mbImpute: an accurate and robust imputation method for microbiome data

Ruochen Jiang [1], Wei Vivian Li [1,2] and Jingyi Jessica Li [1,3,4,*]

Abstract

Microbiome studies have gained increased attention since many discoveries revealed connections between human microbiome compositions and diseases. A critical challenge in microbiome research is that excess false zeros distort taxon abundances, complicate data analysis, and jeopardize the reliability of scientific discoveries. To address this issue, we propose the first imputation method, mbImpute, to identify and recover likely false zeros by borrowing information jointly from similar samples, similar taxa, and optional metadata including sample covariates and taxon phylogeny. Comprehensive simulations verified that mbImpute achieved better imputation accuracy under multiple measures than five state-of-the-art imputation methods designed for non-microbiome data. In real data applications, we demonstrate that mbImpute improved the power and reproducibility of identifying disease-related taxa from microbiome data of type 2 diabetes and colorectal cancer.

Introduction

Microbiome studies explore the collective genomes of microorganisms living in a certain environment such as soil, sea water, animal skin, and human gut. A large number of studies have confirmed the importance of microbiomes in both natural environment and human bodies [1]. For example, new discoveries have revealed the important roles microbiomes play in complex diseases such as obesity [2], diabetes [3], and cancers [4]. These studies have shown the potential of using human microbiomes as biomarkers for disease diagnosis or therapeutic targets for disease treatment [5].

Thanks to the development of high-throughput sequencing technologies in the last decade, researchers have gained access to large-scale and in-depth DNA and RNA sequencing data [6]. Shotgun metagenomic sequencing has been commonly employed to characterize the taxonomic composition of microorganisms in an environment and to study the interplay between microbiomes

[1] Department of Statistics, University of California, Los Angeles, CA 90095-1554
[2] Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Piscataway, NJ 08854
[3] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[4] Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766
[*] To whom correspondence should be addressed. Email: jli@stat.ucla.edu

and host organisms [7–14]. The resulting microbiome whole genome sequencing (WGS) data have enabled the assessment of phylogenetic diversity, interspecies polymorphism, and metabolic pathways. In the processing of microbiome WGS data, sequencing reads are summarized into operational taxonomic units (OTUs). In detail, sequencing reads are first clustered at a pre-specified similarity level independent of taxonomy, with each cluster representing an OTU. Then each OTU is characterized by a representative DNA sequence and a taxonomic lineage, which is obtained by mapping the OTU to a known taxa reference database [7] and a phylogenetic tree, i.e., a branching diagram of evolutionary relationships among species. Typically, OTUs with greater similarities in their representative DNA sequences are placed closer in a phylogenetic tree. After this processing step, a microbiome WGS dataset is converted to an OTU count matrix, whose rows and columns correspond to subjects (i.e., samples) and OTUs (i.e., taxa), respectively.

A critical challenge in microbiome data analysis is the existence of excess zeros in OTU counts. For example, 75.6% of OTU counts are zeros in the data of Zeller et al. [9]. This high level of sparsity is resulted from a mix of true and false zeros. Sample contamination is an important cause of true zeros. For example, a traveler who just returned from a trip might carry some foreign microorganisms, which would only be detected in his/her microbiome sample but have zero counts in other subjects' samples. A prominent reason for the existence of false zeros is the limited sequencing depth constrained by experimental costs. For example, the sequencing cost could increase by 268.9% if researchers increase the sequencing depth from 30x to 100x [15]. This data sparsity issue has challenged the statistical analysis of microbiome data, as most state-of-the-art methods have poor performance on data containing too many zeros. Adding a pseudo-count of one is a common data analysis strategy for handling these excess zeros [16, 17]. However, this simple approach is ad-hoc and does not distinguish true zeros that are biologically meaningful from false zeros. Kaul et al. [18] developed an approach to distinguish between multiple types of zeros and correct only the false zeros due to the limited sequencing depth; however, their correction was still a simple addition of a pseudo-count of one, without accounting for the important fact that the (unobserved) actual counts of these false zeros are unlikely exactly one.

This data sparsity issue has greatly hindered an important microbiome data analysis task, the deferentially abundant (DA) taxon analysis, which is to identify the taxa that exhibit significantly different abundances between two groups of subjects [7]. Microbiome researchers have employed two main approaches to identify DA taxa. The first and most widely-used approach is the Wilcoxon rank-sum test, a nonparametric test that does not have a distributional assumption on data. A major drawback of this approach is that a taxon would be called DA if its zero proportions differ significantly between two groups, but this difference is unlikely biologically meaningful given the prevalence of false zeros. The second approach is to use the zero-inflated negative binomial generalized linear model (ZINB-GLM) to test for the difference of a taxon's mean abundances in two groups [19, 20]. However, a practical concern of using ZINB-GLM is its unrobust performance due to two technical reasons: (1) maximizing the model likelihood is complicated by the existence of local maxima; (2) microbiome data often contain extremely large counts (on the order of 10^8)

that complicate the optimization. Moreover, other microbiome data analysis tasks, such as the construction of a taxon co-occurrence network [21–24], also face the data sparsity challenge. If using the zero-inflated modeling approach, each task calls for a specialized model development, which is often complicated or unrealistic for most microbiome researchers. Hence, a flexible and robust approach is needed to address the data sparsity issue for microbiome research.

Imputation is a widely-used technique to recover missing data and facilitate data analysis. It has various successful applications in many fields such as recommender systems (e.g., the Netflix challenge [25]), image and speech reconstruction [26–28], imputation of unmeasured epigenomics datasets [29], missing genotype prediction in genome-wide association studies [30], and the more recent gene expression recovery in single-cell RNA-sequencing (scRNA-seq) data analysis [31–35]. Microbiome and scRNA-seq data have similar count matrix structures if one considers subjects and taxa as analogs to cells and genes, and both data have excess false zeros. Given the successes of scRNA-seq imputation methods, it is reasonable to hypothesize that imputation will also relieve the data sparsity issue in microbiome data. However, three distinct features of microbiome data make a direct application of existing imputation methods suboptimal. First, microbiome data are often accompanied by metadata including subject covariates and taxon phylogeny, which, however, cannot be used by existing imputation methods. Second, microbiome subjects may not fall into distinct subgroups as cell types, violating a common assumption in many scRNA-seq imputation methods [31, 34, 36]. Third, microbiome data has a much smaller number of subjects (often in hundreds) than the number of cells (often in tens of thousands) in scRNA-seq data, making the deep-learning based imputation methods inapplicable [35, 37]. However, the smaller sample size allows microbiome data to afford an imputation method that focuses more on imputation accuracy than computational time.

Here we propose mbImpute, the first imputation method designed for microbiome data. mbImpute identifies and corrects false zeros and low counts (for ease of terminology, we refer to them as false zeros in the following text) in microbiome OTU count data. The goal of mbImpute is to provide a principled data-driven approach to relieve the data sparsity issue due to excess false zeros. To achieve this, mbImpute leverages three sources of information: OTU count matrix, sample covariates (e.g., subjects' age, gender, and body mass index information), and taxon phylogeny, with the latter two sources optional. mbImpute takes a two-step approach (Fig. 1): it first identifies likely false zeros and second imputes them by borrowing information from similar taxa (determined by both phylogeny and counts), similar samples (in terms of taxon counts), and sample covariates if available (illustration of the imputation step in Supplementary Fig. S1). The imputed data are expected to contain recovered taxon counts and thus facilitate various downstream analyses, such as the identification of DA taxa and the construction of taxon co-existence networks. For microbiome researchers, a clear advantage of using mbImpute, as opposed to handling excess zeros separately in individual analysis tasks, is the greater flexibility in performing data analysis on the imputed data.

Results

mbImpute outperforms non-microbiome imputation methods in recovering missing taxon abundances and empowering DA taxon detection

We first demonstrated the advantages of mbImpute by benchmarking it against five state-of-the-art imputation methods designed for non-microbiome data, including four popular scRNA-seq imputation methods (scImpute [31], SAVER [33], MAGIC [32], and ALRA [34]) and a widely-used general imputation method softImpute [38]. We designed two simulation studies for this benchmarking purpose, and the common goal was to obtain a "complete" microbiome dataset without false zeros, so that imputation accuracy could be evaluated by comparing the imputed data with the complete data. The first simulation used synthetic complete data from a generative model fitted to a real dataset [13], and the second, more realistic simulation took a sub-dataset with fewer than 15% zeros as the complete data from a real dataset [14]. In both simulations (see Supplementary), false zeros were introduced into the complete data by mimicking the observed zero patterns in real datasets, resulting in the zero-inflated data. After applying the six methods to the zero-inflated data in both simulations, we compared their imputation accuracy in three aspects: (1) the mean squared error (MSE) between the imputed data and the complete data, (2) the Pearson correlation between each taxon's abundances in the imputed data and those in the complete data, and (3) the mean and standard deviation of each taxon's abundances in the imputed data vs. those in the complete data. Fig. 2a-d illustrate the comparison results, which indicate that mbImpute achieves the best overall performance in all the three aspects. In particular, Fig. 2c-d and Supplementary Fig. S2 show that the imputed data by mbImpute best resemble the complete data, verifying the advantage of mbImpute in recovering missing taxon abundances in microbiome data.

