Multivariable Association Discovery in Population-scale Meta-omics Studies

Himel Mallick[1,2], Ali Rahnavard[3], Lauren J. McIver[1,2], Siyuan Ma[1,2], Yancong Zhang[1,2], Long H. Nguyen[1,4,5], Timothy L. Tickle[2], George Weingart[1,2], Boyu Ren[1,2], Emma H. Schwager[1,2], Suvo Chatterjee[6], Kelsey N. Thompson[1], Jeremy E. Wilkinson[1], Ayshwarya Subramanian[1,2], Yiren Lu[1], Levi Waldron[7], Joseph N. Paulson[8], Eric A. Franzosa[1,2], Hector Corrada Bravo[9], Curtis Huttenhower[1,2*]

[1]Biostatistics Department, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA

[2]The Broad Institute, 415 Main Street, Cambridge, MA 02142, USA

[3]Computational Biology Institute, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC 20052, USA

[4]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02144, USA

[5]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02144, USA

[6]Epidemiology Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA

[7]Department of Epidemiology and Biostatistics, CUNY School of Public Health, NY 10027, USA

[8]Department of Biostatistics, Product Development, Genentech, Inc., South San Francisco, CA 94080, USA

[9]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA

*Correspondence to chuttenh@hsph.harvard.edu

Abstract

It is challenging to associate features such as human health outcomes, diet, environmental conditions, or other metadata to microbial community measurements, due in part to their quantitative properties. Microbiome multi-omics are typically noisy, sparse (zero-inflated), high-dimensional, extremely non-normal, and often in the form of count or compositional measurements. Here we introduce an optimized combination of novel and established methodology to assess multivariable association of microbial community features with complex metadata in population-scale observational studies. Our approach, MaAsLin 2 (Microbiome Multivariable Associations with Linear Models), uses general linear models to accommodate a wide variety of modern epidemiological studies, including cross-sectional and longitudinal designs, as well as a variety of data types (e.g. counts and relative abundances) with or without covariates and repeated measurements. To construct this method, we conducted a large-scale evaluation of a broad range of scenarios under which straightforward identification of meta-omics associations can be challenging. These simulation studies reveal that MaAsLin 2's linear model preserves statistical power in the presence of repeated measures and multiple covariates, while accounting for the nuances of meta-omics features and controlling false discovery. We also applied MaAsLin 2 to a microbial multi-omics dataset from the Integrative Human Microbiome (HMP2) project which, in addition to reproducing established results, revealed a unique, integrated landscape of inflammatory bowel disease (IBD) across multiple time points and omics profiles.

Keywords: Human Microbiome, Metagenomics, Differential Abundance Analysis, Multivariable Association, Microbiome Epidemiology, Longitudinal Analysis

Introduction

Human-associated microbiota has been convincingly linked to the development and progression of a wide range of complex, chronic conditions including inflammatory bowel diseases (IBD), obesity, diabetes, cancer, and cardiovascular disorders[1,2]. Due to recent advances in multiple high-throughput functional profiling technologies, research has expanded well beyond bacteria-specific 16S rRNA gene amplicon profiles to multi-omics surveys, i.e. non-bacterial, metagenomic, metatranscriptomic, metabolomic, and metaproteomic measurements assessed simultaneously in the same biological specimens[3,4]. Additionally, due to diminishing sequencing costs, longitudinal, within-subject study designs are becoming increasingly common, especially when assessing the microbiome in population health[5,6]. These large, complex data contain abundant information to enable microbe-, gene-, and compound-specific hypothesis generation at scale. However, robust quantitative methods to do so at scale can still be challenging to implement without excessive false positives - one of the main hurdles in accurate translational applications of the microbiome to human health.

One of the primary limitations of leveraging such population-wide multi-omics surveys is thus computational, in part due to the complexity and heterogeneity of microbial community data that have made reliable application of statistical methods difficult. In particular, best practices to guard against spurious discoveries in meta-omics datasets remain scarce[7-14]. High-throughput meta-omics datasets have specific characteristics that complicate their analyses: high-dimensionality, count and compositional data structure, sparsity (zero-inflation), over-dispersion, and hierarchical, spatial, and temporal dependence, among others. To combat these challenges, specialized methods implemented in usable, reproducible software are needed to accurately characterize microbial communities within large human population studies, while maintaining both sensitivity and specificity.

Early advances in microbiome epidemiology focused on omnibus testing for identifying overall associations between aggregate microbiome structure and host or environmental phenotypes and covariates (e.g. disease status, diet, antibiotics or medication usage, age, etc.). Many of these rely on permutation-based procedures for moderated significance testing[11]. These methods assess whether overall community patterns of variation are associated with the covariates of interest, but fail to provide feature-level inference to enable follow-up characterization (where a feature can be any profiled omics abundance, e.g. taxa, genes, pathways, chemicals, etc.) To facilitate actionable outcomes, it is important to discern feature-specific associations at the highest possible resolution. This has led to the development of a variety of per-feature (or feature-wise) association testing methods, most of which are based on similar statistical frameworks, differing primarily in (i) the choice of normalization or transformation, (ii) observation model or likelihood, and (iii) the associated statistical inference[11]. As compared to omnibus testing approaches, per-feature methods (i) identify associations for each individual feature-metadata pair, (ii) facilitate feature-wise covariate adjustment, and (iii) call out specific features (as opposed to complex combinations of features implicated in associations in omnibus testing), leading to increased interpretability for translational and basic biological applications.

Despite a rich literature on feature-wise association testing for microbial communities, methods that can accommodate a wide variety of modern epidemiological study designs remain scarce. For instance, many early methods do not explicitly account for the sparsity observed in microbial meta-omics observations, and only a few scale beyond routine univariate (differential abundance) analyses without becoming overly susceptible to false positive or false negative results[7,11]. Furthermore, most methods for microbiome data do not explicitly adjust for repeated measures and multiple covariates in a unified statistical framework, a lack of which can have a profound (and typically anti-conservative) impact on subsequent epidemiological inference.

Here, we address these issues by providing a flexible approach to identify multivariable associations in large, heterogeneous meta-omics datasets. We have implemented this method as MaAsLin 2 (Microbiome Multivariable Associations with Linear Models, with software version 2.0 released with this study), a successor to MaAsLin 1[15,16]. Unlike MaAsLin 1's single-model framework based on applications of arcsine square root-transformed linear model following Total Sum Scaling (TSS) normalization[15,16], MaAsLin 2 has evaluated and combined the best set of analysis steps to facilitate high-precision association discovery in microbiome epidemiology studies. It provides a coherent paradigm through a multi-model framework with arbitrary coefficients (phenotypes and covariates) and contrasts of interest, along with support for data exploration, normalization, and transformation options to aid in the selection of appropriate data-

and design-driven statistical techniques for analyzing microbial multi-omics data. In this study, we also conducted a large-scale synthetic evaluation of a broad range of circumstances under which straightforward identification of meta-omics features can be challenging. These simulation studies revealed that MaAsLin 2 preserves statistical power in the presence of repeated measurements and multiple covariates while accounting for the nuances of meta-omics features and, critically, controlling false discovery rates. We concluded with an application to novel biomarker discovery in multiple omics datasets from the Integrative Human Microbiome Project (iHMP or HMP2[6]). The implementation of MaAsLin 2, associated documentation and tutorial, and example data sets are freely available in the MaAsLin 2 R/Bioconductor software package at https://huttenhower.sph.harvard.edu/maaslin2.

Results

MaAsLin 2 methodology and validation

MaAsLin 2 provides a comprehensive multi-model system for performing multivariable association testing in microbiome profiles - taxonomic, functional, or metabolomic - with analysis modules for preprocessing, normalization, transformation, and data-driven statistical modeling to tackle the challenges of microbial multi-omics (compositionality, overdispersion, zero-inflation, variable library size, high-dimensionality, etc.; **Fig. 1A**). The MaAsLin 2 implementation requires two inputs: (i) microbial feature abundances (e.g. taxa, genes, transcripts, or metabolites) across samples, in either counts or relative counts; and (ii) environmental, clinical, or epidemiological phenotypes or covariates of interest (together “metadata”). Both metadata and microbial features are first processed for missing values, unknown data values, and outliers. If indicated, microbial measurements are then normalized and transformed to address variable depth of coverage across samples. Feature standardization is optionally performed, and a subset or the full complement of metadata are used to model the resulting quality-controlled microbial features and define p-values for each metadata association per feature using one of a wide range of possible multivariable models. After all features are evaluated, p-values are adjusted for multiple hypothesis testing and a table summarizing statistically significant associations is reported. While the default MaAsLin 2 implementation uses a log-transformed linear model on TSS-normalized quality-controlled data, the software supports several other statistical models including GLM (e.g. Negative Binomial[17]), zero-adjusted models (e.g. Compound Poisson[18,19], ZINB[20]), and multiple normalization/transformation schemes under one estimation umbrella. In the presence of repeated measures, MaAsLin 2 additionally identifies covariate-associated microbial features by appropriately modeling the within-subject (or environment) correlations in a mixed model paradigm, while also accounting for inter-individual variability by specifying between-subject random effects in the model. A variety of summary and diagnostic plots are also provided to visualize the top results.

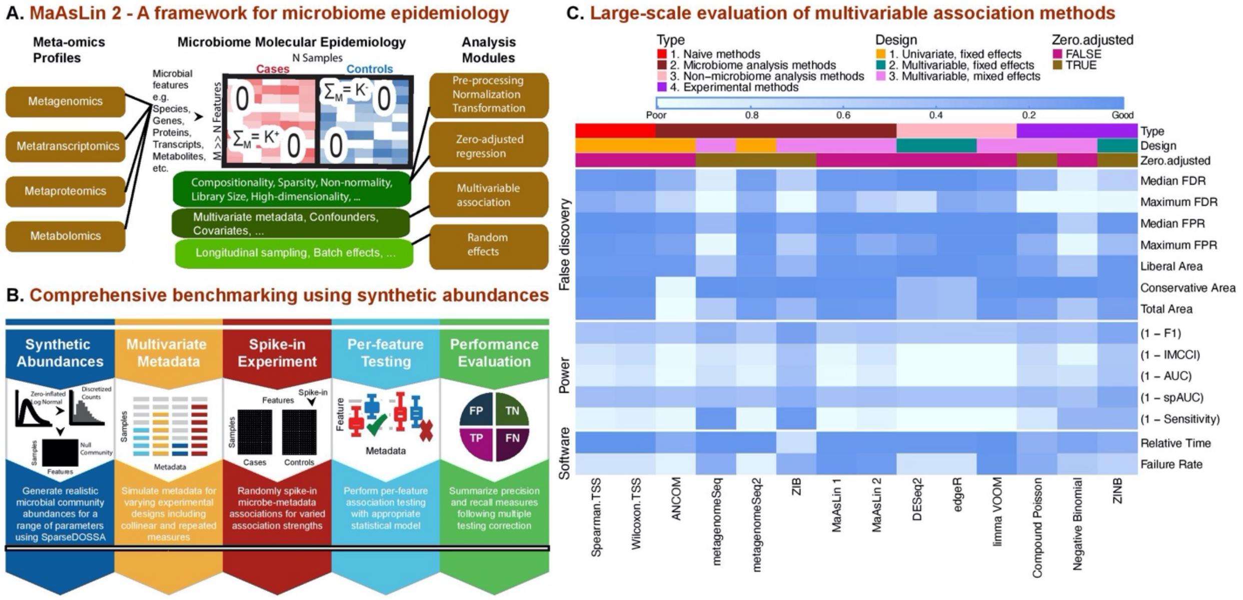


Figure 1: MaAsLin 2 for feature-wise association of microbial communities with phenotypes. A) MaAsLin 2 is a statistical method for association analysis of microbial community meta-omics profiles. It comprises several steps, including data transformation, multivariable inference, multiple hypothesis test correction, and visualization. These are based on a set of flexible and computationally efficient linear models, while accounting for the nuances of microbiome data, repeated measures, and multiple covariates. **B)** Comprehensive benchmarking of multivariable methods for microbiome epidemiology. To identify appropriate methods for associating microbiome features with health outcomes and other covariates, we assessed up to 84 combinations of normalization/transformation, zero-inflation, and regression models (**Supplementary Fig. S1A**). These were applied to synthetic data using a hierarchical model (SparseDOSSA, http://huttenhower.sph.harvard.edu/sparsedossa) to generate realistic, model-agnostic datasets with varying scopes and effect sizes of microbiome associations**.** Individual per-feature association methods were performed repeatedly to evaluate

method-specific recall and precision measures. **C)** Association method performance summary across major evaluation criteria. Three aspects of performance were considered: (i) false discovery, (ii) sensitivity, and (iii) computational efficiency. Evaluation metrics are shown (in rows) for the resulting microbial multivariable association methods (both state-of-the-art and novel), averaged over all simulation parameters (**Supplementary Fig. S1B**). The top-performing methods (as measured by average F1 score) from each class of models (**Supplementary Fig. S1C**) are shown (in columns). With the exception of Spearman and Wilcoxon that maintained best performance on TSS-normalized data, all methods exhibited superior performance with no/default normalization (ANCOM, metagenomeSeq, metagenomeSeq2, DESeq2, edgeR, MaAsLin 1, MaAsLin 2, limma VOOM, ZIB) or library size normalization in which log-transformed library size is included as an offset in the associated GLM likelihood (Compound Poisson, Negative Binomial, ZINB). Top colored boxes represent method characteristics including the capability to handle zero-inflation and random effects. Based on synthetic evaluations, MaAsLin 2 includes optimized default models for epidemiological testing in microbial multi-omics data.

To identify model components appropriate for MaAsLin 2's microbiome feature association testing and to objectively benchmark current association methods, we assessed realistic synthetic datasets generated by SparseDOSSA (full details of individual association methods, as well as simulation parameters, are described in **Methods** and are available online at https://github.com/biobakery/maaslin2_benchmark). Briefly, SparseDOSSA is a synthetic data generation routine that models biologically plausible synthetic data from diverse template microbiome profiles by taking into account (i) feature-feature, (ii) feature-metadata, and (iii) metadata-metadata correlations, superseding previous efforts by including multiple covariates and longitudinal designs (**Methods**). As compared to previous simulation schemes, SparseDOSSA allows multivariable spike-in both in the presence and absence of repeated measures, as well as arbitrary covariance structure in the metadata design matrix.

For this study, we carried out several spike-in experiments to induce and test controlled associations, as governed by configurable simulation parameters (**Supplementary Fig. S1**). When used for this purpose, SparseDOSSA first generates null microbial community features containing no significant association patterns using a Bayesian hierarchical model independently of metadata features (**Fig. 1B**, **Methods**). In addition to varying sample size and feature dimension, a broad range of metadata and experimental designs are then considered, including repeated measures and univariate and multivariate covariates (both continuous and binary) of varying dimension and effect size (**Supplementary Fig. S1A**). Specifically, in each instance, we varied sample sizes from small (10) to large (200) for a fixed feature size (up to 500), and within each sample size, the effect size parameter was again varied from modest (e.g. <2-fold differences) to strong (10-fold). In each simulation, 10% of features (and 20% of metadata for multivariable scenarios) were modified as an in-silico spike-in. Precision and recall measures were averaged over 100 simulation runs (**Supplementary Fig. S1B, Methods)**. All methods were corrected for multiple hypothesis testing using standard approaches for FDR control, declaring significant associations at a target of FDR 0.05. For a fair comparison, a basic, model-free filtering step to remove low-abundance features was performed before statistical modeling for all methods. Methods unable to process specific simulation configurations due to high computational overhead or slow convergence were omitted for those cases.

