


# PhyImpute and UniFracImpute: two imputation approaches incorporating phylogeny information for microbial count data

Qianwen Luo<sup>1</sup>, Shanshan Zhang<sup>2</sup>, Hamza Butt<sup>3</sup>, Yin Chen<sup>4</sup>, Hongmei Jiang<sup>5</sup>, Lingling An <sup>1,2,3,\*</sup>

<sup>1</sup>Department of Biosystems Engineering, University of Arizona, Tucson, AZ 85721, United States

<sup>2</sup>Interdisciplinary Program in Statistics and Data Science, University of Arizona, Tucson, AZ 85721, United States

<sup>3</sup>Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ 85721, United States

<sup>4</sup>Department of Pharmacology and Toxicology, School of Pharmacy, University of Arizona, Tucson, AZ 85721, United States

<sup>5</sup>Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208, United States

\*Corresponding author. E-mail: [anling@arizona.edu](mailto:anling@arizona.edu)

## Abstract

Sequencing-based microbial count data analysis is a challenging task due to the presence of numerous non-biological zeros, which can impede downstream analysis. To tackle this issue, we introduce two novel approaches, PhyImpute and UniFracImpute, which leverage similar microbial samples to identify and impute non-biological zeros in microbial count data. Our proposed methods utilize the probability of non-biological zeros and phylogenetic trees to estimate sample-to-sample similarity, thus addressing this challenge. To evaluate the performance of our proposed methods, we conduct experiments using both simulated and real microbial data. The results demonstrate that PhyImpute and UniFracImpute outperform existing methods in recovering the zeros and empowering downstream analyses such as differential abundance analysis, and disease status classification.

**Keywords:** imputation; phylogenetic tree; microbiome; metagenomics

## Introduction

Microbiome studies investigate the genomes and dynamic interactions of microorganisms inhabiting environments such as the human gut, soil, and ocean to gain a deeper understanding of their crucial ecological roles [1]. Advances in next-generation sequencing technologies have allowed an accurate quantification of microbial communities through feature abundance tables [1, 2]. However, analyzing these tables can be challenging due to their sparse nature, characterized by a substantial proportion of zeros [2–5].

The zeros in microbiome data can be divided into three categories by origin: biological, sampling, and technical zeros [5]. Biological zeros indicate the absence of a microbial feature in the system. Conversely, sampling and technical zeros are non-biological in nature. Sampling zeros occur when sequencing reads are limited, leading to the undercounting of low abundance features in skewed microbial compositions [5]. Technical zeros result from technical bias, typically introduced during sample preparation in sequencing experiments, such as inefficient steps in DNA extraction and PCR amplification [4]. The high proportion of zeros in microbiome data, also known as sparsity, poses challenges for modeling and can impact the performance in pipelines such as differential abundance (DA) and network analysis [3]. The proportion of zeros in sequencing-based microbiome data can be as high as 89% [6]. High sparsity can violate the assumptions

for statistical tests and regression analysis methods, leading to false associations or reduced power when traditional statistical methods are directly applied [2].

Sparsity is a common issue in both microbiome and single-cell RNA-seq (scRNA-seq) data. For scRNA-seq data, many imputation methods have been successfully developed to deal with this high sparsity, for example, scDoc [7], scImpute [8], SAVER [9], MAGIC [10], and softImpute [11]. The imputation methods for scRNA-seq data can be categorized into three broad and often overlapping approaches [12, 13]: model-based imputation methods, data-smoothing methods, and data-reconstruction methods. Model-based methods use probabilistic models to identify technical zeros and usually impute the technical zeros only while leaving other observed expression levels unchanged. For data-smoothing methods, similarity between cells will be measured first based on gene expression profiles or the relationship in a graph. Then, the gene expression values for each cell will be adjusted by smoothing or diffusing the gene expression values in similar cells. The third category, data-reconstruction methods, first defines a latent space representation of the cells using either low-rank matrix factorization, which captures the linear relationship, or deep-learning methods, which capture the non-linear relationships. Then, the data matrix will be reconstructed from a low-rank or simplified representation. Hou et al. (2020) systemically evaluated 18 scRNA-seq imputation methods, which cover three categories of the methods, to assess their accuracy

Received: May 2, 2024. Revised: November 16, 2024. Accepted: December 5, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

and usability [13]. They found that MAGIC, kNN-smoothing, and SAVER outperform the other methods most consistently. Dai et al. (2022) compared 12 imputation methods belonging to the two categories of model-based and deep learning-based methods using both simulated and real scRNA-seq datasets [14]. They concluded that each method has its own advantages and disadvantages, and there is no outstanding method. Cheng et al. (2023) evaluated 11 imputation methods using real scRNA-seq data and recommended SAVER and NE [15] for downstream analyses such as cell clustering and marker gene analysis [16].

Although both microbiome and scRNA-seq datasets share the characteristic of containing a large proportion of zeros, the key difference lies in the scale of the datasets. Microbiome datasets typically consist of a few hundred samples, whereas scRNA-seq datasets often contain tens of thousands of cells. This disparity makes methods based on deep learning and matrix factorization, which are generally designed for scRNA-seq and require large numbers of cells to perform optimally, less suitable for microbiome count data. Furthermore, microbiome data differ from scRNA-seq data in terms of content. The microbiome count table represents the relative abundance of microbial features (e.g. species or Operational Taxonomic Units) in various samples, while scRNA-seq data are concerned with the gene expression in different cells. Moreover, microbiome data also have other information, such as the phylogenetic relationship among microbial features, which can be used to empower the analysis. Therefore, directly applying scRNA-seq imputation methods to microbiome data may be impractical or result in a loss of power, highlighting the need for a microbiome-specific imputation method.

A recently proposed imputation method tailored specifically for microbiome data, known as mbImpute [17], was developed based on the scRNA-seq imputation method scImpute [8]. mbImpute first identifies the type of zeros and determines which taxa require imputation using a likelihood ratio test. It then utilizes information from similar samples, similar taxa, optional sample metadata from sample covariates, and phylogenetic information in the imputation [8, 17], using the same model as that in scImpute [8]. Compared with other scRNA-seq imputation methods, mbImpute has better performance in recovering missing taxon abundances for microbiome data and empowering DA taxon identification [17].

A limitation of mbImpute is that it defines phylogenetic distance as the number of branches connecting two taxa, ignoring the branch lengths between them [17]. Microbial taxa with similar biological functions are usually evolutionarily related, making phylogenetic trees an essential tool for investigating microbial composition's association with biological/environmental factors [18–20]. A variety of distance metrics using branch length information have been proposed, such as the UniFrac distance matrix [21]. UniFrac calculates the dissimilarity in microbial community composition between samples by mapping their phylogenetic composition onto a phylogenetic tree and comparing them based on branch length. However, UniFrac only considers the presence/absence of the microbial taxa and ignores the abundance information. The Weighted-UniFrac distance [22] is a variation that remedies this limitation by assigning weights to taxa proportionally to their abundances in the two samples being compared. Nonetheless, branch length noise can still affect these methods. To address this issue, Chen et al. proposed a generalized UniFrac distance that weights the branch length by both relative difference and its importance indicated by the branch proportion [23]. Therefore, incorporating branch length information from a

phylogenetic tree can potentially enhance the effectiveness of the microbiome-specific imputation method.

In this study, we propose two new imputation methods, PhyImpute and UniFracImpute, which utilize the Poisson-Negative Binomial (PNB) model and incorporate branch length information from a phylogenetic tree to identify and impute non-biological zeros by borrowing information from similar microbial samples. The Poisson distribution is relatively straightforward, with only one parameter determining its shape. It can be used to account for the probability of zero counts and low counts in microbiome data. However, microbiome data can exhibit overdispersion, where the variance exceeds the mean. In such instances, mixture models combining Poisson and Negative Binomial (PNB) are more appropriate for accurately representing microbiome count data. This mixture model has been successfully used in imputing single cell count data [10], where Poisson distribution captures the excess zeros, and Negative Binomial is used to represent the expression. Thus, the PNB model offers a balance between capturing overdispersion, excess zeros, and providing interpretable parameters. We assess the performance of PhyImpute and UniFracImpute using synthetic studies, 16S simulation studies, and real microbial studies. The proposed methods are compared with mbImpute, a microbiome-specific imputation method, and five other imputation methods designed for scRNA-seq data. These results suggest that our proposed methods provide promising approaches for imputing missing values in microbiome data.

