A broken promise: microbiome differential abundance methods do not control the false discovery rate

peer-reviewed author version

Hawinkel, Stijn; Mattiello, Federico; BIJNENS, Luc & THAS, Olivier (2019) A broken promise: microbiome differential abundance methods do not control the false discovery rate. In: BRIEFINGS IN BIOINFORMATICS, 20(1), p. 210-221.

DOI: 10.1093/bib/bbx104
Handle: http://hdl.handle.net/1942/28963

A broken promise: microbiome differential abundance methods do not control the false discovery rate

Stijn Hawinkel[1], Federico Mattiello[1], Luc Bijnens[2,3] and Olivier Thas[1,4]

[1]*Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University,* [2]*Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson,* [3]*Center for Statistics, Hasselt University,* [4] *National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia*

Corresponding author: Stijn Hawinkel (stijn.hawinkel@ugent.be), BioStat group, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University

Key words: Microbiome; differential abundance; simulation; taxa correlation networks; false discovery rate

Stijn Hawinkel is a PhD student in biostatistics in the BioStat group, Department of Mathematical Modelling, Statistics and Bioinformatics at Ghent University, Belgium.
Federico Mattiello is a former PhD student in biostatistics in the BioStat group, Department of Mathematical Modelling, Statistics and Bioinformatics at Ghent University, Belgium, now working at Roche, Basel, Switzerland.
Luc Bijnens is a senior director at Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson and a professor of statistics at the Center for Statistics at Hasselt University, Belgium.
Olivier Thas is a professor of biostatistics at the Department of Mathematical Modelling, Statistics and Bioinformatics at Ghent University, Belgium, and an honorary professor at the National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia.

Abstract

High-throughput sequencing technologies allow easy characterization of the human microbiome, but the statistical methods to analyze microbiome data are still in their infancy. Differential abundance methods aim at detecting associations between the abundances of bacterial species and subject grouping factors. The results of such methods are important to identify the microbiome as a prognostic or diagnostic biomarker or to demonstrate efficacy of pro- or antibiotic drugs. Because of a lack of benchmarking studies in the microbiome field, no consensus exists on the performance of the statistical methods. We have compared a large number of popular methods through extensive parametric and non-parametric simulation as well as real data shuffling algorithms. The results are consistent over the different approaches and all point to an alarming excess of false discoveries. This raises great doubts about the reliability of discoveries in past studies and imperils reproducibility of microbiome experiments. To further improve method benchmarking we introduce a new simulation tool that allows to generate correlated count data following any univariate count distribution; the correlation structure may be inferred from real data. Most simulation studies discard the correlation between species, but our results indicate that this correlation can negatively affect the performance of statistical methods.

Key words: Microbiome; differential abundance; simulation; taxa correlation networks; false discovery rate

Stijn Hawinkel is a PhD student in biostatistics in the BioStat group, Department of Mathematical Modelling, Statistics and Bioinformatics at Ghent University, Belgium.

Federico Mattiello is a former PhD student in biostatistics in the BioStat group, Department of Mathematical Modelling, Statistics and Bioinformatics at Ghent University, Belgium, now working at Roche, Basel, Switzerland.

Luc Bijnens is a senior director at Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson and a professor of statistics at the Center for Statistics at Hasselt University, Belgium.

Olivier Thas is a professor of biostatistics at the Department of Mathematical Modelling, Statistics and Bioinformatics at Ghent University, Belgium, and an honorary professor at the National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia.

Introduction

The human microbiome is known to contribute to key body functions such as food digestion [1], resistance to infection [2], maturation of the immune system [3, 4] and anatomic development [2]. On the other hand, (local) disturbances of the microbiome are associated with disease statuses such as gut inflammation [2], vaginosis [5], diabetes [6, 7] and periodontal disease [8]. The microbiome is usually characterized by sequencing one specific marker gene (usually the 16S rRNA gene), clustering these reads into consensus sequences, which are referred to as Operational Taxonomic Units (OTUs), and mapping them to a reference database to link the sequences to the microbial species. This results in a count table listing the number of times each species was detected per sample [9, 10]. In microbiome studies the OTU level is the lowest level of microbial identification, but often data analysis focuses on a higher taxonomic level. For this reason, we will often use the term *taxon* to refer to the target of the differential abundance analysis.

Development of biomarkers based on the mi-

crobiome composition, as well as evaluation of microbiome-targeting drugs, require singling out of the specific taxa that are most strongly associated with a certain grouping factor of interest (e.g. disease status). This problem was coined the detection of "differential abundance" [11, 12], after the concept of differential *expression* in genomics. With the current sequencing technologies, the total number of reads per sample is a technical artefact, unrelated to its biological composition. As a result these technologies do not allow inference on absolute abundances of the taxa in the sample, and differential abundance is usually defined as a difference in mean *relative* abundances (taxon count relative to total count of all taxa in the sample) between groups [12]. However, an increase of some taxa's relative abundances in response to a physiological change, automatically results in changes in the relative abundances of all other taxa. To tackle this so-called "compositionality", robust normalization techniques [11, 13, 14, 15, 16] as well as methods based on ratios rather than differences of relative abundances (*ANCOM* [17] and *ALDEx*2 [18]), have been developed. They attempt to only detect the minority of the taxa that initially responded to the physiological change, although no clear mathematical definition exists that discriminates between this group and the taxa that undergo changes in relative abundances as a consequence of changes in other taxa.

Some differential abundance methods for microbiome data have been developed [10, 11, 17, 18], but they are not widely used and no consensus exists on their performance due to lack of benchmarking studies in the microbiome field. In practice, simple statistical tests such as two-sample t-tests [19] and Wilcoxon rank sum tests [20, 21] are frequently employed. It has recently been suggested that methods developed for RNA-Seq data could easily be adapted to microbiome data, since both data types are in essence read count data mapped to a reference database [12].

The differentially abundant taxa are usually of great scientific interest. These are taxa that can serve as a disease biomarker or that respond to a treatment. For this reason, sensitive methods are needed for identifying as many of these differentially abundant taxa as possible. However, microbiomes typically consist of hundreds to thousands of different taxa, of which only a minority is expected to be differentially abundant. It is therefore equally important to restrict the number of taxa that are falsely reported as differentially abundant. Suppose that in a microbiome study 1000 taxa have to be tested for differential abundance, for which 100 taxa are truly differentially abundant. For a method that has a sensitivity of 60% and a specificity equal to 95.5%, we expect approximately $100 * 0.6 + 0.045 * 900 \approx 100$ taxa to be declared differentially abundant: 40 false discoveries and 60 true discoveries. In terms of sensitivity/specificity trade-off this particular method appears to perform well. However, this is not relevant with respect to how researchers and the scientific community work with the results of this study. Typically, only the discoveries (significant taxa) are reported in the literature and only these taxa are considered in follow-up studies. However, 40% of the taxa reported in this study are not differentially abundant in reality.

For this reason, researchers often choose for controlling the false discovery rate (FDR), which is the expected fraction of non-differentially abundant taxa among the taxa called significant (the discoveries) [22]. An FDR method typically consists in adjusting the individual p-values and calling taxa differentially abundant if their adjusted p-values fall below a certain threshold. The correction method and the threshold are constructed such that the FDR equals a user specified level, which is referred to as the nominal FDR level. The FDR is thus a measure of the reliability of the discoveries and as such it is of primary importance that the true FDR does not exceed the nominal FDR by much, otherwise the reproducibility of the study is jeopardized. An excess of false discoveries may lead to more costly but fruitless follow up studies on

taxa that are actually not differentially abundant.

One strategy for evaluating methods for differential abundance detection is to reanalyze a sample with a golden standard method (e.g. rt-qPCR or microarray) [23, 24]. Such golden standard method is typically not available and comparative studies therefore rely predominantly on simulations or methods involving randomly shuffling real datasets.

