OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbaf051
https://doi.org/10.1093/bib/bbaf051
Review

Semisynthetic simulation for microbiome data analysis

Kris Sankaran[1], Saritha Kodikara[2], Jingyi Jessica Li[3,4,5], Kim-Anh Lê Cao[2]

[1]Department of Statistics, University of Wisconsin-Madison, 1300 University Ave, Madison,WI 53703, United States
[2]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Building 184/30 Royal Parade, Melbourne, VIC 3052, Australia
[3]Department of Statistics and Data Science, University of California, Los Angeles, 520 Portola Plaza, Los Angeles, CA 90095, United States
[4]Department of Human Genetics, University of California, Los Angeles, 695 Charles E Young Dr S, Los Angeles, CA 90095, United States
[5]Department of Biostatistics, University of California, Los Angeles, 650 Charles E. Young Dr S, Los Angeles, CA 90095, United States

*Corresponding author. Kris Sankaran. E-mail: ksankaran@wisc.edu

Abstract

High-throughput sequencing data lie at the heart of modern microbiome research. Effective analysis of these data requires careful preprocessing, modeling, and interpretation to detect subtle signals and avoid spurious associations. In this review, we discuss how simulation can serve as a sandbox to test candidate approaches, creating a setting that mimics real data while providing ground truth. This is particularly valuable for power analysis, methods benchmarking, and reliability analysis. We explain the probability, multivariate analysis, and regression concepts behind modern simulators and how different implementations make trade-offs between generality, faithfulness, and controllability. Recognizing that all simulators only approximate reality, we review methods to evaluate how accurately they reflect key properties. We also present case studies demonstrating the value of simulation in differential abundance testing, dimensionality reduction, network analysis, and data integration. Code for these examples is available in an online tutorial (https://go.wisc.edu/8994yz) that can be easily adapted to new problem settings.

Keywords: simulation; microbiome; power analysis; methods selection; methods assessment

Introduction

Microbial communities play a central role in human and ecological health. Advances in sequencing technology and statistical methods have made it possible to characterize these communities at unprecedented levels of precision [1–3]. However, statistical methods have to contend with microbiome-specific challenges, like sparse read coverage, large library size variations, uncertainties in the resolved taxa, and overdispersion [4–6]. These difficulties intensify as scientific goals grow more complex. This is especially evident as microbiome data analysis shifts away from descriptive studies toward modeling that often requires experimental designs with multiple batches, longitudinal sampling, or complementary assays [7–9]. In this context, effective use of statistical methods hinges on many steps: framing precise questions, preparing suitable data, applying appropriate methods, and delivering accurate interpretations.

Given the uncertainties present in real data, guiding analysis using simulations where the ground truth is known can be tremendously helpful. Simulation has a long history in shaping microbiome data analysis. For example, McMurdie and Holmes [4] applied rarefaction to datasets simulated from a negative binomial model to clarify its impact on downstream inferences. Similarly, Kodikara et al. [10] employed an autoregressive model to assess method performance for longitudinal studies. In fact, most methodological research requires benchmarking on simulated data. Despite these strong precedents, the field is only beginning to formalize, evaluate, and share reusable simulators. In particular, recent advances have proposed simulators that leverage existing *template data*—real experimental data whose patterns the simulator should mimic—thus simultaneously reducing development time while improving fidelity. In the remainder of this review, words in italic font are listed in our glossary of terms, Table 1.

This review introduces readers to emerging trends in simulation, walks through example use cases, and distills best practices. While research on simulation methods often emphasizes realism of the simulated data, their applications are typically only discussed at a high level. Here, we instead explore in-depth applications of simulated data. We first review existing packages and highlight their potential pitfalls (see Table 2). We then illustrate how researchers can get the most out of their microbiome data through simulation for various analytical tasks, ranging from effective experimental designs to data analysis strategies. In particular, this review focuses on three use cases: power analysis, methods benchmarking, and reliability analysis.

Power analysis is an important step in experimental design that informs the sample size required to detect different effect sizes, thereby enabling studies to be done with minimal resources, and without compromising scientific integrity and rigor [11]. Several power calculators have been proposed for microbiome studies [12–15]. However, these tools lack the flexibility of full simulators, which can be adapted to problem- and data-specific contexts. Most power calculators are limited to case-control designs and, without access to template data, they may make questionable distributional assumptions. In contrast, researchers using simulators can ensure their models reflect important properties of template

Received: October 15, 2024. **Revised:** December 19, 2024. **Accepted:** January 23, 2025

Table 1. A glossary of simulation terms

Term	Definition
Covariate	A biological or experimental characteristic of a sample that is thought to impact community profiles and is used to train the simulator.
Copula	A statistical approach for modeling correlation in non-Gaussian data, often used to model associations between taxa in a community.
De Novo Simulator	A simulator whose mechanism is defined without reference to external, template data.
Factor Model	A model that induces correlation across features through latent variables. Like copulas, these are often used to induce associations between taxa.
Fit-for-Purpose	Approaches that assess simulation quality by referring to small subsets of taxa.
Global Evaluation	Approaches that assess simulation quality by referring to the entire community (all taxa).
Marginal Distribution	A probability model that reflects abundance patterns for a specific taxon.
Multivariate Model	A probability model that reflects abundance patterns across the entire community (all taxa).
Outcome-Specific	Evaluation criteria that target the simulator's downstream application.
Reliability Analysis	The use of simulated data to calibrate interpretations of real data analysis.
Semisynthetic Data	The output from a simulator that has been designed to mimic external, template data.
Template Data	Previously gathered experimental data that can be used to train a simulator.

data and can compare various design choices, like class imbalance and sampling times, that go beyond sample size alone.

Benchmarking is essential for identifying the most suitable method for a specific experimental design and data type [16]. Simulation is useful to benchmark methods when ground truth is scarce. For example, in batch effect integration, it is important to preserve biological effects, but these can be difficult to pinpoint without simulations that provide some ground truth (illustrated in the section "Batch effect ection"). Further, benchmarking new methods on various plausible datasets provides more insight than simply identifying the best performer in a single simulation scenario. To this end, we can train a simulator to emulate real data and then alter it to reflect multiple hypothetical scenarios. Even when ground truth exists, simulations help gauge robustness to dataset perturbations (e.g. reducing the true effect sizes), and testing many simulated scenarios is often easier than collecting multiple real datasets. Formal simulation techniques provide a more automatic approach to benchmarking and allow researchers to focus on data analysis and method selection, rather than spending time programming simulators from scratch.

Reliability analysis can be enabled through simulation. Many modern data analysis workflows lack sufficient theory to guide practice, and simulations offer valuable sanity checks. For example, simulation studies have found that sparse canonical correlation analysis (sCCA) can result in high false discovery rates [17, 18]. Although these studies were motivated by neuroscience, sCCA is also widely used in microbiome data integration [19, 20]. In a similar spirit, the section "Omics data integration" highlights how data integration can introduce unexpected artifacts into dimensionality reduction visualizations. Hence, simulations help prevent misinterpretations that could misdirect research priorities.

Comparing these use cases highlights key distinctions. Power analysis is used to ensure studies have sufficient power (reduce Type II error), while reliability analysis helps to guard against incorrect conclusions (reduce Type I error). Benchmarking evaluates both types of error across a collection of methods and problem settings. By considering multiple methods and contexts, a benchmark can yield more generalizable, rather than application-specific, conclusions. Despite these distinctions, what power analysis, benchmarking, and reliability analysis have in common is that they attempt to guide decisions during the measurement and analysis process. Simulation is useful because it provides a systematic way to formalize the trade-offs between decisions, making it possible to test the strength of specific approaches against a battery of "what if?" questions.

This review makes the following contributions:

1) **Overview of simulation workflows**: We describe the main ingredients of modern simulation algorithms and associated assumptions. This background ensures that methods are not treated as black boxes and guides their effective use.
2) **Evaluation criteria**: We outline how to assess whether simulated data match the properties of previously observed template data, and how simulated data can inform methodological improvements.
3) **Case studies**: We present realistic case studies (see Table 3) demonstrating how simulation can support power analysis, benchmark competing methods, and guarantee reliable conclusions.

This review is accompanied by an online tutorial (https://go.wisc.edu/8994yz). Each chapter of the tutorial starts with a dataset discussion, walks through the process of designing simulators, and applies the resulting models to address common questions about microbiome experimental design, method benchmarking, and result interpretation.

Simulation workflows

Simulators vary widely, and their realism and relevance to downstream tasks is context-dependent. To help navigate this landscape, we first categorize simulators and then outline the workflow for building and refining them.

De novo and template-based simulators. We first note the distinction between *de novo* and template-based simulators. De novo simulators require users to specify parameters related to experimental design and distributional properties. These parameters cannot be automatically matched to real, template data. For example, in the splatter simulator [21], users can include treatment and batch effects, but these are drawn from the simulator's internal generation mechanism. In contrast, template-based simulators first estimate the impact of sample-level *covariates* using real data. Though this does require an initial template, it can lead to more realistic synthetic data. These data are often called *semisynthetic*, reflecting the influence of the template. As many

Table 2. An overview of packages that are available for microbiome community simulation. Models differ according to mechanisms they use to estimate community structure and the scope of applicability to alternative experimental designs. For applications to new environments or designs, the evaluation techniques described in the section 'Evaluation' can be used to compare candidates and gauge simulator fidelity

Package name	Probability mechanism	Experimental Designs	Notes
CAMISIM [34]	Models taxon abundances with a log-normal distribution and does not account for featurewise correlation.	Supports separate modes for case-control and time series designs, including simulation of technical replicates.	Used in the CAMI benchmarking challenge [51, 52] and can output read-level data (Illumina, PacBio, or ONT).
DeepMicroGen [35]	Combines recurrent network and GAN architectures to model multivariate changes over time.	Supports longitudinal designs with regular spacing.	Parameters are difficult to interpret, limiting potential for control.
MB-GAN [36]	Trains a GAN to match a template dataset, capturing both marginal and multivariate structures.	Mimics the design of the template data, but cannot be adapted to alternative designs.	Primarily for Centered Log Ratio (CLR) transformed data but applicable to small sample sizes.
metaSparSim [37]	Uses a hierarchical model to sample biological composition (Gamma distribution) and technical variability (multivariate hypergeometric distribution).	Does not include covariates during parameter estimation but allows changes post-estimation to reflect experimental perturbations.	Preserves sparsity and mean-variance profiles for individual taxa, with cross-taxa correlation quality depending on Gamma parameters.
miaSim [48]	Simulates population dynamics based on various interaction models, including the generalized Lotka-Volterra model.	Supports time series designs with potential environmental interventions.	Accompanied by an interactive Shiny application but cannot be fitted to a template dataset.
microbiomeDASim [32]	Draws community profiles from a zero-inflated, truncated multivariate normal distribution.	Designed for longitudinal experiments with user-specified trends (e.g. quadratic and "hockey-stick").	Only applicable to normalized abundance tables.
MIDASim [31]	Estimates marginals from empirical or generalized gamma models, correlating presence-absence patterns and abundances via a copula.	Supports case-control designs with varying sample sizes, sequencing depths, and taxa presence and abundance.	Scalable for both estimation and simulation, with support for template inputs and post-estimation modifications.
scDesign3 [43]	Applies GAMLSS to each feature, linking them with a Gaussian or Vine copula.	Can include covariates during estimation, enabling simulation of arbitrary designs.	Originally developed for single-cell applications, with similar models used for microbiome data [53–55].
SimulateMSeq [46]	Estimates separate parametric models across subsamples of a reference dataset, so model complexity grows with reference sample size.	Supports inclusion of covariates and confounder variables.	Originally developed to benchmark the robustness and scalability of differential abundance methods.
SparseDOSSA 2 [33]	Simulates features from a zero-inflated log-normal distribution, linking them with a Gaussian copula with a regularized precision matrix.	Does not incorporate covariates during estimation, but allows post-estimation changes that use them.	Previously validated on human gut and vaginal microbiome data.
ZINB-WaVE [30]	Simulates data from a zero-inflated negative binomial model, incorporating known covariates and low-rank latent variation.	Can include covariates during estimation, enabling simulation of arbitrary designs.	Originally developed for single-cell applications, with similar models used for microbiome data [53–55].

Table 3. Summary of case study datasets and analysis goals

Study name	Number of samples	Number of features	Analysis task	Objective
ATLAS [65]	883	89	Differential abundance testing	To select the appropriate differential abundance method for a given experimental design and data type.
Type 1 Diabetes [66]	101	427	Power analysis and multivariate method	To compare complex experimental designs and multivariate analysis approaches.
American gut microbiome [67]	261	45	Benchmarking network inference	To evaluate different network inference approaches under diverse network structures.
Anaerobic digestion [68]	75	231	Batch effect correction	To give insights into the behavior of batch integration methods.
Sepsis in ICU patients [69]	57	180 bacteria, 18 fungi, 42 viruses	Omics data integration	To uncover whether different data modalities (bacteria, fungal, virus) should be analyzed together or separately.