## Materials and methods

PhyImpute and UniFracImpute are two proposed imputation methods for microbiome data. They aim to calculate the probability of non-biological zeros, incorporate branch lengths from the phylogenetic tree to calculate similarities between samples, and recover non-biological zeros by borrowing information from similar samples. The steps involved in these methods are explained in detail below:

Step 1: calculation of the sample-to-sample distance matrix.

Here, we propose two approaches, PhyImpute and UniFracImpute, to compute the similarity or distance between each pair of microbial samples based on the microbial abundance data. These methods incorporate phylogenetic tree information and account for the estimated probabilities of being non-biological zeros. The sample-to-sample distance matrix will be used to find similar microbial samples using the k-nearest neighbor approach in Step 2.

1-(1) PhyImpute: it is based on a modified cosine similarity between vectors in the context of a phylogenetic tree. The cosine similarity is a measure of similarity between two vectors. Given two vectors  $x$  and  $z$ , it measures the cosine of the angle between them:

$$\cos\theta = \frac{x \cdot z}{\|x\| \cdot \|z\|} = \frac{\sum_{k=1}^L x_k z_k}{\sqrt{\sum_{k=1}^L x_k^2} \sqrt{\sum_{k=1}^L z_k^2}} \quad (1)$$

where  $x_k$  and  $z_k$  represent the  $k$ th component of the vector  $x$  and  $z$ , respectively, and  $L$  is the length of the vectors. Here, the cosine similarity is used in positive space, and its value falls within the interval  $[0, 1]$ . When the value equals 1, two vectors are exactly the same. Conversely, when the value equals 0, two vectors are orthogonal or not correlated. Notably, the proposed method incorporates the non-biological zeros' probability and branch length from the

phylogenetic tree into the cosine similarity calculation:

$$s_{ij} = \frac{\sum_{g=1}^n w_g * b_g * y_{gi} * y_{gj}}{\sqrt{\sum_{g=1}^n w_g * b_g * y_{gi}^2} * \sqrt{\sum_{g=1}^n w_g * b_g * y_{gj}^2}} \quad (2)$$

where  $s_{ij}$  is the similarity between two microbial samples/communities  $i$  and  $j$ ,  $b_g$  is the branch length for feature/taxon  $g$  in the phylogenetic tree, and  $y_{gi}$  and  $y_{gj}$  are the taxa proportions descending from the branch  $g$  for community  $i$  and  $j$ , respectively. The weight  $w_g$  is determined based on the probabilities of non-biological zeros for the count  $y_{gi}$  and  $y_{gj}$ , which are estimated using a PNB mixture model [7]. That is, for each feature  $g$ , the probability of being non-biological zero in the community  $i$  is expressed as follows:

$$d_{gi} = \frac{\hat{\pi}_{gi} * f_{\text{Pois}}(y_{gi})}{\hat{\pi}_{gi} * f_{\text{Pois}}(y_{gi}) + (1 - \hat{\pi}_{gi}) * f_{\text{NB}}(y_{gi})} \quad (3)$$

where  $\pi_{gi}$  is the probability that an observed count belongs to a Poisson distribution,  $f_{\text{Pois}}$  and  $f_{\text{NB}}$  are probability density functions for Poisson and Negative Binomial distributions, respectively. All the parameters in the PNB model, including the parameters from the two distributions  $f_{\text{Pois}}$  and  $f_{\text{NB}}$  are estimated using the EM algorithm.

Then in the phylogenic-based cosine similarity, we define the weight  $w_g = 1$  when both  $d_{gi}$  and  $d_{gj}$  obtained from the PNB models are greater than 0.5 or both  $d_{gi}$  and  $d_{gj}$  are less than 0.5; otherwise,  $w_g$  = average of the two probabilities. This allows for a more nuanced and accurate measure of similarity that takes into account the relationships and structure between the two vectors within the context of the phylogenetic tree.

1-(2) UniFracImpute: the distance between two microbial samples/communities is calculated based on the weighted UniFrac distance. UniFrac (qualitative) and weighted UniFrac (quantitative) are widely used phylogenetic distance metrics for comparing biological communities [17, 18]. These metrics calculate the distance between pairs of samples, where all features found in both samples are placed on a phylogenetic tree. UniFrac measures the percentage of observed branch length unique to either sample [17]. When two communities are the same (i.e. identical), the UniFrac distance between them (i.e. the unique fraction of the tree) is 0. In contrast, for two communities that do not share any common microbes, the whole evolutionary tree of one sample is different from the other. In this case, the UniFrac distance equals 1. Weighted UniFrac takes into account the relative abundances of the microbial taxa in the samples, providing a more accurate measure of the similarity between microbial communities [18]:

$$\text{weighted UniFrac} = \frac{\sum_{g=1}^n b_g |p_g^A - p_g^B|}{\sum_{g=1}^n b_g (p_g^A + p_g^B)} \quad (4)$$

where  $b_g$  is the length of the branch  $g$ , and  $p_g^A$  and  $p_g^B$  are the taxa proportions descending from the branch  $g$  for communities A and B, respectively. In the proposed work, we combine the non-biological zeros' probability into the similarity measures  $S_{n \times n} = [s_{ij}]$ ,

$$s_{ij} = 1 - \frac{\sum_{g=1}^n b_g * w_g * |y_{gi} - y_{gj}|}{\sum_{g=1}^n b_g * w_g * (y_{gi} + y_{gj})} \quad (5)$$

where  $i$  and  $j$  represent the indices of two microbial samples,  $b_g$  is the branch length of the taxon  $g$ ;  $y_{gi}$  and  $y_{gj}$  are the taxa proportions descending from the branch  $g$  for sample  $i$  and  $j$ , respectively. Define the weight  $w_g = 1$  when both  $d_{gi}$  and  $d_{gj}$

obtained by PNB models are greater than 0.5 or both  $d_{gi}$  and  $d_{gj}$  are less than 0.5; Otherwise,  $w_g$  is equal to the average of  $d_{gi}$  and  $d_{gj}$ . The modified sample-to-sample similarity is more accurate because it incorporates information from the abundance, branch length, and the probability of observing non-biological zeros.

Step 2: Imputation for non-biological zeros using the k-nearest neighbor approach.

Based on the values  $d_{gi}$  and  $d_{gj}$  derived from the PNB models, non-biological zeros are defined as instances where one probability is greater than 0.5, and the other is less than 0.5. Then to recover values for identified non-biological zeros, we use a k-nearest neighbor (KNN) algorithm to impute values by borrowing information from similar samples [10]. The KNN algorithm works by finding the  $k$  closest neighbors to the target samples  $i$ , and then imputing the non-biological zeros based on the  $k$  most similar samples whose indices are saved in a set  $V$ . The imputed value for taxon  $g$  in sample  $i$  can be calculated as follows:

$$\text{Imp}_{gi} = \frac{\sum_{k \in V} s_{ik} * y_{gk}}{\sum_{k \in V} s_{ik}} \quad (6)$$

where  $s_{ik}$  is the similarity between the target sample  $i$  and candidate sample  $k$  from the first step (either by PhyImpute or by UniFracImpute), and  $V$  is the set of  $K$  nearest samples. In microbiome data studies where the sample size is typically small, we suggest using a default value of  $k = 5$ .

## Synthetic datasets

We first evaluate the performance of the proposed methods using synthetic datasets. The following three microbial datasets and phylogenetic tree information are used to generate synthetic datasets.

3-(1) Dataset 1 is based on a study of Type II diabetes (T2D) [24]. After applying a filter to remove taxa with zero abundance in over 95% of the samples, there are 53 samples and 193 features. We utilize the same strategy as mbImpute to simulate 'complete' data without non-biological zeros, and then subsequently introduce artificial zeros into the taxa to create zero-inflated data. Firstly, we apply the gamma-normal distribution to estimate the parameters of the missing value percentages for each taxon ( $z_g$ ) and taxon's abundance ( $\mu_g$ ) for the set  $\Omega$  which contains the features that are unlikely to be non-biological zeros, according to mbImpute. Hence, these features do not need imputation. Secondly, we determine the percentage of artificial zeros added into the abundance of taxon  $j$  in the complete data by randomly selecting a value from the interval of

$$\left\{ z_g : \left( \mu_j \in \mu_j^{\text{comp}} - \frac{\max(\mu_j^{\text{comp}}) - \min(\mu_j^{\text{comp}})}{3}, \mu_j^{\text{comp}} + \frac{\max(\mu_j^{\text{comp}}) - \min(\mu_j^{\text{comp}})}{3} \right) \right\} \quad (7)$$

where  $\mu_j^{\text{comp}}$  represents the average abundance of a taxon  $j$  across samples. Lastly, the Bernoulli distribution is employed to create zero indicators and then the zero-inflated dataset is generated.