In *parametric simulations*, data are repeatedly generated from a particular parametric distribution with known parameter values. This approach has the advantage of being very tractable and readily allows evaluation of the performance of differential abundance methods because the truth, which is completely described by the underlying distributions with their parameters, is known. However, the distributional assumptions may be wrong and unrealistic. Moreover, mostly only the marginal univariate count distributions of the taxa are specified, ignoring possible correlations between species' abundances (save for one exception [16]). Simulating taxon-by-taxon implies the assumption of independence between counts of different taxa, which biologically makes little sense. In reality, bacteria living in the same niche do not grow independently. Generating data with a correct correlation structure matters, because even though statistical testing for differential abundance typically occurs taxon-by-taxon, the estimation of normalization factors [11, 13, 15, 16, 25] and variance [13, 25, 26], as well as the subsequent FDR multiplicity correction [22, 27, 28] may work on the ensemble of taxa. As a result the correlation structure may affect the behaviour of the statistical testing procedure, and hence it is important not only to specify the marginal univariate count distributions, but also the correlation between taxa so as to achieve a correct joint distribution of the abundances. This issue has been largely ignored in the literature. Many studies that propose new methods present results from parametric simulation under the independence assumption, and often the simulations use the same distribution as the one their new method is built upon. As a consequence, it may be expected that many of the simulation results reported in the literature are too optimistic and show a bias towards the preference of the new methods proposed in the respective papers.

Non-parametric simulation, which has recently been used for RNA-Seq data [24, 29, 30], resamples or modifies counts of real datasets for constructing synthetic datasets. As a consequence, it incorporates realistic levels of noise and retains realistic marginal distributions and correlation structures, resulting in realistic joint distributions. Hence these methods do not rely on simple assumptions about the marginal distributions [24, 31, 32], and thus reduce the risk of evaluating methods based on misspecfied simulation models. On the other hand, when starting from real data it is not possible to control most of the parameter settings.

Finally, a third stream of simulation methods, which are here referred to as *real data shuffling*, encompasses all methods that leave the original counts intact and only manipulate a grouping factor by repeatedly randomly permuting the grouping factor labels. This approach yields realistic data, albeit only under the null hypothesis that none of the taxa are differentially abundant, and it requires strong assumptions on sample homogeneity to evaluate method performance. See figure 1 for an overview of data generation paradigms used.

We performed a large scale simulation study using parametric and non-parametric simulation and real data shuffling for evaluating a wide range of methods from the microbiome and the RNA-Seq world. To mimic various realistic data structures as closely as possible, the simulations were based on real human microbiome datasets from several body sites as templates [20, 33, 34]. These templates were combined with combinations of popular normalization, differential abundance testing and multiplicity correction strategies and with a range of relevant sample sizes and fold changes. In this way we cover a broader range of scenarios for microbiome differential abundance testing than in

previous comparative studies [11, 12, 32, 35, 36]. Moreover, some of our simulation techniques result in more realistic evaluations than was presented in earlier work. The most important findings are presented in the main text of this paper, and the more extensive results are provided in the Supplementary Material.

The results are surprisingly consistent over the different approaches and all raise an alarming concern about the complete failure of the methods to control the FDR at the nominal level. This is of direct importance to past and present microbiome research, because failure to control the FDR gives researchers a false sense of certainty about the reliability of their discoveries. Similar tendencies were seen in previous studies for some RNA-Seq methods [17, 29, 30, 31, 35, 37, 38], but the violations were not as severe and consistent as the ones we found in the microbiome setting.

To tackle the problem of unrealistic correlation structures in parametric simulation, we proposed a simulation technique that allows to specify the correlation structure in combination with any desired marginal univariate count distribution for generating count datasets. The mimicry of realistic correlation structures is made possible through recent advances in the estimation of correlation networks of microbiome sequencing data [39]. Our results reveal that the existence of a correlation structure can negatively affect the performance of the statistical methods evaluated in this paper.

Materials and methods

Computations were run on a Dell laptop, on two servers with 12 respectively 30 cores and on the high performance computing facilities of VSC (the Flemish Supercomputer Center). All analyses were implemented and run with the R programming language versions 3.2.3, 3.3.0 and 3.3.1. R-packages (and version numbers) used were *parallel* (3.2.3, 3.3.0 and 3.3.1), *phyloseq* (1.12.2, 1.14.0 and 1.16.0), *HMP* (1.4.3), *MASS* (7.3.45), *SpiecEasi* (0.1), *TailRank* (3.1.0), *fdrtool* (1.2.15), *SimSeq* (1.4.0), *edgeR* (3.10.5, 3.12.1 and 3.14.0), *DESeq*2 (1.8.2, 1.10.0 and 1.12.4), *ggplot*2 (2.1.0), *metagenomeSeq* (1.12.1 and 1.14.2), *plyr* (1.8.4), *reshape*2 (1.4.1), *AUC* (0.3.0), *samr* (2.0), *ALDEx*2 (1.0.0, 1.2.0 and 1.4.0), *biom* (0.3.12), *rhdf*5 (2.16.0), *Biostrings* (2.40.0), *rmarkdown* (1.0), *knitr* (1.14), *psych* (1.6.9), *VGAM* (1.0.2), *grid* (3.3.1), *Hmisc* (4.0.1) and *bigmemory* (4.5.19). All R-code is available at `http://users.ugent.be/~shawinke/ABrokenPromise/`.

Datasets

Data were used from the Human Microbiome Project (HMP, V13 region of the 16S rRNA gene) [33], the American Gut Project (AGP) [34] and a study on the use of the colorectal microbiome as a biomarker for cancer, referred to as the Zeller data [20]. Only the metagenomics sequencing data of the Zeller data were used in this paper. For simulation purposes, only the 1000 most abundant taxa (by summing each taxon's counts over the samples) were used. Prior to any analysis, taxa with a prevalence lower than 5% (i.e. taxa with a non-zero count in less than 5% of the samples) were trimmed.

Parametric simulation

Distributions and parameters

In general terms the parametric simulation of abundances of taxa required the specification of the joint distribution of the abundances of these taxa. Three methods for the construction of a joint distribution are considered:

1. **Direct specification of a joint multivariate distribution**: The Dirichlet-multinomial is a multivariate distribution and thus characterizes the marginal distributions (beta binomial) as well as the correlation structure. It gives a constant overdispersion parameter to all marginal beta binomial distributions and small negative correlations between all taxa [40].

2. **Specification of the marginal univariate distributions and assuming independence between the taxa**: The negative binomial distribution was used to generate counts for each taxon separately.

3. **Separate specification of the marginal univariate distributions and the correlation structure between the taxa**: The negative binomial and beta binomial were considered as marginal distributions, but now a realistic correlation structure between the taxa was imposed. This new simulation method starts from correlation networks estimated from real datasets using the *SpiecEasi* R-package [39] (see figure 3, and sections 4 and 3 in the Supplementary Material for details).

Parameter values

All parameters for parametric simulations were estimated from the AGP and HMP template datasets. From the HMP data only three body sites were used: the tongue dorsum, stool and mid vagina. The parameters of the negative binomial distribution were estimated by maximum likelihood, and those of the Dirichlet-multinomial by the method of moments. The Dirichlet multinomial distributions imposes beta-binomial distributions on its margins with one common dispersion parameter [40]. Therefore the parameter values (including the common dispersion parameter) estimated for the Dirichlet multinomial model were also used to generate the beta binomial data. As a result of this choice the data generated with the beta-binomial (and the Dirichlet multinomial) have larger variance and larger frequency of zeroes than under the negative binomial, which models the observed zero frequency and variance more accurately (see section 2 in the Supplementary Material). The parameter values used for the parametric simulation were randomly sampled from the pool of estimated parameter values; this was repeated in each simulation run. For the negative binomial distribution, the mean and overdispersion parameters were sampled from the same taxon so as to preserve any mean-dispersion relationship. The total numbers of counts per sample, also known as library sizes, were sampled with replacement from the observed library sizes of the respective template datasets. Sample sizes of 5, 25 and 100 samples per group were considered for the simulation study.