We next demonstrated that mbImpute is a robust method. The core of mbImpute is to borrow three-way information from similar samples, similar taxa, and sample covariates to impute false zeros in microbiome data (see Methods). In the second simulation, we broke up similar samples in real data when we selected the complete data, a situation not optimal for mbImpute; however, mbImpute still outperforms existing imputation methods (Fig. 2a-b). To further test for the robustness of mbImpute, we designed a third simulation study including four simulation schemes, where information useful for imputation was encoded in sample covariates only, samples only, taxa only, or three sources together (see Supplementary). Fig. 2e shows that mbImpute effectively recovers false zeros and reduces the MSE after applied to the zero-inflated data under every scheme. These results verify the robustness of mbImpute in selectively leveraging useful information for imputation.

We designed the fourth simulation to mimic a typical microbiome study that aims to identify DA taxa between two subject groups. We simulated data for 300 taxa in 120 subjects, 60 per

group (see Supplementary). Fig. 2f shows the two-dimensional visualization of the complete data (without false zeros), zero-inflated data (with false zeros), and imputed data (after mbImpute was applied to zero-inflated data). Compared with the zero-inflated data, the 120 subjects are more clearly separated into two groups after imputation. We next performed the DA taxon analysis to verify that imputation can boost the power of detecting DA taxa from the zero-inflated data. Specifically, we applied three state-of-the-art DA methods: the Wilcoxon rank-sum test, ANCOM [17], and ZINB-GLM [39]. Among the available DA methods, the Wilcoxon rank-sum test is the most widely-used in microbiome studies [9–14], ANCOM is the most-cited microbiome-specific DA method [17], and ZINB-GLM was found as the most desirable count-model-based method in a comparative study [19]. We also implemented the imputation-empowerd DA analysis: applying an imputation method to the zero-inflated data, and then identifying DA taxa from the imputed data. We included two imputation methods: mbImpute and softImpute. We chose softImpute as the benchmark imputation method in this DA analysis for two reasons: first, softImptue is a general imputation method unspecific to a particular data type; second, softImpute was observed to have good performance in the first two simulations (Fig. 2a-d). After imputation, we employed the two-sample t test for DA taxon detection, because each taxon's logarithmic transformed counts (in the complete data) follow a Normal distribution in each subject group (see Supplementary); thus if imputation is effective, the Normal distributions should be recovered and the t test should be more powerful than the Wilcoxon test. To evaluate the accuracy of DA taxon detection, we used the DA taxa detected by the t test on the complete data as the ground truth. Then we calculated the precision, recall and F1 score of each method by comparing its detected DA taxa to the ground truth. Under the p-value threshold of 0.1 (Fig. 2g left), the two imputation-empowered DA methods achieve better recall and F_1 scores than the three existing DA methods. Although ANCOM has the highest precision, it has the lowest and close-to-zero recall, suggesting that it finds too few DA taxa. Between mbImpute and softImpute, results under this p-value threshold do not draw a clear conclusion: mbImpute has a better precision but a worse recall, and the two methods have similar F1 scores. To thoroughly compare the five methods, we plotted their performance at varying thresholds in receiver operating characteristic (ROC) curves (Fig. 2g right), which show that mbImpute has the largest area under the curve (AUC) and outperforms the three DA methods and softImpute.

mbImpute improves the reliability of identifying microbiome markers for type 2 diabetes

We applied three existing DA methods (the Wilcoxon rank-sum test, ANCOM, and ZINB-GLM) and the mbImpute-empowered t test to two type 2 diabetes (T2D) datasets [13, 14]. Fig. 3a shows the number of taxa identified as T2D-enriched in two datasets by each DA method. The Venn diagrams show that, the mbImpute-empowered t test identifies the second most T2D-enriched taxa (around three times of those identified by the Wilcoxon test, more than six times of those identified by

ANCOM, and 82% of those identified by ZINB-GLM). Compared to the other three methods, the mbImpute-empowered t test achieves the best reproducibility: 12% of its identified 123 T2D-enriched taxa are found in both datasets, compared to 11% by Wilcoxon, 5.4% by ANCOM, and 5.3% by ZINB-GLM (see Supplementary Tables S1-S4 for the T2D-enriched taxa identified by each method and Supplementary Fig. S3 for data visualization pre- and post-imputation). Note that we chose the two-sample t test instead of the Wilcoxon rank-sum test after imputation because (1) we observed that taxon abundance distributions are approximately Normal after imputation (Supplementary Fig. S4), and (2) it is more interpretable to test if a taxon has two different means (by the t test) than to test if this taxon has different distributions (by the Wilcoxon test).

We further focused on four genera: Streptococcus, Lactobacillus, Clostridium, and Actinomyces, which have been previously reported as enriched in T2D [40–46] (see Supplementary for the literature evidence). In these four genera, the mbImpute-empowered t test discovers strain-level taxa that are DA and highly enriched in T2D samples but missed by the Wilcoxon test applied to the raw data, as shown in Fig. 3b. Moreover, we observed an interesting phenomenon: some Clostridium strain-level taxa (Fig. 3b left panel, the third genus from the top) are no longer detected as enriched in T2D samples after imputation, seemingly violating our claim that mbImpute empowers DA taxon identification. However, by examining the abundance distributions of such taxa, *Clostridium symbiosum* and *Clostridium citroniae* for example (Fig. 3b right panel top row), we found that their non-zero abundance distributions are hardly distinguishable between the T2D and control samples, suggesting that they are not informative markers for T2D. Nonetheless, the Wilcoxon test identifies them as DA in the raw data because they have different zero proportions between the T2D and control samples. This result shows that mbImpute can help reduce likely false positive DA taxa identified due to excess false zeros. See Supplementary for a discussion on statistical definitions of DA taxa.

We further compared mbImpute with softImpute using *Clostridium symbiosum* and *Clostridium citroniae* as examples. We observed that mbImpute retains well the distributions of non-zero abundances (Fig. 3b right panel middle row), while softImpute alters the distributions by introducing artificial spikes and shrinking the variance (Fig. 3b right panel bottom row). Such distortion of abundance distributions may mislead the DA analysis. Indeed, we found that the softImpute-empowered t test identifies *Clostridium symbiosum* and *Clostridium citroniae* as DA due to the artificial distortion by softImpute. A possible reason is that softImpute is a low-rank matrix factorization method, which imputes missing matrix entries by assuming a global low-rank matrix structure. In contrast, mbImpute focuses more on local structures, i.e., how a matrix entry depends on other entries in the same row or column. The fact that mbImpute better preserves non-zero abundance distributions makes it a more reliable imputation method than softImpute for microbiome data.

mbImpute increases the power and reproducibility of identifying microbiome markers for colorectal cancer

Colorectal cancer (CRC) is one of the most frequently diagnosed cancer and a leading cause of cancer mortality worldwide [9, 10]. We applied three existing DA methods (the Wilcoxon rank-sum test, ANCOM, and ZINB-GLM) and the mbImpute-empowered t test to four CRC datasets [9–12]. Fig. 3c shows the number of taxa identified as CRC-enriched in each dataset by each DA method. The Venn diagrams show that, across the four datasets, the mbImpute-empowered t test identifies the second most enriched taxa (more than twice of those identified by the Wilcoxon test, more than four times of those identified by ANCOM, and more than 72% of those identified by ZINB-GLM). Compared to the other three methods, the mbImpute-empowered t test again achieves the best reproducibility: 42% of its identified 198 CRC-enriched taxa are found in at least two datasets, compared to 31% by Wilcoxon, 24% by ANCOM, and 34% by ZINB-GLM. It is also worth noting that the mbImpute-empowered t test identifies eight CRC-enriched taxa in all the four datasets, while the other three methods identify at most two (see Supplementary Tables S5-S10 for the identified CRC-enriched taxa by each method and Supplementary Fig. S5 for data visualization pre- and post-imputation).

We then focused on five genera: Fusobacterium, Peptostreptococcus, Prevotella, Gemella, and Streptococcus, which have been previously reported as enriched in CRC [47–52] (see Supplementary for the literature evidence). In these five genera, the mbImpute-empowered t test discovers strain-level taxa that are DA and highly enriched in CRC samples but missed by the Wilcoxon test applied to the raw data, as shown in Fig. 3d. We use *Peptostreptococcus anaerobius* as an example to demonstrate the effectiveness of mbImpute in empowering DA taxon identification. This strain-level taxon is identified as enriched in CRC samples by the Wilcoxon test in only two out of the four raw datasets generated by different labs. By closely examining the taxon abundance distributions (Fig. 3d right panel top row), we observed that non-zero abundances consistently have higher densities in the CRC samples than in the control, suggesting that this taxon should have been identified as DA in all the four datasets. The Wilcoxon test misses it in Zeller *et al.*'s and Vogtmann *et al.*'s data because the dominance of zeros obscures the differences between non-zero abundances in the CRC and control samples. However, these non-zero abundances are informative for distinguishing the CRC samples from the control; that is, if we detect this taxon with a high abundance in a patient, we should be aware of the potential implication of CRC and perform further diagnosis. mbImpute helps amplify the non-zero signals by reducing likely false zeros (Fig. 3d right panel bottom row), thus empowering the identification of this taxon as DA in all the four datasets (i.e., smaller p-values after imputation, Fig. 3d right panel).