3-(2) Dataset 2 is based on QinJ\_2012\_T2D\_control dataset [25]. We calculate the percentage of zeros for each taxon and extract a sub-dataset containing less than 15% zeros, which we use as the complete data. This sub-dataset consists of 50 samples and 145 features. To generate zero-inflated data, we add false zeros to the complete data, mimicking the pattern of the zeros observed in the real data.

3-(3) Dataset 3 is based on a colorectal cancer study of Zeller\_2014 [26]. By applying the same procedure as for the

QinJ\_2012\_T2D\_control dataset, a total of 50 samples and 134 features are preserved, while additional zeros are included.

## Simulated dataset

To thoroughly assess the effectiveness of the proposed methods in DA analysis, we utilize a simulator called sparseDOSSA [27] to generate simulated 16S rRNA count data. For this simulated dataset, there are 300 taxa in 100 samples under two conditions where 60 taxa were predefined as truly differentially abundant.

## Five real microbiome datasets

The proposed methods are evaluated for their performance in classification analysis using five publicly available real microbiome datasets. For each of these datasets, there are two groups of microbial samples such as diseased group and control group. Support Vector Machines (SVMs) and Random Forest [28] are utilized for classification analysis, and the Precision-Recall Area Under the Curve (PR-AUC) was calculated. The details of the datasets used are summarized below.

5-(1) NielsenHB dataset: Nielsen et al. [29] investigated 435 taxa at the species level with phylogenetic information across 206 samples, comprising of 63 control samples and 143 samples from patients with Inflammatory Bowel Disease (IBD).

5-(2) QinJ dataset: Qin et al. [25] conducted a two-stage metagenome-wide association study to assess and characterize the gut microbiota of individuals with type 2 diabetes. With phylogenetic information, we analyze 139 samples, comprising 89 controls and 50 patients with T2D on 365 taxa at the species level.

5-(3) YuJ dataset: Yu et al. [24] assessed the diagnostic capability of fecal metagenomes in detecting colorectal cancer (CRC). We focus on 417 species with phylogenetic information, for 53 controls and 75 patients with CRC.

5-(4) TettAJ dataset: Tett et al. [25] conducted a study on psoriasis. We focus on 114 taxa at the species level, analyzing 48 (37 controls and 11 patients) samples with phylogenetic information.

5-(5) ShiB dataset: Shi et al. [30] conducted metagenomic shotgun sequencing to characterize the subgingival microbiome in periodontitis patients before and after treatment. The analysis focuses on 181 species taxa with phylogenetic tree information for 48 oral samples (24 periodontitis and 24 scaling and root planning).

## Performance evaluation on imputation accuracy

To evaluate the performance of the proposed imputation methods in imputation accuracy on the three synthetic datasets, we employ four criteria to compare the imputed data with the complete data:

1. Mean squared error (MSE) between imputed data and complete data, is calculated as follows:

$$MSE = \frac{1}{mn} \sum_{i=1}^n \sum_{g=1}^m (Comp_{gi} - Imp_{gi})^2$$

where  $Comp_{gi}$  and  $Imp_{gi}$  are the complete value and the imputed value for the  $g$ th feature/taxon in the  $i$ th sample, and  $n$  and  $m$  are the total number of samples and the total number of features/taxa, respectively,

2. Pearson correlation of each taxon between the imputed data and the complete data,
3. Mean and standard deviation (SD) of the taxon abundance; and comparison of the mean abundance after imputation to the true mean from complete data using linear regression,

4. Library size (i.e. the total number of counts from all features in a sample) and the Wasserstein distance between the imputed and complete data.

## Performance evaluation on DA analysis

To further evaluate the effects of proposed imputation methods on downstream analysis such as DA analysis, we perform a comparative study using eight state-of-the-art methods for DA analysis based on simulated 16S rRNA data using sparseDOSSA [27]. Many DA tests have been developed for microbiome data. In a recent study [31], the performance of 14 DA testing methods were compared on 38 16S rRNA gene datasets with two sample groups. While the results depend on data pre-processing, ALDEx2 and ANCOM-II are found to produce the most consistent results across studies and agree best with the intersection of results from different approaches. The DA methods we employ in this study are ALDEx2 [32], ANCOM-II [33], corncob [34], Deseq2 [35], MaAsLin2 [36], Omnibus [37], metagenomeSeq [38], and Wilcoxon test [39]. The recall, precision, F1 score, number of true positives, and number of false positives are employed to measure their performance:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where true positives (TP) represent the number of features correctly detected as differentially abundant features (DAFs), false positives (FP) represent the number of features incorrectly detected as DAFs, and false negatives (FN) represent the number of features incorrectly detected as not belonging to DAFs.

## Results

### Overview

We present new approaches for imputing microbiome count data that leverage phylogenetic information to calculate the probability of zeros and low counts that are unlikely to be biological zeros, referred to as non-biological zeros. Using this probability, we compute the sample-to-sample similarity (as shown in Fig. 1) and recover the value for possible non-biological zeros based on this similarity.

### PhyImpute and UniFracImpute improve the performance of recovering non-biological zeros

To evaluate the imputation accuracy of PhyImpute and UniFracImpute, we first applied them on three synthetic datasets: Karlsson et al. [24], Qin et al. [25], and Zeller et al. [26] and compare the imputed data against the complete data. Details regarding the generation of these datasets are provided in the Methods section. The original datasets are downloaded from the curatedMetagenomicData [40] Bioconductor package. Given to the small number of samples in microbial studies, deep-learning-based imputation methods designed for scRNA-seq data are not well-suited for microbial count data. Therefore, we compare our methods with mbImpute [17], to date, which is the only available imputation method for static microbiome data, and several commonly used model-based and smoothing-based imputation methods for



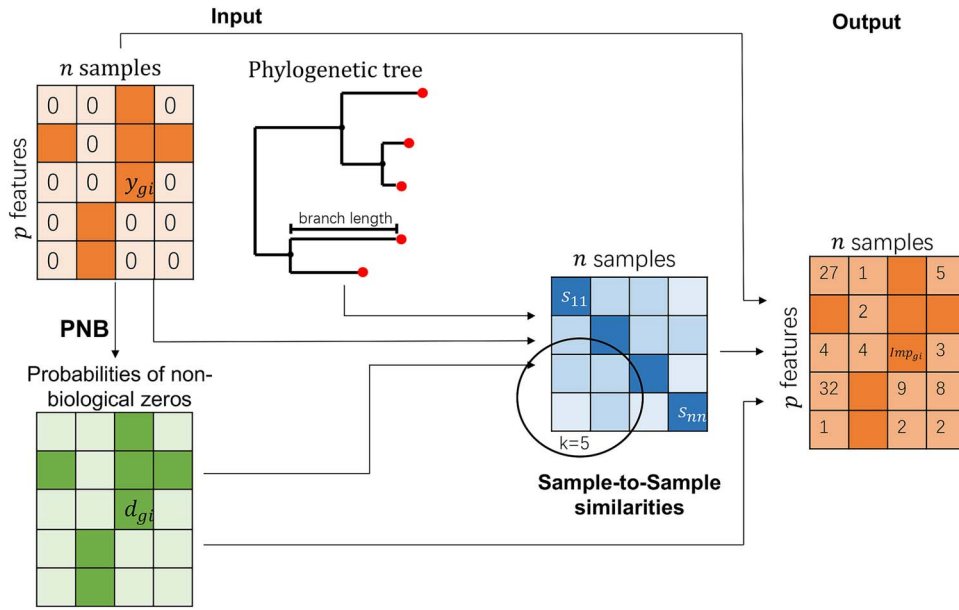


Figure 1. Illustration of microbiome data imputation workflow. First, the probability of non-biological zeros is estimated for each taxon in each sample from the raw count using the PNB mixture model. Second, raw count, non-biological zeros probability, and branch length from the phylogenetic tree are used to compute sample-to-sample similarity. Third, abundance levels of the detected non-biological zeros are imputed.

