Evaluation of microbiome association models under realistic and confounded conditions

Jakob Wirbel[1,*], Morgan Essex[2,3,4,*], Sofia Kirke Forslund[1,2,3,4,5,†], Georg Zeller[1,†]

1. Structural and Computational Biology Unit (SCB), EMBL Heidelberg, Germany
2. Experimental and Clinical Research Center, a cooperation of the Max-Delbrück Center for Molecular Medicine and Charité-Universitätsmedizin, Berlin, Germany
3. Max-Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany
4. Charité-Universitätsmedizin Berlin, Berlin, Germany
5. German Center for Cardiovascular Research (DZHK), partner site Berlin

* both authors contributed equally to the work
† correspondence should be addressed to sofia.forslund@mdc-berlin.de or zeller@embl.de

Abstract

Testing for differential abundance is a crucial task in metagenome-wide association studies, complicated by technical or biological confounding and a lack of consensus regarding statistical methodology. Here, we developed a framework for benchmarking differential abundance testing methods based on implanting signals into real data. This strategy yields a ground truth for benchmarking while retaining the statistical characteristics of real metagenomic data, which we quantitatively validated in comparison to previous approaches. Our benchmark revealed dramatic issues with elevated false discovery rates or limited sensitivity for the majority of methods with the exception of limma, linear models and the Wilcoxon test. When additionally modeling confounders, we observed these issues to be exacerbated, but also that linear mixed-effect models or the blocked Wilcoxon test effectively address them. Exploratory analysis of cardiometabolic disease cohorts illustrates the confounding potential of medications and the need to consider confounders to prevent spurious associations in real-world applications.

149 words

Keywords: Microbiome, Benchmark, Metagenomics, Differential abundance, Confounding

Introduction

The human gut microbiome is increasingly understood to perform critical roles in host physiology and immunity, and has therefore been extensively mined for biomarkers related to host health and disease states. The composition of gut microbial communities is highly variable between individuals[1], yet clinical metagenome-wide association studies (MWAS) strive to overcome inter-individual differences in associating taxonomic and functional features with grouping variables such as disease phenotypes or lifestyle factors. Researchers have performed MWAS in the context of numerous diseases, including but not limited to inflammatory bowel diseases[2,3], gastrointestinal cancers[4,5], and cardiometabolic diseases[6]. To identify group-specific microbiome alterations, MWAS typically perform statistical tests for differential abundance (DA) on each microbial feature independently, borrowing both nomenclature and entire methods from differential gene expression analysis. DA methods applied in MWAS loosely fall into four categories: a) methods adapted from RNA-Seq analysis, b) generalized linear models, c) non-parametric statistical tests, and finally d) methods developed specifically for microbiome data.

While significant microbiome disease associations have been reported in many studies, some meta-analyses and cross-disease comparisons have suggested many of these associations to be unspecific or confounded[7–9]. Microbiome composition varies not only with host health and disease states, but also with myriad other host and environmental factors (covariates) collectively estimated to explain nearly 20% of microbiome variation[10]. Lifestyle and physiological covariates demonstrating the largest effects on microbial communities include medication regimens[11,12], stool quality, geography, and alcohol consumption frequency[13]. Technical differences such as stool sample collection and DNA extraction method, if present, often outweigh biological factors of interest in terms of explained variation, and can therefore hamper meta-analyses if unaccounted for[5,14,15].

Although the unique statistical challenges intrinsic to analysis of high-throughput metagenomic data are well-described by now[16–18], there is no consensus about the most appropriate DA procedures in the literature[19–25]. In principle, this is the purpose of benchmarking studies, which try to assert how methods perform under varying yet controlled conditions in order to establish points of reference for method behavior that cannot be discerned from single applications. To achieve such conditions, simulated data is typically generated from parametric models, which cleanly specify the differentially abundant features required for performance evaluation as a ground truth.

A high-level concern with benchmarking is the phenomenon of over-optimistic performance evaluation observed in newly introduced bioinformatic methods, whereby data, results, and competitor tools in a benchmarking study (frequently published in the same manuscript as the new method) suffer from selection and reporting bias[26]. The lack of consensus on how to simulate data in microbiome research, however, is a much more fundamental problem; an evaluation of simulation methods on the basis of their resemblance to real experimental data, and the impact this has on downstream applications in benchmarks, has not yet been conducted.

Here, we quantitatively assess the degree to which parametric simulations employed in previous DA benchmarks lack biological realism, and show that the choice of simulation framework can explain the divergent recommendations regarding DA methods made in previous studies. To address these shortcomings, we propose a novel simulation technique using spike-ins of differential abundance between two groups (imitating a case-control MWAS) into real data, and extend it to additionally include confounded effects. Based on

these more realistic simulations, we perform a comprehensive benchmarking study of widely used DA methods revealing an alarmingly common inability to control the false discovery rate (FDR). Under confounded conditions, full FDR control is unattainable by any method, but we show how to adjust some methods so that they suffer a lot less from these issues, and explore the merits of doing so in a real clinical dataset.

Results

Assessment of realism for parametric simulations of microbiome data

As a first step, we aimed to explicitly evaluate how data generated from previous simulation frameworks compare with real metagenomic data. To do so, we simulated taxonomic profiles using the source code employed in previous benchmarks[20,22,23,27], whereby case-control datasets were repeatedly generated with differentially abundant features introduced under varying effect sizes (see **Methods**). Simulation parameters were estimated in each case from the same baseline dataset of healthy adults[28]. We observed the data simulated with every one of the tested parametric models to be very different from real data as visualized by principal coordinate analysis (see **Fig. 1a**). Additionally, there was a large discrepancy between the feature variance and sparsity of simulated profiles and what is observed in real metagenomic data (see **Fig. 1b** and **SFig. 1**), where the multinomial method in particular dramatically underestimated feature variance. Finally we trained machine learning classifiers to distinguish between real and simulated samples, which was possible with almost perfect accuracy in nearly all cases, except for data generated from sparseDOSSA[27] (**Fig. 1c**). This classification attempt was motivated by the fact that machine learning, commonly employed in MWAS to detect biomarkers, can reveal even subtle differences between groups and is generally more sensitive than ordination-based analyses[9]. Overall, each of the assessed parametric simulation frameworks failed to produce realistic metagenomic data.

Signal implantation yields realistic benchmarking datasets

To devise a simulation framework which would generate data that closely recapitulates key characteristics of metagenomic data, we opted to manipulate real baseline data as little as possible, by implanting a known signal with pre-defined effect size into a small number of features using randomly selected groups (see **Methods**). As the baseline dataset, we chose a cohort consisting of healthy adults without obvious biological groupings, into which we repeatedly implanted a differentially abundant signal by multiplying the counts in one group with a constant (abundance scaling) and/or by shuffling a certain percentage of non-zero entries across groups (prevalence shift, see **Methods**). The main advantage of this proposed signal implantation approach as a foundation for benchmarking is that it generates a clearly defined ground truth of DA features while retaining key characteristics of real data. In particular, feature variance and sparsity (see **Fig. 1b**) are preserved, which is reflected in both the principal coordinate projection (**Fig. 1a**) and in the more sensitive machine learning classification task (**Fig. 1c**).

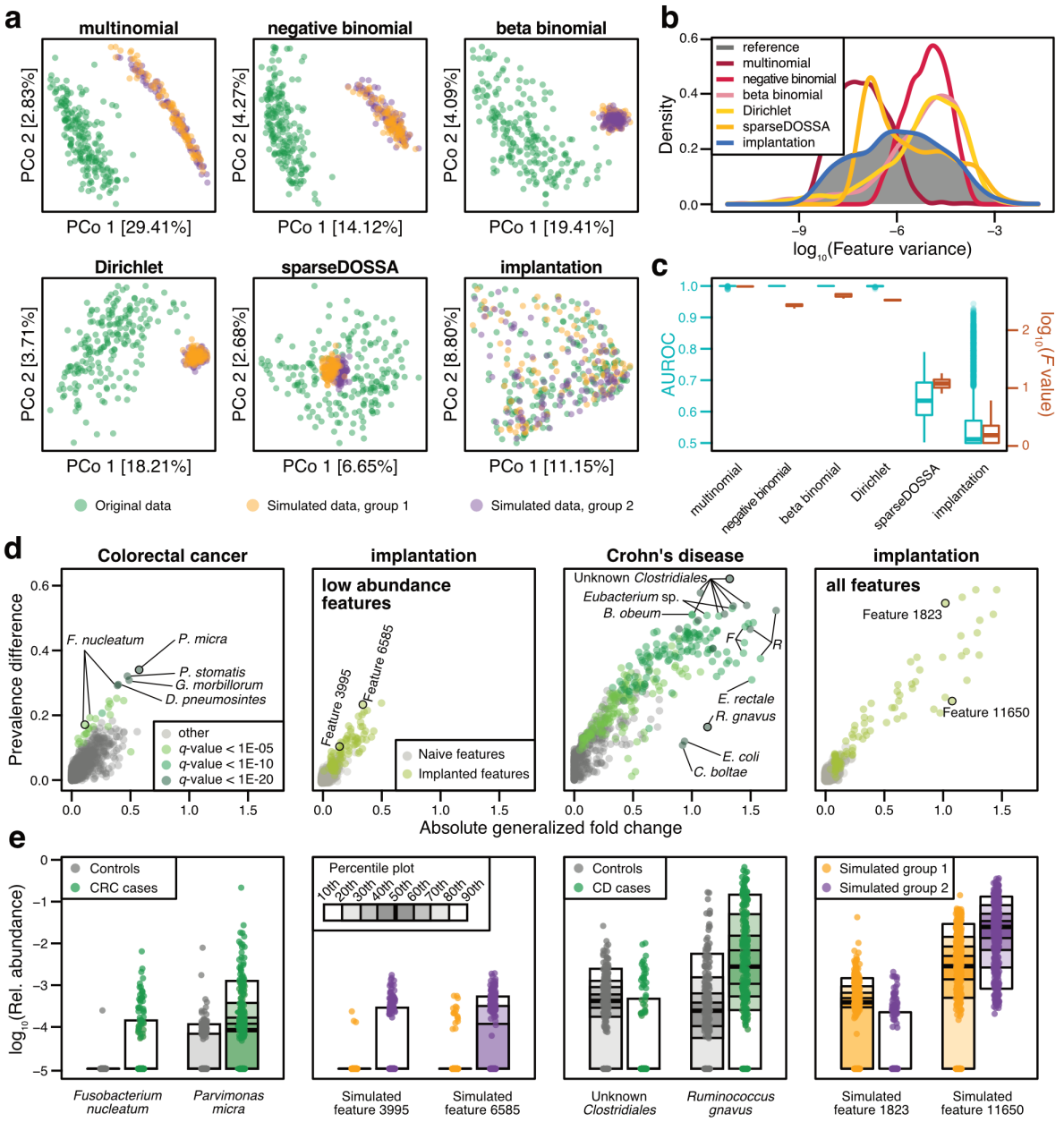


Figure 1: Signal implantation, but not parametric simulations, can reproduce key characteristics of metagenomic data and realistic disease effects

a) Principal coordinate projections on log-Euclidean distances for real samples (from Zeevi et al.[28], which served as a baseline data set) and representative samples of data simulated in a case-control setting (group 1 and 2) using different parametric models or signal implantation. For each method, the results from a single repetition and a fixed effect size are shown (see **Methods**, abundance scaling factor of 2 and prevalence shift of 0.3, if applicable). **b)** The log-transformed feature variance is shown for the real and simulated data from the same simulated data as in a). **c)** The area under the receiver operating curve (AUROC) values from machine learning models to distinguish between real and simulated samples are shown across all simulated data sets in cyan. As complementary information, the log-transformed F values from PERMANOVA tests are shown in brown. **d)** The absolute generalized fold change[5] and the absolute difference in prevalence across groups is shown for all features in colorectal cancer (CRC) and Crohn's disease (CD). As a comparison, the same values are displayed for two data sets simulated using signal implantation (abundance scaling factor of 2, prevalence shift of 0.1), with implantations either into all features or only low-abundance features (see **Methods**). Well-described disease-associated features are highlighted (*F* = *Faecalibacterium*, *R* = *Ruminococcus*) and selected bacterial taxa and simulated features are shown as percentile plot in **e)**.

