare different methods, we vary the access of confounding variables (Figure 4(d)), the number of data sets (Figure S8(a)), signal strength (Figure S8(b)), and sample

size per data set (Figure S8(c)). Similar to the previous set of experiments, the comparisons show that the batch effects can largely perturb the community detection results at the sample level, and thus, suitable data integration is key to understanding the subpopulation of samples. In particular, the sample embedding in MetaDICT offers a concise representation of each sample and can effectively separate sample clusters (Figure 4(b)). These results again confirm the embedding efficiency in MetaDICT as it performs better than other methods. In addition, the clustering at both taxa and sample levels can naturally lead to a biclustering method for integrative analysis. The results shown in Figure S6 suggest that embedding in MetaDICT is a reliable approach to unraveling the hidden communities at both taxa and sample levels.

2.4 MetaDICT Achieves Reliable Integrative Analysis.

This section includes several numerical experiments to study how the data integration impacts the downstream integrative analysis, such as differential abundance analysis and outcome prediction. We first explore the performance of commonly used differential abundance tests on the integrated data set. Specifically, we consider the same four data integration methods as previous sections and three commonly used differential abundance tests: the standard t-test, RDB (Wang, 2023b), and LinDA (Zhou et al., 2022). In the integrative analysis, the outcome of interest in differential abundance tests is also included as an observed covariate in different data integration methods. To evaluate the performance, we consider four experiment settings: 1) there are some differential abundant taxa and the outcome is independent of the batch (Figure S9(a)); 2) there are some differential abundant taxa and the outcome is confounded with the batch (Figure 5(b)); 3) there is no differential abundant taxon when there is no distribution shift in outcome across studies (Figure 5(c) and S9(b)); 4) there is no differential abundant taxon when there is a distribution shift in outcome across studies (Figure 5(c) and S9(b)). When the outcome is independent of the batch, most data integration methods can successfully correct the batch effects, and thus, the differential abundance tests on the integrated data set can control the false discovery well and maintain a decent power. However, if the outcome of interest is a confounding variable that relies on the batch, batch effect correction becomes challenging in existing data integration methods, resulting in an inflated false discover rate and reduced power in all three differential abundance tests. Comparing the false discovery frequency of each taxon with their batch effect disturbance level suggests that the false discoveries in differential abundance analysis mainly result from remaining uncorrected batch effects (Figure 5(a)). Due to exploring the intrinsic structure of microbiome data, MetaDICT can better correct batch effect and thus lead to more reliable differential abundance analysis than existing methods when the outcome of interest is confounded with the batch. Therefore, MetaDICT is a good choice of data integration method when integrative differential abundance analysis is interesting.

The next downstream task considered in this section is outcome prediction. We consider two of the most popular classifiers in the integrative analysis: k-nearest neighbor classifier (k-NN) and random forest. When evaluating the performance, we aim to answer the follow-

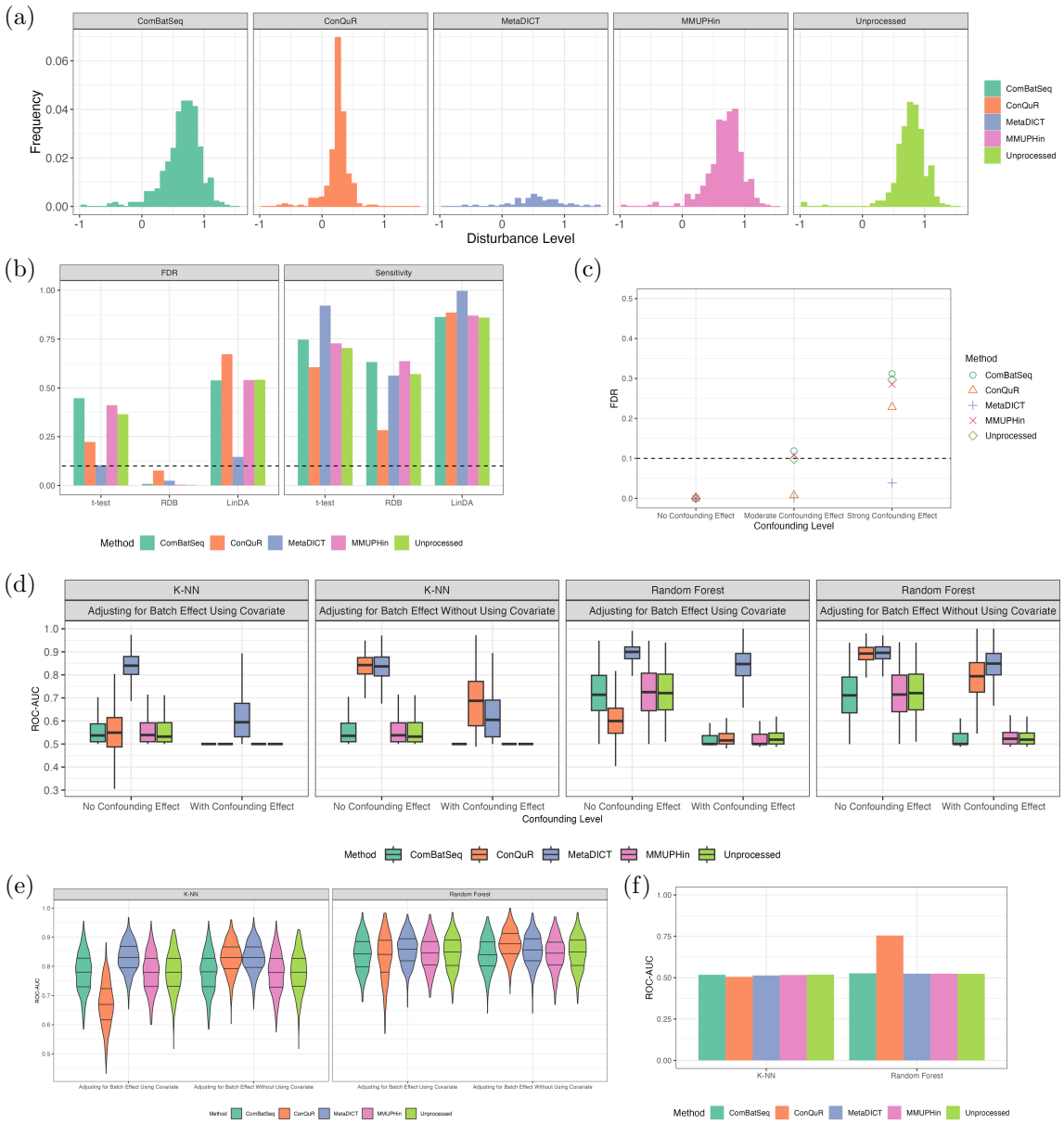

Figure 5: **MetaDICT improves downstream integrative analyses.** Figure (a) shows the frequency of false discovery at the taxa level against their disturbance level due to batch effects. Figure (b) presents the performance of differential abundance tests when data are integrated via different methods. Figure (c) assesses false discovery rate inflation when the covariate of interest is independent of microbial composition but confounded with batch. Figure (d) compares k-NN and random forest accuracy when the training and testing data sets have different measurement efficiency. Figure (e) compares the accuracy of classifiers when the training and testing data sets come from the same integrated data set. Figure (f) shows the prediction accuracy when the outcome of interest is independent of microbial compositions but is used in the data integration. These experiments suggest that MetaDICT enhances the accuracy of downstream analysis, including differential abundance tests and outcome prediction.