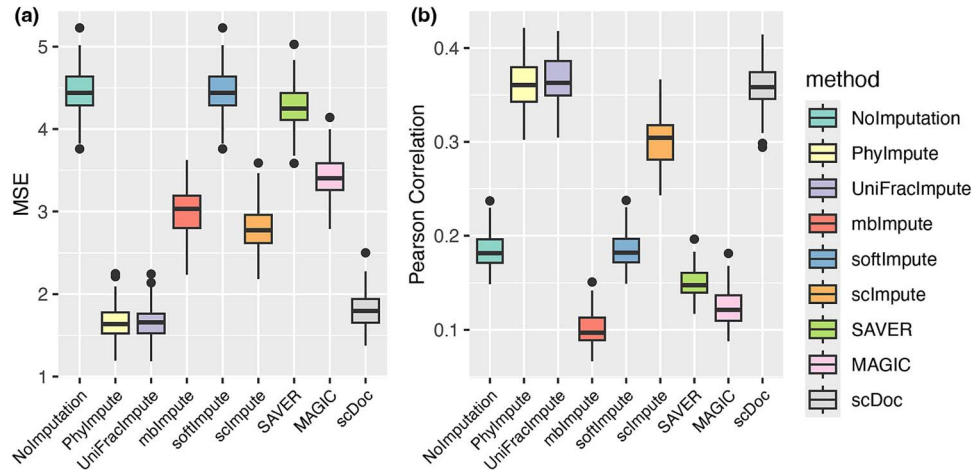


Figure 2. Evaluation metrics for the synthetic dataset of Karlsson *et al.* (a) MSE, and (b) Pearson correlation between imputed data and complete data. Colors represent different imputation methods.

scRNA-seq count data, including scDoc [7], scImpute [8], SAVER [9], MAGIC [10], and softImpute [11].

Figures 2 and 3 show the results of Karlsson *et al.*'s synthetic dataset. The results of Qin *et al.* and Zeller *et al.*'s datasets are available in the supplemental materials (S1–S4, S9–S11). Our proposed methods, PhyImpute and UniFracImpute, outperform other imputation methods across all three datasets, exhibiting smaller MSE values. We calculated Pearson correlations (Fig. 2) based on the raw data on the log-scale. Microbiome count data on the log-scale resemble a continuous normal distribution and microbiome datasets usually follow a log-normal distribution [41]. Both PhyImpute and UniFracImpute demonstrate higher correlation with the ground truth. Additionally, plots of taxon mean versus taxon SD (Fig. 3a) reveal that the proposed methods more effectively recover zeros compared to other imputation methods or the raw data with no imputation. In contrast, methods such as MAGIC, softImpute, and SAVER show much lower mean and SD of the taxa. Scatter plots comparing imputed versus true mean also demonstrate that the proposed methods outperform

others, closely following the 45° line (Fig. 3b). The  $R^2$  and  $p$ -value of models for Fig. 3(b) are shown in the supplementary file (S5). Furthermore, when assessing the Wasserstein distance between the sample library size of imputed data and complete data, our methods exhibit a smaller distance (Fig. 3c), indicating that imputed data more closely resemble the original data. Overall, the proposed imputation methods show promising results for all three synthetic studies.

### PhyImpute and UniFracImpute empower DA analysis

In the downstream DA analysis, we compare the results using imputed data by PhyImpute and UniFracImpute with those using raw data (i.e. no imputation) and using imputed data by mbImpute. Higher precision, recall, and F1 score (Fig. 4 and S12–S14) are expected if the statistically significant results closely align with the predefined DA. We found that our proposed methods perform well using ALDEx2, ANCOM-II, Corncob, DESeq2, and MaAsLin2. However, metagenomeSeq and Omnibus, which use zero-inflated

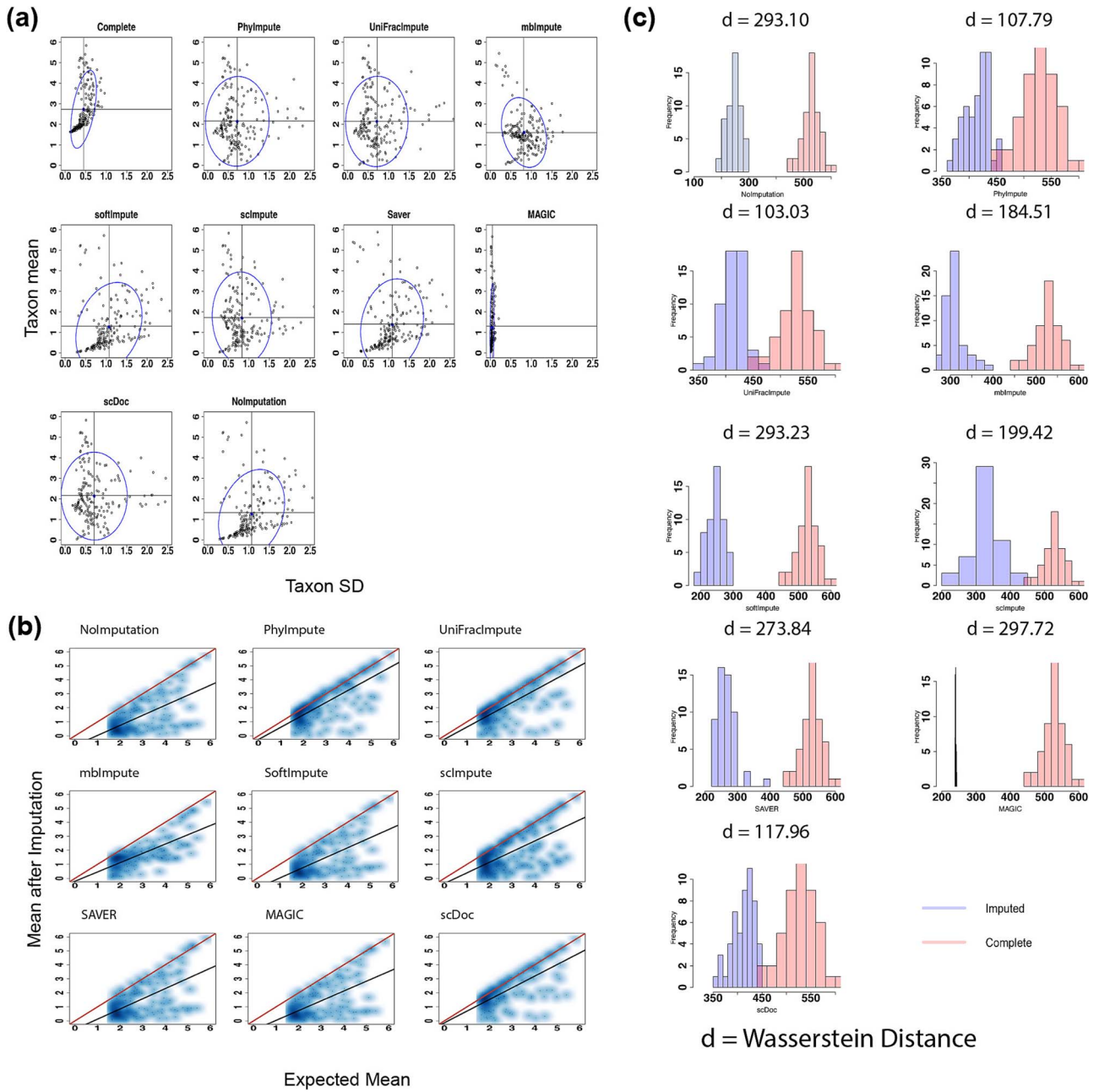


Figure 3. Feature visualization and sample library size for the synthetic dataset of Karlsson et al. (a) For each taxon, the mean and SD of its abundances are calculated for the complete data, the raw data without imputation, and the imputed data by each imputation method; (b) cloud plot for the taxon means after imputation versus the true mean for each imputation method, with an added linear regression line showing the relationship between imputed and true means, and a reference line with a slope of 1 representing perfect agreement; (c) histogram of the sample library size (bp) for the imputed and complete data with Wasserstein distance between the two distributions shown at the top of each plot.

