Computational and Structural Biotechnology Journal 21 (2023) 4804–4815

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Mini-review

Data pre-processing for analyzing microbiome data – A mini review

Ruwen Zhou [a], Siu Kin Ng [a], Joseph Jao Yiu Sung [a,b], Wilson Wen Bin Goh [a,c,d,*], Sunny Hei Wong [a,b,**]

[a] Lee Kong Chian School of Medicine, Nanyang Technological University, 11 Mandalay Road, 308232, Singapore
[b] Department of Gastroenterology and Hepatology, Tan Tock Seng Hospital, National Healthcare Group, 11 Jalan Tan Tock Seng, 308433, Singapore
[c] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore
[d] Center for Biomedical Informatics, Nanyang Technological University, 59 Nanyang Drive, 636921, Singapore

ARTICLE INFO

Keywords:
Microbiome Data
Data Preprocessing
Normalization
Batch Effect
16S rRNA Sequencing

ABSTRACT

The human microbiome is an emerging research frontier due to its profound impacts on health. High-throughput microbiome sequencing enables studying microbial communities but suffers from analytical challenges. In particular, the lack of dedicated preprocessing methods to improve data quality impedes effective minimization of biases prior to downstream analysis. This review aims to address this gap by providing a comprehensive overview of preprocessing techniques relevant to microbiome research. We outline a typical workflow for microbiome data analysis. Preprocessing methods discussed include quality filtering, batch effect correction, imputation of missing values, normalization, and data transformation. We highlight strengths and limitations of each technique to serve as a practical guide for researchers and identify areas needing further methodological development. Establishing robust, standardized preprocessing will be essential for drawing valid biological conclusions from microbiome studies.

1. Introduction

The human microbiome consists of diverse communities of bacteria, viruses, and fungi that inhabit various parts of the human body, including the gut, skin, and lungs. Owing to its profound impact on human health, the microbiome has been extensively studied in recent years [1,2]. Recent research suggests that microbial dysbiosis plays an important role in intestinal diseases, such as colorectal cancer (CRC) [3] and inflammatory bowel diseases (IBD) [4]. Nevertheless, our understanding of the human microbiome remains incomplete. Comprehensively analyzing microbial composition and dynamics is essential for unraveling the effects of microbiome on human health. Comprehensive bioinformatic analyses of microbiome datasets can provide key insights into microbiota disturbances, enhancing knowledge of disease mechanisms and guiding therapeutic development.

The complex nature of microbiome data necessitates effective preprocessing to ensure robust downstream analyses. While numerous preprocessing methods exist, a comprehensive review of these techniques is lacking. This paper aims to address this gap by providing an overview for current preprocessing methods of microbiome data analysis. We compare their strengths and limitations of these methods to offer practical guidelines for their application.

Microbiome profiling involves bulk sequencing to identify microorganisms in each sample. Common techniques include 16S ribosomal RNA (rRNA) gene sequencing for prokaryotic species, internal transcribed spacer (ITS) sequencing for fungal species, or shotgun metagenomic sequencing. The 16S rRNA gene contains both conserved and nine hypervariable regions (V1-V9), enabling it to serve as a fingerprint

Abbreviations: ALR, Additive log-ratio; ASVs, Amplicon sequence variants; BDMMA, Bayesian Dirichlet-multinomial regression meta-analysis; BIOM, BIological Observation Matrix; CLR, Centered log-ratio; CRC, Colorectal cancer; CoDa, Compositional Batch Effects Correction; CSS, Cumulative-sum scaling; GAN, Generative adversarial network; IBD, Inflammatory bowel diseases; ITS, Internal transcribed spacer; ILR, Isometric log-ratio; Limma, Linear Models for Microarray Data; LIGER, Linked inference of genomic experimental relationships; ANCOM-BC, Microbiome with Bias Correction; MVI, Missing value imputation; MNN, Mutual Nearest Neighbors; OTUs, Operational taxonomic units; PLSDA, Partial Least Squares Discriminant Analysis; PCA, Principal Component Analysis; RNN, Recurrent neural network; RUV, Remove Unwanted Variation; SVA, Surrogate Variable Analysis; TCGA, The Cancer Genome Atlas; TSS, Total-Sum Scaling; TMM, Trimmed Mean of M-value; ZINB, Zero-inflated negative binomial; ZIP, Zero-inflated Poisson.

* Correspondence to: Lee Kong Chian School of Medicine, Nanyang Technological University, 59 Nanyang Drive, 636921, Singapore.
** Corresponding author at: Lee Kong Chian School of Medicine, Nanyang Technological University, 11 Mandalay Road, 308232, Singapore.
E-mail addresses: wilsongoh@ntu.edu.sg (W.W.B. Goh), sunny.wong@ntu.edu.sg (S.H. Wong).

https://doi.org/10.1016/j.csbj.2023.10.001
Received 14 July 2023; Received in revised form 1 October 2023; Accepted 1 October 2023
Available online 4 October 2023

for microbial identification. Analysis of microbiome data can be performed directly on the amplicon sequence variants (ASVs) obtained from high-throughput sequencing. Sequencing inaccuracies can occur at a rate of 0.1–1.5 % per nucleotide for Illumina sequencing [5,6]. Similar sequences are typically clustered into operational taxonomic units (OTUs) using a 97 % similarity cutoff through an OTU picking process [7]. This OTU table provides the basis for downstream analysis. In contrast, shotgun metagenomic sequencing profiles the full host and microbial community composition in a given sample [8]. It allows for a broader examination of the genetic material present. The data can be used to identify the microbial species present and elucidate the functional genes and metabolic capabilities of the microbial community. For ITS sequencing, the ITS1 and ITS2 regions, are highly variable among different fungal species, thus serving as a reliable marker for fungal identification and phylogenetic analysis [9]. While this technique is mainly geared towards fungal taxonomy, it does not provide direct insight into the functional or metabolic aspects of the fungal community. Recent advances in long-read sequencing technologies, such as nanopore (Oxford Nanopore) and single-molecule real time (Pacific Bioscience), allow for the generation of much longer reads, from tens to hundreds of thousands of bases. The combination of both shotgun and long-read sequencing technologies can improve resolution, accuracy, and completeness of functional profiling, metagenome binning, and assembly [10]. Despite the power of metagenomics, 16S rRNA gene sequencing remains popular, especially for profiling microbes in samples with low microbial biomass, like tissue biopsies.

Regardless of the sequencing technologies or OTU picking strategies, microbiome data is commonly represented as a tabular abundance matrix (such as an OTU/ASV table or functional feature table) where samples are organized as columns and microbial species, or features are arranged as rows. Microbiome datasets have several typical characteristics. Typically, microbiome data are over-dispersed, meaning that the abundances of features are highly variable. The data is also high-dimensional, with potentially thousands of features being profiled. Finally, microbiome data exhibit high sparsity, with the abundance matrix filled with many zeros – often up to 90 % [11]. This sparsity occurs because many species are only present at low abundance or completely absent in each sample. For instance, the fecal microbiome consists of billions of microbes from thousands of distinct phylogenetic lineages [12], but only a subset of this diversity is present within each individual. The combination of over-dispersion, high dimensionality, and sparsity poses challenges for statistical analysis and subsequent interpretation.

The compositional nature of microbial profiles poses challenges for detecting rare taxa. The measurement sensitivity for identifying low abundance microbes depends on the amplification method, sequencing technology, or sequencing depth. Taxa below detection threshold are assigned zero abundance in compositional data, even though they may actually be present at low levels. Thus, zeros can originate from two sources: technical zeros, which stem from limited measurement sensitivity, and biological zeros, which represent true microbial absence. An overabundance of technical zeros can lead to inflated sparsity in the compositional data matrix. Without appropriate preprocessing, the excessive zeros can introduce biases and impair the performance of downstream statistical and machine learning methods on raw microbiome data [13,14].

