Analysing microbiome intervention design studies: Comparison of alternative multivariate statistical methods

Maryia Khomich (maryia.khomich@uib.no)
University of Bergen https://orcid.org/0000-0002-6840-5739
Ingrid Måge (ingrid.mage@nofima.no)
Nofima https://orcid.org/0000-0003-0364-0225
Ida Rud
Nofima https://orcid.org/0000-0002-1758-292X
Ingunn Berget
Nofima https://orcid.org/0000-0003-1027-1472

Research Article

Keywords: gut microbiome, dietary intervention trials, differential abundance, multivariate, ANOVA, ASCA, FFMANOVA, ALDEx2, ANCOM, DESeq2, PERMANOVA, ANOSIM, SIMPER

Posted Date: September 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-910076/v1

Analysing microbiome intervention design studies: Comparison of alternative multivariate statistical methods

Maryia Khomich[1,2*], Ingrid Måge[3*], Ida Rud[1] and Ingunn Berget[3]

[1]Department of Food safety and quality, Division Food Science, Nofima – Norwegian Institute of Food, Fisheries and Aquaculture Research, P.O. Box 210, 1431 Ås, Norway

[2]Department of Clinical Science, University of Bergen, P.O. Box 7804, 5020 Bergen, Norway

[3]Department of Raw materials and process optimisation, Division Food Science, Nofima – Norwegian Institute of Food, Fisheries and Aquaculture Research, P.O. Box 210, 1431 Ås, Norway

[*]Corresponding authors

Maryia Khomich: marykhomich@gmail.com, maryia.khomich@uib.no; ORCID: 0000-0002-6840-5739

Ingrid Måge: ingrid.mage@nofima.no; ORCID: 0000-0003-0364-0225

Ida Rud: ida.rud@nofima.no; ORCID: 0000-0002-1758-292X

Ingunn Berget: ingunn.berget@nofima.no; ORCID: 0000-0003-1027-1472

Abstract

The diet plays a major role in shaping gut microbiome composition and function in both humans and animals, and dietary intervention trials are often used to investigate and understand these effects. A plethora of statistical methods for analysing the differential abundance of microbial taxa exists, and new methods are constantly being developed, but there is a lack of benchmarking studies and clear consensus on the best multivariate statistical practices. This makes it hard for a biologist to decide which method to use. We compared the outcomes of generic multivariate ANOVA (ASCA and FFMANOVA) against statistical methods commonly used for community analyses (PERMANOVA and SIMPER) and methods designed for analysis of count data from high-throughput sequencing experiments (ALDEx2, ANCOM and DESeq2). The comparison is based on both simulated data and five published dietary intervention trials representing different subjects and study designs. We found that the methods

testing differences at the community level were in agreement regarding both effect size and statistical significance. However, the methods that provided ranking and identification of differentially abundant operational taxonomic units (OTUs) gave incongruent results, implying that the choice of method is likely to influence the biological interpretations. The generic multivariate ANOVA tools have the flexibility needed for analysing multifactorial experiments and provide outputs at both the community and OTU levels; good performance in the simulation studies suggests that these statistical tools are also suitable for microbiome data sets.

Keywords

gut microbiome, dietary intervention trials, differential abundance, multivariate, ANOVA, ASCA, FFMANOVA, ALDEx2, ANCOM, DESeq2, PERMANOVA, ANOSIM, SIMPER

Introduction

The microbiome has emerged as an important link to health and disease [1]. Microbiome analysis methods are rapidly advancing, in particular in areas such as compositional data analysis, multi-omics and data integration [2, 3]. A clear understanding of the type of data being analysed is crucial, given the growing number of studies uncovering the key role of microbiome, its composition and functions following diet intervention or medical treatment [4]. At present, analysis of complex microbial data benefits from adapting the multivariate statistical toolbox from ecology and environmental sciences, and a proper choice of statistical tools is becoming increasingly important [5-7]. However, a lack of benchmarking studies and clear consensus on the best multivariate statistical practices make comparisons across microbiome data sets difficult [2, 8]. New methods are often tested by simulation studies, but there is always a concern that simulations can be biased towards the tested statistical model and cannot mimic the complexity of real microbiome data [6, 9]. Moreover, newly introduced tools are often optimised, whereas the comparison of several statistical methods implies the use of standard or default parameters [6, 9]. It is therefore of interest to compare existing methods on real data sets of different complexity, in addition to simulation studies, to better understand how choice of method affects the results.

Different statistical methods have different properties, and the choice of method should depend on the scientific question, experimental design, data characteristics and expected

relationships among the variables. Furthermore, the choice of method is often biased by the research groups' tradition and familiarity with specific "toolboxes". Main differences between existing statistical approaches for analysing microbiome data are related to: (1) explorative versus confirmative; (2) univariate versus multivariate; (3) parametric versus nonparametric; (4) linear versus nonlinear; (5) compositional versus non-compositional; (6) distance-based versus count/abundance-based; and (7) incorporating phylogenetic information into the analysis or not [10-12]. Here, we explore statistical methods for analysing microbiome data from designed experiments with a focus on dietary intervention trials. In contrast to observational studies, these are usually small in sample size but performed in (semi)-controlled environments and tailored to a specific research hypothesis. The studies often include multiple experimental factors, possibly with more than two levels, and it is therefore natural to turn to analysis of variance (ANOVA)-like methods. Notably, most published analytical tools in microbiome research are essentially univariate [6], which led us to the conclusion that comparison of alternative multivariate statistical tools is sorely missing. From a biologist's point of view, it is also important that the methods are easy to interpret, both at the multivariate (microbial community) and univariate (microbial taxa or operational taxonomic units, OTUs) levels (see Fig 1).

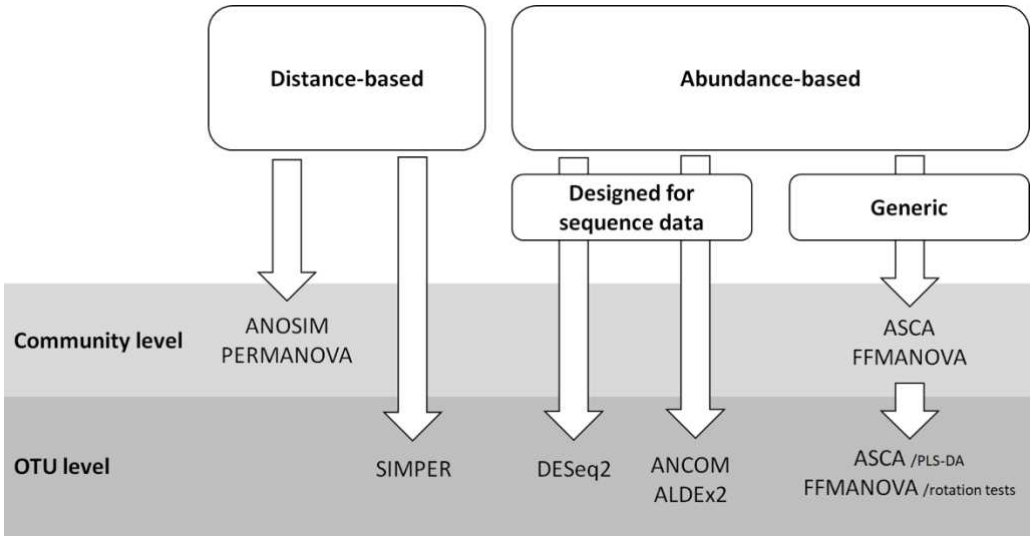

Fig 1. A diagram of statistical methods used in the study.

Distance-based methods

The distance-based methods are multivariate since multiple variables (microbial OTUs) are used to calculate pairwise distances between samples. Among distance-based methods,

permutational multivariate analysis of variance (PERMANOVA) is the most widely used and more powerful than the analysis of similarities (ANOSIM) to detect changes in community structure [13-15]. Both methods may be implemented with any dissimilarity metric. Among abundance-based beta diversity indices, Bray-Curtis is the most common choice for count data [16, 17]. The most widely applied phylogenetic beta diversity indices are UniFrac-type metrics [17-19]. However, UniFrac is unsuitable as a distance metric for studies with a small sample size, which is usually the case for dietary intervention trials [20, 21]. Both PERMANOVA and ANOSIM test differences at the community level but do not provide any information at the OTU level. Similarity percentage analysis (SIMPER) works at the univariate level by computing the relative contribution of each analysed microbial taxon (i.e. OTU) to the overall average Bray-Curtis dissimilarities by pairwise comparison of two or more groups [15]. To the best of our knowledge, no such method exists for the other distance metrics.

Distance-based methods have their strengths and weaknesses that are important to account for beforehand. ANOSIM cannot deal with multifactorial designs, and both ANOSIM and PERMANOVA may have problems detecting differences unless they are present in taxa with high variability [22]. Newer methods aimed at assigning more interpretable effect sizes are under constant development [6, 23]. For example, the more flexible PERMANOVA-S is an extension to existing distance-based methods that can adjust for covariates and simultaneously incorporate multiple distance metrics [24]. However, these methods do not consider covariance/correlation between microbial species, and they encounter significant power loss if *all* microbial species are used for distance calculations [25].

Abundance-based methods

The abundance-based methods can be either univariate (analysing each OTU individually) or multivariate (focusing on covariance structure between OTUs). There are two main approaches to deal with the special nature of abundance data: (1) application of methods that consider the distribution of count data or (2) compositional data analysis (CoDa) based on log-ratio transformed count data [26, 27]. Statistical methods designed for high-throughput sequencing data are ANOVA-like differential expression analysis (ALDEx2) [26], analysis of composition of microbiomes (ANCOM) [28], edgeR [29] and DESeq/DESeq2 [30]. edgeR and DESeq2 model count data directly using generalized linear models with the negative binomial distribution and the logistic link, respectively, whereas ALDEx2 and ANCOM use the log-ratio transformation prior to univariate assessment of statistical significance for individual OTUs.

DESeq2 and edgeR are based on the same modelling approach but differ in normalisation, outlier handling, and other adjustable parameters; these methods had similar performance in simulation studies [30]. Thus, we decided to include only one of the methods – DESeq2 – because differences between DESeq2 and edgeR are at a different conceptual level rather than the other methods discussed. ALDEx2 uses a Dirichlet-multinomial probability distribution to estimate abundances from count data and calculates the false discovery rate (FDR) based on Monte Carlo simulations (see Fig. 3 in [26] for details). In ANCOM, the compositional nature of the data is considered by testing the log-ratio for all pairs of OTUs, and then counting the number of tests where the log-ratio is significantly different from zero. This number (W-stat) can be used to obtain a ranking of OTUs most likely to differ between the groups. The newly published ANCOM-BC corrects the bias induced by differences in sampling fractions and provides p-values and confidence intervals for the differential abundance of each OTU [31]. The FDR and power were shown to be similar for both ANCOM and ANCOM-BC, and therefore we limit the present study to ANCOM.