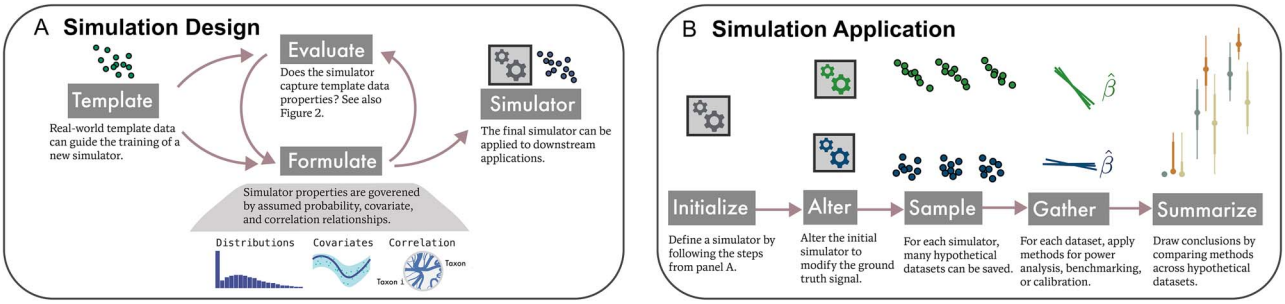


Figure 1. (A) Simulation design is an iterative process, where choices of probability distribution, experimental and biological covariates, and correlation structure can be refined according to evaluation criteria that draw attention to differences between real and template data. See Fig. 2 for example criteria. (B) Given a satisfactory simulator, the same workflow applies to power analysis, benchmarking, and reliability analysis. The initial simulator can be modified to reflect hypothetical experimental or biological settings, like changes in the signal effect size or the sample size. Outputs from competing approaches can be gathered and interpreted to guide experimental design and analysis.

public catalogs of 16S, metagenomics, and metabolomics data are now available [22, 23], it has become easier for researchers to access relevant template data for various possible analysis. We caution that different research communities have used different terms to describe the concept of template-based simulation. For example, "semisynthetic" is used in metabolomics [24, 25] and microbiome studies [26, 27], while "reference-based" is common in single-cell genomics [28, 29]. Although de novo and template-based simulators differ in their requirements, similar factors guide their application. We will review these workflow considerations next.

Formulation and application. In the formulation phase (Fig. 1A), a simulation model is created by adjusting parameters until the generated data pass a series of evaluation checks. In the application phase (Fig. 1B), several altered versions of the simulator can be defined according to the simulation study's goals. For example, to characterize false discovery rates, we can introduce synthetic negative control features designed to lack associations with the outcome. For each altered version of the simulator, we can generate multiple datasets, apply candidate analysis strategies, and gather summary statistics quantifying their performance. We provide practical examples of both phases in the "Case Studies" section.

The formulation phase relies on concepts from probability, statistical estimation, regression, and multivariate analysis. Probabilistic models enable sampling of new data with appropriate distributional properties, *multivariate models* induce plausible associations across multiple taxa, and regression methods ensure that simulations reflect biological or experimental influences.

Probability distributions

The choice of probability distribution dictates many properties of the generated data. Distributions must be carefully selected, as properties that might be easy to manipulate in one distributional family might be difficult to modify in another. For example, sparsity can be more easily modified in a zero-inflated versus ordinary negative binomial distributions. We first review univariate, then multivariate distributions.

Univariate distributions for simulation. For a specific taxon, we need to decide whether to simulate counts, proportions, or real numbers. Count distributions can be used to simulate amplicon sequence variant or metagenomics data before having applied any transformations. Common count distributions include the Poisson, negative binomial, and hypergeometric distributions. The Poisson distribution arises when counting events that, though individually unlikely, become common due to repeated opportunities for them to occur. For example, a stretch of DNA is unlikely to align to any given read, but given enough reads, we would expect a Poisson number of them to align. However, in real data, this often gives a poor approximation—the variance in observed counts is often higher than the Poisson can capture. Such overdispersion is more appropriately modeled using the negative binomial distribution. Finally, if the data exhibit more zeros than either Poisson or negative binomial distribution allows, then zero-inflation can be employed to introduce additional sparsity, as in zero-inflated negative binomial (ZINB) used in the single-cell RNA-seq simulator ZINB-WaVE [30]. Alternatively, presence-absence can be modeled separately from abundances, as in the simulator MIDASim [31].

When data are transformed, count distributions no longer apply. Depending on the transformation, different distributions can be used to model nonnegative, interval, or arbitrary real values. For nonnegative values, options include the truncated normal, log-normal, or variants of gamma distributions. The truncated and log-normal distributions enforce nonnegativity by either truncating or exponentiating a normal distribution. These distributions are used in microbiomeDASim [32], SparseDossa 2 [33], and CAMISIM [34]. Proportions within the interval [0, 1] can be represented by beta, Dirichlet, or generalized Gamma distribution, which include parameters for mean and concentration along the boundaries. The simulator MIDASim uses this approach to model relative abundance-transformed data. If data have been transformed to include both positive and negative real values, then they can often be modeled with normal or Student's T distributions, the latter being more appropriate when outliers are present. Further, such transformations can enable generative adversarial network models to perform well, as in simulators DeepMicroGen [35] and MB-GAN [36]. Continuous distributions can also be used as a preliminary sampling step for hierarchical count models. For example, the means of a count model can first be modeled as a Gamma distribution, allowing estimates for rare taxa to borrow strength from more abundant taxa. This hierarchical approach is used by metaSparsSim [37], which draws Gamma marginal distributions for individual taxa before sampling from a multivariate hypergeometric distribution for all taxa jointly.

Multivariate distributions for simulation. It is important that simulators generate realistic community profiles, capturing relational, multivariate structure and not just univariate, per-taxon abundances. Two common strategies for modeling these associations are (i) learning a multivariate normal covariance

in an appropriately transformed space or (ii) incorporating shared latent variation across features (taxa). Strategy (i) uses the multivariate normal distribution's covariance parameter to control the relationship between features. These multivariate normal samples can then be transformed into sequencing-specific data distributions. For example, this approach is adopted by simulators using the Logistic-normal Multinomial distribution [38, 39]. This method induces correlations across features by first sampling from a correlated, multivariate normal distribution. This correlated vector is transformed into a vector of proportions using a log-ratio transformation. Given an overall sequencing depth, simulated reads are then allocated to individual taxa based on these proportions. Alternatively, a correlated multivariate normal distribution can be transformed into a multivariate distribution with known univariate marginals, a process known as *copula modeling* [40, 41]. Copula models can be used even when the marginals are drawn from different distributions. This makes copulas easily adaptable, and they are used by simulators MIDASim [31], SparseDOSSA 2 [33], and single-cell simulators scDesign2 [42] and scDesign3 [43].

In contrast, strategy (ii) ties features together by assuming a latent, low-dimensional vector. The shared source of variation induces downstream correlations. This is often accomplished through variations of *factor models*. For example, ZINB-WaVE simulates entries from a model whose parameters vary according to a low-dimensional latent vector representing unobserved sample-level properties. Similarly, Latent Dirichlet Allocation assumes that samples are drawn from a multinomial distribution whose mean depends on a low-dimensional community membership vector [44]. In both univariate and multivariate settings, interpretable parameterizations can support modifications of the simulation mechanism. These modifications are helpful for validating workflows through computational negative or positive controls. For example, the mean parameter of a negative binomial model can be adjusted to reflect stronger or weaker treatment effects. In contrast, flexible machine-learning-based multivariate simulators, like MB-GAN and DeepMicroGen, can be challenging to alter in this way. Even if their simulated data are realistic, it is difficult to control them and enforce desired constraints.

Accounting for experimental and biological factors

How can we model differences across experimental or biological conditions? For example, a taxon's abundance may change gradually over time or may be a marker of disease status. In this case, histograms of individual taxon abundances may reveal complex patterns, like separate peaks for disease and healthy groups, which cannot be captured by models that assume independent draws from the same probability distribution. To simulate these patterns, we can specify distributional parameters based on sample characteristics. In addition to producing more realistic data, conditioning on sample covariates enables more precise control. For example, simulators that take into account time or disease status can generate data at a denser sequence of timepoints or different sample sizes for healthy versus disease patients, both of which can guide experimental design and method benchmarking.

Many simulators are tailored to specific experimental designs. For example, MIDASim [31] and SparseDOSSA 2 [33] support simulation from case-control designs. In both simulators, some taxa share parameters across case and control groups, representing synthetic negative controls. To create a subset of taxa that differ across groups, representing synthetic positive controls, taxa parameters can be allowed to vary. The greater the difference between parameters, the stronger the true effect. These simulators are particularly useful for power analysis and benchmarking for differential abundance studies [45]. To gauge robustness of differential abundance testing to unmeasured confounders, the simulators from references [46, 47] support the inclusion of confounders that are correlated with the covariates of interest. Other simulators have been created specifically for longitudinal designs. For example, microbiomeDASim [32] allows taxonomic abundance to vary over time based on various plausible trends, to mimic brief disruptions or gradual development. Similarly, miaSim [48], CAMISIM [34], and DeepMicroGen [35] are designed to capture longitudinal dynamics.

Though these simulators streamline work with specific experimental designs, it can be useful to generalize and map arbitrary sample-level variables onto distributional properties. This is especially valuable in multifactorial experiments, where multiple biological or experimental characteristics can jointly influence measurements. Such generalization also allows modeling interactions or nesting between variables. For example, treatment and control groups may have divergent temporal trajectories, or treatments might have different effects across cohorts. For these applications, parameters can be linked to sample covariates using regression techniques. For example, in the scDesign3 simulator for single-cell and spatial omics data [43], mean and dispersion parameters are modeled as functions of sample characteristics using generalized additive models of location, scale, and shape (GAMLSS) [49, 50]. This flexibility allows simulated data to vary smoothly over temporal or spatial coordinates. Similarly, in the ZINB-WaVE simulator for single-cell RNA-seq data [30], sample-level covariates can influence the mean, dispersion, and zero-inflation parameters through a linear model beyond feature-level properties, sample covariates can modify multivariate relationships. For example, scDesign3 allows different copula covariances to be used within prespecified groups of samples.

Evaluation

Evaluating synthetic data is a crucial step in data simulation to assess whether the generated synthetic data closely mirror the statistical properties of the original data. Without proper evaluation, synthetic data may not accurately represent the underlying distribution, potentially leading to biased models or incorrect conclusions.

In evaluating synthetic data, three primary types of utility measures can be used: *fit-for-purpose measures*, *global (broad) utility measures*, and *outcome-specific (narrow) utility measures* [56] (see Fig. 2). Fit-for-purpose measures provide an initial evaluation of whether the synthetic data appear reasonably close to the real data on relevant low-dimensional views, and they can be used to improve the simulation approach [56]. In contrast, global utility measures aim to assess the multivariate characteristics of the data. Outcome-specific utility measures aim to quantify similarity in analysis results or specific model parameters between real and synthetic data.