Implanted DA features are similar to real-world disease effects

To compare implanted DA features to those observed in real MWAS data in terms of their effect sizes, we focused on two diseases with well-established microbiome alterations, namely colorectal cancer (CRC)[4,5] and Crohn's disease (CD)[2,3]. In two separate meta-analyses (see **Methods**), we calculated the generalized fold change as well as the difference in prevalence between controls and the respective cases for each microbial feature (see **Fig. 1de**). The effect sizes in CRC were generally found to be much lower than in CD, which is consistent with machine learning results in both diseases (mean AUROC for the distinction between cases and controls: 0.92 in CD and 0.81 in CRC, see ref[9]). For instance, the well-described CRC marker *Fusobacterium nucleatum* exhibits a rather moderate increase in abundance in CRC, but a strongly increased prevalence. This observation, generalizable also to many other established microbial disease biomarkers, motivated the inclusion of the prevalence shift as an additional type of effect size for the proposed implantation framework.

Depending on the type and strength of effect size used to implant DA features, the simulated datasets included effects that closely resemble those observed in the CRC and CD case-control datasets (**Fig. 1ed**). In particular, simulated abundance shifts with a scaling factor between groups <=10 were the most realistic and thus used for subsequent analyses (**SFig. 2**).

Performance evaluation of differential abundance testing methods

To benchmark the performance of various DA testing methods under realistic conditions, fifteen DA tests (see **Methods** for a list) were applied featurewise across all simulated datasets including repeated sampling with varying effect sizes. Different sample sizes were created by repeatedly selecting random samples from each simulated group in a balanced manner, and each test was applied to the exact same sets of samples (see **Methods**). In general, we aimed to use the recommended data preprocessing steps for each method, but also explored different normalization techniques when appropriate (see **SFig. 3**).

The *P* values resulting from each of the included DA methods were adjusted for multiple hypothesis testing with the Benjamini-Hochberg procedure[29] and recall and precision (equivalent to 1 - false discovery rate (FDR)) were estimated using *P*=0.05 as cutoff. Additionally, a receiver operating characteristic (ROC) analysis was carried out to evaluate how accurately the returned *P* values could distinguish between ground truth and background features. In the ideal case of *P* values for all ground truth features being smaller than for any of the background features, the area under the ROC curve (AUROC) will be one; for random *P* values an AUROC of 0.5 is expected.

In this benchmark, most methods failed to consistently control the FDR at the nominal 5% level, especially for sample sizes under 200 (displayed for a single representative effect size in **Fig. 2a-d** and **SFig. 4** for other effect sizes). In the most extreme case, the *fitZig* method from metagenomeSeq (*mgs*), only around 20% of features identified as significantly differentially abundant between groups were correctly predicted (FDR of 80%), and this behavior was observed across many sample and effect sizes (see **SFig. 4**). Theoretically, a possible explanation for these high FDR values could be that the methods are not well-calibrated for microbiome data and report universally low *P* values, while still able to distinguish between ground truth and background features. Such cases would result in high AUROC values and could theoretically be addressed by changing the *P* value cutoff for significance. Of the methods with insufficient FDR control, only *edgeR* showed comparably high AUROC values and could therefore be recalibrated with a lower *P* value cutoff

(although determining this cutoff in practice is not straightforward), whereas other methods (*mgs*, *ZIBSeq*, *corncob* or *DESeq2*) could not even be theoretically recalibrated to low FDR and high sensitivity.

On the other hand, six DA methods managed to consistently provide reasonable FDR control across all tested settings (FDR never exceeding twice the nominal value, even at smaller sample sizes). Of those, *ALDEx2*, *distinct*, and the Kolmogorov-Smirnov (*KS*) test resulted in comparably low recall and AUROC values (mean AUROC<0.75 over all repetitions), indicating that they are relatively insensitive in identifying true DA features. The other three methods, *limma*, the *Wilcoxon* test, and the linear model (*LM*) manage to control the FDR at comparably high sensitivity values (**Fig. 2e**), thereby emerging as reliable testing frameworks suitable for the analysis of microbiome data.

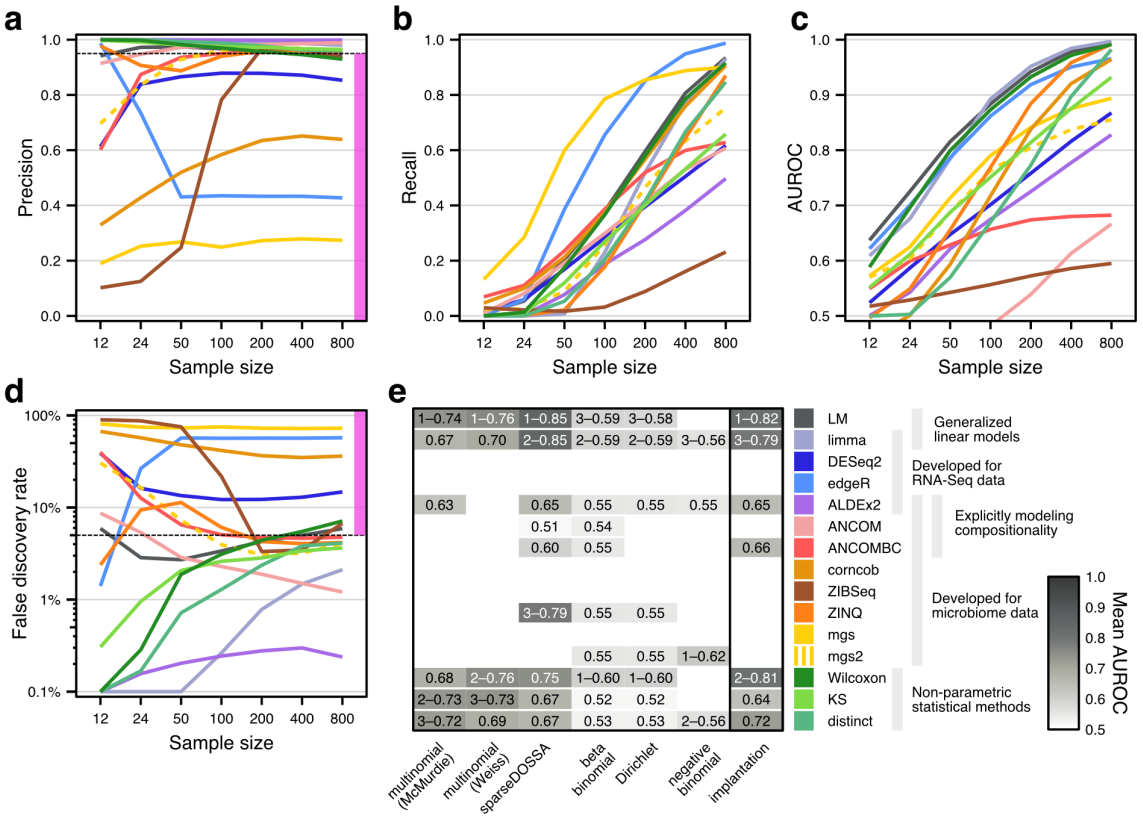


Figure 2: Performance evaluation of differential abundance testing methods
a) For a signal implantation simulation with a single, moderate effect size combination (abundance scaling factor of 2, prevalence shift of 0.2, all features eligible for implantation), the mean precision **a)**, mean recall **b)**, the mean AUROC values **c)**, and the mean log-transformed FDR (equivalent to the inverse of precision) are shown for all included DA testing methods (see also **Methods**) across different sample sizes. For precision and FDR, the nominally expected value (5% FDR or 95% precision) is indicated by a dotted black line and the area of insufficient FDR control is highlighted by a magenta bar on the side. **e)** The mean AUROC values across the sample sizes 50, 100, and 200, all repetitions, and all effect sizes are depicted in the heat map for the different simulation strategies. Cells are not coloured if the mean FDR across all settings exceeded 10% in at least 10% of cases.

Variations of the differential abundance testing benchmark

To estimate the effect of different normalization strategies on the performance of DA methods, we additionally explored common normalization techniques, whenever appropriate. One of the most commonly employed normalization techniques is rarefaction, which is

compatible with all DA methods that use counts as input. On average, rarefaction led to a reduction of sensitivity (lower recall and lower AUROC values) across all tested methods without improving precision (see **SFig. 3**). *Limma* and the *LM* were most affected by some normalizations, which in extreme cases resulted in substantial improvements or decreases of AUROC or recall values, which suggests that appropriate normalization should be taken into account for these methods, namely the TSS-log transformation for the *LM* and TSS-arcsin for *limma* (see **SFig. 3** and **Methods**).

Some of the tested DA methods were developed to explicitly model the compositionality of microbiome data, since compositional effects can theoretically lead to spurious associations[16,30]. In our framework, signals were implanted to be balanced between groups, thereby minimizing compositional effects (see Weiss et al.[22] and **Methods**). To further test DA methods under conditions of stronger compositional effects, we created simulations in which we implanted signals into one group only (see **Methods**), leading to abundance shifts also for background features (see **SFig. 5**). Under these conditions, all tested methods lost FDR control with increasing effect sizes, with *ANCOM* and *ALDEx2* being least affected (FDR around 20% at an abundance shift of 5, FDR > 40% for other methods, **SFig. 5**). However, the (relative) FDR control of these methods was not matched by gains in sensitivity, highlighting a general challenge to disentangle true from spurious signals given extreme compositionality effects.

Lastly, our benchmarking results were found to strongly depend on the simulation method used for data generation, indicating that the distributional assumptions made by these can in turn lead to biased conclusions (see **Fig. 2e** and **SFig. 6**). For example, the benchmarking study by Weiss et al.[22] proposed using the *ANCOM* method for DA testing. While this method shows apparently superior performance when multinomial data distributions are assumed (as done in Weiss et al[22] and reproduced here), this is in stark contrast to benchmarking with the realistic signal implantation setting, where *ANCOM* did not perform significantly better than random guessing for the identification of ground truth DA features (*P* value=0.34, one-sample *t*-test) for sample sizes up to 200. Similarly, the publication by Hawinkel et al.[23] concluded that almost all methods fail to control the FDR. This conclusion, based on metagenomic data simulated using a negative binomial distribution, could be faithfully reproduced here (**Fig. 2b** and **SFig. 6**), but was not supported by the results of our more realistic benchmarking.

Extending the simulation framework to include different types of confounding

In recent years, awareness of confounding as a prevalent issue in MWAS has increased[13,31]. Confounding occurs when covariates other than the main variable of interest explain variation in microbial composition or individual taxon abundances. When not accounted for, confounding can lead to spurious associations that do not replicate in independent datasets. As a well-known example, associations between gut microbiome composition and type 2 diabetes (T2D) from two different studies were later identified to be mainly caused by metformin treatment in a subset of T2D patients[7]. In addition, study or batch effects are prominent in metagenomic data due to non-standardized data generation procedures, which generate technical variation between datasets[32,33]. This is a particular issue for meta-analyses, when multiple MWAS of the same disease are compared.

Whether technical or biological in nature, the confounding potential of a covariate can be assessed by calculating its association with the main variable of interest (e.g. the disease status label). In the case of a covariate which is also binary, this can be quantified by the phi coefficient[34] which functionally ranges from 0 to 1. Large values indicate a class imbalance in

the joint distribution of both variables, implying potential overlap in their explanatory power. For example, disease cohorts are often highly medicated in contrast to mostly unmedicated healthy control groups[35] and therefore display very high phi coefficients between disease labels and intake of standard drug treatments (**SFig. 7**).
To create confounded conditions for our benchmark, we parameterized our signal implantation method to incorporate a bias term analogous to the phi coefficient and used it to generate DA features that were imbalanced between groups with respect to an additional binary variable (see **Methods** and **SFig. 7**). To evaluate how DA methods perform under such conditions, we set up scenarios mimicking broad (affecting a large number of features) or narrow (affecting a smaller subset of features) confounding effects separately, and applied both naive (tested in previous benchmark) and covariate-aware DA tests to compare their performance. We focused on those methods which were found to control the FDR at concomitant high sensitivity in our previous benchmark (*limma*, the *LM*, and the *Wilcoxon* test), and which could be explicitly adjusted for covariates.