Data preprocessing techniques involving filtering and transformation, are important [15]. Raw data often contain inconsistencies and errors that can bias conclusions. For instance, batch effects arise when different groups of samples are processed at different times or under different conditions. In addition, microbiome data might be incomplete due to technical constraints, as limitations in sequencing depth or detection sensitivity mean some microbes are not fully captured. To mitigate these issues, data preprocessing is applied to improve data quality before analysis. Preprocessing helps address missing data, reduce technical noise and biases, and filter out uninformative features. We focus on typical data preprocessing workflows (Fig. 1) for microbiome sequencing data. Key steps include quality filtering, batch effect correction, imputation of missing values, data normalization, and transformation (Fig. 2). The strengths and limitations of these steps are summarized in Table 1, whereas the comparison between characteristics of transcriptomic data and microbiome data is in Table 2. While microbiome data analysis is the primary focus of this review, many of the preprocessing methods can be applied to other data types such as transcriptomic [16] and proteomic [17]. Metatranscriptomics focuses on analyzing the collective set of RNA transcripts to understand which genes are being actively expressed, and metaproteomics examines the profile of expressed proteins. They are gaining attention to decipher the active biological processes occurring within microbiome communities [18–20]. As these technologies mature, robust preprocessing will be crucial to derive significant biological insights. Careful data curation before analysis can enhance the reliability, interpretability, and comparability of the results.

2. Pre-processing methods

Pre-processing of raw sequencing data is essential before generating the abundance matrices. Preprocessed, denoising, dereplication, and taxonomic assignment of 16S rRNA gene sequences are often performed using two widely used packages, QIIME2 [42] or Mothur [43]. They generate abundance tables and taxonomies in standard formats such as the BIological Observation Matrix (BIOM) format or tabular plain text format. In contrast, diverse workflows exist for processing shotgun metagenomic data. Packages such as MetaPhlAn [44] and MEGAN [45] were developed to taxonomically profile metagenomic reads after quality control and host filtering.

1) Filtering

Microbiome data analysis typically begins with raw sequences in plain text format such as FASTQ or FASTA. The FASTQ format is used to store both the raw sequence data and its associated base quality scores, whereas FASTA files are usually lack quality scores. The initial preprocessing step usually involves quality control of raw sequencing reads, during which all platform-specific sequencing adapters and bases with a low base calling score (Phred score) are trimmed from the sequencing reads. Read pairs with a low average Phred score are also discarded prior to downstream analysis. Tools such as FastQC [21], Trimmomatic [22] or Cutadapt [23] are widely used for this step.

Abundance matrices and taxonomic assignments for amplicon data and metagenomics data can be generated by 16S rRNA analytic pipelines such as QIIME2 [42] or Mothur [43], or packages such as MetaPhlAn [44], MEGAN [45], Kraken [46], or Kaiju [47]. For 16S rRNA data, QIIME2 and Mothur are commonly used for tasks like filtering, denoising, dereplication, and taxonomic assignment. The resulting abundance matrix and taxonomic assignment are typically in BIOM format or tabular plain text format. Metagenomic data, on the other hand, are processed using different workflows and pipelines such as MetaPhlAn [44] and MEGAN [45]. Before utilizing these tools, quality filtering and the removal of host-derived reads are usually performed. The abundance matrix undergoes two filtering methods: sample filtering and OTU/feature filtering.

Sample filtering involves removing samples with significantly lower library sizes (total number of reads) compared to other samples in the dataset. These low-quality samples may be affected by poor sample or DNA quality, or technical errors, which could introduce biases or inaccuracies in downstream analysis [48]. It is important to conduct sample filtering to ensure a reliable representation of the microbial community without sacrificing too much statistical power.

OTU or feature filtering, on the other hand, focuses on dropping OTUs or features with low abundance or prevalence. This step is crucial to improve the reliability of analysis and can have a significant impact

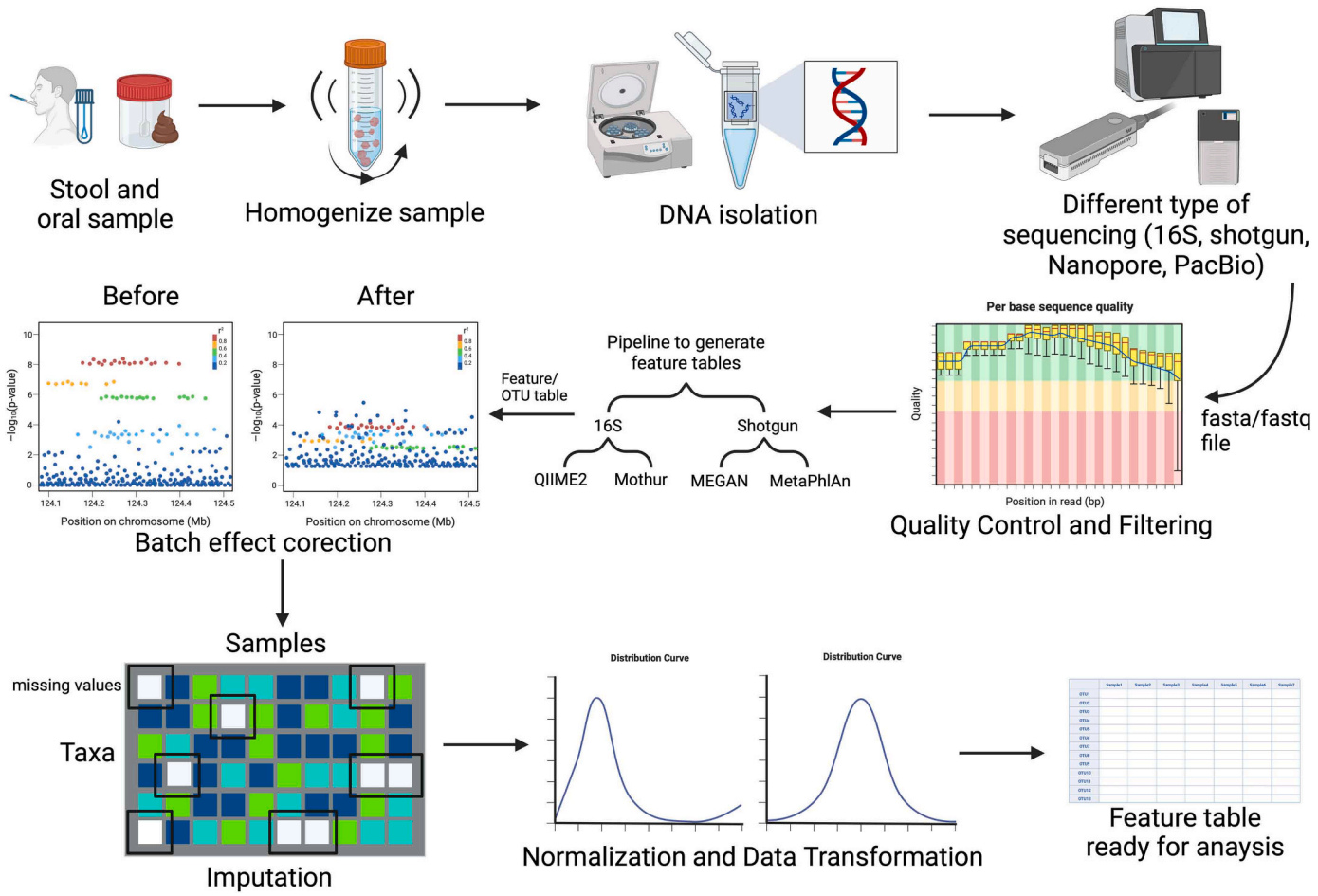

Fig. 1. Flow chart from DNA extraction to data preprocessing steps This process encompasses the extraction of microbiome DNA from human stool or oral samples, followed by 16S or shotgun sequencing. Subsequently, standard pipelines are employed to generate feature tables, which are then processed for filtering, batch effect correction, missing value imputation, and normalization.

on the outcomes. Filtering can be based on count values or their relative abundance. Prevalence filtering, also known as singleton filtering, involves discarding features observed in fewer than a specific number of samples. This is often determined by a threshold value, typically set between 0.01 % and 0.1 %, to determine whether a feature is considered present in a sample (Table 6). The rationale behind this filtering is to remove non-informative OTUs or features that are present only in a few samples and may not contribute significantly to the biological or pathological process of interest. Similar techniques are used in microarray data analysis, such as those implemented in the R package genefilter [24]. In addition to quantitative OTU filtering, filtering can also be applied using qualitative criteria. For example, samples with low microbial biomass or potential contamination from extraction kits, reagents, laboratory, or clinical environments can be identified and removed. Taxa detected in negative controls can be considered as contaminants and be removed from the sample data. Tools such as Decontam [49] can help identify contaminating taxa that inversely correlate with sample DNA concentration. Taxa commonly found in commercially available kits and reagents should be removed [50]. Any taxa that show a strong correlation with these contaminating taxa or batch factors should be removed from downstream analysis.