Introduction of differential abundance

To create differentially abundant taxa, we pursued two different strategies. In the first approach (referred to as "without compensation") the mean relative abundance of a randomly sampled fraction of 10% of the taxa was multiplied with a given fold change in one of the groups, and the ensemble of mean relative abundances was renormalized to sum to 1 prior to random number generation. This corresponds to the scenario for which the absolute abundance of a small group of taxa responds to a physiological change. Even though this procedure modifies the mean relative abundances of all taxa, a microbiologist would only want to detect the small group that initially reacted to the physiological change. For this reason significant results for other taxa will be considered as false discoveries. In a second approach (referred to as "with compensation"), the mean relative abundance of 10% of the taxa was changed such a way that it did not affect the relative abundances of the remaining 90%. In particular, to introduce a fold change FC, the mean relative abundances of a fraction $\frac{1}{FC+1}$ of the 10% (with relative abundances summing to a) were multiplied by FC, the remaining $\frac{FC}{FC+1}$ (with relative abundances summing to b) had its mean relative abundance multiplied by $\frac{a}{b}(1 - FC) + 1$ so that the total of all mean relative abundances equaled 1 again. The fold changes were set to 1 (no differential abundance), 1.5, 3 and 5 (differential abundance).

Outliers

In a separate set of simulations, datasets were generated by adding outliers to data generated with the negative binomial distribution. Outliers were introduced based on the pattern of Pearson residuals from a negative binomial distribution fit of the

real data. A count was considered an outlier when its Pearson residual was larger than or equal to 5. Samples with smaller library sizes contain more outliers. To preserve the relationships between the number of outliers and library size, the number of outliers added was derived from the observed fraction of outliers in the following way. The number of outliers effectively introduced in a sample i, n_i^{out}, was determined by binomial sampling with the observed outlier fraction as success probability and the total number of taxa as size parameter. The magnitude of the Pearson residuals was determined by sampling from the observed Pearson residuals of outliers for that particular body site. Outliers were introduced for n_i^{out} randomly chosen taxa by replacing their original simulated count x_{ij} (the count in sample i for taxon j) by

$$x_{ij}^{new} = r_{ij}^{Pearson} * \sqrt{\rho_j s_i (1 + \rho_j s_i * \phi_j)} + \rho_j s_i \quad (1)$$

with $r_{ij}^{Pearson}$ the observed Pearson residual and ϕ_j, s_i and ρ_j the estimated dispersion, library size and mean relative abundance, respectively. These new values x_{ij}^{new} were rounded to the nearest integers and negative values were set to zero.

Parameter combinations

Per unique combination of sample size (5, 25 and 100), fold change (1, 1.5, 3, and 5), distribution (Negative binomial with and without outliers, with or without correlation, correlated beta-binomial and Dirichlet-multinomial) and body site (stool (AGP), stool (HMP), tongue dorsum (HMP) and mid vagina (HMP)), 250 simulation runs were executed. Prior to statistical analyses, taxa with a prevalence lower than 5% were trimmed from the generated datasets, as one would do with a real dataset.

Non-parametric simulation

The non-parametric simulation paradigm implemented in the R-package *SimSeq* enables generating synthetic datasets by resampling from a real dataset. *SimSeq* proceeds as follows: first a factor that is believed to be associated to some of the taxa abundances is selected, and all taxa are tested for differential abundance between the factor-defined groups. For each taxon the method computes a local false discovery rate (lfdr). Next, a number of taxa that are supposed to be differentially abundant is specified by the user. These differentially abundant taxa are subsequently sampled from the taxa with an lfdr below a cut-off of 5% with sampling probabilities equal to one minus the lfdr's. A user-specified number of non-differentially abundant taxa is next sampled with uniform weights from the taxa with an lfdr above the cut-off [29]. The original SimSeq method uses the Wilcoxon rank sum test for delivering the lfdr's, but this may favour this statistical test in subsequent method evaluations. Therefore, the lfdr's are here obtained as follows. First the lfdr's are estimated from t-tests, Wilcoxon rank sum tests, *edgeR*, *DESeq2* and *metagenomeSeq* (the latter two with TSS normalization, the others with their own normalization method), and subsequently the lfdr's are averaged by calculating their geometric mean. One minus this geometric mean was used as weight in the resampling procedure. The grouping factors used to test for differential abundance were IBD (Crohn's disease or ulcerative colitis) [41], penicillin use in the last year [42] for the AGP data and sex [43] for the HMP and AGP data. For the Zeller data, sex and cancer diagnosis were used to determine the groups. Subjects with diagnoses "healthy" and "small adenoma" were pooled for this purpose. For the AGP and HMP data, sample sizes of 5, 25 and 75 were used, for the Zeller data 5, 25 and 41.

Real data shuffling

This method starts from a subset of samples from a real dataset. The subset is selected so that it is homogeneous with respect to covariates that are believed to affect the microbiome composition. All samples in the subset are assumed to have on average the same composition. Next, this subset is evenly split at random into two artificial mock groups. The taxa are subsequently tested for dif-

ferential abundance between the mock groups. The process of randomly splitting and hypothesis testing is repeated many times. Since all taxa on average have equal means in both mock groups, because of the random splitting, all discoveries can be considered to be false and the results can be used to estimate the specificity [30, 32, 44]. We can also use these simulation results to assess the distribution of the p-values under the null hypothesis. An excess of small p-values compared to a uniform distribution can cause an excess of false discoveries [16, 36, 45].

The AGP dataset was subsetted to include only female Caucasians who declared not to be pregnant or gluten intolerant, not to have used penicillin recently, not to have used selective antibiotics in the last year, and not to have diabetes or IBD. These covariates (sex [46, 47], race [42], pregnancy [48], gluten intolerance [43], penicillin or selective antibiotics use [49], diabetes [7, 50] and IBD [41]) are thought to be sources of variability in the microbiome. The HMP data were subsetted to include only female subjects that were sequenced at the Washington university genome center. The splitting procedure was repeated 250 times, the samples sizes were 5, 25 and 100 per mock group for the AGP data and 5 and 25 per mock group for the HMP data.

Evaluation-verification method

The concept of the evaluation-verification method is to exploit real datasets with grouping factors that are believed to be associated to taxon composition. Since the truly differentially abundant taxa are unknown in this setting, a large proportion of the available samples is set apart as a *verification set*. Because of the large number of samples, it is assumed that the results of differential abundance tests on this set can be considered as the truth. The differential abundance detection methods are then also applied to the remaining samples, constituting the much smaller *evaluation set*. The splitting of the data into a verification and an evaluation set must happen randomly, with the restriction that the distribution of the grouping factors must be the same in the two sets. The test results obtained from the verification set are used to evaluate of the performance of the methods on the smaller evaluation set. The splitting into evaluation and verification sets is repeated many times. Hence, on average both sets are drawn from the same population [32]. Since it can be expected that test results on the evaluation and verification sets are more similar when the same test method is applied to both sets than when different test methods are used, the verification sets were analyzed with the other test methods under study. To limit the number of comparisons, only the method's default normalization method was applied. For t-tests and Wilcoxon rank sum tests, both total sum scaling and rarefying were used, because these are both popular normalization methods for these tests. The same grouping factors were used as for the non-parametric simulation with SimSeq.

Evaluation sets were constructed by selecting 5 or 25 subjects from each group for the AGP data and 5 or 20 for the metagenomics data. The corresponding verification sets were then constructed with the remaining samples, and were always at least 3.5 times as large as the evaluation sets. The splitting procedure was repeated 250 times in each setting.