To compare the detection power of various methods in identifying true positive feature associations, we first comprehensively evaluated published metagenomic tools and representative methods for bulk RNA-seq analysis within each simulation scenario. These methods were combined with several microbiome-appropriate normalization, transformation, and linkage models (**Supplementary Fig. S1C, Methods**). In particular, we considered six distinct categories of methods in our evaluations: (i) published methods specifically designed for microbial community, such as metagenomeSeq[21], ANCOM[14,22], and ZIB[23,24], (ii) published bulk RNA-seq differential expression methods, such as DESeq2[25], edgeR[26], and limma VOOM[27,28]; (iii) existing generalized linear model (GLM) count models, such as the negative binomial[17], (iv) methods based on linear models, such as limma[29] and “pure” linear models (LMs); (v) representative zero-adjusted methods from the microbiome and scRNA-seq literature such as the compound Poisson[18,19] and the zero-inflated negative binomial (ZINB[20,30]); and finally (vi) traditional, simplistic non-parametric methods, such as Spearman correlation and Wilcoxon tests. Of note, many of these methods can only compare two groups (i.e. a single binary metadatum), and not all are compatible with continuous and multivariate metadata, resulting in a distinct set of comparable methods for each experimental design.

Our first consideration in designing MaAsLin 2 for microbiome epidemiology was to ensure that both current statistical theory and practical issues were considered during the analysis of microbiome multi-omics data. To this end, we rigorously characterized various finite-sample properties of different association methods focusing on three broadly defined aspects: (i) false discovery, (ii) detection power, and (iii) software implementation, with multiple performance indicators for each category (**Fig. 1C**). Rather than focusing on a single evaluation metric like the Area Under the Curve (AUC) or the False Positive Rate (FPR), we ranked methods based on a combination of metrics (**Methods**), many not considered in previous benchmarking. To summarize each evaluation criteria, a normalized continuous score ranging between 0 and 1 was assigned (**Methods**). Methods were then eliminated based on the presence of 'red flags' with respect to at least one evaluation criteria, i.e. extreme departures from the best possible value. Finally, metrics that are mainly descriptive rather than quantitative were also evaluated (e.g. the ability to handle complex metadata designs, zero-inflation, or repeated measures) to achieve a final consensus. This tiered strategy not only allowed us to select a set of "best" methods based on the fewest 'red flags' across all scenarios, but also to identify a method that is (i) sufficiently robust to false discovery control and detection power, (ii) scalable to large multi-omics datasets, and (iii) accommodating of most modern epidemiological designs and microbial data types.

Notably, previous benchmarking in this area has only focused on differential abundance testing without the simultaneous consideration of multiple covariates and repeated measures[7-9]. Additionally, with the exception of Hawinkel et al.[7], these benchmarking efforts lacked important considerations to the extent that they either (i) did not consider FDR as a metric of evaluation[9,31,32] or (ii) misreported false positive rate as FDR[8] (**Methods**). While a majority of these studies made a final recommendation based on the traditional AUC metric or a combination of sensitivity and specificity, we argue that without considering the FDR-controlling behavior of a method, the AUC values alone are too optimistic to draw any meaningful conclusions about its practical utility. In other words, particularly for biological follow-up, high precision among the most confident (lowest recall) predictions is essential. To this end, our large-scale benchmarking enables a progressive unification of traditional and practically important evaluation metrics by providing a comprehensive connected view of microbiome multivariable association methods, especially in the context of modern study designs, multiple covariates, and repeated measures.

Overall, our simulation study revealed that virtually all high-sensitivity methods with an overoptimistic median AUC, especially those targeted to microbial communities, exhibited a highly inflated average FDR (**Fig. 1C**). A similar pattern was observed for other AUC-like measures such as F1 score and Matthew's correlation coefficient (MCC). On the other end of the spectrum, compositionality-corrected methods such as ANCOM exhibited an extreme departure from 'good' performance with respect to several criteria including sensitivity and p-value calibration, as measured by both Conservative and Total Area[7] (**Methods**). Overall, these simulations reveal that while there is no single method that outperforms others in all scenarios, MaAsLin 2 was the only multivariable method tested that controlled FDR with the fewest 'red flags' across scenarios (**Fig. 1C).**

This initial phase of our study thus expands the understanding of statistical association methods appropriate for microbial community data under varying study designs, and it especially highlights the inability of many common methods to control false discoveries. This is of critical importance to past and present microbiome association methods, as failure to control the FDR causes uncertainty in both scientific reproducibility and interpretability. Based on these evaluations, a linear model with TSS normalization and log transformation was adopted as the default model in MaAsLin 2, and the software provides several flexible options to apply a combination of other normalization, transformation, and statistical methods to customize specific analysis tasks. The implementation of MaAsLin 2, associated documentation, and example data sets are freely available in R/Bioconductor and at https://huttenhower.sph.harvard.edu/maaslin2.

MaAsLin 2 controls false discovery rate while maintaining power in differential abundance analysis

Differential abundance testing for microbial community features (taxa, pathways, etc.) is one of the most commonly used strategies to identify features that differ between sample categories such as cases and controls. Despite a large number of developments in the area, a lack of consensus on the most appropriate statistical method has been a major concern[11]. To model experimental designs of this type, we used synthetic count data with spiked-in features differentially abundant between two defined groups of samples. In particular, we multiplied the mean relative abundance of a randomly sampled fraction of 10% of the features with a given effect size (fold change) in one of the groups and renormalized the ensemble of relative abundances to a unit sum to create features differentially abundant between groups. We repeated this procedure for each unique combination of sample size (10, 20, 50, 100, 200), feature dimension (100, 200, 500), and fold change (1, 2, 5, and 10), each time summarizing performance over 100 simulation runs (**Methods**). Before model fitting, features with a low prevalence (<10%) were trimmed from the generated data sets.

As in our overall evaluation (**Fig. 1C**), we observed marked differences between the FDR-controlling behavior of different methods in the simple case of single binary metadata and non-longitudinal design, in some cases exceeding 75% (**Fig. 2**). Among the methods with good, robust FDR control, only those based on linear models achieved moderate power, whereas, for methods such as DESeq2 and edgeR, the FDR control came at the cost of reduced power. Among other methods, practically all count and zero-inflated models, as well as newer methods based on log-ratios such as ANCOM, struggled to correctly control the FDR at the intended (nominal) level, and the best performance in this class of methods was obtained by metagenomeSeq2, Compound Poisson, and ZINB (as measured by the F1 score). Many of the remaining methods were too liberal, with metagenomeSeq and Negative Binomial standing out with a large number of false positive findings. Overall, linear models (LMs) remained critically the only class of methods tested that has good control of FDR across study designs, and they all exhibited a boost in statistical power with increased sample size and association strength (**Supplementary Fig. S2).**

We also evaluated the average FPR of these methods by recording the fraction of tested unassociated (negative) features that were deemed significant following FDR correction. Nearly all methods controlled the FPR well below the imposed level (**Supplementary Fig. S3**). Relatedly, we employed a previously proposed metric called “departure from uniformity” (i.e. departure from uniformity of p-value under the null), which, complementary to FPR, quantifies the liberal or conservative area between observed and theoretical quantiles of a uniform distribution[7] (**Methods**). As expected, methods with high average false discovery rates, including zero-inflated and count models, showed extreme departures from uniformity in the liberal direction, whereas conservative methods such as DESeq2 and edgeR showed the same in the opposite direction, suggesting extreme violation of uniformly distributed p-values under the null hypothesis (**Supplementary Fig. S4**). While these results raise potential concerns about the FDR-controlling behaviors of most existing methods, LM-based approaches did not exhibit this trend. In general, tools based on linear models (such as limma) performed very similarly when calibrated with MaAsLin 2's default model parameters, as expected, but not with their recommended default parameters (**Supplementary Fig. S3**). Additionally, their options for handling sparsity and compositionality were generally not appropriate for microbiome data. Amplicon, metagenomic taxonomic, and functional profiles each show distinct count and zero-inflation properties, for example, that are best handled by a multi-model system. As such, in addition to the binary metadata design, we repeated the above simulation experiments for univariate continuous metadata as well, which led to similar conclusions (**Supplementary Fig. S5**), further supporting MaAsLin 2’s default model’s performance across metadata types and experimental designs.

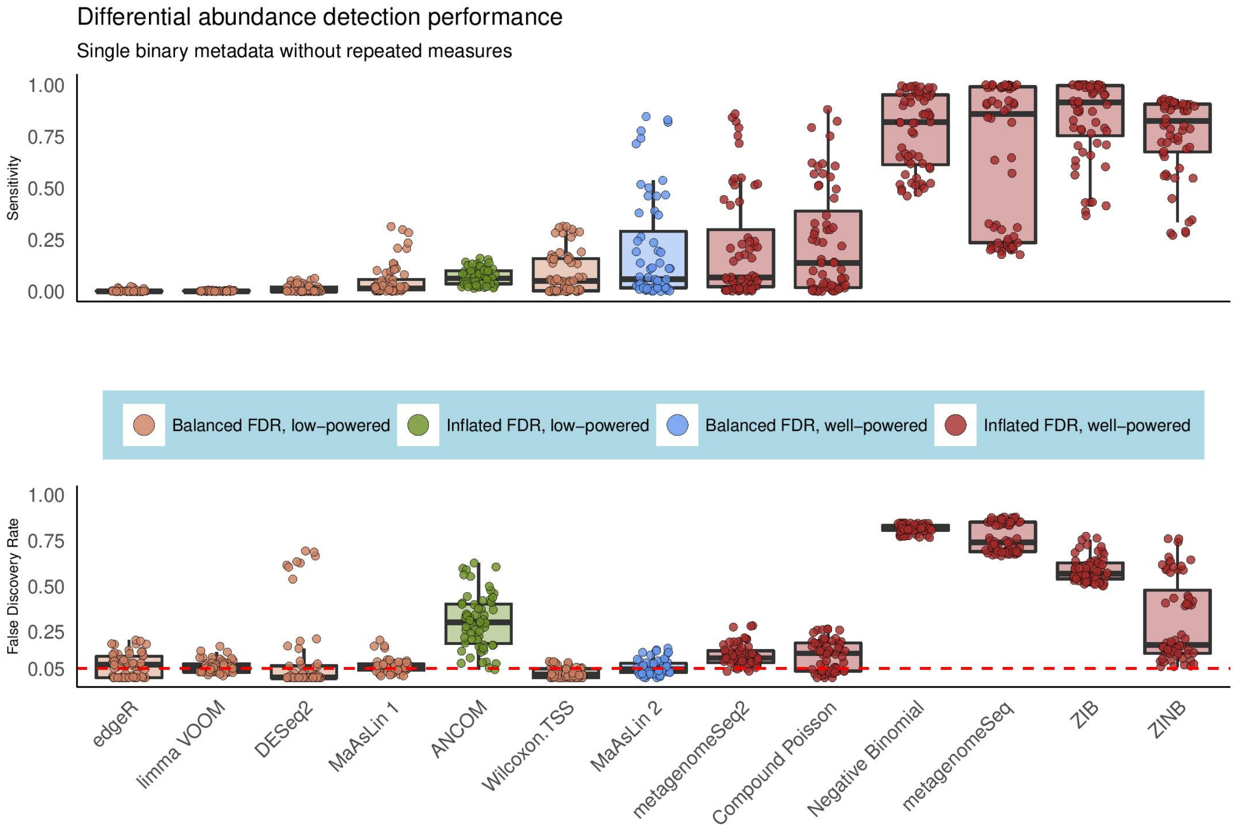


Figure 2: MaAsLin 2 controls false discovery rate while maintaining power in differential abundance analysis of microbial communities. To assess models' behaviors during differential abundance analysis, we simulated 100 independent datasets per parameter combination, each containing a single binary metadatum and a fixed number of true positive features (10% of features differentially abundant) for varying association strengths and sample sizes (**Supplementary Fig. S1C**). We then evaluated the ability of different microbiome association methods to recover these associations using a variety of performance metrics and summarized the results across runs (**Methods**). Both sensitivity and false discovery rates (FDR) are shown for the best-performing method from each class of models (as measured by average F1 score; **Methods**; full results in **Supplementary Figs. S2-S5).** Compared to zero-inflated and count-based approaches, MaAsLin 2's linear model formulation consistently controlled false discovery rate at the intended nominal level while maintaining moderate sensitivity. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

As a final evaluation, we assessed the impact of various normalization schemes on the associated statistical modeling, evaluating all combinations of normalizations appropriate for each applicable method (**Supplementary Fig. S1C, Methods**). Focusing on the best-performing linear models, we found that model-based normalization schemes such as relative log expression (RLE[33]) as well as data-driven normalization methods such as the trimmed mean of *M*-values (TMM[34]) and cumulative sum scaling (CSS[21]) led to good control of FDR, but they also led to a dramatic reduction in statistical power (**Supplementary Figs. S3, S5).** In contrast, TSS showed the best balance of performance among all tested normalization procedures, leading to more powerful detection of differentially abundant features. These results have potential implications for other analyses in addition to differential abundance testing, as normalization is usually the first critical step before any analysis of microbiome data, and an inappropriate normalization method may severely impact post-analysis inference. In summary, our synthetic evaluation indicates that TSS normalization, although simplistic in nature, may be superior to other normalization schemes especially in the context of feature-wise differential abundance testing (and more generally for multivariable association testing, as described later), in addition to community-level comparisons as previously described[35].

MaAsLin 2 facilitates multivariable association discovery in population-scale epidemiological studies

Moving beyond univariate comparisons, we next assessed MaAsLin 2's performance in multivariable association testing in comparison to other methods. Although widespread in microarray and gene expression literature, multivariable analysis methods have remained underdeveloped in microbial community studies. From an epidemiological point of view, coefficients from a covariate-adjusted regression model are arguably more interpretable than its individual, unadjusted counterparts. As a result, major conclusions from existing benchmarking studies geared towards univariate associations are not generalizable to this broader setting, where challenges such as zero-inflation and multiple testing are likely to be exacerbated, especially in relation to multiple rounds of independently conducted univariate analyses as commonly practiced.

To introduce multivariable associations into synthetically generated microbial feature profiles, we supplemented each "sample" with multiple covariates consisting of both binary and continuous metadata, either independent or correlated (**Supplementary Fig. S1A, Methods**). In each of these datasets, 10% randomly selected features were modified ("spiked") to be associated with randomly chosen 20% metadata features with a given magnitude (effect size). After spiking in, samples were rescaled to their original (simulated) sequencing depth. As before, we repeated this procedure for each unique combination of sample size (10, 20, 50, 100, 200), feature dimension (100, 200, 500), and effect size (1, 2, 5, and 10), each time summarizing performance over 100 simulation runs.