Given advances in machine learning (ML) algorithms and models [51,52,53], steps like feature filtering and contamination detection of microbiome data could be improved by ML-enabled methods. However, the use of these methods for feature filtering, whether supervised or unsupervised, should be carefully evaluated. A recent study focusing on analyzing tumor-associated microbial signatures from The Cancer Genome Atlas (TCGA) data was criticized for employing biased supervised filtering and normalization methods, which introduced distinctive signatures into their findings [54,55].

Proper removal of low-quality samples or features is crucial in microbiome data analysis. However, the selection of threshold values or filtering methods should be undertaken with caution to avoid introducing biases and spurious findings. Sometimes, the choice of a filtering threshold is subjective, but it should be based on the study design and biological understanding of the microbial community. Solely relying on abundance levels for setting the threshold can result in exclusion of rare but relevant features. It is essential for researchers to critically evaluate the rationale behind their choice of filtering criteria, to prevent the inadvertent exclusion of important and informative features.

2) Batch Effect Correction

Batch effects are sources of variation that can arise from technical artefacts or confounding biological variables [56]. If not addressed, batch effects will introduce unwanted variations and confounding factors that align with the groups being compared, and consequently obscure true signals or lead to false positives. Microbial-associated variables of interest in case-control and longitudinal studies often confound with batch surrogates, either environmental [57] or host-related factors [58]. Therefore, it is important to detect and correct for batch factors during data pre-processing.

Batch effects can be minimized through proper study designs with balanced sample sizes across groups or batches and consistent

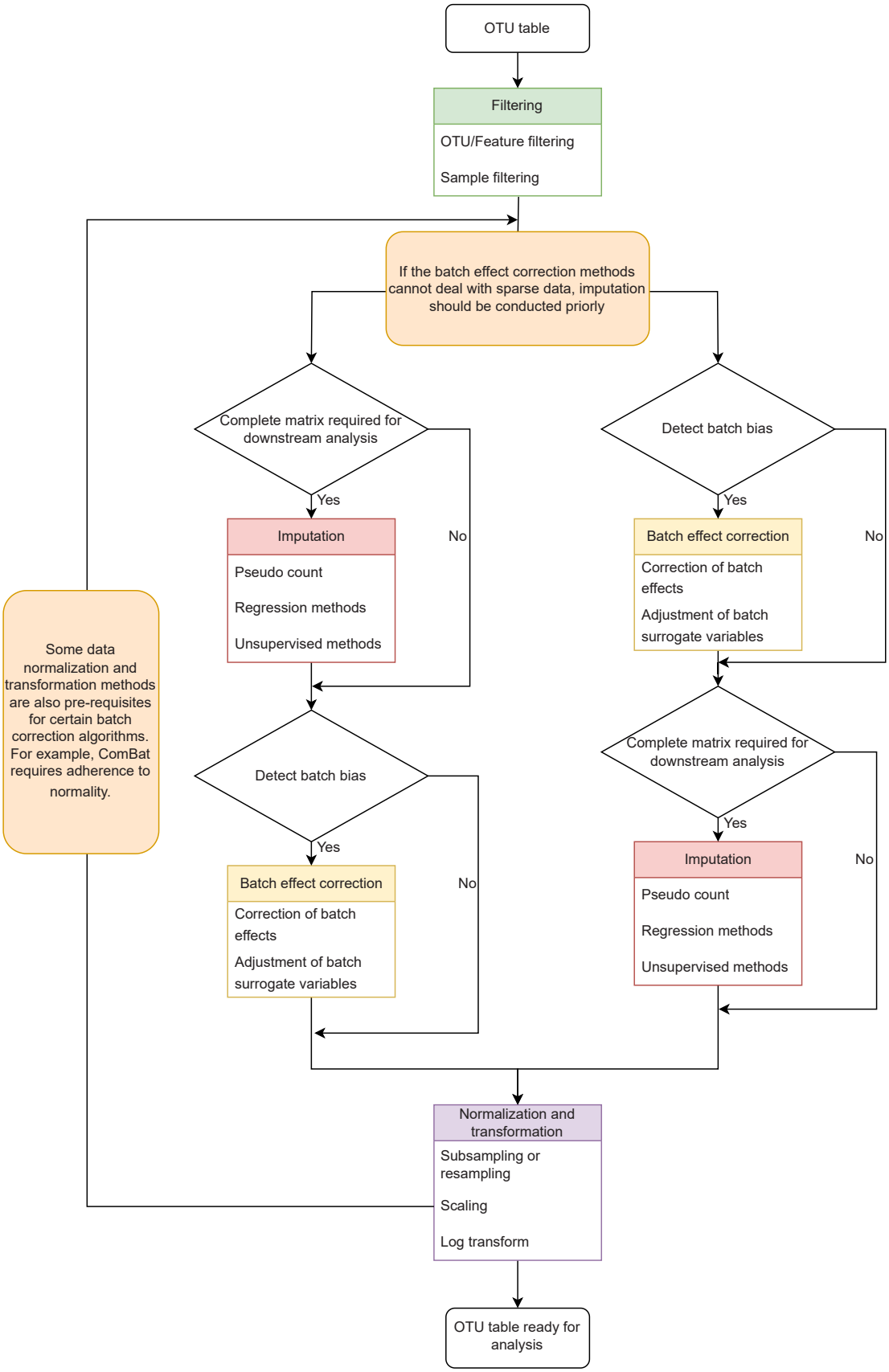

Fig. 2. Flow chart to summarize the workflow of data preprocessing Imputation sometimes need to be conducted before batch effect correction, as some batch effect correction methods cannot handle sparse data. Additionally, certain data normalization and transformation methods serve as prerequisites for specific batch correction algorithms. For instance, ComBat necessitates adherence to normality.

Table 1
Common data preprocessing steps of microbiome data.

Preprocessing Step	Description	Methods	Advantages	Limitation
Quality Control & Filtering	Removes low quality sequences and potential contaminants and filters out low-abundance OTUs and samples。	FastQC [21] Trimmomatic [22] Cutadapt [23] genefilter [24]	Reduces biases in the data and improves reliability of downstream analysis.	Might result in loss of some real but rare sequences.
Batch Effect Correction	Adjusts for systematic differences in data due to non-biological factors like different sequencing runs or different batches.	ComBat [25] Limma [26] DESeq2 [27] Bayesian Dirichlet-multinomial regression [28] SVA [29] Harmony [30] MNN [31] LIGER [32] PLSDA-batch [33] ConQuR [34]	Improves comparability across samples and reduces potential confounding.	May inadvertently remove some real biological differences if not carefully applied.
Imputation	Decides how to handle zero counts in the OTU table, imputing their likely values.	DESeq2 [27] phyloseq [35] k-NN [36] random forest [37] mbImpute [38]	Addresses the issue of missing data.	Different methods make different assumptions and may not always reflect reality.
Normalization	Adjusts for differing sequencing depth across samples.	Rarefaction ANCOM-BC [39] CSS [40] TMM [41] TSS	Makes samples directly comparable.	Rarefaction discards data and can reduce sensitivity. Other normalization methods might not be suitable for all data types.
Data Transformation	Converts count data to a different scale, often to meet assumptions of downstream statistical methods.	Log-ratio based transformation	Helps meet assumptions of downstream analysis and can reduce impact of high-abundance taxa.	Can distort data and mask real biological differences.

Table 2
Comparison of host transcriptomic and metagenomic data set.