Performance evaluation under confounded conditions
To simulate broad confounding resembling for example study effects encountered in meta-analyses or treatment with broad-spectrum antibiotics, we combined the samples from two baseline datasets of healthy adults[28,36] and implanted DA features into the datasets. When we generated the sample sets for testing, a low bias parameter, meaning proportional sampling from both studies, led to minimal confounding (see **Fig. 3a**). Larger bias values resulted in study differences that became more and more aligned with the group label, making it increasingly challenging to discern true DA features from study differences (see **Fig. 3b**). Relying again on the meta-analyses datasets for CRC and CD, we additionally verified the relevance of study effects in real study comparisons, although we noted that a balanced study design could mitigate extreme study confounding in most cases (see **SFig. 8**).
When applying DA tests to the simulated data, we found that without adjustment for the study covariate, all DA tests exhibited an increased FDR, even in the presence of moderate study confounding (median FDR increased 12-fold over the nominal value for *Wilcoxon* and the *LM*, or 60%), which further surged under strong study confounding (up to 16-fold, or 80% FDR, see **Fig. 3c**). We next adjusted the DA methods by including the study covariate in the respective test formula (see **Methods** for details) and observed that the *LM* and the *Wilcoxon* test performed nearly as well as under non-confounded conditions, apart from a moderate loss of sensitivity (median recall of 50% for strong confounding vs 60% for no confounding for the *LM*, see **Fig. 3c**) and a moderate loss of precision for the *LM* (only notable with extremely strong confounding, see also **SFig. 9**). One notable exception was *limma*, which showed a relatively high median FDR (around 60% under moderate confounding) even after adjustment for the study covariate.

To simulate confounding imitating a factor with a strong, but narrow effect on the microbiome, we again began with data from Zeevi et al.[28] and implanted DA features into random groups (mimicking the disease label) as described before. Then, we additionally created an orthogonal confounder label (corresponding to a medication or other biological factor) and implanted distinct sets of DA features into these groups (see **Methods** for details). By again varying the bias parameter while generating the set of samples for testing, we could simulate different levels of narrow confounding which increased as disease and confounder labels became more and more aligned (see **Fig. 4a** and **SFig. 9**).

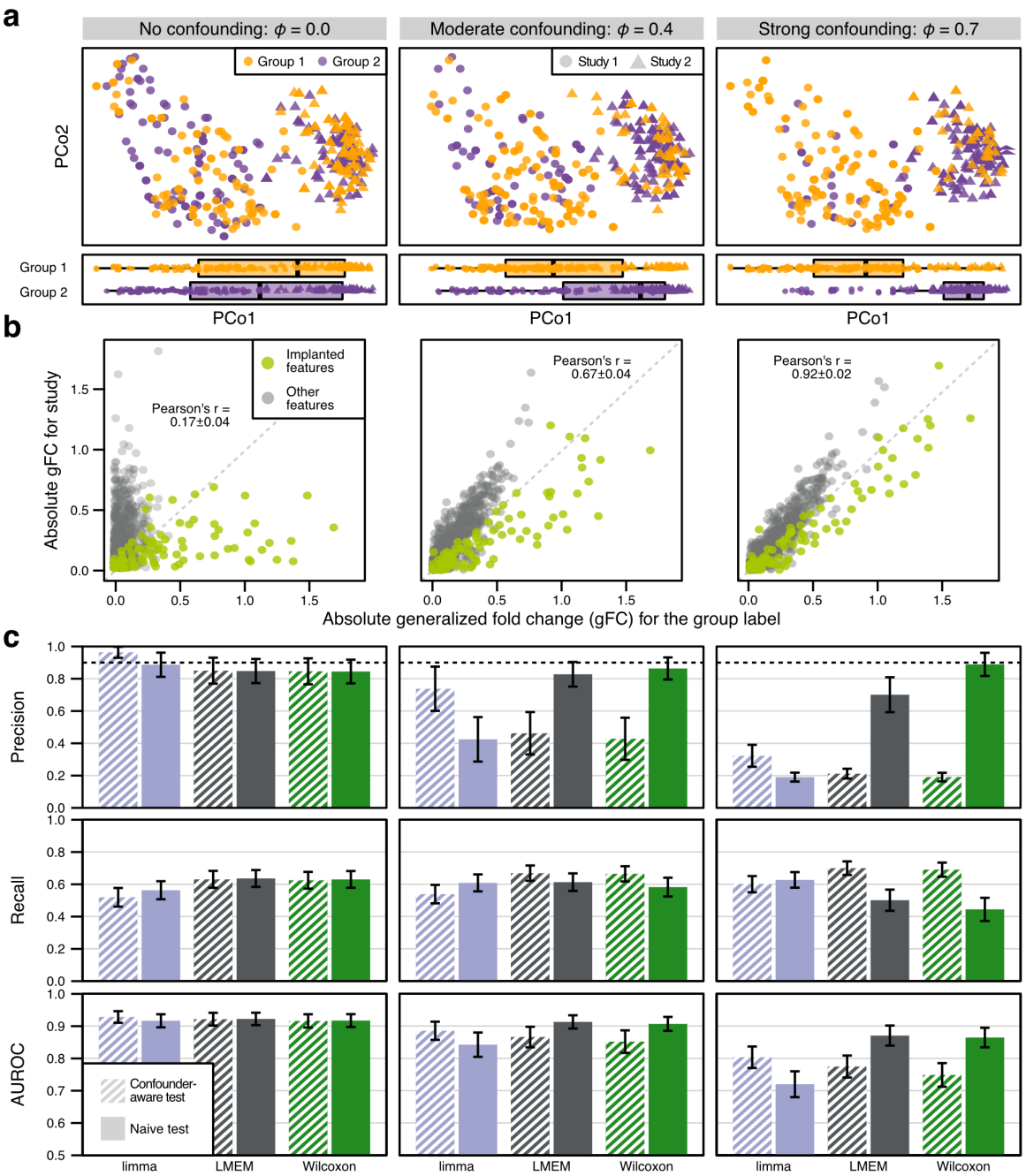


Figure 3: Broad confounding leads to a dramatic reduction of precision which can partly be alleviated by methods able to model confounders

a) Principal coordinate projections for simulated data (abundance scaling factor of 2, prevalence shift of 0.3, all features eligible for implantation, a single representative repeat shown) using two different studies as input. The difference between studies is most prominent in the first principal coordinate. Beneath the projections, the values of the first principal coordinate are shown across the two simulated groups. On the left, samples are selected for both groups proportionally from both studies, leading to a setup with minimal confounding. Towards the right side, the proportion of samples of Study 1 to be selected into group 1 increases, leading to stronger confounding affecting most bacterial taxa (broad confounding). **b)** Generalized fold change (gFC) calculated for the label is contrasted to the gFC calculated for differences between studies across all bacterial taxa for the same repeat as shown in a). Implanted features are highlighted in green. **c)** Mean precision, recall and AUROC for sample size 200 and the same effect sizes as shown in a) were computed for included DA methods, using either naive (hatched bars) and confounder-aware (solid bars) tests. Error bars indicate standard deviation around the mean for all repeats.

Under these narrowly confounded conditions mimicking biological effects, the results of the DA testing with *limma*, the *LM*, and the *Wilcoxon* test were generally similar compared to the results obtained from broad confounding conditions imitating study effects (see **Fig. 4b**). One difference was that the loss of FDR control was not as severe given narrow confounding (10-fold increase from the nominal FDR for naive tests under strong confounding versus 16-fold for broad confounding, or 50% versus 80%) and that the confounder-adjusted *limma* showed better precision (at a notable loss of sensitivity). Lastly, the FDR control of the confounder-adjusted *LM* was better under strongly confounded conditions, while the *LM* also exhibited the highest sensitivity for the detection of true DA features (see also **SFig. 9**). Overall, these results suggest that measured confounders can be effectively adjusted for in MWAS when explicitly modeled as such.

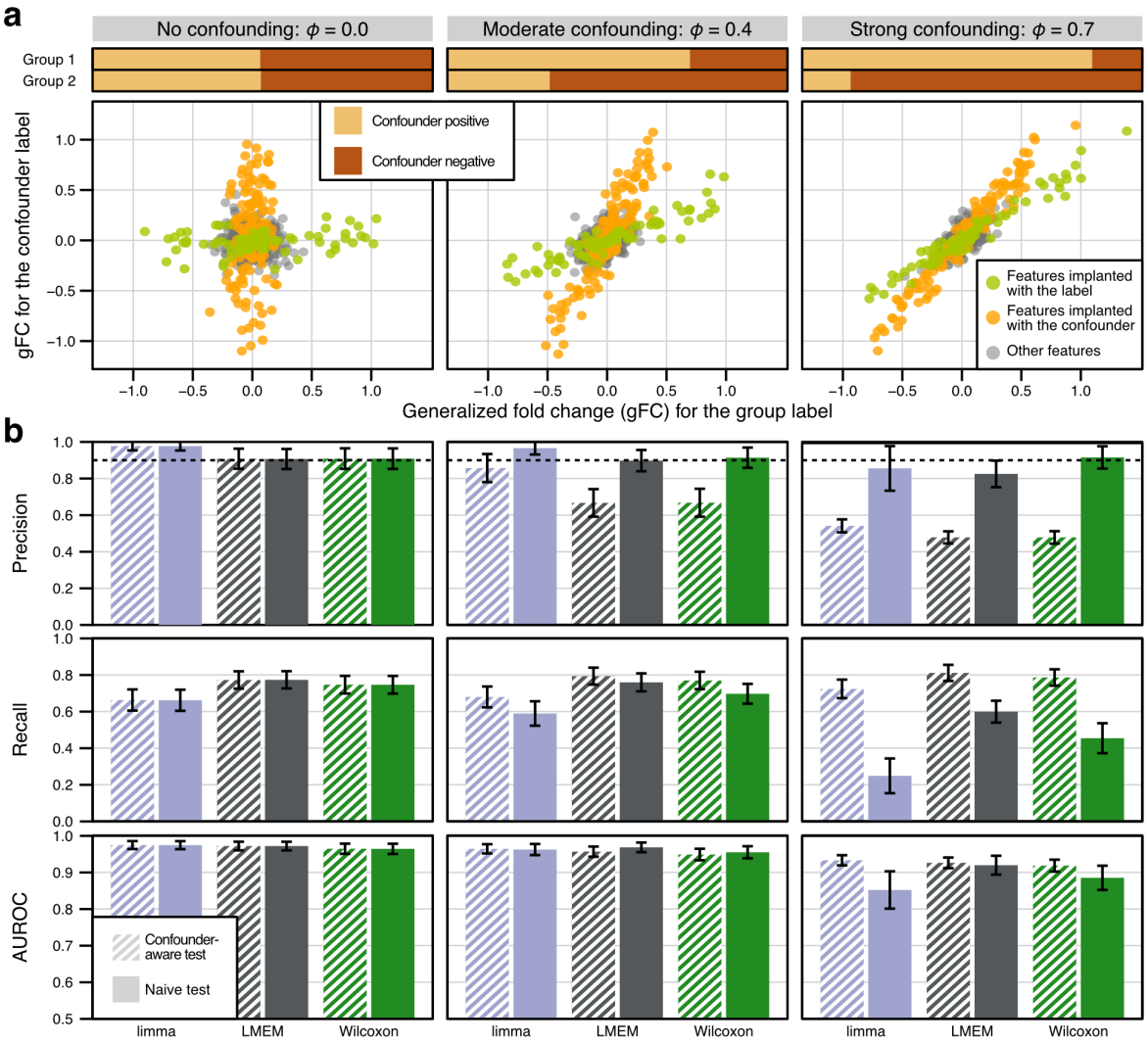


Figure 4: Loss of precision and recall under narrow confounding can be addressed by confounder-aware DA tests

Using a single dataset, DA features were implanted for both a main disease label (as described above) and additionally for an independent binary confounder label, imitating narrow effects (i.e. affecting a limited number of taxa) by a medication, for example. **a)** Generalized fold change (gFC) calculated for the label is contrasted to the gFC calculated for differences between confounder values across all bacterial taxa (abundance scaling factor of 2, prevalence shift of 0.3, all features eligible for implantation, a single representative repeat shown). The top bars visualize the confounder strength by showing the proportion of confounder-positive samples across both groups. Implanted features are highlighted in green and features implanted for the confounder label in orange. **b)**

Mean precision, recall and AUROC for sample size 200 and the same effect sizes as shown in a) were computed for included DA methods, using either naive (hatched bars) and confounder-aware (solid bars) tests. Error bars indicate standard deviation around the mean for all repeats.