The results from this set of simulations revealed that MaAsLin 2's default linear model had the highest sensitivity among the methods that controlled the FDR at the target level, which also remained consistent at larger sample sizes and stronger effect sizes (**Fig. 3**). We also observed an improvement in performance when TSS normalization was employed (as compared to no normalization) but did not observe similar improvement for other normalization methods (**Supplementary Fig. S6)**. As before, zero-inflated and count models failed to control the FDR at the nominal level, in the sense that the actual FDR was always above the nominal threshold used for identifying significant features - a phenomenon that was surprisingly consistent regardless of the metadata covariance structure (**Supplementary Fig. S7**). Taken together, these findings further confirm that MaAsLin 2's default linear model is able to detect relevant associations across a broad range of metadata designs, facilitating population-level analyses of microbial communities.

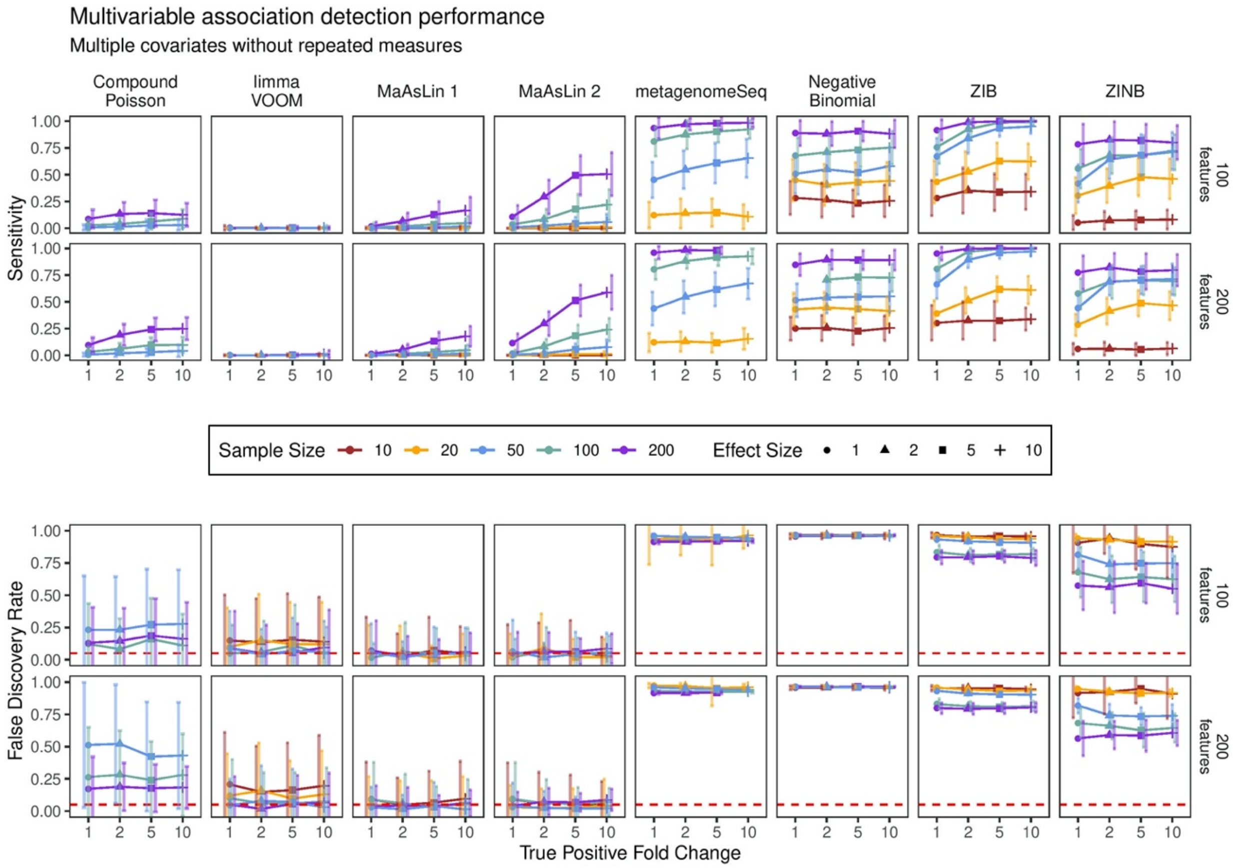


Figure 3: MaAsLin 2 facilitates multivariable association discovery in large-scale human epidemiological and other microbial community studies. Synthetic datasets containing five "metadata" with varying types of induced feature associations were analyzed using a variety of multivariable approaches (**Supplementary Fig. S1C)**. As measured by power (recall) and false discovery rate (FDR), MaAsLin 2's default linear model outperformed other methods in controlling FDR while maintaining power across true-positive fold-change values, regardless of the total number of features. As expected, MaAsLin 2 has better power for stronger effect sizes, eventually attaining the highest power among all FDR-controlling methods. Red line parallel to the x-axis is the nominal FDR. Values are averages over 100 iterations for each parameter combination. The x-axis (effect size) within each panel represents the linear effect size parameter; a higher effect size represents a stronger association. For visualization purposes, the best-performing methods from each class of models (as measured by average F1 score; **Methods**; full results in **Supplementary Figs. S6-S7**) are shown. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

MaAsLin 2 enables targeted microbiome hypothesis testing in the presence of repeated measures

To further validate MaAsLin 2 for longitudinal (or other repeated measures) microbiome data, we modified our simulation scheme to introduce subject-specific random effects - a key feature of modern microbiome population studies[36]. To this end, we tested MaAsLin 2 and related methods on two types of study designs. The first comprised univariate binary metadata designed to be challenging by the inclusion of non-independence of the data across time points. Second, we also simulated more realistic datasets using multiple independent covariates specific to longitudinal microbiome studies. In both these regimes, realistic data were generated using SparseDOSSA each with five time points[24], as in previous studies, but after introducing within-subject correlations and between-subject random effects drawn from a multivariate normal distribution (**Methods**). It is to be noted that the set of evaluable models is greatly reduced from the previous set of cross-sectional association tests, as methods not capable of assessing repeated measures were discarded.

Using these longitudinal synthetic "microbial communities," we compared the estimation and inference from MaAsLin 2 with those of the existing methods, which revealed that MaAsLin 2 had much lower false discovery rates than alternatives including ZIB (**Fig. 4, Supplementary Figs. S8-S11**), a method specifically designed for microbiome longitudinal data. Both ZIB and MaAsLin 2's linear mixed-effects models are capable of identifying covariate-associated features by jointly modeling all time points. However, the computational overhead of ZIB is significantly higher than that of MaAsLin 2, which is prominent even for small datasets (**Supplementary Fig. S12)**. Notably, although not nearly as severe as count-based and zero-inflated models, MaAsLin 2 had a slightly inflated FDR in the univariate repeated measures scenario (**Fig. 4A**) but not in the multivariable scenario (**Fig. 4B)**. Among other methods, GLM-based methods such as Negative Binomial and Compound Poisson performed similarly to their non-longitudinal counterparts for both normalized and non-normalized counts (**Supplementary Figs. S8-S9**). This remained consistent for both univariate continuous metadata (**Supplementary Fig. S10**) as well as multiple, correlated covariates (**Supplementary Fig. S11**). Overall, these results suggest that MaAsLin 2's linear mixed effect model consistently provides lower false discovery rates across metadata designs and can effectively aid in testing differential abundance and multivariable association of longitudinal microbial communities.

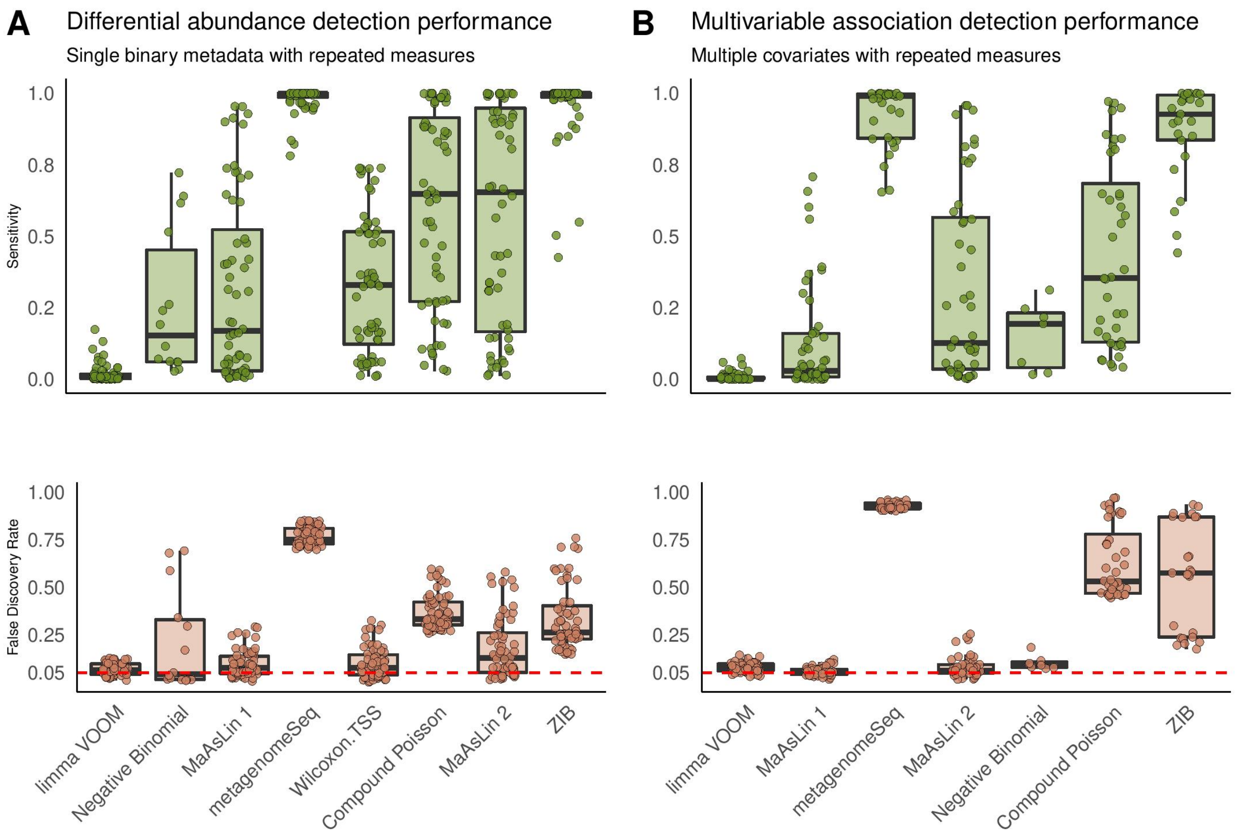


Figure 4: MaAsLin 2 enables targeted microbial feature testing in the presence of repeated measures. Results on simulated data comprising SparseDOSSA-derived compositions with five repeated measures per sample. The FDR is close to the target 0.05 level for MaAsLin 2's default linear model but not for zero-inflated and count models. As before, MaAsLin 2's linear model is consistently better powered than both negative binomial and limma VOOM at comparable FDR values, which remains consistent for both univariate continuous metadata (**A**) and multivariable mixed metadata designs (**B)** (a combination of continuous and binary covariates with five metadata features; **Methods,** full results in **Supplementary Figs. S8-S11**). The red line parallel to the x-axis is the given threshold for FDR in multiple testing. Within each panel, methods are sorted by increasing order of average F1 score across all associated simulation parameters in each setting.

Multi-omics associations from the Integrative Human Microbiome Project

We applied MaAsLin 2 to identify relevant microbial features associated with the inflammatory bowel diseases (IBD) using longitudinal multi-omics data from the Integrative Human Microbiome Project (iHMP or HMP2[36]). The HMP2 Inflammatory Bowel Disease Multi-omics (IBDMDB) dataset included 132 individuals recruited in five US medical centers with Crohn's disease (CD), ulcerative colitis (UC), and non-IBD controls, followed longitudinally for one year with up to 24 time points each (**Methods**).

Integrated multi-omics profiling of the resulting 1,785 stool samples generated a variety of data types including metagenome-based taxonomic profiles as well as metagenomic and metatranscriptomic functional profiles, producing one of the largest publicly available microbial multi-omics datasets. Metagenomes and metatranscriptomes were functionally profiled using HUMAnN 2[37] to quantify MetaCyc pathways[38], and taxonomic profiles from metagenomes were determined using MetaPhlAn 2[39] (**Methods**). For each of these data modalities (i.e. taxonomic profiles and DNA/RNA pathways), independent filtering was performed before downstream testing to reduce the effect of zero-inflation on the subsequent inference. In particular, features for which the variance across all samples was very low (below half the median of all feature-wise variances) or with >90% zeros were removed[36]. To further remove the effect of redundancy in pathway abundances (explainable by at most a single taxon), only features (both DNA and RNA) with low correlation with individual microbial abundances (Spearman correlation coefficient <0.5) were retained.

We first used the IBDMDB to perform an additional semi-synthetic evaluation of association methods' performance in "real" data, specifically when attempting to associate randomized, null microbial taxonomic profiles to covariates. To this end, we permuted all samples 1,000 times to construct shuffled "negative control" datasets, each time assessing the number of significant associations (unadjusted p <0.05) for each applicable method. These were averaged across iterations to derive the expected number of null associations per method (which should remain near-zero under usual circumstances). In particular, we fit (i) a baseline model as a function of IBD diagnosis (a categorical variable with non-IBD as the reference group) while adjusting for enrollment age (as a continuous covariate) and antibiotic use (as a binary covariate), and (ii) a mixed effects model (with subject as random effects) with IBD dysbiosis state as an additional time-varying covariate in addition to the time-invariant covariates considered in the baseline model. Consistent with prior simulations, we found that several methods produced inflated empirical type I error rates (**Supplementary Fig. S13)**. This conclusion remained unchanged across varying significance thresholds, and as a result, we did not further apply these methods to the non-permuted data. Relevantly and importantly, linear models did not suffer from this problem, providing additional support for MaAsLin 2's robustness to false positive findings.

To dissect dysbiotic changes in IBD at greater resolution, we applied MaAsLin 2 to each individual microbial feature type (i.e. species and DNA/RNA pathways) to test association with IBD phenotype while controlling for IBD dysbiosis state, diagnosis, age, and antibiotic use (**Fig. 5**; **Methods**). Nominal p-values for UC- and CD-specific effects were subjected to multiple hypothesis testing correction using the Benjamini-Hochberg method[40] with an FDR threshold of 0.25. MaAsLin 2 identified a comparable number of significant associations with those initially reported by the IBDMDB[36]. Among microbial species, MaAsLin 2's default linear model identified 206 significant associations, among which 150 (72.8%) overlapped with the original study (**Supplementary Fig. S14**). MaAsLin 2 also reported many significant associations that were not discovered in the original study (**Supplementary Dataset S1)**. For instance, we observed a significant increase in *Bacteroides ovatus* in both UC and CD dysbiotic patients that was not previously captured, as well as detecting (with MaAsLin 2's increased power) specific depleted *Roseburia* species (*R. inulinivorans* and *R. hominis*) not captured by the previous analysis. Notably, top hits from both MaAsLin 2 and the original study yielded nearly identical rankings across data types, which broadly manifested as a characteristic increase in facultative anaerobes at the expense of obligate anaerobes, in agreement with the previously observed depletion of butyrate producers such as *Faecalibacterium prausnitzii* in IBD (**Fig. 5A**).