	Host transcriptomic data	Metagenomic data
Matrix Type	Dense matrix with numerical values	Sparse matrix with large number of zeros
Typical Values	Continuous values representing gene expression levels	Counts data representing abundance of features in the samples
Value Characteristics	Values may vary widely in scale	Zero-inflated and overdispersion, usually convert to relative abundance
Nature of Data	Non-compositional data	Compositional data, values can be counts or relative abundance that sums up to a constant
Dimensionality	High: many genes, potentially fewer samples, around tens of thousands to over 100,000 different transcripts/splicing isoforms	High: many microbial species potentially fewer samples, depend on the number of bacterial species, around few thousands to tens of thousands
Challenges	Managing variation in scale across genes and samples, dealing with high dimensionality	Dealing with sparsity and zero-inflation, handling compositional nature of data, dealing with high dimensionality

experimental protocols such as sample processing [59]. However, even with these measures in place, exploratory analysis is still necessary to uncover hidden batch effects or covariates that may impact the effects of interest. Ordination techniques such as Principal Component Analysis (PCA) and clustering methods can be used to qualitatively diagnose the contribution of individual variables to the explained variance. However, these methods may not be directly applicable to microbiome data, due to unbalanced designs and confounding between batch effects and treatment effects [33].

Managing batch effects can be carried out by two strategies: correction of batch effects and adjustment of batch surrogate variables. The first correction strategy aims to remove batch effects from the raw data to generate corrected data for downstream analysis. Methods such as univariate ComBat [25], removeBatchEffect [26], percentile normalization [60], multivariate RUVIII [61], and SVD [62] can be employed for this purpose. These methods assume no interaction between batch effects and factors of interest. Some of these methods, such as ComBat and RUVIII, require the data to be complete, continuous, normally distributed and with negative controls across batches. The second adjustment strategy aims to estimate unknown batch variables and incorporate them as covariates in linear models. Methods include Surrogate Variable Analysis (SVA) [29], Remove Unwanted Variation (RUV) [63], and Bayesian Dirichlet-multinomial regression meta-analysis (BDMMA) [28] can be used for this approach. SVA identifies surrogate variables that represent unwanted variation, while RUV estimates unwanted variance with the inclusion of negative control across batches. Additionally, batch effect correction methods developed for single-cell RNA sequencing (scRNA-seq) data, such as Harmony [30], Mutual Nearest Neighbors (MNN) [31], and LIGER [32], can be adopted to microbiome data analysis with some modifications.

Many methods for batch effect correction were originally developed for transcriptomics and may not be fully optimized for microbiome data analysis. Improper use of these methods can introduce biases or artifacts [64,65], and may fail to effectively remove batch effects while confounding biological signals of interest. To address these challenges, several methods have been proposed. PLSDA-batch utilizes Partial Least Squares Discriminant Analysis (PLSDA) [33] and considers the sparsity of microbiome data by incorporating weighted group sizes to deal with unbalanced batches. ConQuR [34], based on a two-part quantile regression model, demonstrates improved performance in reducing batch variability compared to existing methods and has the potential to

Box 1
Batch effect correction methods.

ComBat works by estimating batch-specific means ($\widehat{\mu_{ik}}$) and variances ($\widehat{\sigma_{ik}}^2$) and then shrinks these estimates towards the overall mean ($\widehat{\mu_i}$) and variance ($\widehat{\sigma_i}^2$):

$$X_{ij}^{'} = \frac{X_{ij} - \widehat{\mu_{ik}}}{\widehat{\sigma_{ik}}} \times \widehat{\sigma_i} + \widehat{\mu_i}$$

SVA is based on the identification of surrogate variables that capture the unwanted variation in the data [29]

$$E = X \cdot B + S \cdot C + \varepsilon$$

where E is the expression matrix, with known covariate X and surrogate variable S, C is the matrix of coefficients for the surrogate variables.

Harmony uses a matrix factorization approach to estimate the bias and adjust the data by aligning all possible batch effects.

MNN identifies sample pairs that are nearest neighbors between different batches as these pairs are more likely to influence each other and cause batch effect. MNN can then estimate batch variations based on these pairs.

LINGER distinguishes shared factors and dataset-specific factors through integrative non-negative matrix factorization (iNMF). Clusters of samples are then normalized by these factor loadings.

extend its applicability to metagenomic data. While missing value imputation (MVI) algorithms are increasing being used in microbiome data analysis, recent works demonstrated that inappropriate use of MVI can obscure batch effects [66,67]. Therefore, a promising direction for the future of microbiome data analysis may involve integrating imputation and batch effect correction into a comprehensive package or pipeline, considering the inherent data sparsity. Similar to HarmonizR [68], incorporating appropriate missing value handling prior to batch effect correction could harness more reliable data for downstream analysis. Given that the factors contributing to batch effects in microbiome data are often multifaceted and interdependent, a thorough assessment of the batch variables prior to correction is crucial. (Table 3).

3) Imputation

Imputation is a critical technique used in different fields of data science analysis to estimate missing values. However, imputing missing values in microbiome data poses unique challenges due to the abundance of zeros resulting from the absence or low presence of certain microbiome taxa.

The prevalence of zeros in microbiome data presents several difficulties that can impact downstream analysis. Firstly, the large number of zeros can lead to information loss, as taxa may not be accurately detected due to limited sequencing depth. Secondly, the sparsity can introduce bias into statistical analyses, inflating the false-positive rate and reducing statistical power [69]. Lastly, the high sparsity can also make the results difficult to interpret. Therefore, it is crucial for researchers to carefully select appropriate statistical methods to handle this sparsity, such as zero-inflated negative binomial (ZINB) and zero-inflated Poisson (ZIP) distributions. These distributions can model the excess zeros in microbiome data. Despite these challenges, with careful and suitable analysis, valuable insights into the microbiome community and its roles in human diseases can still be obtained.

The zeros in microbiome data can be classified into two categories: biological and technical. Biological zeros represent the true absence of a taxon in a corresponding microbiome sample due to biological process. In contrast, technical zeros arise from limitations of measurement, such as insufficient sequencing depth [70]. The imputation strategy for microbiome data aims to differentiate between biological zeros and technical zeros and insert estimated values for the latter.

A traditional approach to handle missing values is to add a pseudo-count [71]. Although this strategy is straight-forward and simple, it can introduce bias. Imputation methods can be categorized as parametric and non-parametric. Parametric methods assume a specific distribution of data, such as a normal distribution that permits the use of regression modelling to predict missing values [40,72]. However, microbiome data is often not normally distributed, and the zero-inflated negative binomial (ZINB) distribution is more commonly used to account for the high number of zeros and the overdispersion. Parametric imputation can be implemented using tools like DESeq2 [27] and phyloseq [35]. However, parametric models may not be able to capture complex patterns in the data [73]. Non-parametric methods, on the other hand, make fewer assumptions and have been widely used in imputations, including k-Nearest Neighbors (k-NN) imputation [36] and random forest imputation [37] to predict missing values. Nevertheless, imputation on microbiome data has limitations in capturing the complexity of microbial interactions and non-linear relationships. Some imputation techniques may have unrealistic assumptions of data distribution. Additionally, imputation methods may fail to estimate missing values of rare taxa given limited information.

While imputation has historically been overlooked in microbiome data analysis, its importance in the data preprocessing pipeline is increasing. Although it is challenging due to the high data sparsity, specific imputation method for microbiome data has been developed [38]. The method aims to distinguish between true zeros and technical zeros by assuming different data distributions: Gamma distribution for the non-biological or technical zeros, and normal distribution for the taxonomical abundances including the true zeros with a presumed missing value rate. DeepMicroGen is another newly developed method to impute missing values in longitudinal microbiome studies [74]. It utilizes a generative adversarial network (GAN) model with recurrent neural network (RNN)-based components. When used appropriately, imputation can provide a more complete dataset of the microbiome (Table 4). Future studies should consider improving current imputation techniques or developing new methods specifically for microbiome data. This will enable researchers to better capture the richness of information that microbiome data holds.