Gaussian and zero-inflated negative binomial regression models, respectively, may not be suitable for imputed data.

We observe that PhylImpute and UniFracImpute perform very well using ALDEx2 and ANCOM-II which is consistent with the findings in a recent systematic study where these two methods produce consistent results across 38 studies amongst 14 DA testing methods [31]. Additionally, we evaluated the number of true and false positive features across all eight DA methods, both with and without imputation, as illustrated in Fig. 5 and S15. Our findings indicate that PhylImpute and UniFracImpute enhance the DA analysis when paired with ALDEx2, ANCOM, Corncob, and DESeq2. For the remaining three methods (metagenomeSeq, Omnibus, and Wilcoxon), UniFracImpute shows lower true

positives but also lower false positives. MaAsLin2 detects more true positives but higher false positives in PhylImpute.

### PhylImpute and UniFracImpute enhance the performance of disease status classification

Next, we apply the proposed methods to five real microbiome datasets, which are all generated using the whole genome shotgun sequencing technologies and can be downloaded from the curatedMetagenomicData Bioconductor package [40], for disease status classification. Three of these studies focus on stool samples [25, 29, 42], one study analyzes skin samples [43], and one study examines oral samples [30]. The specifics of each dataset details are summarized in the Materials section.

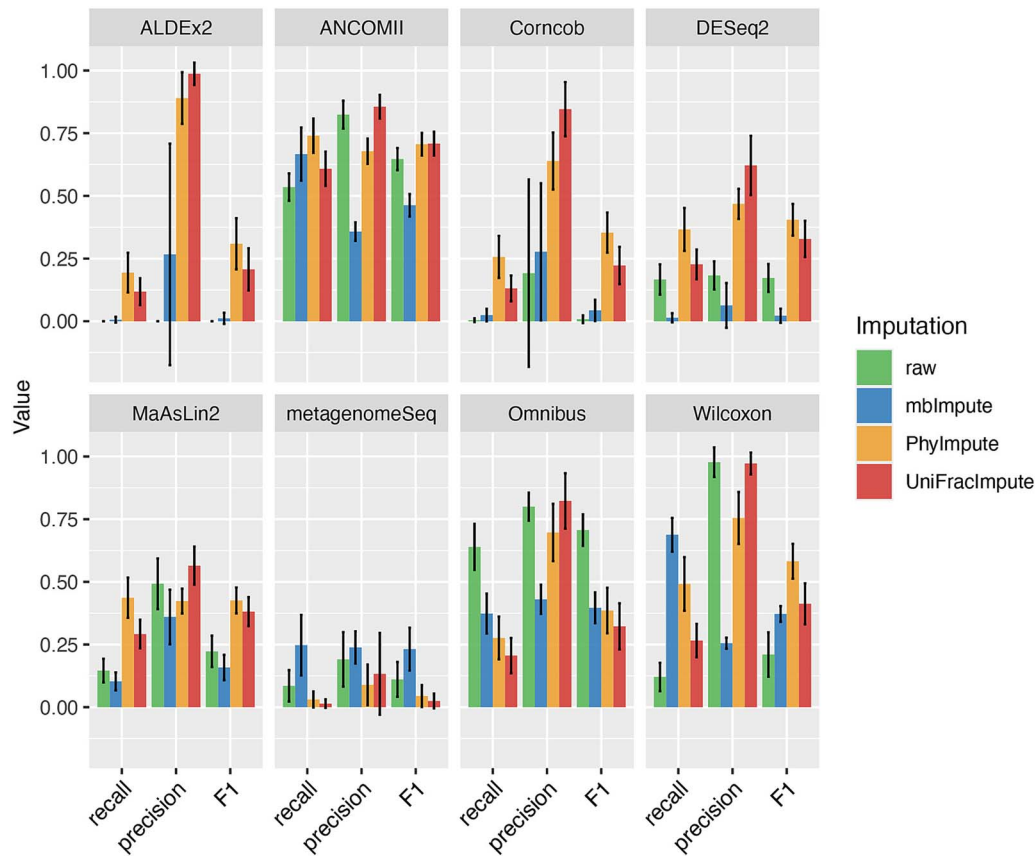


Figure 4. Accuracy measurements for eight DA methods coupled with different imputation methods in the 16S simulation studies using an FDR threshold of 0.05.

We first identify DA features between the two conditions in each of the five datasets. DESeq2 is often chosen for DA analysis, especially in contexts like RNA-seq or microbiome studies. A recent evaluation paper found that DESeq2 was one of the top-performing tools overall among those tested [44]. It is also valued for its ease of use. We explore the DA features identified by DESeq2 in the raw data and in data that are imputed using various methods, which are relevant for diagnosing the status. These identified DA features, along with the diagnosis status serving as classification labels, are used to evaluate classification results (Fig. 6 and S16 for SVM method). Data imputed using PhyImpute or UniFracImpute show slightly higher prediction accuracy compared to mbImpute or no imputation, particularly in the QinJ, ShiB, TettAJ, and YuJ studies. The Random Forest results are provided in Supplementary S18. Given that no single imputation method consistently outperformed across all studies, we included these results in the supplementary materials for the readers' benefit. Additionally, results are comparable in the NielsenHB study. These indicate that proposed methods generally improve the performance of classification analysis in microbiome studies. Additionally, we compile a list of unique DA features discovered in the imputed-DESeq2 analysis, which consistently align with findings from previous literature in terms of pathway enrichment analyses (Table 1).

Microbiome profiling has emerged as an effective approach for studying host-microbiome interactions. In this study, we utilize the identified DA features to conduct enrichment analysis using MicroPattern [45] on three selected datasets (NielsenHB, ShiB, and TettAJ), and the results are presented in Fig. 7 and Supplementary file (S6–S7 and S17). In Fig. 7, the identified DA features are linked

Table 1. Selected list of differentially abundant features detected uniquely in the imputed datasets using PhyImpute and UniFracImpute methods.

Datasets	DA features
NielsenHB	<i>Flavonifractor plautii</i> [47], <i>Dorea formicigenerans</i> [48], <i>Citrobacter freundii</i> [49]
QinJ	<i>Clostridium symbiosum</i> [25], <i>Roseburia intestinalis</i> [25], <i>Coprococcus comes</i> [50], <i>Lactobacillus amylovorus</i> [51], <i>Lactobacillus plantarum</i> [51], <i>Bacteroides xylanisolvens</i> [52]
ShiB	<i>Veillonella atypica</i> [53], <i>Streptococcus salivarius</i> [54], <i>Streptococcus sanguinis</i> [55], <i>Actinomyces oris</i> [56], <i>Capnocytophaga granulosa</i> [55], <i>Actinomyces sp oral taxon 448</i> [57]
TettAJ	<i>Escherichia coli</i> [57]
YuJ	<i>Fretibacterium fastidiosum</i> [58], <i>Bacteroides stercoris</i> [59]

to their respective diseases. For instance, in the study focusing on IBD, as shown in Fig. 7(a), the DA features identified through the proposed imputation methods establish a connection between Irritable Bowel Syndrome (IBS) and Enterocolitis Necrotizing Disease. IBD and IBS are two common chronic gastrointestinal disorders that exhibit significant overlap in terms of symptoms, pathophysiology, and treatment. This suggests that despite being at opposite ends of the spectrum, they might be a single disease entity [46].