Differential abundance testing

Normalization

Normalization factors "Trimmed mean of M-values" (TMM), "Relative log-expression" (RLE) and "Cumulative sum scaling" CSS were estimated with the default settings of the *edgeR*, *DESeq2* and *metagenomeSeq* packages respectively. The SAM normalization method from the *samr* package was slightly modified to adapt to the high fraction of zeroes encountered in microbiome data. All taxa with Chi-squared goodness of fit statistics between the 20th and 80th percentile were used to estimate the normalization factor, and a pseudocount of 1 was added to the estimated normalization factor

to avoid estimated normalization factors equal to 0. For TSS or "total sum scaling", library sizes were used as normalization factors. As specified in the *edgeR* help [14], TMM normalization factors were multiplied by the library sizes to obtain the final normalization factors. Normalization type "none" means normalization factors equal to 1 were used. Rarefying was done by random subsampling counts with replacement to the smallest library size in the dataset. Subsequently, normalization factors equal to 1 were applied to these rarefied data. For two-sample t-tests and Wilcoxon rank sum tests, normalization was achieved by dividing the observed counts by the normalization factors. For methods based on generalized linear models (*edgeR*, *DESeq2*, *limma − voom* and *metagenomeSeq*) the normalization factors were used as offsets. *ALDEx2* uses the geometric mean as an inherent normalization technique and was not combined with the other normalization factors [18]. For *edgeR* only TSS, TMM and rarefying normalization were used.

Statistical tests for differential abundance

Differential abundance detection always occurred between two groups, without correction for any other covariate than the normalization factors. The following methods were applied: two-sample t-test and Wilcoxon rank sum test, permutation t-test and permutation Wilcoxon rank sum test [51], *ALDEx2*, *edgeR*, *DESeq2*, *metagenomeSeq*, *SAMseq* and *limma − voom*. In *edgeR* the robust option was used, in all other packages default settings were used.

Multiple testing correction

Multiple testing corrections were done by the Benjamini-Hochberg [22], Benjamini-Yekutieli [27] and local false discovery rate [28, 52] procedures. *SAMseq* employs its own plug-in multiple testing correction and directly returns differentially abundant taxa given a nominal FDR level [53]. The nominal false discovery rate was set at 5% for all analyses.

Performance evaluation

The results from the simulations where differential abundance is expected (parametric simulation under H_1, the non-parametric simulation and the evaluation-verification method) were evaluated for sensitivity, false discovery rate (FDR), specificity, area under the ROC-curve (AUC) and for the non-parametric simulation also departure from uniformity of the p-values of the non-differentially abundant taxa in both the liberal and the conservative direction (see below). With FN the number of false negatives, FP the number of false positives, TP the number of true positives and FP the number of false positives among the test results of a single simulated dataset, we define:

$$\begin{aligned} Sp &= \frac{TN}{TN+FP} \\ Se &= \frac{TP}{TP+FN} \\ FDP &= \frac{FP}{TP+FP}. \end{aligned} \tag{2}$$

When no taxa were declared significant (i.e. TP=FP=0) the FDP (false discovery proportion) was set to 0. An estimate of the false discovery rate (FDR) was obtained by averaging the values of FDP over the 250 simulation runs. Similarly, sensitivity and specificity were calculated as the average of Se and Sp, respectively, over the 250 simulation runs. The AUC is a measure of classifier accuracy that varies between 50% for a random classifier and 100% for a perfect classifier. The AUC was calculated as the area under the ROC-curve, obtained by plotting the sensitivity versus 1-specificity for varying p-value thresholds.

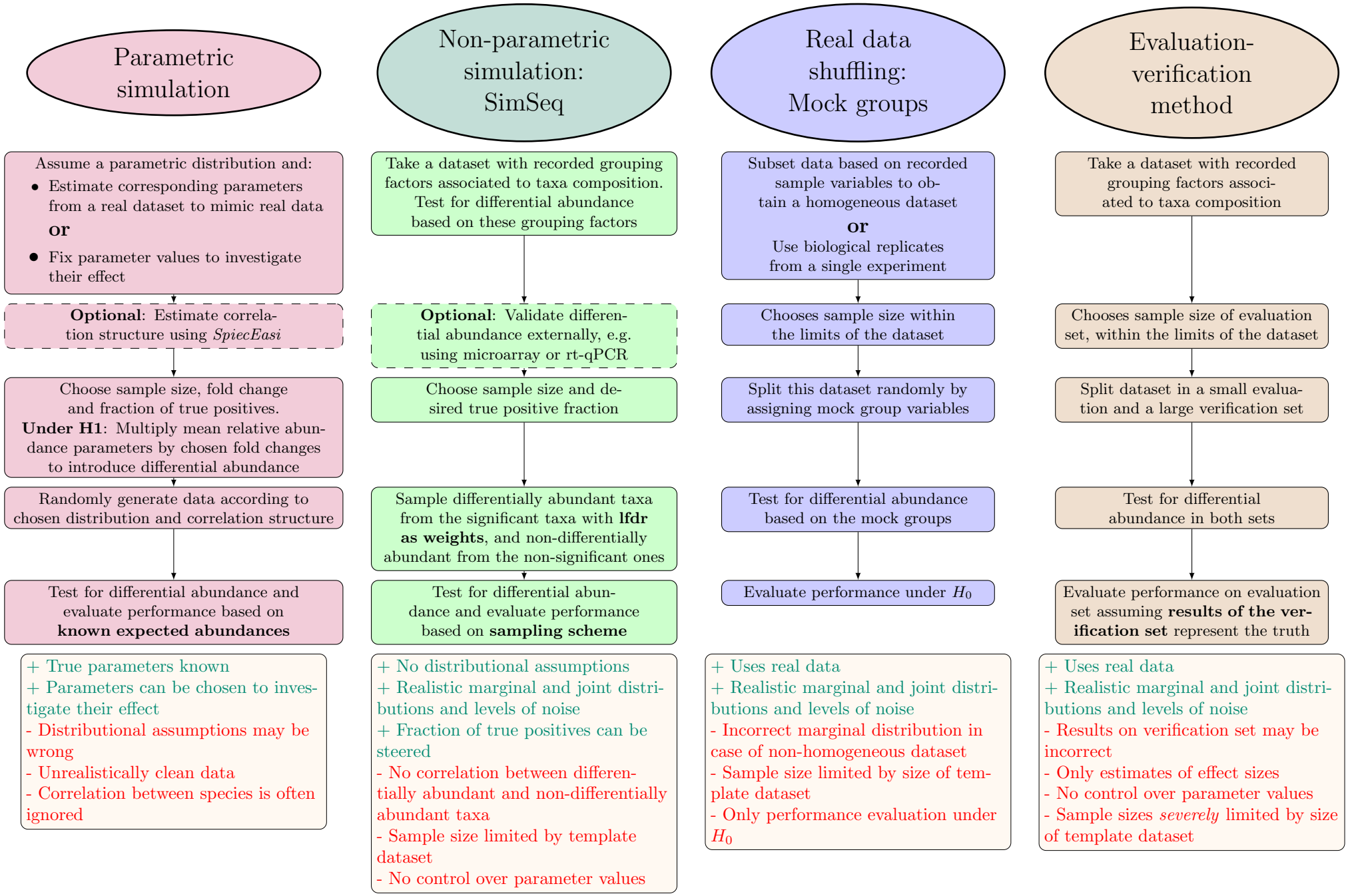

Figure 1: Overview of data generation paradigms used in this study. The data generation steps are represented as a flowchart from to to bottom, below the strengths (in green) and weaknesses (in red) of every method are listed.

The departure of the p-values from uniformity was quantified as follows. For each simulation scenario and for each taxon for which the null hypothesis holds true (i.e. non-differentially abundant), p-values obtained from the simulations were used for the construction of a QQ-plot for the uniform distribution: sorted p-values were plotted against the corresponding quantiles of the uniform distribution. We are mainly interested in p-value distributions that are stochastically smaller than uniform, because they can lead to inflated type I errors (i.e. liberal hypothesis tests). To quantify the departures from uniformity into this direction, twice the mean distance between the diagonal line and the points in the QQ-plot below the diagonal was computed (see figure 2). This measure will further be called the "liberal area" and can range from 0 when there are no p-values smaller than expected, to 1 for extreme departure from uniformity in the liberal direction. Analogously, also the "conservative area" was calculated as the distance between the diagonal line and points above it. Both calculated areas can be averaged over all taxa to get a summary statistic for each simulation scenario. The results from the datasets where no differential abundance was expected (parametric simulation under H_0 and the real data shuffling method) were evaluated for specificity and for the departure of the p-value distributions from uniformity in liberal and conservative directions.