As an additional validation, we next re-analyzed the HMP2 taxonomic and functional profiles using a zero-adjusted model (implemented in MaAsLin 2 as the compound Poisson). While this maintained type-I error control in our shuffled data validation (as did the default linear model, **Supplementary Fig. S13),** it was generally less desirable due to FDR inflation in simulations

(**Figs. 2-4**). In terms of the number of differentially abundant features detected, both the default linear model and the compound Poisson model performed similarly, with a significant overlap between the top hits identified by each method (**Supplementary Fig. S15**). Among other methods, ZIB and limma VOOM also maintained good Type I error control in these experiments, although again both underperformed along other axes in our simulation studies. These results further strengthen the finding that a combination of controlled parametric simulations and 'negative control' experiments based on data shuffling are useful together in identifying methods for real-world applications, as the lack of either can lead to misleading (and irreproducible) conclusions across independent evaluations[7]. This also highlights the flexibility of MaAsLin 2's multi-analysis framework, wherein researchers are well-served with multiple (i) normalization schemes, (ii) statistical models, (iii) multiplicity adjustments, (iv) multiple fixed and random effects specifications, and (v) in-built visualization and pre-processing options, facilitating seamless application of methods across diverse experimental designs under a single estimation umbrella.

Finally, in addition to taxonomic associations, MaAsLin 2 also detected 492 and 58 significant functional associations for metagenomic (DNA) and metatranscriptomic (RNA) pathways, respectively (**Supplementary Datasets S2-S3),** among which 358 (72.7%) and 39 (67.2%) overlapped with the original study (**Supplementary Fig. S14**). While the original analysis of these data included only community-wide functional profiles, we extended this by considering metagenomic and metatranscriptomic functional profiles at both whole-community and species-stratified levels in order to quantify overall dysbiotic functions while simultaneously assigning them to specific taxonomic contributors. In particular, this considers a per-feature DNA covariate model, in which per-feature normalized transcript abundance is treated as a dependent variable, regressed on per-feature normalized DNA abundances along with other regressors in the model (**Methods)**. Surprisingly, bioinformatics and statistics for metatranscriptomics are not yet standardized, and our results indicate that subtle model variations can produce substantially different results, due to the interactions between two compositions (DNA and RNA relative abundances, **Supplementary Dataset S4).** This novel modeling strategy thus led to the discovery of several novel transcript associations relative to the original study.

In a majority of these pathways, functional perturbations were driven by shifts in their characteristic contributing taxa (**Fig. 5B**). For example, the most significant DNA pathways enriched in CD patients were characteristic of facultative anaerobes such as *Escherichia coli*, which are broadly more abundant during inflammation. These included pathways such as synthesis of the enterobactin siderophore, lipid A, and sulfate reduction. A second set of enriched pathways was depleted due to the loss of microbes such as *F. prausnitzii*, a particularly prevalent organism that, when abundant, tended to contribute the majority of all enriched pathways it encodes in this cohort (e.g. synthesis of short-chain fatty acids and various amino acids).

With the increased sensitivity of this analysis for species-stratified pathways, the overwhelming majority of significant metagenomic differences were attributable solely to the most differential individual organisms, as expected (**Supplementary Datasets 5-6**). Essentially every pathway reliably detectable in *E. coli* was enriched during CD, UC, or both, and most *F. prausnitzii* pathways depleted, along with many pathways from other gut microbes common in “health” (*Bacteroides vulgatus*, *B. ovatus*, *B. xylanisolvens*, *B. caccae*, *Parabacteroides* spp., *Eubacterium rectale*, several *Roseburia* spp., and others). Interestingly, since both more potentially causal “driver” pathways, along with all other “passenger” pathways encoded by an affected microbe, are detected by this more sensitive stratified analysis, it can be in many ways more difficult to interpret than the non-stratified, community-wide, cross-taxon metagenomic responses to broad ecological conditions such as immune activity, gastrointestinal bleeding, or oxygen availability.

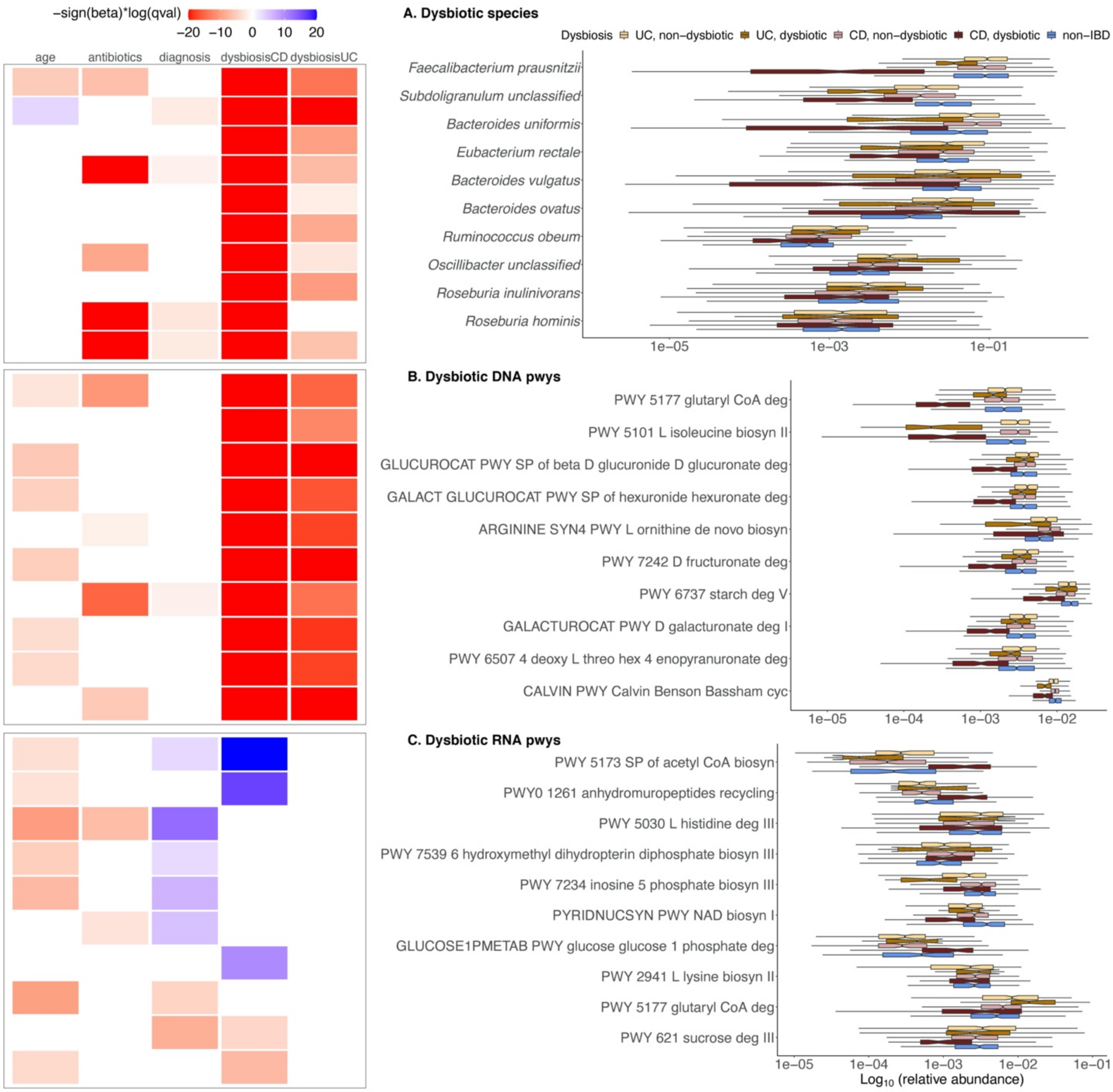


Figure 5: Multi-omics associations from the Integrative Human Microbiome Project. A) Top 10 significant associations (FDR < 0.25) detected by MaAsLin 2's default linear model (significance and coefficients in **Supplementary Datasets S1-S3**.). All detected associations are adjusted for subjects and sites as random effects and for other fixed-effects metadata including the subject's age, diagnosis status (CD, UC, or non-IBD), disease activity (defined as median Bray-Curtis dissimilarity from a reference set of non-IBD samples), and antibiotic usage. Representative significant associations with dysbiosis state from each 'omics profile are shown: species (**B**), metagenomic (DNA) pathways (**C**), and metatranscriptomic (RNA) pathways (**D**). Values are log-transformed relative abundances with half the minimum relative abundance as pseudo count; full results in **Supplementary Datasets S1-S3.**

Conversely, differentially abundant microbe- and pathway-specific transcript levels highlighted a much more specific and dramatic shift toward oxidative metabolism, away from anaerobic fermentation, and towards Gram-negative (often *E. coli*) growth during inflammation (**Fig. 5C**)[41]. Many of these processes were either more extreme during (e.g. gluconeogenesis) or unique to (e.g. glutathione utilization) active CD, as compared to UC. CD and UC responses were opposed in a small minority of cases (e.g. glutaryl-CoA degradation). When stratified among contributing taxa, these differences were almost universally attributable to a few key species, particularly an increase in *E. coli* activity during inflammation and decreases of *F. prausnitzii* transcript representation. Condition-specific transcriptional changes were occasionally contributed (or not) by “passenger” *Bacteroides* spp. (*B. fragilis*, *B. xylanisolvens*, *B. dorei*) instead. Note that these differences include pathways more likely to be “causal” in some sense, as significant transcriptional changes were generally a subset of those detected due to whole-taxon shifts in DNA content (including housekeeping pathways such as general amino acid or nucleotide biosynthesis). These findings further support the importance of disease-specific transcriptional microbial signatures in the inflamed gut relative to metagenomic profiles of functional potential, suggesting that a potential loss of species exhibiting altered expression profiles in disease may have more far-reaching consequences than suggested by their genomic abundances alone.

Discussion

A longstanding goal of microbial community studies, be they for human epidemiology or environmental microbiomes, is to identify microbial features associated with phenotypes, exposures, health outcomes, and other important covariates in large, complex experimental designs. This parallels other methods for high-throughput molecular biology, but microbial community multi-omics must account for properties such as variable sequencing depth, zero-inflation, overdispersion, mean-variance dependency, measurement error, and the importance of repeated measures and multiple covariates. To this end, we have developed and validated a highly flexible, integrated framework utilizing an optimized combination of novel and well-established methodology, MaAsLin 2. This accommodates a wide variety of modern study designs ranging from within-subject, longitudinal to between-subject, cross-sectional, diverse covariates, and a range of quality control and statistical analysis modules to identify statistically significant as well as biologically relevant associations in a reproducible framework. The embedding of these strategies in the paradigm of generalized linear and mixed models enables the treatment of both simple and quite complex designs in a unified setting, improving the power of microbial association testing while controlling false discoveries. To validate this framework, we have extensively evaluated its performance alongside a set of plausible methods for differential abundance analysis in a wide range of scenarios spanning simple univariate to complex multivariable with varying scopes and effect sizes of microbiome associations. Finally, we applied MaAsLin 2 to identify disease-associated features by leveraging the HMP2's multi-omics profile of the IBD microbiome, confirming known associations and suggesting novel ones for future validation.

A unique aspect of our synthetic evaluation of microbial community feature-wise association methods while developing MaAsLin 2 is their comprehensive assessment in the presence of multiple covariates and repeated measures, an increasingly common characteristic of modern study designs. To identify covariate-associated microbial features from longitudinal, non-independent measurements, it is necessary to jointly model data from all time points and appropriately account for the within-subject correlations while allowing for multiple covariates. This is particularly critical in the human microbiome, where baseline between-subject differences can be far greater than those within-subjects over time, or of the effects of phenotypes of interest. To the best of our knowledge, the synthetic evaluation presented here is the first to consider such aspects of large-scale microbiome epidemiology in statistical benchmarking. This enabled us to investigate key aspects of published methods that would be difficult to generalize from univariate comparisons alone[7-9]. Note that the resulting conclusion is largely independent of the association models being evaluated, as the synthetic data were generated from an additional, completely external model (i.e. the zero-inflated log-normal, **Methods**), which is fundamentally different from any of the evaluated parametric methods. Our simulation results thus complement the findings of previous studies in several important aspects. Consistent with previous reports, nearly all zero-inflated models suffer from poor performance (i.e. inflated false positives and higher computation costs), here in both univariate and multivariable scenarios with or without repeated measures. This calls for methodological advancements in statistical modeling of zero-inflated data, as existing theory seems to differ very surprisingly from practice when implemented by established optimization algorithms and applied to noisy data.

One noteworthy finding of our evaluation is that a random effect implementation of the same underlying statistical model can lead to different substantive conclusions than its fixed-effects counterpart. This was particularly evident for the negative binomial case, where a substantially better control of FDR (albeit inflated) was observed for the random effect analog. Interestingly, the negative binomial model (with or without zero inflation) is in many ways considered the most “appropriate” model for count-based microbial community profiles, but we observed extremely inconsistent behavior for negative binomial and ZINB implementations during our evaluation, as also observed in previous findings[42]. In particular, our negative binomial evaluation used the *glm.nb()* function from the *MASS* R package[43] for fixed-effects and the *glmer()* function from the R package *lme4*[44] for random effects, whereas the ZINB evaluation used the *zeroinfl()* function from the R package *pscl*[45]*.* This additionally highlights the potential reproducibility concerns induced by differences in algorithms, implementations, and computational environments even for the same underlying model, suggesting that great caution should be taken when interpreting multiple implementations of the same statistical model for challenging microbial community settings in the absence of an experimentally validated gold standard.

In agreement with previous studies, we confirmed that most RNA-seq differential expression analysis tools tend to provide suboptimal performance when applied unmodified to zero-inflated microbial community profiles. In particular, count-based models, due to their strong parametric assumptions on the distributions or parametric specifications of the mean-variance dependency, tend to have inflated FDR when the assumptions are violated. In sharp contrast to previous claims, however, compositionality-corrected methods such as ANCOM[14,22] as well as specialized normalization and transformation methods such as CLR[46] did not improve performance over non-compositional approaches[8,47], consistent with recent findings that compositional methods may not always outperform non-compositional methods[32]. Importantly, these conclusions hold regardless of the nature of the modeling paradigm (i.e. univariate vs. multivariable), thus providing a generalizable benchmark for future evaluation studies of applied microbiome association methods. Though we primarily focused on data generated in microbial community surveys, many of our conclusions are extendible to similar zero-inflated count data arising in other applications such as single-cell RNA-seq. Taken together, these simulation results revealed that further investigation into the causes of the failure of FDR correction and development of specialized false positive-controlling methods are important upcoming challenges in microbiome statistical research.

Limitations of the current MaAsLin 2 method include, first, its restriction to associating one feature at a time. While this strategy has the advantage of being straightforward to interpret, implement, and parallelize, it sacrifices inferential accuracy by ignoring any correlation structure among features. This can certainly exist due both to compositionality and to biology and will differ e.g. between taxonomic features (related by phylogeny) vs. functional ones (such as pathways). A potential extension would be to adopt an additional multivariate framework that allows modeling simultaneously rather than sequentially, thus improving power by borrowing strength across non-independent features. Second, as revealed by our synthetic evaluation, not surprisingly, linear models remain underpowered in detecting weak effects among microbial communities, especially when accompanied by a small sample size. This is in some ways a necessary consequence of the restrictions of current microbiome measurement technologies, and it emphasizes the importance of an informed power analysis before study planning to ensure an optimal sample size with adequate detection power. Finally, and relatedly, it is not straightforward to incorporate any type of graph structure knowledge such as phylogeny or pathway-based functional roles into the per-feature linear model framework. Bayesian linear models can potentially improve on this by including such information through a suitable prior distribution.

