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 non-biological 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 non-biological 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], pulmonary disease [4, 5], and cancers [6]. These studies have shown the potential of using human microbiomes as biomarkers for disease diagnosis or therapeutic targets for disease treatment [7].

The development of high-throughput sequencing technologies has advanced microbiome studies in the last decade [8]. Microbiome studies primarily use two sequencing technologies: the 16S ribosomal RNA (rRNA) amplicon sequencing and the shotgun metagenomics sequencing. The former specifically sequences 16S rRNAs, which can be used to identify and distinguish microbes

[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

[9]. The 16S sequencing reads are either clustered into operational taxonomic units (OTUs) [10] or mapped to amplicon sequence variants (ASVs) for a higher resolution [11, 12]. The latter, often referred to as whole-genome sequencing (WGS), sequences all DNAs in a microbiome sample, including whole genomes of microbial species and host DNAs [10, 13–19], and its sequencing reads are mapped to known microbiome genome databases to quantify the abundance of each microbial species. Despite the vast differences between these two technologies, 16S and WGS data can both be processed into a similar data format about abundances of microbes in biological samples: a taxon count matrix with rows as samples (which often correspond to subjects) and columns as taxa (i.e., OTUs for 16S rRNA data and species for WGS data), and each entry corresponds to the number of reads mapped to a taxon in a sample. It is worth noting that the total read count per sample, i.e., the sum of entries in a row of the count matrix, differs by five orders of magnitude between the two technologies: $\sim 10^3$ per sample for 16S rRNA data and $\sim 10^8$ for WGS data [20].

A critical challenge in microbiome data analysis is the existence of excess zeros in taxon counts, a phenomenon prevalent in both 16S rRNA and WGS data [20]. The excess zeros belong to three categories by origin: biological, technical, and sampling zeros [21]. Biological zeros represent true zero abundances of non-existent taxa in samples. In contrast, both technical and sampling zeros are non-biological zeros with different origins: technical zeros arise from pre-sequencing experimental artifacts (e.g., DNA degradation during library preparation and inefficient sequence amplification due to factors such as GC content bias) [22], while sampling zeros are due to limited sequencing depths. Although WGS data have much larger per-sample total read counts than 16S data have, they still suffer from excess zeros because they sequence more nucleic acid sequences (microbial genomes instead of 16S rRNAs) and widespread host DNA contamination reduces the effective sequencing depths for microbial genomes [23–25].

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 to zeros is a common, simple approach [26, 27], but it is known to be ad-hoc and suboptimal as it cannot not distinguish biological zeros from technical and sampling zeros [28, 29]. Kaul et al. [30] developed an approach to distinguish these three types of zeros and only correct the sampling zeros; however, their correction is still a simple addition of a pseudo-count of one, ignoring the fact that the (unobserved) actual counts of these sampling zeros may not be exactly one.

In particular, this data sparsity issue has greatly hindered the differentially abundant (DA) taxon analysis, which is to identify the taxa that exhibit significantly different abundances between two groups of samples [13]. Microbiome researchers employ two major types of statistical methods to identify DA taxa. Methods of the first type are based on parametric models [7, 26, 31–38]. For example, the zero-inflated negative Binomial generalized linear model (ZINB-GLM) is used in [7, 31, 32], the DESeq2-phyloseq method uses the negative Binomial regression [33, 34], and the metagenomeSeq method uses the zero-inflated Gaussian model [35]. However, the different parametric model assumptions do not always fit data well [39]. Methods of the second type perform non-parametric statistical tests that do not assume specific distributions, and widely-used methods include the Wilcoxon rank-sum test [14–19] and ANCOM [27]. A major drawback of these non-parametric methods is that a taxon would be called DA if its zero proportions differ significantly between two groups of samples, but this difference is unlikely biologically meaningful due to the prevalence of technical and sampling zeros. Both types of methods consider taxon abundances at one of three scales: counts [7, 31, 32, 34], log-transformed counts [35], and proportions (i.e., each taxon's count is divided by the total of all taxa's counts in a sample) [26, 27, 36–38]. We note that excess zeros would negatively affect taxon abundances at all the three levels.

In addition to DA taxon analysis, other microbiome data analyses, such as the construction of taxon interaction networks [40–43], are also impeded by 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 [44]), image and speech reconstruction [45–47], imputation of unmeasured epigenomics datasets [48], missing genotype prediction in genome-wide association studies [49], and the more recent gene expression recovery in single-cell RNA-sequencing (scRNA-seq) data analysis [50–54]. Microbiome and scRNA-seq data have similar count matrix structures if one considers samples and taxa as analogs to cells and genes, and both data have excess non-biological 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. Although there are methods utilizing matrix completion in the microbiome field, their main purpose is to perform community detection or dimension reduction instead of imputation [55, 56]. Two distinct features of microbiome data make direct application of existing imputation methods suboptimal. First, microbiome data are often accompanied by metadata including sample covariates and taxon phylogeny, which, however, cannot be used by existing imputation methods. In particular, phylogenetic information is known to be valuable for microbiome data analysis [57–64], as taxa closely related in a phylogeny are likely to have similar functions and abundances in samples [65–68]. Second, microbiome data has a much smaller number of samples (often in hundreds) than the number of cells (often in tens of thousands) in scRNA-seq data, making those deep-learning based imputation methods inapplicable [54, 69]. On the other hand, 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, including both 16S and WGS data. mbImpute identifies and corrects zeros and low counts that are unlikely biological (for ease of terminology, we will refer to them as non-biological zeros in the following text) in microbiome taxon count data. The goal of mbImpute is to provide a principled data-driven approach to relieve the data sparsity issue due to excess non-biological zeros. To achieve this, mbImpute leverages three sources of information: a taxon count matrix, sample covariates (e.g., sample library size and subjects' age, gender, and body mass index), and taxon phylogeny, with the latter two sources optional. mbImpute takes a two-step approach (Fig. 1): it first identifies likely non-biological 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 (see 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 interaction networks. Microbiome researchers can use mbImpute to avoid the hassles of handling excess zeros in individual analysis tasks and to enjoy the flexibility of building up data analysis pipelines.

Results

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

As there are no imputation methods for microbiome data, we benchmarked mbImpute against five state-of-the-art imputation methods designed for non-microbiome data, including four popular scRNA-seq imputation methods (scImpute [50], SAVER [52], MAGIC [51], and ALRA [53]) and a widely-used general imputation method softImpute [70]. We designed two simulation studies, and the common goal was to obtain a "complete" microbiome dataset without non-biological zeros, so that imputation accuracy could be evaluated by comparing the imputed data with the complete data. The first study simulated complete data from a generative model that was fitted to a real WGS dataset of type 2 diabetes (T2D) samples [18], and the second, more realistic simulation study took a sub-dataset with fewer than 15% zeros as the complete data from another real WGS dataset of T2D samples [19]. In both simulation studies (see Supplementary), non-biological 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 studies, we compared these methods' 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 Wasserstein distance between the distribution of taxa's abundance means and standard deviations in the imputed data and that 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 non-biological zeros in microbiome data (see Methods). In the aforementioned second simulation study, we broke up similar samples in the real T2D WGS 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). Supplementary Fig. S3a shows that, after applied to the zero-inflated data, mbImpute effectively recovers non-biological zeros and reduces the MSE under every scheme. These results verify the robustness of mbImpute in selectively leveraging information useful for imputation.