The drawback of univariate methods is that they treat all taxa as independent variables without considering the covariance between the OTUs. Such methods may fail to detect community-level differences [32]. A classical generalisation of ANOVA to multiple variables (MANOVA) cannot be used when the number of variables exceeds the number of samples, as it suffers from the problem of a singularity of covariance matrices and assumptions that are not fulfilled [33, 34]. Novel statistical ANOVA-like methods include fifty-fifty multivariate analysis of variance (FFMANOVA) [35] and ANOVA-simultaneous component analysis (ASCA) [33]. Both methods are based on principal component analysis (PCA), and they can handle multiple collinear responses. In FFMANOVA, the multivariate effects are estimated by a modified variant of classical MANOVA, and OTU-level p-values are obtained by rotation tests which adjust the p-values for multiple testing [36]. For ASCA, the multivariate effects are calculated from combined sums-of-squares from all OTUs, and significance is assessed by permutation testing. ASCA also provides scores and loadings related to each experimental factor, which can be visualised in the same way as for PCA to better understand covariance patterns within the data. The contribution of each OTU can be quantified by the loadings or by partial least squares discriminant analysis (PLS-DA) for pairwise comparisons. ASCA has recently gained popularity in metabolomics [37-39], and both ASCA and FFMANOVA have successfully been applied to microbiome data [40-44].

Linear discriminant analysis effect size (LEfSe) is a stepwise approach that combines univariate analysis with multivariate discriminant analysis [45]. LEfSe has found wide application in microbiome research due to its easy to-use-and-interpret visualization [46, 47], but it is not adapted to experimental designs with several multilevel factors and is therefore not considered in this study.

Method comparison

An overview of the different methods compared in this study is given in Fig 1 and Table 1. ANOSIM and PERMANOVA provide results only at the community level, while SIMPER, DESeq2, ANCOM and ALDEx2 report results for single OTUs. ASCA and FFMANOVA are generic methods and the only methods that provide results at both the community and OTU level.

The aim of the method comparison was to investigate how different strategies for statistical modelling affect biological inference. At the community level, methods were compared with respect to effect sizes (expressed as percentage of explained variance) and corresponding p-values. At the OTU level, comparison of methods is complex because some methods provide results for an omnibus test of differences between factor levels (FFMANOVA and ANCOM), whereas the other methods provide ranking for specific pairwise comparisons (ASCA and PLS-DA, SIMPER) or contrasts/model coefficients (ALDEx2 and DESeq2). Even so, a biologist will make inferences based on the output provided by the chosen method, and in this context, it is relevant to compare the ranking statistics although the tests are not the same. In our study, the ranking of OTUs was compared by Spearman's rank correlation and by investigation of scatterplots between the different ranking metrics. For the simulated data, where we know which OTUs are differentially abundant, True Positive Rate (TPR) and True Negative Rate (TNR) were also evaluated.

We focused exclusively on designed experiments, which are usually smaller in sample size and are more controlled in contrast to observational studies. We used five published data sets as a basis for the comparisons (S1 Table). The following criteria for studies to be included were considered: (1) at least two-factorial experimental design with a minimum of two-factor levels; (2) either human or animal gut microbiome surveys; and (3) a taxonomic assignment at the OTU level reported. Diet is the main factor of interest in all five studies, and we restricted our comparisons to this factor.

184 The simulated data were based on data set 1 (S1 Table) using the same study design and
185 OTU counts as a starting point. Four different scenarios were simulated to investigate how the
186 methods perform in situations with varying effect sizes and different numbers of differentially
187 abundant OTUs. In all scenarios, one of the diet levels was manipulated to be significantly
188 different from the others, and there was no effect of the second experimental factor (dose). See
189 Methods section for further details.
190

Table 1. An overview of statistical methods and their properties.

Method	Method name	Number of experimental factors allowed	Parametric	Multivariate	Univariate	Provides output at community level	Statistics for ranking OTUs	Reference
ALDEx2	ANOVA-like differential expression tool for high-throughput sequencing data	any	yes	no	yes	no	p-values or effect sizes	[26]
ANCOM	Analysis of composition of microbiomes	main factor + covariates	yes	no	yes	no	W-stat for the main variable	[28]
ANOSIM	Analysis of similarities	one	no	yes	no	yes	no	[15]
ASCA	ANOVA-simultaneous component analysis	any	yes	yes	no	yes	loadings or PLS-DA regression coefficients	[33]
DESeq2	Differential gene expression analysis based on the negative binomial distribution	any	yes (GLM)	no	yes	no	p-values or effect sizes (coefficients)	[30]
FFMANOVA	Fifty-fifty multivariate ANOVA	any	yes	yes	yes (rotation tests)	yes	p-values	[35]
PERMANOVA	Permutational multivariate analysis of variance	any	no	yes	no	yes	no	[14]
SIMPER	Similarity percentage	two-group comparison	no	yes	no	no	permutation p-values	[15]

ANOVA – analysis of variance; GLM – generalized linear model; OTU – operational taxonomic unit; PLS-DA – partial least squares discriminant analysis.

Results

Community level

Four of the methods can be used to test the association between diet and the overall microbiome composition, the distance-based ANOSIM and PERMANOVA, and abundance-based ASCA and FFMANOVA. The results for each of the four simulated scenarios are shown in Fig 2, and the results across real data sets are summarised in Table 2.

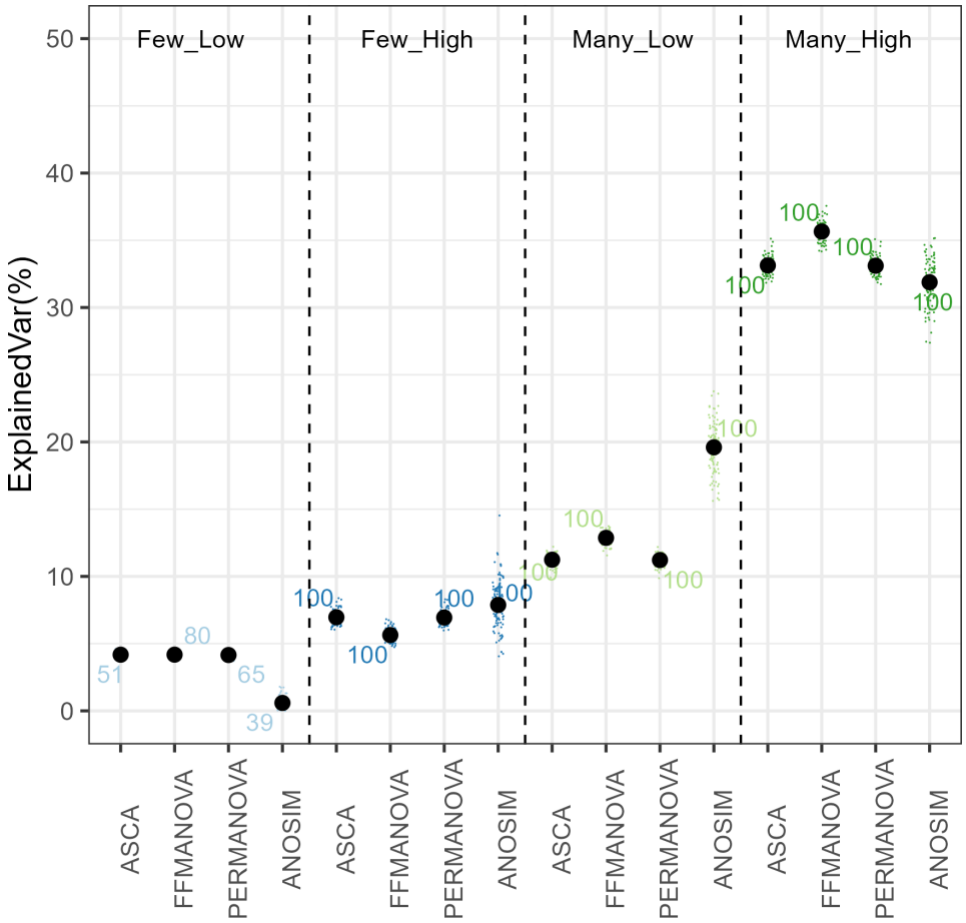

Fig 2. Explained variance for simulated data and the relative number of simulations where the simulated effect was detected.

204 **Table 2. Community-level method comparison across five experimental data sets.**

Factor/ predictor	FFMANOVA[clr]		ASCA[clr]		PERMANOVA[clr]		ANOSIM[3, clr]		Factor/ predictor
	Effect size (explained variance), %	p-value[1]	Effect size (explained variance), %	p-value[2]	Effect size (explained variance), %	p-value[2]	Effect size (explained variance), %	p-value[2]	
Moen et al., 2016 (data set 1) Model: OTU ~ fiber*dose							**Moen et al., 2016 (data set 1) Model: OTU ~ fiber_dose**		
fiber	34.31	< 0.001	37.39	< 0.001	37.26	0.001	82.19	0.001	**fiber**
dose	3.96	< 0.001	4.14	< 0.001	4.12	0.001			**dose**
fiber:dose	5.85	< 0.001	5.99	< 0.001	5.96	0.002			**fiber:dose**
residuals	55.75		52.69		52.67		17.81		**residuals**
Lai et al., 2016 (data set 2) Model: OTU ~ diet*exercise							**Lai et al., 2016 (data set 2) Model: OTU ~ diet_exercise**		
diet	27.31	< 0.001	30.50	< 0.001	30.85	0.001	99.11	0.001	**diet**
exercise	12.18	< 0.001	14.08	< 0.001	13.99	0.001			**exercise**
diet:exercise	7.93	< 0.001	7.51	< 0.001	7.47	0.002			**diet:exercise**
residuals	52.31		47.74		47.70		0.89		**residuals**
Le Sciellour et al., 2018 (data set 3) Model: OTU ~ diet*period + subject							**Le Sciellour et al., 2018 (data set 3) Model: OTU ~ diet_period**		
diet	1.78	< 0.001	1.99	< 0.001	2.14	0.001	13.33	0.001	**diet**
period	3.29	< 0.001	3.45	< 0.001	3.51	0.001			**period**
diet:period	1.32	< 0.001	1.41	0.006	1.36	0.009			**diet:period**
subject	26.53	< 0.001	26.42	0.073	27.28	0.046	86.67		**subject**
residuals	66.13		65.71		65.71				**residuals**
Wang et al., 2016 (data set 4) Model: OTU ~ diet + time + diet:time + subject							**Wang et al., 2016 (data set 4) Model: OTU ~ diet_time**		
diet	2.49	0.067	2.11	0.038	2.27	0.036	4.32	0.143	**diet**
time	2.00	0.007	1.96	0.082	1.77	0.246			**time**

diet:time	5.15	0.824	4.41	0.777	4.45	0.701			**diet:time**
subject	50.44	< 0.001	54.23	< 0.001	64.30	0.001	95.68		**subject**
residuals	31.23		27.25		27.21				**residuals**
Birkeland et al., 2020 (data set 5) Model: OTU ~ treatment:day + subject							**Birkeland et al., 2020 (data set 5) Model: OTU ~ treatment_day**		
treatment:day	1.75	< 0.001	1.27	0.107	1.27	0.132	-0.03	0.998	**treatment:day**
subject	69.38	< 0.001	73.85	0	73.85	0.001			**subject**
residuals	28.88		24.87		24.87				**residuals**

Distance-based ANOSIM and PERMANOVA and abundance-based ASCA and FFMANOVA were compared with respect to effect sizes (expressed as percentage of explained variance) and corresponding p-values.