Significant features detected in the ShiB and TettAJ studies are associated with periodontal diseases and psoriasis, respectively, which align with the original focus of these studies. However, in QinJ and YuJ studies, the DA features could not establish

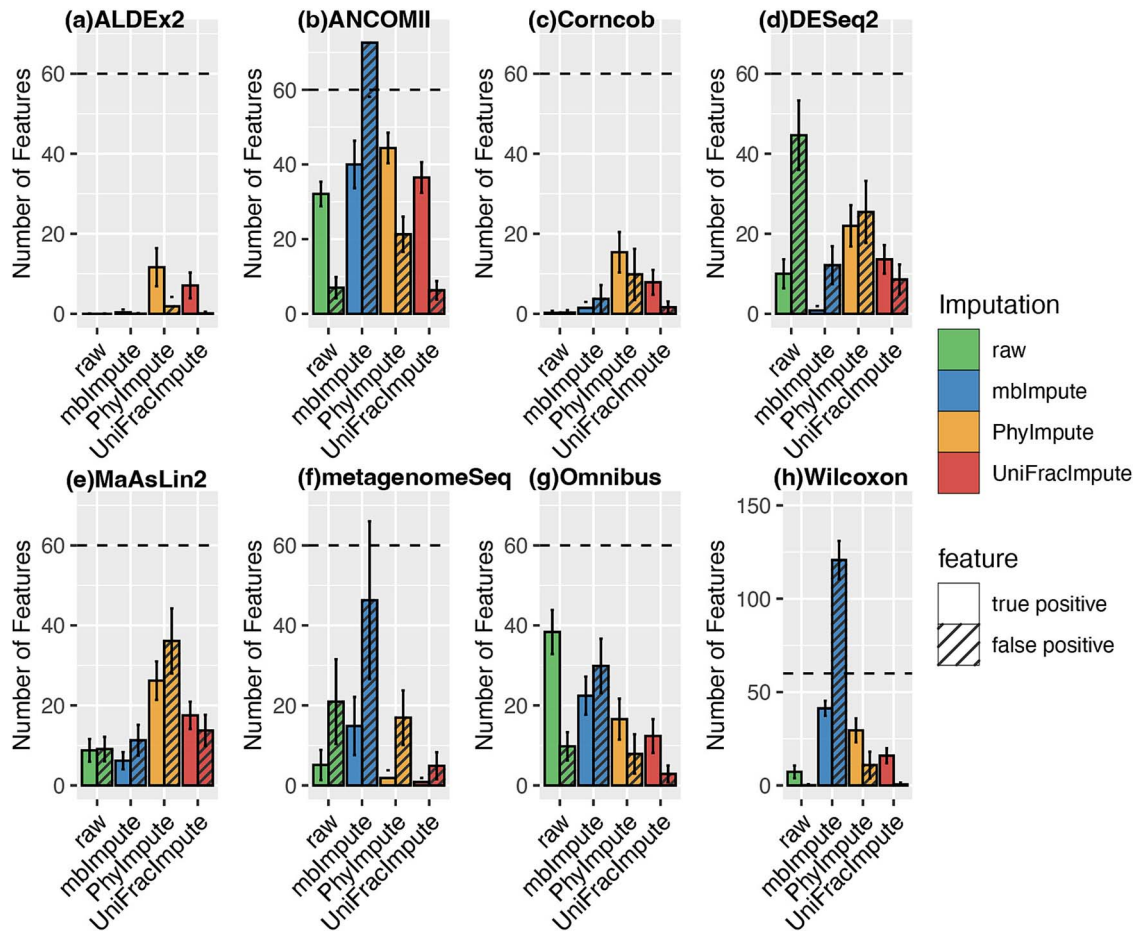


Figure 5. The number of true and false positive features detected from eight DA methods coupled with different imputation methods for the simulation study with 300 features and 100 replicated simulations. Each bar represents the mean number of features that are detected as differentially abundant at an FDR threshold level of 0.05, and the error bars represent 1 SD calculated from 100 replications.

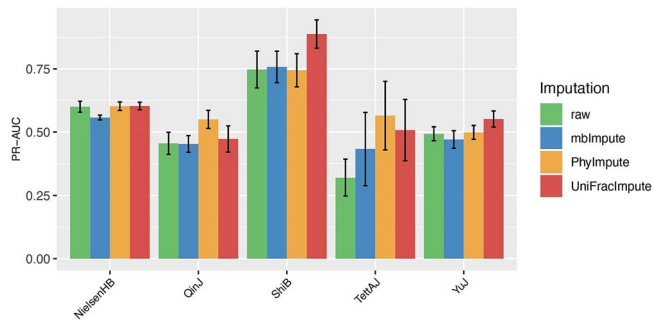


Figure 6. Comparison of the classification accuracy of imputation methods on five real datasets. The bar plots show classification accuracy from different imputation methods, measured by a five-fold cross-validated PR-AUC by the SVMs for predicting diagnosis status in five real datasets.

a link between diabetes mellitus type 2 and colorectal cancer, as indicated in the supplementary materials. The DA features identified in the NielsenHB (stool study), ShiB (oral study), and TetAJ (skin study) datasets, after imputation using PhylImpute and UniFracImpute, provide a meaningful functional interpretation.

## Discussion

In our proposed work, we utilize the probability of non-biological zeros, along with an abundance table, and the total branch length

of the phylogenetic tree. This allows us to calculate the sample-to-sample similarity and recover the values of detected non-biological zeros by leveraging information from  $K$  similar samples. For most microbiome count data, the default  $K=5$  is sufficient, but for larger sample sizes, users can opt for a larger  $K$  value. Our method incorporates modifications to the weighted UniFrac distance and cosine similarity in our method. The sample-to-sample similarity calculated in our approaches incorporates feature weight based on the probability of non-biological zeros. This enhances the accuracy of the sample-to-sample similarity calculations and improves the recovery of non-biological zeros. It is important to note that a limitation of both PhylImpute and UniFracImpute arises when a phylogenetic tree is unavailable. In such cases, we assume that all branches have the same length of 1, and UniFracImpute functions similarly to scDoc [7].

In the simulation and real data analysis, the PNB distribution is used to model the microbiome data with excess zeros, as well as to estimate the probability of non-biological zeros. Alternatively, the estimated probability of non-biological zeros using the zero-inflated log-normal (ZILN) distribution, available through the metagenomeSeq [38] package in R Bioconductor, can also be used for this purpose. The performance comparison of various methods is shown in Fig. S8 in the Supplementary file. As demonstrated, metrics such as MSE and Pearson correlation coefficients are similar for both the PNB and ZILN models and both outperform other imputation methods. In practice, users can explore different models, such as PNB and ZILN, to estimate the