We designed the fourth simulation to mimic a typical microbiome WGS study that aims to identify DA taxa between two sample groups. We simulated data for 300 taxa in 120 samples, 60 per group (see Supplementary). Supplementary Fig. S3b shows the two-dimensional visualization of the complete data (without non-biological zeros), zero-inflated data (with non-biological zeros), and imputed data (after mbImpute was applied to zero-inflated data). Compared with the zero-inflated data, the 120 samples 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, and ZINB-GLM. Among the available DA methods, the Wilcoxon rank-sum test is the most widely-used in microbiome studies [14–19], ANCOM is one of the most-cited microbiome-specific DA method [27], and ZINB-GLM was found as the most desirable count-model-based method in a comparative study [31]. 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 identification, because each taxon's logarithmic transformed counts (in the complete data) follow a Normal distribution in each sample 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 identification, 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 F$_1$ score of each method by comparing its detected DA taxa to the ground truth. Under the p-value threshold of 0.1 (Supplementary Fig. S3c 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 F_1 scores. To thoroughly compare the five methods, we plotted their performance at varying thresholds in receiver operating characteristic (ROC) curves (Supplementary Fig. S3c right), which show that mbImpute has the largest area under the curve (AUC) and outperforms the three DA methods and softImpute.

To further evaluate the performance of mbImpute on 16S rRNA sequencing data, we used a 16S simulator `sparseDOSSA` [71] to generate abundances of 150 taxa in 100 samples under two conditions (see Supplementary). Among these 150 taxa, 45 are predefined as truly DA taxa. We applied six existing DA methods, including the Wilcoxon rank-sum test, the two-sample t-test, ANCOM, ZINB/NB-GLM, DESeq2-phyloseq, and metagenomeSeq. (Note that ZINB-GLM is applied to the zero-inflated data, while NB-GLM is applied to the imputed data because the imputed data are no longer zero inflated.) To evaluate the accuracy of DA taxon identification, we calculated the precision, recall and F_1 score of each method, with or without using mbImpute as a preceding step, by comparing each method's detected DA taxa to the truly DA taxa. Under the p-value threshold of 0.1, the mbImpute-empowered DA methods consistently have better F_1 scores than those of the same DA methods without imputation. In particular, mbImpute improves both precision and recall rates of four DA methods: the t-test, ZINB/NB-GLM, DESeq2-phyloseq, and metagenomeSeq (Fig. 2e).

mbImpute improves the reproducibility and reliability of identifying T2D microbiome markers

To demonstrate that mbImpute can benefit the identification of DA taxa in real microbiome data, we applied the six DA methods to two T2D WGS datasets: Qin et al. and Karlsson et al., with or without using mbImpute as a preceding step. We observed that taxon abundance distributions are approximately Normal after imputation (Supplementary Fig. S4). We analyzed the identified T2D-enriched taxa in two aspects. First, we examined the overlap of these identified taxa by each method, with or without mbImpute, between the two datasets. Fig. 3a shows that mbImpute improves the reproducibility of all these DA methods, whose identified T2D-enriched taxa have increased overlaps after mbImpute is used (see Venn diagrams in Supplementary Fig. S5). Second, we investigated whether the T2D-enriched taxa identified in one dataset are reliable biomarkers for predicting T2D in another dataset. Towards this goal, we trained a random forest classifier [72] on one dataset with features as the T2D-enriched taxa identified from the other dataset. Then we calculated the 5-fold cross-validation accuracy, which reflects the reliability of the identified T2D-enriched taxa as biomarkers. Fig. 3b shows that mbImpute improves this reliability for all the DA methods but ANCOM, whose accuracy stays unchanged after mbImpute. The improvement is especially significant for the Wilcoxon rank-sum test, ZINB/NB-GLM, DESeq2-phyloseq, and metagenomeSeq. For example, the classification accuracy of the T2D-enriched taxa identified by DESeq2-phyloseq increases from 62% without mbImpute to 75% with mbImpute.

As a positive control, we also evaluated the classification accuracy when no DA methods are used but random forest automatically selects predictive features from all taxa. Encouragingly, we found that the accuracy becomes comparable to the positive control when ZINB/NB-GLM and DESeq2-phyloseq are used after mbImpute. Our results demonstrate that mbImpute improves the reproducibility of DA taxon identification between two T2D datasets, and that the identified DA taxa after mbImpute have better cross-dataset predictive power.

Further, We focused on four genera: *Streptococcus*, *Lactobacillus*, *Clostridium*, and *Actinomyces*, which have all been previously reported as enriched in T2D [73–79] (see Supplementary for the literature evidence). In these four genera, the mbImpute-empowered t-test discovers species-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. 4a. Moreover, we observed an interesting phenomenon: some Clostridium species taxa (Fig. 4a 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 can empower DA taxon identification as we observed in the fourth simulation. However, by examining the abundance distributions of such taxa, *Clostridium symbiosum* and *Clostridium citroniae* for example (Fig. 4a 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 non-biological zeros. See Supplementary for a discussion on statistical definitions of DA taxa.

We then compared mbImpute with softImpute using *Clostridium symbiosum* and *Clostridium citroniae* as examples. We observe that mbImpute retains well the distributions of non-zero abundances (Fig. 4a right panel middle row), while softImpute alters the distributions by introducing artificial spikes and shrinking the variance (Fig. 4a 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 preserves distributional characteristics of taxa's non-zero abundances and recovers downsampling zeros

In the T2D WGS data analysis, we have found that mbImpute can well maintain the distributions of taxa's non-zero abundances. To further verify the property of mbImpute in preserving characteristics of non-zero abundances, we examined pairwise taxon-taxon relationships in the two T2D WGS datasets: Qin et al. and Karlsson et al. For a pair of taxa, we estimated two Pearson correlations based on the raw data: one using all the samples ("raw all-sample correlation") and the other only using the samples where both taxa have non-zero abundances ("raw non-zero-sample correlation"). We also estimated a Pearson correlation based on the imputed data by mbImpute, using all the samples ("imputed all-sample correlation"). As shown in Fig. 5, there are vast differences between the raw all-sample correlations and the corresponding raw non-zero-sample correlations. However, the imputed all-sample correlations much resemble the corresponding raw non-zero-sample correlations, suggesting that mbImpute well preserves pairwise taxon-taxon correlations encoded in taxa's non-zero abundances.

We also explored the linear relationship of each taxon pair using the standard major axis (SMA) regression, which, unlike the least-squares regression, treats two taxa symmetric and considers randomness in both taxa's abundances. For a pair of taxa, we performed two SMA regressions on the raw data: one using all the samples ("raw all-sample regression") and the other using only the samples where both taxa have non-zero abundances ("raw non-zero-sample regression"). We also performed the SMA regression on the imputed data by mbImpute, using all the samples ("imputed all-sample regression"). Fig. 5 shows that the raw all-sample regressions and the raw non-zero-sample regressions return strongly different lines. Especially, the two lines between the two taxa *Eubacterium sirasum* and *Ruminococcus obeum* in the Karlsson et al. data (Fig. 5b bottom left) exhibit slopes of opposite signs. In contrast, the imputed all-sample regressions output lines with similar slopes to those of the raw non-zero-sample regressions. This result again confirms mbImpute's capacity to preserve characteristics of taxa's non-zero abundances in microbiome data.

Our results echo existing concerns about spurious taxon-taxon correlations estimated from microbiome data due to excess non-biological zeros [80, 81]. In other words, taxon-taxon correlations cannot be accurately estimated from raw data. Without imputation, an intuitive approach is to use taxa's non-zero abundances to estimate taxon-taxon correlations; however, this approach reduces the sample size for estimating each taxon pair's correlation, as the samples with zero abundances for either taxon would not be used, and it also makes different taxon pairs' correlation estimates rely on different samples. To address these issues, mbImpute provides another approach: its imputed data allow taxon-taxon correlations to be estimated from all the available samples. We have verified this mbImpute approach by the fact that the correlation estimates from the imputed data resemble those from the non-zero abundances in the raw data.