ing two questions: 1) how does the downstream classifier perform when the batch effect is not corrected appropriately? 2) is using the same data set for data integration and classifier

training safe? We consider two common prediction settings in practice to answer the first question. In the first setting, we design an experiment similar to the settings in transfer learning, where the training and testing data sets come from different studies and thus have different measurement efficiencies (Figure 5(d)). The results suggest that the batch effect correction between training and testing data sets is essential for building a reliable and generalizable classifier. In particular, MetaDICT and ConQuR lead to accurate classifiers in k-NN and random forest while other methods perform similarly to unprocessed data. Besides the transfer learning setting, we also consider the second setting, where the training and testing data sets are randomly drawn from the same integrated data set (Figure 5(e)). The results show that the classifier trained by the data set integrated by MetaDICT can achieve better performance in k-NN as they can integrate the data set more effectively. To test the potential double-dipping issues, we design an experiment with an outcome fully independent of microbial sequencing count and included it as a covariate in the data integration (Figure 5(f)). The results in Figure 5(f) show that most data integration methods are safe when the same data set is used for data integration and classifier training. Among these methods, ConQuR is more likely to achieve an over-accurate classifier in the random forest but not k-NN. A similar observation is also noted in the original ConQuR paper (Ling et al., 2022). The results indicate that the double-dipping issue could be mitigated when suitable combinations of data integration methods and classifiers are used. In all the above experiments, MetaDICT consistently demonstrates robust performance in data integration, and the resulting integrated data set can significantly improve the accuracy of the downstream analysis.

2.5 Meta-analysis of CRC microbiome via MetaDICT

This section applies MetaDICT to an integrative analysis of five colorectal cancer (CRC) metagenomic studies to further demonstrate the practical merit (Table S1). The integrative analysis aims to study the generalizable association between microbiome alterations and colorectal cancer. Although focusing on the same disease, these five studies were conducted in different countries, including the United States (US), France (FR), Austria (AT), China (CN) and Germany (DE). To reduce the technical variation, the sequencing data from these five studies were processed using consistent bioinformatics tools for taxonomic profiling. More details on sequencing and bioinformatics analysis can be found in Wirbel et al. (2019) and Supplementary Material. In the integrative analysis, we include age, gender, BMI, country, and disease status as the covariates. In particular, the study is completely confounded with the country since each study is conducted in different countries. Before conducting integrative analysis, we first explore the effect of different studies on the microbial composition and other commonly observed covariates. The results in Figure 6(a) show that the study variable had a dominant effect on the microbial profiles. Since the study is completely confounded with the county of each sample, it is difficult to tell from the explorative analysis that such a effect is due to the difference among countries or the technical variations in different studies, such as different sampling procedures, sample storage, and DNA extraction methods. In addition, there is a clear distribution shift of the observed covariates across studies (Fig-

ure S10(a)). These observations suggest that the potential batch effects could invalidate the integrative analysis, and there is a need for appropriate batch correction that can account for confounding variables and preserve biological variations. We apply MetaDICT to integrate these five studies. Because the design of MetaDICT allows for the separation of the effect of countries and batch effects, the effects of study on microbial composition are significantly reduced while the variation due to different countries is maintained (Figure 6(a)). Furthermore, the effect of disease status on the microbial composition became more significant in the MetaDICT-corrected data set, indicating good preservation of true biological variation (Figure 6(b)).

In the integrative analysis, we first study the underlying structure of microbial communities via the taxa embedding in MetaDICT. As shown in Figure 6(c), the taxa embedding in MetaDICT generates a microbial interaction network and leads to 30 distinct detected microbial communities. In the microbial interaction network, 40% of the total edges are intra-phylum edges, while the expected frequency is 26% if by chance only, suggesting that the taxa from the same phylum tend to be closely connected in the network. Besides taxonomic similarities, the microbial interaction network derived from the embedding can also reflect functional similarities between taxa. For example, two subnetworks within communities 2 and 9 are butyrate generators: one subnetwork has a hub taxon *Butyricicoccus* and other taxa, like *Faecalibacterium*, *Blautia*, *Eubacterium*, *Dorea*, *Lachnospiraceae*, and the other subnetwork includes *Anaerostipes*, *Anaeromassilibacillus*, and *Pseudoflavonifractor*, which possess the genetic pathways necessary for converting pyruvate and acetyl-CoA into butyrate (Medvecky et al., 2018). Another highlighted example is a subnetwork within community 1 that includes several oral and periodontal pathogens, including *Porphyromonas*, *Peptostreptococcus*, *Parvimonas*, *Gemella*, *Tannerella*, *Lachnoanaerobaculum*, and *Solobacterium* (Hampelska et al., 2020; Sabrie et al., 2023; Ternes et al., 2020; Zwezerijnen-Jiwa et al., 2023). In addition to capturing known taxonomic and functional similarities, the detected microbial communities also suggest several interesting observations. Specifically, besides oral pathogens, community 1 also includes several other genera linked to other diseases, such as pathogens involved in opportunistic infection (*Anaerococcus*, *Helcococcus*, and *Providencia*) (Murphy and Frick, 2013; Lotte et al., 2015; Wie, 2015), and pathogens causing bacterial vaginosis (*Mobiluncus*) (Schwebke and Lawing, 2001), suggesting possible synergistic relationships among these pathogens. In addition, several genera from order *Eggerthellales*, including *Eggerthella* and *Adlercreutzia*, are likely to exhibit with butyrate generators in community 9, which could be explained by the hypothesis that they, as acetate consumers, could compete with microbes that convert acetate into butyrate (Noecker et al., 2023). The above observations suggest that the taxa embedding in MetaDICT can capture the taxonomic and functional similarities among taxa and offer new insights into microbial communities.

Besides taxa embedding, the sample embedding in MetaDICT effectively stratifies the samples based on their microbial profiles. We detect four sample subpopulations after applying the community detection algorithm to the sample embedding in MetaDICT. As shown in Figure 6(g), the significant difference in microbial composition across subpopulations is driven by the abundance of *Bacteroides*, *Prevotella*, and *Clostridium*, which are commonly