4) Normalization

Microbiome data is inherently compositional, and the application of standard statistical methods to such data can yield unreliable results

Table 3
Common methods for batch effect correction.

Method	Original Data Type	Theoretical Framework	Details on Theoretical Framework	Advantages	Limitation	R Package	Reference
ComBat	None	Empirical Bayes methods, location and scale adjustments	Assume that data is normal-like distributed. Bayesian framework borrows information from all the OTUs/features to estimate the batch effect parameters, which is beneficial when the sample size is small.	Can correct batch effects in presence of confounding variables, widely used, effective	Potential overcorrection for large sample sizes, does not perform well with sparse count data	sva	[25]
Limma	Transcriptomics	Linear modeling, empirical Bayes statistics	Fit linear models to expression data. Each OTU/feature is individually modeled, but information is borrowed across features using empirical Bayes methods. Require log-transformed to ensure a normal-like distribution	Provides robustness to outliers, uses empirical Bayes moderation, powerful statistical framework	Less appropriate for zero-inflated datasets, such as single-cell RNA-seq	limma	[26]
DESeq2	Transcriptomics	Negative binomial distribution, Wald significance tests	Negative binomial distribution captures the over-dispersion observed in microbiome count data.	Appropriate for count data, estimates variance-mean dependence, controls false discovery rate	Not designed specifically for batch effect correction, potential for over-dispersion	DESeq2	[27]
Bayesian Dirichlet-multinomial regression	Microbiome data	Dirichlet-multinomial regression model and adopt the Bayesian framework	Investigate the impact of over-dispersion on the association detection performance of BDMMA by varying the degree of baseline dispersion. BDMMA shows better performance than the other methods.	Reduce the number of false discoveries Able to identify the small set of taxa that are truly associated with the phenotypes with very low false discovery rates	Require the batch information to be known for all the samples.	BDMMA	[28]
SVA	Transcriptomics	Surrogate Variable Analysis, data decomposition	Assume that the most substantial component of variability is not explained by the model (i.e., the SVs). They represent unwanted or technical variation.	Corrects for known and unknown confounding variables, flexible application. Design for the high-dimensional data	Less straightforward to use, not ideal for zero-inflated data	sva	[29]
Harmony	Transcriptomics	Iterative harmonic alignment, dimensionality reduction	The principle is that within each batch, OTUs can be grouped into clusters, and these clusters should be similar across batches. By adjusting the centers of these clusters across batches, Harmony corrects for batch effects.	Intended for high-dimensional data, allows integration of multiple datasets, good performance with single-cell data	Requires high computational resources, not ideal for low-dimensional data	harmony	[30]
MNN	Single-cell transcriptomics	k-nearest neighbors, data alignment	For each OTU/feature in one batch, the algorithm finds its nearest neighbors in another batch. Assume that batch effects are additive, half of the difference of these pairs are considered as an estimate of the batch effect.	Effective for single-cell RNA-seq data, can identify mutual nearest neighbors across batches	Might not perform well with large differences between batches, requires good preprocessing	scran, scater	[31]
LIGER	Single-cell transcriptomics	Nonnegative Matrix Factorization, clustering	Use non-negative matrix factorization (NMF) to decompose each dataset into two matrices: one representing shared and unique factors (W matrix) and another representing cell or sample loadings for these factors (H matrix).	Allows joint analysis of multiple datasets, can identify shared and dataset-specific factors	Relatively complex to use, requires careful parameter tuning, high computational demand	LIGER	[32]
PLSDA-batch	Microbiome data	Partial Least Squares Discriminant Analysis	A relaxed assumption about data distribution and thus is more suitable for microbiome data	Preserve treatment variation; include group size weight to handle unbalanced batch × treatment designs; include variable selection	Require pre-defined batch group information; constructed based on a linear combination of variables	PLSDAbatch	[33]

(*continued on next page*)

Table 3 (*continued*)

Method	Original Data Type	Theoretical Framework	Details on Theoretical Framework	Advantages	Limitation	R Package	Reference
				when estimating treatment components			
ConQuR	Microbiome data	Non-parametric modeling with two-party quantile regression model	A logistic model determines the likelihood of the taxon's presence, and quantile regression model's percentiles of the read count distribution given the taxon is present.	Account for zero-inflated, over-dispersed microbiome data	Not working if batch completely confounds the key variable	ConQuR	[34]

Table 4
Methods for imputation of missing values in microbiome data.

Method	Theoretical Framework	Advantages	Limitation	R package	Reference
Pseudo count	Replace the missing value with a pseudo count or small constant	Easy to implement and interpretation	Introduce biases	None	None
Average/median of the feature	Compute the average or median to replace the missing value	Easy to implement and interpretation	Introduce biases	None	None
mbImpute	After distinguishing biological zeros and technical zeros, impute technical zeros with statistical modeling	Specifically designed for microbiome data. Empowered DESeq2-phyloseq has better performance in selecting predictive taxa for disease conditions	Highly depend on the assumption of data distribution. Some real-world datasets may not fulfil the requirement	mbImpute	[38]
DeepMicroGen	Recurrent neural network-based GAN model	Specifically designed for microbiome data. Discriminator can differentiate the actual and the imputed values from the generator and predict the timepoint of each sample	Assume that the samples are generated with the same time intervals. If subjects are sampled at irregular time intervals, DeepMicroGen may not perform well.	DeepMicroGen	[74]

[11]. To address this challenge, normalization techniques are employed in microbiome data to transform the data to a common and comparable scale. Normalization is important due to over-dispersion and uneven sampling depths, and it should be performed prior to downstream analysis of microbiome data [75]. The effectiveness of normalization can be influenced by different factors, including the choice of normalization methods and sample size. The presence of rare OTUs can also impact the selection normalization method, as some methods are sensitive to low-abundance taxa. Therefore, it is critical to carefully select appropriate method that suits the specific dataset.

Among the normalization methods used in microbiome research, rarefaction is the most employed. Samples with different sequencing depths are subsampled to the same depth based on the minimum sequencing depth across samples, or by evaluating rarefaction curves [76]. To avoid under-sampling, it is important to remove samples with low library sizes prior to rarefaction. Scaling is another simple and commonly used normalization method. It involves multiplying each count by a ratio determined based on a quantile of the data. Commonly used scaling methods include Cumulative-sum scaling (CSS) [40], Trimmed Mean of M-value (TMM) [41], or Total-Sum Scaling (TSS). Since microbiome data often contains many zeros, different approaches are used to handle these zeros. Some researchers choose to retain the zeros [77], while some others assign a small value to all the zeros [78]. Both approaches are part of the standard pipeline for analyzing microbiome data.

Several log-ratio based normalization methods have been proposed for microbiome data. These include log-transformations such as additive log-ratio (ALR), isometric log-ratio (ILR) and centered log-ratio transformation (CLR). Quantile normalization [79] is another useful technique that involves adjusting the distribution of each sample's data to match a common reference distribution. Given a matrix X with columns as samples and rows as OTUs/features, the quantile normalization proceeds as follows:

a. Rank the data in each column.
b. For each rank, calculate the average value across all columns.
c. Replace the data in each column with the average values corresponding to the rank from step b.

If the data distribution is not considered, normalization methods can potentially distort the relationships between samples and features. Therefore, some novel methods like Analysis of Compositions of Microbiome with Bias Correction (ANCOM-BC) were specifically developed for more complex compositional data [39]. ANCOM-BC focuses on differential abundance analysis and incorporates a bias correction term in linear regression to account for biases caused by different library size. This approach is similar to the log-ratio transformation employed in other methods like ANCOM[71] and DESeq2 [27] for analyzing compositional data.

Normalization is a standard pre-processing step for various types of data, including transcriptome and microbiome data. However, the high sparsity observed in abundance matrices can pose challenges for some normalization approaches. For instance, log-transformations including CLR, ALR and ILR cannot be directly applied to zero counts. There are several ways to address this issue, such as adding a small constant to all counts (pseudo-count) as mentioned in imputation methods. Consequently, it is important to select an appropriate normalization method that can account for different library sizes.