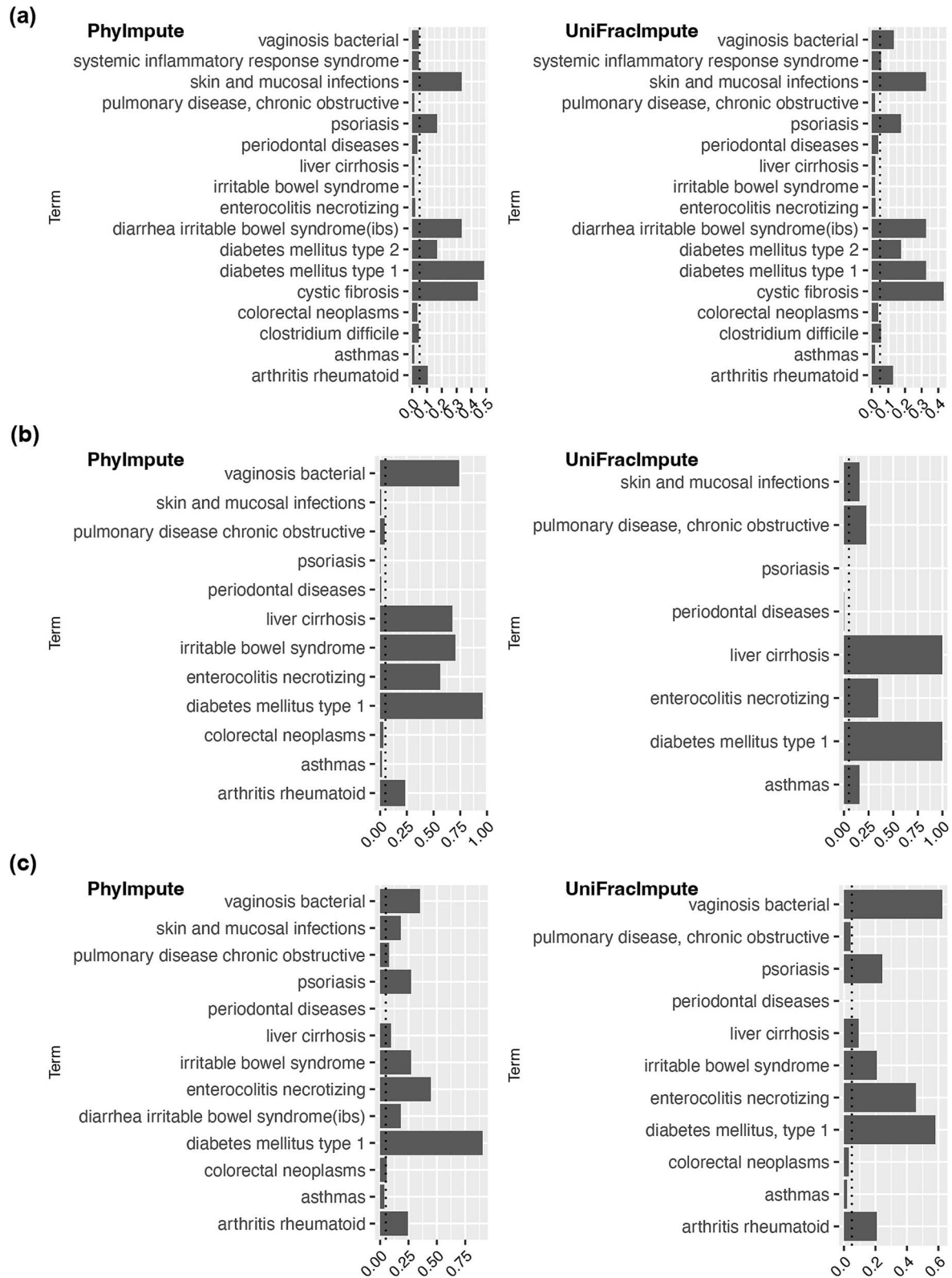


Figure 7. Enrichment analysis of differentially abundant features for three selected datasets, namely (a) NielsenHB (stool study), (b) TettAJ (skin study), and (c) ShiB (oral study). The differentially abundant features are detected based on imputed data using PhylImpute and UniFracImpute, respectively. The x-axis is the FDR value, and the vertical dashed line indicates the FDR level of 0.05.

probabilities of non-biological zeros and choose an appropriate model for their data. Identifying the most appropriate and robust model for microbiome data, however, remains a challenging and ongoing area of research.

The proposed imputation methods are able to take the raw read matrix as input and produce a count matrix with the same dimensions as output. This output count matrix eliminates the influence of many non-biological zeros and can be seamlessly

integrated with downstream analyses, such as DA analysis and classification. Through simulation studies, the proposed imputation methods have proven their effectiveness in recovering non-biological zeros and have shown significant utility in the downstream analysis of microbial count data. The results obtained from three synthetic microbial datasets and 16s rRNA simulation data highlight that the proposed imputation methods consistently outperform mbImpute and existing scRNA-seq imputation methods. They exhibit superior accuracy based on various evaluation metrics, including MSE, Pearson correlation, the mean and SD of each taxon, and Wasserstein distance. To evaluate the performance of the imputation methods in DA analysis, we use simulated 16s rRNA data with predefined DAFs and conducted DA analysis using eight state-of-the-art methods. We assess the precision, recall, F1-score, number of false positives, and number of false negatives, and find that the proposed imputation methods yield promising results across most DA methods. However, it's important to note that PhyImpute and UniFracImpute may not be suitable for certain DA analysis methods, such as metagenome and Omnibus, which employ zero-inflated models.

When comparing the computing time of our proposed methods, PhyImpute and UniFracImpute, to existing imputation methods, we observe that our approaches offer a balance between accuracy and computational efficiency. Our methods run faster than mbImpute, but much slower than scRNA-seq imputation methods such as MAGIC and softImpute. Using synthetic data (53 samples and 193 features) as an example, PhyImpute and UniFracImpute take 28.92 minutes and 68.4 minutes, respectively. In comparison, mbImpute takes 168.6 minutes, MAGIC and softImpute only take less than 10 seconds, scDoc takes 2.73 minutes, scImpute takes 7.1 minutes, and SAVER takes 16.47 minutes using Apple M1 Pro, 16GB memory.

Furthermore, we apply the proposed approaches, combined with the DA testing approach, to five real microbiome datasets, which include samples from various anatomical locations, such as the human gut, skin, and oral cavity. The results of this analysis showcase that DAFs detected using imputed data with DESeq2 empower the accuracy for disease status classification. Importantly, the DAFs detected in the analysis are consistent with previous biological findings regarding their association with the respective disease.

In this study, we explored methods for imputing non-biological zeros in microbiome data, addressing the challenges posed by sparsity and zero-inflation. Both PhyImpute and UniFracImpute have shown effective performance using the PNB model, and we have provided an alternative approach using ZILN model. Moving forward, longitudinal microbiome studies have become increasingly popular to capture the temporal dynamics of microbial communities. Our future research will focus on refining imputation methods that are tailored specifically for longitudinal microbiome data. This work will involve developing models that incorporate temporal dependencies, allowing us to better estimate missing values in a way that respects the natural progression of microbial communities over time.

#### Key Points

- We developed two imputation methods, PhyImpute and UniFracImpute, to deal with the excess zeros issue in microbiome data.

- The methods have been consistently effective across synthetic studies, 16S simulation studies, and real data applications, demonstrating their robustness in recovering non-biological zeros.
- PhyImpute and UniFracImpute significantly improve downstream analyses, such as DA analysis and classification tasks, making them valuable tools for microbiome data processing.

## Authors' contributions

LA and QL conceived the study. QL and LA designed the methods and algorithms. QL, SZ, and HB performed the simulation studies. QL, YC, HJ, and LA contributed to the real data analyses. All authors drafted, reviewed, and revised the manuscript.

## Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

This work has been partially supported by the United States Department of Agriculture (ARZT-1361620-H22-149 to L.A.) and the National Institute of Health (R01AI149754 and R01ES027013 to Y.C.).

## Data availability

The code is available at [github.com/anlingUA/PhyImpute\\_UniFracImpute](https://github.com/anlingUA/PhyImpute_UniFracImpute)

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## References

1. Berg G, Rybakova D, Fischer D. et al. Microbiome definition revisited: old concepts and new challenges. *Microbiome* 2020;**8**:103. <https://doi.org/10.1186/s40168-020-00875-0>.
2. Li H, Li H. Introduction to special issue on statistics in microbiome and metagenomics. *Stat Biosci* 2021;**13**:197–9. <https://doi.org/10.1007/s12561-021-09307-5>.
3. Kaul A, Mandal S, Davidov O. et al. Analysis of microbiome data in the presence of excess zeros. *Front Microbiol* 2017;**8**:2114.
4. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. *Elife* 2019;**8**:8. <https://doi.org/10.7554/eLife.46923>.