Fit-for-purpose measures typically involve checking the univariate and bivariate distributions of observed and synthetic data. We divided these measures into two main types: univariate and bivariate evaluations (Fig. 2A1-A2). In the former, we focus on whether the marginal distributions of variables in the real and synthetic data match. In the latter, we focus on whether the pairwise relationships of variables in the synthetic data resemble

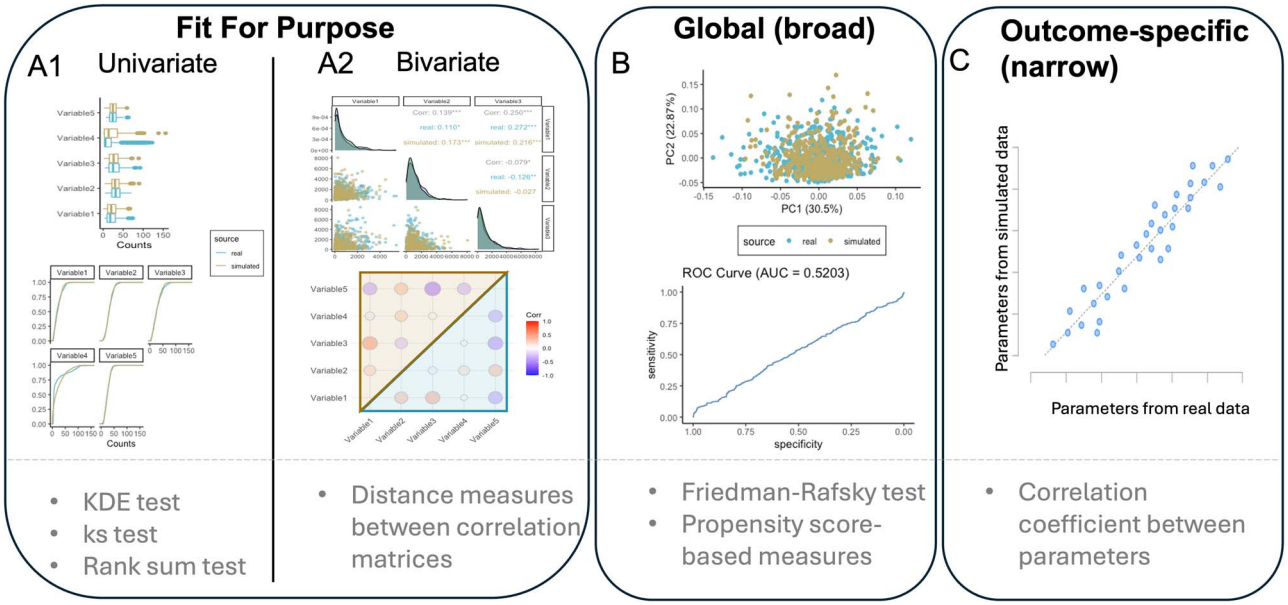


Figure 2. Utility measures for synthetic data, including visual comparison (top) and quantitative measures (bottom). (A) Fit-for-purpose, (B) global (broad), and (C) outcome-specific (narrow) utility measures. Fit-for-purpose and global utility measures both evaluate the similarity between synthetic and real data without considering any analytical objective. In contrast, outcome-specific utility assess the similarity of analysis results or specific model parameters using both real and synthetic data.

those in the real data, typically through bivariate distributions and correlations. We further divide evaluation criteria into graphical and quantitative measures. For example, graphical measures include side-by-side box plots or empirical cumulative density plots (univariate setting) and pairwise scatter plots or correlation heat maps (bivariate setting). Quantitative measures include the Kolmogorov–Smirnov (KS) test, the Wilcoxon Rank-Sum test, or independence tests based on kernel density estimation [57] (univariate setting) and the distance measures between correlation matrices (bivariate setting). However, while fit-for-purpose measures offer an initial assessment of synthetic data quality, they do not account for the multivariate nature of the data.

Global utility measures are built upon fit-for-purpose measures to evaluate the complex, multivariate nature of the data. We again divide these into graphical and quantitative measures (Fig. 2B). For example, graphical measures include Principal Component Analysis (PCA) plots, which jointly project synthetic and real data into a 2D space, or receiver operating characteristic (ROC) curves, which use binary classification to separate synthetic from real data, to determine if the synthetic and real data can be distinguished. Quantitative measures include the Friedman-Rafsky test [58] or propensity score-based [59] techniques. Global utility measures provide an aggregated similarity between simulated and real data. However, they do not guarantee that specific analyses on real and simulated data will yield similar results, as these measures do not consider a specific analysis goal [56, 60, 61].

Outcome-specific utility measures are designed to assess the simulated data for a particular analysis goal. Since there are multiple analytical approaches, these utility measures can vary significantly. For example, if the focus is on fitting a multiple regression model between sequencing features and a biological condition of interest, then we can compare regression coefficients obtained when fitting the regression to either simulated or real data. In this situation, a scatter plot can serve as a graphical measure, and the correlation value between the parameters estimated from real and simulated data can serve as a quantitative measure (Fig. 2C). However, if the objective is to perform a correlation analysis, like those which underlie microbiome network models, then the evaluation should focus on comparing the correlation matrices between the features in the real and the simulated data (for an example of this comparison, see Section 'Power analysis for multivariate methods'). No single utility measure is universally applicable. Therefore, performing several utility measures based on the specific objective of the simulation will help modify certain aspects of the simulator, such as the selection of distributions. Further, even once a realistic simulator is obtained, it should be used with care. For example, conclusions based on too few simulation replicates may not withstand scrutiny. More generally, analysis of how simulators are used in practice have found that researchers are systematically biased toward simulation experiments that cast their methods in a favorable light [62–64]. We hope that advances in simulator design and evaluation help reduce the cost of implementing thorough experiments that, more than acting as a rubber stamp, give insight into how to carry out complex data analysis.

Case studies

Statistical power for univariate differential abundance methods

Motivation. Differential abundance analysis is a cornerstone of microbiome research. It is often used to highlight taxa that are associated with disease or that respond to interventions. The research community has developed various testing methods to account for specific characteristics of microbiome data, such as zero-inflation, overdispersion, library size differences, compositionality, and small sample sizes [10, 31, 70]. However, assumptions that are reasonable in one microbiome system (e.g. the gut) may not necessarily translate to others (e.g. marine). Applying a method in an inappropriate context can lead to excess false

positives or reduced power. Moreover, unlike classical two-sample tests or linear models, differential abundance methods do not come with closed-form formulas for calculating statistical power. Therefore, semisynthetic simulation can give insights into the properties of differential abundance methods for specific data analysis applications.

We compared three common differential abundance methods commonly used in microbiome studies: DESeq2 [71], limma-voom [72], and ANCOM-BC2 [73]. ANCOM-BC2 was developed specifically for microbiome data, whereas DESeq2 and limma-voom were originally designed for RNA-seq data. The goal of this simulation is to select the most appropriate method for a specific dataset based on statistical power across sample sizes.

Data. We analyzed data from the ATLAS study [65]. This study profiled the gut microbiomes of $n = 1006$ healthy adult Europeans. Notably, it discovered microbiome "tipping points," which are unstable community composition profiles that could "tip" into more stable states. The competing stable states were related to factors like age and body mass index (BMI). We filtered down to the 89 most abundant genera in the 883 subjects with known region membership and BMI within the lean, overweight, or obese categories.

Simulation and evaluation. Our simulator follows scDesign3's modeling assumptions, applying a ZINB GAMLSS regression where both the mean and dispersion for each taxon can vary as a function of the BMI category. For the top 10 most abundant taxa, we found that the boxplot quartiles and the observed shifts across BMI groups were captured well in the simulated data (Fig. 3A1). We compared the abundance of all genera in simulated and real data for each BMI group using a kernel density-based two-sample test [57, 74]. At least 70% of genera were not statistically distinguishable (Fig. 3A2). As evident by these evaluations, the simulated data were generally similar to the original data. However, for some genera with positively skewed distributions and a high number of outliers, such as *Prevotella melaninogenica et rel*, the simulation failed to capture the high degree of skewness or generate outliers, hence the significant differences between the real and simulated data returned by some of the tests.

Data analysis results. We simulated synthetic negative controls by re-estimating a subset of genera so that the ZINB means and dispersions did not depend on BMI. These genera were declared not significant from a Wilcoxon Rank-Sum test of association with BMI (P-values corrected for multiple testing with 0.1 FDR level). Such weakly associated genera have previously been used to define pseudo-negative controls [75], whereas in our simulator, these genera are genuine negative controls. These simulation design considerations, as well as those for all other case studies, are summarized in Table 4. We then simulated datasets with sample sizes ranging from 50 to 1200, and applied DESeq2, limma-voom, and ANCOM-BC2 using an FDR-control level of 0.1. On average, DESeq2 had better power compared to the other two approaches, and all three approaches provided valid FDR control (Fig. 3B). The power increased most rapidly from 50 to 337 samples.

Summary. Differential abundance testing requires a delicate balance. On one hand, parametric assumptions, like those for overdispersion or zero inflation, can improve test sensitivity. On the other hand, inappropriate assumptions can lead to invalid results. Generic benchmarking studies can help identify the appropriateness of each method for a given experimental design and data type but can only provide blanket recommendations. In contrast, we showed that simulations let us design "self-service" benchmarks where we can run our own *in silico* experiments to inform a statistical analysis workflow with more problem-specific choices. An additional benefit is that we can ask how certain changes to the data (e.g. increasing the sample size) might affect method performance.

Power analysis for multivariate methods

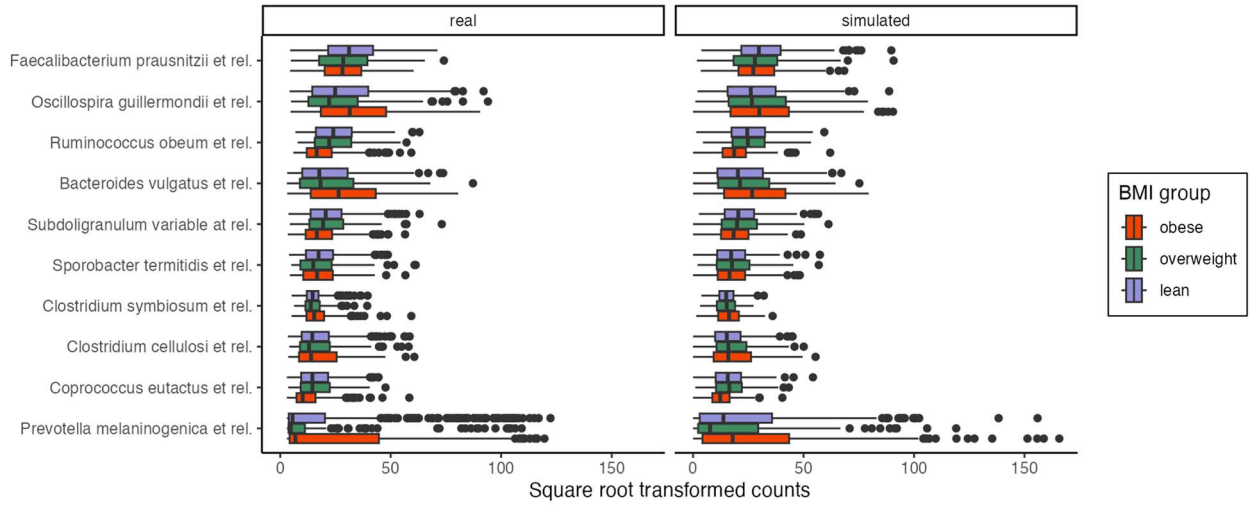
Motivation. Differential analysis methods can draw attention to significant taxa but can overlook important correlation structure within microbiome communities. To shed light on this more global structure, multivariate analysis methods play an essential role [76]. Theoretically characterizing the statistical efficiency of multivariate methods is an active area of research [77, 78], and practical sample size guidance typically relies on simulation [79, 80]. However, effective community-wide simulation is more challenging than what is required for differential abundance analysis, because we must pay close attention to the quality of the estimated correlations. In this case study, we explore how semisynthetic data can inform sample size calculations in an application that uses sparse partial least squares discriminant analysis (SPLS-DA) [81]. SPLS-DA is a classification method that makes use of correlations between input features to ensure stable predictions in small sample size settings. In the process, it computes a dimensionality reduction of the data analogous to PCA, but with the explicit goal of separating classes.

Data. We revisited the Type 1 Diabetes (T1D) study from Gavin et al. [66], who identified metaproteomic patterns in the microbiomes of T1D patients from a cohort of $n = 101$ study participants. We filtered down to the 427 proteins present in at least 70% of either the T1D ($n = 51$) or control ($n = 50$) groups. Following the original study's data preprocessing, we applied a centered log-ratio (CLR) transformation to the measured protein relative abundances and then used these as predictors in an SPLS-DA with T1D status as the outcome. We set the SPLS-DA hyperparameters to 5 PLS dimensions and a selection of 30 predictors. On this dataset, 10 repetitions of 5-fold cross validation yield a holdout area under the receiver operating characteristic curve (AUROC) of 0.667 ± 0.037.

Simulation and evaluation. We fitted a Gaussian GAMLSS simulator, allowing means and variances for all proteins to depend on T1D status. Since SPLS-DA models the relationships across all proteins, we first applied a Gaussian copula using the standard sample covariance to attempt to reflect the true correlation structure. Contrary to the last section, we focus here on the quality of the simulated sample correlations rather than the univariate simulation quality (which is covered in our online tutorial section 3.3).