Normalization methods usually assume that the factors causing variations in sequencing depth or technical biases have the same impact among all samples. The normalization process aims to uniformly transforms the data across all samples. However, in certain cases, this assumption may not hold true. The normalization process may not capture different types of biases among samples. Therefore, it is important to carefully consider the selection of normalization methods that better accommodate the characteristics of the microbiome data and provide a more accurate representation of the true biological diversity within the samples (Table 5).

5) Data integration in microbiome data

Box 2
Log-ratio normalization methods formula.

Given a sample with different features: $x = (x_1, x_2, \ldots, x_n)$

$$\mathrm{ALR(x)} = \left[\log\left(\frac{x_1}{x_n}\right), \log\left(\frac{x_2}{x_n}\right), \ldots, \log\left(\frac{x_{n-1}}{x_n}\right)\right]$$

Where x_n is chosen as the reference part.

In ILR, we first need to divide the composition into two groups: numerator and denominator groups: $(x_1, \ldots, x_c) and (x_{c+1}, \ldots, x_n)$.

$$ILR(x) = \sqrt{\frac{d1}{d1+d2}}\log\left(\frac{\sqrt[d1]{x1 \ldots x_c}}{\sqrt[d2]{x_{c+1} \ldots x_n}}\right)$$

Where $d1$, $d2$ are the number of elements in the numerator and denominator groups respectively.

CLR transformation takes the logarithm of each element divided by the geometric mean of all elements. For a composition vector x with positive components $[x_1, x_2, \ldots., x_n]$.

$$CLR(x_i) = \left[\log\left(\frac{x_1}{\sqrt[n]{(x_1 x_2 \ldots x_n)}}\right), \log\left(\frac{x_2}{\sqrt[n]{(x_1 x_2 \ldots x_n)}}\right), \ldots, \log\left(\frac{x_n}{\sqrt[n]{(x_1 x_2 \ldots x_n)}}\right)\right]$$

Table 5
Common methods for normalization and data transformation.

Method	Theoretical Framework	Advantages	Limitation	R package	Reference
Rarefaction	Random sampling	Simple, easy to implement	Loss of data, increased false negatives	phyloseq	[35]
Cumulative-sum scaling	Statistical modeling	Robust to various library sizes	Sensitive to outliers, may introduce bias	metagenomeSeq	[40]
Trimmed Mean of M-value	Linear scaling	Minimizes log-fold changes between samples	Requires counts to be of similar distribution	edgeR	[41]
Total-Sum Scaling	Constant sum scaling	Easy to interpret	Sensitive to outliers	phyloseq	[35]
Additive log-ratio	Log-ratio based transformation	Can reveal relative changes between features	Require selection of denominator component and may cause the loss of one dimension	compositions	[80]
Isometric log-ratio	Log-ratio based transformation	Removes the unit-sum constraint	Computationally intense and harder to interpret	compositions	[80]
Centered log-ratio	Log-ratio based transformation	Centered at zeros	Highly sensitive to zeros	compositions	[80]
ANCOM-BC	Bayesian framework	Can deal with zero-inflated data	Requires sufficient sample size	ANCOMBC	[39]

Data integration involves merging data from multiple sources to gain a more comprehensive understanding of datasets. In the context of microbiome studies, where different platforms, techniques, and studies are often used, effective data integration is crucial. Batch effect correction and data transformation are important steps to harmonize the datasets. Batch effect correction identifies technical variations to ensure comparability between datasets, while data transformation aligns datasets from different studies to a common scale. Microbiome data presents statistical analysis challenges due to zero inflation and the compositional nature of abundance data. Transformations are applied to modify the data structure, making it more suitable for statistical tests and ensuring the validity of the analyses.

There are numerous methods available for data integration. Meta-analysis is a robust approach for pooling results from multiple studies. There are R packages that allow for the meta-analysis of microbiome data from different studies, combining effect sizes to determine overall patterns. MetamicrobiomeR employs the Generalized Additive Models for location, scale and shape using a zero-inflated beta family to analyze microbiome relative abundance datasets [81]. MANTA (Microbiota And phenotype correlation Analysis platform) is a data integration platform that allows users to test the correlation between the microbiome abundance and other phenotypic variables, including dietary habits and lifestyle parameters [82]. NetMoss can minimize batch effects and identify strong biomarkers that may be overlooked by abundance-based methods [83]. Instead of combining the raw data, another approach is to analyze the summarized results from other studies. MMUPHin (Meta--Analysis Methods with a Uniform Pipeline for Heterogeneity in microbiome studies) incorporates batch and study effect correction, meta-analyzed differential abundance testing and population structure discovery [84].

In addition to discussing the combination of data from different sources, we have reviewed various research papers focused on microbiome studies (Table 6). These papers have adopted different preprocessing methods in their analysis of microbiome data. While filtering and normalization are consistently performed, there is variability in the specific techniques used. Imputation is under-utilized, possibly due to the absence of reliable and unbiased techniques. Moreover, there is a growing trend in adopting integrated pipeline packages capable of performing multiple preprocessing steps. These observations highlight the need to further optimize and standardize preprocessing methodology for microbiome studies.

3. Conclusion

Issues of over-dispersal, high sparsity, and high dimensionality in microbiome data pose significant challenges for downstream analyses. Failing to address these challenges can result in misleading outcomes. In this paper, we examined several common approaches for data preprocessing. These play a crucial role in mitigating biases and improving the interpretability of microbiome data. However, it is important to acknowledge that they have limitations. Therefore, future research should prioritize improving existing methods and exploring novel approaches that can effectively handle the unique characteristics of microbiome data.

Table 6
Different combinations of preprocessing steps in some published studies.

Study Design	Filtering	Batch effect correction	Imputation	Normalization
Perspective study; Metagenomics of fecal samples from 94 melanoma patients treated with anti-PD-1 [85]	Dataset was filtered to include at least 5000 reads per sample	ComBat	Pseudo count	Log transformed and quantile normalized
Randomized controlled trial; metagenomic and 16s rRNA sequencing of fecal and mucosal microbiome from subjects with or without probiotics [86]	16s: OTUs with relative abundance <0.1 % Metagenomics: fecal samples with less than 500k and mucosal samples with less than 100k bacterial reads were removed	None	None	16s: rarefied to a depth of 10,000 reads All samples: normalized to baseline
Re-analysis of whole-genome and whole-transcriptome sequencing data from The Cancer Genome Atlas (TCGA) to identify the patterns of tumor-associated microbial signatures [54]	Samples with missing metadata (ethnicity, ICD10 codes, DNA/RNA analyte amounts, or FFPE status) OTUs likely to be contaminants were removed	Supervised normalization correction (SNM)	None	Log transformation of detected microbial read counts into count per million using the Voom algorithm Quantile normalization for SNM correction
Meta-analysis of published studies on inflammatory bowel disease [84]	None	Combat (MMUPHin)	Pseudo count	Quantile normalized
Randomized study; 16s rRNA sequencing of gut microbiome from infants with early-onset neonatal sepsis and healthy infants from two separate sites [87]	Abundance-based filtering selects OTUs present at a confident level of detection (0.1 % relative abundance) in at least two samples	None	None	Scaling: To calculate counts per million in order to normalize read counts for library size, counts were divided by a normalization factor of the library size divided by 1000,000.
Randomized, double-blind, placebo-controlled clinical trial; 16s rRNA sequencing of gut microbiome from 169 participants received probiotics *Lactobacillus rhamnosus* or placabo [88]	Quality filtering with DADA2	None	None	Total Sum Scaling (TSS) normalization and Centered log-ratio (CLR) transformation.
Randomized controlled trial; 16s rRNA sequencing of fecal microbiome from participants randomized to a high protein diet or normal protein diet [89]	Amplicon sequence variants were filtered if not present in at least 15 % of all samples.	DESeq2	DESeq2	DESeq2
Clinical trial; 16s rRNA sequencing of infant stool samples [90]	Keep taxa with 5 counts in a minimum of 5 % of samples	phyloseq	phyloseq	phyloseq: Total Sum Scaled (TSS)
Randomized, double-blind, placebo-controlled crossover intervention trial; 16s rRNA sequencing of fecal microbiome from participants underwent supplemental bacteriophage antimicrobial treatment [91]	Phylotypes with prevalence less than 75 % were removed	None	None	Rarefaction
Case-controlled study; 16s rRNA sequencing of fecal microbiome from normal controls and patients with dermatitis [92]	DADA2	Multivariate linear regression was carried out to adjust the clinical variables and batch effects.	mbDenoise and mbImpute	Centered log-ratio transformation (CLR)
Perspective study; metagenomic sequencing of skin microbiome from patients with atopic dermatitis [93]	Quality-filtered reads (median of 3 million reads per metagenome)	PLSDA-batch	None	Trimmed Mean of M-value (TMM)

In recent years, computational biology has placed greater emphasis on addressing key challenges such as data imputation and batch effect correction. Robust batch effect correction methods are essential when integrating data from multiple sources. Continuous efforts are being made to develop new methodologies that can effectively tackle these issues. By employing appropriate methods for data pre-processing, the analysis of microbiome data will enable researchers to obtain a more accurate and comprehensive understanding of the intricate role of microbiome in human health and disease.