[1]based on the 50-50 F-test, 999 permutations

[2]based on 999 permutations

[3]based on a combined factor with no interaction in the model (limitation of ANOSIM)

[clr]centred log-ratio transformed data as input

Simulated data. As expected, the multivariate effect size (explained variance) is lowest (around 5%) for the *"Few-Low"* scenario and highest (30-35%) for the *"Many-High"* scenario. The multivariate effect was significant in 100% of the simulations in three scenarios with the highest effect size. For the *"Few-Low"* scenario, FFMANOVA performed best by detecting the effect in 80% of the data sets. For the *"Many"* simulations explained variance was slightly higher with FFMANOVA than with ASCA and PERMANOVA, whereas for the *"Few-High"* simulations the opposite trend was observed. ANOSIM was less consistent than the other methods, with higher within-scenario variation and higher differences in effect size between the two *"Low"* scenarios.

Data set 1. The effect of dietary fibres inulin (IN), cellulose (CE) or brewers spent grain (BSG) on the overall caecal microbiota composition in mice from a study by Moen *et al.* [42] accounted for 34-37% of the explained variance according to the FFMANOVA, ASCA and PERMANOVA. In general, the three methods produced similar results, with slightly smaller p-values by FFMANOVA and ASCA.

Data set 2. Lai *et al.* [48] investigated the effect of diet (the main variable), exercise and their interaction on the overall faecal microbiota in sedentary and exercised mice fed high fat or normal fat diet (four groups in total). Similarly, all tested methods, except ANOSIM, produced congruent results (effect of diet 27-31%), with smaller p-values by FFMANOVA and ASCA.

Data set 3. In a longitudinal study by Le Sciellour *et al.* [49] the authors tested the effect of dietary fibre content on faecal microbiota in growing-finishing pigs fed alternately a low-fibre and a high-fibre diet during four successive 3-week periods. In this survey, the effect of diet was small (2%) compared to the effect of diet in data sets 1 and 2. Similarly, FFMANOVA and ASCA reported slightly lower p-values than PERMANOVA, but all three methods agreed on the effect of diet.

Data set 4. In a longitudinal study by Wang *et al.* [50] the objectives were to determine the impact of beta glucan on the composition of faecal microbiota in mildly hypercholesterolemic individuals. The individuals received for 5 weeks either a treatment containing 3 g high molecular weight (HMW), 3 g low molecular weight (LMW), 5 g LMW barley beta glucan or wheat and rice (control group) [50]. The effect of diet accounted for ~2% of the explained variance reported by FFMANOVA, ASCA and PERMANOVA. Diet was significant on a 5% level for PERMANOVA and ASCA ($p = 0.036$ and $p = 0.038$, respectively) and on a 10% level for FFMANOVA ($p = 0.067$). However, different conclusions were drawn

with respect to time where significant result at the community level was obtained only for FFMANOVA ($p = 0.007$).

Data set 5. Birkeland *et al.* [44] assessed the effect of prebiotic fibres or a control supplement on faecal microbiota composition in human subjects with type two diabetes. The interaction effect of treatment and day accounted for 1-2% of the explained variance according to the FFMANOVA, ASCA and PERMANOVA. Significant results at the community level were obtained only for FFMANOVA ($p < 0.001$).

To summarise, PERMANOVA, FFMANOVA and ASCA gave almost identical results regarding effect sizes and statistical significance across studies (Table 2). They all revealed that there was a considerable difference in effect sizes of the main factor of interest (diet) between animal (2-37%) and human (1-2%) dietary interventions, with effect sizes being very small in human studies. In addition, three of the studies had crossover designs allowing for estimation of interindividual variation. This variation was considerably higher for trials involving human subjects (54-74%) compared to the animal study (26%). ANOSIM provided the most different results and was not able to reveal the same biological insights since the multifactorial nature of the studies cannot be taken into consideration by this approach.

OTU level

Six of the methods, namely SIMPER, ASCA, FFMANOVA, ANCOM, ALDEx2 and DESeq2 can be used to make biological inferences for individual OTUs. The methods give different outputs which can be used to identify differentially abundant OTUs and/or rank the OTUs according to effect sizes (see Methods for details).

Simulated data. The True Positive Rate (TPR) for the four simulated scenarios is shown in Fig 3. The True Negative Rate (TNR) was close to 100% for all methods and is therefore not shown. ASCA provided the overall best results in terms of the TPR in all four scenarios. FFMANOVA, SIMPER and ANCOM were all highly sensitive in the scenario with few significant OTUs and a high effect size (*"Few-High"*). FFMANOVA was also highly sensitive in the scenario with many significant OTUs and a high effect size (*"Many-High"*); SIMPER and ANCOM performed best in the scenario with few significant OTUs with a lower effect size (*"Few-Low"*). Both ALDEx2 and DESeq2 detected very few OTUs in any of the scenarios and therefore had very low TPR.

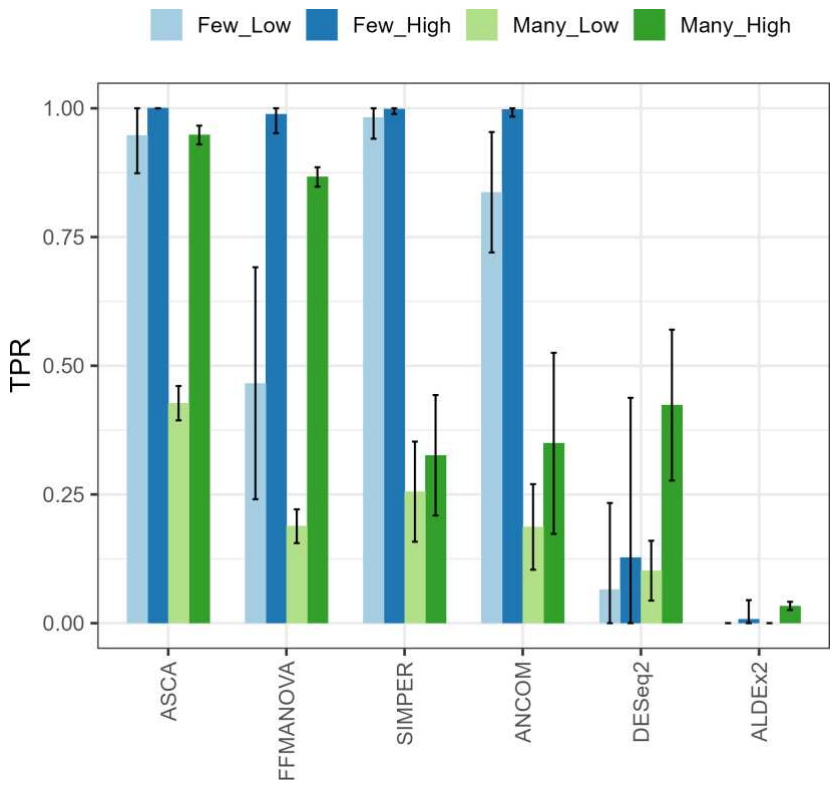

Fig 3. Sensitivity (True Positive Rate) for the four scenarios in the simulation study.

Experimental data. Summary tables and scatterplots comparing ranking for different methods on the experimental data sets are given as S1 File and S1 Fig, respectively. The number of significant OTUs detected by the methods is summarised in S2 Table. As expected, many significant OTUs were discovered in the studies with large multivariate effect sizes (data sets Moen and Lai), whereas few OTUs were found in the studies with low multivariate effect sizes. The highest number of significant OTUs was identified by FFMANOVA, with almost twice the numbers detected by ALDEx2. ANCOM differed considerably between the study designs. ASCA recovered fewer OTUs than the other methods for the Moen data set, but more OTUs than the other methods for the other data sets.

Correlation between the methods. Agreement between the methods was investigated by calculating Spearman's rank correlation between all pairs of output metrics (Fig 4). In the simulated data, FFMANOVA and ASCA had higher agreement than any other pair of methods, with correlations ranging from 0.6 to 0.75. In addition, the correlations were generally higher for scenarios with many differentially abundant OTUs for all methods. The results from the experimental data sets also showed that FFMANOVA and ASCA had highest agreement, with correlations ranging from 0.4 to 0.9. However, correlations varied considerably between the data sets. Highest agreements were observed for the animal studies, which are more controlled,

and where interindividual variation is smaller compared to studies involving human subjects. The Moen and Wang data sets had a lower correlation between the methods than the other data sets for many comparisons, which could be explained by the fact that there are more than two levels of the diet factor, and omnibus tests will therefore not completely agree with the pairwise comparisons.