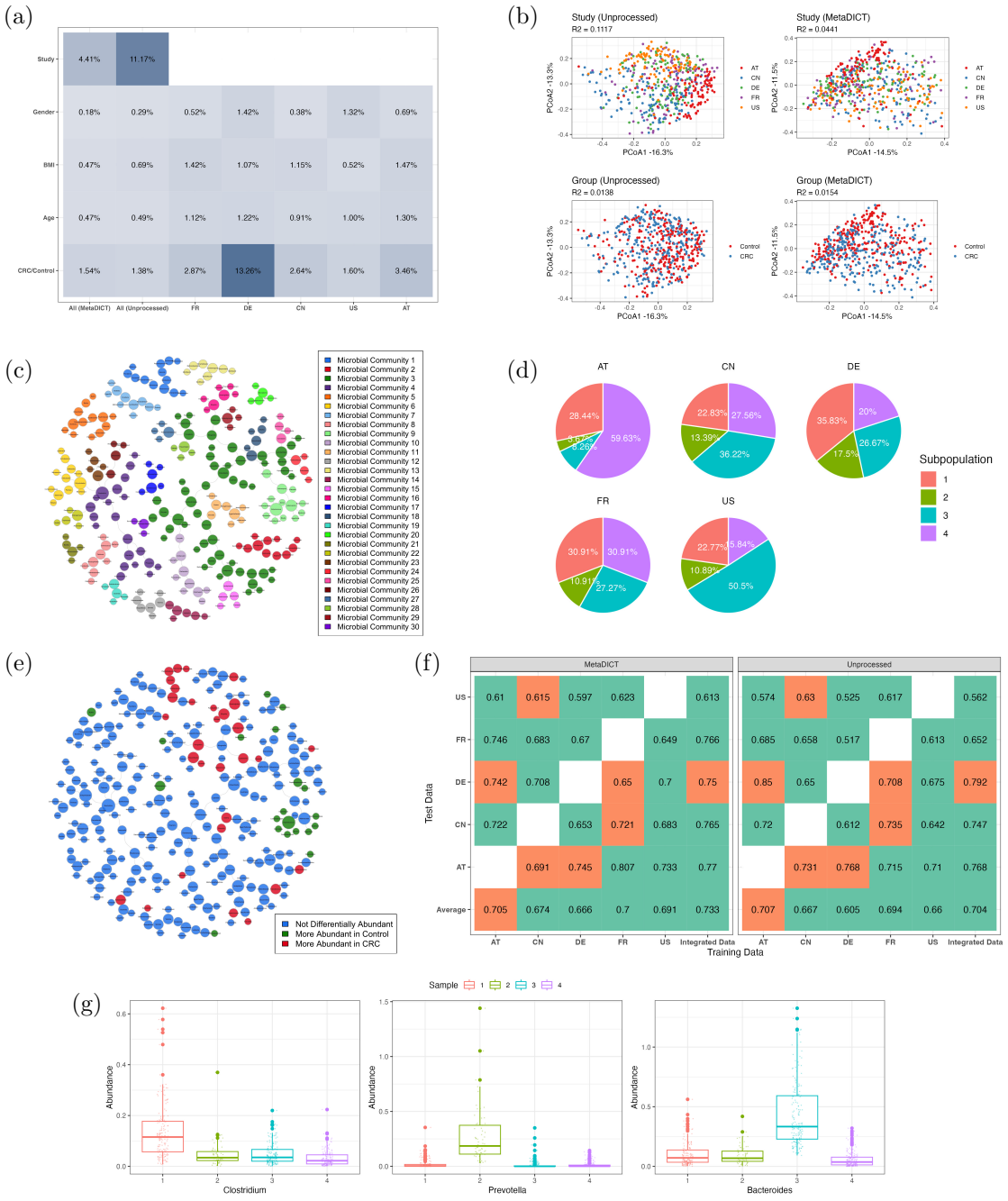

Figure 6: **Meta-analysis of CRC microbiome via MetaDICT.** Figure (a) shows each covariate's contribution to the microbiome variability when PERMANOVA is applied to each study and integrated data set. Figure (b) compares PCoA plots before and after batch effect correction using MetaDICT. Figures (c) and (e) display the microbial interaction network derived from the taxa embedding in MetaDICT, where the community colors genera in (c) and the results of differential abundance analysis color genera in (e). Figure (d) presents the composition of subpopulations in each study, while Figure (g) compares the abundance of three genera. Figure (f) demonstrates the accuracy of the random forest model trained on the data set from each study and integrated data set (green represents the case when accuracy on the MetaDICT-corrected data set is better, while orange represents the case when accuracy on the unprocessed data set is better).

used signatures to characterize enterotypes, i.e., the classification of people defined by the types of bacteria in their gut microbiome (Arumugam et al., 2011; Wu et al., 2011; Costea et al., 2018). Comparing the proportions of these subpopulations in each country suggests that a high proportion of US samples are from subpopulation 3 while DE has more samples from subpopulation 1 (Figure 6(d)). Since enterotype is usually associated with long-term diets, the above observations could result from the fact that the diet in the US includes more high-fat and low-fiber intake, and people in DE take more fermented food than in other countries. Moreover, as expected, subpopulation 2 is more abundant in CN and DE since these two countries have carbohydrate-rich diets (Feng et al., 2015b). In addition, AT exhibits a significantly different composition of enterotypes compared to other countries, as the average age and BMI in that study are greatly higher than in other countries (Figure S10(a)). These results demonstrate the effectiveness of sample embedding in classifying the microbial profiles.

Next, we study the association between the microbial profiles and CRC status via the integrated data set by MetaDICT. More concretely, we apply LinDA to identify differential abundance genera after adjusting the effect of observed covariates. When controlling the false discovery rate at 10%, 50 genera are detected on the integrated data set, while much fewer genera are detected on each data set (Figure S11), highlighting the increased power of integrative analysis. It is interesting to observe that the differentially abundant genera are well clustered in the microbial interaction network derived by the taxa embedding in MetaDICT (Figure 6(e)), indicating that the taxa embedding can provide insight into the pathogenic mechanism of the microbiome. In particular, most differentially abundant genera are grouped in communities 1, 9, 16, and 28, making it convenient to interpret the results from a perspective of functionality in genera. Specifically, several oral pathogens, including *Porphyromonas*, *Peptostreptococcus*, *Parvimonas*, *Gemella*, and *Solobacterium*, in the microbial community 1 are more abundant in CRC samples than control ones, suggesting a close connection between oral microbiome and progression of colorectal cancer (Flemer et al., 2018; Mo et al., 2022). Microbial community 16 includes three differentially abundant genera, i.e., *Fusobacterium*, *Bilophila*, and *Alistipes*, that can produce hydrogen sulfide, a key signaling biomolecule in colorectal cancer (Basic et al., 2017; Tilg et al., 2018; Lin et al., 2023). *Flavonifractor* and *Tyzzerella* from microbial community 28, well-known biomarkers of colorectal cancer (Yang et al., 2021; Wu et al., 2021), are reported as differentially abundant genera. While the differentially abundant genera from the previous three communities associate with colorectal cancer positively, the ones from microbial community 9 show a negative association. The differentially abundant genera from microbial community 9 can mostly produce butyrate, which is a primary energy source for colonocytes and protects against colorectal cancer and inflammatory bowel diseases (Lopez-Siles et al., 2017; Luo et al., 2023). Besides these well-grouped genera, we also discover several isolated differentially abundant genera that can provide insight into the difference between CRC and control samples. For example, three oral-origin genera, *Eikenella*, *Anaeroglobus*, and *Rothia*, that are usually linked to periodontitis disease (Karim et al., 2013; Bao et al., 2017; Mazurel et al., 2023) show a significant association with the development of colorectal cancer, further underscoring the close relationship between the oral microbiome and colorectal cancer

(Zepeda-Rivera et al., 2024). The above discoveries highlight the key role of MetaDICT in increasing power and interpretability for downstream statistical analysis.

Lastly, we aim to use the integrated data set to train a classifier that uses microbial profiles to predict CRC status in disease diagnosis. To compare the generalizability, we evaluate the accuracy of the classifier (measured by ROC-AUC) on the data set from one study and train the classifier on the individual data set from other studies or the integrated data set of other studies. The results of unprocessed and MetaDICT-corrected data sets are summarized in Figure 6(f). Due to the interplay between batch effects and heterogeneity of data sets, the classifiers trained by the unprocessed data cannot generalized well, e.g., the classifier trained on DE can only achieve an accuracy of around 50% in FR (almost randomly guessing). After correcting batch effects by MetaDICT, resulting classifiers' generalizability is mostly improved. For instance, when the training data are from DE and testing data are from FR, prediction accuracy increases from 51.7% to 67% after MetaDICT processes the data sets. As expected, the classifier training on the integrated data set is more accurate than training on the individual data set. Again, the performance of the resulting classifier is enhanced when the batch effects are corrected on the integrated data set. These comparisons underscore the importance of batch effect correction in increasing the generalizability and prediction accuracy of integrative analysis.