Several aspects of microbiome epidemiology remain to be investigated both biologically and computationally, in addition to the challenges addressed here. For example, while it is possible to obtain strain-level resolution from metagenomic sequencing data, strain variants are generally unique to particular individuals rather than broadly carried by many people, presenting unique challenges for strain-level multi-omics. From a computational point of view, this calls for further refinements to MaAsLin 2's methodology when applied to strain-resolved community profiles. In addition, the modeling framework adopted here can only inform undirected associations, and hence cannot be used to infer causation. Advanced methods from other molecular epidemiology fields such as causal modeling and mediation analysis methods can help overcome these issues[48]. Another opportunity for future extension of our method is the integration of established missing data imputation methods across features and metadata, a common pitfall in many molecular epidemiology studies[36]. Combined, such extensions will lead to further improvement in downstream inference, allowing researchers to investigate a range of hypotheses related to differential abundance and multivariable association.

As currently implemented, MaAsLin 2 is designed to be applicable to most human and environmental microbiome study designs, including cross-sectional and longitudinal. Clearly, these can also be extended to additional designs, such as nested case-control and case-cohorts. It is to be noted that MaAsLin 2's capability extends well beyond association analysis. For instance, MaAsLin 2's multi-analysis framework has been used in the context of meta-analysis[49], and the extracted residuals and random effects from a MaAsLin 2 fit can be used for further downstream analysis (e.g. as has been done in the original HMP2 study for cross-measurement correlation analysis[36]). By adhering to a flexible mixed effects framework, MaAsLin 2 is able to analyze multiple groups and time points jointly with other associated covariates, which allows formulation of both fixed effects (for cross-sectional associations) and random effects (for within-subject correlations) in a single unified framework. This is particularly appropriate for non-longitudinal studies (those with a small number of repeated measures e.g. multiple tissues or families), or from sparse and irregular longitudinal data from many subjects (e.g. with unequal number of repeated measurements per subject, as commonly encountered in population-scale epidemiology). This aspect could also be extended in the future, based on the increasing availability of dense time-series profiles appropriate for non-linear trajectory-based methods from Bayesian nonparametrics, such as Gaussian processes, particularly in the presence of multiple covariates[5,50]. Finally, methods need to be developed to accommodate the increasing availability of microbiome-host interactomics and electronic health records in population-scale microbiome-wide epidemiology[6], moving beyond observational discovery toward translational applications of the human microbiome. In summary, the methodology presented here provides a starting point for more efficient identification of microbial associations from large microbial community studies, offering practitioners a wide set of analysis strategies with state-of-the-art inferential power for the human microbiome and other complex microbial environments.

Methods

Data for differential feature model evaluations

Synthetic null community abundances

Realistic null community data were generated using the SparseDOSSA[51] (Sparse Data Observations for the Simulation of Synthetic Abundances) hierarchical model (http://huttenhower.sph.harvard.edu/sparsedossa). SparseDOSSA is a newly developed simulator designed to model the fundamental characteristics of real microbial communities (e.g. zero-inflation, compositionality, etc.) and to simulate new, realistic metagenomic count data with known feature-feature and feature-metadata correlations and provide a gold standard to enable benchmarking of statistical metagenomics methods, superseding previous efforts by including multiple covariates and longitudinal designs.

Briefly, SparseDOSSA's Bayesian model captures microbial features (taxon, gene, or pathway abundances) as truncated, zero-inflated log-normal distributions, the parameters of which are hierarchically derived from a parent log-normal distribution. SparseDOSSA estimates feature-specific parameters by fitting to a real-world template dataset, and generates synthetic features from zero-inflated, truncated log-normal distributions based on both fitted and user-defined parameters on a per feature basis (**Supplementary Fig. S1A**). All feature-specific parameters, namely the log-mean, zero-inflation proportion, truncation point, and log-variance are empirically determined to resemble the template dataset's properties. After sampling, the samples are rounded to the nearest integer to mimic count data. A combined dataset of the RISK[52], PRISM[15], pouchitis[16], and NLIBD[53] gut microbiomes, totaling several thousand samples, was used as empirical microbiome template data for the simulations reported in this study. To mimic realistic variation in library size, sequencing depth was generated from a lognormal distribution with average sequencing depth 50,000, resulting in approximately 30-fold to 100-fold variation in sequencing depth.

Synthetic metadata generation

Simulated metadata matrices in simple univariate cases (UVA, UVB) were generated with continuous values from a standard normal distribution. For the univariate binary case (UVB), we additionally dichotomized the continuous variable by coding samples in the bottom and top half of the distribution as 0 and 1, respectively. For multivariate cases (MVA, MVB), we repeated the above discretization for multiple metadata by first generating from a multivariate normal distribution, and concurrently binarizing half of the metadata features at random. We considered two frequently encountered correlation structures for the multivariate cases: independent and AR (1) with coefficient 0.5, which correspond to MVA and MVB, respectively. Additionally, we considered repeated measures by incorporating random effects in these cross-sectional design matrices. To that end, we generated a simple blocking variable that is normally distributed (with mean 0 and variance 1) across subjects but invariant within subjects, representing a single random effect factor such as subject or time point (block size determined by the simulation parameters as reported in **Supplementary Fig. S1A**). Subsequently, we added this as an additional covariate to the fixed-effects metadata to impose correlations within the blocks, mimicking longitudinal studies. For multivariable cases (MVA, MVB), the number of covariates is fixed to 5. Similarly, for the repeated measures settings, T = 5 time points per subject is considered.

Multivariable spike-ins of synthetic feature-metadata associations

To introduce associations between features and metadata, we used SparseDOSSA's default additive spike-in procedure. Following Weiss et al.[8], we implement the spike-ins in a balanced way across all metadata to avoid compositional bias. Briefly, SparseDOSSA standardizes both (microbial) features and metadata and randomly chooses (microbial) null features and metadata without replacement. The standardization procedure ensures that the spiked-in features are not dominated by the values of the target metadata but rather distributed similarly to the real data. Next, the standardized non-zero abundances of the selected features are modified by adding a linear combination of all spiked-in standardized metadata, in which a real-valued effect size parameter (**Supplementary Fig. S1A**) governs the strength of association for each associated feature-metadata pair. To create differentially abundant features, a randomly sampled fraction of 10% of the features are spiked-in to be associated with the metadata. In the multivariable case, 20% of the metadata are randomly selected to be associated with the 10% 'differentially abundant' features.

Multivariable association test evaluation

Preprocessing, normalizations, and transformations

We considered several commonly used normalization procedures including Total Sum Scaling (TSS), Trimmed Mean of *M*-values (TMM[34]), Relative Log Expression (RLE[33]), and Cumulative Sum Scaling (CSS[21]) (**Supplementary Fig. S1C**). For TSS normalization, raw counts were converted into relative abundances by scaling each sample by the total sum (across features). For the remainder, we used the default settings of the edgeR[26], DESeq2[25], and metagenomeSeq[21] R packages, respectively.

In addition to the above normalization procedures, several parametric transformations were also considered. When appropriate, these variance-stabilizing transformations aim at improving parametric estimation models in the presence of violated data assumptions (such as normality and homoscedasticity). These include logit and arcsine square root (AST) for TSS-normalized or proportional relative abundance data, and log transformation (**Supplementary Fig. S1C**). For both log and logit transformations, undefined values were replaced with zeroes (equivalent to adding a small pseudo count of 1 to the zero observations before applying the log transformation). Among other normalization/transformation methods, a 'Default/None' category was also considered which represents either (i) default normalization/transformation employed by the associated software or (ii) no normalization/transformation, or (iii) library size normalization in the form of a GLM offset modeling. Prior to applying any normalization and transformation, a basic filtering was performed to prune features absent in >90% of samples. As in previous benchmarking[7,8,10], correction for multiple testing was performed using the Benjamini-Hochberg[40] FDR threshold of 0.05.

Statistical methods

We selected several commonly used methods for differential abundance and multivariable association testing along with a set of experimental methods to apply on the synthetic datasets, using a combination of statistical model and normalization/transformation schemes for each method as appropriate (**Supplementary Fig. S1C**). All tests were conducted using the statistical software R and parallelized using custom bash scripts in a high-performance computing environment (full source code available at: https://github.com/biobakery/maaslin2_benchmark). The selected statistical models (abbreviations in parentheses) are as follows:

- ANCOM: following Weiss et al.[8], we used the default implementation of ANCOM[14] using the *ANCOM()* function call with default settings. Unlike other methods, ANCOM does not report p-values but instead returns logical indicators of whether a feature is differentially abundant based on a test statistic W. It is to be noted that in the presence of multiple covariates, ANCOM does not return statistically significant feature-metadata pairs with respect to every covariate in the model, making it infeasible for our multivariable setting. Also, we did not test the ANCOM method in repeated measures settings as it was too slow and unstable for assessment, as noted elsewhere[32].
- metagenomeSeq: for the fixed effects, counts were first CSS-normalized with the default quantile supplied by the *cumNormStat()* function and the (log-transformed) CSS-normalized counts were subjected to final testing using *fitZig*()[21]. For random effects, *useMixedModel was* set to TRUE in the *fitZig*() function call.
- metagenomeSeq2: same as metagenomeSeq[21], except that the final testing was done using the *fitFeatureModel*() function.
- DESeq2: for fixed effects, following Thorsen et al.[9], geometric means were first calculated manually from the raw counts and supplied to the *estimateSizeFactors()* function before

calling the *DESeq()* function for final testing[25]. Random effects modeling compatible with our setting is currently not supported by the DESeq2 software[54].

- edgeR: for fixed effects, following Thorsen et al.[9], normalization factors were calculated with TMM, which was followed by common and tagwise dispersion estimation steps, before invoking the standard test with the *exactTest()* function[26]*.* Random effects modeling compatible with our setting is currently not supported by the edgeR software[54].
- limma: the default functionality of *lmFit()* was applied to the feature counts[29]. Repeated measures were handled using the *duplicateCorrelation()* function before calling *lmFit(),* in combination with appropriate normalization/transformation (**Supplementary Fig. S1C**).
- limma VOOM: same as limma, except that features were subjected to a voom transformation before applying limma[27,28].
- limma2: same as limma, except that library size or scale factor is included as an additional covariate, in combination with appropriate normalization/transformation (**Supplementary Fig. S1C**).
- Wilcoxon: the built-in R function *wilcox.test()* using default parameters was applied to the features in combination with appropriate normalization/transformation (**Supplementary Fig. S1C**).
- Spearman: the built-in R function cor*.test()* was applied to the features in combination with appropriate normalization/transformation (**Supplementary Fig. S1C**).
- Linear model (LM): the built-in R function *lm()* with default settings was used in combination with appropriate normalization/transformation (**Supplementary Fig. S1C**).
- Linear model (LM2): same as LM, except that library size or scale factor is included as an additional covariate in the model, in combination with appropriate normalization/transformation (**Supplementary Fig. S1C**).

- Negative binomial (negbin): we used the *glm.nb()* function from the *MASS* package[43] and the *glmer.nb()* function from the *lme4* package[44] for fixed and random effects respectively. In both cases, we used the logarithm of library size (for no normalization) or scaling factor (for other normalization schemes such as CSS, RLE, and TMM) as offset.
- Zero-inflated Negative Binomial (ZINB): for fixed effects, we used the *zeroinfl()* function from the *pscl* package[45] with the logarithm of library size (for no normalization) or scaling factor (for other normalization schemes such as CSS, RLE, and TMM) as offset. In the absence of a robust random effect implementation of the same, the ZINB method was not considered in the repeated measures settings.
- Zero-inflated Beta (ZIB): following Peng et al.[23], we used the *gamlss()* function from the R package *gamlss*[55] for fixed effects and the *ZIBR()* function from the *ZIBR* R package for random effects[24]. In both cases, the features are TSS-normalized before statistical testing.
- Compound Poisson (CPLM): we used the *cpglm()* function from the *cplm* package[56] and the *glmmPQL()* function from the *MASS* package[56] for fixed and random effects respectively. In both cases, we used the logarithm of library size (for no normalization) or scaling factor (for other normalization schemes such as CSS, RLE, and TMM) as offset. No offset was used when combined with the TSS-normalized relative counts.
- MaAsLin 1: we used the default TSS-normalized, arcsine square root-transformed linear model without gradient boosting[15,16].
- MaAsLin 2: we used the default TSS-normalized, log-transformed linear model with half the minimum relative abundance as pseudo count.

Naming convention

The nomenclature for the model/normalization/transformation combinations for each method is described in the following set of rules:

1. For published methods with default parameters, there is no additional identifier following the name of the method, indicating default or no normalization/transformation. These include ANCOM, metagenomeSeq, metagenomeSeq2, limma, limma2, limma VOOM, DESeq2, edgeR, and ZIB.
2. Similarly, for experimental methods with custom normalization/transformation schemes, no additional identifier simply indicates either no normalization (for non-GLM methods such as LM) or library size normalization (for specific GLMs such as Negative Binomial, Compound Poisson, and ZINB).
3. Finally, for methods with additional identifiers, method name is always accompanied by a normalization scheme (after the first dot) which is followed by a transformation (after the second dot) except in cases where either no normalization or no transformation is applied. As an example, limma.CSS.LOG denotes a default limma model followed by CSS normalization and log transformation. Similarly, LM.CLR denotes a vanilla linear model followed by a CLR transformation and no normalization, whereas, ZINB.TMM denotes a zero-inflated negative binomial model with TMM normalization and no transformation, and so on and so forth.

Performance evaluation

Several performance metrics were considered for evaluation, all derived from some combination of the elements from the confusion matrix: false positives (FPs), true positives (TPs), true negatives (TNs), and false negatives (FNs). These include measures related to (i) statistical power, (ii) false discovery, and (iii) software implementation and scope, all as averages over 100 simulation runs (**Supplementary Fig. S1B**). Several measures were considered for statistical power - Sensitivity, Area Under the Curve (AUC), and scaled partial AUC (spAUC). The AUC was calculated as the area under the ROC curve, obtained by plotting the sensitivity versus 1-specificity for the varying p-value threshold. spAUC was calculated as the partial area over the high specificity range (0, 0.20), rescaled to mimic the interpretation of AUC (i.e. 0.5 for a random guess and 1 for a perfect classifier using p-values to discriminate between spiked and non-spiked features). The R package *ROCR*[57] was used to calculate both these AUC measures. We also considered Matthew's correlation coefficient as well as F1 scores as alternate accuracy measures of performance.