Discerning robust from confounded associations in real datasets

To further explore the role of adjusting for confounders in real microbiome data, we applied naive and covariate-aware association tests to gut metagenomic samples from cardiometabolic diseases in the MetaCardis cohort[12,35,37]. The strongest confounding potential was seen for chronic coronary artery disease (CCAD) and commonly-indicated medications taken by a large fraction of these patients, especially statins and aspirin (ϕ=0.89 and ϕ=0.9, respectively), as well as type 2 diabetes (T2D) and metformin (ϕ=0.72, see **SFig. 7**). Four linear models were built for each disease-drug combination across all species-level taxonomic features, and the resulting coefficients and *P* values were used to classify each feature association with respect to both drug intake and disease status (see **Methods**). As expected, large phi coefficients manifested as a strong linear relationship between naive drug and disease associations with taxa (as in **Fig. 4a**). Inclusion of random effects in the adjusted models diminished this relationship and exposed drug- or disease-specificity in numerous individual taxon associations (**Fig. 5a**).

In contrast to CCAD (and other diseases), T2D exhibited larger, more significant naive associations with more taxa (**Fig. 5b**). A little less than half of these were confounded by metformin treatment, as identified by the loss of a significant association with T2D and retention of a significant association with metformin in the adjusted models for a given taxon (**Fig. 5b** and **Methods**). Adjusting for antibiotic intake, on the other hand, did not significantly reduce the number of T2D-associated taxa, but generally reduced their coefficient size and significance. CCAD-associated taxa were sensitive to adjustment by multiple different drugs, including antibiotics, better reflecting the complexity and variation typically seen in disease MWAS. In general, covariate-aware linear models reduced the number of significant disease associations by disentangling disease- and drug-associated taxa, similar to their ability to distinguish true positives from confounder positives in our confounded benchmarks.

Metformin and proton pump inhibitors (PPIs) were among the largest drug effects observed in our analysis. Whereas most metformin-associated taxa were also naively T2D-associated, most PPI-associated taxa were not. This is in line with previous reports characterizing metformin intake as a marker of T2D disease severity[7,35,38], and PPI-associated gut microbiota changes as a disease-independent “spike-in” of mainly oral commensals[39,40]. To further resolve disease-associated taxa, we cross-referenced the feature classifications we obtained from our models across multiple drugs as a robustness analysis. For example, PPI-adjusted models resulted in disease associations, several of which were metformin-confounded; however, integrating information across both covariate-aware models revealed a more robust subset of disease-associated taxa (**Fig. 5c**). The consistent number of background features across models also demonstrated that adjusting for covariates specifically reduced would-be spurious associations, especially in cases where confounding potential was high. Taken together, this analysis suggests that linear mixed-effect models are a flexible and scalable method to statistically improve robustness of findings in MWAS.

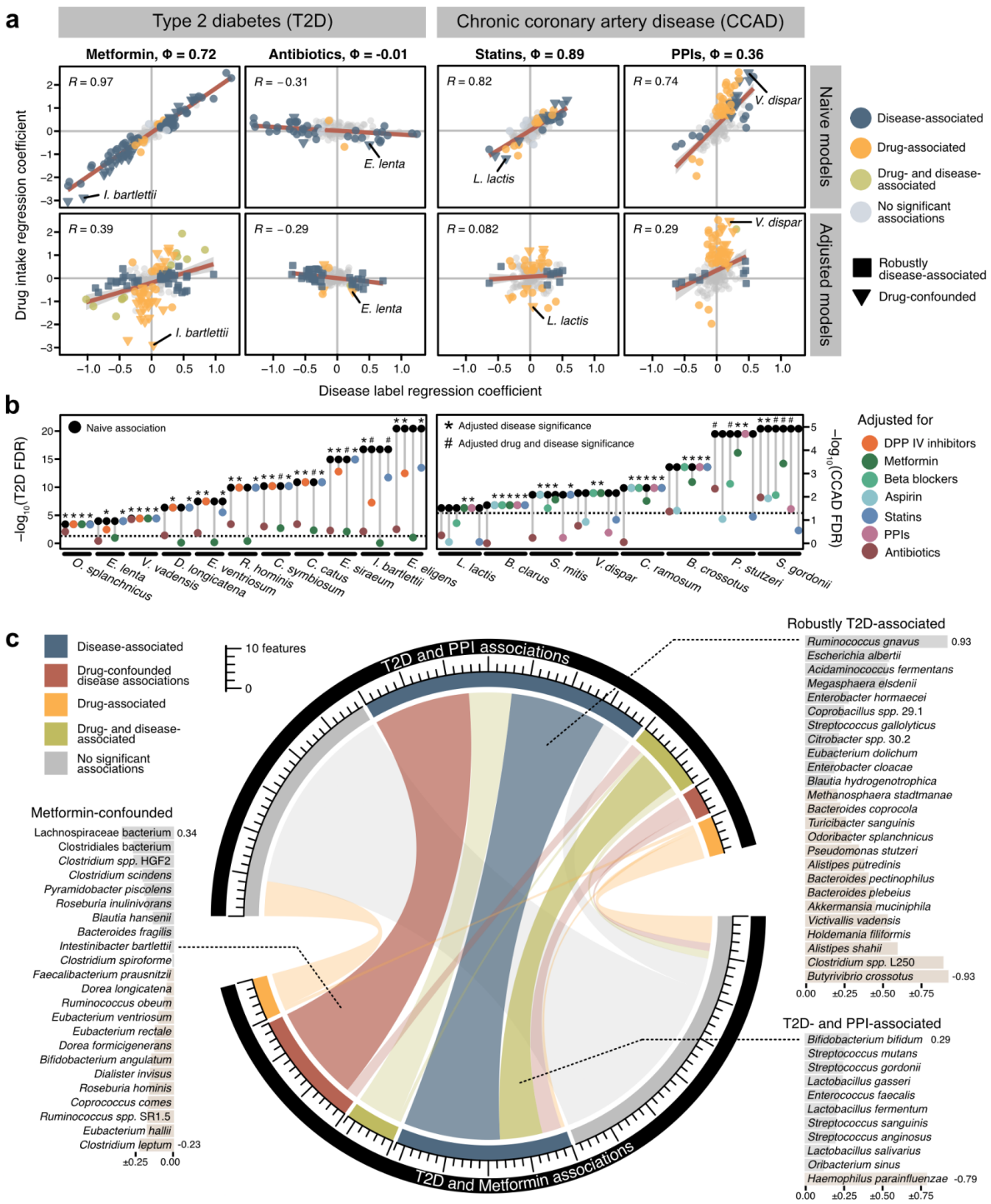


Figure 5: Linear models are capable of disentangling drug- and disease-associated microbial features
a) Regression coefficients from a subset of disease-drug combinations comparing naive linear models to mixed-effect models for all bacterial taxa. Adjusted models included a second term (either drug intake or disease status for x and y axes, respectively) as a random effect, which diminished the strong linear dependence between naive model coefficients (if present) and revealed drug-specific effects in some features. **b)** Subset of features displaying the largest number of significant disease associations across different drug-adjusted models (most robust, indicated with an asterisk or a hash) or displaying the largest reductions in disease coefficient significance upon adjustment (most confounded, below 5% FDR threshold line). **c)** Comparison of feature classifications (see **Methods**) from the metformin- and PPI-adjusted disease association models across all bacterial taxa. Integrating information across models restricts disease associations to a more robust subset and reveals drug-confounded associations. Adjusted T2D coefficients are shown in light grey or light brown bars behind species names (indicating enrichment in T2D or control group, respectively).

Discussion

Clinical interest in the microbiome has produced myriad studies which apply differential abundance tests to detect associations with host phenotypes, including many common diseases, or response to treatment[41]. To address this fundamental statistical task in MWAS[25], numerous DA methods have been developed encompassing a broad range of assumptions and hypotheses, which unsurprisingly vary in their performance when applied to real data[42]. We argue that previous benchmarks[19–24] have failed to produce consensus, since their parametric simulations were unvalidated and ultimately unrealistic. We addressed this by proposing a novel implantation framework for the generation of simulated taxonomic profiles based on minimal modifications to real metagenomic data, which we extended to incorporate effects resembling technical and biological confounders frequently encountered in MWAS. To our knowledge, our benchmark is the first to comprehensively evaluate DA methods using simulated data that has itself been evaluated on its resemblance to real clinical microbiome data. We empirically verified that our framework, unlike previously employed simulation methods based on parametric distributions, retains essential metagenomic data properties, most importantly feature sparsity and variance, and produces data that are virtually indistinguishable from real samples. Yet, our framework provides the flexibility to specify effect and sample sizes as well as confounder effects needed for an extensive evaluation against a ground truth, which was not the case in previous benchmarks built upon real datasets[23,24,42].

To resolve the question of which DA methods are best suited for gut microbiome data, we impartially benchmarked widely used DA methods using our feature implantation framework. Evaluating each DA test on nearly one million simulated data sets, we found that the majority of methods yielded an excess of false-positives (mean FDR exceeding twice the adjusted value at least once for 8/15 methods at a data set size between 50 and 200), which was generally worse at smaller sample sizes (N<100). Notable exceptions were the non-parametric *Wilcoxon* test, *limma*, and *LMs*, all of which were found to properly control the FDR while retaining high sensitivity across a range of sample and effect size. LEfSe[43], one of the most popular DA analysis packages for microbiome data, uses the Wilcoxon test as part of a more complicated procedure to estimate feature effect sizes (an evaluation of which is beyond the scope of this study). Our results strongly suggest that these DA methods should be preferred over the other tests evaluated here, including *ALDEx2*, *distinct*, and the *Kolmogorov-Smirnov* test, which did control the FDR, but were much less sensitive. DA methods borrowed from RNA-seq analysis, with the exception of *limma*, were among those with the highest FDRs. These methods were originally developed for few replicates with much lower dispersion than what is observed for fecal metagenomes of different human individuals. Our conclusion that these methods are unsuitable for microbiome data directly contradicts the results of a previous benchmark[20], a discrepancy that can be explained by the use of the multinomial simulation in that study, which strongly underestimated the variance of real microbiome data (see **Fig. 1b**). Surprisingly, most methods developed specifically for metagenomic data were found to have comparably low power and high false discovery rates (with the exception of *mgs2*) across the range of dataset sizes most commonly seen in recent MWAS (see **Fig. 2b**). On a positive note, more DA tests (including both *ANCOM* and *ANCOMBC*, *ZIBseq*, *ZINQ*, and *mgs2*) controlled the FDR at the nominal level when applied to larger samples (N>=200 per group). Overall, however, our findings

strongly suggest that many MWAS studies may have reported a substantial fraction of spurious microbiome-disease associations.