Discussion

A critical challenge in WGS microbiome data analysis is how to perform statistical inference of taxon abundance from highly sparse and noisy data. Our proposed method, mbImpute, will address this challenge and facilitate microbiome data analysis. mbImpute works by correcting false zeros and retaining taxa's original non-zero abundance distributions after imputation. As the first imputation method designed for microbiome data, mbImpute is shown to outperform multiple state-of-the-art imputation methods developed for other data types. Regarding applications of mbImpute, we demonstrate that the mbImpute-empowered DA analysis has advantages over the existing DA methods in three aspects. First, mbImpute increases the power of DA taxon identification by recovering the taxa that are missed by the existing methods (due to excess zeros) but should be called DA (i.e., having non-zero abundances exhibiting different means between two subject groups). Second, mbImpute reduces the false positive taxa, which are identified by the existing methods (due to different proportions of zeros) but should not be called DA (i.e., having similar non-zero abundances between two subject groups). Third, mbImpute improves the reproducibility of DA taxon identification across studies. Furthermore, we found literature evidence for the DA taxa identified as enriched in T2D or CRC samples after mbImpute was applied, supporting the application potential of mbImpute in revealing microbiome markers for disease diagnosis and therapeutics.

In statistical inference, a popular and powerful technique is the use of indirect evidence by borrowing information from other observations, as seen in regression, shrinkage estimation, empirical Bayes, among many others [53]. Imputation follows the indirect evidence principle, where the most critical issue is to decide what observations to borrow information from so as to increase data accuracy instead of introducing excess biases. To achieve this, mbImpute employs penalized regression to selectively leverage similar samples, similar taxa, and sample covariates to impute likely false zeros, whose identification also follows the indirect evidence principle by incorporating sample covariates into consideration. mbImpute also provides a flexible framework to make use of microbiome metadata: it selectively borrows metadata information when available, but it does not rely on the existence of metadata (see Methods).

In the comparison of mbImpute with softImpute, a general matrix imputation method widely used in other fields, we observed that softImpute's imputed taxon abundances exhibit artificial spikes and smaller variances than those of the original non-zero abundances, possibly due to its low-rank assumption. In contrast, mbImpute is a regression-based method that focuses more on local matrix structures, and we found that it retains well the original non-zero abundance distributions. We will investigate the methodological differences between mbImpute and softImpute in a future study.

We also observed that, after mbImpute was applied, the imputed taxon abundances exhibit a bell-shaped distribution across samples on the logarithmic scale, suggesting that statistical methods based on the Normal distributional assumptions might become suitable and applicable.

For example, we have shown that the two-sample t test works well with the imputed data in the identification of DA taxa. In addition to DA analysis, another possible use of the imputed microbiome data is to construct a taxon-taxon co-existence network, to which network analysis methods may be applied to find taxon modules and hub taxa.

Online Methods

mbImpute methodology

Here we describe mbImpute, a statistical method that corrects prevalent false zeros in microbiome data. As an overview, mbImpute takes an OTU count matrix as input, pre-processes the data, identifies the likely false zeros and imputes them based on the input count matrix, sample metadata, and taxon phylogeny, and finally outputs an imputed count matrix.

Notations and data pre-processing

We denote the sample-by-taxon OTU count matrix as $\mathbf{M} = (M_{ij}) \in \mathbb{N}^{n \times m}$, where n is the number of samples and m is the number of taxa. We denote the sample covariate matrix (i.e., metadata) as $\mathbf{X} \in \mathbb{R}^{n \times q}$, where q denotes the number of covariates plus one (for the intercept). In addition, we define a phylogenetic distance matrix of taxa as $\mathbf{D} = (D_{jj'}) \in \mathbb{N}^{m \times m}$, where $D_{jj'}$ represents the number of edges connecting taxa j and j' in the phylogenetic tree.

mbImpute begins by filtering out the counts for each taxon with mostly zeros to avoid imputing true zeros due to environmental contamination. This filtering step is exactly the same as how Kaul et al. [18] define structural zeros, i.e., true zeros. More specifically, taxon j would be filtered out if the 95% confidence interval of its expected non-zero proportion does not cover zero:

$$\tilde{p}_j - 1.96\sqrt{\frac{\tilde{p}_j(1-\tilde{p}_j)}{n}} > 0\,,$$

where $\tilde{p}_j$ is the observed non-zero proportion of taxon j. This filtering step is called the binomial test. In the mbImpute package, users can choose the filtering threshold.

To account for the varying library sizes (i.e., total counts) of samples, mbImpute first normalizes the count matrix $\mathbf{M}$ by row. The normalized count matrix is denoted as $\mathbf{M}^{(\mathcal{N})} = (M_{ij}^{(\mathcal{N})}) \in \mathbb{N}^{n \times m}$, where

$$M_{ij}^{(\mathcal{N})} = 10^6 \cdot \frac{M_{ij}}{\sum_{j'=1}^{m} M_{ij'}}\,.$$

After this normalization, every sample has a total count of 10^6.

Next, mbImpute applies the logarithmic transformation to the normalized counts so as to reduce the effects of extremely large counts [54]. The resulted log-transformed normalized matrix

is denoted as $\mathbf{Y} = (Y_{ij}) \in \mathbb{N}^{n \times m}$, with

$$Y_{ij} = \log_{10}\left(M_{ij}^{(\mathcal{N})} + 1.01\right),$$

where the value 1.01 is added to make $Y_{ij} > 0$ to avoid the occurrence of infinite values in a later parameter estimation step, following Li and Li [31, 55]. This logarithmic transformation allows us to fit a continuous probability distribution to the transformed data, thus simplifying the statistical modeling. In the following text, we refer to $\mathbf{Y}$ as the sample-by-taxon abundance matrix.

mbImpute step 1: identification of taxon abundances that need imputation

mbImpute assumes that each taxon's abundances (across samples within a subject group), i.e., a column in $\mathbf{Y}$, follow a mixture model. The model consists of two components: a Gamma distribution for the taxon's false zero and low abundances and a Normal distribution for the taxon's actual abundances, with the Normal mean incorporating sample covariate information. Specifically, mbImpute assumes that the abundance of taxon j in sample i, Y_{ij}, follows the following mixture distribution:

$$Y_{ij} \sim p_j \cdot \Gamma\left(\alpha_j, \beta_j\right) + (1 - p_j) \cdot \mathcal{N}\left(X_{i\cdot}^{\mathsf{T}}\gamma_j, \sigma_j^2\right),$$

where $p_j \in (0, 1)$ is the missing rate, i.e., the probability that taxon j is falsely undetected, $\Gamma\left(\alpha_j, \beta_j\right)$ denotes the Gamma distribution with shape parameter $\alpha_j > 0$ and rate parameter $\beta_j > 0$, and $\mathcal{N}\left(X_{i\cdot}^{\mathsf{T}}\gamma_j, \sigma_j^2\right)$ denotes the Normal distribution with mean $X_{i\cdot}^{\mathsf{T}}\gamma_j$ and standard deviation $\sigma_j > 0$. In other words, with probability p_j, Y_{ij} is a missing value that needs imputation; with probability $1 - p_j$, Y_{ij} does not need imputation but reflects the actual abundance of taxon j in sample i. mbImpute models the Normal mean parameter as a linear function of sample covariates: $X_{i\cdot}^{\mathsf{T}}\gamma_j$, where $X_{i\cdot} \in \mathbb{R}^q$ denotes the i-th row in the covariate matrix $\mathbf{X}$, i.e., the covariates of sample i, and $\gamma_j \in \mathbb{R}^q$ represents the q-dimensional covariate effect vector of taxon j. This allows a taxon to have similar expected (actual) abundances in samples with similar covariates.