Among false discovery metrics, maximum and average of several commonly used metrics including False Discovery Rate (FDR) and False Positive Rate (FPR) were considered. When no features were declared significant (i.e. TP = FP = 0), the false discovery rate (FDR) was set to 0. Notably, Weiss et al.[8] misreported false positive rate as FDR, as evident from the supplemental R code of that paper (Additional files 9 and 10 of Weiss et al[8]). In order to avoid any ambiguity, we provide the analytical expressions of the above-mentioned measures (except AUC and spAUC) as follows:

$$\text{FDR}\ (1-\ \text{Precision})\ = \frac{\text{FP}}{\text{FP}+\text{TP}}$$

$$\text{FPR}\ (1-\ \text{Specificity}) = \frac{\text{FP}}{\text{FP}+\text{TN}}$$

$$\text{Sensitivity (Power or Recall)}\ = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F1 score}\ = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$$

$$\text{Matthew's correlation coefficient (MCC)}\ = \frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{\sqrt{(\text{TP}+\text{FP})*(\text{TP}+\text{FN})*(\text{TN}+\text{FP})*(\text{TN}+\text{FN})}}$$

Following Hawinkel et al.[7], an alternative measure based on the p-value distribution under the null, 'Departure from Uniformity', was also considered. Briefly, to quantify the departures from uniformity into liberal (or conservative) direction, twice the mean distance between the diagonal line and the points in the QQ plot below (or above) the diagonal was computed. We called these measures ‘Liberal Area’ and ‘Conservative Area’, respectively. Both calculated areas are averages over all features, and they both range from 0 to 1. A combined metric called ‘Total Area’ that defines departure in either direction (defined as Total Area = Liberal Area + Conservative Area) was also computed.

Finally, we calculated computational time and convergence aspects of different methods based on their available implementation. Following Soneson and Robinson[58], we record the actual time required to run each method using a single core and normalize all times for a given data set instance so that the maximal value across all methods is 1 (as reported in **Fig. 1C**). Thus, a 'relative' computational time of 1 for a given method and a given data set instance means that this method was the slowest one for that particular instance, and a value of, for example, 0.1 means that the time requirement was 10% of that for the slowest method. Similarly, we estimated the ‘relative’ convergence failure rates for each method, as before, with the worst method as a reference.

Analysis of the iHMP (HMP2) IBDMDB multi-omics dataset

Study design, data, and quality control

Data were obtained from the Integrative Human Microbiome Project (HMP2 or iHMP), which is described in detail in Lloyd-Price et al.[36] and available through the Inflammatory Bowel Disease Multi-omics Database (IBDMDB, http://ibdmdb.org). Briefly, subjects included in this cohort were recruited from five academic medical centers across the US: three pediatric sub-cohorts including Cincinnati Children’s Hospital, Massachusetts General Hospital (MGH) Pediatrics, and Emory University Hospital, and two adult sub-cohorts including MGH and Cedars-Sinai Medical Center. 132 subjects were followed for one year each to generate integrated longitudinal molecular profiles of host and microbial activity during disease (up to 24 time points each; in total 2,965 stool, biopsy, and blood specimens). Self-collected stool samples were transported in ethanol fixative before storage at -80 C until DNA extraction.

Multiple measurement types were generated from many individual stool specimens, including 305 samples that contain all stool-derived measurements and 791 metagenome-metatranscriptome pairs. Metagenomic data generation and processing were performed at the Broad Institute. After standard sequence- and sample-level quality control as described in Lloyd-Price et al.[36], species-level taxonomic abundances were inferred for all samples using MetaPhlAn 2[39] and functional profiling was performed by using HUMAnN 2[37]. The resulting data types including metagenome-based taxonomic abundances and pathway abundance profiles for both metagenomics and metatranscriptomics (summarized as structured pathways from MetaCyc[59]) were used as inputs for MaAsLin 2 analysis.

Significance testing with shuffled data

In order to quantify whether MaAsLin 2 and other multivariable association methods identified more significant associations than expected by chance (i.e. when all the shared signal between features and metadata are broken), we repeatedly shuffled the metadata sample labels, applied multivariable association methods on the randomized data to link features to metadata, and compared the number of statistically significant associations obtained with these randomized data to the number of statistically significant associations obtained with the original data based on the unadjusted p-values. For a comprehensive comparison of both count and noncount models in this experiment, prior to data shuffling, we multiplied the species-level taxonomic abundances by 5% of the filtered read counts as a “proxy” for the underlying raw sequencing count data. The procedure was repeated 1,000 times to estimate the null distribution of the detection performance in both baseline and longitudinal models (with the exception of Compound Poisson mixed effects model which was repeated 100 times to save computation time). While the baseline model included all time-invariant covariates (age, antibiotic use, IBD diagnosis), the longitudinal model also included subjects as random effects with an additional time-variant fixed effect i.e. IBD dysbiosis state, as stated below.

Statistical analysis of species, DNA pathways, and RNA pathways

For both species and DNA pathways (whole-community and species-stratified), we regressed the log-transformed relative abundances (with half the minimum relative abundance as pseudo count, the default in MaAsLin 2) using the following per-feature linear mixed-effects model:

feature ~ (intercept) + diagnosis + diagnosis/dysbiosis + antibiotic use + consent age + (1 | recruitment site) + (1 | subject).

Additionally, we modeled the log-transformed relative abundances of the whole-community and species-stratified RNA pathways (with half the minimum relative abundance per feature as pseudo count) using the similar linear mixed-effects model as before, while additionally adjusting the corresponding DNA pathways abundance as a continuous covariate to filter out the influence from gene copies:

RNA feature ~ (intercept) + diagnosis + diagnosis/dysbiosis + antibiotic use + consent age + DNA feature + (1 | recruitment site) + (1 | subject)

That is, in each per-feature multivariable model, recruitment sites and subjects were included as random effects to account for the correlations in the repeated measures (denoted by (1 | recruitment site) and (1 | subject) respectively) and the transformed abundances of each feature was modeled as a function of diagnosis (a categorical variable with non-IBD as the reference group) and dysbiosis state as a nested binary variable (with non-dysbiotic as reference) within each IBD phenotype (UC, CD, and non-IBD), while adjusting for consent age as a continuous covariate, and antibiotics as a binary covariate. Nominal p-values were adjusted for multiple hypothesis testing with a target false discovery rate of 0.25 with this FDR chosen to match the original study.

Data Availability

The iHMP dataset is publicly available at the IBDMDB website (https://ibdmdb.org) and the HMP DACC web portal (https://www.hmpdacc.org/ihmp/). The processed HMP2 datasets analyzed in this manuscript are also available as **Supplementary Datasets S1-S6**.

Implementation and Software Availability

The implementation of MaAsLin 2 is publicly available with source code, documentation, tutorial data, and as an R/Bioconductor package at http://huttenhower.sph.harvard.edu/maaslin2. The software packages used in this work are free and open source, including bioBakery[60] methods available via http://huttenhower.sph.harvard.edu/biobakery as source code, cloud-compatible images, and installable packages. Analysis scripts using these packages to generate figures and results from this manuscript (and associated usage notes) are available from https://github.com/biobakery/maaslin2_benchmark. The following R packages were used to generate the manuscript figures: ComplexHeatmap[61], ggalluvial[62], ggplot2[63], UpSetR[64], and cowplot[65].

Acknowledgements

This work was funded in part by US National Institutes of Health grants U19AI110820 (to Owen White), R01HG005220 (to Rafael Irizarry), and R24DK110499 and U54DK102557 (CH).

Supplementary Materials

Supplementary Figure S1: Details of simulation parameters, evaluation metrics, and benchmarking methods. A. Four broad metadata designs commonly encountered in microbiome epidemiology for varying sample size, effect size, and feature dimensions are considered: UVA (Single continuous metadata), UVB (Single binary metadata), MVA (Multiple independent metadata), and MVB (Multiple correlated metadata). For each of this broad metadata design, both cross-sectional and longitudinal cases are evaluated (**Methods**). **B.** Three aspects of performance are considered: (i) false discovery, (ii) sensitivity, and (iii) scope and computational efficiency of the associated software, each comprising multiple evaluation metrics (**Methods**). **C.** A combination of statistical models, normalization, and transformation schemes are employed to the synthetic datasets for a variety of association methods, leading up to 84 combinations of normalization/transformation, zero-inflation, and regression models.

Supplementary Figure S2: Full summary of detection performance for varying effect size, sample size, and feature dimensions in the simple case of univariate binary metadata without repeated measures. Both sensitivity and false discovery rates (FDR) are shown for the best-performing methods from each class of methods (as measured by average F1 score). Values are averages over 100 iterations for each parameter combination. The x-axis (effect size) within each panel represents the linear effect size parameter; a higher effect size represents a stronger association. For visualization purposes, the best-performing methods from each class of models (as measured by average F1 score) are shown. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting. All methods were parallelized using custom bash scripts in a high-performance computing environment and methods unable to process specific simulation configurations due to high computational overhead or slow convergence were omitted for those cases.

Supplementary Figures S3: Meta-summary of detection performance in the simple case of univariate binary metadata without repeated measures. Detection performance measures (Sensitivity, FPR, FDR) for all methods are provided. Values are averages over all parameter combinations each summarized over 100 iterations. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figures S4: Meta-summary of p-value calibration performance in the simple case of univariate binary metadata without repeated measures. P-value calibration measures as measured by 'departure from uniformity' (Liberal Area, Conservative Area, Total Area) for all methods are provided. Values are averages over all parameter combinations. Values are averages over all parameter combinations each summarized over 100 iterations. Redd line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figure S5: Full summary of detection performance for varying effect size, sample size, and feature dimensions in the simple case of univariate continuous metadata without repeated measures. Both sensitivity and false discovery rates (FDR) are shown for the best-performing methods from each class of methods (as measured by average F1 score). Values are averages over 100 iterations for each parameter combination. The x-axis (effect size) within each panel represents the linear effect size parameter; a higher effect size represents a stronger association. For visualization purposes, the best-performing methods from each class of models (as measured by average F1 score) are shown. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting. All methods were parallelized using custom bash scripts in a high-performance computing environment and methods unable to process specific simulation configurations due to high computational overhead or slow convergence were omitted for those cases.

Supplementary Figures S6: Meta-summary of detection performance in the presence of multiple independent covariates without repeated measures. Detection performance measures (F1 score, Matthew's correlation coefficient, FDR) for all methods are provided. Values are averages over all parameter combinations each summarized over 100 iterations. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figure S7: Full summary of detection performance for varying effect size, sample size, and feature dimensions in the presence of multiple independent covariates without repeated measures. Both sensitivity and false discovery rates (FDR) are shown for the best-performing methods from each class of methods (as measured by average F1 score). Values are averages over 100 iterations for each parameter combination. The x-axis (effect size) within each panel represents the linear effect size parameter; a higher effect size represents a stronger association. For visualization purposes, the best-performing methods from each class of models (as measured by average F1 score) are shown. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting. All methods were parallelized using custom bash scripts in a high-performance computing environment and methods unable to process specific simulation configurations due to high computational overhead or slow convergence were omitted for those cases.

Supplementary Figures S8: Meta-summary of detection performance in the presence of repeated measures and univariate binary metadata. Detection performance measures (Sensitivity, FPR, FDR) for all methods are provided. Values are averages over all parameter combinations each summarized over 100 iterations. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figures S9: Meta-summary of detection performance in the presence of repeated measures and multiple independent covariates. Detection performance measures (Sensitivity, FPR, FDR) for all methods are provided. Values are averages over all parameter combinations each summarized over 100 iterations. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figures S10: Meta-summary of detection performance in the presence of repeated measures and univariate continuous metadata. Detection performance measures (Sensitivity, FPR, FDR) for all methods are provided. Values are averages over all parameter combinations each summarized over 100 iterations. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figures S11: Meta-summary of detection performance in the presence of repeated measures and multiple correlated covariates. Detection performance measures (Sensitivity, FPR, FDR) for all methods are provided. Values are averages over all parameter combinations each summarized over 100 iterations. Red line parallel to the x-axis is the target threshold for FDR in multiple testing. Methods are sorted by increasing order of average F1 score across all simulation parameters in this setting.

Supplementary Figure S12. **Runtime of association methods**. CPU time (in minutes) is shown for all models faceted by feature dimension (100, 200, 500) and colored by metadata design (i.e.

univariate and multivariable) in both cross-sectional (top) and longitudinal (bottom) settings. Values are averages over 100 iterations for each parameter combination. All methods were parallelized using custom bash scripts in a high-performance computing environment and methods unable to process specific simulation configurations due to high computational overhead or slow convergence were omitted for those cases.

Supplementary Figure S13. **Performance of multivariable association methods on negative training data across a range of significance levels**. MaAsLin 2's default linear model produced a consistently lower proportion of significant associations in negative training data (or repeatedly shuffled training set) (averaged over 1,000 permutations) than the positive training (unshuffled) counterpart in both baseline and longitudinal models (**Methods**).

Supplementary Figure S14: **Statistically significant overlap of detected features by MaAsLin 2 and those found in the original study**. Contingency tables describing the intersection of detected features between MaAsLin 2 and Lloyd-Price et al.[36] for various data modalities in the IBDMDB dataset are shown.

Supplementary Figure S15: **Overlap of detected taxonomic features by various MaAsLin models**. Upset plot describing the intersection of detected taxonomic features between various MaAsLin 2 models in the IBDMDB dataset reveal significant overlap across methods. A similar pattern was observed for functional profiles.

Supplementary Datasets S1-S6: MaAsLin 2 associations between HMP2 multi-omics features and covariates. List of statistically significant associations (FDR<0.25) between IBD disease phenotype (with non-IBD as reference), IBD dysbiosis state (with non-dysbiotic as reference), age, and antibiotic use with multiple data modalities (**S1**: species, **S2**: unstratified DNA pathways, **S3:** unstratified RNA pathways, **S4:** pathway RNA/DNA ratios, **S5**: stratified DNA pathways, **S6**: stratified RNA pathways) using a multivariable linear mixed effects model (**Methods**). Features are sorted by minimum FDR-adjusted p-values. For each feature, coefficient estimates and test statistics and the associated two-tailed p-values are also reported. For each data modality, input features and metadata are also provided.

References

1 Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N Engl J Med* **375**, 2369-2379, doi:10.1056/NEJMra1600266 (2016).

2 Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr Opin Gastroenterol* **31**, 69-75, doi:10.1097/mog.0000000000000139 (2015).

3 Franzosa, E. A. *et al.* Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat Rev Microbiol* **13**, 360-372, doi:10.1038/nrmicro3451 (2015).

4 Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83, doi:10.1186/s13059-017-1215-1 (2017).

5 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).

6 iHMPConsortium. The Integrative Human Microbiome Project. *Nature* **569**, 641-648, doi:10.1038/s41586-019-1238-8 (2019).

7 Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief Bioinform* **20**, 210-221, doi:10.1093/bib/bbx104 (2019).

8 Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27, doi:10.1186/s40168-017-0237-y (2017).

9 Thorsen, J. *et al.* Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4**, 62, doi:10.1186/s40168-016-0208-8 (2016).

10 McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* **10**, e1003531, doi:10.1371/journal.pcbi.1003531 (2014).

11 Mallick, H. *et al.* Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol* **18**, 228, doi:10.1186/s13059-017-1359-z (2017).

12 Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* **17**, 78, doi:10.1186/s12864-016-2386-y (2016).

13 Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. Variability in Metagenomic Count Data and Its Influence on the Identification of Differentially Abundant Genes. *J Comput Biol* **24**, 311-326, doi:10.1089/cmb.2016.0180 (2017).