A simple histogram of pairwise correlations between features in simulated data showed greater variability compared to the observed data, suggesting that the simulation could be improved using a regularized covariance estimator. We therefore used the adaptive thresholding covariance matrix estimator of Cai and Liu [82] (higher thresholds apply stronger regularization, while lower thresholds reduce bias). To choose the threshold, we evaluated the correlation quality across a range of candidate values from 0.001 to 0.2 using two metrics: the KS statistic between observed and simulated histograms of pairwise correlations and the Frobenius norm between observed and simulated correlation matrices. The KS statistic formalizes our histogram check, while the Frobenius norm measures the differences between these matrices. We found that the KS statistic was minimized at 0.14, while the Frobenius error was minimized at 0.03. To balance these two metrics, we chose a threshold of 0.1. To further

A1



A2

BMI Group	Lean	Overweight	Obese
Proportion of insignificant tests	0.72	0.80	0.71

B

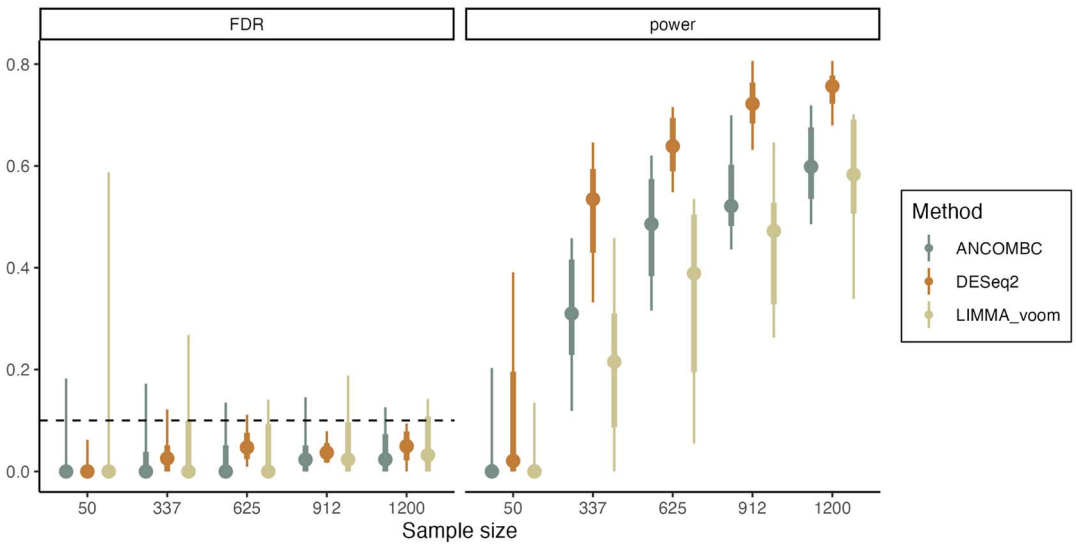


Figure 3. Differential abundance analysis in the ATLAS study. (A1) Comparison between real and synthetic data for the 10 most abundant taxa. A ZINB model achieves satisfactory performance in matching quartiles of the observed data. (A2) Quantitative evaluation between real and synthetic data. Proportion of non-rejected tests from the kernel density-based global two-sample comparison test for each feature across BMI groups in real and simulated data. Insignificant test results indicate that the kernel density estimates for the given feature in each BMI group cannot distinguish between the real and simulated data. (B) Realized power and false discovery rates for the DESeq2, limma-voom, and ANCOM-BC2 methods applied to the simulated data. Large samples lead to higher and less variable experimental power. FDR control is maintained on average.

check the simulator, we created pairwise scatter plots (Fig. 4A) and heatmaps of real and simulated sample covariance matrices (Fig. 4B). The heatmaps show similar blocks of positive or negative correlation. For the pairwise scatter plots, we filtered to a subset of pairs with moderate positive correlations between 0.72 to 0.8 in the real data, as checking all 90K pairs is impossible. The scatter plots of real and simulated data overlapped well. The only notable differences were the streaks of exact zeros in the real data, a reflection of the zeros present before CLR transformation. Overall, the revised simulator with an adaptive covariance estimator seemed sufficient for a power analysis with SPLS-DA.

We next altered the fraction denoted π_1 of nonnull proteins whose means and standard deviations depend on T1D status.

Table 4. Summary of simulation design considerations for the case studies. Abbreviations: power analysis (P), benchmarking (B), reliability analysis (R)

Analysis task	Kept from the template	Simulator modifications	Justification
Differential abundance testing (P, B)	BMI effects for the ground truth differentially abundant taxa.	Varies the proportion of ground truth differentially taxa and overall sample size.	BMI effects must be removed to create synthetic negative control taxa.
Power analysis and multivariate method (P)	T1D effects for the ground truth differentially taxa; community covariance.	Varies the proportion of ground truth DA taxa and the overall sample size.	Prediction performance is known to depend on the sample size and the proportion of relevant features.
Benchmarking network inference (B)	BMI and sequencing depth effects for all taxa.	Alters the ground truth covariance structures are substituted into the Gaussian copula.	Covariance is a common target for network inference methods, but no single edge pattern is universally applicable.
Batch effect correction (R)	Batch and treatment effects for all taxa; community covariance.	Introduces a treatment group with a weak effect.	Overintegration is especially problematic in settings with weak signals.
Omics data integration (R)	Sepsis and antibiotics effects for taxa in the ITS and Virome assays; community covariance.	In one setting, sepsis and antibiotics effects are removed from the 16S assay.	Dimensionality reduction plots should not introduce spurious associations.

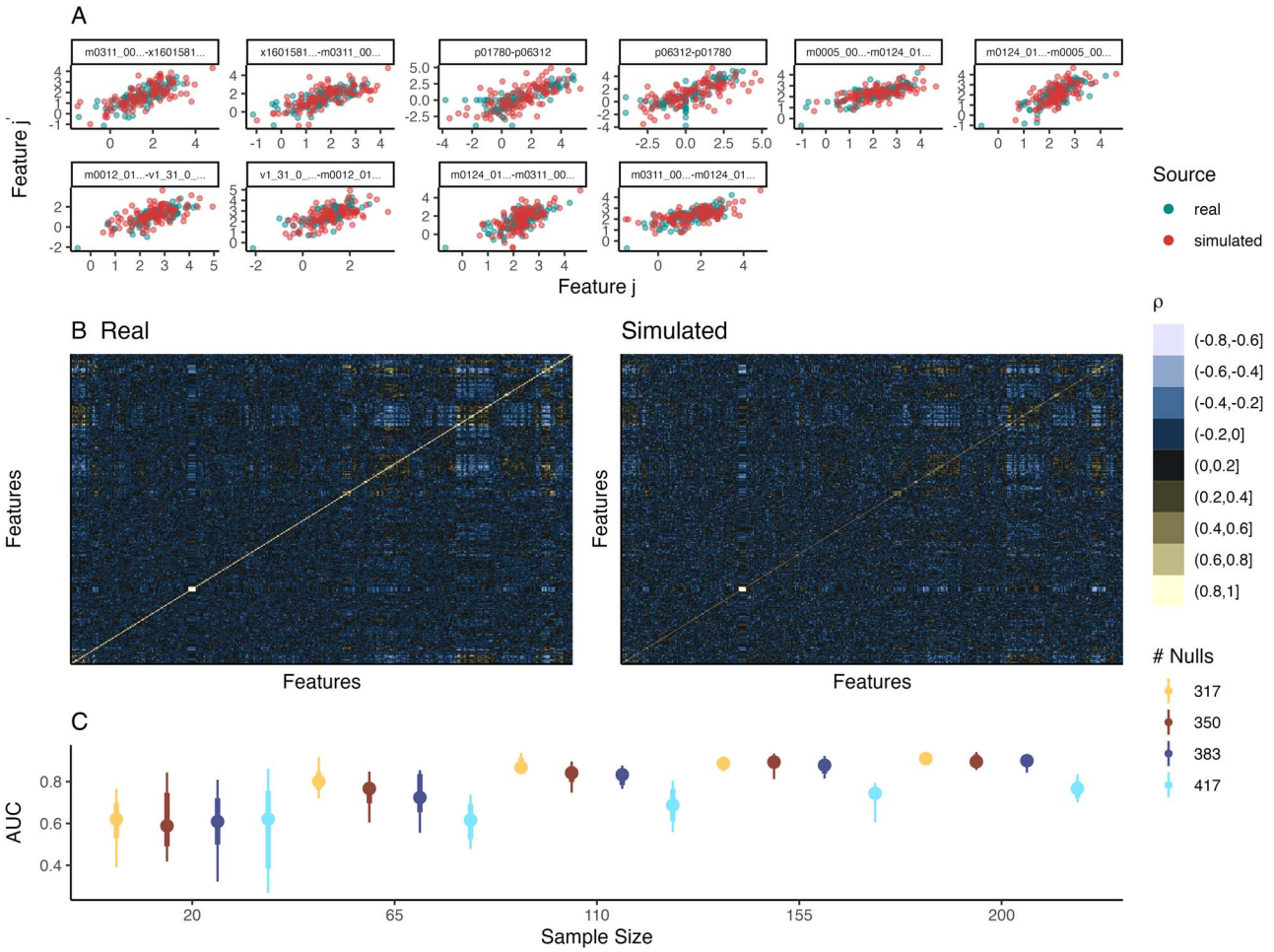


Figure 4. SPLS-DA power analysis for Type 1 Diabetes association. (A) Pairwise scatter plots for proteins with real-data correlation ranging from [0.72, 0.8]. Except for streaks at 0, a Gaussian copula appears to preserve associations between pairs of proteins. (B) A heatmap of the correlation matrices across all proteins in the real and simulated data. Blocks of positively and negatively correlated proteins appear to be preserved in the simulation. (C) Prediction accuracy of SPLS-DA across simulation settings. The x and y axes correspond to sample size and AUROC, respectively. Color corresponds to the number of null features, $427 \times (1 - \pi_1)$, which governs the rate of power increase as a function of sample size.

We chose the nonnull proteins according to their P-values from a Wilcoxon Rank-Sum test on the original experimental data. We considered the most significant proteins as true signals in the simulator. We then applied SPLS-DA with the pre-specified hyperparameters to semisynthetic data of varying sample sizes, with $\pi_1 \in \{\frac{10}{427}, \frac{30}{427}, \frac{50}{427}, \ldots, \frac{110}{427}\}$, giving coverage of settings where

differentially abundant taxa range from rare (< 2.5% of taxa) to relatively common (> 25% of taxa).

Data analysis results. When considering 20 samples total, we found that the proportion π_1 of true nonnull proteins has little effect on average AUROC, which is only slightly better than random (Fig. 4C). Larger π_1 was associated with less variable performance. When $\pi_1 = \frac{110}{427}$, performance increased quickly with sample size, plateauing at 110 samples. In contrast, for $\pi_1 = \frac{10}{427}$, performance increased more gradually, with room for improvement even at 200 samples. Interestingly, performance was relatively stable from $\pi_1 = \frac{44}{427}$ to $\frac{110}{427}$. Altogether, this analysis suggested that if more than a few dozen proteins are thought to be associated with the outcome, then between 110 and 155 samples is sufficient. For fewer proteins, either a larger sample size or an alternative analysis strategy should be considered. A sense of the true proportion π_1 of nonnull proteins can be gauged from the distribution of *P*-values in the template data, using estimators like those introduced in [83–85]. For these data, the true signal appeared to be weak. For example, Storey's estimator returned $\hat{\pi}_1 < 0$.

Summary. We showed how semisynthetic data can help determine sample size for SPLS-DA applied to microbial proteomics. While no formulas for statistical power exist for this method, simulation allowed us to study how both experimental (sample size) and biological (proportion of nonnull proteins) factors influence its performance. Further, we detailed the process of evaluating the simulator's correlation structure, demonstrating how visualizations and metrics could be used to iteratively refine a model.

Benchmarking network inference

Motivation. Network models give a holistic view of interactions in microbial ecosystems. They can identify tightly connected subcommunities and keystone taxa [86–88]. Unfortunately, validating these models is difficult, since determining ground-truth edges typically requires low-throughput experimentation such as knockout studies [89]. This creates a barrier to benchmarking, both for their use in specific studies and for evaluating new methodologies. Simulation can address these issues by providing ground-truth edges.

We compared two methods: SpiecEasi [90], a graphical lasso-based algorithm tailored to compositional data, and the Ledoit-Wolf estimator [91], a covariance estimator created for high-dimensional but low-rank data. A priori, we may expect SpiecEasi to perform better on microbiome data, since it was specifically designed for this purpose. However, this comparison has not been previously reported, and it is also unclear if potentially improved accuracy justifies the additional time required to solve the SpiecEasi optimization problem. To help answer this question, we can use simulations with known covariance structures.

Data. We analyzed the data from the American Gut Microbiome (AGM) project, a citizen science initiative where participants submit stool samples and complete detailed diet and health surveys [67]. The study has revealed associations between survey responses and microbiome composition. We considered a subset of 261 samples available through the SpiecEasi package [92], after excluding samples with fewer than 1000 reads. We filtered down to a "core" gut microbiome [93] of the 45 taxa with an abundance of over 100 in at least 50 samples.