In addition, based on the T2D WGS dataset generated by Qin et al., we verified mbImpute's ca-

Removal rate	40%	70%
% of downsampling zeros indentified	95.83% ± 0.46%	92.83% ± 0.92%
Pearson correlation before imputation	0.7565 ± 0.0023	0.5261 ± 0.0016
Pearson correlation after imputation	0.8747 ± 0.0100	0.7582 ± 0.0235

Table 1: Effectiveness of mbImpute in indentifying zeros due to downsampling of Qin et al.'s T2D WGS dataset. Two downsampled datasets with removal rates 40% and 70% were constructed. The first row lists the percentages of downsampling zeros identified by mbImpute; the second row lists the Pearson correlations between each of the two downsampled matrices and the original matrix (on the log scale) before imputation; the third row lists the Pearson correlations (on the log scale) after mbImpute was used. For each number, we included its error margin as the 1.96 times of the corresponding standard error over 10 replications of downsampling.

pacity to identify non-biological zeros generated by downsampling. In each sample (i.e., each row in the sample-by-taxon count matrix), we assigned every taxon a sampling probability proportional to its count, i.e., the larger the count, the more likely the taxon is to be sampled; based on these probabilities, we sampled 60% or 30% of the non-zero taxon counts, and we set the unsampled counts to zeros (corresponding to a removal rate of 40% or 70%). After mbImpute is applied to the downsampled count matrices, we found that mbImpute correctly identifies 95.83% and 92.83% of the newly introduced non-biological zeros under the two downsampling schemes. Before imputation, the Pearson correlations between the two downsampled matrices and the original matrix (on the log scale) are 0.76 and 0.53. After applying mbImpute to all the three matrices, the correlations are increased to 0.87 and 0.76 (Table 1). This result confirms the effectiveness of mbImpute in recovering zeros due to downsampling.

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 [14, 15]. We applied the six DA methods to four CRC datasets: Zeller et al., Feng et al., Vogtmann et al. , and Yu et al., with or without using mbImpute as a preceding step. We also evaluated two aspects of DA taxon identification as we did in the aforementioned T2D analysis: the number of identified DA taxa identified in at least two datasets and the across-dataset classification accuracy when the identified DA taxa are used as features. Fig. 3c shows that mbImpute improves the reproducibility of all these DA methods, whose identified CRC-enriched taxa have increased overlaps after mbImpute is used (see diagrams in Supplementary Fig. S6). Then we investigated whether the CRC-enriched taxa identified in one dataset are reliable biomarkers for predicting CRC in another dataset. We used the same procedures as in the T2D analysis to obtain the classification results based on random forest. Fig. 3d shows that mbImpute improves the across-dataset classification accuracy for all the six DA methods. We again set a positive control by allowing random forest to automatically select predictive features from all taxa, and we found that the accuracy of all DA methods becomes comparable to or even surpasses the positive control after mbImpute is used.

We then focused on five genera: *Fusobacterium*, *Peptostreptococcus*, *Prevotella*, *Gemella*, and *Streptococcus*, which have been previously reported as enriched in CRC [82–87] (see Supplementary for the literature evidence). In these five genera, the mbImpute-empowered t-test discovers species-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. 4b. We use *Peptostreptococcus anaerobius* as an example to demonstrate the effectiveness of mbImpute in empowering DA taxon identification. This species 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 this taxon's abundance distributions (Fig. 4b 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 fails to identify it in Zeller et al.'s and Vogtmann et al.'s data because the dominance of zeros obscures the differences between the 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 non-biological zeros (Fig. 4b 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. 4b right panel).

mbImpute increases the similarity of microbial community structure between 16S rRNA and WGS data

We further show that mbImpute can enhance the similarity of taxon-taxon correlations inferred from micrbiome data measured by two technologies—16S rRNA sequencing and WGS. We used two microbiome datasets of healthy human stool samples: a 16S rRNA dataset from the Human Microbiome Project [88] and a WGS dataset from the control samples in Qin et al. We compared genus-level taxon-taxon correlations between these two datasets, and we did the comparison again after applying mbImpute. Fig. 6 shows that mbImpute increases the similarity between the taxon correlation structures in the two datasets. Before imputation, the Pearson correlation between the two correlation matrices (one computed from 16S rRNA taxon abundances and the other from WGS taxon abundances) is 0.59; mbImpute increases the correlation to 0.64. In particular, we observe three taxon groups (highlighted by magenta, green, and purple squares in Fig. 6) supported by both 16S rRNA and WGS data after imputation. Notably, in the magenta squares, *Acidaminococcus* has correlations with *Dialister* and *Blautia* only after imputation, and this result is consistent with the literature: *Acidaminococcus* and *Dialister* are both reported to have low abundances in healthy human stool samples [89]; *Acidaminococcus* and *Blautia* are both associated with risks of T2D and obesity, lipid profiles, and homeostatic model assessment of insulin resistance [90]. The green squares contain three bile-tolerant genera: *Alistipes*, *Bilophila*, and *Bacteroides* [91]. The raw 16S and WGS data only reveal the correlation between *Bacteroides* and *Alistipes*, but mbImpute recovers the correlations *Bilophila* has with *Alistipes* and *Bacteroides*. The purple squares indicate a strong correlation between *Sutterella* and *Prevotella* after imputation, yet this correlation is not observed in raw WGS data. We verified this correlation in the MACADAM database [92], which contains metabolic pathways associated with microbes. Out of 1260 pathways, *Sutterella* and *Prevotella* are associated with 154 and 278 pathways, respectively, and 122 pathways are in common; Fisher's exact test finds that the overlap is statistically significant (p-value $< 2.2 \times 10^{-16}$), suggesting that *Sutterella* and *Prevotella* are indeed functionally related. Overall, our results indicate that mbImpute can facilitate meta-analysis of 16S and WGS data by alleviating the hurdle of excess non-biological zeros.

Discussion

A critical challenge in microbiome data analysis is statistical inference of taxon abundance from highly sparse and noisy data. Our proposed method, mbImpute, will address this challenge and facilitate analysis of both 16S and WGS data. mbImpute works by correcting non-biological zeros and retaining taxa's 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 sample 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 sample groups). Third, mbImpute improves the reproducibility of DA taxon identification across studies and the consistency of microbial community detection between 16S and WGS data. 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.

There has been a long-standing concern about sample contamination in microbiome sequencing data, e.g. contamination from DNA extraction kits and laboratory reagents [1, 3]. Existing studies have attempted to address this issue via calibrated sequencing operations [2, 3, 6] and computational methods [4,5]. We recommend researchers to perform contamination removal before applying mbImpute. Moreover, by its design, mbImpute is robust to certain types of sample contamination that result in outlier taxa and samples. For each outlier taxon, mbImpute would borrow little information from other taxa to impute this outlier taxon's abundances. Similarly, mbImpute is robust to the existence of outlier samples that do not resemble any other sample.

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 [93]. Imputation follows the indirect evidence principle, where the most critical issue is to decide what observations to borrow information from so as to improve data quality 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 non-biological 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.