14 Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* **26**, 27663, doi:10.3402/mehd.v26.27663 (2015).

15 Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* **13**, R79, doi:10.1186/gb-2012-13-9-r79 (2012).

16 Morgan, X. C. *et al.* Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol* **16**, 67, doi:10.1186/s13059-015-0637-x (2015).

17 Zhang, X. *et al.* Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics* **18**, 4, doi:10.1186/s12859-016-1441-7 (2017).

18 Sharpton, T. *et al.* Development of Inflammatory Bowel Disease Is Linked to a Longitudinal Restructuring of the Gut Metagenome in Mice. *mSystems* **2**, doi:10.1128/mSystems.00036-17 (2017).

19 Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. *mSystems* **4**, doi:10.1128/mSystems.00332-18 (2019).

20 Xinyan, Z., Himel, M. & Nengjun, Y. Zero-inflated negative binomial regression for differential abundance testing in microbiome studies. *Journal of Bioinformatics and Genomics*, 1-1 (2016).

21 Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**, 1200-1202, doi:10.1038/nmeth.2658 (2013).

22 Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of Microbiome Data in the Presence of Excess Zeros. *Front Microbiol* **8**, 2114, doi:10.3389/fmicb.2017.02114 (2017).

23 Peng, X., Li, G. & Liu, Z. Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. *J Comput Biol* **23**, 102-110, doi:10.1089/cmb.2015.0157 (2016).

24 Chen, E. Z. & Li, H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32**, 2611-2617, doi:10.1093/bioinformatics/btw308 (2016).

25 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

26 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

27 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).

28 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

29 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3, doi:10.2202/1544-6115.1027 (2004).

30 Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol* **19**, 24, doi:10.1186/s13059-018-1406-4 (2018).

31 McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217, doi:10.1371/journal.pone.0061217 (2013).

32 Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol* **21**, 191, doi:10.1186/s13059-020-02104-1 (2020).

33 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

34 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25, doi:10.1186/gb-2010-11-3-r25 (2010).

35 McKnight, D. T. *et al.* Methods for normalizing microbiome data: an ecological perspective. *Methods in Ecology and Evolution* **10**, 389-400 (2019).

36 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).

37 Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**, 962-968, doi:10.1038/s41592-018-0176-y (2018).

38 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* **46**, D633-d639, doi:10.1093/nar/gkx935 (2018).

39 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902-903, doi:10.1038/nmeth.3589 (2015).

40 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289-300 (1995).

41 Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* **3**, 337-346, doi:10.1038/s41564-017-0089-z (2018).

42 Hawinkel, S., Rayner, J. C. W., Bijnens, L. & Thas, O. Sequence count data are poorly fit by the negative binomial distribution. *PLoS One* **15**, e0224909, doi:10.1371/journal.pone.0224909 (2020).

43 Venables, W. N. & Ripley, B. D. *Modern applied statistics with S-PLUS*. (Springer Science & Business Media, 2013).

44 Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67** (2015).

45 Zeileis, A., Kleiber, C. & Jackman, S. Regression models for count data in R. *Journal of statistical software* **27**, 1-25 (2008).

46 Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**, 139-160 (1982).

47 Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**, 2224, doi:10.3389/fmicb.2017.02224 (2017).

48 VanderWeele, T. J. Mediation Analysis: A Practitioner's Guide. *Annu Rev Public Health* **37**, 17-32, doi:10.1146/annurev-publhealth-032315-021402 (2016).

49 Ma, S. *et al.* Population Structure Discovery in Meta-Analyzed Microbial Communities and Inflammatory Bowel Disease. *bioRxiv* (2020).

50 Gibson, T. E. & Gerber, G. K. Robust and scalable models of microbiome dynamics. *arXiv preprint arXiv:1805.04591* (2018).

51 Ren, B. S., E; Tickle, T; Huttenhower, C sparseDOSSA: Sparse Data Observations for Simulating Synthetic Abundance. R package version 1.12.0. (2020).

52 Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382-392, doi:10.1016/j.chom.2014.02.005 (2014).

53 Imhann, F. *et al.* Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* **67**, 108-119, doi:10.1136/gutjnl-2016-312135 (2018).

54 Cui, S., Ji, T., Li, J., Cheng, J. & Qiu, J. What if we ignore the random effects when analyzing RNA-seq data in a multifactor experiment. *Stat Appl Genet Mol Biol* **15**, 87-105, doi:10.1515/sagmb-2015-0011 (2016).

55 Stasinopoulos, D. M. & Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. (2007).

56 Zhang, Y. Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models. *Statistics and Computing* **23**, 743-757 (2013).

57 Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-3941, doi:10.1093/bioinformatics/bti623 (2005).

58 Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**, 255-261, doi:10.1038/nmeth.4612 (2018).

59 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **42**, D459-471, doi:10.1093/nar/gkt1103 (2014).

60 McIver, L. J. *et al.* bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235-1237, doi:10.1093/bioinformatics/btx754 (2018).

61 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849, doi:10.1093/bioinformatics/btw313 (2016).

62 Brunson, J. C. ggalluvial: Layered Grammar for Alluvial Plots. *Journal of Open Source Software* **5**, 2017 (2020).

63 Wickham, H. *ggplot2: elegant graphics for data analysis*. (springer, 2016).

64 Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics* **20**, 1983-1992 (2014).

65 Wilke, C. O., Wickham, H. & Wilke, M. C. O. Package ‘cowplot’. *Streamlined Plot Theme and Plot Annotations for ‘ggplot2* (2019).

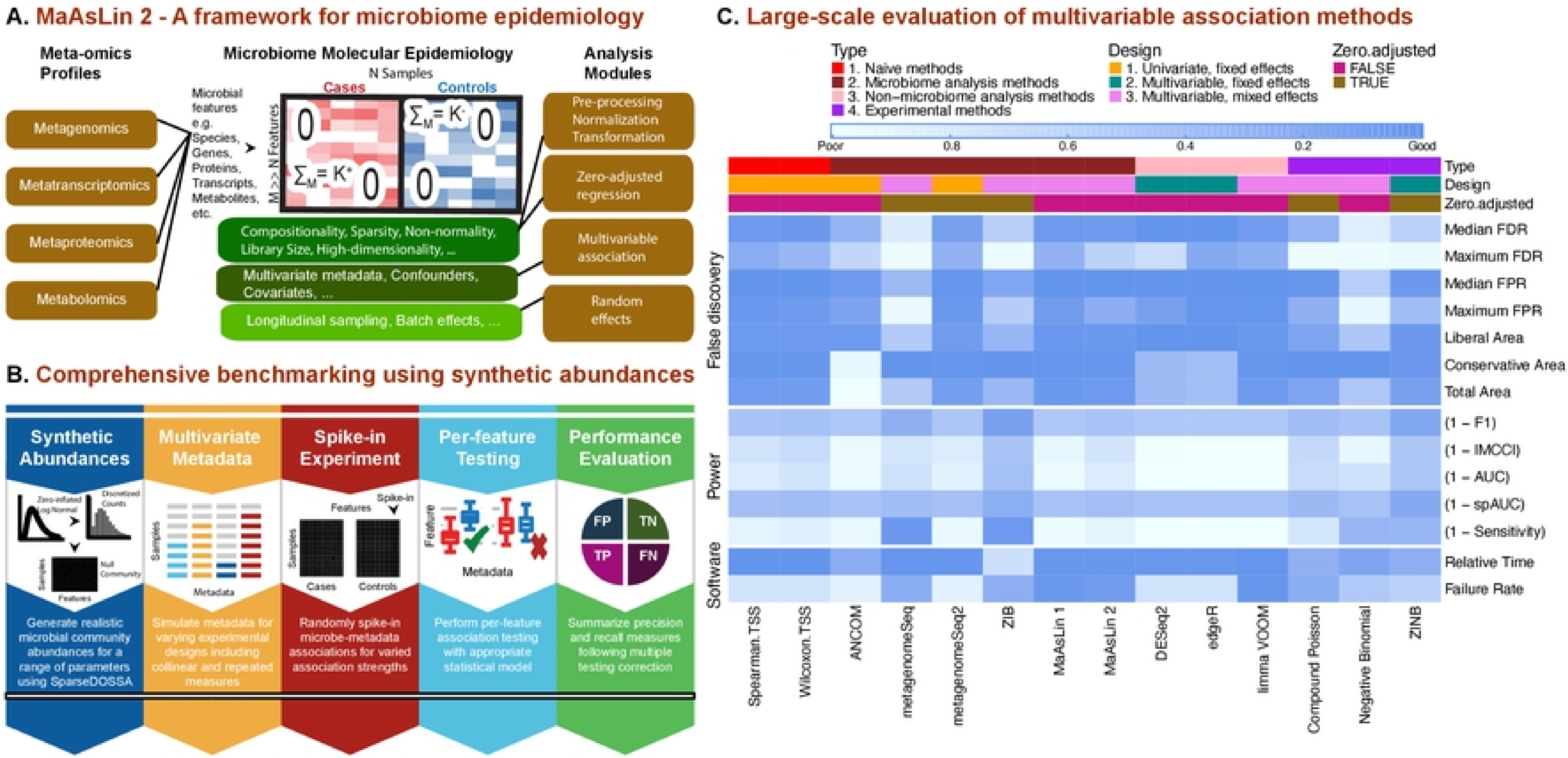

A. MaAsLin 2 - A framework for microbiome epidemiology
Meta-omics Profiles
Metagenomics
Metatranscriptomics
Metaproteomics
Metabolomics
Microbial features e.g. Species, Genes, Proteins, Transcripts, Metabolites, etc.
Microbiome Molecular Epidemiology
N Samples
Cases
Controls
M >> N Features
0
$\Sigma_M = K^*$
Compositionality, Sparsity, Non-normality, Library Size, High-dimensionality, ...
Multivariate metadata, Confounders, Covariates, ...
Longitudinal sampling, Batch effects, ...
Analysis Modules
Pre-processing Normalization Transformation
Zero-adjusted regression
Multivariable association
Random effects
B. Comprehensive benchmarking using synthetic abundances
Synthetic Abundances
Zero-inflated Log Normal
Discretized Counts
Samples
Null Community
Features
Generate realistic microbial community abundances for a range of parameters using SparseDOSSA
Multivariate Metadata
Samples
Metadata
Simulate metadata for varying experimental designs including collinear and repeated measures
Spike-in Experiment
Features
Spike-in
Samples
Cases
Controls
Randomly spike-in microbe-metadata associations for varied association strengths
Per-feature Testing
Feature
Metadata
Perform per-feature association testing with appropriate statistical model
Performance Evaluation
FP
TN
TP
FN
Summarize precision and recall measures following multiple testing correction
C. Large-scale evaluation of multivariable association methods
Type
1. Naive methods
2. Microbiome analysis methods
3. Non-microbiome analysis methods
4. Experimental methods
Design
1. Univariate, fixed effects
2. Multivariable, fixed effects
3. Multivariable, mixed effects
Zero.adjusted
FALSE
TRUE
Poor
0.8
0.6
0.4
0.2
Good
Type
Design
Zero.adjusted
False discovery
Median FDR
Maximum FDR
Median FPR
Maximum FPR
Liberal Area
Conservative Area
Total Area
Power
(1 – F1)
(1 – IMCCI)
(1 – AUC)
(1 – spAUC)
(1 – Sensitivity)
Software
Relative Time
Failure Rate
Spearman.TSS
Wilcoxon.TSS
ANCOM
metagenomeSeq
metagenomeSeq2
ZIB
MaAsLin 1
MaAsLin 2
DESeq2
edgeR
limma VOOM
Compound Poisson
Negative Binomial
ZINB