Simulation and evaluation. We fitted a ZINB GAMLSS model to these data using BMI and log sequencing depth as covariates. The BMI term was included because the original AGM study [67] found a significant association between BMI and microbiome composition. The sequencing depth term allows samples with deeper sequencing to have larger means on average across all taxa. Note that, though we convert to compositions before applying SpiecEasi, our simulation operates on the scale of absolute abundances, and this could influence conclusions. In simulation, there is often the question of where in the measurement process to generate data. This example simulates counts of identified taxa, but we could have alternatively simulated sequencing reads (further upstream) or compositional data (further downstream). While simulations grounded earlier in the measurement process apply more widely, they may be harder to match to template data and use within the intended analysis.

The evaluations comparing real and simulated data are available in section 3.2 of the accompanying online tutorial. Here, we focus on benchmarking estimator accuracy across several network structures beyond those observed in AGM. We define several ground-truth correlation structures representing different statistical regimes: block, banded, Duncan-Watts small-world, scale-free, and Erdoős-Rényi random graph structures [94]. The block and banded covariances were defined directly, while others were derived from Gaussian graphical models with the respective structure, shown in Fig. 5A. For example, the scale-free network had several hub taxa with high covariance across many neighbors, while the Erdoős-Rényi network connected all pairs of taxa with equal probability. Note that while our focus here is on covariance matrix estimation, a similar study using known, structured precision matrices could also be implemented.

Data analysis results. Across network regimes, the Ledoit-Wolf estimator had larger bias but lower variance compared to SpiecEasi (Fig. 5B). Since both estimators are regularized, they exhibited some bias toward correlations with smaller magnitudes. The Ledoit-Wolf estimator performed better in the banded and block settings, while the two approaches were comparable in other cases. All methods were challenged in the Erdoős-Rényi setting, often declaring large correlations for taxa with no true relationship. Interestingly, in the block covariance scenario, SpiecEasi estimated many exact zero correlations as slightly negative. This is a consequence of SpiecEasi's compositional assumption, which induces negative correlation across taxa. In spite of this artifact, the block covariance setting appeared to support more efficient estimation than any of the other covariance settings we considered. Thus, if we assume that a community has a block correlation structure, then fewer samples may be required. This is consistent with general statistical theory, which argues that low-dimensional block structure can greatly simplify high-dimensional covariance estimation [95].

Summary. We showed that simulations offer a useful lens for comparing network inference methods across diverse network structures while maintaining realistic abundance distributions. Starting from a single template dataset, we were able to simulate according to several types of ground truth correlation, enabling us to identify settings where methods are more likely to fail or succeed.

Batch effect correction

Motivation. In large microbiome studies, it is often difficult to guarantee uniform data collection and processing for all samples, leading to systematic differences across experimental groups, often referred to as batch effects [96, 97]. For example, storing samples at different temperatures might change the abundances of taxa whose marker gene sequences degrade more rapidly at some temperatures. Failing to address these batch effects can limit power by masking treatment effects. They can also

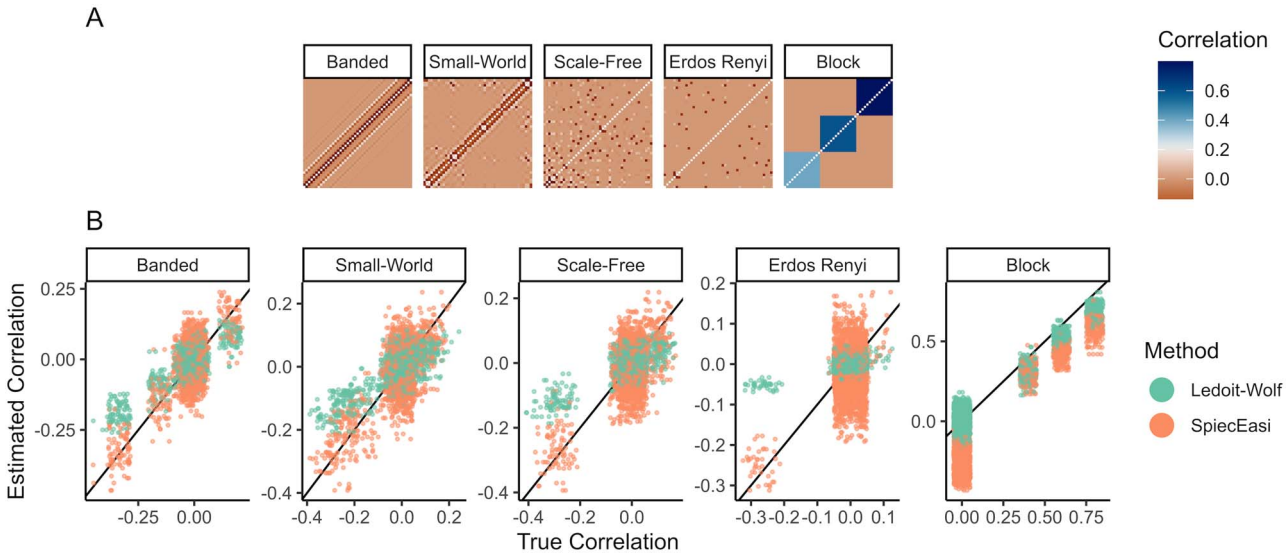


Figure 5. Benchmarking network methods for American Gut Microbiome project. (A) Ground truth correlation matrices used in the simulation study and (B) estimates made using SpiecEasi and the Ledoit-Wolf estimator. Entries (j, j') in the matrices of (A) provide the copula correlation between taxa j and j' before transformation using the ZINB model. From (B), SpiecEasi tends to have lower bias, since it is generally centered along the ground truth black line, but higher variance, since its vertical spread is larger. This figure also highlights settings where estimation tends to be easier (block and banded) and more difficult (Erdoős-Rényi).

compromise validity by introducing spurious differences across treatment groups [98]. Hence, a common preprocessing step in microbiome data analysis is to apply batch effect correction to standardize data across batches [99, 100].

Despite their increasingly widespread adoption in microbiome studies, batch effect correction methods must be applied carefully. Failing to account for the correction in downstream differential tests can lead to miscalibrated P-values [101] or performance estimates [102]. Further, it can be difficult to balance underintegration, where batch effects persist post-correction, against overintegration, where aggressive batch effect correction eliminates meaningful biological variation [103]. It is often unclear in advance whether these issues will arise for a given dataset or batch effect correction method. By defining ground-truth batch and biological effects, simulations give a way to evaluate batch effect correction methods and their impact on downstream inferences. Moreover, they allow quick comparison of methods across experimental designs and biological scenarios, allowing more complete evaluation than is possible in isolated benchmark datasets.

Data. We analyzed a study of anaerobic digestion (AD). AD is a biodegradation process underlying many bioenergy production technologies. The study Chapleur et al. [68] profiled the microbiomes from AD samples treated with phenol, a micropollutant that affects biodegradation efficiency and stability, posing challenges for large-scale industrial deployment. We use a subset of 75 samples and 231 genera discussed in Wang and Lê Cao [96], which focused on changes in community composition under two phenol concentrations. Since obtaining samples is time-consuming, the experiment was carried out over five sessions, leading to batch effects visible in the PCA plot in Fig. 6A.

Simulation and evaluation. The purpose of this simulation experiment is to compare the performance of RUV-III [104] and ComBat [105], two batch-effect correction methods, when applied to data like those in the AD study. We used the CLR-transformed AD dataset as our simulation template. For each taxon, we applied a Gaussian GAMLSS and copula model with batch and treatment status as covariates. Given the relatively small sample size, we used an adaptive thresholding covariance matrix estimator [82] within the Gaussian copula. PCA on data simulated from this model revealed that the simulator could recapitulate the observed batch effects. We can evaluate this more precisely using a narrow utility evaluation, comparing the RUV-III batch effect regression coefficients from the real and simulated data. For each taxon, a regression coefficient was calculated using AD as the factor of interest. The regression coefficients from the simulated data closely matched those from the real data, with an overall correlation of 0.89 (Fig. 6B).

Batch effect correction methods are known to be sensitive to imbalance between true biological groups [97]. We next explored whether this could pose a challenge in AD studies if we considered an additional phenol concentration treatment level. We simulated a hypothetical scenario where a stronger treatment had been applied to a small subset of samples. Each batch was assigned 15 samples at the reference concentration ($t = 0$) and the previously observed comparison group ($t = 1$), but only six at a new, hypothetical treatment (set to $t = 1.8$). Relative to the reference, this new treatment level is anticipated to perturb the microbiome community in the same direction as the original $t = 1$ group, but to a greater extent. Since this treatment group is less frequently sampled, there is a risk that its effect might be masked by overly aggressive batch integration.

Data analysis results. We compared the results of RUV-III and ComBat applied to data simulated from our altered experimental design (Fig. 6C). The simulator generated data with clear batch and treatment effects (first row). Although we observed a lack of replication structure for some of the batches in Fig. 6A, the simulated data preserved sufficient batch effects to evaluate batch effect correction methods. We then compared the PCA projections when RUV-III and ComBat were applied to the simulated data. RUV-III requires users to specify pseudocontrol features, which are assumed to be unaffected by biological factors, and the experiment's replication structure. We set pseudocontrols to be taxa that were not significantly associated (q =

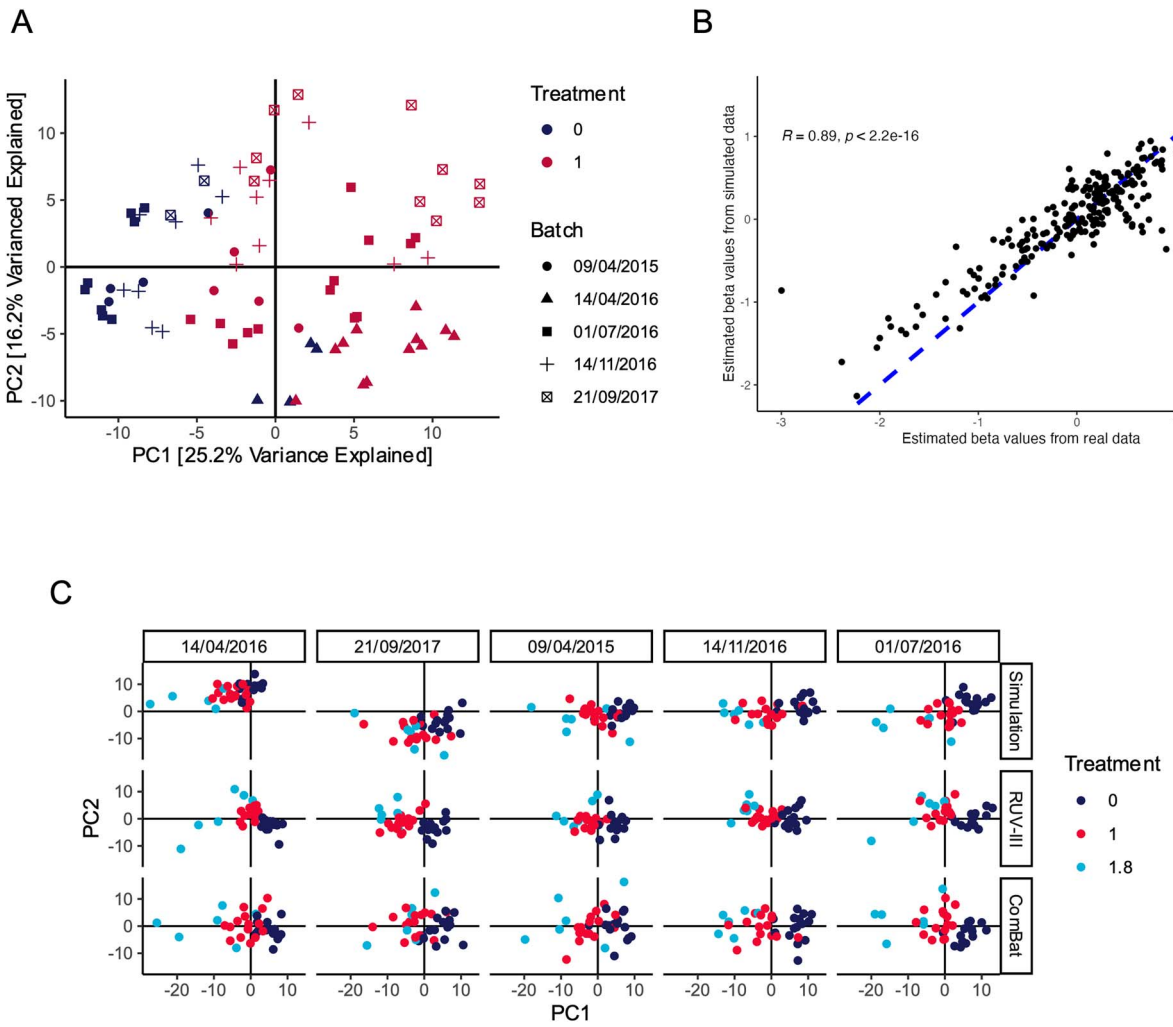


Figure 6. Assessing batch effect correction methods in an AD study. (A) A principal components plot of the original AD dataset reveals significant batch-to-batch variation. For example, all experimental samples generated on 14 April 2016 are shifted toward the bottom-right quadrant. Note that the technical replication structure is also visible, for example, the clusters of untreated samples from 01 July 2016 on the left-hand side. Though the two phenol concentration treatments can still be distinguished from one another, removing the batch effect can improve power. (B) Narrow evaluation using a scatterplot and correlation coefficient to compare the RUV-III batch effect regression coefficients for each taxon in real and simulated data. (C) Original and batch corrected data in the simulation with a new, less frequently sampled treatment group ($t = 1.8$) introduced. Batches (columns) have been sorted from the lowest to the largest average PC1. Both RUV-III and ComBat successfully remove systematic differences between batches, but in some cases, ComBat removes true differences between the $t = 1$ and $t = 1.8$ groups. Note that the principal components are derived separately for each row.