5. Silverman JD, Roche K, Mukherjee S. et al. Naught all zeros in sequence count data are the same. *Comput Struct Biotechnol J* 2020;**18**:2789–98. <https://doi.org/10.1016/j.csbj.2020.09.014>.
6. Calgaro M, Romualdi C, Waldron L. et al. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol* 2020;**21**:1–31. <https://doi.org/10.1186/s13059-020-02104-1>.
7. Ran D, Zhang S, Lytal N. et al. scDoc: correcting drop-out events in single-cell RNA-seq data. *Bioinformatics* 2020;**36**:4233–9. <https://doi.org/10.1093/bioinformatics/btaa283>.
8. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 2018;**9**:997. <https://doi.org/10.1038/s41467-018-03405-7>.
9. Huang M, Wang J, Torre E. et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42. <https://doi.org/10.1038/s41592-018-0033-z>.
10. van Dijk D, Sharma R, Nainys J. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**:716–729.e27e27. <https://doi.org/10.1016/j.cell.2018.05.061>.
11. Hastie T. et al. Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research* 2015;**16**:3367–3402.
12. Lähnemann D, Köster J, Szczurek E. et al. Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**:1–35. <https://doi.org/10.1186/s13059-020-1926-6>.
13. Hou W, Ji Z, Ji H. et al. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol* 2020;**21**:1–30. <https://doi.org/10.1186/s13059-020-02132-x>.
14. Dai C, Jiang Y, Yin C. et al. scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods. *Nucleic Acids Res* 2022;**50**:4877–99. <https://doi.org/10.1093/nar/gkac317>.
15. Wang B, Pourshafeie A, Zitnik M. et al. Network enhancement as a general method to denoise weighted biological networks. *Nat Commun* 2018;**9**:3108. <https://doi.org/10.1038/s41467-018-05469-x>.
16. Cheng Y, Ma X, Yuan L. et al. Evaluating imputation methods for single-cell RNA-seq data. *BMC Bioinform* 2023;**24**:302. <https://doi.org/10.1186/s12859-023-05417-7>.
17. Jiang R, Li WV, Li JJ. mbImpute: an accurate and robust imputation method for microbiome data. *Genome Biol* 2021;**22**:192. <https://doi.org/10.1186/s13059-021-02400-4>.
18. Kim KJ, Park J, Park SC. et al. Phylogenetic tree-based microbiome association test. *Bioinformatics* 2020;**36**:1000–6. <https://doi.org/10.1093/bioinformatics/btaz686>.
19. Plantinga AM, Chen J, Jenq RR. et al. pldist: ecological dissimilarities for paired and longitudinal microbiome association analysis. *Bioinformatics* 2019;**35**:3567–75. <https://doi.org/10.1093/bioinformatics/btz120>.
20. Wu C, Chen J, Kim J. et al. An adaptive association test for microbiome data. *Genome Med* 2016;**8**:56. <https://doi.org/10.1186/s13073-016-0302-3>.
21. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005;**71**:8228–35. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>.
22. Lozupone CA, Hamady M, Kelley ST. et al. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 2007;**73**:1576–85. <https://doi.org/10.1128/AEM.01996-06>.
23. Chen J, Bittinger K, Charlson ES. et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 2012;**28**:2106–13. <https://doi.org/10.1093/bioinformatics/bts342>.
24. Karlsson FH, Tremaroli V, Nookaew I. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 2013;**498**:99–103. <https://doi.org/10.1038/nature12198>.
25. Qin J, Li Y, Cai Z. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60. <https://doi.org/10.1038/nature11450>.
26. Zeller G, Tap J, Voigt AY. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 2014;**10**:766. <https://doi.org/10.15252/msb.20145645>.
27. Ma S, Ren B, Mallick H. et al. A statistical model for describing and simulating microbial community profiles. *PLoS Comput Biol* 2021;**17**:e1008913. <https://doi.org/10.1371/journal.pcbi.1008913>.
28. Hollister EB, Oezguen N, Chumpitazi BP. et al. Leveraging human microbiome features to diagnose and stratify children with irritable bowel syndrome. *J Mol Diagn* 2019;**21**:449–61. <https://doi.org/10.1016/j.jmoldx.2019.01.006>.
29. Nielsen HB, Almeida M, Juncker AS. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 2014;**32**:822–8. <https://doi.org/10.1038/nbt.2939>.
30. Shi B, Chang M, Martin J. et al. Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. *MBio* 2015;**6**:e01926–14. <https://doi.org/10.1128/mBio.01926-14>.
31. Nearing JT, Douglas GM, Hayes MG. et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat Commun* 2022;**13**:342. <https://doi.org/10.1038/s41467-022-28034-z>.
32. Fernandes AD, Reid JNS, Macklaim JM. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2014;**2**:15. <https://doi.org/10.1186/2049-2618-2-15>.
33. Mandal S, van Treuren W, White RA. et al. Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microb Ecol Health Dis* 2015;**26**:27663. <https://doi.org/10.3402/mehd.v26.27663>.
34. Martin BD, Witten D, Willis AD. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann Appl Stat* 2020;**14**:94–115. <https://doi.org/10.1214/19-aos1283>.
35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. <https://doi.org/10.1186/s13059-014-0550-8>.
36. Mallick H, Rahnavard A, McIver LJ. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput Biol* 2021;**17**:e1009442. <https://doi.org/10.1371/journal.pcbi.1009442>.
37. Chen J, King E, Deek R. et al. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* 2018;**34**:643–51. <https://doi.org/10.1093/bioinformatics/btx650>.
38. Paulson JN, Stine OC, Bravo HC. et al. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013;**10**:1200–2. <https://doi.org/10.1038/nmeth.2658>.
39. Wilcoxon F. Individual Comparisons by Ranking Methods. In: Kotz S, Johnson NL. (eds), *Break throughs in Statistics: Methodology and Distribution*, pp. 196–202. New York, NY: Springer, 1992.
40. Pasolli E, Schiffer L, Manghi P. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* 2017;**14**:1023–4. <https://doi.org/10.1038/nmeth.4468>.
41. Zhou H, He K, Chen J. et al. LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome Biol* 2022;**23**:95. <https://doi.org/10.1186/s13059-022-02655-5>.

42. Yu J, Feng Q, Wong SH. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 2017;**66**:70–8. <https://doi.org/10.1136/gutjnl-2015-309800>.
43. Tett A, Pasolli E, Farina S. et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ Biofilms Microbiomes* 2017;**3**:14. <https://doi.org/10.1038/s41522-017-0022-5>.
44. Calgaro M, Romualdi C, Waldron L. et al. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol* 2020;**21**:191. <https://doi.org/10.1186/s13059-020-02104-1>.
45. Ma W, Huang C, Zhou Y. et al. MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes. *Sci Rep* 2017;**7**:40200. <https://doi.org/10.1038/srep40200>.
46. Abdul Rani R, Raja Ali RA, Lee YY. Irritable bowel syndrome and inflammatory bowel disease overlap syndrome: pieces of the puzzle are falling into place. *Intest Res* 2016;**14**:297–304. <https://doi.org/10.5217/ir.2016.14.4.297>.
47. Pisani A, Rausch P, Ellul S. et al. P685 gut microbiota in patients with inflammatory bowel disease during remission. *J Crohn's Colitis* 2021;**15**:S604–5. <https://doi.org/10.1093/ecco-jcc/jjab076.805>.
48. Schirmer M, Garner A, Vlamakis H. et al. Microbial genes and pathways in inflammatory bowel disease. *Nat Rev Microbiol* 2019;**17**:497–511. <https://doi.org/10.1038/s41579-019-0213-6>.
49. Ricanek P, Lothe SM, Frye SA. et al. Gut bacterial profile in patients newly diagnosed with treatment-naïve Crohn's disease. *Clin Exp Gastroenterol* 2012;**5**:173–86. <https://doi.org/10.2147/CEG.S33858>.
50. Graessler J, Qin Y, Zhong H. et al. Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: Correlation with inflammatory and metabolic parameters. *Pharmacogenomics J* 2013;**13**: 514–22. <https://doi.org/10.1038/tpj.2012.43>.
51. Gurung M, Li Z, You H. et al. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* 2020;**51**:102590. <https://doi.org/10.1016/j.ebiom.2019.11.051>.
52. Bakir-Gungor B, Bulut O, Jabeer A. et al. Discovering potential taxonomic biomarkers of type 2 diabetes from human gut microbiota via different feature selection methods. *Front Microbiol* 2021;**12**:628426. <https://doi.org/10.3389/fmicb.2021.628426>.
53. Mashima I, Theodorea CF, Thaweboon B. et al. Identification of Veillonella species in the tongue biofilm by using a novel one-step polymerase chain reaction method. *PloS One* 2016;**11**:e0157516. <https://doi.org/10.1371/journal.pone.0157516>.
54. Jansen PM, Abdelbary MMH, Conrads G. A concerted probiotic activity to inhibit periodontitis-associated bacteria. *PloS One* 2021;**16**:e0248308. <https://doi.org/10.1371/journal.pone.0248308>.
55. Colombo AP, Boches SK, Cotton SL. et al. Comparisons of subgingival microbial profiles of refractory periodontitis, severe periodontitis, and periodontal health using the human oral microbe identification microarray. *J Periodontol* 2009;**80**:1421–32. <https://doi.org/10.1902/jop.2009.090185>.
56. Pichaphop C, Apiwattanakul N, Wanitkun S. et al. Bacterial endocarditis caused by Actinomyces oris: first reported case and literature review. *J Investig Med High Impact Case Rep* 2020;**8**:2324709620910645. <https://doi.org/10.1177/2324709620910645>.
57. Zhang Y, Shao C, Zhang T. et al. Periodontal and peri-implant microbiome dysbiosis is associated with alterations in the microbial community structure and local stability. *Front Microbiol* 2021;**12**:785191. <https://doi.org/10.3389/fmicb.2021.775570>.
58. Zhao L, Cho WC, Nicolls MR. Colorectal cancer-associated microbiome patterns and signatures. *Front Genet* 2021;**12**:787176. <https://doi.org/10.3389/fgene.2021.787176>.
59. Lucas C, Barnich N, Nguyen HTT. Microbiota, Inflammation and Colorectal Cancer. *International Journal of Molecular Sciences* 2017;**18**:1310.