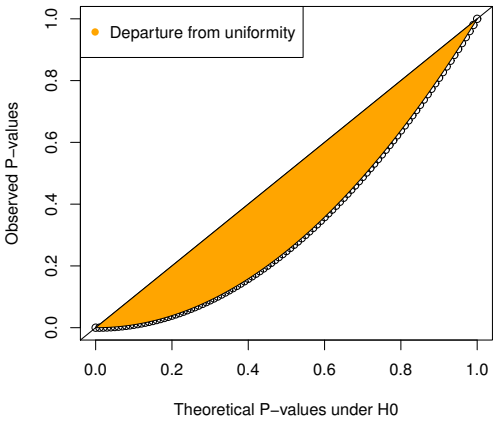

Figure 2: Illustration of how departures from uniformity of the p-value distribution under the null hypothesis can be quantified through the area between observed and theoretical, uniform quantiles. Here the case of a p-value distribution stochastically smaller than uniform is shown.

Results and discussion

Generation of correlated count data

Microbiome or RNA-Seq parametric simulation studies almost invariably generate counts taxon-by-taxon (or gene-by-gene), which requires only the specification of their marginal univariate distributions. This implicitly assumes independence between counts of different taxa. However, bacteria in a community interact with each other and as a result their abundances are expected to be correlated. To accurately mimic the joint distribution of a microbiome dataset one needs to correctly specify the marginal distribution of each taxon separately as well as the correlation structure. A multivariate count model such as the Dirichlet multinomial [54] comes with a complete joint distribution but makes strong and unrealistic assumptions on the correlation structure. Alternatively, we propose to estimate the correlation network from real data and combine it with marginal count distributions of choice. The correlation network is estimated with the sparse network estimation methodology implemented in the *SpiecEasi* R package [39]. Next, correlated count data are generated with the "Normal to anything" approach implemented in the same R package. This exists in the generation of correlated multivariate normal data with the desired dimensions and correlation structure (as estimated from real data), conversion to the copula space with the normal cumulative distribution function, and finally transformation to the desired marginal count distributions through the use of the corresponding quantile functions. This data generating pipeline is schematically represented in figure 3; for further details see the Supplementary Material, section 3.

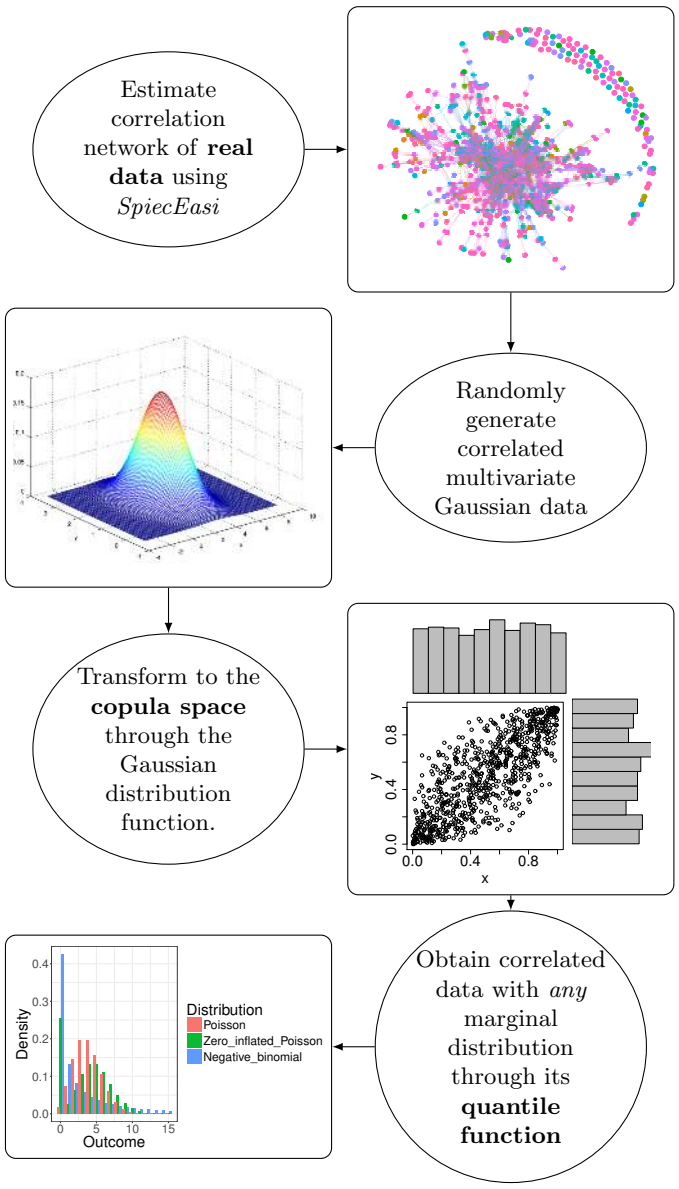

Figure 3: Schematic representation of the pipeline for generating count data with a realistic correlation structure. Correlation networks are estimated from the data using *SpiecEasi*, after which count data with this correlation structure are generated with the normal-to-anything approach. See text for details.

Sensitivity

Under parametric simulation with compensation the sensitivity increases with sample size and effect size for all methods except for $SAMseq$ and $DESeq2$ (see figure 4 and Supplementary figures S4- S6). $metagenomeSeq$ is the most powerful method at sample sizes of 25 and 100 (with a sensitivity of 50% and higher); $DESeq2$ and $SAMseq$ are the only methods with a modest sensitivity at a sample size of 5. For $SAMseq$ the sensitivity is higher under the negative binomial distribution when the counts are correlated. The results under parametric simulation without compensation are very similar, with only $edgeR$ achieving slightly lower powers (see Supplementary figures S1- S3). In non-parametric simulation and the evaluation-verification method, also $DESeq2$ and $SAMseq$ are the only methods with sensitivity at a sample size of 5; $edgeR$ and $metagenomeSeq$ are most sensitive at sample sizes of 25 and 100 (see Supplementary figures S7- S14). The sensitivity varies considerably among the taxa, depending on their relative abundance and frequency of zeroes (see Supplementary figures S82-S85).

False discovery rate

In the parametric simulation with compensation, the false discovery rate (FDR) is seen to decrease with sample size for $DESeq2$, $limma-voom$ and $SAMseq$, but to increase with sample size for $edgeR$ and $metagenomeSeq$. The t-test and Wilcoxon rank sum test control the FDR below the nominal level in all settings (see Supplementary figures S18 - S20). The FDR of $DESeq2$, $edgeR$, $metagenomeSeq$ and $SAMseq$ exceed the nominal level of 5% by a large margin in many settings, no matter which FDR multiplicity correction is applied. For $limma-voom$ this only happens when the Benjamini-Hochberg multiplicity correction is applied. The results under parametric simulation without compensation are very similar, except that now $edgeR$ has a lower FDR, and the Wilcoxon rank sum test and $DESeq2$ have a higher FDR (see figure 4 and Supplementary figures S15- S17) This suggests that some of the false discoveries of the latter two methods are due to the compositionality effect, which is not present in the setting "with compensation".