0.05) with concentration in an initial multiple testing screen that used a linear model with batch and concentration as covariates. Replication structure was set to the biological sample ID. Both methods centered all batches around the origin, as expected from a batch effect correction method. However, the $t = 1.8$ group was less clearly separated from the $t = 1$ group in the ComBat output compared to either the original simulation or the RUV-III corrected data. For a more quantitative analysis, we trained a linear discriminant analysis to predict the treatment group from the first two principal components, yielding a classification accuracy of 73.9% on the original simulation, 78.3% on the ComBat correction, and 88.3% on the RUV-III correction. These classifications performance results suggest that ComBat may overintegrate imbalanced data, while RUV-III preserves more subtle treatment differences.

Summary. We illustrated how simulation can give insight into the behavior of candidate batch integration methods in a way that is tailored to specific data analysis contexts, with the opportunity to create new treatment scenarios not present in existing real-data benchmarks.

Omics data integration

Motivation. A single assay can only give a partial view of a microbial ecosystem. For example, 16S rRNA sequencing data characterize bacterial community composition within a sample, but other kingdoms, metabolites, and host cells shape microbiome properties. To capture these features, different sequencing methods conducted on the same samples (e.g. ITS for fungi) or profiling techniques (e.g. NMR for metabolites) are needed. By analyzing the complementary views offered from diverse datasets, it is possible to develop a more holistic understanding of an ecosystem's biology [20, 106, 107].

Effectively analyzing these data, however, is challenging, with no consensus on how integration strategies should be deployed across contexts. The statistical task is not simply multivariate but also multiassay, linking tables influenced by diverse biological or

experimental factors. It is necessary to decide on which datasets to integrate, how each table should be normalized, and which integration method might be appropriate. Simulation can give a controlled, simplified setting within which to compare analysis strategies and can help to predict method performance under realistic biological scenarios. That is, simulation can guide reliability analysis for data integration.

Integration methods often return a dimensionality reduction plot designed to uncover shared covariation across assays, analogous to how PCA arranges samples to describe variation within a single assay. It is critical that these methods faithfully preserve between-assay and between-sample relationships. Ideally, data integration dimensionality reduction will highlight genuine, shared axes of variation across data sources while avoiding the appearance of false associations between unrelated data sources. We design a simulation to assess the extent to which these plots can misleadingly suggest similarities in inherently "unalignable" assays [108]. This multi-assay analysis parallels the multi-batch problem discussed in the previous case study—instead of studying overintegration across batches, we consider overintegration across assays.

Data. We re-analyzed the data from Haak et al. [69], who studied how the gut microbiome is altered during sepsis in ICU patients. Sepsis can be triggered by microbial infection unrelated to bacteria (e.g. from the *Candida* fungus). Therefore, each sample was profiled using ITS amplicon and Virome sequencing in addition to 16S rRNA sequencing. Since sepsis is treated with antibiotics, the study included healthy patients undergoing antibiotics treatment. The study included 20 healthy controls, 23 sepsis patients, 5 healthy patients on antibiotics, and 9 ICU patients without sepsis. After applying the same filtering as [69], we obtained measurements for 180 bacterial genera, 18 fungal genera, and 42 viruses.

Simulation and evaluation. We considered the multiblock SPLS-DA method to integrate these class-labeled sequencing datasets [109], and simulated from the scenario where the class label was relevant only for a subset of tables. This could occur if any of the three kingdoms assayed were unrelated to sepsis or antibiotics. An ideal multiassay integration in this setting would recover the similarities in the abundance profiles across kingdoms (e.g. revealing shared clusters of samples) while avoiding introduction of false relationships with sepsis or antibiotic status. Note that multiblock SPLS-DA is the multiassay extension of SPLS-DA (discussed in the section 'Power analysis for multivariate methods') and replaces the SPLS-DA objective with a weighted average of covariances across pairs of tables. We applied a Gaussian GAMLSS simulator to each assay, as they were already normalized. For the ITS and Virome assays, we allowed means and variances for each feature to depend on a categorical feature encoding both ICU and antibiotic use. For the 16S data, we used two simulators: one which conditioned taxonomic abundances on ICU/antibiotics category and one which did not. We joined all tables using a Gaussian copula with an adaptive covariance estimator to handle high-dimensionality. Therefore, in the ground truth simulation, the marginal (taxon-level) and covariance (community-level) structure from the template data are maintained. However, while the sepsis and antibiotic factors continue to influence all the ITS and Virome measurements, this relationship has been deliberately removed from one of the two simulated 16S datasets.

This design allows us to study the extent to which integration can spuriously introduce sepsis or antibiotic associations into the 16S data. Specifically, we then compared multiblock SPLS-DA output on the original and both versions of the simulated data (Fig. 7).

Data analysis results. We find that, even after removing all association between 16S abundances and the class label, the multiblock SPLS-DA projections for the 16S block still separated by class, albeit more weakly than before. This is a consequence of the analysis' multiblock nature: strong class associations from other tables can be artificially introduced into the 16S table. Indeed, part of the objective of the multiblock SPLS-DA algorithm is to maximize similarity in projections across all tables. In this case, in the real data, the class differences in the 16S data were much larger than those seen in this null scenario. Thus, simulation can guarantee reliable conclusions for complex integration tasks.

Further, to decide on which tables should be analyzed together, or separately, we calculated alignment measures [108] on the real data and compared them with a synthetic null where the tables are not alignable. We study the distribution of canonical correlations in both the observed and synthetic negative control data when integrating the Virome and 16S data together (Fig. 8). The control was defined so that the two tables are known to have no correlation within the underlying copula model. The results showed the presence of slight, but systematic, shared variation: observed canonical correlations were consistently larger than those in the synthetic null for the top dimensions, but the increase was quite modest. This suggests that substantial variation is isolated within the tables in a way that an integrated analysis cannot capture, motivating the use of within-table analyses alongside integrative methods.

Summary. Analogous to using negative controls in biological experimentation, we created synthetic negative controls in a computational workflow for data integration. This allowed us to gauge evidence for potential discoveries in real data, and prioritize either simultaneous or table-specific analysis based on departures from the synthetic negative control setting.

Discussion

We have reviewed the potential to apply simulators to microbiome study design and analysis. We have considered advances that make simulators more broadly applicable than their predecessors and have provided reproducible case studies showing how they can offer ground truth for evaluating FDR and power (sections "Benchmarking differential abundance methods and statistical power for univariate methods" and "Power analysis for multivariate methods"), help identify regimes where competing methods differ (sections "Benchmarking network inference" and "Batch effect correction"), and ensure valid interpretation of statistical outputs (section "Omics data integration").

We have emphasized methods for evaluating simulation quality, introducing vocabulary for distinguishing between fit-for-purpose, narrow, and global metrics. These metrics can be used both to compare simulation packages, like those listed in Table 2, and to refine initial simulator formulations. Moreover, this review has identified criteria for determining a simulator's relevance to specific applications, including distributional assumptions, ability to match a template, handling of multivariate relationships, and incorporation of experimental or biological effects. Between simulators, there are trade-offs between faithfulness, controllability, and generality of application, and the most appropriate approach will depend on the relative importance of these factors.

A key insight of modern simulation is that models can be trained on real experimental data from related contexts. A

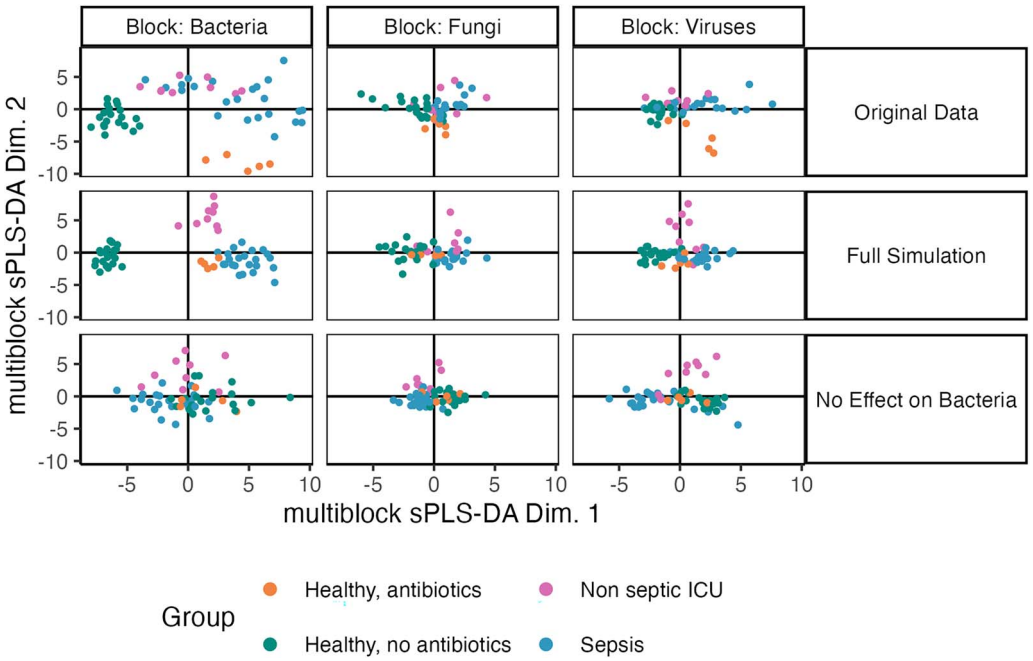


Figure 7. Using simulation to understand properties of omics data integration methods in a transkingdom analysis. Columns correspond to three amplicon sequencing datasets for transkingdom analysis of sepsis from Haak et al. [69]. Multiblock SPLS-DA projections from real and simulated data are given in the first two rows. The bottom rows show a perturbed simulation where all associations in the 16S data have been removed. The fact that between-group differences are still visible reflects the cross-table reduction and serves as a null reference for reliability analysis.

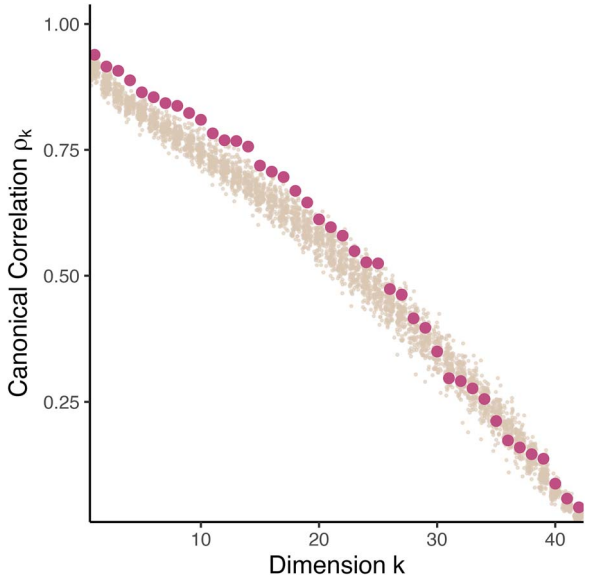


Figure 8. Comparison between canonical correlations in real (red) and reference null (beige) samples in a transkingdom analysis. For the first 20 dimensions, the real canonical correlations are slightly larger than those in the synthetic null. These dimensions are likely to reflect true, if modest, shared variation across tables.

semisynthetic approach can enhance simulation quality and requires less implementation effort compared to designing a de novo mechanism. However, it is worth cautioning that even a simulator that perfectly emulates a template dataset may still be unhelpful if the template is not a good match to the motivating problem. There is a trade-off between using template data from pilot experiments, which are small but highly representative of future conditions, and large public databases, which offer more data but may not directly relate to the problem of interest.