This issue is further exacerbated by confounding factors, for which awareness is growing in recent years[7,11,13,31,44]. While medication can be an obvious confounder for disease cohorts[35], one study based on the American Gut Project dataset[13] identified various lifestyle and physiological parameters, for example alcohol intake or bowel movement quality, as additional sources of heterogeneity. The proposed solution, matching for all potential confounders between groups, is however unattainable for most MWAS that are usually limited in their sample size. As a more straightforward alternative, we explored confounder-aware DA tests by extending our signal implantation framework to model both narrow (strong effects on a small number of taxa, imitating for example medication[35]) and broad (mimicking technical variation affecting the majority of taxa) confounding (see **Methods**). Reassuringly, inclusion of the confounder in the respective DA models mostly restored unconfounded performance. When explicitly adjusted for a measured covariate, the blocked Wilcoxon test, followed by linear mixed-effect models (LMEMs), most tightly controlled the FDR while retaining high power even under strong confounding; however, as it is limited to blocking a single discrete covariate, this test is less flexible than LMEMs, which can accommodate multiple covariates and be configured to handle nested or longitudinal study designs.

In our benchmark, we implemented simple LMEMs from the *lmerTest* R package[45]. Statistically analogous implementations to ours are available through MaAsLin2[46], which also includes several count-based linear models, and SIAMCAT[9], which provides not only confounder-aware DA tests but also dedicated functionality to check for confounders in the metadata. Similar to LEfSe, *metadeconfoundR*[35,47,48] employs non-parametric tests to first screen for naive associations, but goes a step further to construct multiple mixed-effect models (from the *lme4* package[49]) and apply iterative nested model testing procedures to further classify feature robustness or susceptibility to confounding. Analogous logic may be found in the vibration-of-effects paradigm[50,51]. In our analysis of drug confounding in T2D, we demonstrated the importance of integrating information across covariate-aware association models to reveal robust disease-associated microbial features. Given the limiting prerequisite that covariates need to be recorded for such approaches, methods that account for broadly-manifesting unmeasured confounders, as with population structure in genome-wide association studies[52,53], could help to minimize confounding variation in MWAS as well. As more attention is paid to the complexity of factors at play in clinical MWAS, the need for DA tools to not only consider potential confounders in their association models, but also to implement high-throughput robustness checks, is increasingly apparent.

In our view, the unsatisfactory performance of a wide range of DA methods and the persistent danger of unchecked confounding in MWAS warrant a community effort to develop and benchmark more robust methodology. To assist researchers in developing and validating new DA methods, or establishing benchmarks for other microbiota and study designs, both the signal implantation framework and the benchmarking analysis are designed to be easily extensible and available as open source code (see **Methods**). Ultimately, community-driven benchmarking efforts similar to DREAM challenges[54] or the critical assessment of metagenome interpretation (CAMI)[55] project could accelerate the much-needed consolidation of statistical methodology for microbiome research.

Methods

The codebase for the presented results is split into two projects. The first one, an R package called SIMBA (Simulation of Metagenomic data with Biological Accuracy), provides the modular functionality to i) simulate metagenomic data for a benchmarking project, ii) perform reality checks on the simulated data, iii) run differential abundance (DA) testing methods, and finally iv) evaluate the results of the tests. The second project, BAMBI (Benchmarking Analysis of MicroBiome Inference methods), is a collection of scripts that produce the presented analyses, consisting mostly of functions to automate and parallelize the execution of SIMBA functions relying on the *batchtools* package[56]. Both projects are available through Github and will enable other researchers to explore a similar benchmarking setting for other baseline datasets, other biomes, and additional DA testing methods. As part of the respective Github repositories, we included vignettes to showcase the functionality with toy examples. However, the simulation files and results of the statistical tests are available through Zenodo in order to provide reproducibility and direct comparison of new methods with the presented benchmark.

Data preprocessing

The dataset from Zeevi et al.[28] was used as a baseline for the simulations in most cases. Additionally, the TwinsUK dataset[36] was included in the broad confounder simulations (mimicking study effects). The MetaCardis dataset[35] was used to explore drug confounding in real data. Raw data were downloaded from ENA (PRJEB11532 for Zeevi and ERP010708 and TwinsUK) and profiled using the mOTUs2 profiler, version 2.5[57]. The resulting taxonomic profiles were filtered within SIMBA for prevalence (at least 5% in the complete dataset) and abundance (relative abundance of at least 1e-04). For the MetaCardis data, cell count adjusted, quantitative mOTUs profiles generated by the procedure outlined in Vandeputte et al.[58] were downloaded from an internal data hub. In the case of repeated samples per patient, SIMBA selects only the first time point for each patient.

Parametric methods for the simulation of metagenomic data

To simulate metagenomic data on the basis of parametric methods, the implementations employed in previous benchmarking efforts were adapted into SIMBA. Data were simulated under multinomial distributions using code from both McMurdie and Holmes[20] and Weiss et al.[22], since the functions to include differentially abundant features differed slightly between the two benchmarks. If not indicated otherwise, results for multinomial simulations were based on the implementation from Weiss et al., since the effect sizes were closer to real effects (see **SFig. 2**). The publication from Hawinkel et al.[23] included simulations based on the negative binomial, the beta binomial, and the Dirichlet distribution, which were likewise included in SIMBA. As in the original publication[23], correlations across bacterial taxa were estimated using SPIEC-EASI[17], since the correlation structure was needed for the beta binomial and could optionally be considered for the negative binomial simulations. Lastly, to simulate data as described in Ma et al.[27], SIMBA relies on the dedicated functions in the sparseDOSSA R package.

For each of the parametric simulation methods, the required parameters were estimated on the filtered Zeevi dataset. A dataset of equal size was simulated to include two different groups into which differentially abundant features were added as described in the respective original publications. For the multinomial simulations from McMurdie and Holmes as well as

for the sparseDOSSA approach, features were scaled in abundance after the simulation was completed. In the case of the other simulation methods, the underlying parameters were adjusted with a scaling factor before the simulation. A range of effect sizes (abundances scaled by multipliers of 1, 1.25, 1.5, 2, 5, 10, and 20) was explored and for each effect size, a total of 20 repetitions were simulated per simulation method. At an abundance scaling factor of 1, no effects were introduced into the data and therefore those repeats can serve as internal negative controls.

Signal implantation into real data

To create benchmarking datasets through signal implantation, differentially abundant features were implanted using the Zeevi dataset as a baseline. In each repetition, the original samples were randomly split into two groups, which served as the positive and negative groups. Differential abundance effects were implanted into a set of randomly selected features both via scaling abundances (same effect sizes as the parametric simulations) as well as by shifting prevalences (0.0, 0.1, 0.2, and 0.3).

For the abundance scaling, the count values in one group were multiplied with the scaling factor to increase the abundance. The prevalence shifts were implemented by identifying non-zero counts in one group and exchanging a specific percentage of those with occurrences of zero abundances in the other group (if possible), thereby creating a difference in prevalence across the groups. The implantation of signals alternated between the two groups in order to not introduce a systematic difference in total count number across groups (inspired by the considerations in Weiss et al.[22]). For each combination of effect sizes (abundance scaling and prevalence shift), 100 repetitions were simulated.

In each repetition, 10% of features were randomly selected for signal implantation. The set of features eligible to be selected could vary (see **SFig. 2**): *all* - all taxa were equally likely to be selected to carry a signal, *low* - only low abundance features (the 75th percentile across all samples not exceeding 0), *high* - only high abundance features (the median abundance across all samples higher than 0), *abundance* - the probability of a taxon to be selected is proportional to the mean abundance across all samples, and *inverse_abundance* - the probability of a taxon to be selected is inversely proportional to the mean abundance. Since the effect sizes from other schemes yielded unrealistic effect sizes, the downstream analyses were only carried out for the *all* and *low* implantations.

As a last step, the resulting generalized fold change[5] between the groups for all implanted features was recorded. Features with a fold change lower than 0.001 (resulting mostly from low-prevalence features being selected for implantation) were rejected and not recorded as implanted signals.

To generate simulations that mimic compositional effects, the signal implantation was carried out as described with the modification that signals were implanted into one group only (not alternating between groups). Then, the number of counts for each sample was scaled down to the original value of the unaltered sample by rarefaction using the *vegan* R package[59].

Reality assessment for simulated data

To determine how well a simulated metagenomic dataset approximated real data, several metrics were calculated by SIMBA. For each repetition of each simulation, sample sparsity and feature variance were recorded together with differences in prevalence and the generalized fold change[5] between groups. Additionally, the separation between original and simulated samples in principal coordinate space was evaluated using PERMANOVA as implemented in the *vegan* package[59]. As a complementary approach, a machine learning

model was trained to classify real and simulated samples using the SIAMCAT R package[9] and the AUROC of the cross-validated model was recorded.

Included DA testing methods and normalization procedures

To evaluate the performance of various DA testing methods, the R implementation of each method was incorporated into SIMBA using the recommended preprocessing, if applicable. The following methods were included in the benchmark (usually available through an R package of the same name): the *Wilcoxon* test, the Kolmogorov-Smirnov (*KS*) test and linear models (*LM*, all available within the base R distribution), *limma*[60], *edgeR*[61], *DESeq2*[62], *ALDEx2*[63], *metagenomeSeq*[19], *ZIBSeq*[64], *corncob*[65], *ZINQ*[66], *distinct*[67], *ANCOM*[68], and *ANCOMBC*[69]. For metagenomeSeq, two different models can be fitted within the same R package, which are included here as *mgs* (using the *fitZig* function) and *mgs2* (using the *fitFeatureModel* function), analogously to Weiss et al.[22]. For *ANCOM*, no dedicated R package is available from the original publication and the standard implementation is prohibitively slow, thus the implementation available through Lin et al.[70] was used (see **SFig. 10**).

All methods were run with different ways of normalizing the data: *pass* (no normalization) and *rarefaction* (rarefaction of counts to the 25th percentile of the total counts across samples) were applied to all methods. For the *Wilcoxon* test, the *KS* test, the *LM*, and *limma*, additional normalization methods were explored, namely *clr* (centered log ratio transform), *rclr* (robust centered log ratio transform), *TSS* (total sum scaling), *TSS.log* (total sum scaling, followed by log10 transformation of the data), and *TSS.arcsin* (total sum scaling, followed by the arcsine square root transformation), *rarefaction-TSS* (rarefaction followed by total sum scaling), *rarefaction-TSS.log* (rarefaction followed by total sum scaling and log10 transformation of the data). Lastly, the *ZIBSeq* DA test included an internal option for a *sqrt* normalization.

Benchmarking of DA testing methods at different sample sizes

To simulate different cohort (sample) sizes, SIMBA randomly selected *n* samples out of the two groups equally for each combination of effect size and each repetition. These samples were saved via their indices such that each method was applied to the exact same data. Seven different sample sizes were explored (12, 24, 50, 100, 200, 400, and 800) and 50 sets of test indices were created for each. For the evaluation of a single DA method, a total of 980,000 unique configurations were generated and used as input (7 abundance shifts x 4 prevalence shifts x 100 simulation repeats x 7 sample sizes x 50 repeats).

Each method was applied to each bacterial taxon in succession using the previously indexed samples. The *P* values across all taxa were recorded and adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure[29]. Since *ANCOM* does not return *P* values, its primary outputs (*W* values) were converted to be comparable to *P* values by transforming them to range between 0 and 1. The recommended decision threshold for significance in *ANCOM* is equal to 0.7 x number of tested taxa. Therefore, the *W* values above this decision threshold were transformed into 'significant' *P* values (lower than 0.05), whereas all other *W* values were transformed to range between 0.05 and 1 in the *P* value space. The ranking of the *W* values was conserved in this transformation.

To evaluate the performance of each method, SIMBA checked the *P* values from each testing scenario for how well bacterial taxa with differential abundance were detected. An AUROC was calculated with the *P* values as a predictor and the false discovery rate (FDR) was recorded with 0.05 serving as the decision threshold.