The intuition behind this model is that taxon j's actual abundance in a sample (i.e., subject) is drawn from a Normal distribution, whose mean depicts the expected abundance given the sample covariates. However, due to the under-sampling issue in sequencing, false zero or low counts could have been introduced into the data, creating another mode near zero in taxon j's abundance distribution. mbImpute models that mode using a Gamma distribution with mean α_j/β_j, which is close to zero.

mbImpute fits this mixture model to taxon j's abundances using the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood estimates $\hat{p}_j$, $\hat{\alpha}_j$, $\hat{\beta}_j$, $\hat{\gamma}_j$, and $\hat{\sigma}_j^2$. Supplementary Fig. S6 shows four examples where the fitted mixture model well captures the bimodality of an individual taxon's abundance distribution. However, some taxa are observed to have an abundance distribution containing a single mode that can be well modelled by a Normal distribution. When that occurs, the EM algorithm would encounter a convergence issue. To fix this, mbImpute uses a likelihood ratio test (LRT) to first decide if the Gamma-Normal mixture model fits to taxon j's

abundances significantly better than a Normal distribution $Y_{ij} \sim \mathcal{N}\left(X_{i\cdot}^{\mathsf{T}}\eta_j, \omega_j^2\right)$ does. Given the maximum likelihood estimates $\hat{\eta}_j$ and $\hat{\omega}_j^2$ and under the assumption that Y_{ij}'s are all independent, the LRT statistic of taxon j is:

$$\Lambda_j = -2\ln \frac{\prod_{i=1}^n f_{\mathcal{N}}\left(Y_{ij}; X_{i\cdot}^{\mathsf{T}}\hat{\eta}_j, \hat{\omega}_j^2\right)}{\prod_{i=1}^n \left[\hat{p}_j \cdot f_{\Gamma}\left(Y_{ij}; \hat{\alpha}_j, \hat{\beta}_j\right) + (1-\hat{p}_j)\cdot f_{\mathcal{N}}\left(Y_{ij}; X_{i\cdot}^{\mathsf{T}}\hat{\gamma}_j, \hat{\sigma}_j^2\right)\right]} ,$$

which asymptotically follows a Chi-square distribution with 3 degrees of freedom (because the mixture model has three more parameters than in the Normal model) under the null hypothesis that the Normal model is the correct model. If the LRT p-value ≤ 0.05, mbImpute uses the mixture model to decide which of taxon j's abundances need imputation. Specifically, mbImpute decides if Y_{ij} needs imputation based on the estimated posterior probability that Y_{ij} comes from the Gamma component:

$$d_{ij} = \frac{\hat{p}_j \cdot f_{\Gamma}\left(Y_{ij}; \hat{\alpha}_j, \hat{\beta}_j\right)}{\hat{p}_j \cdot f_{\Gamma}\left(Y_{ij}; \hat{\alpha}_j, \hat{\beta}_j\right) + (1-\hat{p}_j)\cdot f_{\mathcal{N}}\left(Y_{ij}; X_{i\cdot}^{\mathsf{T}}\hat{\gamma}_j, \hat{\sigma}_j^2\right)} ,$$

where $\Gamma(\cdot; \hat{\alpha}_j, \hat{\beta}_j)$ and $f_{\mathcal{N}}(\cdot; X_{i\cdot}^{\mathsf{T}}\hat{\gamma}_j, \hat{\sigma}_j^2)$ represent the probability density functions of the estimated Gamma and Normal components in the mixture model. Otherwise, if the LRT p-value > 0.05, mbImpute concludes that none of taxon j's abundances need imputation and sets $d_{1j} = \cdots = d_{nj} = 0$.

Based on the d_{ij}'s, mbImpute defines a set Ω of (sample, taxon) pairs whose abundances are unlikely missing and thus do not need imputation:

$$\Omega = \{(i,j) : d_{ij} < d_{\text{thre}}, i = 1, \ldots, n; j = 1, \ldots, m\} ,$$

and a complement set Ω^c containing other (sample, taxon) pairs whose abundances need imputation:

$$\Omega^c = \{(i,j) : d_{ij} \geq d_{\text{thre}}, i = 1, \ldots, n; j = 1, \ldots, m\} .$$

Although $d_{\text{thre}} = 0.5$ is used as the default threshold on d_{ij}'s to decide the abundances that need imputation, mbImpute is fairly robust to this threshold choice because most d_{ij}'s are concentrated around 0 or 1. We show this phenomenon in Supplementary Fig. S7, which displays the distribution of all the d_{ij}'s in the data from Zeller et al. [9], Feng et al. [10], Yu et al. [11], Vogtmann et al. [12], Qin et al. [14], and Karlsson et al. [13].

To summarize, mbImpute does not impute all zeros in the OTU count matrix; instead, it first identifies the abundances that are likely missing using a mixture-modelling approach, and it then only imputes these values in the next step.

mbImpute step 2: imputation of the missing taxon abundances

In step 1, mbImpute identifies a set Ω of the (sample, taxon) pairs whose abundances do not need imputation. To impute the abundances in Ω^c, mbImpute first learns inter-sample and inter-taxon

relationships from Ω by training a predictive model for Y_{ij}, the abundance of taxon j in sample i. The rationale is that taxon j should have similar abundances in similar samples, and that in every sample, the taxa similar to taxon j should have abundances similar to taxon j's abundance. In addition, sample covariates are assumed to carry predictive information of taxon abundances. Hence, for interpretability and stability reasons, mbImpute uses a linear model to combine the predictive power of similar taxa, similar samples, and sample covariates:

$$Y_{ij} = Y_{i\cdot}^{\mathsf{T}}\kappa_j + Y_{\cdot j}^{\mathsf{T}}\tau_i + X_{i\cdot}^{\mathsf{T}}\zeta_j + \epsilon_{ij}\,,$$

where $Y_{i\cdot} \in \mathbb{R}^m$ denotes the m taxa's abundances in sample i, $Y_{\cdot j} \in \mathbb{R}^n$ denotes taxon j's abundances in the n samples, $X_{i\cdot} \in \mathbb{R}^q$ denotes sample i's covariates (including the intercept), and ϵ_{ij} is the error term. The parameters to be estimated include $\kappa_j \in \mathbb{R}^m$, $\tau_i \in \mathbb{R}^n$ and $\zeta_j \in \mathbb{R}^q$, $i = 1, \ldots, n$; $j = 1, \ldots, m$. Note that κ_j represents the m taxa's coefficients (i.e., weights) for predicting taxon j, with the j-th entry set to zero, so that taxon j would not predict itself; τ_i represents the n samples' coefficients (i.e., weights) for predicting sample i, with the i-th entry set to zero, so that sample i would not predict itself; ζ_j represents the coefficients of sample covariates for predicting taxon j. In the model, the first term $Y_{i\cdot}^{\mathsf{T}}\kappa_j$ borrows information across taxa, the second term $Y_{\cdot j}^{\mathsf{T}}\tau_i$ borrows information across samples, and the third term $X_{i\cdot}^{\mathsf{T}}\zeta_j$ borrows information from sample covariates. The total number of unknown parameters is $m(m-1)+n(n-1)+mq$, while our data $\mathbf{Y}$ and $\mathbf{X}$ together have $nm+nq$ values only. Given that often $m \gg n$, the parameter estimation problem is high dimensional, as the number of parameters far exceeds the number of data points. mbImpute performs regularized parameter estimation by using the Lasso-type ℓ_1 penalty, which leads to good prediction and simultaneously selects predictors (i.e., similar samples and similar taxa) to ease interpretation. That is, mbImpute estimates the above parameters by minimizing the following loss function:

$$L\left(\{\kappa_j, \zeta_j\}_{j=1}^m, \{\tau_i\}_{i=1}^n\right) := \sum_{(i,j)\in\Omega}\left[Y_{ij} - \left(Y_{i\cdot}^{\mathsf{T}}\kappa_j + Y_{\cdot j}^{\mathsf{T}}\tau_i + X_{i\cdot}^{\mathsf{T}}\zeta_j\right)\right]^2 + \lambda\left(\sum_{j=1}^m\sum_{j'\neq j}^m D_{jj'}^{\psi}|\kappa_{jj'}| + \sum_{i=1}^n\sum_{i'\neq i}^n|\tau_{ii'}|\right)\,,$$

where $\lambda, \psi \geq 0$ are tuning parameters chosen by cross-validation, $D_{jj'}$ represents the phylogenetic distance between taxa j and j', $\kappa_{jj'}$ represents the j'-th element of κ_j, and $\tau_{ii'}$ represents the i'-th element of τ_i. Here $D_{jj'}^{\psi}$, i.e., $D_{jj'}$ to the power of ψ, represents the penalty weight of $|\kappa_{jj'}|$.The intuition is that if two taxa are closer in the phylogenetic tree, they are more closely related in evolution and tend to have more similar DNA sequences and biological functions [56, 57], and thus we want to borrow more information between them. For example, if $D_{j_1j_2} > D_{j_1j_3}$, i.e., taxa j_1 and j_2 are farther away than taxa j_1 and j_3 in the phylogenetic tree, then the estimate of $\kappa_{j_1j_2}$ will be more likely shrunk to zero than the estimate of $\kappa_{j_1j_3}$, and mbImpute would use taxon j_3's abundance more than taxon j_2's to predict taxon j_1's abundance. The tuning parameter ψ is introduced because the distance $D_{jj'}$, the number of edges connecting taxa j and j', may not be the best penalty weight for prediction purpose. Choosing ψ by cross-validation is expected to

enhance the predication accuracy.

mbImpute performs the estimation using the R package `glmnet` [58] and obtains the parameter estimates: $\hat{\kappa}_j \in \mathbb{R}^m$, $\hat{\tau}_i \in \mathbb{R}^n$, and $\hat{\zeta}_j \in \mathbb{R}^q$, $i = 1, \ldots, n$; $j = 1, \ldots, m$. Finally, for $(i, j) \in \Omega^c$, the abundance of taxon j in sample i is imputed as:

$$\hat{Y}_{ij} = Y_{i\cdot}^{\mathsf{T}} \hat{\kappa}_j + Y_{\cdot j}^{\mathsf{T}} \hat{\tau}_i + X_{i\cdot}^{\mathsf{T}} \hat{\zeta}_j \,,$$

and mbImpute does not alter Y_{ij} if $(i, j) \in \Omega$.