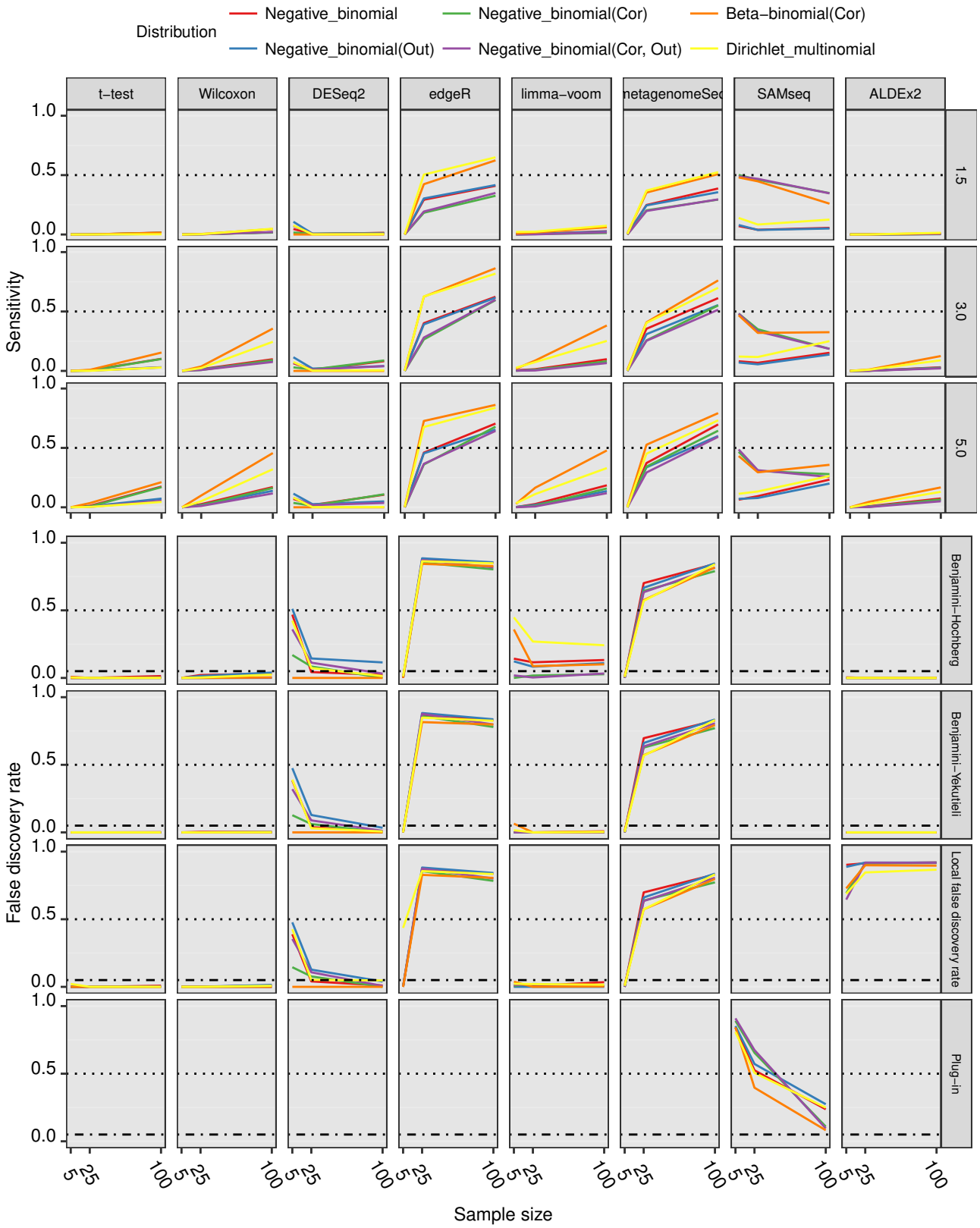

Figure 4: Results from parametric simulations with compensation with the stool samples of the American Gut project as template. Top: Sensitivity after Benjamini-Hochberg multiple testing correction (plug-in correction for SAMseq). Bottom: false discovery rate with a fold change of 3. Top panels indicate the testing method used, right panels the fold change applied (top) and the multiple testing correction applied (bottom). Colours indicate the distribution through which the counts were generated. Normalization occurred through the geometric mean for $ALDEx2$ and with library sizes for the other methods. Dotted horizontal lines indicate 50%, dot-dashed lines indicate nominal FDR rate of 5% (bottom). No lines are drawn when the multiplicity correction method was not applied Out: outliers added, Cor: Counts generated with estimated correlation networks. Error bars are omitted for clarity; for the sensitivity the interquartile ranges lie between 0 and 0.24 for t-test, Wilcoxon rank sum test, DESeq2 and ALDEx2; reach up to 0.58 for edgeR, Wilcoxon rank sum test and SAMseq and up to 0.66 for metagenomeSeq. For the FDR the interquartile ranges range are 0 for t-test and Wilcoxon rank sum test, and lie between 0 and 0.93 for all other methods

Under non-parametric simulation the FDR also decreases with increasing sample sizes for $DESeq2$, $limma-voom$ and $SAMseq$, and increases with sample sizes for $edgeR$ and $metagenomeSeq$. Only $limma-voom$, the t-test and the Wilcoxon rank sum test control the FDR at or below the nominal level (see Supplementary figures S22 - S24).In the evaluation-verification method, similar patterns are observed. In all settings for which a method has a sensitivity above 5%, the nominal level of the FDR of 5% is exceeded. In the setting of the evaluation-verification method, even the Wilcoxon rank sum test does exceed the nominal FDR (see Supplementary figures S25 - S28).

P-value distribution under the null hypothesis

In most scenarios we evaluated, the false discovery rate control strategies fail to control the FDR at the nominal level. This could be caused by incorrect raw p-values obtained by the statistical methods. Even a small number of incorrect p-values can throw off the overall false discovery rate control because FDR control methods take the ensemble of all p-values as input. Some FDR control methods even work when the hypotheses tests are correlated [27], but they all assume uniformity of the p-value distributions under H0 (or at least no tests with p-values stochastically smaller than uniform).More *small* p-values than expected under uniformity indicates that the test is too liberal, which may cause an excess of false discoveries. For each taxon, this is quantified in the simulation studies through the *liberal area*. On the other hand, with small sample sizes and with discrete count data, the p-value distribution can also exhibit discreteness and show more large p-values than expected under uniformity (quantified by the *conservative area*). With such tests present, FDR control methods may result in conservative control of the FDR. When both types of departure from uniformity are present, FDR may deviate from its nominal level in either direction.

As can be seen from figure 5, the p-value distributions of $edgeR$, $ALDEx2$, $limma-voom$, t-test and $metagenomeSeq$ depart most from uniformity in the liberal direction in parametric simulation under H0 (see also Supplementary figures S56 - S57). The departures are larger when the counts of the negative binomial distribution are correlated for $edgeR$, but smaller for t-test and $metagenomeSeq$. With the mock variable method the departures from uniformity are more severe, but here also t-test, $limma-voom$, $edgeR$ and $metagenomeSeq$ have the strongest departures from uniformity in the liberal direction (see Supplementary figure S58). Very similar patterns are observed for the null taxa under non-parametric simulation with SimSeq (Supplementary figures S59 - S60). This confirms previous findings of non-uniformity of p-values [24] and may partly explain the excess of false discoveries for these methods. The departure from uniformity in the liberal direction varies considerably between the taxa; for all the aforementioned methods and $DESeq2$ there are subgroups of taxa with a considerable liberal area. Taxa with large liberal areas are more often reported as false positives (see Supplementary figures S96 - S98). The liberal area is largest for taxa with small abundances and intermediate frequencies of zero counts (see Supplementary figures S86 - S95). For taxa with very high numbers of zeroes, there is either no departure from uniformity or a departure in the conservative direction, and most methods have a lower power (see Supplementary figures S99 - S107 and S82 - S85). This indicates that if taxa have too many zero counts, then their p-value distributions become very discrete and more large p-values than expected are obtained.

Specificity

In parametric simulations, all methods except for $metagenomeSeq$, $edgeR$, $DESeq2$ and $SAMseq$ are almost 100% specific (see Supplementary figures S29 - S31). Under non-parametric simulation and with real data shuffling and the evaluation-verification method, very similar patterns can be observed (see Supplementary figures S34 - S41).