Generating confounders through biased resampling

To simulate the effects of a confounding variable, the signal implantation method was extended to include a secondary binary variable, similar to the main binary groups. In the narrow confounding simulations, an additional set of non-overlapping features (different than those implanted into the main binary groups) was implanted as described above, resulting in two orthogonal sets of ground truth features. Alternatively, for the broad confounding simulations which combined data from two different studies, this variable denoted the actual study membership (Zeevi[28] or TwinsUK[36]).

The confounding effect was generated when selecting samples for testing: in the different simulation repetitions, the probability of a sample to be selected for the positive group was contingent on the value of the secondary variable and a bias parameter. At a bias of 0.5, the probability for a sample to be selected for the positive group was not influenced by the secondary variable and the resulting groups were balanced (all four cells of the 2x2 contingency table have the same probability of 0.25). At a bias of 1, only the samples also bearing positive values for the secondary variable were selected for the positive group (increased probabilities along diagonal of 2x2 contingency table). Modulation of the bias parameter, and thereby the resulting class balance of samples selected with respect to both binary variables, enabled a systematic shift between the groups to be introduced into most taxa.

Confounder-aware DA testing

For the confounded benchmarking, only tests with a reasonable performance in the not-confounded setting were run through SIMBA, namely the *Wilcoxon* test, the *LM*, and *limma*. All tests could also be adjusted by the confounder covariate, usually by inclusion in the test formula. For the *Wilcoxon* test, confounder-aware testing was performed using the blocked *Wilcoxon* test implemented in the *coin* package[71] and for the *LM*, the confounder covariate was included as a random effect in the model formula. The significance of the original study variable was then tested by fitting the model using the *lmer* function within the *lmerTest* package[45]. The evaluation procedure was otherwise unchanged compared to the benchmarking without confounding.

Effect size assessment in real case-control datasets

To compare simulated data to real case-control microbiome studies, we collected datasets for two diseases with a well-described microbiome signal. For colorectal cancer (CRC), we included the data from five studies[5,72–75] across three continents, which were the basis for an earlier meta-analysis that identified consistent and predictive microbial biomarkers for CRC[5]. For Crohn's disease (CD), we similarly included five case-control studies[3,76–79] that had been analyzed previously[9]. For CD, the data were restricted to the first measurement for each individual, whenever applicable. The data from all studies were taxonomically profiled via mOTUs2 (version 2.5, ref[57]) and features were filtered for at least 5% prevalence in at least three of the studies. Differences in prevalence across groups and the generalized fold change were calculated for each microbial feature as previously described[5] and the significance of enrichment was calculated using the blocked *Wilcoxon* test from the *coin* package in R[71].

Confounder and robustness analysis in the MetaCardis data

To deepen our understanding of confounding variable effects found in real clinical microbiome data, we evaluated a subset of drug-disease combinations from the MetaCardis cohort, which were preprocessed as described above. Each disease subcohort was combined with the control group to constitute a case-control dataset, and the phi coefficient was calculated using a custom implementation of the standard formula[34] with respect to each binary drug intake metadata variable (see **SFig. 7a**). For each bacterial taxon, two naive linear models were built using the base R *lm* function which modeled bacterial abundance as a function of either disease status or drug intake only, and two corresponding covariate-aware models were built using the *lmer* function from the *lmerTest* package[45], and additionally incorporated drug intake or disease status as a random effect, respectively. Significances of the resulting coefficients were adjusted for multiple testing according to the Benjamini-Hochberg procedure[29] and used to classify associations.
For a given disease-drug combination, taxa which were significantly associated with the disease status and the drug intake in all four models were assigned a "drug- and disease-associated" status. Likewise, taxa bearing a significant association with drug intake in both naive and drug-adjusted models but lacking a significant disease association in the drug-adjusted model were considered to be "drug-confounded" if naively disease-associated, and "drug-associated" if not. Finally, taxa bearing a significant disease association in both naive and drug-adjusted models which did not possess a significant association with drug intake in the drug-adjusted model were classified as (robustly) "disease-associated".

Author contributions

G.Z. and S.K.F. conceived the study and supervised the work. J.W. and M.E. implemented the software and performed the statistical analyses with guidance from G.Z. and S.K.F.. J.W., M.E., and G.Z. designed the figures with input from S.K.F.. J.W., M.E., and G.Z. wrote the manuscript with contributions from S.K.F. All authors discussed and approved the final manuscript.

Code availability

The software package to simulate metagenomic data (SIMBA) is available on Github: https://github.de/zellerlab/SIMBA. Similarly, the repository containing the scripts to run a benchmark (BAMBI) is also available on Github: https://github.de/zellerlab/BAMBI.

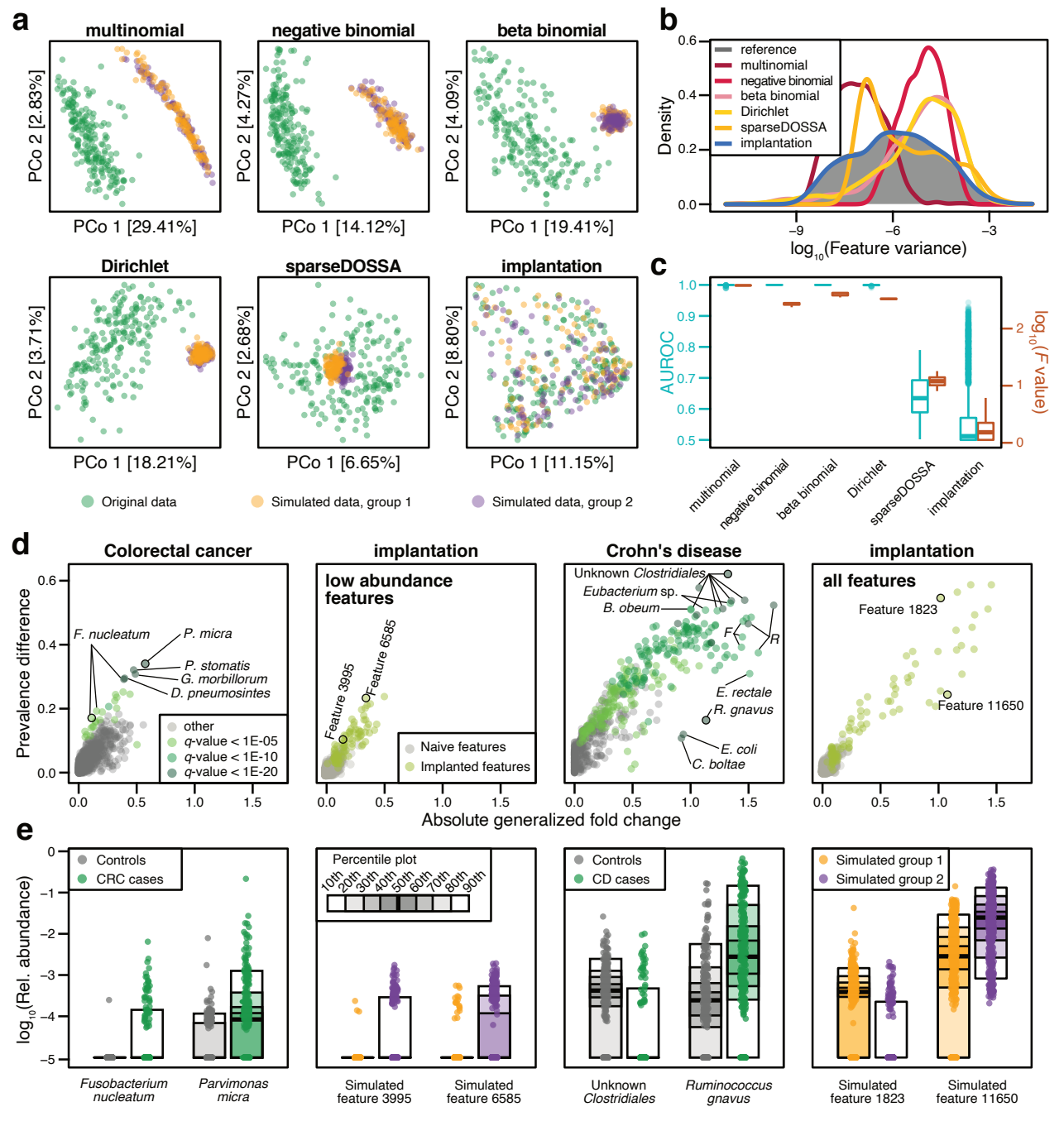
References

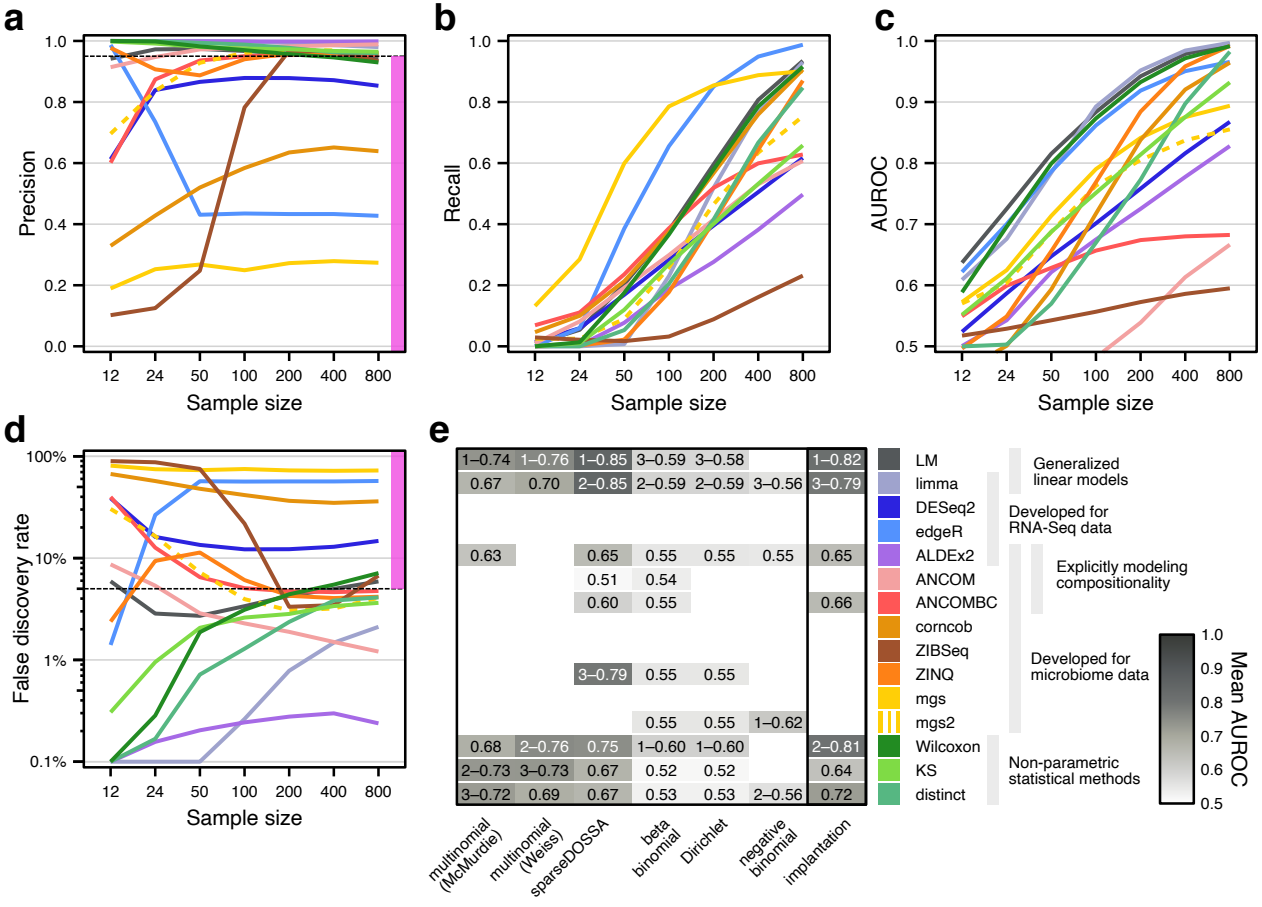
1. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16**, 73 (2015).
2. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
3. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293–305 (2019).
4. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
5. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
6. Li, J. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* **5**, 14 (2017).
7. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
8. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).
9. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 93 (2021).
10. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
11. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
12. Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**, 310–315 (2020).
13. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
14. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
15. Bartolomaeus, T. U. P. *et al.* Quantifying technical confounders in microbiome studies. *Cardiovasc. Res.* **117**, 863–875 (2021).
16. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* vol. 8 (2017).
17. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
18. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
19. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
20. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).
21. Thorsen, J., Brejnrod, A., Mortensen, M. & Rasmussen, M. A. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in …. *Microbiome* (2016).
22. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
23. 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 (2019).