Note that mbImpute does not require the availability of the sample covariate matrix $\mathbf{X}$ or the phylogenetic tree. In the absence of sample covariates, the loss function becomes

$$L\left(\{\kappa_j\}_{j=1}^m, \{\tau_i\}_{i=1}^n\right) := \sum_{(i,j)\in\Omega} \left(Y_{ij} - \left(Y_{i\cdot}^{\mathsf{T}} \kappa_j + Y_{\cdot j}^{\mathsf{T}} \tau_i\right)\right)^2 + \lambda \left(\sum_{j=1}^m \sum_{j'\neq j}^m D_{jj'}^{\psi} |\kappa_{jj'}| + \sum_{i=1}^n \sum_{i'\neq i}^n |\tau_{ii'}|\right) ,$$

minimizing which returns the parameter estimates: $\hat{\kappa}_j \in \mathbb{R}^m$ and $\hat{\tau}_i \in \mathbb{R}^n$, $i = 1, \ldots, n$; $j = 1, \ldots, m$. Finally, for $(i, j) \in \Omega^c$, the abundance of taxon j in sample i is imputed as:

$$\hat{Y}_{ij} = Y_{i\cdot}^{\mathsf{T}} \hat{\kappa}_j + Y_{\cdot j}^{\mathsf{T}} \hat{\tau}_i \,,$$

and mbImpute does not alter Y_{ij} if $(i, j) \in \Omega$. In the absence of the phylogenetic tree, mbImpute sets $D_{jj'} = 1$ for all $j \neq j' \in \{1, \ldots, m\}$.

When m is large, mbImpute does not estimate $m(m-1) + n(n-1) + mq$ parameters but uses the following strategy to increase its computational efficiency. For each taxon j, mbImpute selects the k taxa closest to it (excluding itself) in phylogenetic distance and sets the other $(n-k)$ taxa's coefficients in κ_j to zero. This strategy reduces the number of parameters to $mk + n(n-1) + mq$ and the computational complexity from $O(m^2)$ to $O(m)$.

In summary, mbImpute step 2 includes two phases: training on Ω and prediction (imputation) on Ω^c, as illustrated in Supplementary Fig. S1.

Imputation methods

We compared mbImpute with five existing imputation methods designed for non-microbiome data: softImpute and four scRNA-seq imputation methods (scImpute, SAVER, MAGIC, and ALRA). All these imputation methods take a count matrix as input and ouput an imputed count matrix with the same dimensions.

softImpute

We used `R` package `softImpute` (version 1.4) and the following command to impute an OTU count matrix (a sample-by-taxon matrix):

```
complete(OTU_count_matrix, softImpute(OTU_count_matrix, rank.max = cv.rankmax))
```

where `rank.max` was chosen by 10-fold cross-validation.

scImpute

We used `R` package `scImpute` (version 0.0.9) with the input as a taxon-by-sample count matrix (transpose of the matrix in Fig. 1):

```
scimpute(count_path = "OTU_count_matrix_trans.csv", Kcluster = 1, out_dir = "sim_imp")
```

where `OTU_count_matrix_trans.csv` is the input file containing the transposed OTU count matrix.

SAVER

We used `R` package `SAVER` (version 1.1.2) with the input as a taxon-by-sample count matrix (transpose of the matrix in Fig. 1):

```
saver(t(OTU_count_matrix), ncores = 1, estimates.only = TRUE)
```

MAGIC

We used `Python` package `MAGIC` (version 2.0.3) and the following commands to impute an OTU count matrix:

```
magic_op = magic.MAGIC()
magic_op.set_params(n_pca = 40)
magic_op.fit_transform(OTU_count_matrix)
```

ALRA

We applied `R` functions `normalize_data`, `choose_k`, and `alra`, which were released on Aug 10, 2019 at `https://github.com/KlugerLab/ALRA`, and the following commands to impute an OTU count matrix:

```
normalized_mat = normalize_data(OTU_count_matrix)
k_chosen = choose_k(normalized_mat, K = 49, noise_start = 44)$k
alra(normalized_mat, k = k_chosen)$A_norm_rank_k_cor_sc
```

DA analysis methods

In both simulation and real data studies, we compared the mbImpute-empowered t test and the softImpute-empowered t test, which apply to log-transformed taxon abundances, with three existing DA methods: the Wilcoxon rank-sum test, ANCOM, and ZINB-GLM, which apply to taxon counts. Each method calculates a p-value for each taxon and identifies the DA taxa by setting a p-value threshold to control the false discovery rate (FDR). See Supplementary for the statistical definitions of DA taxa.

Wilcoxon rank-sum test

We implemented the Wilcoxon rank-sum test using the `R` function `pairwise.wilcox.test` in the package `stats` (version 3.5.1). For each taxon, we performed the test on its counts in two subject groups to obtain a p-value, which suggests if this taxon is DA between the two groups. In simulations, we used the following command to implement a two-sided test for each taxon:

```
pairwise.wilcox.test(x = taxon_counts, g = condition, p.adjust.method = "none")
```

In real data analysis, we used the following command to implement a one-sided test to find if a taxon is disease-enriched (the first condition is the disease condition) and obtained a p-value:

```
pairwise.wilcox.test(x = taxon_counts, g = condition, p.adjust.method = "none",
alternative = "greater")
```

ANCOM

We used the `ANCOM.main` function released on Sep 27, 2019 at `https://github.com/FrederickHuangLin/ANCOM` [17]. Since this function does not provide an option for a one-sided test, we used its default settings and reported its identified DA taxa based on a two-sided test with a significance level 0.1 (`sig = 0.1`), in both simulations and real data analysis. We note that no external FDR control was implemented. Specifically, we used the following command to obtain the result of ANCOM:

```
ANCOM.main(OTU_count_matrix, covariate_matrix, adjusted = F, repeated = F, main.var
= "condition", adj.formula = NULL, repeat.var = NULL, multcorr = 2, sig = 0.1, prev.cut
= 0.90, longitudinal = F)
```

where `OTU_count_matrix` is a sample-by-taxon count matrix and `covariate_matrix` is a sample-by-covariate matrix, same as the input of mbImpute.

ZINB-GLM

We implemented the ZINB-GLM method using the `R` function `zeroinfl` in the package `pscl` (version 1.5.2). For each taxon, the `condition` variable is a group indicator (treatment or control) included as a predictor in the generalized linear model (GLM). The partial Wald test was used to test if the coefficient of the `condition` variable is significantly different from 0. For each taxon, we used the following command to implement the ZINB-GLM method:

```
summary(zinb <- zeroinfl(OTU_count_matrix[,i] ~ condition, dist = "negbin"))
```

In simulations, we used the output two-sided p-value for each taxon. In real data analysis, we were interested in the disease-enriched taxa, so we converted the output two-sided p-value into a one-side p-value as follows:

- If the estimated coefficient is non-negative, we divided the p-value by two;
- otherwise, we set the p-value to 1.

mbImpute-empowered t test and softImpute-empowered t test

For mbImpute-empowered t test, we applied mbImpute (in `R` package `mbImpute`, version 0.0.1) to samples in each subject group and then collected the subject groups together to obtain the imputed data, which have the same dimensions as the original data.

For softImpute-empowered t test, we applied softImpute (in `R` package `softImpute`, version 1.4) to samples in each subject group and then collected the subject groups together to obtain the imputed data, which have the same dimensions as the original data. Specifically, we used the following command to obtain the imputed data for a subject group (condition 1):

```
complete(raw_data_condition1, softImpute(raw_data_condition1, rank.max = cv.rankmax))
```

where `rank.max` was chosen by 10-fold cross-validation.

Then for each taxon, we performed the two-sample t test on the imputed data of the scale

$$\log_{10}(\text{imputed count } + 1.01)$$

instead of the original count matrix to obtain a p-value, which suggests if this taxon is DA between the two groups. In simulations, we used the following command to implement a two-sided test for each taxon:

```
pairwise.t.test(x = taxon_imputed, g = condition, p.adjust.method = "none")
```

In real data analysis, we used the following command to implement a one-sided test to find if a taxon is disease-enriched (the first condition is the disease condition) and obtained a p-value:

```
pairwise.t.test(x = taxon_imputed, g = condition, p.adjust.method = "none",
alternative = "greater")
```

For the Wilcoxon rank-sum test, ZINB-GLM, and mbImpute-empowered or softImpute-empowered t test, after obtaining the p-values of all taxa and collecting them into a vector `p_values`, we adjusted them for FDR control using the `R` function `p.adjust` in the package `stats` (version 3.5.1):

```
p.adjust(p_values, method = "fdr")
```

Then we set the FDR threshold to 0.1 in both simulation and real data analysis. The taxa whose adjusted p-values did not exceed this threshold were called DA. ANCOM directly outputs the DA taxa.