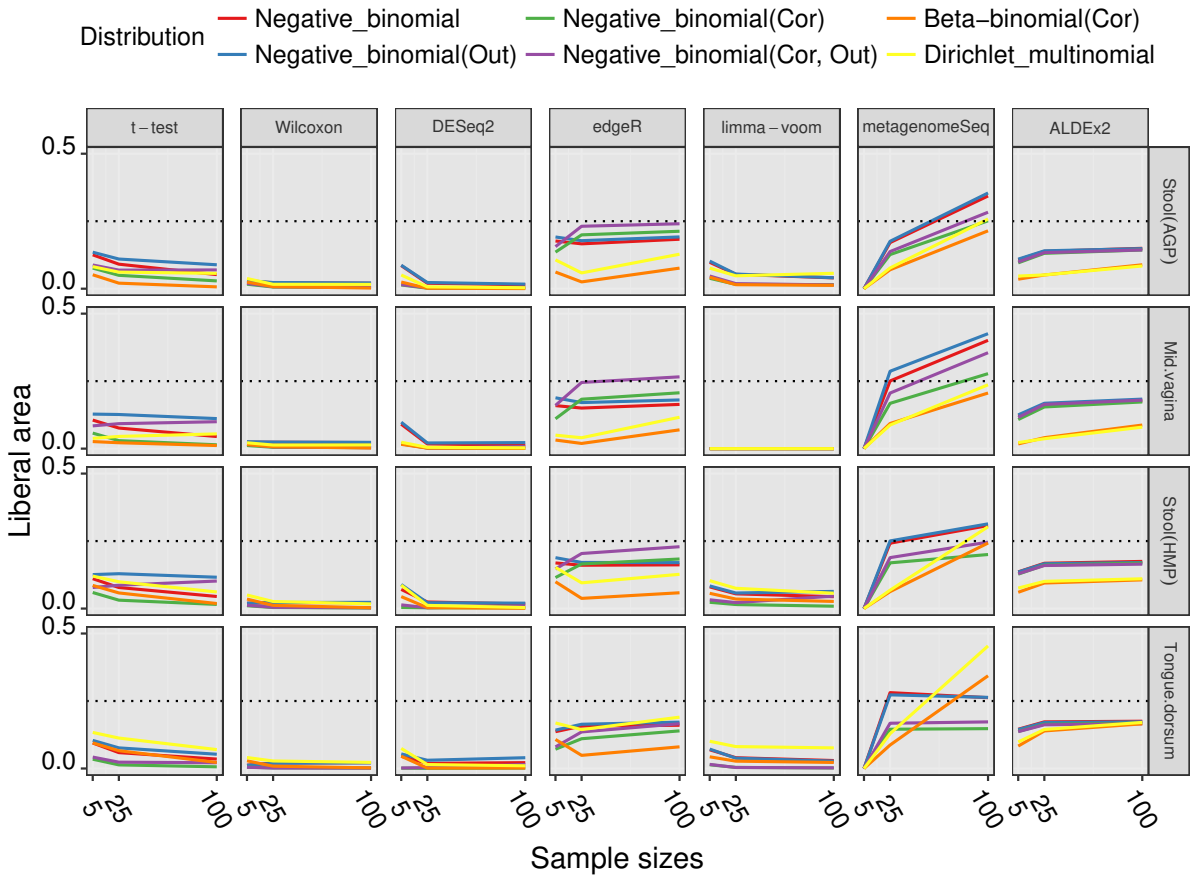

Figure 5: Liberal areas as measure of departures from uniformity of the p-value distribution for parametrically simulated data under the null hypothesis, averaged over the taxa. See section 1.5.1 in the supplementary material for a detailed discussion of how these areas are calculated. Top panels indicate the testing method used, right panels the template datasets and the colour of the lines the distribution by which the data was generated. Geometric mean normalization was used for $ALDEx2$ for other methods library sizes are used as normalization factors. Error bars are omitted for clarity; the dotted horizontal lines indicate 50%. The inter quartile ranges lie between 0 and 0.1 for t-test, between 0 and 0.23 for Wilcoxon rank sum test and ALDEx2, between 0 and 0.34 for edgeR up, and between 0 and 46 for DESeq2, for metagenomeSeq it reaches up to 0.53 and for limma-voom up to 0.95. Out: outliers added, Cor: Counts generated with estimated correlation networks. The horizontal dashed line is a reference line at 0.25.

AUC

In parametric simulations, *edgeR*, *limma − voom* and *metagenomeSeq* achieve the highest AUC values (see Supplementary figures S44- S45). In non-parametric simulations with *SimSeq*, the Wilcoxon rank sum test, *edgeR*, *metagenomeSeq* and *limma − voom* also have the largest AUC values (see Supplementary figures S48- S55). For the evaluation-verification method the AUC values are too variable to draw any meaningful conclusions.

Conclusions

The quick rise of microbiome science necessitates the development of new specialized statistical methodologies. Detection of differences in mean relative abundance or "differential abundance" is often one of the key goals. To decide on which methods are best suited, benchmark studies comparing the different methods are urgently needed. Although for RNA-Seq data many comparative studies have been carried out already on the same

methods [24, 29, 30, 31, 35, 37, 38, 45], microbiome data require special attention because of their higher variance and higher frequency of zeroes and larger sample sizes. When new methods for differential expression or differential abundance are proposed in the literature, their performance is usually evaluated in parametric simulation studies. In the simulation studies of these papers [11, 26, 32, 55] often data are generated with the same distribution as is used in the construction of the testing method itself. Hence, as a self fulfilling prophecy these papers conclude that their proposed methods perform well. For a broad, honest and reliable assessment of the performance of a large set of methods, we used multiple count models in parametric simulation studies. This allows us to see how the methods perform when their distributional assumptions are violated. Complementary to that, non-parametric simulations and use of real data allow for the evaluation of methods under realistic correlation structures between different taxa and realistic levels of noise and confounding. To allow for realistic correlation structures in parametric simulations as well, we proposed a new data generating method. We exploited recent advances in sparse network inference [39] for estimating real correlation networks and generating correlated counts with any marginal univariate count distribution. Our simulations indicate that the correlation between taxa counts can affect the performance of some of the testing procedures.

Several methods have been proposed to address the issue of varying sequencing depths of read count data through *normalization* [11, 13, 14, 15, 16], and some simulation and other studies have tried to establish an optimal choice of normalization factor [12, 45, 56]. In our results, no single normalization method uniformly outperforms the others under all settings. This suggests that the default normalization method of each R package may be used, or that simply the use of the library size as a normalization factor is sufficient.

We found that popular differential abundance methods, such as two sample t-test and Wilcoxon rank sum test, and methods based on log-ratios such as $ALDEx2$, have very low power to detect differential abundance. An even more worrying result obtained with all evaluation approaches is that the false discovery rate (FDR) exceeds the nominal level by a large margin for all methods that do have a sensitivity above 20% ($edgeR$, $SAMseq$, $metagenomeSeq$ and $ANCOM$, see Supplementary section 8 for the latter method's results). The opposite has been claimed before[24], but many studies on RNA-Seq methods have obtained similar results, albeit less extreme than ours [17, 29, 30, 31, 35, 37, 38, 57]. Still this very general result has never been obtained in the microbiome setting, or under such a wide range of conditions that were explored in this paper. In conjunction with these findings our results raise great concern about the reliability and reproducibility of microbiome research. A typical result of a differential abundance analysis is a list of significantly differentially abundant taxa. The researcher set the nominal FDR level such that it reflects the average risk of a finding being a false positive, he is willing to accept; this nominal level hardly ever larger than 10% and is often set to 5%. The researcher (and the scientific community at large) thus expects that only about 5% of the reported significant findings are false. Our results indicate that in reality this promise is often broken and up to more than half of the discoveries are false! As a consequence, researchers should treat the list of significant taxa with great suspicion, particularly when taxa are only marginally significant and the fold-changes and sample sizes are small.