3 Discussion

This paper presents a new data integration method for microbiome studies, called MetaDICT. While existing methods mainly explore the relationship between microbial composition and observed covariates, MetaDICT also utilizes the intrinsic structure of microbiome data via shared dictionary learning. By taking advantage of the intrinsic structure, the new approach can better correct batch effects and preserve biological variation than existing methods, especially when unmeasured confounding variables exist or studies are highly heterogeneous in populations. In addition to batch effect correction, MetaDICT generates the embedding at both taxa and sample levels to unravel the hidden structures of microbiome data. Our comprehensive numerical experiments show that the corrected count tables and embedding offered by MetaDICT can improve the commonly used integrative analysis, such as community detection, differential abundance analysis, and outcome prediction.

MetaDICT explores the assumptions and intrinsic structures in multiplicative batch effects, shared dictionary in absolute abundance, and measurement efficiency's smoothness with respect to the taxa similarity. While these assumptions have been discussed in various aspects of microbiome literature, putting them together in MetaDICT offers a fresh perspective on how intrinsic structure can contribute to batch effect correction and data integration in microbiome data. We believe this new perspective will inspire the development of more exciting data integration methods for microbiome data. Moreover, while MetaDICT is designed for microbiome data, its assumptions are flexible enough to potentially work for other types of data, such as single-cell RNA sequencing data, due to the shared similarities in different sequencing protocols. This versatility opens up a world of possibilities for

MetaDICT's application in a broad range of data sets.

The workflow in MetaDICT represents just one way to utilize the intrinsic structures in microbiome data. However, there are likely more intrinsic structures to be explored or alternative ways to explore the same structure. For instance, the nonconvex formulation could be substituted with convex methods to exploit the shared dictionary in absolute abundance. Another possibility is replacing the graph Laplacian with a total variation on a graph to measure the overall smoothness of measurement efficiency (Sadhanala et al., 2016). Due to the nonconvex formulation, the results in MetaDICT could also be influenced by the choice of optimization algorithm, as these algorithms can introduce implicit bias (Gunasekar et al., 2018). This underscores the possibility for further investigation into more efficient workflows for correcting batch effects and integrating data sets than MetaDICT, inspiring the development of similar methods in the field.

4 Methods

4.1 A Model for Microbial Sequencing Data

In microbiome studies, researchers usually measure the abundance of different microbes in each sample by collecting their microbial sequencing data (Lozupone et al., 2007; Vandeputte et al., 2017). However, the observed sequencing count via commonly used sequencing techniques cannot accurately reflect microbial loads (absolute abundance) in each sample due to the bias introduced in the sequencing procedure (McLaren et al., 2019; Wang, 2023a). To characterize such bias, we consider a simple mathematical model to connect the absolute abundance and observed sequencing count

$$O_{i,j,k} \approx w_{i,k} A_{i,j,k} c_{i,j}, \qquad 1 \le i \le m, \quad 1 \le j \le n_i, \quad 1 \le k \le d, \tag{1}$$

where $O_{i,j,k}$ represents the observed sequencing count of taxon k in sample j from study i, and $A_{i,j,k}$ represents the corresponding microbial loads in the sample. There are m studies, n_i samples in study i, and d different microbial taxa. In the above model, the sampling efficiency $c_{i,j}$ characterizes the sample-specific bias related to technical factors such as sequencing depth (Robinson and Oshlack, 2010; Conesa et al., 2016; Wang, 2023b). On the other hand, $w_{i,k}$ is the measurement efficiency of taxon k in study i. The taxon-specific bias characterized by $w_{i,k}$ can be related to the distinct combinations of chemicals and reagents used in different DNA extraction protocols (Morgan et al., 2010), PCR binding and amplification efficiencies in a distinct agreement between primer design and sequences (Polz and Cavanaugh, 1998), and coverage variability with different sequencing platforms (Harismendy et al., 2009). While the above model seems straightforward, several studies have validated the multiplication effects of sample-specific and taxon-specific bias represented in (1) (Conesa et al., 2016; McLaren et al., 2019).

In each sample, the microbes rarely live isolated but coexist as a complex ecosystem (Woyke et al., 2006; Chaffron et al., 2010). In particular, multiple microbial communities naturally form through interactions such as nutritional cross-feeding, co-colonization, and competition (Faust et al., 2012). This observation indicates that the microbes' abundance could

change in a highly correlated way. We consider a mixed membership model for the microbial abundance profiles to capture such structural patterns in abundance. Specifically, let us write the absolute abundance vector in sample j from study i as $\vec{A}_{i,j} = [A_{i,j,1}, \ldots, A_{i,j,d}]^T \in \mathbb{R}^d$. We can approximate $\vec{A}_{i,j}$ as

$$\vec{A}_{i,j} \approx \sum_{l=1}^{r} R_{i,j,l}\vec{D}_l = D\vec{R}_{i,j}, \tag{2}$$

where each D_l represents a group of microbes of which abundances change in a highly correlated way, and the representation $\vec{R}_{i,j} = [R_{i,j,1}, \ldots, R_{i,j,r}]^T \in \mathbb{R}^r$ characterizes the amplitude of changes along these r groups. The combination of these r groups of microbes, $D = [\vec{D}_1, \ldots, \vec{D}_r] \in \mathbb{R}^{d\times r}$, is called a shared dictionary as it represents the universal structural pattern in microbial abundance across the data sets. Although the shared dictionary is universal across studies, the distribution of the representation $\vec{R}_{i,j}$ could differ for different data sets due to the potential heterogeneity across data sets. While the above type of matrix factorization structure has been widely used in the model for a single microbiome data set (Sankaran and Holmes, 2019; Cao et al., 2020; Kim et al., 2023), we extend it as a universal structure across multiple data sets. The design of the shared dictionary for absolute abundance allows for the separation of the batch effects (captured by $w_{i,k}$) and biological variation in absolute abundance (captured by $\vec{R}_{i,j}$).

The sequence structure of each taxon mainly determines the taxon-specific measurement efficiency $w_{i,k}$. For example, DNA extraction efficiencies are related to the cell walls and membrane structure (Krsek and Wellington, 1999; Carrigg et al., 2007), and both PCR binding and sequencing efficiencies are related to the arrangement or organization of nucleotides, especially the GC contents (Polz and Cavanaugh, 1998; Benjamini and Speed, 2012). This observation suggests that sequences with close structures can exhibit similar capture efficiencies, that is,

$$w_{i,k} \approx w_{i,k'}, \qquad \text{if } k \text{ and } k' \text{ are close.} \tag{3}$$

In other words, the measurement efficiency is smooth with respect to the similarity among taxa. The approximations in (1), (2), and (3) are three key assumptions in our model for microbial sequencing data.