Figure 1

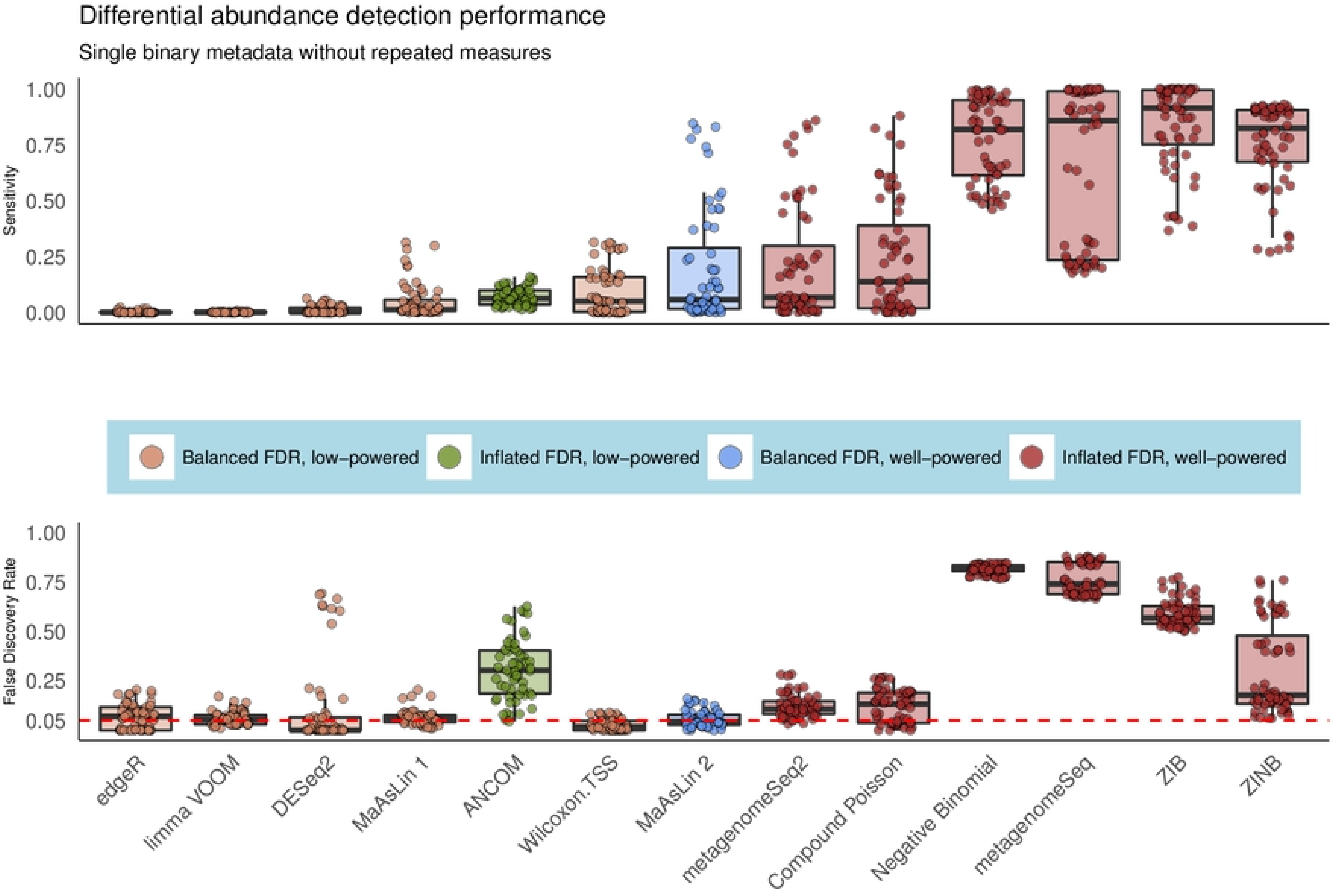
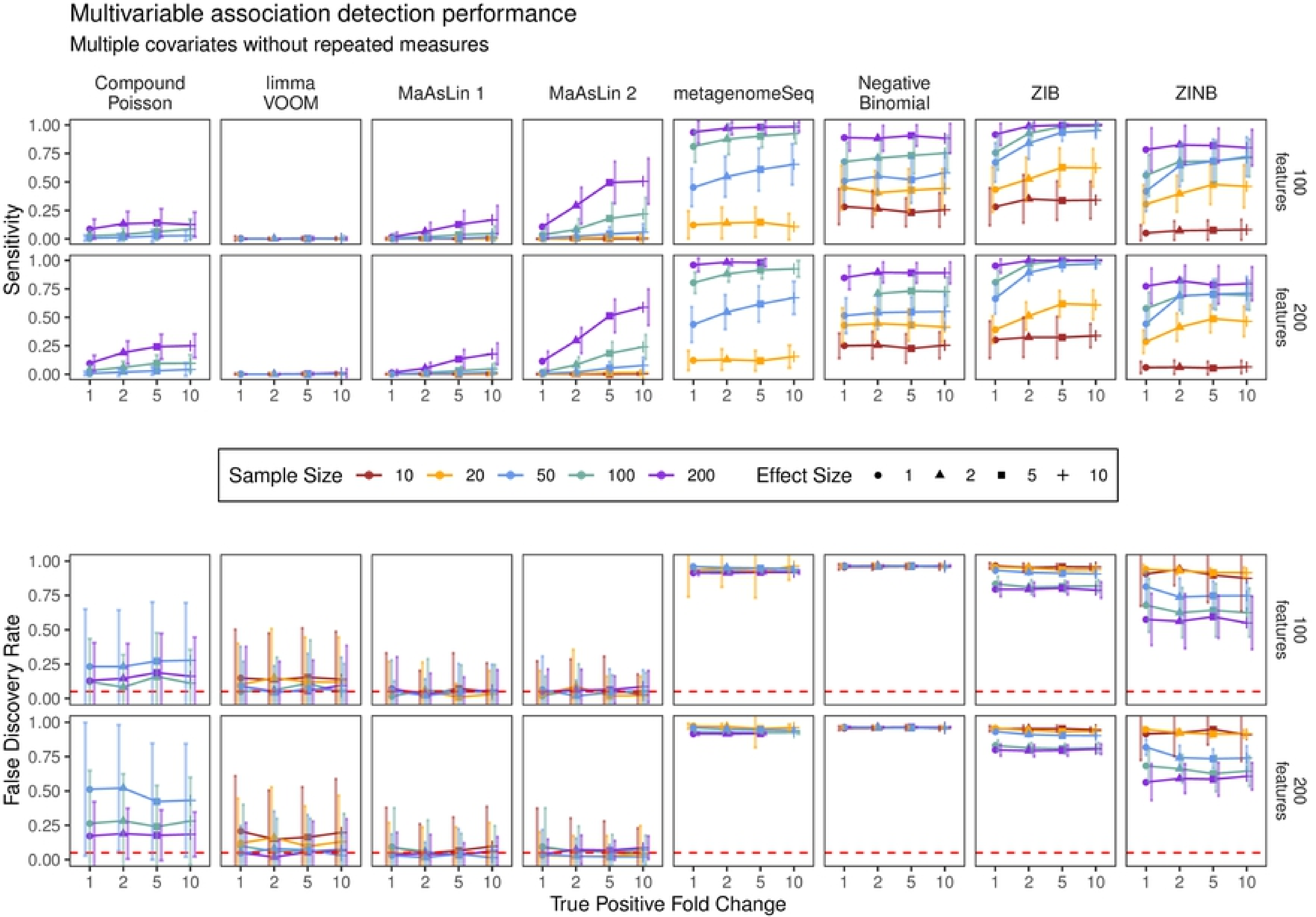

Differential abundance detection performance
Single binary metadata without repeated measures
Sensitivity
1.00
0.75
0.50
0.25
0.00
Balanced FDR, low−powered
Inflated FDR, low−powered
Balanced FDR, well−powered
Inflated FDR, well−powered
False Discovery Rate
1.00
0.75
0.50
0.25
0.05
edgeR
limma VOOM
DESeq2
MaAsLin 1
ANCOM
Wilcoxon.TSS
MaAsLin 2
metagenomeSeq2
Compound Poisson
Negative Binomial
metagenomeSeq
ZIB
ZINB

Figure 2



Multivariable association detection performance
Multiple covariates without repeated measures
Compound Poisson
limma VOOM
MaAsLin 1
MaAsLin 2
metagenomeSeq
Negative Binomial
ZIB
ZINB
Sensitivity
0.00
0.25
0.50
0.75
1.00
100 features
200 features
1
2
5
10
Sample Size
10
20
50
100
200
Effect Size
● 1
▲ 2
■ 5
+ 10
False Discovery Rate
True Positive Fold Change

Figure 3

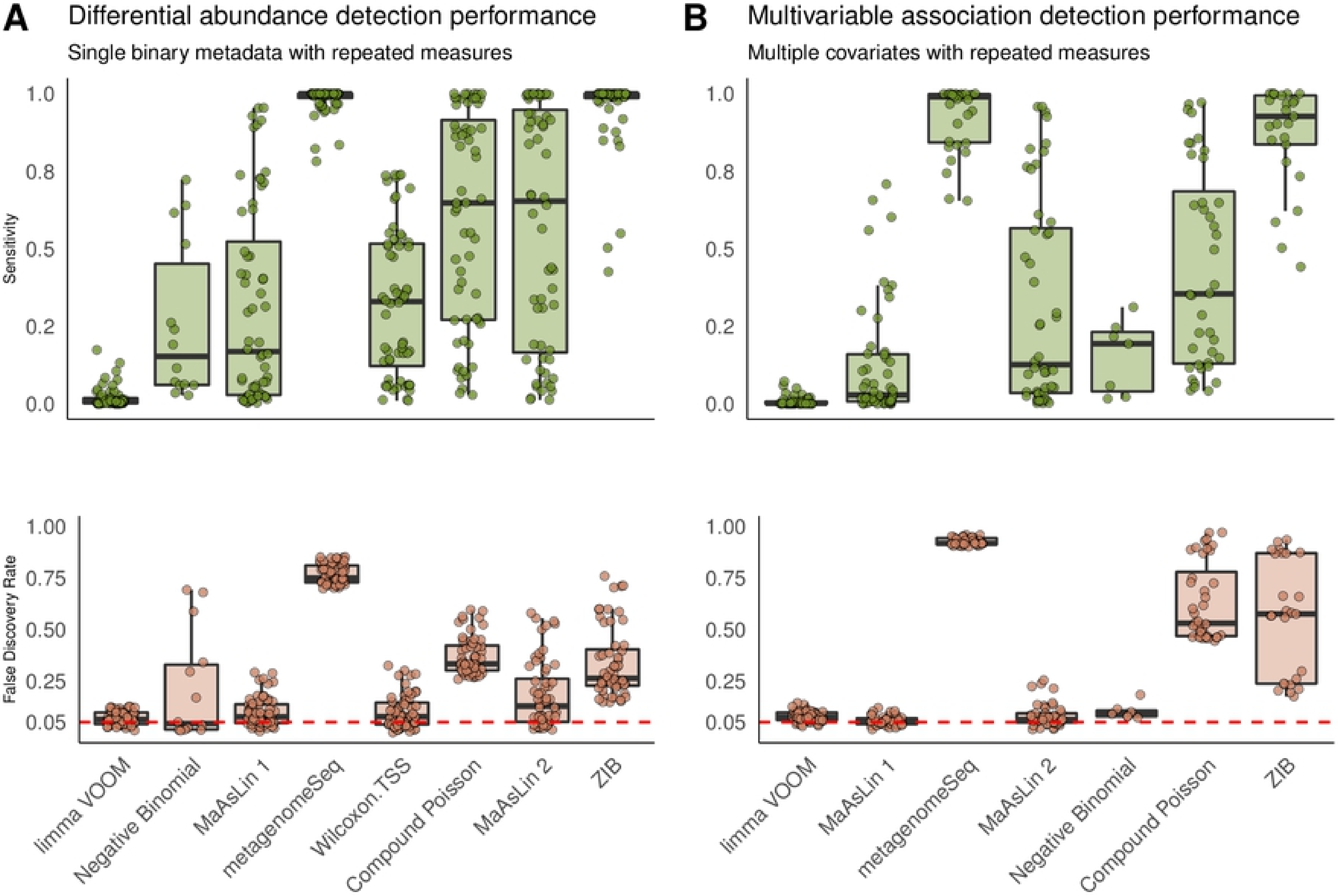
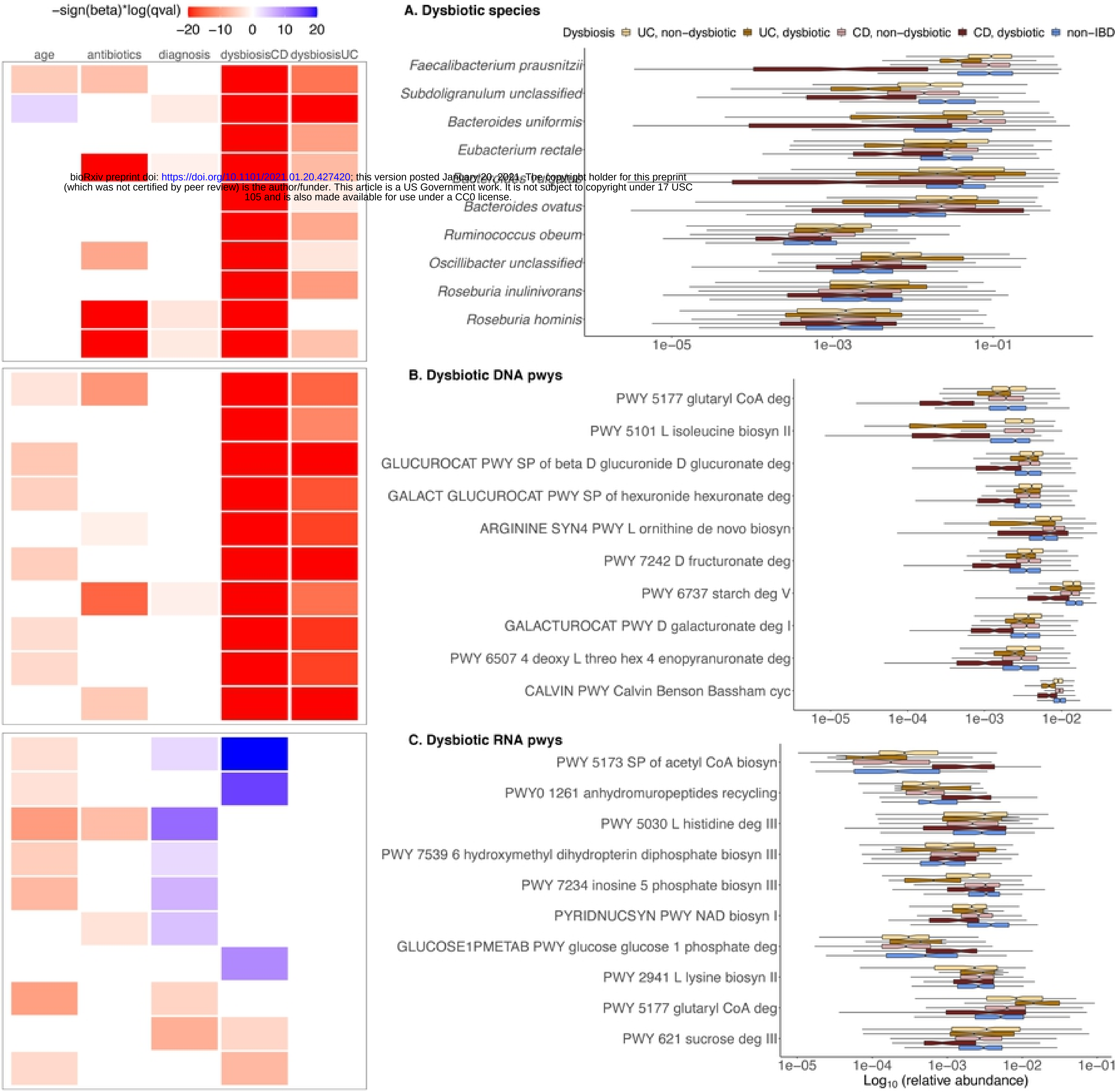

A
Differential abundance detection performance
Single binary metadata with repeated measures
B
Multivariable association detection performance
Multiple covariates with repeated measures
Sensitivity
1.0
0.8
0.5
0.2
0.0
False Discovery Rate
1.00
0.75
0.50
0.25
0.05
limma VOOM
Negative Binomial
MaAsLin 1
metagenomeSeq
Wilcoxon.TSS
Compound Poisson
MaAsLin 2
ZIB
limma VOOM
MaAsLin 1
metagenomeSeq
MaAsLin 2
Negative Binomial
Compound Poisson
ZIB

Figure 4



−sign(beta)*log(qval)
−20 −10 0 10 20
age
antibiotics
diagnosis
dysbiosisCD
dysbiosisUC
A. Dysbiotic species
Dysbiosis
UC, non−dysbiotic
UC, dysbiotic
CD, non−dysbiotic
CD, dysbiotic
non−IBD
Faecalibacterium prausnitzii
Subdoligranulum unclassified
Bacteroides uniformis
Eubacterium rectale
Bacteroides vulgatus
Bacteroides ovatus
Ruminococcus obeum
Oscillibacter unclassified
Roseburia inulinivorans
Roseburia hominis
1e−05
1e−03
1e−01
B. Dysbiotic DNA pwys
PWY 5177 glutaryl CoA deg
PWY 5101 L isoleucine biosyn II
GLUCUROCAT PWY SP of beta D glucuronide D glucuronate deg
GALACT GLUCUROCAT PWY SP of hexuronide hexuronate deg
ARGININE SYN4 PWY L ornithine de novo biosyn
PWY 7242 D fructuronate deg
PWY 6737 starch deg V
GALACTUROCAT PWY D galacturonate deg I
PWY 6507 4 deoxy L threo hex 4 enopyranuronate deg
CALVIN PWY Calvin Benson Bassham cyc
1e−05
1e−04
1e−03
1e−02
C. Dysbiotic RNA pwys
PWY 5173 SP of acetyl CoA biosyn
PWY0 1261 anhydromuropeptides recycling
PWY 5030 L histidine deg III
PWY 7539 6 hydroxymethyl dihydropterin diphosphate biosyn III
PWY 7234 inosine 5 phosphate biosyn III
PYRIDNUCSYN PWY NAD biosyn I
GLUCOSE1PMETAB PWY glucose glucose 1 phosphate deg
PWY 2941 L lysine biosyn II
PWY 5177 glutaryl CoA deg
PWY 621 sucrose deg III
1e−05
1e−04
1e−03
1e−02
1e−01
Log10 (relative abundance)

Figure 5