In publications in which new methods for testing for differential abundance are introduced [11], as well as in comparative simulation studies [12, 23, 58], authors have failed to report on the FDR and only compared methods in terms of the sensitivity/specificity trade-off. However, when only a minority of the taxa are differentially abundant it is possible for the specificity to be close to 1, whereas the true FDR exceeds the nominal level. Therefore sensitiv-

ity and FDR (rather than specificity) should be the two criteria of interest in benchmarking studies.
Our simulations showed that the raw p-values of some taxa are stochastically smaller than uniform under the null-hypothesis for the Wilcoxon rank sum test, *edgeR*, $limma - voom$, $ALDEx2$ and *metagenomeSeq*. This explains at least in part why *edgeR*, $limma - voom$, and *metagenomeSeq* fail to control the FDR at the nominal level. Especially taxa with low abundances and intermediate frequencies of zeroes have p-value distributions stochastically smaller than uniform. These departures from uniformity may reflect the fact that statistical inference is hampered by the skewed count distributions and the many zero abundances encountered in microbiome data. These count distributions are less tractable than the Gaussian distribution, and their associated generalized linear models require much larger sample sizes to result in precise parameter estimates and in approximate normal sampling distributions of the parameter estimators [24]. The latter is crucial for correct p-value calculation.
Further investigation into the causes of the failure of FDR control and to methods that do control the FDR correctly are important challenges in microbiome statistics research in the near future.

Key points

- We conducted an extensive simulation study to assess performance of methods for differential abundance detection in microbiome studies.
- We propose a framework to generate realistically correlated count data through parametric simulation and show that correlation between species' counts does negatively affect the performance of the statistical methods.
- *edgeR*, $DESeq2$, $limma - voom$, $SAMseq$ and *metagenomeSeq* fail to control the false discovery rate at the nominal level, questioning the reliability and reproducibility of the discoveries in microbiome studies.
- The failure to control the false discovery rate is caused by stochastically smaller than uniform p-value distributions, and can be attributed in particular to taxa with low abundances and moderately large number of zeroes.

Acknowledgments

This research relied on the computational resources and services provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.

References

1. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207–214.
2. Sekirov I and Finlay BB. The role of the intestinal microbiota in enteric infection. J Physiol 2009;587:4159–4167.
3. Ivanov II, Atarashi K, Manel N, et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell 2009;139:485–498.
4. Ivanov II and Littman DR. Segmented filamentous bacteria take the stage. Mucosal Immunol 2010;3:209–212.
5. Ravel J, Gajer P, Abdo Z, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 2011;108:4680–4687.
6. Kahrstrom CT. Microbiome: Gut microbiome as a marker for diabetes. Nat Rev Micro 2012;10:733–733.

7. Kostic AD, Gevers D, Siljander H, et al. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression towards Type 1 Diabetes. Cell Host Microbe 2015;17:260 –273.

8. Scher JU, Ubeda C, Equinda M, et al. Periodontal Disease and the Oral Microbiota in New-Onset Rheumatoid Arthritis. Arthritis Rheum 2012;64:3083–3094.

9. Janda JM and Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J Clin Microbiol 2007;45:2761–2764.

10. Morgan XC and Huttenhower C. Chapter 12: Human Microbiome Analysis. PLoS Comput Biol 2012;8:e1002808.

11. Paulson JN, Stine OC, Bravo HC, et al. Robust methods for differential abundance analysis in marker gene surveys. Nat Methods 2013;10:1200–1202.

12. McMurdie PJ and Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 2014;10:e1003531.

13. Anders S and Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106–R106.

14. Robinson MD and Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25–R25.

15. Bullard JH, Purdom E, Hansen KD, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010;11:94–94.

16. Li J, Witten DM, Johnstone IM, et al. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics 2012;13:523–538.

17. Mandal S, Treuren WV, White R, et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecology in Health and Disease 2015;26.

18. Fernandes AD, Reid JN, Macklaim JM, et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014;2:15–15.

19. Paulson JN, Pop M, and Bravo HC. Metastats: An improved statistical method for analysis of metagenomic data. Genome Biol 2011;12:P17 –P17.

20. Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 2014;10:766.

21. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011;12:R60 – R60.

22. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995;57:289–300.

23. Nookaew I, Papini M, Pornputtapong N, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res 2012;40:10084–10097.

24. Rigaill G, Balzergue S, Brunaud V, et al. Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. Briefings in Bioinformatics 2016.

25. Robinson MD and Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881–2887.

26. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 2014;15:1–17.

27. Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 2001;29:1165 –1188.

28. Efron B. Microarrays, Empirical Bayes and the Two-Groups Model. Statist. Sci. 2008;23:1 –22.

29. Benidt S and Nettleton D. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. Bioinformatics 2015;31:2131–2140.

30. Reeb PD and Steibel JP. Evaluating statistical analysis models for RNA sequencing experiments. Front Genet 2013;4:178.

31. Kvam VM, Liu P, and Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American Journal of Botany 2012;99:248–256.

32. Love MI, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

33. The NIH HMP Working Group, Peterson J, Garges S, et al. The NIH Human Microbiome Project. Genome Res 2009;19:2317–2323.

34. https://github.com/biocore/American-Gut/blob/master/data/AG/AG_100nt.txt. 2015.

35. Soneson C and Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 2013;14:91–91.

36. Jonsson V, sterlund T, Nerman O, et al. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics 2016;17:78.

37. Seyednasrollah F, Laiho A, and Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 2013;16:59–70.

38. Burden CJ, Qureshi SE, and Wilson SR. Error estimates for the analysis of differential expression from RNA-seq count data. PeerJ 2014;2.

39. Kurtz ZD, Müller CL, Miraldi ER, et al. Sparse and Compositionally Robust In-

ference of Microbial Ecological Networks. PLoS Comput Biol 2015;11:e1004226.

40. Danaher PJ. Parameter estimation for the dirichlet-multinomial distribution using supplementary beta-binomial data. Communications in Statistics - Theory and Methods 1988;17:1777 –1788.

41. Kostic AD, Xavier RJ, and Gevers D. The Microbiome in Inflammatory Bowel Diseases: Current Status and the Future Ahead. Gastroenterology 2014;146:1489–1499.

42. Looft T, Johnson TA, Allen HK, et al. In-feed antibiotic effects on the swine intestinal microbiome. Proc. Natl. Acad. Sci. USA 2012;109:1691 –1696.

43. Markle JGM, Frank DN, Mortin-Toth S, et al. Sex Differences in the Gut Microbiome Drive Hormone-Dependent Regulation of Autoimmunity. Science 2013;339:1084–1088.

44. White JR, Nagarajan N, and Pop M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol 2009;5:e1000352.

45. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology 2013;14:R95.

46. Marietta EV, Gomez AM, Yeoman C, et al. Low Incidence of Spontaneous Type 1 Diabetes in Non-Obese Diabetic Mice Raised on Gluten-Free Diets Is Associated with Changes in the Intestinal Microbiome. PLoS One 2013;8:e78687.

47. Singh P and Manning SD. Impact of age and sex on the composition and abundance of the intestinal microbiota in individuals with and without enteric infections. Annals of Epidemiology 2016;26:380 –385.

48. Koren O, Goodrich JK, Cullender TC, et al. Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 2012;150:470 –480.

49. Fortenberry JD. The uses of race and ethnicity in human microbiome research. Trends in Microbiology 2013;21:165 –166.

50. Larsen N, Vogensen FK, Berg FWJ van den, et al. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. PLoS ONE 2010;5:e9085.

51. McMurdie PJ and Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 2013;8:e61217.

52. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics 2008;9:1 –14.

53. Li J and Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 2013;22:519–536.

54. La Rosa PS, Brooks JP, Deych E, et al. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS ONE 2012;7:e52078.

55. Zhou X, Lindsay H, and Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Research 2014.

56. Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 2013;14:671–683.

57. Schurch NJ, Schofield P, Gierliński M, et al. arXiv:1505.02017 2015;43:10.

58. Ching T, Huang S, and Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA 2014;20:1684–1696.