4.2 Data Integration via Shared Dictionary Learning

We collect m microbiome data sets from different studies in the data integration. In the ith data set, we observe the microbial sequencing count $\{O_{i,j,k}\}_{1\le j\le n_i, 1\le k\le d}$ from n_i samples and d common taxa and the covariates $\{\vec{X}_{i,j}\}_{1\le j\le n_i}$ from n_i samples, such as age and blood type. We assume that the observed microbial sequencing data follow the model introduced in the previous section (or the three assumptions in (1), (2), and (3)). The model in (1) suggests that our observed sequencing counts are disturbed by the unobserved measurement efficiency, and thus, it is difficult to distinguish whether the variation across multiple data sets results from the confounding effect of measurement efficiency or true biological variation.

Therefore, the main challenge of data integration is to remove the effect of heterogeneous measurement efficiency (also known as batch effects) while maintaining the true biological variation in the microbial loads across different data sets.

We introduce a two-stage method for integrating data sets from multiple studies to address the challenge. The first stage obtains an initial estimator by adjusting the observed covariates, while the second stage further refines the estimation by exploring the shared dictionary of microbial sequencing data. We assume $c_{i,j} = 1$ for the simplicity of analysis in this section, as the sampling efficiency can be normalized by the commonly used normalization methods (Bullard et al., 2010; Paulson et al., 2013; Love et al., 2014; Yuan and Wang, 2023).

Stage 1: Initial Estimation by Covariate Balancing The conventional wisdom in data integration is that the difference in sequencing count distributions is due to batch effects after adjusting the effects of all possible confounding variables, and we shall remove such differences before integrating data sets (Gibbons et al., 2018; Zhang et al., 2020; Ma et al., 2022; Ling et al., 2022; Wang and Lê Cao, 2023). The main assumption behind such a strategy is that the conditional expectations of absolute abundance are the same across different data sets, that is,

$$\mathbb{E}(A_{i,j,k}|\vec{X}_{i,j} = x) = \mathbb{E}(A_{i',j',k}|\vec{X}_{i',j'} = x), \qquad 1 \le k \le d, \quad i \ne i'. \tag{4}$$

Together with the assumption in (1), this assumption naturally leads to

$$\mathbb{E}(O_{i,j,k}|\vec{X}_{i,j} = x)/\mathbb{E}(O_{i',j',k}|\vec{X}_{i',j'} = x) \approx w_{i,k}/w_{i',k}.$$

Consequently, we can adopt treatment effect estimation techniques in causal inference literature to estimate the ratio $w_{i,k}/w_{i',k}$. In particular, the weighting method, one of the most widely used covariate balancing techniques, is particularly suitable for the estimand with the ratio form. The idea of the weighting method is to assign weights $e_{i,j}$ to each sample so that the weighted distributions of covariates $\vec{X}_{i,j}$ are balanced across the data sets (Imbens and Rubin, 2015). There are several ways to estimate weights in the literature, including inverse-probability weighting (Rosenbaum and Rubin, 1983; Robins et al., 2000; Hirano and Imbens, 2001) and balancing weighting (Hainmueller, 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2016; Yu and Wang, 2024). With the estimated weights $e_{i,j}$, we can estimate the ratio $w_{i,k}/w_{i',k}$ by

$$r_{i,i',k} := \widehat{w_{i,k}/w_{i',k}} = \frac{1}{n_i}\sum_j e_{i,j}O_{i,j,k} \Big/ \frac{1}{n_{i'}}\sum_{j'} e_{i',j'}O_{i',j',k}.$$

The definition suggests $r_{i,i,k} = 1$. The ratio estimator $r_{i,i',k}$ offers a straightforward way to correct batch effects and integrate data. Specifically, the corrected and comparable microbial sequencing count can be defined as

$$r_{i,i',k}O_{i',j',k}, \qquad 1 \le i' \le m, \quad 1 \le j' \le n_i, \quad 1 \le k \le d.$$

The above strategy can successfully correct the batch effects when all the confounding variables are observed and the assumption in (4) is satisfied. However, it is common that only a

few covariates are observed across all data sets, and there are several important unobserved confounding variables, such as lifestyle. When the assumption in (4) is invalid, the unobserved confounding variables' effect might be characterized as the batch effect, leading to overcorrection of the batch effect and wrong reduction of the true biological variation. To address such an issue, we propose to refine the above estimation by exploring the intrinsic structure of microbial sequencing data.

Stage 2: Estimation Refinement by Shared Dictionary Learning The first intrinsic structure explored here is the shared dictionary structure introduced in assumption (2). The shared dictionary structure and assumption in (1) suggest that the matrix of observed sequencing count in each data set, $O_i = [O_{i,j,k}]_{1\le k\le d, 1\le j\le n_i} \in \mathbb{R}^{d\times n_i}$, can be factorized as a product of three matrices

$$O_i \approx \mathrm{diag}(\vec{w}_i) D R_i, \qquad 1 \le i \le m,$$

where $\vec{w}_i = (w_{i,1}, \ldots, w_{i,d})^T \in \mathbb{R}^d$ is the vector of measurement efficiency in the ith data set, $\mathrm{diag}(\vec{w}_i) \in \mathbb{R}^{d\times d}$ is a diagonal matrix with diagonal entries as $\vec{w}_i$, and $R_i = [\vec{R}_{i,1}, \ldots, \vec{R}_{i,n_i}] \in \mathbb{R}^{r\times n_i}$ is the representation matrix of ith data set. As one would prefer a concise model, it is natural to assume both D and R_i have some low-rank structure. To capture such shared and low-rank dictionary structure, we can consider the following loss function

$$\mathcal{L}_D(D, \{\vec{w}_i, R_i\}_{1\le i\le m}) = \sum_{i=1}^m \|O_i - \mathrm{diag}(\vec{w}_i) D R_i\|_F^2 + \alpha \left(\sum_{i=1}^m \frac{1}{n_i r} \|R_i\|_F^2 + \frac{1}{dr} \|D\|_F^2 \right),$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. The first term in the above loss function measures the discrepancy between the observation and matrix product, and the second term promotes the low-rank structure in D and $\{R_i\}_{1\le i\le m}$ (Srebro et al., 2004). Another intrinsic structure we can explore here is the smoothness of measurement efficiency as characterized by the assumption in (3). To characterize the similarity of the sequence structure of taxa, we can construct a graph $G = (V, E)$ such that each vertex is a taxon and two taxa are connected if they are similar in the sequence structure. For example, the sequence similarity between taxa can be measured by phylogenetic distance or taxonomic similarity. Given the graph G, we can consider the graph Laplacian to measure the overall smoothness of measurement efficiency with respect to the graph

$$\mathcal{L}_S(\{\vec{w}_i\}_{1\le i\le m}) = \frac{\beta}{d^2} \sum_{i=1}^m \vec{w}_i^T L_G \vec{w}_i = \frac{\beta}{d^2} \sum_{i=1}^m \sum_{(k,k')\in E} L_{G,k,k'} (w_{i,k} - w_{i,k'})^2,$$

where L_G is the graph Laplacian matrix, and $L_{G,k,k'}$ is the weight of the edge (k, k'). Putting these two loss functions together yields the following optimization problem

$$\min_{D, \{\vec{w}_i, R_i\}_{1\le i\le m}} \mathcal{L}_D(D, \{\vec{w}_i, R_i\}_{1\le i\le m}) + \mathcal{L}_S(\{\vec{w}_i\}_{1\le i\le m}), \qquad \text{s.t. } 0 \le w_{i,j} \le 1, \forall\, i, j. \tag{5}$$

In the above optimization problem, the estimation of measurement efficiency is further refined by treating the shared dictionary as an anchor. To solve the above optimization problem, we

can set the results from stage 1 as the initial point so that the estimation for measurement efficiency can be further refined. Since the low-rank matrix factorization has the effect of denoising (Chi et al., 2019), our final corrected and comparable microbial sequencing data are $\hat{D}\hat{R}_i$ for $1 \leq i \leq m$, where $\hat{D}$ and $\hat{R}_i$ are the outputs of the above optimization problem. Notably, the output of the nonconvex optimization problem in (5) highly relies on the choice of initial points, so the estimation in stage 1 is also critical.