To calculate the ROC curve of each method, we applied the `roc` function in the `R` package `pROC` (version 1.15.3) to the method's output taxon p-values. Then we used the `auc` function in the same package to calculate the AUC values. The 95% ROC confidence bands were generated by bootstrapping the taxa.

T2D and CRC datasets

We applied mbImpute to six real microbiome datasets, each corresponding to an independent study on the relationship between microbiomes and the occurrence of a human disease. All these six datasets were generated by the whole genome shotgun sequencing and are available in the `R` package `curatedMetagenomicData` [59]. We compared the disease-enriched DA taxa identified

by each of four DA methods, namely the Wilcoxon rank-sum test, ANCOM, ZINB-GLM, and the mbImpute-empowered t test. Below is the description of the six datasets and our analysis.

Two datasets are regarding T2D [13, 14]. The Karlsson *et al.* data contain 145 fecal samples from 70-year-old European women for studying the relationship between human gut microbiome compositions and T2D status. The samples/subjects are in three groups: 53 women with T2D, 49 women with impaired glucose tolerance (IGT), and 43 women as the normal control (CON). The twelve sample covariates include the study condition, the subject's age, the number of reads in each sample, the triglycerides level, the hba1c level, the ldl (low-density lipoprotein cholesterol) level, the c peptide level, the cholesterol level, the glucose level, the adiponectin level, the hscrp level, and the leptin level. In our analysis, we considered the 344 taxa at the strain level with phylogenetic information available in the `R` package `curatedMetagenomicData`. Qin et al. [14] performed deep shotgun metagenomic sequencing on 369 Chinese T2D patients and non-diabetic controls (CON). The three sample covariates include the study condition, the body mass index, and the number of reads in each sample. We analyzed 469 taxa at the strain level with phylogenetic information. From both datasets, we identified T2D-enriched taxa by comparing the T2D and CON groups.

Four datasets are regarding CRC [9–12]. Zeller et al. [9] and Feng et al. [10] studied CRC-related microbiomes in three conditions: CRC, small adenoma (ADE; diameter < 10 mm), and control (CON). Zeller et al. [9] sequenced the fecal samples of patients across two countries (France and Germany) in these three groups: 191 patients with CRC, 66 patients with ADE, and 42 patients in CON. The sample covariates include the study condition, the subject's age category, gender, body mass index and country, and the number of reads in each sample. We included 486 taxa at the strain level with phylogenetic information. Feng et al. [10] sequenced samples from 154 human subjects aged between 45–86 years old in Australia, including 46 patients with CRC, 47 patients with ADE, and 61 in CON. The sample covariates include the study condition, the subject's age category, gender and body mass index, and number of reads in each sample. We included 449 taxa at the strain level in our analysis. Yu et al. [11] and Vogtmann et al. [12] studied CRC-related microbiomes in two conditions: CRC vs. CON. In detail, Yu et al. [11] sequenced 128 Chinese samples, including 75 patients with CRC and 53 patients in CON. The sample covariates include the study condition and the number of reads in each sample. We studied 417 taxa at the strain level. Vogtmann et al. [12] included 104 subjects from Washington DC and sequenced their fecal samples, including 52 with CRC and 52 in CON. The sample covariates include the study condition, the subject's age category, gender and body mass index, and number of reads in each sample. We included 412 taxa at the strain level. From all the four datasets, we identified CRC-enriched taxa by comparing the CRC and CON groups.

Software and code

The `mbImpute R` package and the code for simulation and real data analysis are available at https://github.com/ruochenj/mbImpute

Acknowledgements

The authors would like to thank Dr. Hongzhe Li at University of Pennsylvania for pointing us to this research direction. The authors also appreciate the comments and feedback from the members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

Funding

This work was supported by the following grants: National Science Foundation DMS-1613338 and DBI-1846216, NIH/NIGMS R01GM120507, PhRMA Foundation Research Starter Grant in Informatics, Johnson & Johnson WiSTEM2D Award, and Sloan Research Fellowship (to J.J.L).

References

[1] Katherine R Amato. An introduction to microbiome analysis for human biology applications. *American Journal of Human Biology*, 29(1):e22931, 2017.

[2] Peter J Turnbaugh, Ruth E Ley, Michael A Mahowald, Vincent Magrini, Elaine R Mardis, and Jeffrey I Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*, 444(7122):1027, 2006.

[3] Buck S Samuel and Jeffrey I Gordon. A humanized gnotobiotic mouse model of host–archaeal–bacterial mutualism. *Proceedings of the National Academy of Sciences*, 103(26): 10011–10016, 2006.

[4] Elaine Holmes, Jia V Li, Thanos Athanasiou, Hutan Ashrafian, and Jeremy K Nicholson. Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends in microbiology*, 19(7):349–359, 2011.

[5] Zhang Xinyan, Mallick Himel, and Yi Nengjun. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *Journal of Bioinformatics and Genomics*, (2 (2)), 2016. ISSN 2530-1381. doi: 10.18454/jbg.2016.2.2.1. URL `http://journal-biogen.org/article/view/12`.

[6] John Besser, Heather A Carleton, Peter Gerner-Smidt, Rebecca L Lindsey, and Eija Trees. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical microbiology and infection*, 24(4):335–341, 2018.

[7] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[8] Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, et al. Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in microbiology*, 7:459, 2016.

[9] Georg Zeller, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*, 10(11), 2014.

[10] Qiang Feng, Suisha Liang, Huijue Jia, Andreas Stadlmayr, Longqing Tang, Zhou Lan, Dongya Zhang, Huihua Xia, Xiaoying Xu, Zhuye Jie, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature communications*, 6:6528, 2015.

[11] Jun Yu, Qiang Feng, Sunny Hei Wong, Dongya Zhang, Qiao yi Liang, Youwen Qin, Longqing Tang, Hui Zhao, Jan Stenvang, Yanli Li, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, 66(1):70–78, 2017.

[12] Emily Vogtmann, Xing Hua, Georg Zeller, Shinichi Sunagawa, Anita Y Voigt, Rajna Hercog, James J Goedert, Jianxin Shi, Peer Bork, and Rashmi Sinha. Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PloS one*, 11(5), 2016.

[13] Fredrik H Karlsson, Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.

[14] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[15] Kirsten JM van Nimwegen, Ronald A van Soest, Joris A Veltman, Marcel R Nelen, Gert Jan van der Wilt, Lisenka ELM Vissers, and Janneke PC Grutters. Is the 1000 genome as near as we think? a cost analysis of next-generation sequencing. *Clinical chemistry*, 62(11):1458–1464, 2016.

[16] Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.

[17] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1):27663, 2015.

[18] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:2114, 2017.

[19] Lizhen Xu, Andrew D Paterson, Williams Turpin, and Wei Xu. Assessment and selection of competing models for zero-inflated microbiome data. *PloS one*, 10(7), 2015.

[20] Jun Chen, Emily King, Rebecca Deek, Zhi Wei, Yue Yu, Diane Grill, and Karla Ballman. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4):643–651, 2018.

[21] M Claire Horner-Devine, Jessica M Silver, Mathew A Leibold, Brendan JM Bohannan, Robert K Colwell, Jed A Fuhrman, Jessica L Green, Cheryl R Kuske, Jennifer BH Martiny, Gerard Muyzer, et al. A comparison of taxon co-occurrence patterns for macro-and microorganisms. *Ecology*, 88(6):1345–1353, 2007.

[22] Albert Barberán, Scott T Bates, Emilio O Casamayor, and Noah Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal*, 6(2):343–351, 2012.

[23] Jarishma K Gokul, Andrew J Hodson, Eli R Saetnan, Tristram DL Irvine-Fynn, Philippa J Westall, Andrew P Detheridge, Nozomu Takeuchi, Jennifer Bussell, Luis AJ Mur, and Arwyn Edwards. Taxon interactions control the distributions of cryoconite bacteria colonizing a high arctic ice cap. *Molecular ecology*, 25(15):3752–3767, 2016.

[24] Ilma Tapio, Daniel Fischer, Lucia Blasco, Miika Tapio, R John Wallace, Ali R Bayat, Laura Ventto, Minna Kahala, Enyew Negussie, Kevin J Shingfield, et al. Taxon abundance, diversity, co-occurrence and network analysis of the ruminal microbiota in response to dietary changes in dairy cows. *PloS one*, 12(7), 2017.

[25] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. Citeseer, 2007.

[26] Sarat C Dass and Vijayan N Nair. Edge detection, spatial smoothing, and image reconstruction with partially observed multivariate data. *Journal of the American Statistical Association*, 98(461):77–89, 2003.

[27] Friedrich Faubel, John McDonough, and Dietrich Klakow. Bounded conditional mean imputation with gaussian mixture models: A reconstruction approach to partly occluded features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872. IEEE, 2009.

[28] Valeria Rulloni, Oscar Bustos, and Ana Georgina Flesia. Large gap imputation in remote sensed imagery of the environment. *Computational Statistics & Data Analysis*, 56(8):2388–2403, 2012.

[29] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33(4):364, 2015.