Moreover, we observed that, similar to each taxon's non-zero abundances, the imputed abundances exhibit a bell-shaped distribution across samples on the logarithmic scale. This suggests that statistical methods utilizing Normal distributional assumptions become suitable and applicable to imputed taxon abundances. 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 interaction network, to which network analysis methods may be applied to find taxon modules and hub taxa [94]. As a preliminary exploration, we constructed Bayesian networks of taxa based on the two T2D datasets Qin et al. and Karlsson et al. after applying mbImpute. Interesting shifts are observed in taxon interactions from control samples to T2D samples (Supplementary Figs. S7–8). For example, two genera, *Ruminococcus* and *Eubacterium*, have interactive species in control samples but not in T2D samples. In future research, differential network analysis methods can be applied to find taxon communities whose interactions differ between two sample groups.

Methods

mbImpute methodology

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

Notations

We denote the sample-by-taxon taxa 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). (By default, mbImpute includes sample library size as a covariate.) 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.

Data pre-processing

mbImpute requires every taxon's counts across samples to be on the same scale before imputation. If this condition is unmet, normalization is needed. However, how to properly normalize microbiome data is challenging, and multiple normalization methods have been developed in recent years [29, 95, 96]. To give users the flexibility of choosing an appropriate normalization method, mbImpute allows users to directly input a normalized count matrix by specifying that the input matrix does not need normalization. Otherwise, mbImpute normalizes samples by library size.

Default normalization (optional) 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.

First, mbImpute filters out taxa that have too few non-zero counts to avoid imputing these taxa's zeros, which are likely biological. This filtering step is exactly the same as how Kaul et al. [30] 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.

Next, mbImpute applies the logarithmic transformation to the normalized counts so as to reduce the effects of extremely large counts [97]. 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 [50, 98]. 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 sample 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 (including sample library size as a covariate). 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. S9 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. S10, which displays the distribution of all the d_{ij}'s in the data from Zeller et al. [14], Feng et al. [15], Yu et al. [16], Vogtmann et al. [17], Qin et al. [19], and Karlsson et al. [18].

To summarize, mbImpute does not impute all zeros in the taxon 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 [99, 100], 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` [101] 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 taxon count matrix (a sample-by-taxon matrix):

```
complete(taxa_count_matrix, softImpute(taxa_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 = "taxa_count_matrix_trans.csv", Kcluster = 1, out_dir = "sim_imp")
```

where taxa_count_matrix_trans.csv is the input file containing the transposed taxon 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(taxa_count_matrix), ncores = 1, estimates.only = TRUE)
```

MAGIC

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

```
magic_op = magic.MAGIC()
magic_op.set_params(n_pca = 40)
magic_op.fit_transform(taxa_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 taxon count matrix:

```
normalized_mat = normalize_data(taxa_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. We further compared five existing DA methods: the Wilcoxon rank-sum test, ANCOM, ZINB/NB-GLM, metagenomSeq and DESeq2-phyloseq, which apply to taxon counts, with or without using mbImpute as a preceding step. 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 sample 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` [27]. 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(taxa_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 `taxa_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(taxa_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.

metagenomeSeq

We used two R packages, metagenomeSeq combined with phyloseq. Specifically, we used the following command to obtain the result:

```
mseq_obj <- phyloseq_to_metagenomeSeq(physeq2)
pd <- pData(mseq_obj)
mod <- model.matrix(∼ 1 + condition, data = pd)
ran_seq <- fitFeatureModel(mseq_obj, mod)
```

where physeq2 is an object created from a count matrix and metadata using the phyloseq package.

DESeq2-phyloseq

We used the DESeq2 package combined with phyloseq. Specifically, we used the following command to obtain the result of DESeq2:

```
Deseq2_obj <- phyloseq_to_deseq2(physeq2, ∼ condition)
results <- DESeq(Deseq2_obj, test="Wald", fitType="parametric")
```

where physeq2 is an object created from a count matrix and metadata using the phyloseq package.

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 sample group and then collected the sample 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 sample group and then collected the sample 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 sample 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. DESeq2-phyloseq uses the Benjamini-Hochberg procedure to control the FDR under 0.1. For metagenomeSeq, we thresholded its FDR adjusted p-values at 0.1.

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` [102]. 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 [18, 19]. 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 species level with phylogenetic information available in the `R` package `curatedMetagenomicData`. Qin et al. [19] 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 species level with phylogenetic information. From both datasets, we identified T2D-enriched taxa by comparing the T2D and CON groups.

Four datasets are regarding CRC [14–17]. Zeller et al. [14] and Feng et al. [15] studied CRC-related microbiomes in three conditions: CRC, small adenoma (ADE; diameter < 10 mm), and control (CON). Zeller et al. [14] 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 species level with phylogenetic information. Feng et al. [15] 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 species level in our analysis. Yu et al. [16] and Vogtmann et al. [17] studied CRC-related microbiomes in two conditions: CRC vs. CON. In detail, Yu et al. [16] 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 species level. Vogtmann et al. [17] included 104 samples 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 species 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] Jakob Stokholm, Martin J Blaser, Jonathan Thorsen, Morten A Rasmussen, Johannes Waage, Rebecca K Vinding, Ann-Marie M Schoos, Asja Kunøe, Nadia R Fink, Bo L Chawes, et al. Maturation of the gut microbiome and risk of asthma in childhood. Nature communications, 9(1):1–10, 2018.

[5] Alexa A Pragman, Hyeun Bum Kim, Cavan S Reilly, Christine Wendt, and Richard E Isaacson. The lung microbiome in moderate and severe chronic obstructive pulmonary disease. PloS one, 7(10):e47305, 2012.

[6] 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.

[7] 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.

[8] 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.

[9] M Luz Calle. Statistical analysis of metagenomics data. Genomics & informatics, 17(1), 2019.

[10] 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.

[11] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: high-resolution sample inference from illumina amplicon data. Nature methods, 13(7):581–583, 2016.

[12] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal, 11(12):2639–2643, 2017.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Matteo Calgaro, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo. Assessment of single cell rna-seq statistical methods on microbiome data. BioRxiv, 2020.

[21] Barak Brill, Amnon Amir, and Ruth Heller. Testing for differential abundance in compositional counts data, with application to microbiome studies. arXiv preprint arXiv:1904.08937, 2019.

[22] Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. BioRxiv, page 477794, 2020.

[23] Joana Pereira-Marques, Hout Anne, Rui Manuel Ferreira, Michiel Weber, Ines Pinto-Ribeiro, Leen-Jan van Doorn, Cornelis Willem Knetsch, and Ceu Figueiredo. Impact of host dna and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. Frontiers in microbiology, 10:1277, 2019.

[24] Microbiome Human. Project consortium 2012. Structure, function and diversity of the healthy human microbiome. Nature, 486:207–214.

[25] Jason Lloyd-Price, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A Brantley Hall, Arthur Brady, Heather H Creasy, Carrie McCracken, Michelle G Giglio, et al. Strains, functions and dynamics in the expanded human microbiome project. Nature, 550(7674):61–66, 2017.

[26] 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.

[27] 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.