4.3 A Practical Workflow of MetaDICT

While the last section introduces a general methodology for data integration, we present a detailed workflow used in all numerical experiments.

1. **Preprocessing** Before correcting batch effects, we apply a popular normalization method to remove the effect of unobserved sampling fraction $c_{i,j}$. We apply RSim (Yuan and Wang, 2023) when the taxa are high-resolution or UQ (Bullard et al., 2010) when the taxa are low-resolution.

2. **Covariate Balancing** In the initial estimation, we choose the inverse-probability weighting to balance the distribution of covariates (Rosenbaum and Rubin, 1983; Robins et al., 2000), where the propensity score is estimated by logistic regression. With estimated weights, the ratio of measurement efficiency is estimated by a weighted average of sequencing count.

3. **Shared Dictionary Learning** In the optimization problem of (5), we use the phylogenetic tree to construct a p-nearest nearest neighbors graph for taxa. The optimization problem is solved by the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) (Byrd et al., 1995), a quasi-Newton method particularly suitable for solving large non-linear optimization problems subject to simple bounds. The initial point for $\{\vec{w}_i\}_{1\leq i\leq m}$ is constructed by the estimation in the previous step, while the initial point for dictionary D and representation matrix $\{R_i\}_{1\leq i\leq m}$ are the singular value decomposition of concatenating corrected microbial sequencing data matrix $O = [\mathrm{diag}(\vec{r}_{i,1})O_1, \ldots, \mathrm{diag}(\vec{r}_{i,m})O_m]$.

In the above workflow, we also need to choose four tunning parameters:

- The parameter p represents the number of nearest neighbors to construct the graph G. In our numerical experiments, we select p from 5 to 10.

- r is an important parameter in the size of D and $\{R_i\}_{1\leq i\leq m}$. In our numerical experiments, we select r as the estimated rank of O. Since we put some penalty in place to promote a low-rank structure, the rank of the resulting D and $\{R_i\}_{1\leq i\leq m}$ might be smaller than r.

- Parameters α and β control the importance of the penalty in the optimization problem. We choose small α and β to ensure a reliable data reconstruction result in our numerical experiments. The effects of α and β are illustrated in Figure S1.

4.4 Representation Learning in MetaDICT

While the proposed method can correct the batch effects robustly, the output from our shared dictionary learning can provide more insight into the microbial sequencing data. Like other matrix factorization methods, the resulting dictionary D and representation matrix $\{R_i\}_{1\leq i\leq m}$ can naturally lead to the embedding of taxa and samples. It is worth noting that the decomposition of D and R_i is unique up to a rotation. We apply Varimax, a technique to find a good rotation, to make the shared dictionary interpretable (Kaiser, 1958).

Embedding of Taxa As discussed in the model for microbial sequencing data, each dictionary column represents a direction in which the microbes are likely to change systemically. This interpretation suggests that we can consider each dictionary row as taxon's embedding. When two rows/representations are similar, the abundance of these two taxa constantly changes similarly. These embeddings of taxa can provide more understanding of taxa. For example, taxa representation can help detect the taxa communities. Specifically, we can construct a k-nearest neighbor graph via the distance matrix of taxa representations and apply some clustering algorithms, such as spectral clustering, the Louvain algorithm (Blondel et al., 2008), or the Walktrap algorithm (Pons and Latapy, 2005). These detected taxa communities can help reveal universal microbial co-occurrence relationships in multiple data sets.

Embedding of Samples Besides the shared dictionary, the representation matrix can provide an important source for understanding these microbial data. In particular, each column of the representation matrix can be interpreted as a sample representation, as it reflects the coordinate in the low-dimensional space spanned by the shared dictionary. Due to the corrected batch effects, these sample representations are cast in a common space and thus are comparable across multiple data sets. Therefore, these representations can be directly useful in the downstream sample analysis, such as sample clustering and classifier construction.

Acknowledgments

The authors acknowledge support from Seed Grant in Personalized Nutrition Initiative and NSF Grants DMS-2113458 and DBI-2243257.

Data Availability

Data set in He et al. (2018) can be downloaded from Qiita (`https://qiita.ucsd.edu/`) under study ID 11757. Data sets in Wirbel et al. (2019) in real data analyses can be found in Zenodo (`https://zenodo.org/`) under the identifier No. 3517209. Data sets in (Duvallet et al., 2017) can be found in Zenodo under the identifier No. 840333.

Code Availability

The R package is available at `https://github.com/BoYuan07/MetaDICT`. All analyses can be found under `https://github.com/BoYuan07/MetaDICT_manuscript_code`.

References

M. Amodio, D. Van Dijk, K. Srinivasan, W. S. Chen, H. Mohsen, K. R. Moon, A. Campbell, Y. Zhao, X. Wang, M. Venkataswamy, A. Desai, V. Ravi, P. Kumar, R. Montgomery, G. Wolf, and S. Krishnaswamy. Exploring single-cell data with deep multitasking neural networks. *Nature Methods*, 16(11):1139–1145, 2019.

M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001.

M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J. Batto, , M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, M. Antolín, F. Artiguenave, H. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, K. Kristiansen, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, P. Bork, and MetaHIT Consortium. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

K. Bao, N. Bostanci, T. Thurnheer, and G. N. Belibasakis. Proteomic shifts in multi-species oral biofilms caused by anaeroglobus geminatus. *Scientific Reports*, 7(1):4409, 2017.

N. Barkas, V. Petukhov, D. Nikolaeva, Y. Lozinsky, S. Demharter, K. Khodosevich, and P. V. Kharchenko. Joint analysis of heterogeneous single-cell rna-seq dataset collections. *Nature Methods*, 16(8):695–698, 2019.

A. Basic, M. Blomqvist, G. Dahlén, and G. Svensäter. The proteins of fusobacterium spp. involved in hydrogen sulfide production from l-cysteine. *BMC Microbiology*, 17:1–10, 2017.

N. T. Baxter, M. T. Ruffin, M. A. Rogers, and P. D. Schloss. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, 8:1–10, 2016.

Y. Benjamini and T. P. Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72–e72, 2012.

V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, 2008.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(1):1–13, 2010.

A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Y. Cao, A. Zhang, and H. Li. Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*, 107(1):75–92, 2020.

C. Carrigg, O. Rice, S. Kavanagh, G. Collins, and V. O'Flaherty. Dna extraction method affects microbial community profiles from soils and sediment. *Applied Microbiology and Biotechnology*, 77:955–964, 2007.

S. Chaffron, H. Rehrauer, J. Pernthaler, and C. Von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7): 947–959, 2010.

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700, 2016.

Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17(1):1–19, 2016.