There is a close connection between semisynthetic simulation and statistical plasmode simulation [110, 111]. Plasmodes are data where aspects of the measured system are known in advance, for example, data that include a spike-in control. Analogously, statistical plasmode simulators are resampling-based methods that introduce a fixed parametric component. For example, a plasmode approach to a linear regression simulator may resample covariates but generate the response using a known coefficient vector. Both plasmode and semisynthetic approaches to simulation use templates to improve simulator realism. However, plasmode models bypass simulation for some variables by directly resampling them from the template.

We expect future simulators to support workflows with novel data modalities and their combinations. For example, while we considered integration across batches and assays, their heterogeneity was mild. Developing simulators that model variation not just across batches but across cohorts, and not just for different amplicon technologies but for entirely different assay types, is a worthwhile direction for further study. Moreover, though we focused on simulating community taxonomic and metabolomic profiles, methods for simulating metagenomics reads have also been proposed [112–114], and incorporating template data into read simulation could allow systematic study of entire processing and analysis pipelines, similar to recent advances in single-cell read simulation [115].

As simulators become easier to develop and apply, the potential for reusing others' work will increase. For example, instead of creating one-off simulators for the power analysis in a grant proposal or the simulation study in a methods paper, researchers could borrow and modify existing simulator definitions and output. Curating repositories of reusable simulators, where the template data and generating mechanisms are specified, would be valuable. While we have focused on researcher-level power analysis, benchmarking, and reliability analysis, simulation also has the potential to resolve field-level controversies. For example, no consensus

has emerged for the analysis of strain-level variation, with some researchers emphasizing their importance to health outcomes [116] and others cautioning against attempts to detect them with existing technology [117, 118]. Simulation can provide the ground truth and control necessary for fine-grained discussion of these issues. Indeed, it has already played an important role in a debate about the use of supervised normalization in microbiome data [119, 120]. As techniques become more powerful and accessible, computational studies using semisynthetic data will become an important part of microbiome research toolkit.

Key Points

- Semisynthetic simulators differ from de novo simulators by being trained on an experimental template, allowing them to more accurately reflect the properties of real-world data, and are an important recent development in approaches to microbiome simulation.
- Aspects of simulator design, like how experimental factors or multivariate associations are modeled, result in trade-offs in simulator generality, faithfulness, and controllability, which are essential considerations when choosing between simulators.
- Despite advances in simulation methodology, all simulators only approximate reality, and it is important to evaluate them according to relevant criteria, including fit-for-purpose, global, or outcome-specific metrics.
- Microbiome simulators can be used for power analysis and benchmarking for a variety of microbiome analysis tasks, including differential abundance testing, network analysis, and data integration.
- We have prepared an online tutorial with public data and reproducible code examples, making it easy to adapt the microbiome simulation concepts reviewed in this article to practical problems encountered in microbiome research.

Funding

This work was supported by the National Institutes of Health/-National Institutes for General Medical Sciences [R01GM152744 to K.S. and R35GM140888 to J.J.L.]; National Institutes of Health [GM112625 to the Computational Genomics Summer Institute, where part of this work took place]; National Science Foundation [DBI-1846216 and DMS-2113754 to J.J.L.]; National Health and Medical Research Council [Investigator Grant GNT2025648 to K.A.L.C.]; and the Maurice H. Belz Fund [travel fellowship to K.S. hosted by K.A.L.C.].

References

1. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 2004;**68**:669–85. https://doi.org/10.1128/MMBR.68.4.669-685.2004
2. Gilbert JA, Blaser MJ, Gregory Caporaso J. *et al*. Current understanding of the human microbiome. *Nat Med* 2018;**24**:392–400. https://doi.org/10.1038/nm.4517
3. Pinto Y, Bhatt AS. Sequencing-based analysis of microbiomes. *Nat Rev Genet* 2024;**25**:829–45. https://doi.org/10.1038/s41576-024-00746-6
4. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;**10**:e1003531. https://doi.org/10.1371/journal.pcbi.1003531
5. Weiss S, Xu ZZ, Peddada S. *et al*. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 2017;**5**:27. https://doi.org/10.1186/s40168-017-0237-y
6. 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
7. Gerber GK. *Longitudinal Microbiome Data Analysis*. Amsterdam: Elsevier, 2015, 97–111.
8. Duvallet C, Gibbons SM, Gurry T. *et al*. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* 2017;**8**:1784. https://doi.org/10.1038/s41467-017-01973-8
9. Knight R, Vrbanac A, Taylor BC. *et al*. Best practices for analysing microbiomes. *Nat Rev Microbiol* 2018;**16**:410–22. https://doi.org/10.1038/s41579-018-0029-9
10. Kodikara S, Ellul S, Cao K-AL. Statistical challenges in longitudinal microbiome data analysis. *Brief Bioinform* 2022;**23**:1–18. https://doi.org/10.1093/bib/bbac273
11. Zhang Y, Hedo R, Rivera A. *et al*. Post hoc power analysis: is it an informative and meaningful analysis? *General psychiatry* 2019;**32**:e100069. https://doi.org/10.1136/gpsych-2019-100069
12. Kelly BJ, Gross R, Bittinger K. *et al*. Power and sample-size estimation for microbiome studies using pairwise distances and permanova. *Bioinformatics* 2015;**31**:2461–8. https://doi.org/10.1093/bioinformatics/btv183
13. Ferdous T, Jiang L, Dinu I. *et al*. The rise to power of the microbiome: power and sample size calculation for microbiome studies. *Mucosal Immunol* 2022;**15**:1060–70. https://doi.org/10.1038/s41385-022-00548-1
14. Rahman G, McDonald D, Gonzalez A. *et al*. Determination of effect sizes for power analysis for microbiome studies using large microbiome databases. *Genes* 2023;**14**:1239. https://doi.org/10.3390/genes14061239
15. Clarke TH, Greco C, Brinkac L. *et al*. MPrESS: an R-package for accurately predicting power for comparisons of 16s rRNA microbiome taxa distributions including simulation by dirichlet mixture modeling. *Microorganisms* 2023;**11**:1166. https://doi.org/10.3390/microorganisms11051166
16. Friedrich S, Friede T. On the role of benchmarking data sets and simulations in method comparison studies. *Biom J* 2024;**66**:e2200212. https://doi.org/10.1002/bimj.202200212
17. Gossmann A, Zille P, Calhoun V. *et al*. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. *IEEE Trans Med Imaging* 2018;**37**:1761–74. https://doi.org/10.1109/TMI.2018.2815583
18. Wenxing H, Lin D, Cao S. *et al*. Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia. *IEEE Trans Biomed Eng* 2018;**65**:390–9.
19. Chen J, Bushman FD, Lewis JD. *et al*. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 2013;**14**:244–58. https://doi.org/10.1093/biostatistics/kxs038
20. Sankaran K, Holmes SP. Multitable methods for microbiome data integration. *Front Genet* 2019;**10**:627. https://doi.org/10.3389/fgene.2019.00627
21. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;**18**:174. https://doi.org/10.1186/s13059-017-1305-0

22. 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
23. Muller E, Algavi YM, Borenstein E. The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *NPJ Biofilms Microbiomes* 2022;**8**:79. https://doi.org/10.1038/s41522-022-00345-5
24. Wieder C, Lai RPJ, Ebbels TMD. Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics* 2022;**23**:481. https://doi.org/10.1186/s12859-022-05005-1
25. Pezoulas VC, Zaridis DI, Mylona E. *et al*. Synthetic data generation methods in healthcare: a review on open-source tools and methods. *Comput Struct Biotechnol J* 2024;**23**:2892–910. https://doi.org/10.1016/j.csbj.2024.07.005
26. Sazal M, Mathee K, Ruiz-Perez D. *et al*. Inferring directional relationships in microbial communities using signed Bayesian networks. *BMC Genomics* 2020;**21**:663. https://doi.org/10.1186/s12864-020-07065-0
27. Maringanti VS, Bucci V, Gerber GK. MDITRE: scalable and interpretable machine learning for predicting host status from temporal microbiome dynamics. *mSystems* 2022;**7**:e0013222. https://doi.org/10.1128/msystems.00132-22
28. Li H, Zhang Z, Squires M. *et al*. Scmultisim: simulation of multimodality single cell data guided by cell-cell interactions and gene regulatory networks biorXiv. 2023.
29. Crowell HL, Morillo SX, Leonardo CS. *et al*. The shaky foundations of simulating single-cell RNA sequencing data. *Genome Biol* 2023;**24**:62. https://doi.org/10.1186/s13059-023-02904-1
30. Risso D, Perraudeau F, Gribkova S. *et al*. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 2018;**9**:284. https://doi.org/10.1038/s41467-017-02554-5
31. He M, Zhao N, Satten GA. MIDASim: a fast and simple simulator for realistic microbiome data. *Microbiome* 2024;**12**.
32. Williams J, Bravo HC, Tom J. *et al*. MicrobiomeDASim: simulating longitudinal differential abundance for microbiome data. *F1000Res* 2020;**8**:1769. https://doi.org/10.12688/f1000research.20660.2
33. 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
34. Fritz A, Hofmann P, Majda S. *et al*. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 2019;**7**:17. https://doi.org/10.1186/s40168-019-0633-6
35. Choi JM, Ji M, Watson LT. *et al*. DeepMicroGen: a generative adversarial network-based method for longitudinal microbiome data imputation. *Bioinformatics* 2023;**39**. https://doi.org/10.1093/bioinformatics/btad286
36. Rong R, Jiang S, Lin X. *et al*. MB-GAN: microbiome simulation via generative adversarial network. *Gigascience* 2021;**10**:1–11. https://doi.org/10.1093/gigascience/giab005
37. Patuzzi I, Baruzzo G, Losasso C. *et al*. metaSPARSim: a 16S rRNA gene sequencing count data simulator. *BMC Bioinformatics* 2019;**20**:416. https://doi.org/10.1186/s12859-019-2882-6
38. Sohn MB, Li H. A GLM-based latent variable ordination method for microbiome samples. *Biometrics* 2018;**74**:448–57. https://doi.org/10.1111/biom.12775
39. Zeng Y, Pang D, Zhao H. *et al*. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *J Am Stat Assoc* 2022;**118**:1–31.
40. Joe H. *Dependence Modeling with Copulas*. London, England: Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 2023.
41. Deek RA, Li H. Inference of microbial covariation networks using copula models with mixture margins. *Bioinformatics* 2023;**39**. https://doi.org/10.1093/bioinformatics/btad413
42. Sun T, Song D, Li WV. *et al*. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol* 2021;**22**:163. https://doi.org/10.1186/s13059-021-02367-2
43. Song D, Wang Q, Yan G. *et al*. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat Biotechnol* 2023;**42**:247–52. https://doi.org/10.1038/s41587-023-01772-1
44. Sankaran K, Holmes SP. Latent variable modeling for the microbiome. *Biostatistics* 2019;**20**:599–614. https://doi.org/10.1093/biostatistics/kxy018
45. Kohnert E, Kreutz C. Computational study protocol: leveraging synthetic data to validate a benchmark study for differential abundance tests for 16s microbiome sequencing data. *F1000Res* 2024;**13**:1180. https://doi.org/10.12688/f1000research.155230.1
46. Yang L, Chen J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome* 2022;**10**:130. https://doi.org/10.1186/s40168-022-01320-0
47. Wirbel J, Essex M, Forslund SK. *et al*. A realistic benchmark for differential abundance testing and confounder adjustment in human microbiome studies. *Genome Biol* 2024;**25**:247. https://doi.org/10.1186/s13059-024-03390-9
48. Gao Y, Şimşek Y, Gheysen E. *et al*. MiaSim: an R/Bioconductor package to easily simulate microbial community dynamics. *Methods Ecol Evol* 2023;**14**:1967–80. https://doi.org/10.1111/2041-210X.14129
49. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C Appl Stat* 2005;**54**:507–54. https://doi.org/10.1111/j.1467-9876.2005.00510.x
50. Mikis, Stasinopoulos D, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) inr. *J Stat Softw* 2007;**23**:1–46. https://doi.org/10.18637/jss.v023.i07
51. Sczyrba A, Hofmann P, Belmann P. *et al*. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat Methods* 2017;**14**:1063–71. https://doi.org/10.1038/nmeth.4458
52. Meyer F, Fritz A, Deng Z-L. *et al*. Critical assessment of metagenome interpretation: the second round of challenges. *Nat Methods* 2022;**19**:429–40. https://doi.org/10.1038/s41592-022-01431-4
53. Ren B, Bacallado S, Favaro S. *et al*. Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis. *Ann Appl Stat* 2020;**14**:494–17. https://doi.org/10.1214/19-AOAS1295
54. Zeng Y, Li J, Wei C. *et al*. MbDenoise: microbiome data denoising using zero-inflated probabilistic principal components analysis. *Genome Biol* 2022;**23**:94. https://doi.org/10.1186/s13059-022-02657-3
55. Jiang S, Xiao G, Koh AY. *et al*. A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* 2019;**22**:522–40. https://doi.org/10.1093/biostatistics/kxz050
56. Drechsler J, Haensch A-C. 30 years of synthetic data. *Stat Sci* 2024;**39**:221–42. https://doi.org/10.1214/24-STS927