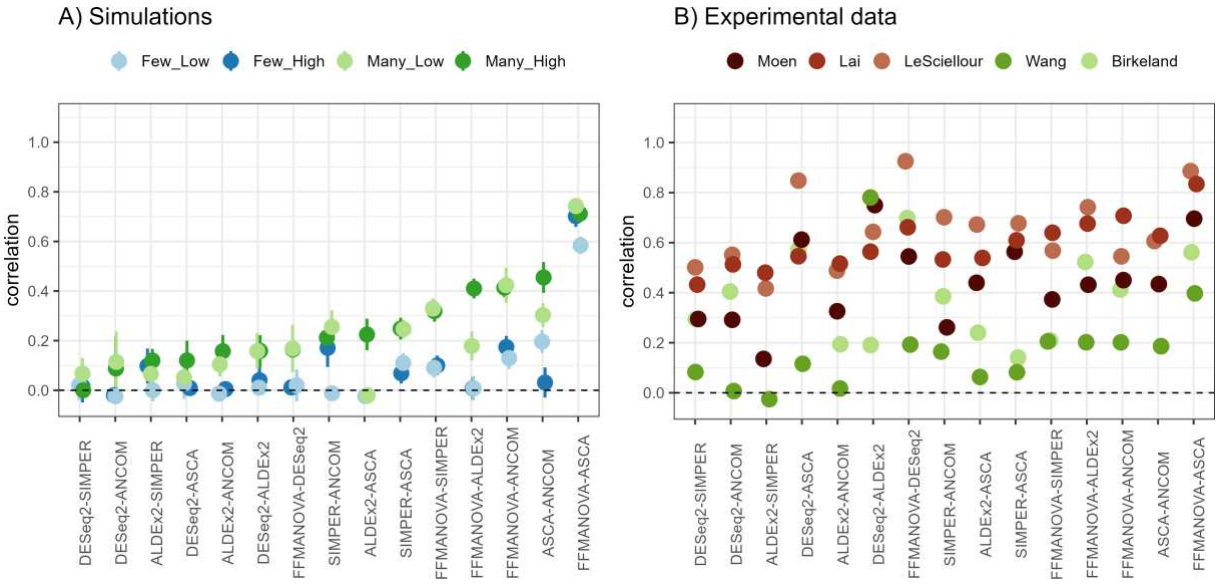

Fig 4. Spearman's correlation (Y-axis) calculated for pairwise comparison of statistical methods (X-axis) for (A) simulated data and (B) five experimental data sets. Each point represents Spearman's rank correlation coefficient between OTU ranking metrics from the two methods compared.

Impact of OTU abundance. Gut microbiome data are typically represented by a few dominant OTUs (relative abundance >1%) and a majority of low-abundant OTUs. One is therefore interested in ranking OTUs independent of their average abundance to detect biologically relevant changes in low-abundant OTUs. For all methods, except SIMPER, correlations in the range 0.1-0.4 were observed between the ranking statistics and the (log) mean relative abundance. However, it was not consistent between the data sets which method gave higher correlations. Correlations between the relative abundance and the ranking statistics were highest for the Moen data set and lowest for the Birkeland data set (S2 Table). ANCOM differed from the other methods as the ranking of OTUs was highly dependent on the abundance (Fig 5 and S1 Fig). In particular, the results indicated that highly abundant OTUs had either very high or very low W-stat, while low-abundant OTUs always had medium-to-low W-stat.

To the best of our knowledge, this finding has not been reported before and shows that ANCOM is not able to identify changes in low-abundant OTUs. In the simulation study, clear differences between the scenarios with *"Few"* or *"Many"* differentially abundant OTUs can be observed (Fig 5A and 5C), and similar patterns are reflected for the experimental data (Fig 5B and 5D). It has been shown that power of ANCOM drops when the number of differentially abundant OTUs exceeds 25% [28]. In our *"Many"* simulations, 50% of the OTUs are differentially abundant, which justifies why ANCOM performed poorly for these scenarios. The dependency on the relative abundance can be more problematic also for the *"Many"* scenarios (Fig 5C) and data sets with many differentially abundant OTUs as, for instance, the Moen data set (Fig 5D).

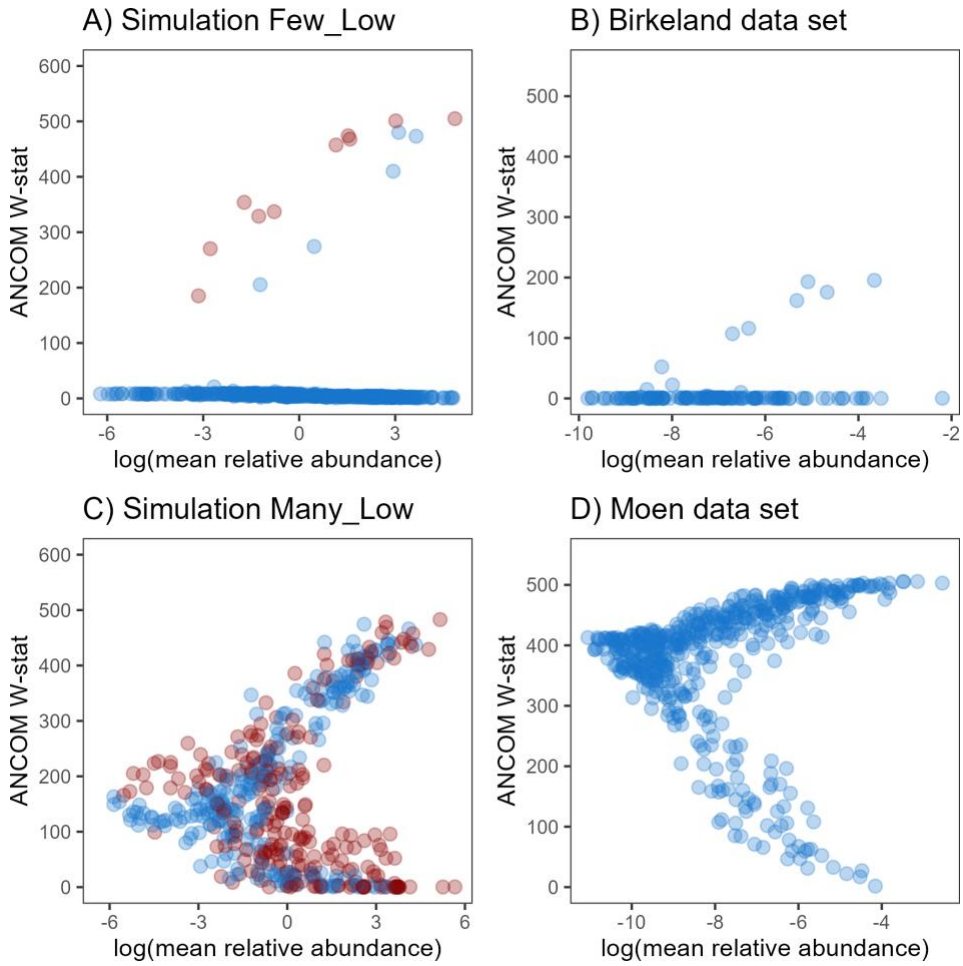

Fig 5. Mean relative abundance (log-scale) plotted versus ANCOM W-stat. (A) The *"Few-Low"* simulation scenario and (B) Birkeland data set, (C) *"Many-Low"* simulation scenario and (D) Moen data set.

Discussion

Diet is considered to be an important driver of microbiome variation [51]. However, in observational population-based studies, diet consistently accounts for only a small proportion of microbiome variation, and this is partly due to large interindividual differences in microbiome composition, small sample sizes and limitations in study designs such as potentially insufficient washout periods in crossover studies [51-55]. In general, higher interindividual variation is observed in gut microbiome of human subjects compared to animal species [56]. This was confirmed by our results, and it can partly explain the lower diet related effect on the gut microbiomes in the two human studies. Variation in how much a diet can influence the microbiome is also dependent on the nutritional differences of the compared diets. Nevertheless, use of animal models to study a causal role of the gut microbiome in health and disease is an established practice although animal models lack the specific interactions present in the complex system of a human organism [57].

Univariate and multivariate analyses provide information at different levels. Biologists will often find that the outputs of univariate analyses are easier to interpret compared to those generated by multivariate analyses, though assumptions are similar for both method types [10]. In general, multivariate methods provide a more holistic overview of differences between samples and account for correlations and interactions between the variables, whereas univariate methods are well suited to point out the differences for specific microbial groups. Therefore, the two levels of analysis provide complementary information, and it is generally of biological interest to report differences at both levels.

FFMANOVA and ASCA consider the covariance between all OTUs. It could therefore be expected that these methods would be better at detecting scenarios with many differentially abundant OTUs due to the *consistency at large* phenomenon, i.e. that many OTUs carry the same information and such methods collectively are able to detect small effects that are not significant at the univariate level [58, 59]. At the community level, these methods performed similarly to the distance-based methods. At the OTU level, they had considerably higher sensitivity to identify true positive OTUs than the other methods in the *"Many"* OTU scenarios.

FFMANOVA and ASCA depend on the relative scaling of the OTUs, while for the distance-based methods it depends on the chosen distance measure. It is a common practice in many areas to scale all variables to equal variance thus giving them an equal weight in the model, but other options are also possible, see, for instance, van den Berg *et al*. [60]. The clr-transformation puts the variables at comparable levels, and the need for scaling is less obvious.

However, the highly abundant OTUs might still have slightly higher variance and scaling should be considered depending on the data characteristics and the biological interpretation. With scaling, all OTUs have the same contribution in the analysis, whereas without scaling the more abundant OTUs will dominate the analyses and the inferences will be related to the more abundant OTUs.

It is a known fact that microbial sequence data are zero-inflated, and rare OTUs should be removed prior to downstream statistical analyses. We have observed that the threshold for filtering out OTUs can significantly affect the results both at the community and OTU levels. This can be exemplified in the human Birkeland data set, where stricter OTU-filtering performed in the present study resulted in no significant treatment effect by ASCA, in contrast to the original publication [44].

The tools tested in the present study vary in flexibility. SIMPER allows only pairwise comparisons and ANOSIM provides multigroup analysis, but it is not suited for multifactorial study designs. The other methods can employ more complex models with multiple factors with varying number of levels, and corresponding interactions. In experiments with repeated measurements, the subject effect is often included as a random factor. Neither of the methods discussed here can do this, hence the subject effect was included as a fixed factor. These aspects should be considered when selecting methods because different study designs might require different types of statistical models and tests. Newer developments of ASCA, namely ASCA+ [61] and LiMM-PCA [62], increase flexibility when there are unbalanced designs or random (i.e. subject) effects, respectively. Even so, the longitudinal modelling of large microbiome data sets in combination with multiple covariates is only starting to emerge [6].

Although Spearman's rank correlation indicated good agreement for the animal studies (Fig 4), little overlap between lists of "significant" results could be detected (results not shown). There can be several reasons for this. One important aspect is that different criteria must be used to define "significance" or generally "importance". FFMANOVA, SIMPER, DESeq2 and ALDEx2 provide p-values, while more heuristic tools must be applied with ASCA and ANCOM. For some methods, such as FFMANOVA and ANCOM, the ranking is related to *all* levels of the experimental factors, whereas the other methods use only pairwise comparisons (SIMPER) or comparison against a selected control/reference level (ALDEx2 and DESeq2). ASCA provides a test related to all levels at the community level, while the PLS-DA step relates to pairwise comparisons at the OTU level. This will introduce a bias for data sets with multilevel factors where more than one level is different from each other. Nevertheless, we chose to

compare these rankings as this is the output that is available to the user. In the simulations only one level of one factor was designed to have an effect, hence the "omnibus" and the "specific" tests are more directly comparable. In experiments with multilevel factors, additional information can be obtained by looking at the clr-difference between the groups of interest in addition to the ranking statistics.