P. I. Costea, F. Hildebrand, M. Arumugam, F. Bäckhed, M. J. Blaser, F. D. Bushman, W. M. De Vos, S. D. Ehrlich, C. M. Fraser, M. Hattori, C. Huttenhower, I. B. Jeffery, D. Knights, J. D. Lewis, R. E. Ley, H. Ochman, P. W. O'Toole, C. Quince, D. A. Relman, F. Shanahan, S. Sunagawa, J. Wang, G. M. Weinstock, G. D. Wu, G. Zeller, L. Zhao, J. Raes, R. Knight, and P. Bork. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1):8–16, 2018.

C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8(1):1784, 2017.

K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology*, 8(7):e1002606, 2012.

Q. Feng, S. Liang, H. Jia, A. Stadlmayr, L. Tang, Z. Lan, D. Zhang, H. Xia, X. Xu, Z. Jie, L. Su, X. Li, X. Li, J. Li, L. Xiao, U. Huber-Schönauer, D. Niederseer, X. Xu, J. Al-Aama, H. Yang, J. Wang, K. Kristiansen, M. Arumugam, H. Tilg, C. Datz, and J. Wang. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications*, 6(1):6528, 2015a.

S. Feng, R.and Du, Y. Chen, S. Zheng, W. Zhang, G. Na, Y. Li, and C. Sun. High carbohydrate intake from starchy foods is positively associated with metabolic disorders: a cohort study from a chinese population. *Scientific Reports*, 5(1):16919, 2015b.

B. Flemer, R. D. Warren, M. P. Barrett, K. Cisek, A. Das, I. B. Jeffery, E. Hurley, O. Micheal, F.s Shanahan, and W. Paul. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*, 67(8):1454–1463, 2018.

E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J . McIver, J. S. Sauk, R. G. Wilson, B. W. Stevens, J. M. Scott, K. Pierce, A. A. Deik, K. Bullock, F. Imhann, J. A. Porter, A. Zhernakova, J. Fu, R. K. Weersma, C. Wijmenga, C. B. Clish, H. Vlamakis, C. Huttenhower, and R. J. Xavier. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*, 4(2):293–305, 2019.

S. M. Gibbons, C. Duvallet, and E. J. Alm. Correcting for batch effects in case-control microbiome studies. *PLoS Computational Biology*, 14(4):e1006102, 2018.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, pages 25–46, 2012.

K. Hampelska, M. M. Jaworska, Z. Babalska, and T. M. Karpiński. The role of oral microbiota in intra-oral halitosis. *Journal of Clinical Medicine*, 9(8):2484, 2020.

O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, and K. A. Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10:1–13, 2009.

Y. He, W. Wu, H. Zheng, P. Li, D. McDonald, H. Sheng, M. Chen, Z. Chen, G. Ji, Z. Zheng, P. Mujagond, X. Chen, Z. Rong, P. Chen, L. Lyu, X. Wang, C. Wu, N. Yu, Y. Xu, J. Yin, J. Raes, R. Knight, W. Ma, and H. Zhou. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine*, 24(10):1532–1535, 2018.

B. Hie, B. Bryson, and B. Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37(6):685–691, 2019.

K. Hirano and G. W. Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278, 2001.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

M. M. Karim, T. Hisamoto, T. Matsunaga, Y. Asahi, Y. Noiri, S. Ebisu, A. Kato, and H. Azakami. Luxs affects biofilm maturation and detachment of the periodontopathogenic bacterium eikenella corrodens. *Journal of Bioscience and Bioengineering*, 116(3):313–318, 2013.

A. Kim, S. Sevanto, E. R. Moore, and N. Lubbers. Latent dirichlet allocation modeling of environmental microbiomes. *PLoS Computational Biology*, 19(6):e1011075, 2023.

I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

M. Krsek and E. Wellington. Comparison of different methods for the isolation and purification of total community dna from soil. *Journal of Microbiological Methods*, 39(1):1–16, 1999.

M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28:1–26, 2008.

E. S. Lander. Array of hope. *Nature Genetics*, 21(1):3–4, 1999.

A. Langdon, N. Crook, and G. Dantas. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Medicine*, 8:1–16, 2016.

H. Lin, Y. Yu, L. Zhu, N. Lai, L. Zhang, Y. Guo, X. Lin, D. Yang, N. Ren, Z. Zhu, and Q. Dong. Implications of hydrogen sulfide in colorectal cancer: Mechanistic insights and diagnostic and therapeutic strategies. *Redox Biology*, 59:102601, 2023.

W. Ling, J. Lu, N. Zhao, A. Lulla, A. M. Plantinga, W. Fu, A. Zhang, H. Liu, H. Song, Z. Li, J. Chen, T. W. Randolph, W. A. Koay, J. R. White, L. J. Launer, A. A. Fodor, K. A. Meyer, and M. C. Wu. Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, 13(1):5418, 2022.

M. Lopez-Siles, S. H. Duncan, L. J. Garcia-Gil, and M. Martinez-Medina. Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics. *The ISME Journal*, 11(4): 841–852, 2017.

R. Lotte, L. Lotte, N. Degand, A. Gaudart, S. Gabriel, M. Ben H'dech, M. Blois, J. Rinaldi, and R. Ruimy. Infectious endocarditis caused by helcococcus kunzii in a vascular patient: a case report and literature review. *BMC Infectious Diseases*, 15:1–7, 2015.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):1–21, 2014.

C. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585, 2007.

Q. Luo, P. Zhou, S. Chang, Z. Huang, and X. Zeng. Characterization of butyrate-metabolism in colorectal cancer to guide clinical treatment. *Scientific Reports*, 13(1):5106, 2023.

S. Ma, D. Shungin, H. Mallick, M. Schirmer, L. H. Nguyen, R. Kolde, E. Franzosa, H. Vlamakis, R. Xavier, and C. Huttenhower. Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using mmuphin. *Genome Biology*, 23(1):208, 2022.

D. Mazurel, M. Carda-Diéguez, T. Langenburg, M. Žiemytė, W. Johnston, C. P. Martínez, F. Albalat, C. Llena, N. Al-Hebshi, S. Culshaw, A. Mira, and B. T. Rosier. Nitrate and a nitrate-reducing rothia aeria strain as potential prebiotic or synbiotic treatments for periodontitis. *npj Biofilms and Microbiomes*, 9(1):40, 2023.

M. R. McLaren, A. D. Willis, and B. J. Callahan. Consistent and correctable bias in metagenomic sequencing experiments. *Elife*, 8:e46923, 2019.

M. Medvecky, D. Cejkova, O. Polansky, D. Karasova, T. Kubasova, A. Cizek, and I. Rychlik. Whole genome sequencing and function prediction of 133 gut anaerobes isolated from chicken caecum in pure cultures. *BMC Genomics*, 19:1–15, 2018.

A. Milanese, D. R. Mende, L. Paoli, G. Salazar, H. Ruscheweyh, M. Cuenca, P. Hingamp, R. Alves, P. I. Costea, L. P. Coelho, T. S. B. Schmidt, A. Almeida, A. L. Mitchell, R. D. Finn, J. Huerta-Cepas, P. Bork, G. Zeller, and S. Sunagawa. Microbial abundance, activity and population genomic profiling with motus2. *Nature Communications*, 10(1):1014, 2019.