57. Duong T, Goud B, Schauer K. Closed-form density-based framework for automatic detection of cellular morphology changes. *Proc Natl Acad Sci* 2012;**109**:8382–7. https://doi.org/10.1073/pnas.1117796109
58. Friedman JH, Rafsky LC. Graph-theoretic measures of multivariate association and prediction. *Ann Stat* 1983;**11**:377–91. https://doi.org/10.1214/aos/1176346148
59. El Emam K, Mosquera L, Hoptroff R. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. Sebastapol: O'Reilly Media, 2020.
60. Karr AF, Kohnen CN, Oganian A. *et al*. A framework for evaluating the utility of data altered to protect confidentiality. *Am Stat* 2006;**60**:224–32. https://doi.org/10.1198/000313006X124640
61. El Emam K, Mosquera L, Fang X. *et al*. An evaluation of the replicability of analyses using synthetic health data. *Sci Rep* 2024;**14**:6978. https://doi.org/10.1038/s41598-024-57207-7
62. Ullmann T, Beer A, Hünemörder M. *et al*. Over-optimistic evaluation and reporting of novel cluster algorithms: an illustrative study. *Adv Data Anal Classif* 2022;**17**:211–38.
63. Pawel S, Kook L, Reeve K. Pitfalls and potentials in simulation studies: questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biom J* 2023;**66**. https://doi.org/10.1002/bimj.202200091
64. Nießl C, Hoffmann S, Ullmann T. *et al*. Explaining the optimistic performance evaluation of newly proposed methods: a cross–design validation experiment. *Biom J* 2023;**66**. https://doi.org/10.1002/bimj.202200238
65. Lahti L, Salojärvi J, Salonen A. *et al*. Tipping elements in the human intestinal ecosystem. *Nat Commun* 2014;**5**:4344. https://doi.org/10.1038/ncomms5344
66. Gavin PG, Mullaney JA, Loo D. *et al*. Intestinal metaproteomics reveals host-microbiota interactions in subjects at risk for type 1 diabetes. *Diabetes Care* 2018;**41**:2178–86. https://doi.org/10.2337/dc18-0777
67. McDonald D, Hyde E, Debelius JW. *et al*. American gut: an open platform for citizen science microbiome research. *mSystems* 2018;**3**. https://doi.org/10.1128/msystems.00031-18.
68. Chapleur O, Madigou C, Civade R. *et al*. Increasing concentrations of phenol progressively affect anaerobic digestion of cellulose and associated microbial communities. *Biodegradation* 2015;**27**:15–27. https://doi.org/10.1007/s10532-015-9751-4
69. Haak BW, Argelaguet R, Kinsella CM. *et al*. Integrative transkingdom analysis of the gut microbiome in antibiotic perturbation and critical illness. *mSystems* 2021;**6**. https://doi.org/10.1128/msystems.01148-20
70. Gloor GB, Macklaim JM, Pawlowsky-Glahn V. *et al*. Microbiome datasets are compositional: and this is not optional. *Front Microbiol* 2017;**8**:2224. https://doi.org/10.3389/fmicb.2017.02224
71. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21. https://doi.org/10.1186/s13059-014-0550-8
72. Law CW, Chen Y, Shi W. *et al*. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:1–17.
73. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. *Nat Commun* 2020;**11**:3514. https://doi.org/10.1038/s41467-020-17041-7
74. Duong T. Ks: kernel density estimation and kernel discriminant analysis for multivariate data inr. *J Stat Softw* 2007;**21**:1–16. https://doi.org/10.18637/jss.v021.i07
75. Olbrich M, Künstner A, Busch H. MBECS: microbiome batch effects correction suite. *BMC Bioinformatics* 2023;**24**:182. https://doi.org/10.1186/s12859-023-05252-w
76. Cao K-AL, Welham ZM. *Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package*. New York: Chapman and Hall/CRC, 2021. https://doi.org/10.1201/9781003026860.
77. Andreella A, Fino L, Scarpa B *et al*. Towards a power analysis for PLS-based methods arXiv. 2024.
78. Johnstone IM, Paul D. PCA in high dimensions: an orientation. *Proc IEEE* 2018;**106**:1277–92. https://doi.org/10.1109/JPROC.2018.2846730
79. Skalski JR, Richins SM, Townsend RL. A statistical test and sample size recommendations for comparing community composition following PCA. *PloS One* 2018;**13**:e0206033. https://doi.org/10.1371/journal.pone.0206033
80. Guerra-Urzola R, Van Deun K, Vera JC. *et al*. A guide for sparse PCA: model comparison and applications. *Psychometrika* 2021;**86**:893–919. https://doi.org/10.1007/s11336-021-09773-2
81. Cao K-AL, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* 2011;**12**:253. https://doi.org/10.1186/1471-2105-12-253
82. Cai T, Liu W. Adaptive thresholding for sparse covariance matrix estimation. *J Am Stat Assoc* 2011;**106**:672–84. https://doi.org/10.1198/jasa.2011.tm10560
83. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Series B Stat Methodology* 2003;**66**:187–205.
84. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 2006;**93**:491–507. https://doi.org/10.1093/biomet/93.3.491
85. Blanchard G, Roquain E. Adaptive false discovery rate control under independence and dependence. *J Mach Learn Res* 2009;**10**:2837–71.
86. Jiang D, Armour CR, Chenxiao H. *et al*. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Front Genet* 2019;**10**:995. https://doi.org/10.3389/fgene.2019.00995
87. Fabbrini M, Scicchitano D, Candela M. *et al*. Connect the dots: sketching out microbiome interactions through networking approaches. *Microbiome Res Rep* 2023;**2**:25. https://doi.org/10.20517/mrr.2023.25
88. Ozminkowski S, Solís-Lemus C. Identifying microbial drivers in biological phenotypes with a Bayesian network regression model. *Ecol Evol* 2024;**14**:e11039. https://doi.org/10.1002/ece3.11039
89. Chevrette MG, Bratburd JR, Currie CR. *et al*. Experimental microbiomes: models not to scale. *mSystems* 2019;**4**. https://doi.org/10.1128/mSystems.00175-19
90. Kurtz ZD, Müller CL, Miraldi ER. *et al*. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 2015;**11**:e1004226. https://doi.org/10.1371/journal.pcbi.1004226
91. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *J Multivar Anal* 2004;**88**:365–411. https://doi.org/10.1016/S0047-259X(03)00096-4
92. Kurtz Z, Mueller C, Miraldi E. *et al*. *SpiecEasi: Sparse Inverse Covariance for Ecological Statistical Inference* 2024. R package version 1.1.3, commit 5f396da85baa114b31c13d9744c05387a1b04c23.

93. Shade A, Handelsman J. Beyond the venn diagram: the hunt for a core microbiome. *Environ Microbiol* 2012;**14**:4–12. https://doi.org/10.1111/j.1462-2920.2011.02585.x
94. Peeters CFW, Bilgrau AE, van Wieringen WN. rags2ridges: a one-stop-ℓ_2-shop for graphical modeling of high-dimensional precision matrices. *J Stat Softw* 2022;**102**:1–32.
95. Ravikumar P, Wainwright MJ, Raskutti G. *et al*. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electron J Stat* 2011;**5**:935–980. https://doi.org/10.1214/11-EJS631
96. Wang Y, Cao K-AL. Managing batch effects in microbiome data. *Brief Bioinform* 2019;**21**:1954–70. https://doi.org/10.1093/bib/bbz105
97. Luecken MD, Büttner M, Chaichoompu K. *et al*. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
98. Leek JT, Scharpf RB, Bravo HC. *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9. https://doi.org/10.1038/nrg2825
99. Gibbons SM, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome studies. *PLoS Comput Biol* 2018;**14**:e1006102. https://doi.org/10.1371/journal.pcbi.1006102
100. Ling W, Jiuyao L, Zhao N. *et al*. Batch effects removal for microbiome data via conditional quantile regression. *Nat Commun* 2022;**13**:5418. https://doi.org/10.1038/s41467-022-33071-9
101. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 2015;**17**:29–39. https://doi.org/10.1093/biostatistics/kxv027
102. Soneson C, Gerster S, Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS One* 2014;**9**:e100335. https://doi.org/10.1371/journal.pone.0100335
103. Zhang Z, Mathew D, Lim T. *et al*. Signal recovery in single cell batch integration bioRxiv.org. *Nature Biotechnology* 2024. https://doi.org/10.1038/s41587-024-02463-1.
104. Molania R, Gagnon-Bartsch JA, Dobrovic A. *et al*. A new normalization for nanostring ncounter gene expression data. *Nucleic Acids Res* 2019;**47**:6073–83. https://doi.org/10.1093/nar/gkz433
105. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27. https://doi.org/10.1093/biostatistics/kxj037
106. Valles-Colomer M, Menni C, Berry SE. *et al*. Cardiometabolic health, diet and the gut microbiome: a meta-omics perspective. *Nat Med* 2023;**29**:551–61. https://doi.org/10.1038/s41591-023-02260-4
107. Chetty A, Blekhman R. Multi-omic approaches for host-microbiome data integration. *Gut Microbes* 2024;**16**:2297860. https://doi.org/10.1080/19490976.2023.2297860
108. Ma R, Sun ED, Donoho D. *et al*. Principled and interpretable alignability testing and integration of single-cell data. *Proc Natl Acad Sci* 2024;**121**:e2313719121. https://doi.org/10.1073/pnas.2313719121
109. Singh A, Shannon CP, Gautier B. *et al*. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;**35**:3055–62. https://doi.org/10.1093/bioinformatics/bty1054
110. Schreck N, Slynko A, Saadati M. *et al*. Statistical plasmode simulations–potentials, challenges and recommendations. *Stat Med* 2024;**43**:1804–25. https://doi.org/10.1002/sim.10012
111. Franklin JM, Schneeweiss S, Polinski JM. *et al*. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal* 2014;**72**:219–26. https://doi.org/10.1016/j.csda.2013.10.018
112. Jia B, Xuan L, Cai K. *et al*. NeSSM: a next-generation sequencing simulator for metagenomics. *PloS One* 2013;**8**:e75448. https://doi.org/10.1371/journal.pone.0075448
113. Johnson S, Trost B, Long JR. *et al*. A better sequence-read simulator program for metagenomics. *BMC Bioinformatics* 2014;**15**:S14. https://doi.org/10.1186/1471-2105-15-S9-S14
114. Gourlé H, Karlsson-Lindsjö O, Hayer J. *et al*. Simulating illumina metagenomic data with InSilicoSeq. *Bioinformatics* 2019;**35**:521–2. https://doi.org/10.1093/bioinformatics/bty630
115. Yan G, Song D, Li JJ. scReadSim: a single-cell RNA-seq and ATAC-seq read simulator. *Nat Commun* 2023;**14**:7482. https://doi.org/10.1038/s41467-023-43162-w
116. Yan Y, Nguyen LH, Franzosa EA. *et al*. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Med* 2020;**12**:71. https://doi.org/10.1186/s13073-020-00765-y
117. Johnson JS, Spakowicz DJ, Hong B-Y. *et al*. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019;**10**:5029. https://doi.org/10.1038/s41467-019-13036-1
118. Senthil Kumar M, Slud EV, Hehnly C. *et al*. Differential richness inference for 16S rRNA marker gene surveys. *Genome Biol* 2022;**23**:166. https://doi.org/10.1186/s13059-022-02722-x
119. Tonkin-Hill G. GitHub - gtonkinhill/TCGA_analysis — github.com. https://github.com/gtonkinhill/TCGA_analysis. 2023. (21 June 2024, date last accessed).
120. Daybog I, Kolodny O. A computational framework for resolving the microbiome diversity conundrum. *Nat Commun* 2023;**14**:7977. https://doi.org/10.1038/s41467-023-42768-4

Briefings in Bioinformatics, 2025, **26(1)**, bbaf051
https://doi.org/10.1093/bib/bbaf051
Review