Moreover, differences in sample collection, sample preparation and sequencing contribute to additional variability, which, in turn, affects the validity of the results [63] and complicates comparisons across studies with similar interventions.

Past benchmarking studies [9, 64, 65] have reported varying results from different tools, which was also confirmed in our study. Currently, there is no consensus for the best existing tool for detecting differentially abundant microbial taxa, and there is no reason to believe that one single method is best in all cases. Based on our simulations, the generic multivariate tools, ASCA and FFMANOVA, performed best. We anticipate that these results will inform future studies with more complex settings where more than one factor has an effect or interactions between experimental factors are included in the model.

In addition to performance, ease of use is an important aspect when selecting the appropriate tool. FFMANOVA and ASCA are based on standard statistical tools, namely PCA and ANOVA. Some of the tools designed for microbiome high-throughput sequencing data, on the other hand, can be difficult for non-statisticians, and it can be questioned if the users are able to interpret all parameters correctly, even if the methods are supported by comprehensive documentation. It is always good scientific practice to compare and report outputs from several methods. There are no standards on how to report multiple modelling results, and there is a high risk of "fishing for significance" when several methods are applied [66]. Before designing the experiment, researchers should be aware of the different properties of the statistical methods and consider whether it is most important not to miss out on any possible findings or to obtain robust results. In the latter case, OTUs should be reported as differentially abundant only if they were flagged as "significant" by several methods [66].

Conclusion

In the present study, we compared the performance of several multivariate ANOVA-like statistical methods taking four simulated scenarios and five real dietary intervention microbiome data sets as examples. At the community level, all the different methods came to

similar conclusions; at the OTU level, the agreement between the methods considerably varied. ANCOM provided output metrics that were dependent on the average abundance, making it impossible to detect differences in low-abundant OTUs. At the OTU level, the ranking of OTUs obtained with different methods correlated better for animal studies than for human studies, possibly due to lower interindividual variation in animal studies. Based on the simulation results we advise applying FFMANOVA and ASCA for overall and pairwise comparisons of microbiome data, respectively, also because these methods provide output at both the community and OTU levels, can handle several design factors, as well as other data types common in microbiome research.

Methods

Experimental design and data characteristics. The dietary intervention data sets were selected based on the study design, with a minimum of two independent variables (for example, diet and dose; see S1 Table for details). Prior to the statistical analyses, the data were filtered to keep the OTUs that were present: (1) with relative abundance more than 0.005% in an individual, and (2) in at least 50% of the individuals and in one of the groups. For each study in total 507, 561, 560, 397 and 216 OTUs passed this filter and were subsequently used in downstream analyses (S1 Table and S2 File). All statistical analyses were performed in R version 4.1.0 [67] unless otherwise specified and were run in 999 permutations.

Simulated data were generated using the R package metaSPARSim [68] developed for simulating 16S rDNA data. The simulation was a two-step process: (1) modelling the abundance (expected counts for each experimental group) using a Gamma distribution; (2) modelling the within-group variability using a Multivariate Hypergeometric (MHG) distribution taking the output from step 1 as input parameters for the distribution. In addition, metaSPARSim contains functions for parameter estimation from observed data. We used data set 1 [42] (S1 Table) to estimate real starting parameters for the simulations. Raw counts were generated for the same number of OTUs as in the experimental data set, and subsequently pre-processed and analysed in the same manner as for the experimental data (described below). Four different scenarios were simulated, with a varying number of the differentially abundant OTUs (*"Few"* versus *"Many"*) and the effect sizes (*"Low"* versus *"High"*). 100 data sets were generated and analysed for each of the four scenarios:

1) *"Few – Low"*: 10 randomly selected OTUs were assigned random log2 fold changes from a uniform distribution with boundaries [3,4].

2) *"Few – High"*: 10 randomly selected OTUs were assigned random log2 fold changes from a uniform distribution with boundaries [8,9].

3) *"Many – Low"*: 254 randomly selected OTUs were assigned random log2 fold changes from a uniform distribution with boundaries [3,4].

4) *"Many – High"*: 254 randomly selected OTUs were assigned random log2 fold changes from a uniform distribution with boundaries [8,9].

Zero-value replacements were done prior to clr-transformation [27, 69] by applying function *cMultRepl* with a setting *method = 'CZM'* in the R package zCompositions [70]. Zero replacement is an ongoing and yet unsolved statistical problem in microbiome research, and newer methods are constantly being developed and applied to both simulated and experimental data sets [71-74].

ANOSIM and **PERMANOVA** were run using the functions *anosim* and *adonis* in the R package vegan [75], with a setting *method = Euclidean* and the clr-transformed data as input.

SIMPER was run on filtered relative abundance data using function *simper* from the R package vegan [75]. The p-values for each OTU between selected pairwise comparisons of diet levels were used as a ranking metric; the p-value represents the probability of getting a larger contribution to the Bray-Curtis dissimilarity in a random permutation of the group factor.

FFMANOVA was performed by using the function *ffmanova* implemented in the R package ffmanova [35] on the clr-transformed and standardised data. Raw p-values were used as a ranking metric.

ASCA was run on the clr-transformed and centred data using an in-house implementation in MATLAB (R2018b, The MathWorks Inc.); p-values at the community level were calculated by permutation tests (n = 999). PLS-DA models [76] with the ASCA diet effect matrix, residuals as a predictor and factor levels as a response were used for pairwise comparison of diet levels. Variable Importance in Prediction (VIP) values [76] were used to identify significant OTUs, and the VIP threshold was set using the Uninformative Variable Elimination (UVE) method [77]. The UVE procedure was repeated 100 times, and OTUs that were above the threshold in >95% of the repetitions were defined as significant. For study designs with two diet levels, the ASCA loadings were used to rank OTUs, while PLS-DA regression coefficients were used for pairwise comparisons of multilevel diet factors.

ANCOM was run by using the ANCOM 2.0 source code implemented in R [28] using filtered relative abundance data as input. The W-stat was used to rank OTUs indicating the

number of significantly different pairwise log-ratios while adjusting for FDR by applying a Benjamini-Hochberg correction at a 0.05 level of significance.

ALDEx2 was run using the functions *aldex.clr* and *aldex.glm* from the R package ALDEx2 v.1.18.0 [26]. We used raw counts as input and p-values for selected factor level contrasts to rank OTUs.

DESeq2 was run using the functions *DESeqDataSetFromMatrix* (to generate object), *DESeq* (for analysis) and *results* (to extract results) from the Bioconductor package version 1.32.0 [30]. We used raw counts as input and raw p-values for selected level contrasts to rank OTUs (default settings).

Factor level comparisons. For study designs with more than two diet groups, only one pairwise comparison was analysed for the methods based on the pairwise group comparison, namely ALDEx2, ASCA and SIMPER. The following pairs with the most contrasting outcomes were compared: (1) BSG group vs. IN group [42]; (2) control group vs. 3 g HMW [50] and (3) Placebo 6 weeks group vs. Fibre 6 weeks group [44].

Differentially abundant OTUs were identified by setting thresholds on the ranking metrics for each method. For the methods providing p-values, the threshold was set to 0.01. For ANCOM, the 60th percentile of the empirical distribution of the W-stat was used as a threshold. For ASCA, OTUs selected in 95 out of 100 UVE-runs were identified as differentially abundant.

True Positive Rate (TPR) and True Negative Rate (TNR) were calculated for the simulated data. TPR (also called *Power* or *Sensitivity*) is calculated as *TP/P*, where *TP* is the number of true differentially abundant OTUs identified by a statistical method and *P* is the number of differentially abundant OTUs defined in the simulation setup. The TNR (also called *Specificity*) is calculated as *TN/N*, where *TN* is the number of true non-differentially abundant OTUs identified by a statistical method and *N* is the corresponding number defined in the simulation setup.

List of abbreviations

ANOVA: Analysis of variance; ALDEx2: ANOVA-like differential expression tool for high-throughput sequencing data; ANCOM: Analysis of composition of microbiomes; ANOSIM: Analysis of similarities; ASCA: ANOVA-simultaneous component analysis; DESeq2: Differential gene expression analysis based on the negative binomial distribution;

FDR: False discovery rate; FFMANOVA: Fifty-fifty multivariate ANOVA; OTU: Operational taxonomic unit; PERMANOVA: Permutational multivariate analysis of variance; PLS-DA: Partial least squares discriminant analysis; SIMPER: Similarity percentage.

Acknowledgements

We thank all the authors who were involved in data generation used in the present study. All data sets are properly cited and referred to in S1 Table.

Funding

This work was funded by Nofima – Norwegian Institute of Food, Fisheries and Aquaculture Research and Foundation for Research Levy on Agricultural Products (Projects RCN 262306 and 262308). The funding sponsors had no role in the design of the study, the collection, analyses and interpretation of data, in the writing of the manuscript or in the decision to distribute the results.

Author contributions statement

IM, IB and IR conceived the project. MK, IM, IB and IR designed the project. MK, IM and IB analysed the data. MK, IM, IB and IR interpreted the results. MK and IB generated the figures and tables. MK was a major contributor in writing the manuscript, with help from IM and IB. All authors reviewed and edited the manuscript and approved the final version.

Competing interests' statement

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable. The study is based on already published data sets listed in S1 Table.

Availability of data and materials

559 All data analysed in the present study are included in Supplementary Information.
560 Custom R scripts to run the experimental and simulated data analyses are available as S3-S5
561 Files.

562

563 ## References

564 1. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis.
565 Annu Rev Stat Appl. 2015;2:73-94. doi: https://doi.org/10.1146/annurev-statistics-010814-
566 020351.

567 2. Blanco-Míguez A, Fdez-Riverola F, Sánchez B, Lourenço A. Resources and tools for
568 the high-throughput, multi-omic study of intestinal microbiota. Brief Bioinform.
569 2019;20(3):1032-56. doi: https://doi.org/10.1093/bib/bbx156.

570 3. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best
571 practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410-22. doi:
572 https://doi.org/10.1038/s41579-018-0029-9.

573 4. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome
574 high-throughput sequencing data. Can J Microbiol. 2016;62(8):692-703. doi:
575 https://doi.org/10.1139/cjm-2015-0821.