[30] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

[31] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):1–9, 2018.

[32] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

[33] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542, 2018.

[34] George C Linderman, Jun Zhao, and Yuval Kluger. Zero-preserving imputation of scrna-seq data using low-rank approximation. *bioRxiv*, page 397588, 2018.

[35] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*, 10(1):1–14, 2019.

[36] Mengjie Chen and Xiang Zhou. Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. *Genome biology*, 19(1):1–15, 2018.

[37] Cédric Arisdakessian, Olivier Poirion, Breck Yunits, Xun Zhu, and Lana X Garmire. Deep-impute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome biology*, 20(1):1–14, 2019.

[38] T Hastie and R Mazumder. softimpute: Matrix completion via iterative soft-thresholded svd. *R package version*, 1:p1, 2015.

[39] Simon Jackman. pscl: Classes and methods for r. developed in the political science computational laboratory, stanford university. department of political science, stanford university, stanford, ca. r package version 1.03. 5. *http://www. pscl. stanford. edu/*, 2010.

[40] RCV Casarin, A Barbagallo, T Meulman, VR Santos, EA Sallum, FH Nociti, PM Duarte, MZ Casati, and RB Gonçalves. Subgingival biodiversity in subjects with uncontrolled type-2 diabetes and chronic periodontitis. *Journal of periodontal research*, 48(1):30–36, 2013.

[41] Ninh T Nguyen, Xuan-Mai T Nguyen, John Lane, and Ping Wang. Relationship between obesity and diabetes in a us adult population: findings from the national health and nutrition examination survey, 1999–2006. *Obesity surgery*, 21(3):351–355, 2011.

[42] Ivana Semova, Juliana D Carten, Jesse Stombaugh, Lantz C Mackey, Rob Knight, Steven A Farber, and John F Rawls. Microbiota regulate intestinal absorption and metabolism of fatty acids in the zebrafish. *Cell host & microbe*, 12(3):277–288, 2012.

[43] Ruth E Ley, Fredrik Bäckhed, Peter Turnbaugh, Catherine A Lozupone, Robin D Knight, and Jeffrey I Gordon. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences*, 102(31):11070–11075, 2005.

[44] Marlene Remely, Simone Dworzak, Berit Hippe, Jutta Zwielehner, E Aumüller, Helmut Brath, and Alexander Haslberger. Abundance and diversity of microbiota in type 2 diabetes and obesity. *J Diabetes Metab*, 4(253):2, 2013.

[45] Nadja Larsen, Finn K Vogensen, Frans WJ Van Den Berg, Dennis Sandris Nielsen, Anne Sofie Andreasen, Bente K Pedersen, Waleed Abu Al-Soud, Søren J Sørensen, Lars H Hansen, and Mogens Jakobsen. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS one*, 5(2), 2010.

[46] Yaohua Yang, Qiuyin Cai, Wei Zheng, Mark Steinwandel, William J Blot, Xiao-Ou Shu, and Jirong Long. Oral microbiome and obesity in a large study of low-income and african-american populations. *Journal of oral microbiology*, 11(1):1650597, 2019.

[47] Emma Allen-Vercoe and Christian Jobin. Fusobacterium and enterobacteriaceae: important players for crc? *Immunology letters*, 162(2):54–61, 2014.

[48] Tingting Wang, Guoxiang Cai, Yunping Qiu, Na Fei, Menghui Zhang, Xiaoyan Pang, Wei Jia, Sanjun Cai, and Liping Zhao. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME journal*, 6(2):320–329, 2012.

[49] Na Wu, Xi Yang, Ruifen Zhang, Jun Li, Xue Xiao, Yongfei Hu, Yanfei Chen, Fengling Yang, Na Lu, Zhiyun Wang, et al. Dysbiosis signature of fecal microbiota in colorectal cancer patients. *Microbial ecology*, 66(2):462–470, 2013.

[50] Santosh Dulal and Temitope O Keku. Gut microbiome and colorectal adenomas. *Cancer journal (Sudbury, Mass.)*, 20(3):225, 2014.

[51] Geicho Nakatsu, Xiangchun Li, Haokui Zhou, Jianqiu Sheng, Sunny Hei Wong, William Ka Kai Wu, Siew Chien Ng, Ho Tsoi, Yujuan Dong, Ning Zhang, et al. Gut mucosal

microbiome across stages of colorectal carcinogenesis. *Nature communications*, 6:8727, 2015.

[52] Iradj Sobhani, Julien Tap, Françoise Roudot-Thoraval, Jean P Roperch, Sophie Letulle, Philippe Langella, Gerard Corthier, Jeanne Tran Van Nhieu, and Jean P Furet. Microbial dysbiosis in colorectal cancer (crc) patients. *PloS one*, 6(1), 2011.

[53] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[54] Inés Martínez, James M Lattimer, Kelcie L Hubach, Jennifer A Case, Junyi Yang, Casey G Weber, Julie A Louk, Devin J Rose, Gayaneh Kyureghian, Daniel A Peterson, et al. Gut microbiome composition is linked to whole grain-induced immunological improvements. *The ISME journal*, 7(2):269–280, 2013.

[55] Wei Vivian Li and Jingyi Jessica Li. A statistical simulator scdesign for rational scrna-seq experimental design. *Bioinformatics*, 35(14):i41–i50, 2019.

[56] Jamie Waese, Nicholas J Provart, and David S Guttman. Topo-phylogeny: Visualizing evolutionary relationships on a topographic landscape. *PloS one*, 12(5):e0175895, 2017.

[57] Jian Xiao, Li Chen, Stephen Johnson, Yue Yu, Xianyang Zhang, and Jun Chen. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Frontiers in microbiology*, 9:1391, 2018.

[58] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

[59] Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, et al. Accessible, curated metagenomic data through experimenthub. *Nature methods*, 14(11):1023, 2017.

Figures

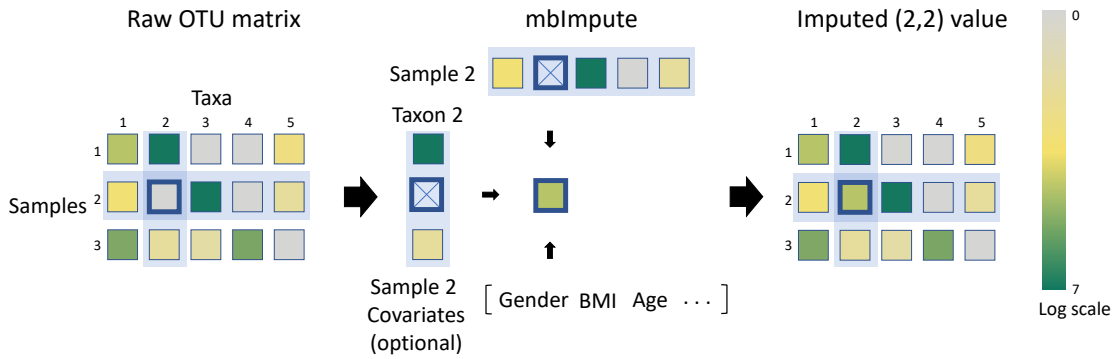

Figure 1: An illustration of mbImpute. After mbImpute identifies likely false zeros, it imputes them (e.g. the abundance of taxon 2 in sample 2) by jointly borrowing information from similar samples, similar taxa, and sample covariates if available (details in Methods).