S. Mo, H. Ru, M. Huang, L. Cheng, X. Mo, and L. Yan. Oral-intestinal microbiota in colorectal cancer: inflammation and immunosuppression. *Journal of Inflammation Research*, pages 747–759, 2022.

J. L. Morgan, A. E. Darling, and J. A. Eisen. Metagenomic sequencing of an in vitro-simulated microbial community. *PloS One*, 5(4):e10209, 2010.

E. C. Murphy and I. Frick. Gram-positive anaerobic cocci–commensals and opportunistic pathogens. *FEMS Microbiology Reviews*, 37(4):520–553, 2013.

C. Noecker, J. Sanchez, J. E. Bisanz, V. Escalante, M. Alexander, K. Trepka, A. Heinken, Y. Liu, D.n Dodd, I. Thiele, B. C. DeFelice, and P. J. Turnbaugh. Systems biology elucidates the distinctive metabolic niche filled by the human gut microbe eggerthella lenta. *PLoS Biology*, 21(5):e3002125, 2023.

J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 2013.

M. F. Polz and C. M. Cavanaugh. Bias in template-to-product ratios in multitemplate pcr. *Applied and Environmental Microbiology*, 64(10):3724–3730, 1998.

P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, pages 284–293, 2005.

M. E Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, 11:1–9, 2010.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

N. A. Salad Sabrie, S. Halani, F. Maguire, P. Aftanas, R. Kozak, and N. Andany. Lachnoanaerobaculum orale bacteremia in a patient with acute myeloid leukemia and stomatitis: An emerging pathogen. *IDCases*, 33:e01837, 2023.

V. Sadhanala, Y. Wang, and R. J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. *Advances in Neural Information Processing Systems*, 29, 2016.

K. Sankaran and S. P. Holmes. Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614, 2019.

J. R. Schwebke and L. F. Lawing. Prevalence of mobiluncus spp among women with and without bacterial vaginosis as detected by polymerase chain reaction. *Sexually Transmitted Diseases*, 28(4):195–199, 2001.

N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17, 2004.

D. Ternes, J. Karta, M. Tsenkova, P. Wilmes, S. Haan, and E. Letellier. Microbiome in colorectal cancer: how to get from meta-omics to mechanism? *Trends in Microbiology*, 28 (5):401–423, 2020.

H. Tilg, T. E. Adolph, R. R. Gerner, and A. R. Moschen. The intestinal microbiota in colorectal cancer. *Cancer Cell*, 33(6):954–964, 2018.

P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457 (7228):480–484, 2009.

D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, S. Vermeire, G. Falony, and J. Raes. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681):507–511, 2017.

E. Vogtmann, X. Hua, G. Zeller, S. Sunagawa, A. Y. Voigt, R. Hercog, J. J. Goedert, J. Shi, P. Bork, and R. Sinha. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PloS One*, 11(5):e0155362, 2016.

S. Wang. Multiscale adaptive differential abundance analysis in microbial compositional data. *Bioinformatics*, 39(4):btad178, 2023a.

S. Wang. Robust differential abundance test in compositional data. *Biometrika*, 110(1): 169–185, 2023b.

Y. Wang and K. Lê Cao. Plsda-batch: a multivariate framework to correct for batch effects in microbiome data. *Briefings in Bioinformatics*, 24(2):bbac622, 2023.

J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177 (7):1873–1887, 2019.

S. Wie. Clinical significance of providencia bacteremia or bacteriuria. *The Korean Journal of Internal Medicine*, 30(2):167, 2015.

J. Wirbel, P. T. Pyl, E. Kartal, K. Zych, A. Kashani, A. Milanese, J. S. Fleck, A. Y. Voigt, A. Palleja, R. Ponnudurai, S. Sunagawa, L. P. Coelho, P. Schrotz-King, E. Vogtmann, N. Habermann, E. Niméus, A. M. Thomas, P. Manghi, S. Gandini, D. Serrano, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, L. Waldron, A. Naccarati, N. Segata, R. Sinha, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork, and G. Zeller. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine*, 25(4):679–689, 2019.

T. Woyke, H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin, and N. Dubilier. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950–955, 2006.

G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

Y. Wu, N. Jiao, R. Zhu, Y. Zhang, D. Wu, A. Wang, S. Fang, L. Tao, Y. Li, S. Cheng, X. He, P. Lan, C. Tian, N. Liu, and L. Zhu. Identification of microbial markers across populations in early detection of colorectal cancer. *Nature Communications*, 12(1):3063, 2021.

Y. Yang, L. Du, D. Shi, C. Kong, J. Liu, G. Liu, X. Li, and Y. Ma. Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature Communications*, 12(1):6757, 2021.

T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.

H. Ye, X. Zhang, C. Wang, E. L. Goode, and J. Chen. Batch-effect correction with sample remeasurement in highly confounded case-control studies. *Nature Computational Science*, 3(8):709–719, 2023.

J. Yu, Q. Feng, S. Wong, D. Zhang, Q. yi Liang, Y. Qin, L. Tang, H. Zhao, J. Stenvang, Y. Li, X. Wang, X. Xu, N.g Chen, W. Wu, J. Al-Aama, H. J. Nielsen, P. Kiilerich, B. Jensen, T. Yau, Z. Lan, H. Jia, J. Li, L. Xiao, T. Lam, S. Ng, A. Cheng, V. Wong, F. Chan, Xun Xu, H. Yang, L. Madsen, C. Datz, H. Tilg, J. Wang, N. Brünner, K. Kristiansen, J. Arumugam, M.and Sung, and J. Wang. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1):70–78, 2017.

R. Yu and S. Wang. Treatment effects estimation by uniform transformer. In *The Twelfth International Conference on Learning Representations*, 2024.

B. Yuan and S. Wang. Rsim: A reference-based normalization method via rank similarity. *PLoS Computational Biology*, 19(9):e1011447, 2023.

J. P. Zackular, M. A. Rogers, M. T. Ruffin IV, and P. D. Schloss. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research*, 7(11):1112–1121, 2014.

G. Zeller, J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann, R. Hercog, M. Koch, A. Luciani, D. R. Mende, M. A. Schneider, P. Schrotz-King, C. Tournigand, J. Tran Van Nhieu, T. Yamada, J. Zimmermann, V. Benes, M. Kloor, C. M. Ulrich, M. von Knebel Doeberitz, I. Sobhani, and P. Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):766, 2014.

M. Zepeda-Rivera, S. S. Minot, H. Bouzek, H. Wu, A. Blanco-Míguez, P. Manghi, D. S. Jones, K. D. LaCourse, Y. Wu, E. F. McMahon, S. Park, Y. K. Lim, A. G. Kempchinsky, A. D. Willis, S. L. Cotton, S. C. Yost, E. Sicinska, J. Kook, F. E. Dewhirst, N. Segata, S. Bullman, and C. D. Johnston. A distinct fusobacterium nucleatum clade dominates the colorectal cancer niche. *Nature*, 628(8007):424–432, 2024.

Y. Zhang, G. Parmigiani, and W. E. Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078, 2020.

H. Zhou, K. He, J. Chen, and X. Zhang. Linda: linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, 23(1):95, 2022.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

F. H. Zwezerijnen-Jiwa, H. Sivov, P. Paizs, K. Zafeiropoulou, and J. Kinross. A systematic review of microbiome-derived biomarkers for early colorectal cancer detection. *Neoplasia*, 36:100868, 2023.