576 5. Lê Cao K-A, Costello M-E, Lakis VA, Bartolo F, Chua X-Y, Brazeilles R, et al.
577 MixMC: a multivariate statistical framework to gain insight into microbial communities. PLoS
578 One. 2016;11(8):e0160169. doi: https://doi.org/10.1371/journal.pone.0160169.

579 6. Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental
580 design and quantitative analysis of microbial community multiomics. Genome Biol.
581 2017;18(1):228. doi: https://doi.org/10.1186/s13059-017-1359-z.

582 7. Waldron L. Data and statistical methods to analyze the human microbiome. mSystems.
583 2018;3(2):e00194-17. doi: https://doi.org/10.1128/mSystems.00194-17.

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

587 9. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and
588 microbial differential abundance strategies depend upon data characteristics. Microbiome.
589 2017;5(1):27. doi: https://doi.org/10.1186/s40168-017-0237-y.

10. Paliy O, Shankar V. Application of multivariate statistical techniques in microbial ecology. Mol Ecol. 2016;25(5):1032-57. doi: https://doi.org/10.1111/mec.13536.

11. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, et al. Balance trees reveal microbial niche differentiation. mSystems. 2017;2(1):e00162-16. doi: https://doi.org/10.1128/mSystems.00162-16.

12. Wang S, Cai TT, Li H. Hypothesis testing for phylogenetic composition: a minimum-cost flow perspective. Biometrika. 2021;108(1):17-36. doi: https://doi.org/10.1093/biomet/asaa061.

13. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 2001;26(1):32-46. doi: https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.

14. Anderson MJ, Walsh DC. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? Ecol Monogr. 2013;83(4):557-74. doi: https://doi.org/10.1890/12-2010.1.

15. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Aust J Ecol. 1993;18(1):117-43. doi: https://doi.org/10.1111/j.1442-9993.1993.tb00438.x.

16. Schroeder PJ, Jenkins DG. How robust are popular beta diversity indices to sampling error? Ecosphere. 2018;9(2):e02100. doi: https://doi.org/10.1002/ecs2.2100.

17. Wong RG, Wu JR, Gloor GB. Expanding the UniFrac toolbox. PLoS One. 2016;11(9):e0161196. doi: https://doi.org/10.1371/journal.pone.0161196.

18. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12):8228-35. doi: https://doi.org/10.1128/AEM.71.12.8228-8235.2005.

19. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011;5(2):169-72. doi: https://doi.org/10.1038/ismej.2010.133.

20. Roldán Ahumada JA, Avendaño Garrido ML. A commentary on diversity measures UniFrac in very small sample size. Evol Bioinform. 2019;15:1176934319843515. doi: https://doi.org/10.1177%2F1176934319843515.

21. Schloss PD. Evaluating different approaches that test whether microbial communities have the same structure. ISME J. 2008;2(3):265-75. doi: https://doi.org/10.1038/ismej.2008.5.

620 22. Warton DI, Wright ST, Wang Y. Distance-based multivariate analyses confound
621 location and dispersion effects. Methods Ecol Evol. 2012;3(1):89-101. doi:
622 https://doi.org/10.1111/j.2041-210X.2011.00127.x.

623 23. Randolph TW, Zhao S, Copeland W, Hullar M, Shojaie A. Kernel-penalized regression
624 for analysis of microbiome data. Ann Appl Stat. 2018;12(1):540-66. doi:
625 https://dx.doi.org/10.1214/17-AOAS1102.

626 24. Tang Z-Z, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial
627 community composition that accommodates confounders and multiple distances.
628 Bioinformatics. 2016;32(17):2618-25. doi: https://doi.org/10.1093/bioinformatics/btw311.

629 25. Wu C, Chen J, Kim J, Pan W. An adaptive association test for microbiome data. Genome
630 Med. 2016;8(1):56. doi: https://doi.org/10.1186/s13073-016-0302-3.

631 26. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB.
632 Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S
633 rRNA gene sequencing and selective growth experiments by compositional data analysis.
634 Microbiome. 2014;2(1):15. doi: https://doi.org/10.1186/2049-2618-2-15.

635 27. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are
636 compositional: and this is not optional. Front Microbiol. 2017;8:2224. doi:
637 https://doi.org/10.3389/fmicb.2017.02224.

638 28. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of
639 composition of microbiomes: a novel method for studying microbial composition. Microb Ecol
640 Health Dis. 2015;26(1):27663. doi: https://doi.org/10.3402/mehd.v26.27663.

641 29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for
642 differential expression analysis of digital gene expression data. Bioinformatics.
643 2010;26(1):139-40. doi: https://doi.org/10.1093/bioinformatics/btp616.

644 30. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
645 RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1-21. doi:
646 https://doi.org/10.1186/s13059-014-0550-8.

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

649 32. Saccenti E, Hoefsloot HC, Smilde AK, Westerhuis JA, Hendriks MM. Reflections on
650 univariate and multivariate analysis of metabolomics data. Metabolomics. 2014;10(3):361-74.
651 doi: https://doi.org/10.1007/s11306-013-0598-6.

652 33. Smilde AK, Jansen JJ, Hoefsloot HC, Lamers R-JA, Van Der Greef J, Timmerman ME.
653 ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed
654 metabolomics data. Bioinformatics. 2005;21(13):3043-8. doi:
655 https://doi.org/10.1093/bioinformatics/bti476.

656 34. Stahle L, Wold S. Multivariate analysis of variance (MANOVA). Chemometr Intell Lab
657 1990;9:127-41. doi: https://doi.org/10.1016/0169-7439(90)80094-M.

658 35. Langsrud Ø. 50–50 multivariate analysis of variance for collinear responses. J R Stat
659 Soc - Ser D Stat. 2002;51(3):305-17. doi: https://doi.org/10.1111/1467-9884.00320.

660 36. Langsrud Ø. Rotation tests. Stat Comput. 2005;15(1):53-60. doi:
661 https://link.springer.com/content/pdf/10.1007/s11222-005-4789-5.pdf.

662 37. Combrink M, Du Preez I, Ronacher K, Walzl G, Loots DT. Time-dependent changes in
663 urinary metabolome before and after intensive phase tuberculosis therapy: a
664 pharmacometabolomics study. OMICS. 2019;23(11):560-72. doi:
665 https://doi.org/10.1089/omi.2019.0140.

666 38. Gómez-Canela C, Prats E, Lacorte S, Raldúa D, Piña B, Tauler R. Metabolomic changes
667 induced by nicotine in adult zebrafish skeletal muscle. Ecotox Environ Safe. 2018;164:388-97.
668 doi: https://doi.org/10.1016/j.ecoenv.2018.08.042.

669 39. Trimigno A, Khakimov B, Savorani F, Tenori L, Hendrixson V, Čivilis A, et al.
670 Investigation of Variations in the Human Urine Metabolome amongst European Populations:
671 An Exploratory Search for Biomarkers of People at Risk-of-Poverty. Mol Nutr Food Res.
672 2019;63(1):1800216. doi: https://doi.org/10.1002/mnfr.201800216.

673 40. Bjerke GA, Rudi K, Avershina E, Moen B, Blom H, Axelsson L. Exploring the brine
674 microbiota of a traditional Norwegian fermented fish product (Rakfisk) from six different
675 producers during two consecutive seasonal productions. Foods. 2019;8(2):72. doi:
676 https://doi.org/10.3390/foods8020072.

677 41. Måge I, Steppeler C, Berget I, Paulsen JE, Rud I. Multi-way methods for understanding
678 longitudinal intervention effects on bacterial communities. bioRxiv: 363630v1 [Preprint].
679 2018:[cited 1 Sept 2021]. Available from: https://www.biorxiv.org/content/10.1101/363630v1.

680 42. Moen B, Henjum K, Måge I, Knutsen SH, Rud I, Hetland RB, et al. Effect of dietary
681 fibers on cecal microbiota and intestinal tumorigenesis in azoxymethane treated A/J Min/+
682 mice. PLoS One. 2016;11(5):e0155402. doi: https://doi.org/10.1371/journal.pone.0155402.

683 43. Moen B, Røssvoll E, Måge I, Møretrø T, Langsrud S. Microbiota formed on attached
684 stainless steel coupons correlates with the natural biofilm of the sink surface in domestic
685 kitchens. Can J Microbiol. 2016;62(2):148-60. doi: https://doi.org/10.1139/cjm-2015-0562.

686 44. Birkeland E, Gharagozlian S, Birkeland KI, Valeur J, Måge I, Rud I, et al. Prebiotic
687 effect of inulin-type fructans on faecal microbiota and short-chain fatty acids in type 2 diabetes:
688 a randomised controlled trial. Eur J Nutr. 2020;59:3325-38. doi:
689 https://doi.org/10.1007/s00394-020-02282-5.

690 45. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic
691 biomarker discovery and explanation. Genome Biol. 2011;12(6):R60. doi:
692 https://doi.org/10.1186/gb-2011-12-6-r60.

693 46. Chi C, Xue Y, Lv N, Hao Y, Liu R, Wang Y, et al. Longitudinal gut bacterial
694 colonization and its influencing factors of low birth weight infants during the first 3 months of
695 life. Front Microbiol. 2019;10:1105. doi: https://doi.org/10.3389/fmicb.2019.01105.

696 47. Girard C, Tromas N, Amyot M, Shapiro BJ. Gut microbiome of the canadian arctic
697 Inuit. mSphere. 2017;2(1):e00297-16. doi: https://doi.org/10.1128/mSphere.00297-16.

698 48. Lai Z-L, Tseng C-H, Ho HJ, Cheung CK, Lin J-Y, Chen Y-J, et al. Fecal microbiota
699 transplantation confers beneficial metabolic effects of diet and exercise on diet-induced obese
700 mice. Sci Rep. 2018;8(1):15625. doi: https://doi.org/10.1038/s41598-018-33893-y.

701 49. Le Sciellour M, Labussière E, Zemb O, Renaudeau D. Effect of dietary fiber content on
702 nutrient digestibility and fecal microbiota composition in growing-finishing pigs. PLoS One.
703 2018;13(10):e0206159. doi: https://doi.org/10.1371/journal.pone.0206159.

704 50. Wang Y, Ames NP, Tun HM, Tosh SM, Jones PJ, Khafipour E. High molecular weight
705 barley β-glucan alters gut microbiota toward reduced cardiovascular disease risk. Front
706 Microbiol. 2016;7:129. doi: https://doi.org/10.3389/fmicb.2016.00129.