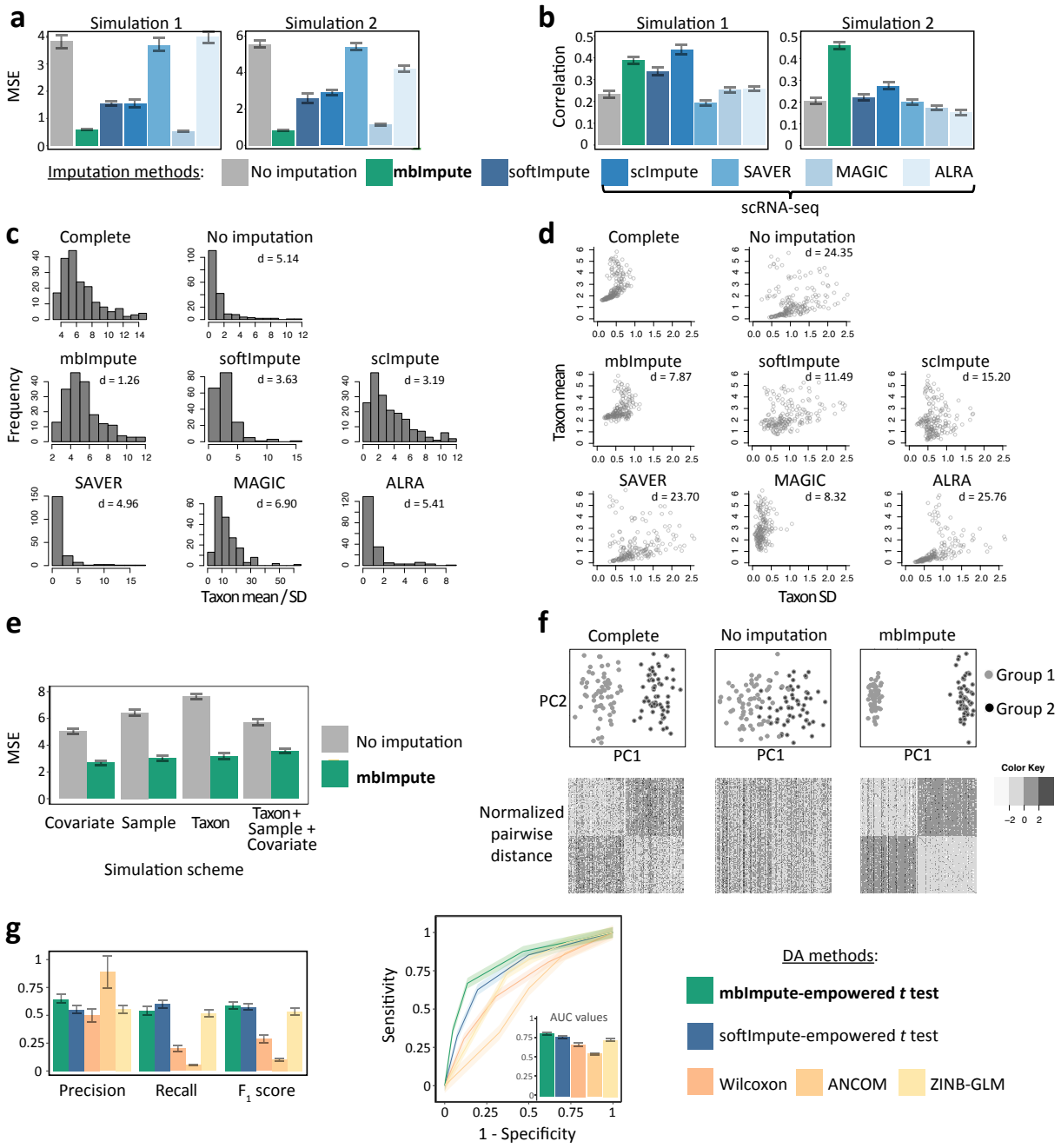

Figure 2: mbImpute outperforms state-of-the-art imputation methods designed for non-microbiome data and enhances the detection of DA taxa. **(a)** Mean squared error (MSE) and **(b)** mean Pearson correlation of taxon abundances between the complete data and the zero-inflated data ("No imputation," the baseline) or the imputed data by each imputation method (mbImpute, softImpute, scImpute, SAVER, MAGIC, and ALRA) in Simulations 1 and 2 (see Supplementary). **(c)-(d)** For each taxon, the mean and standard deviation (SD) of its abundances were calculated for the complete data, the zero-inflated data, and the imputed data by each imputation method in Simulation 1; (c) shows the distributions of the taxon mean / SD and the Wasserstein distance between every distribution and the complete distribution; (d) shows the taxa in two coordinates, mean vs. SD, and the Euclidean distance between the taxa in every (zero-inflated or imputed) dataset and the complete data in these two coordinates. **(e)** MSE between the complete data and the zero-inflated data or the imputed data by mbImpute under four simulation schemes in Simulation 3. **(f)** Data in Simulation 4. Top: the complete data, the zero-inflated data, and the imputed data by mbImpute in the first two principal components (PC1 and PC2). Bottom: normalized pairwise Euclidean distances between samples are shown for each dataset; each row of the distance matrix was normalized to have zero mean and unit standard deviation. **(g)** DA taxon detection results in Simulation 4. Left: precision, recall and F_1 scores of three existing DA methods (Wilcoxon rank-sum test, ANCOM, and ZINB-GLM), and two imputation-empowered t test (softImpute and mbImpute). Right: ROC curves and corresponding AUC values of the five methods.

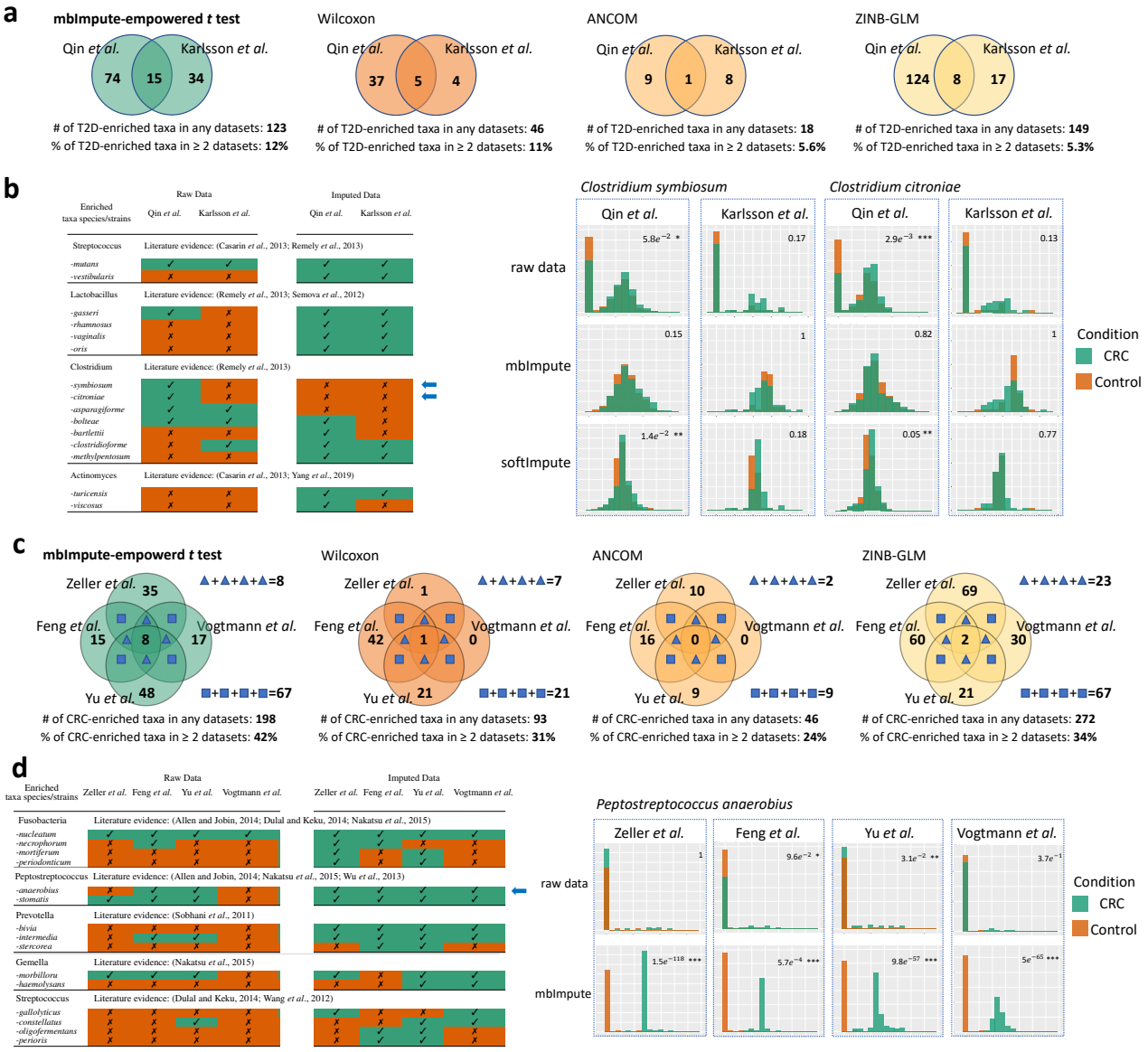

Figure 3: mbImpute increases the power and reproducibility of DA taxon identification in real datasets. Panels (a)-(b) and (c)-(d) are regarding T2D and CRC microbiome data, respectively. **(a)** Venn diagrams showing the overlap of T2D-enriched taxa in two datasets identified by four DA methods applied to raw data: the Wilcoxon test, ANCOM, and ZINB-GLM, and the mbImpute-empowered t test. **(b)** Example T2D-enriched taxa identified by the Wilcoxon test on the raw data vs. the t test on the imputed data by mbImpute. Literature evidence supporting the enrichment of these genera is listed. Check marks and crosses indicate the enriched and non-enriched taxa identified in each of the two datasets. Two strains in the Clostridium genus are marked by left arrows, and their abundance (log-transformed, see Methods) distributions are plotted for the raw data (top), the imputed data by mbImpute (middle), and the imputed data by softImpute (bottom). **(c)** Venn diagrams showing the overlap of CRC-enriched taxa in four real datasets identified by the same four methods as in (a). **(d)** Example CRC-enriched taxa identified by the Wilcoxon test on the raw data vs. the t-test on the imputed data by mbImpute. Literature evidence supporting the enrichment of these genera is listed. Check marks and crosses indicate the enriched and non-enriched taxa identified in each of the four datasets. *Peptostreptococcus anaerobius* is marked by a left arrow, and its log-transformed abundance distributions are plotted for the raw data (top) and the imputed data by mbImpute (bottom).