[28] Matthew CB Tsilimigras and Anthony A Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of epidemiology, 26(5):330–335, 2016.

[29] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome, 5(1):27, 2017.

[30] 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.

[31] 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.

[32] 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.

[33] Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. PloS one, 8(4):e61217, 2013.

[34] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology, 15(12):550, 2014.

[35] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. Nature methods, 10(12):1200–1202, 2013.

[36] Xiaoling Peng, Gang Li, and Zhenqiu Liu. Zero-inflated beta regression for differential abundance analysis with metagenomics data. Journal of Computational Biology, 23(2): 102–110, 2016.

[37] Timothy W Randolph, Sen Zhao, Wade Copeland, Meredith Hullar, and Ali Shojaie. Kernel-penalized regression for analysis of microbiome data. The annals of applied statistics, 12 (1):540, 2018.

[38] Zhigang Li, Katherine Lee, Margaret R Karagas, Juliette C Madan, Anne G Hoen, A James O'malley, and Hongzhe Li. Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. Statistics in biosciences, 10 (3):587–608, 2018.

[39] Stijn Hawinkel, Federico Mattiello, Luc Bijnens, and Olivier Thas. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Briefings in bioinformatics, 20(1):210–221, 2019.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

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

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] Cameron Martino, James T Morton, Clarisse A Marotz, Luke R Thompson, Anupriya Tripathi, Rob Knight, and Karsten Zengler. A novel sparse compositional technique reveals microbial perturbations. MSystems, 4(1), 2019.

[56] Yun Cai, Hong Gu, and Toby Kenney. Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. Microbiome, 5(1):110, 2017.

[57] László Zsolt Garamszegi. Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice. Springer, 2014.

[58] Liam J Revell. phytools: an r package for phylogenetic comparative biology (and other things). Methods in ecology and evolution, 3(2):217–223, 2012.

[59] Steven W Kembel, Peter D Cowan, Matthew R Helmus, William K Cornwell, Helene Morlon, David D Ackerly, Simon P Blomberg, and Campbell O Webb. Picante: R tools for integrating phylogenies and ecology. Bioinformatics, 26(11):1463–1464, 2010.

[60] David Orme, R Freckleton, G Thomas, T Petzoldt, S Fritz, N Isaac, et al. The caper package: comparative analysis of phylogenetics and evolution in r. R package version, 5(2):1–36, 2013.

[61] Gregory B Gloor and Gregor Reid. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Canadian journal of microbiology, 62(8): 692–703, 2016.

[62] Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics, 14(2):244–258, 2013.

[63] Tao Wang and Hongyu Zhao. Constructing predictive microbial signatures at multiple taxonomic levels. Journal of the American Statistical Association, 112(519):1022–1031, 2017.

[64] Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. Bioinformatics, 33(18):2873–2881, 2017.

[65] Alex D Washburne, James T Morton, Jon Sanders, Daniel McDonald, Qiyun Zhu, Angela M Oliverio, and Rob Knight. Methods for phylogenetic analysis of microbiome data. Nature microbiology, 3(6):652–661, 2018.

[66] T Michael Anderson, Marc-André Lachance, and William T Starmer. The relationship of phylogeny to community structure: the cactus yeast community. The American Naturalist, 164(6):709–721, 2004.

[67] Campbell O Webb, David D Ackerly, Mark A McPeek, and Michael J Donoghue. Phylogenies and community ecology. Annual review of ecology and systematics, 33(1):475–505, 2002.

[68] Evan Weiher and Paul A Keddy. Assembly rules, null models, and trait dispersion: new questions from old patterns. Oikos, pages 159–164, 1995.

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

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

[71] B Ren, E Schwager, TL Tickle, and C Huttenhower. Sparsedossa: Sparse data observations for simulating synthetic abundance. 2016.

[72] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. R news, 2(3):18–22, 2002.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] Eric Z Chen and Hongzhe Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics, 32(17):2611–2617, 2016.