Funding

The authors would like to acknowledge the NTU Start Up Grant (021337–00001, 021281–00001), Centre for Microbiome Medicine, Wang Lee Wah Memorial Fund, Singapore National Supercomputing Centre (12002587 / 12003632), Singapore Ministry of Education (MOE) Tier 1 Academic Research Fund (RG37/22), the National Research Foundation Singapore under its Clinician Scientist Individual Research Grant (CIRG23jan-0004 / MOH-001353) administrated by the Singapore Ministry of Health's National Medical Research Council (MOH-NMRC) for support of this work.

CRediT authorship contribution statement

Ruwen Zhou: Writing - original draft, Writing - review & editing, Data curation, Conceptualization. **Siu Kin Ng**: Writing - review & editing, Conceptualization. **Joseph Jao Yiu Sung**: Writing - review & editing, Conceptualization. **Wilson Wen Bin Goh**: Writing - review & editing, Supervision, Conceptualization. **Sunny Hei Wong**: Writing - review & editing, Supervision, Conceptualization, Funding acquisition.

Conflict of Interest

The authors of this manuscript declare no conflict of interests.

Acknowledgement

This research is supported by the NTU Start Up Grant (021337–00001, 021281–00001), Centre for Microbiome Medicine, Wang Lee Wah Memorial Fund, Singapore National Supercomputing Centre (12002587 / 12003632), Singapore Ministry of Education (MOE) Tier 1 Academic Research Fund (RG37/22), the National Research Foundation Singapore under its Clinician Scientist Individual Research Grant (CIRG23jan-0004 / MOH-001353) administrated by the Singapore Ministry of Health's National Medical Research Council

(MOH-NMRC).

References

[1] Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med 2018;24(4):392–400. https://doi.org/10.1038/nm.4517.

[2] Clemente JC, Ursell LK, Parfrey LW, Knight R. The impact of the gut microbiota on human health: an integrative view. Cell 2012;148(6):1258–70. https://doi.org/10.1016/j.cell.2012.01.035.

[3] Rebersek M. Gut microbiome and its role in colorectal cancer. BMC Cancer 2021;21(1):1325. https://doi.org/10.1186/s12885-021-09054-2.

[4] Ren L, Ye J, Zhao B, Sun J, Cao P, Yang Y. The role of intestinal microbiota in colorectal cancer. Front Pharmacol 2021;12:674807. https://doi.org/10.3389/fphar.2021.674807.

[5] Glenn TC. Field guide to next-generation DNA sequencers: field guide to next-gen sequencers. Mol Ecol Resour 2011;11(5):759–69. https://doi.org/10.1111/j.1755-0998.2011.03024.x.

[6] Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinforma 2016;17(1):125. https://doi.org/10.1186/s12859-016-0976-y.

[7] Johnson JS, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun 2019;10(1):5029. https://doi.org/10.1038/s41467-019-13036-1.

[8] Gill SR, et al. Metagenomic analysis of the human distal gut microbiome. Science 2006;312(5778):1355–9. https://doi.org/10.1126/science.1124234.

[9] Schoch CL, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for *Fungi*". Proc Natl Acad Sci 2012;109(16):6241–6. https://doi.org/10.1073/pnas.1117018109.

[10] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17(6):333–51. https://doi.org/10.1038/nrg.2016.49.

[11] Weiss S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 2017;5(1):27. https://doi.org/10.1186/s40168-017-0237-y.

[12] Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 2016;14(8):e1002533. https://doi.org/10.1371/journal.pbio.1002533.

[13] Weiss S, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 2017;5(1):27. https://doi.org/10.1186/s40168-017-0237-y.

[14] Pasolli E, et al. Accessible, curated metagenomic data Through ExperimentHub. Nat Methods 2017;14(11):1023–4. https://doi.org/10.1038/nmeth.4468.

[15] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLOS Comput Biol 2015;11(5):e1004226. https://doi.org/10.1371/journal.pcbi.1004226.

[16] Shakya M, Lo C-C, Chain PSG. Advances and challenges in metatranscriptomic analysis. Front Genet 2019;10:904. https://doi.org/10.3389/fgene.2019.00904.

[17] Heintz-Buschart A, Wilmes P. Human gut microbiome: function matters. Trends Microbiol 2018;26(7):563–74. https://doi.org/10.1016/j.tim.2017.11.002.

[18] Zhang Y, et al. Metatranscriptomics for the human microbiome and microbial community functional profiling. Annu Rev Biomed Data Sci 2021;4(1):279–311. https://doi.org/10.1146/annurev-biodatasci-031121-103035.

[19] Long S, et al. Metaproteomics characterizes human gut microbiome function in colorectal cancer. NPJ Biofilms Micro 2020;6(1):14. https://doi.org/10.1038/s41522-020-0123-4.

[20] Abu-Ali GS, et al. Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat Microbiol 2018;3(3):356–66. https://doi.org/10.1038/s41564-017-0084-4.

[21] S. Andrews, "FastQC." in FastQC: a quality control tool for high throughput sequence data. [Online]. Available: ⟨https://www.bioinformatics.babraham.ac.uk/projects/fastqc/⟩.

[22] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.

[23] M. Martin, Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads, doi: https://doi.org/10.14806/ej.17.1.200.

[24] R. Gentleman, V. Carey, W. Huber, and F. Hahne, genefilter. in genefilter: methods for filtering genes from high-throughput experiments. 2023.

[25] Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinforma 2020;2(3):lqaa078. https://doi.org/10.1093/nargab/lqaa078.

[26] Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. e47–e47 Nucleic Acids Res 2015;43(7). https://doi.org/10.1093/nar/gkv007.

[27] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

[28] Dai Z, Wong SH, Yu J, Wei Y. Batch effects correction for microbiome data with Dirichlet-multinomial regression. Bioinformatics 2019;35(5):807–14. https://doi.org/10.1093/bioinformatics/bty729.

[29] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;28(6):882–3. https://doi.org/10.1093/bioinformatics/bts034.

[30] Korsunsky I, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019;16(12):1289–96. https://doi.org/10.1038/s41592-019-0619-0.

[31] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36(5):421–7. https://doi.org/10.1038/nbt.4091.

[32] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. e17 Cell 2019;177(7):1873–87. https://doi.org/10.1016/j.cell.2019.05.006.

[33] Wang Y, Cao K-ALê. PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data. Brief Bioinform 2023;24(2):bbac622. https://doi.org/10.1093/bib/bbac622.

[34] Ling W, et al. Batch effects removal for microbiome data via conditional quantile regression. Nat Commun 2022;13(1):5418. https://doi.org/10.1038/s41467-022-33071-9.

[35] McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 2013;8(4):e61217. https://doi.org/10.1371/journal.pone.0061217.

[36] Liao SG, et al. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinforma 2014;15(1):346. https://doi.org/10.1186/s12859-014-0346-6.

[37] Moritz S, Bartz-Beielstein T. imputeTS: Time series missing value imputation in R. R J 2017;9(1):207. https://doi.org/10.32614/RJ-2017-009.