707 51. Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, et
708 al. US immigration westernizes the human gut microbiome. Cell. 2018;175(4):962-72.e10. doi:
709 https://doi.org/10.1016/j.cell.2018.10.029.

710 52. Hughes RL, Kable ME, Marco M, Keim NL. The role of the gut microbiome in
711 predicting response to diet and the development of precision nutrition models. Part II: results.
712 Adv Nutr. 2019;10(6):979-98. doi: https://doi.org/10.1093/advances/nmz049.

713 53. Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, et
714 al. Daily sampling reveals personalized diet-microbiome associations in humans. Cell Host
715 Microbe. 2019;25(6):789-802. e5. doi: https://doi.org/10.1016/j.chom.2019.05.005.

716 54. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al.
717 Environment dominates over host genetics in shaping human gut microbiota. Nature.
718 2018;555(7695):210-5. doi: https://doi.org/10.1038/nature25973.

719 55. So D, Whelan K, Rossi M, Morrison M, Holtmann G, Kelly JT, et al. Dietary fiber
720 intervention on gut microbiota composition in healthy adults: a systematic review and meta-
721 analysis. Am J Clin Nutr. 2018;107(6):965-83. doi: https://doi.org/10.1093/ajcn/nqy041.

722 56. Nagpal R, Wang S, Solberg Woods LC, Seshie O, Chung ST, Shively CA, et al.
723 Comparative microbiome signatures and short-chain fatty acids in mouse, rat, non-human
724 primate, and human feces. Front Microbiol. 2018;9:2897. doi:
725 https://doi.org/10.3389/fmicb.2018.02897.

726 57. Baker DH. Animal models in nutrition research. Nutr J. 2008;138(2):391-6. doi:
727 https://doi.org/10.1093/jn/138.2.391.

728 58. Hui BS, Wold HOA. Consistency and consistency at large of partial least squares
729 estimates. In: Jöreskog KG, Wold HOA, editors. Systems under indirect observation, part II.
730 Amsterdam: North-Holland Publ. Co.; 1982. p. 119-30.

731 59. Schneeweiss H. Consistency at large in models with latent variables. In: Haagen K,
732 Bartholomew DJ, Deistler M, editors. Statistical modelling and latent variables. Amsterdam:
733 Elsevier; 1993. p. 299-320.

734 60. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ.
735 Centering, scaling, and transformations: improving the biological information content of
736 metabolomics data. BMC Genomics. 2006;7(1):142. doi: https://doi.org/10.1186/1471-2164-7-
737 142.

738 61. Thiel M, Féraud B, Govaerts B. ASCA+ and APCA+: Extensions of ASCA and APCA
739 in the analysis of unbalanced multifactorial designs. J Chemom. 2017;31(6):e2895. doi:
740 https://doi.org/10.1002/cem.2895.

741 62. Martin M, Govaerts B. LiMM-PCA: Combining ASCA+ and linear mixed models to
742 analyse high-dimensional designed data. J Chemom. 2020;34(6):e3232. doi:
743 https://doi.org/10.1002/cem.3232.

744 63. Hughes RL, Marco ML, Hughes JP, Keim NL, Kable ME. The role of the gut
745 microbiome in predicting response to diet and the development of precision nutrition models—
746 Part I: overview of current methods. Adv Nutr. 2019;10(6):953-78. doi:
747 https://doi.org/10.1093/advances/nmz022.

748 64. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential
749 abundance methods do not control the false discovery rate. Brief Bioinform. 2019;20(1):210-
750 21. doi: https://doi.org/10.1093/bib/bbx104.

751 65. Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, et al.
752 Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S
753 rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome.
754 2016;4(1):62. doi: https://doi.org/10.1186/s40168-016-0208-8.

755 66. Boulesteix AL, Binder H, Abrahamowicz M, Sauerbrei W. On the necessity and design
756 of studies comparing statistical methods. Biometrical J. 2017;60(1):216-8. doi:
757 https://doi.org/10.1002/bimj.201700129.

758 67. R Development Core Team. R: A language and environment for statistical computing.
759 R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 2021.

760 68. Patuzzi I, Baruzzo G, Losasso C, Ricci A, Di Camillo B. metaSPARSim: a 16S rRNA
761 gene sequencing count data simulator. BMC Bioinform. 2019;20(9):1-13. doi:
762 https://doi.org/10.1186/s12859-019-2882-6.

763 69. Aitchison J. The analysis of compositional data. London: Chapman and Hall; 1986. 416
764 p.

765 70. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for
766 multivariate imputation of left-censored data under a compositional approach. Chemometr
767 Intell Lab. 2015;143:85-96. doi: https://doi.org/10.1016/j.chemolab.2015.02.019.

768 71. Brill B, Amir A, Heller R. Testing for differential abundance in compositional counts
769 data, with application to microbiome studies. arXiv:190408937 [Preprint]. 2019:[cited 1 Sept
770 2021]. Available from: https://arxiv.org/pdf/1904.08937.pdf.

771 72. Costea PI, Zeller G, Sunagawa S, Bork P. A fair comparison. Nat Methods.
772 2014;11(4):359. doi: https://doi.org/10.1038/nmeth.2897.

773 73. Wang S. Robust differential abundance test in compositional data. arXiv: 210108765
774 [Preprint]. 2021:[cited 1 Sept 2021]. Available from: https://arxiv.org/pdf/101.08765.pdf.

74. Lubbe S, Filzmoser P, Templ M. Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemometr Intell Lab. 2021;210:104248. doi: https://doi.org/10.1016/j.chemolab.2021.104248.

75. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community ecology package (R package version 2.5-5) https://cran.r-project.org/ and https://github.com/vegandevs/vegan/. 2019.

76. Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C. Multi-and megavariate data analysis basic principles and applications. Malmö, Sweden: Umetrics Academy; 2013.

77. Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. Anal Chem. 1996;68(21):3851-8. doi: https://doi.org/10.1021/ac960321m.

Figures

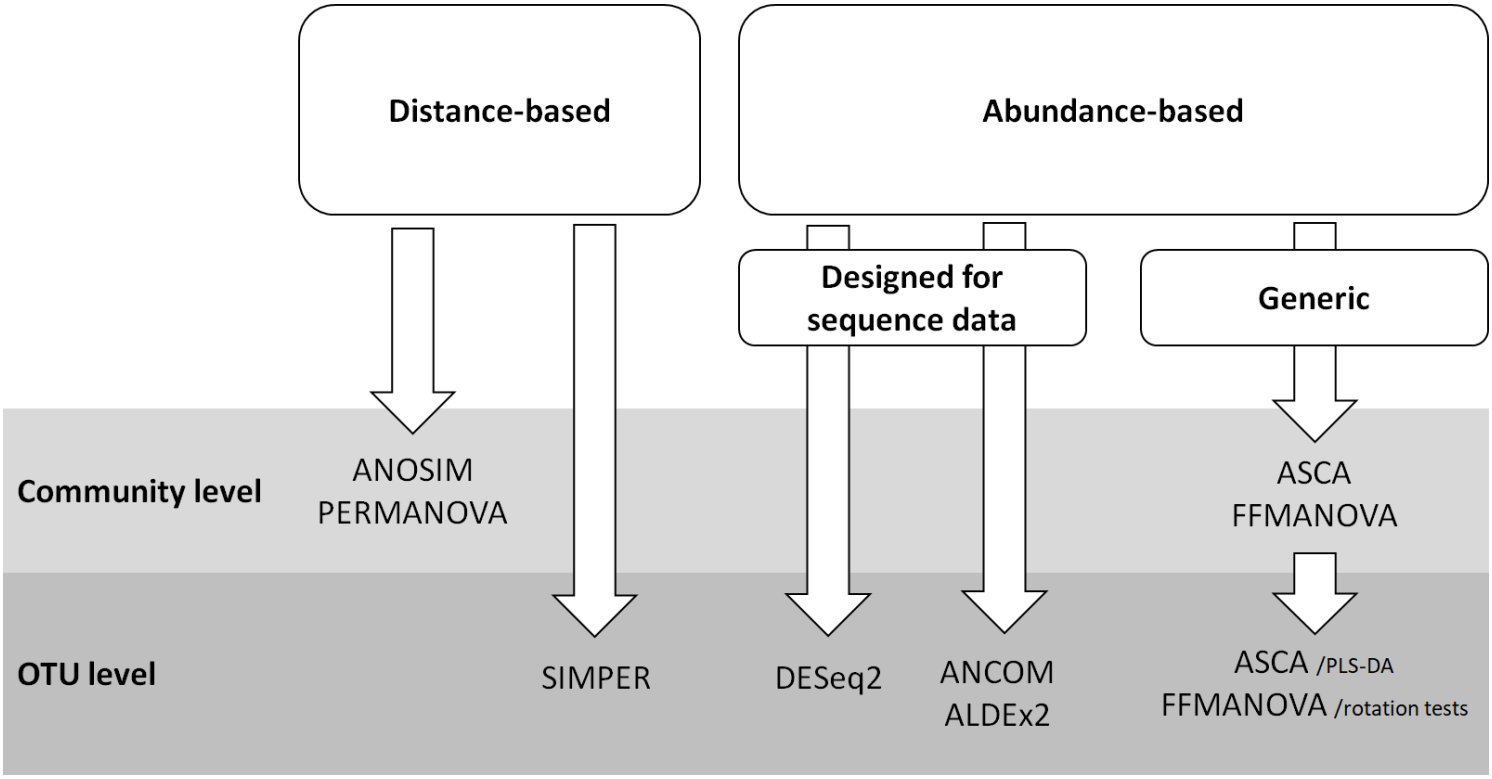

Figure 1

A diagram of statistical methods used in the study.