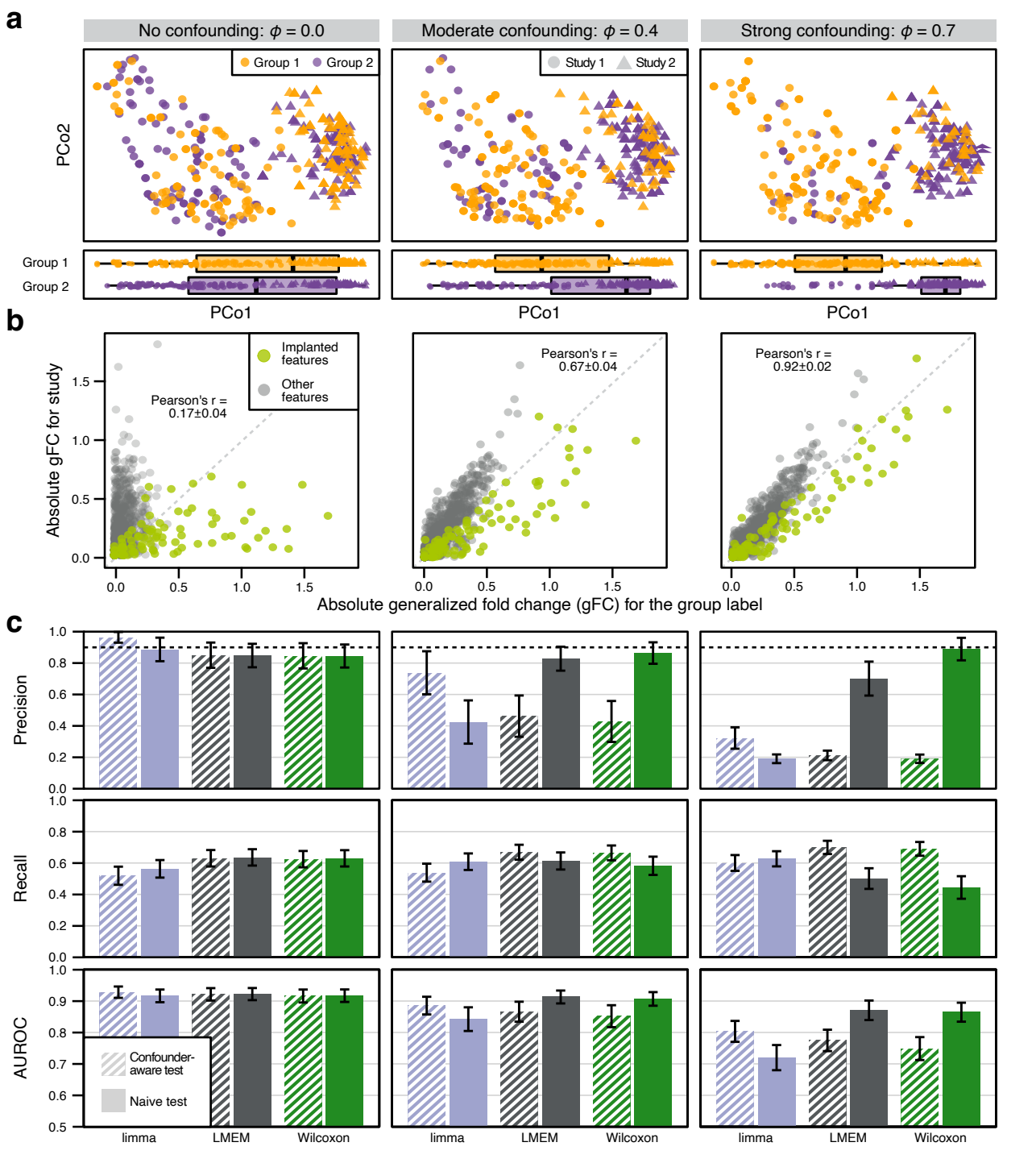
24. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol.* **21**, 191 (2020).
25. Kleine Bardenhorst, S. *et al.* Data Analysis Strategies for Microbiome Studies in Human Populations-a Systematic Review of Current Practice. *mSystems* **6**, (2021).
26. Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R. & Boulesteix, A.-L. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biol.* **22**, 152 (2021).
27. Ma, S., Ren, B., Mallick, H., Moon, Y. S. & Schwager, E. A Statistical Model for Describing and Simulating Microbial Community Profiles. *bioRxiv* (2021).
28. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
29. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
30. Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G. & Raes, J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* **12**, 3562 (2021).
31. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association to Modulation. *Cell* **172**, 1198–1215 (2018).
32. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
33. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nature Biotechnology* **35**, 1077–1086 (2017).
34. Guilford, J. P. The phi coefficient and chi square as indices of item validity. *Psychometrika* **6**, 11–19 (1941).
35. Forslund, S. K. *et al.* Combinatorial, additive and dose-dependent drug–microbiome associations. *Nature* **600**, 500–505 (2021).
36. Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst* **3**, 572–584.e3 (2016).
37. Molinaro, A. *et al.* Imidazole propionate is increased in diabetes and associated with dietary patterns and altered microbial ecology. *Nat. Commun.* **11**, 5881 (2020).
38. Wu, H. *et al.* The Gut Microbiota in Prediabetes and Diabetes: A Population-Based Cross-Sectional Study. *Cell Metab.* **32**, 379–390.e3 (2020).
39. Jackson, M. A. *et al.* Proton pump inhibitors alter the composition of the gut microbiota. *Gut* **65**, 749–756 (2016).
40. Vich Vila, A. *et al.* Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nat. Commun.* **11**, 362 (2020).
41. Lee, K. A. *et al.* Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nat. Med.* **28**, 535–544 (2022).
42. Nearing, J. T., Douglas, G. M., Hayes, M. G. & MacDonald, J. Microbiome differential abundance methods produce disturbingly different results across 38 datasets. *bioRxiv* (2021).
43. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
44. Forslund SK*, Chakaroun R*, Zimmermann-Kogadeeva M*, Markó L*, Aron-Wisnewsky J*, Nielsen T* et al. (in press) Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* (2021).
45. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles* **82**, 1–26 (2017).
46. Mallick, H., Rahnavard, A., McIver, L. J., Ma, S. & Zhang, Y. Multivariable association discovery in population-scale meta-omics studies. *Biorxiv* (2021).
47. TillBirkner. *TillBirkner/metadeconfoundR: MetadeconfoundR Release for Documentation of the MetaDrugs Analysis as Part of the MetaCardis Consortium*. (2021). doi:10.5281/zenodo.4721079.

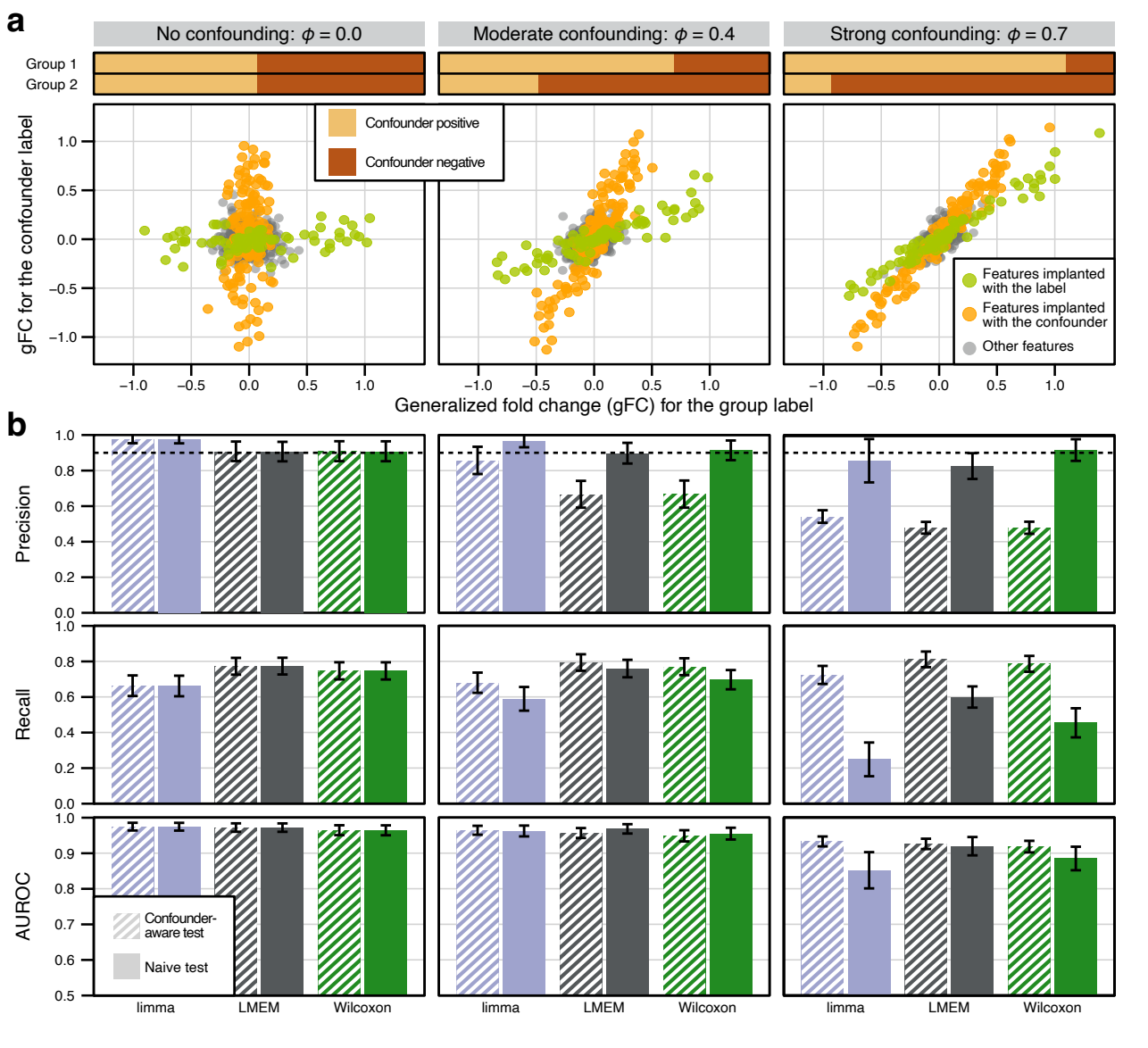
48. Fromentin, S. *et al.* Microbiome and metabolome features of the cardiometabolic disease spectrum. *Nat. Med.* **28**, 303–314 (2022).
49. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
50. Tierney, B. T. *et al.* Leveraging vibration of effects analysis for robust discovery in observational biomedical data science. *PLoS Biol.* **19**, e3001398 (2021).
51. Tierney, B. T. *et al.* Systematically assessing microbiome–disease associations identifies drivers of inconsistency in metagenomic research. *PLoS Biol.* **20**, e3001556 (2022).
52. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
53. Fusi, N., Stegle, O. & Lawrence, N. D. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.* **8**, e1002330 (2012).
54. Prill, R. J., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K. & Stolovitzky, G. Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.* **4**, mr7 (2011).
55. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
56. Lang, M., Bischl, B. & Surmann, D. batchtools: Tools for R to work on batch systems. *J. Open Source Softw.* **2**, 135 (2017).
57. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
58. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
59. Oksanen, J. *et al.* Package 'vegan'. *Community ecology package, version* **2**, 1–295 (2013).
60. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
61. 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 (2010).
62. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
63. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE* vol. 8 e67019 (2013).
64. Peng, X., Li, G. & Liu, Z. Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. *J. Comput. Biol.* **23**, 102–110 (2016).
65. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.* **14**, 94–115 (2020).
66. Ling, W. *et al.* Powerful and robust non-parametric association testing for microbiome data via a zero-inflated quantile approach (ZINQ). *Microbiome* **9**, 181 (2021).
67. Tiberi, S., Crowell, H. L., Samartsidis, P., Weber, L. M. & Robinson, M. D. Distinct: A novel approach to differential distribution analyses. *bioRxiv* (2020) doi:10.1101/2020.11.24.394213.
68. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
69. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
70. Lin, F. H. HuangLin/ANCOM: third release of ANCOM. *Zenodo10* **5281**, (2019).
71. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. A Lego System for Conditional Inference. *Am. Stat.* **60**, 257–263 (2006).
72. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
73. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma

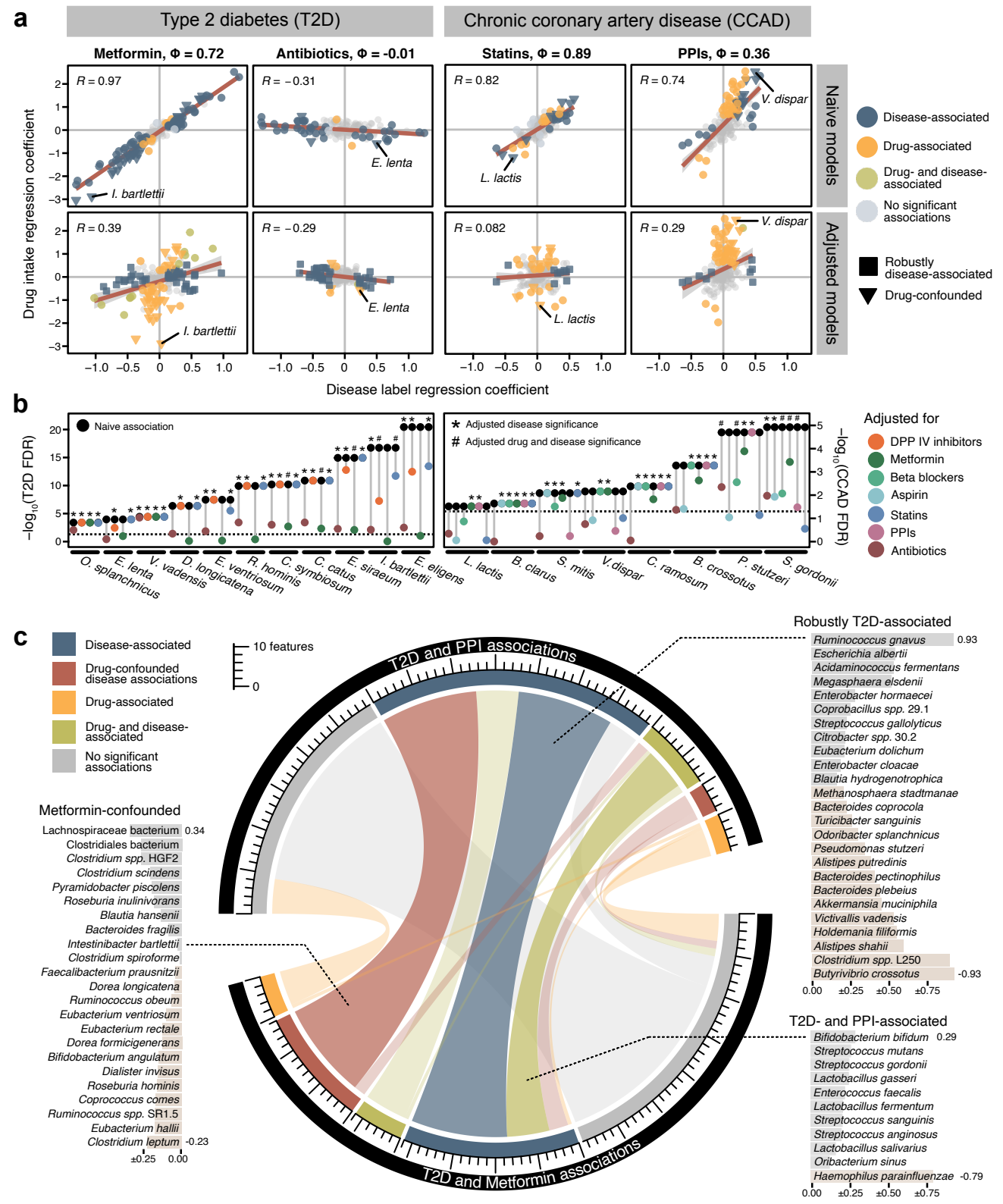
sequence. *Nat. Commun.* **6**, 6528 (2015).
74. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
75. Vogtmann, E. *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* **11**, e0155362 (2016).
76. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
77. Lewis, J. D. *et al.* Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489–500 (2015).
78. He, Q. *et al.* Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1–11 (2017).
79. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).


a
multinomial
negative binomial
beta binomial
Dirichlet
sparseDOSSA
implantation
PCo 2 [2.83%]
PCo 1 [29.41%]
PCo 2 [4.27%]
PCo 1 [14.12%]
PCo 2 [4.09%]
PCo 1 [19.41%]
PCo 2 [3.71%]
PCo 1 [18.21%]
PCo 2 [2.68%]
PCo 1 [6.65%]
PCo 2 [8.80%]
PCo 1 [11.15%]
Original data
Simulated data, group 1
Simulated data, group 2
b
reference
multinomial
negative binomial
beta binomial
Dirichlet
sparseDOSSA
implantation
Density
log10(Feature variance)
c
AUROC
log10(F-value)
multinomial
negative binomial
beta binomial
Dirichlet
sparseDOSSA
implantation
d
Colorectal cancer
implantation
Crohn's disease
implantation
Prevalence difference
F. nucleatum
P. micra
P. stomatis
G. morbillorum
D. pneumosintes
other
q-value < 1E-05
q-value < 1E-10
q-value < 1E-20
low abundance features
Feature 3995
Feature 6585
Naive features
Implanted features
Unknown Clostridiales
Eubacterium sp.
B. obeum
F
R
E. rectale
R. gnavus
E. coli
C. boltae
all features
Feature 1823
Feature 11650
Absolute generalized fold change
e
log10(Rel. abundance)
Controls
CRC cases
Fusobacterium nucleatum
Parvimonas micra
Percentile plot
10th
20th
30th
40th
50th
60th
70th
80th
90th
Simulated feature 3995
Simulated feature 6585
Controls
CD cases
Unknown Clostridiales
Ruminococcus gnavus
Simulated group 1
Simulated group 2
Simulated feature 1823
Simulated feature 11650


a
Precision
Sample size
b
Recall
Sample size
c
AUROC
Sample size
d
False discovery rate
Sample size
e
multinomial (McMurdie)
multinomial (Weiss)
sparseDOSSA
beta binomial
Dirichlet
negative binomial
implantation
LM
limma
DESeq2
edgeR
ALDEx2
ANCOM
ANCOMBC
corncob
ZIBSeq
ZINQ
mgs
mgs2
Wilcoxon
KS
distinct
Generalized linear models
Developed for RNA-Seq data
Explicitly modeling compositionality
Developed for microbiome data
Non-parametric statistical methods
Mean AUROC


a
No confounding: ϕ = 0.0
Moderate confounding: ϕ = 0.4
Strong confounding: ϕ = 0.7
Group 1
Group 2
Study 1
Study 2
PCo2
PCo1
b
Implanted features
Other features
Pearson's r = 0.17±0.04
Pearson's r = 0.67±0.04
Pearson's r = 0.92±0.02
Absolute gFC for study
Absolute generalized fold change (gFC) for the group label
c
Precision
Recall
AUROC
Confounder-aware test
Naive test
limma
LMEM
Wilcoxon


a
No confounding: $\phi = 0.0$
Moderate confounding: $\phi = 0.4$
Strong confounding: $\phi = 0.7$
Group 1
Group 2
Confounder positive
Confounder negative
gFC for the confounder label
Features implanted with the label
Features implanted with the confounder
Other features
Generalized fold change (gFC) for the group label
b
Precision
Recall
AUROC
Confounder-aware test
Naive test
limma
LMEM
Wilcoxon


a
Type 2 diabetes (T2D)
Chronic coronary artery disease (CCAD)
Metformin, Φ = 0.72
Antibiotics, Φ = -0.01
Statins, Φ = 0.89
PPIs, Φ = 0.36
R = 0.97
R = −0.31
R = 0.82
R = 0.74
R = 0.39
R = −0.29
R = 0.082
R = 0.29
I. bartlettii
E. lenta
L. lactis
V. dispar
Naive models
Adjusted models
Drug intake regression coefficient
Disease label regression coefficient
Disease-associated
Drug-associated
Drug- and disease-associated
No significant associations
Robustly disease-associated
Drug-confounded
b
Naive association
* Adjusted disease significance
# Adjusted drug and disease significance
−log10(T2D FDR)
−log10(CCAD FDR)
Adjusted for
DPP IV inhibitors
Metformin
Beta blockers
Aspirin
Statins
PPIs
Antibiotics
O. splanchnicus
E. lenta
V. vadensis
D. longicatena
E. ventriosum
R. hominis
C. symbiosum
C. catus
E. siraeum
I. bartlettii
E. eligens
L. lactis
B. clarus
S. mitis
V. dispar
C. ramosum
B. crossotus
P. stutzeri
S. gordonii
c
Disease-associated
Drug-confounded disease associations
Drug-associated
Drug- and disease-associated
No significant associations
10 features
0
T2D and PPI associations
T2D and Metformin associations
Robustly T2D-associated
Ruminococcus gnavus 0.93
Escherichia albertii
Acidaminococcus fermentans
Megasphaera elsdenii
Enterobacter hormaecei
Coprobacillus spp. 29.1
Streptococcus gallolyticus
Citrobacter spp. 30.2
Eubacterium dolichum
Enterobacter cloacae
Blautia hydrogenotrophica
Methanosphaera stadtmanae
Bacteroides coprocola
Turicibacter sanguinis
Odoribacter splanchnicus
Pseudomonas stutzeri
Alistipes putredinis
Bacteroides pectinophilus
Bacteroides plebeius
Akkermansia muciniphila
Victivallis vadensis
Holdemania filiformis
Alistipes shahii
Clostridium spp. L250
Butyrivibrio crossotus -0.93
0.00 ±0.25 ±0.50 ±0.75
Metformin-confounded
Lachnospiraceae bacterium 0.34
Clostridiales bacterium
Clostridium spp. HGF2
Clostridium scindens
Pyramidobacter piscolens
Roseburia inulinivorans
Blautia hansenii
Bacteroides fragilis
Intestinibacter bartlettii
Clostridium spiroforme
Faecalibacterium prausnitzii
Dorea longicatena
Ruminococcus obeum
Eubacterium ventriosum
Eubacterium rectale
Dorea formicigenerans
Bifidobacterium angulatum
Dialister invisus
Roseburia hominis
Coprococcus comes
Ruminococcus spp. SR1.5
Eubacterium hallii
Clostridium leptum -0.23
±0.25 0.00
T2D- and PPI-associated
Bifidobacterium bifidum 0.29
Streptococcus mutans
Streptococcus gordonii
Lactobacillus gasseri
Enterococcus faecalis
Lactobacillus fermentum
Streptococcus sanguinis
Streptococcus anginosus
Lactobacillus salivarius
Oribacterium sinus
Haemophilus parainfluenzae -0.79
0.00 ±0.25 ±0.50 ±0.75