[38] Jiang R, Li WV, Li JJ. mbImpute: an accurate and robust imputation method for microbiome data. Genome Biol 2021;22(1):192. https://doi.org/10.1186/s13059-021-02400-4.

[39] Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nat Commun 2020;11(1):3514. https://doi.org/10.1038/s41467-020-17041-7.

[40] Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 2013;10(12):1200–2. https://doi.org/10.1038/nmeth.2658.

[41] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.

[42] Bolyen E, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019;37(8):852–7. https://doi.org/10.1038/s41587-019-0209-9.

[43] Schloss PD, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75(23):7537–41. https://doi.org/10.1128/AEM.01541-09.

[44] Blanco-Míguez A, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol 2023. https://doi.org/10.1038/s41587-023-01688-w.

[45] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007;17(3):377–86. https://doi.org/10.1101/gr.5969107.

[46] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46.

[47] Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;7(1):11257. https://doi.org/10.1038/ncomms11257.

[48] Bokulich NA, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 2013;10(1):57–9. https://doi.org/10.1038/nmeth.2276.

[49] Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 2018;6(1):226. https://doi.org/10.1186/s40168-018-0605-2.

[50] Salter SJ, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12(1):87. https://doi.org/10.1186/s12915-014-0087-z.

[51] A.L. Bluma, "Selection of relevant features and examples in machine".

[52] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97(1–2):273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.

[53] Dormann CF, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 2013;36(1):27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

[54] Poore GD, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 2020;579(7800):567–74. https://doi.org/10.1038/s41586-020-2095-1.

[55] Gihawi A, et al. Major data analysis errors invalidate cancer microbiome findings. Cancer Biol Prepr 2023. https://doi.org/10.1101/2023.07.28.550993.

[56] Schloss PD. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. mBio 2018;9(3):e00525–18. https://doi.org/10.1128/mBio.00525-18.

[57] Wang Y, LêCao K-A. Managing batch effects in microbiome data. Brief Bioinform 2020;21(6):1954–70. https://doi.org/10.1093/bib/bbz105.

[58] Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. Nature 2020;587(7834):448–54. https://doi.org/10.1038/s41586-020-2881-9.

[59] Leek JT, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11(10):733–9. https://doi.org/10.1038/nrg2825.

[60] Gibbons SM, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome studies. PLoS Comput Biol 2018;14(4):e1006102. https://doi.org/10.1371/journal.pcbi.1006102.
[61] Jacob L, Gagnon-Bartsch JA, Speed TP. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Biostatistics 2016;17(1):16–28. https://doi.org/10.1093/biostatistics/kxv026.
[62] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci 2000;97(18):10101–6. https://doi.org/10.1073/pnas.97.18.10101.
[63] J.A. Gagnon-Bartsch, L. Jacob, T.P. Speed, "Removing Unwanted Variation from High Dimensional Data with Negative Controls".
[64] Goh WWB, Yong CH, Wong L. Are batch effects still relevant in the age of big data? Trends Biotechnol 2022;40(9):1029–40. https://doi.org/10.1016/j.tibtech.2022.02.005.
[65] Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol 2017;35(6):498–507. https://doi.org/10.1016/j.tibtech.2017.02.012.
[66] Goh WWB, Hui HWH, Wong L. How missing value imputation is confounded with batch effects and what you can do about it. Drug Discov Today 2023;28(9):103661. https://doi.org/10.1016/j.drudis.2023.103661.
[67] Hui HWH, Kong W, Peng H, Goh WWB. The importance of batch sensitization in missing value imputation. Sci Rep 2023;13(1):3003. https://doi.org/10.1038/s41598-023-30084-2.
[68] Voß H, et al. HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. Nat Commun 2022;13(1):3523. https://doi.org/10.1038/s41467-022-31007-x.
[69] Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol 2022;23(1):31. https://doi.org/10.1186/s13059-022-02601-5.
[70] Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol 2017;8:10.
[71] Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis 2015;26(0). https://doi.org/10.3402/mehd.v26.27663.
[72] Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics 2016;32(17):2611–7. https://doi.org/10.1093/bioinformatics/btw308.
[73] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc Ser A Stat Soc 1995;158(3):419. https://doi.org/10.2307/2983440.
[74] Choi JM, Ji M, Watson LT, Zhang L. DeepMicroGen: a generative adversarial network-based method for longitudinal microbiome data imputation. Bioinformatics 2023;39(5):btad286. https://doi.org/10.1093/bioinformatics/btad286.
[75] Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. NPJ Biofilms Micro 2020;6(1):60. https://doi.org/10.1038/s41522-020-00160-w.
[76] Gotelli N, Colwell R. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 2001;4:379–91. https://doi.org/10.1046/j.1461-0248.2001.00230.x.
[77] McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: an ecological perspective. Methods Ecol Evol 2019;10(3):389–400. https://doi.org/10.1111/2041-210X.13115.
[78] Korthauer K, et al. A practical guide to methods controlling false discoveries in computational biology. Genome Biol 2019;20(1):118. https://doi.org/10.1186/s13059-019-1716-1.
[79] Townes FW, Irizarry RA. Quantile normalization of single-cell RNA-seq read counts without unique molecular identifiers. Genome Biol 2020;21(1):160. https://doi.org/10.1186/s13059-020-02078-0.
[80] Van Den Boogaart KG, Tolosana-Delgado R. 'compositions': a unified R package to analyze compositional data. Comput Geosci 2008;34(4):320–38. https://doi.org/10.1016/j.cageo.2006.11.017.
[81] Ho NT, Li F, Wang S, Kuhn L. metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models. BMC Bioinforma 2019;20(1):188. https://doi.org/10.1186/s12859-019-2744-2.
[82] Chen Y-A, et al. MANTA, an integrative database and analysis platform that relates microbiome and phenotypic data. PLoS One 2020;15(12):e0243609. https://doi.org/10.1371/journal.pone.0243609.
[83] Xiao L, Zhang F, Zhao F. Large-scale microbiome data integration enables robust biomarker identification. Nat Comput Sci 2022;2(5):307–16. https://doi.org/10.1038/s43588-022-00247-8.
[84] Ma S, et al. Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. Genome Biol 2022;23(1):208. https://doi.org/10.1186/s13059-022-02753-4.
[85] McCulloch JA, et al. Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. Nat Med 2022;28(3):545–56. https://doi.org/10.1038/s41591-022-01698-2.
[86] Zmora N, et al. Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. e21 Cell 2018;174(6):1388–405. https://doi.org/10.1016/j.cell.2018.08.041.
[87] Reyman M, et al. Effects of early-life antibiotics on the developing infant gut microbiome and resistome: a randomized trial. Nat Commun 2022;13(1):893. https://doi.org/10.1038/s41467-022-28525-z.
[88] Aljumaah MR, Bhatia U, Roach J, Gunstad J, Azcarate Peril MA. The gut microbiome, mild cognitive impairment, and probiotics: a randomized clinical trial in middle-aged and older adults. Clin Nutr 2022;41(11):2565–76. https://doi.org/10.1016/j.clnu.2022.09.012.
[89] Dong TS, et al. A high protein calorie restriction diet alters the gut microbiome in obesity. Nutrients 2020;12(10):3221. https://doi.org/10.3390/nu12103221.
[90] Gilley SP, et al. Associations between maternal obesity and offspring gut microbiome in the first year of life. Pediatr Obes 2022;17(9):e12921. https://doi.org/10.1111/ijpo.12921.
[91] Febvre H, et al. PHAGE study: effects of supplemental bacteriophage intake on inflammation and gut microbiota in healthy adults. Nutrients 2019;11(3):666. https://doi.org/10.3390/nu11030666.
[92] Wang Y, et al. Unique gut microbiome signatures among adult patients with moderate to severe atopic dermatitis in Southern Chinese. Int J Mol Sci . 2023;vol. 24(16):12856. https://doi.org/10.3390/ijms241612856.
[93] Saheb Kashaf S, et al. Staphylococcal diversity in atopic dermatitis from an individual to a global scale. e6 Cell Host Microbe 2023;31(4):578–92. https://doi.org/10.1016/j.chom.2023.03.010.