[81] Mehdi Layeghifard, David M Hwang, and David S Guttman. Disentangling interactions in the microbiome: a network perspective. Trends in microbiology, 25(3):217–228, 2017.

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

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project. Nature, 449(7164):804–810, 2007.

[89] Kameron Y Sugino, Nigel Paneth, and Sarah S Comstock. Michigan cohorts to determine associations of maternal pre-pregnancy body mass index with pregnancy and infant gastrointestinal microbial communities: late pregnancy and early infancy. PloS one, 14 (3):e0213733, 2019.

[90] Qian Yang, Shi Lin Lin, Man Ki Kwok, Gabriel M Leung, and C Mary Schooling. The roles of 27 genera of human gut microbiota in ischemic heart disease, type 2 diabetes mellitus, and their risk factors: a mendelian randomization study. American Journal of Epidemiology, 187 (9):1916–1922, 2018.

[91] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature, 505(7484): 559–563, 2014.

[92] Malo Le Boulch, Patrice Déhais, Sylvie Combes, and Géraldine Pascal. The macadam database: a metabolic pathways database for microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups. Database, 2019, 2019.

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

[94] R Poudel, A Jumpponen, Dan C Schlatter, TC Paulitz, BB McSpadden Gardener, Linda L Kinkel, and KA Garrett. Microbiome networks: a systems framework for identifying candidate microbial assemblages for disease management. Phytopathology, 106(10):1083–1096, 2016.

[95] Li Chen, James Reeve, Lujun Zhang, Shengbing Huang, Xuefeng Wang, and Jun Chen. Gmpr: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ, 6:e4600, 2018.

[96] Ohad Manor and Elhanan Borenstein. Musicc: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome biology, 16(1):53, 2015.

[97] 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.

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

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

[100] 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.

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

[102] 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

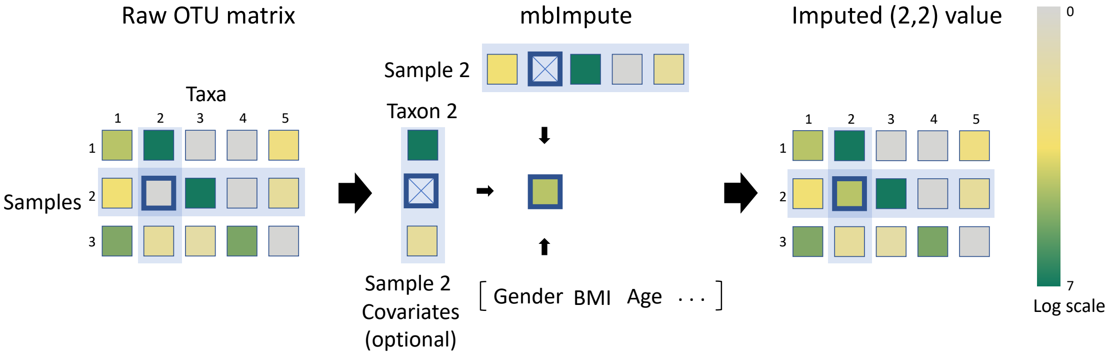


Figure 1: An illustration of mbImpute. After mbImpute identifies likely non-biological zeros, it imputes them (e.g. the abundance of taxon 2 in sample 2) by jointly borrowing information from similar samples, similar taxa, and sample covariates if available (details in Methods).

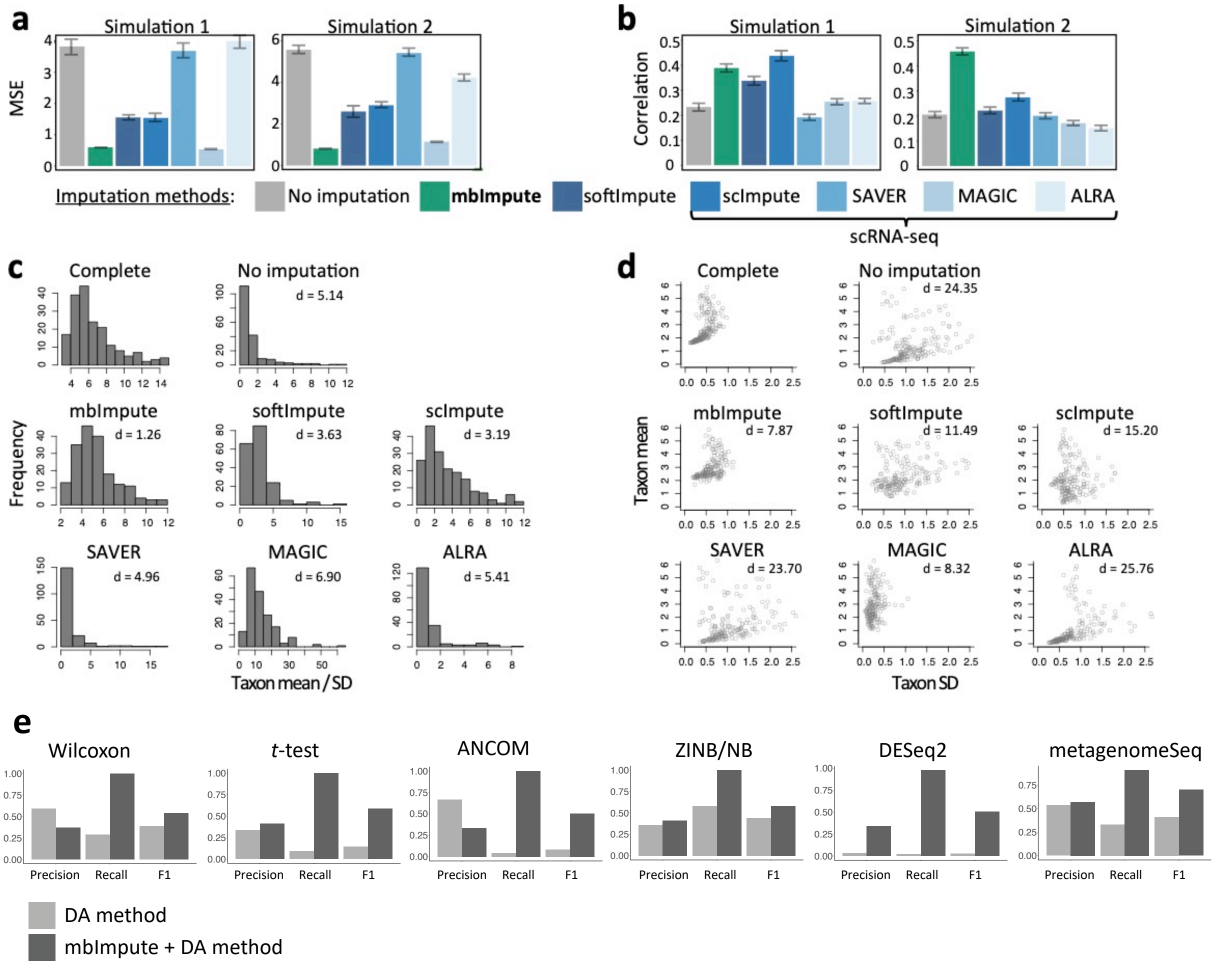


Figure 2: mbImpute outperforms state-of-the-art imputation methods designed for non-microbiome data and enhances the identification of DA taxa. (a) Mean squared error (MSE) and **(b)** mean Pearson correlation of taxon abundances between the complete data and the zero-inflated data ("No imputation," 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)** Accuracy (Precision, recall and F_1 scores) of six DA methods (Wilcoxon rank-sum test, t-test, ANCOM, ZINB/NB-GLM, DESeq2-phyloseq, metagenomeSeq) on raw data (light color) and imputed data by mbImpute (dark color) in Simulation 4.

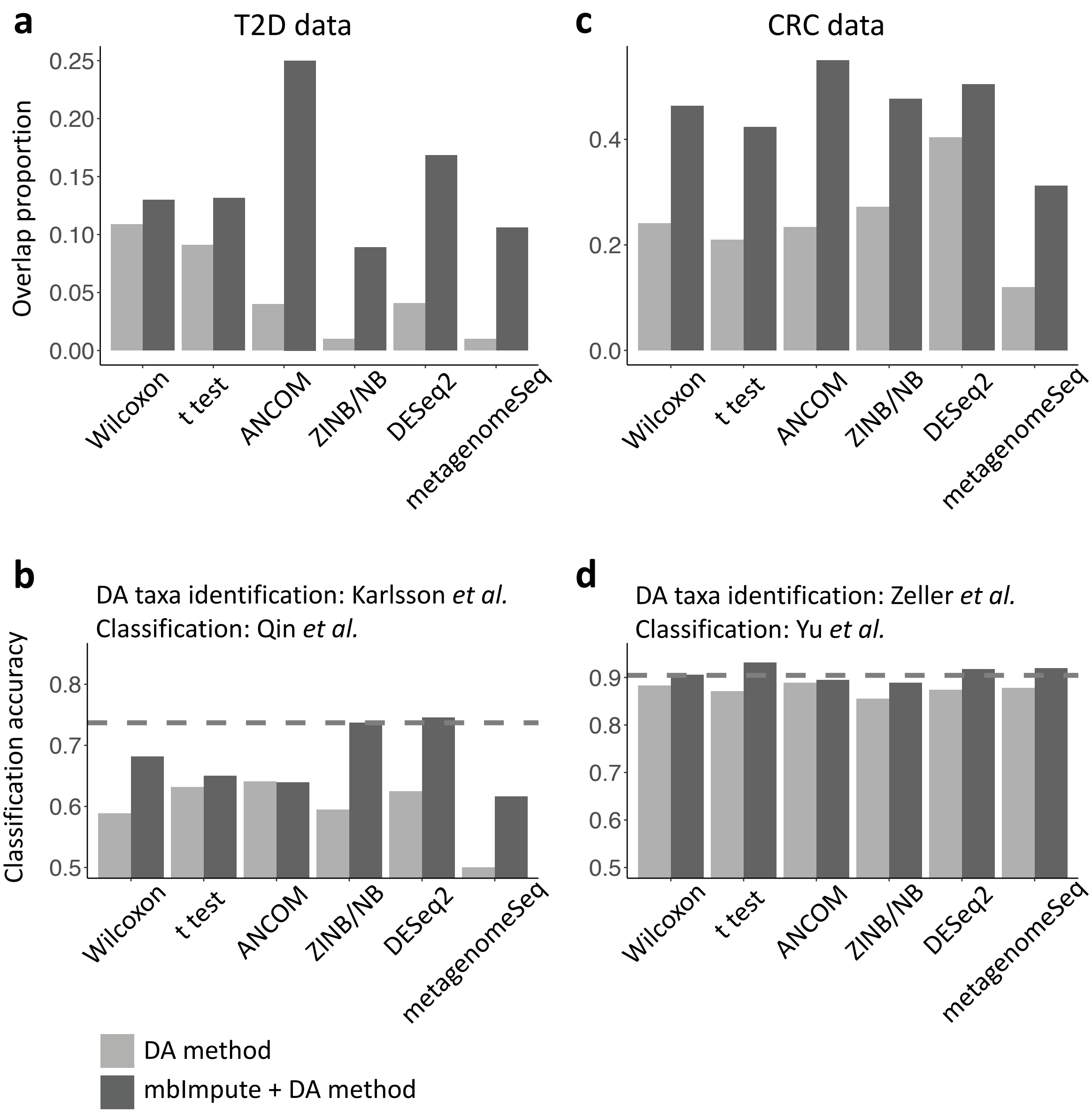
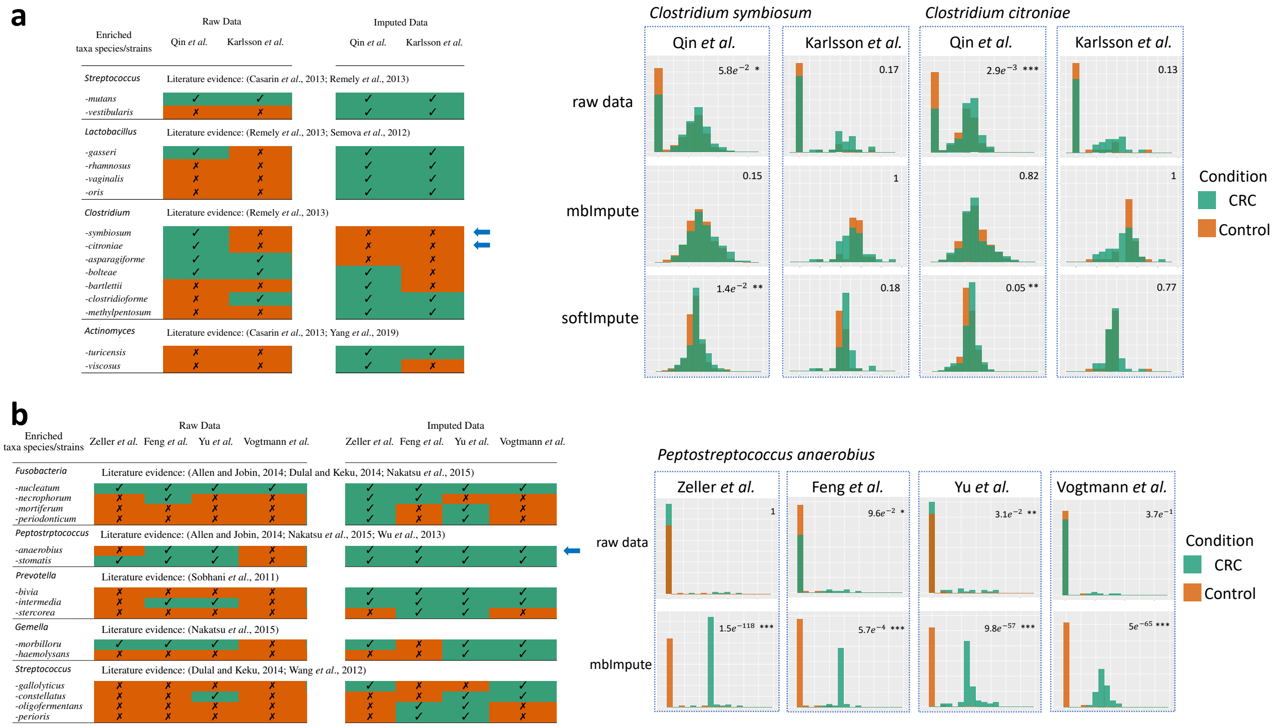


Figure 3: mbImpute increases the reproducibility of DA taxon identification and the accuracy of sample classification in cross-data studies. **(a)** The overlapping proportion (taxa identified as DA in both of the datasets / total number of taxa identified in either of the datasets) of identified T2D-enriched taxa between two T2D datasets [18, 19] for six DA methods, Wilcoxon rank-sum test (Wilcoxon), t-test, ANCOM, ZINB/NB-GLM (ZINB/NB), DESeq2-phyloseq (DESeq2), metagenomeSeq, before (light color) and after imputation (dark color). **(b)** The proportion of CRC-enriched taxa identified in at least two datasets among four CRC data [14–17] by the six DA methods before (light color) and after imputation (dark color). **(c)** The barplots show classification accuracy of prediction using random forest algorithm on the T2D status of Qin et al. by using the identified DA taxa in Karlsson et al. using six DA methods before (light color) and after imputation (dark color). The dotted horizontal line shows the prediction accuracy using random forest that automatically selects predictive features from all the taxa in Qin et al. to predict T2D statues. **(d)** The barplots show classification accuracy of prediction using random forest on the T2D status of Yu et al. by using the identified DA taxa in Zeller et al. using six DA methods before (light color) and after imputation (dark color). The dotted horizontal line shows the prediction accuracy using random forest that automatically selects predictive features from all the taxa in Yu et al. to predict CRC statues.



a

Enriched taxa species/strains	Raw Data		Imputed Data		
	Qin *et al.*	Karlsson *et al.*	Qin *et al.*	Karlsson *et al.*	
Streptococcus	Literature evidence: (Casarin *et al*., 2013; Remely *et al*., 2013)				
-mutans	✓	✓	✓	✓	
-vestibularis	✗	✗	✓	✓	
Lactobacillus	Literature evidence: (Remely *et al*., 2013; Semova *et al*., 2012)				
-gasseri	✓	✗	✓	✓	
-rhamnosus	✗	✗	✓	✓	
-vaginalis	✗	✗	✓	✓	
-oris	✗	✗	✓	✓	
Clostridium	Literature evidence: (Remely *et al*., 2013)				
-symbiosum	✓	✗	✗	✗	⬅
-citroniae	✓	✗	✗	✗	⬅
-asparagiforme	✓	✓	✗	✗	
-bolteae	✓	✓	✓	✗	
-bartlettii	✗	✗	✓	✗	
-clostridioforme	✗	✓	✓	✓	
-methylpentosum	✗	✗	✓	✓	
Actinomyces	Literature evidence: (Casarin *et al*., 2013; Yang *et al*., 2019)				
-turicensis	✗	✗	✓	✓	
-viscosus	✗	✗	✓	✗	

b

Enriched taxa species/strains	Raw Data				Imputed Data				
	Zeller *et al.*	Feng *et al.*	Yu *et al.*	Vogtmann *et al.*	Zeller *et al.*	Feng *et al.*	Yu *et al.*	Vogtmann *et al.*	
Fusobacteria	Literature evidence: (Allen and Jobin, 2014; Dulal and Keku, 2014; Nakatsu *et al*., 2015)								
-nucleatum	✓	✓	✓	✓	✓	✓	✓	✓	
-necrophorum	✗	✓	✗	✗	✓	✓	✗	✗	
-mortiferum	✗	✗	✗	✗	✓	✗	✓	✗	
-periodonticum	✗	✗	✗	✗	✓	✗	✓	✗	
Peptostrptococcus	Literature evidence: (Allen and Jobin, 2014; Nakatsu *et al*., 2015; Wu *et al*., 2013)								
-anaerobius	✗	✓	✓	✗	✓	✓	✓	✓	⬅
-stomatis	✓	✓	✓	✗	✓	✓	✓	✓	
Prevotella	Literature evidence: (Sobhani *et al*., 2011)								
-bivia	✗	✗	✗	✗	✓	✓	✓	✓	
-intermedia	✗	✓	✓	✗	✓	✓	✓	✓	
-stercorea	✗	✗	✗	✗	✗	✓	✓	✗	
Gemella	Literature evidence: (Nakatsu *et al*., 2015)								
-morbilloru	✓	✓	✓	✗	✓	✗	✓	✓	
-haemolysans	✗	✗	✗	✗	✗	✗	✓	✓	
Streptococcus	Literature evidence: (Dulal and Keku, 2014; Wang *et al*., 2012)								
-gallolyticus	✗	✗	✗	✗	✓	✗	✗	✓	
-constellatus	✗	✗	✓	✗	✗	✗	✓	✓	
-oligofermentans	✗	✗	✗	✗	✗	✓	✓	✗	
-perioris	✗	✗	✗	✗	✗	✓	✓	✗	

Figure 4: mbImpute increases the power and reproducibility of DA taxon identification in T2D and CRC WGS datasets. (a) 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 speciess 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). **(b)** 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).

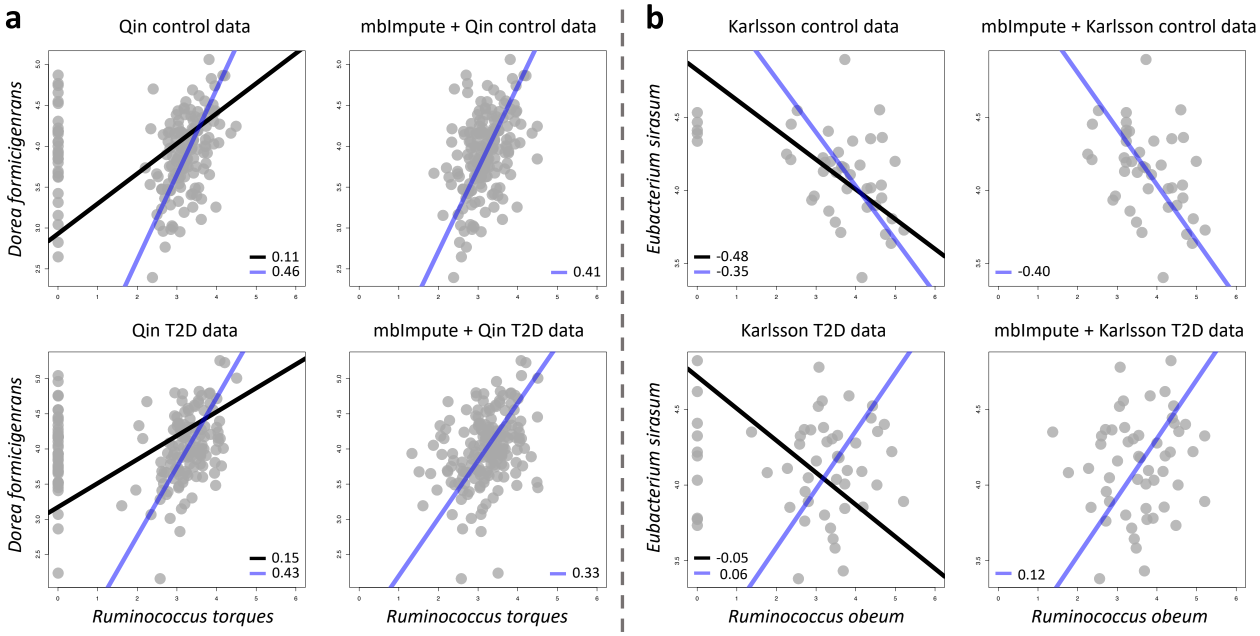


Figure 5: mbImpute preserves distributional characteristics of taxa's non-zero abundances. (a) Top: two scatter plots show the relationship between the abundances of *Dorea formicigenerans* and *Ruminococcus torques* in Qin et al.'s control samples, with or without using mbImpute as a preceding step. The left plot shows two standard major axis (SMA) regression lines and two corresponding Pearson correlations based on the raw data (balck: based on all the samples; blue: based on only the samples where both taxa have non-zero abundances). The right plot shows the SMA regression line (blue) and the Pearson correlation using all the samples in the imputed data. Bottom: two scatter plots for the same two taxa in Qin et al.'s T2D samples, with lines and legends defined the same as in the Top panel. **(b)** Four scatter plots show the SMA regression lines and correlations between *Eubacterium sirasum* and *Ruminococcus obeum* in Karlsson et al.'s control and T2D samples, with lines and legends defined the same as in (a).

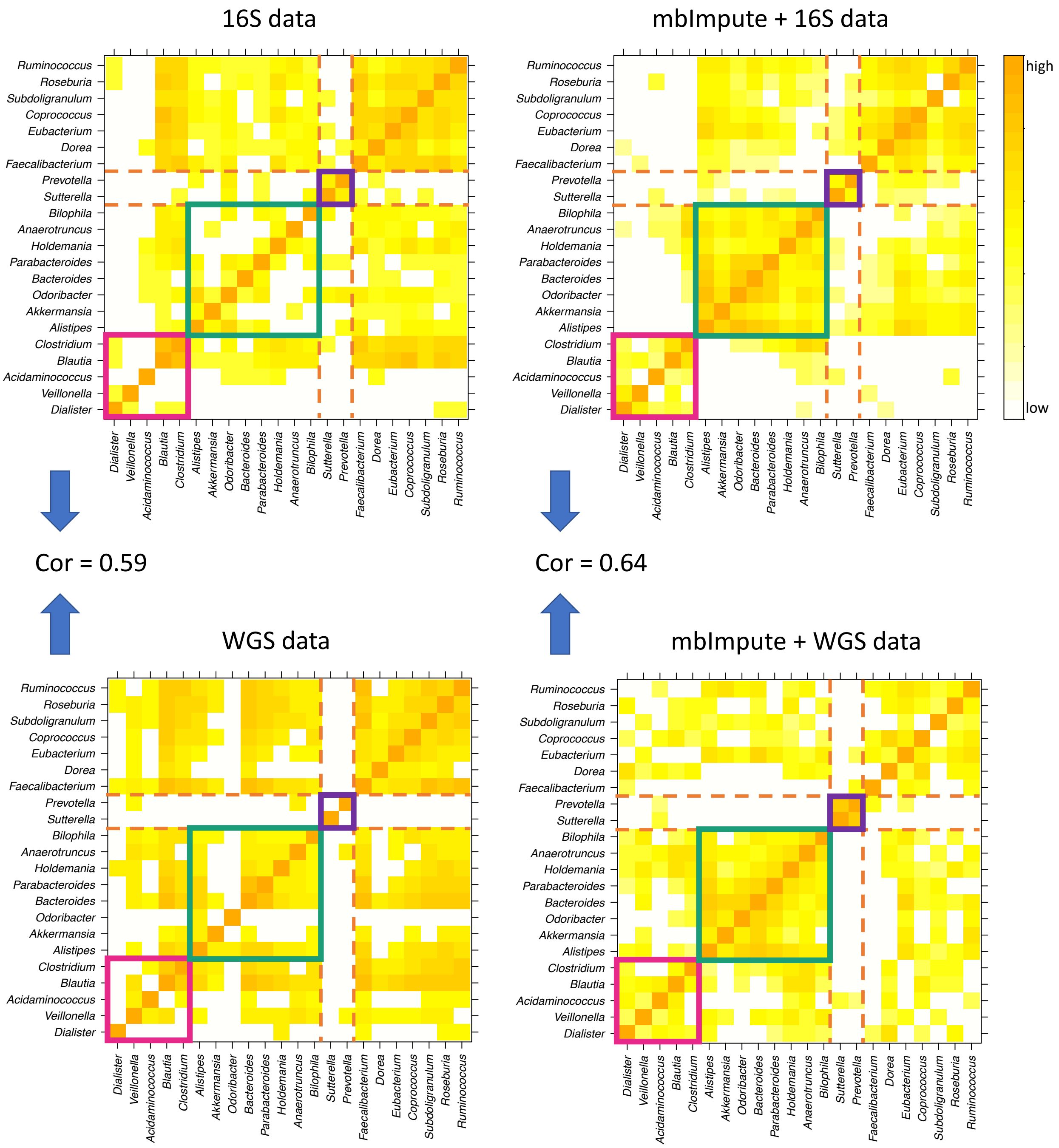


Figure 6: mbImpute improves the consistency in estimating taxon-taxon correlations between 16S and WGS data of microbiome composition in the healthy human stool samples. Four Pearson correlation matrices are calculated based on genus-level taxa's abundances in 16S and WGS data, with or without using mbImpute as a preceding step. Before imputation, the Pearson correlation between the two correlation matrices is 0.59, and this correlation increases to 0.64 after imputation. For illustration purposes, each heatmap shows square roots of Pearson correlations, with the bottom 40% of values truncated to 0. The magenta, green, and purple squares highlight three taxon groups, each of which contains strongly correlated taxa and is consistent between the 16S and WGS data after imputation.