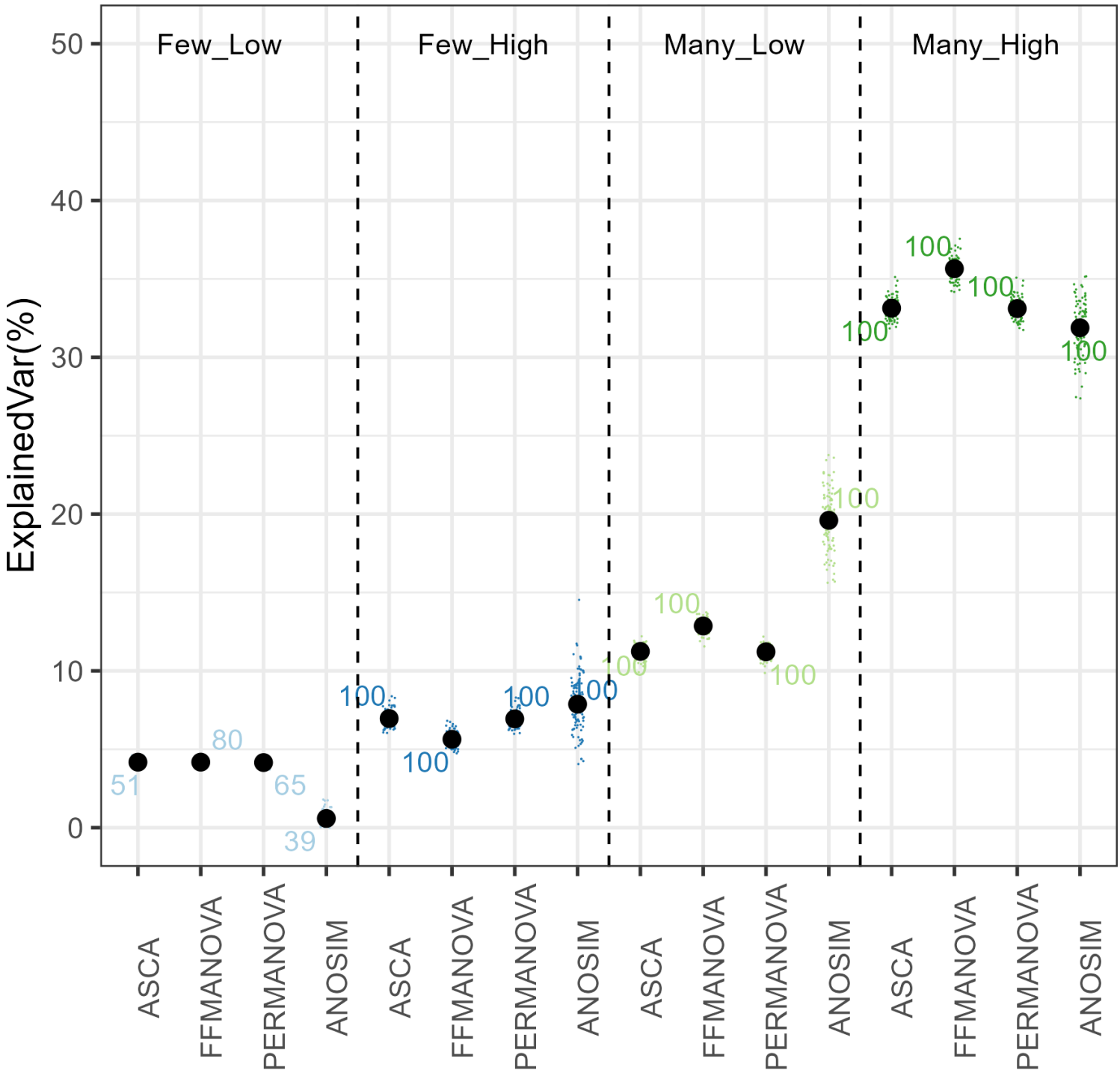

Figure 2

Explained variance for simulated data and the relative number of simulations where the simulated effect was detected.

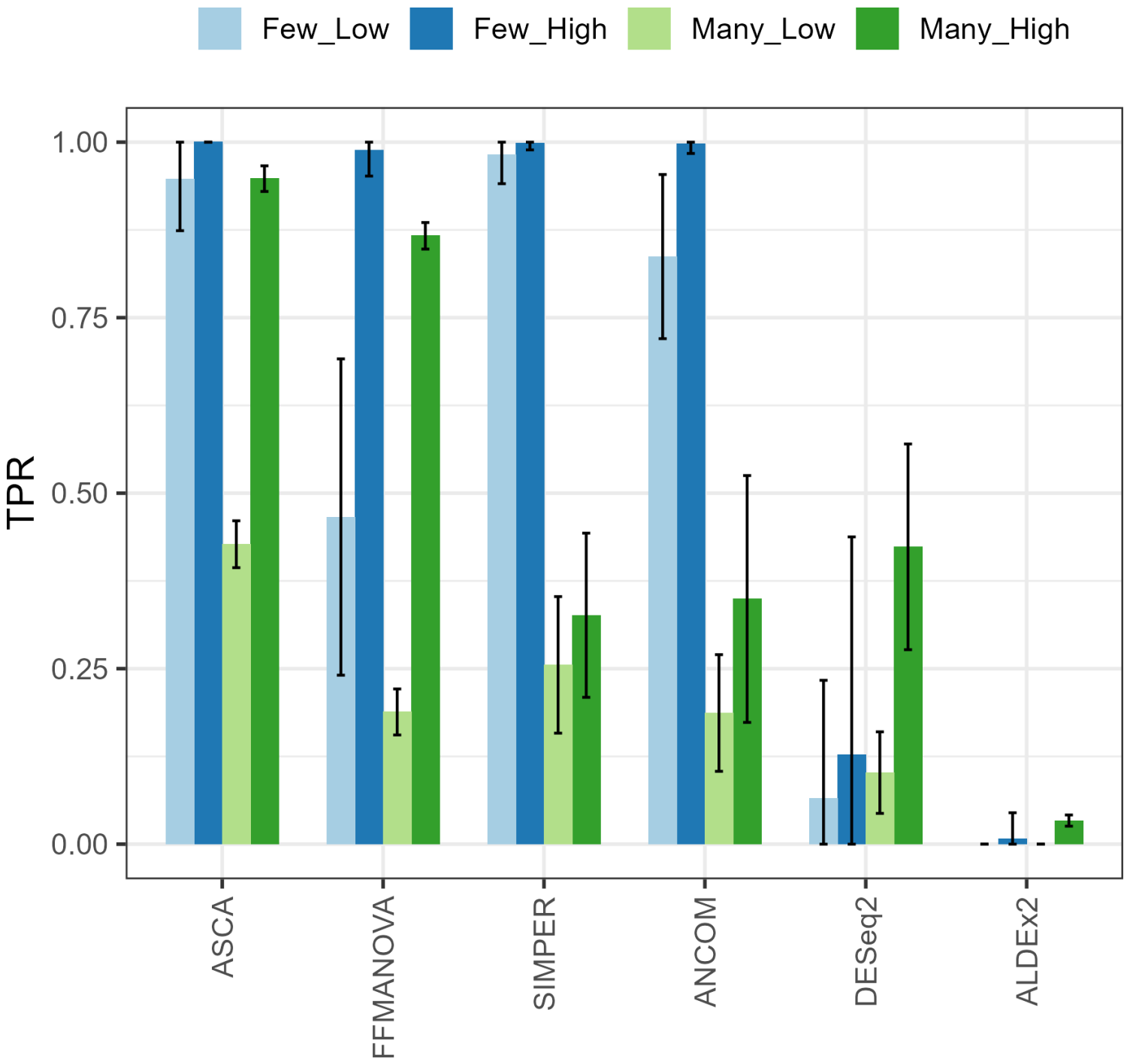

Figure 3

Sensitivity (True Positive Rate) for the four scenarios in the simulation study.

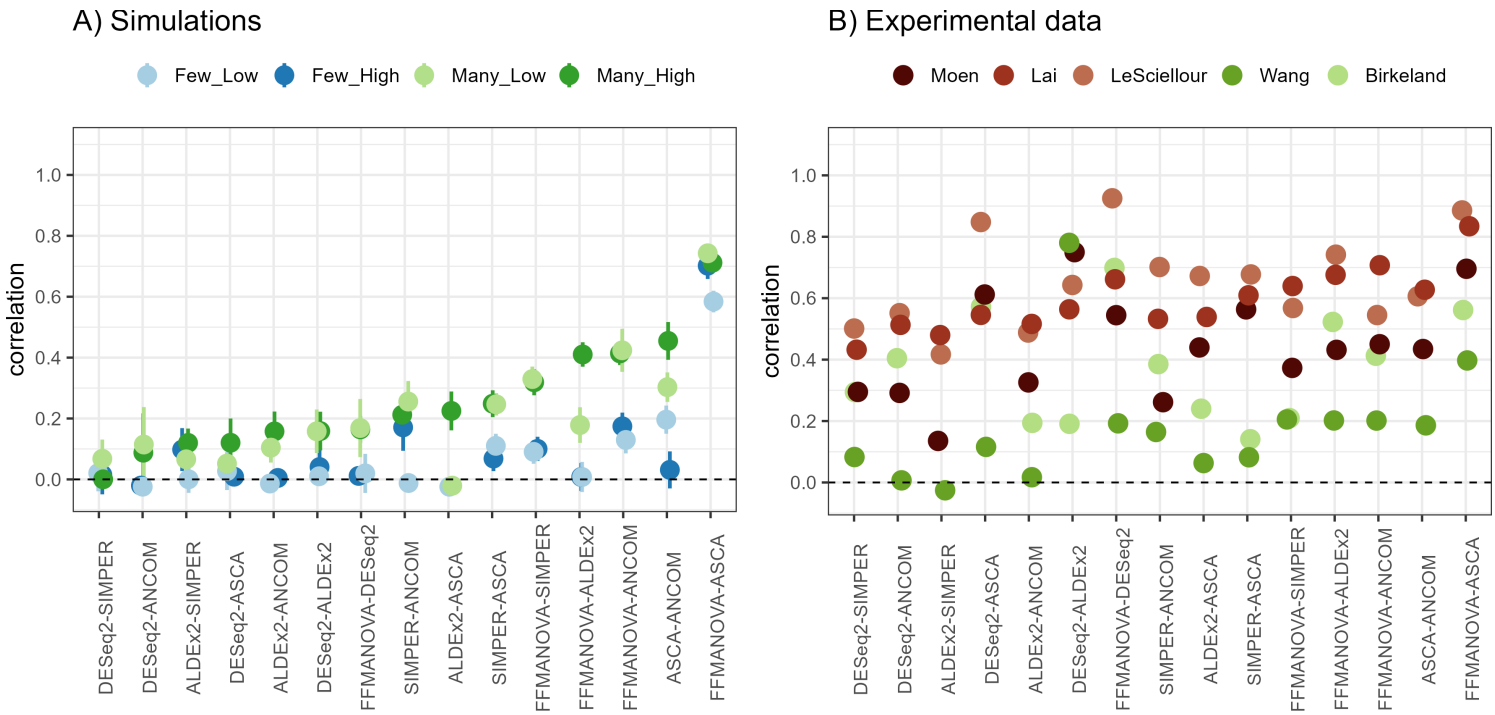

Figure 4

Spearman's correlation (Y-axis) calculated for pairwise comparison of statistical methods (X-axis) for (A) simulated data and (B) five experimental data sets. Each point represents Spearman's rank correlation coefficient between OTU ranking metrics from the two methods compared.

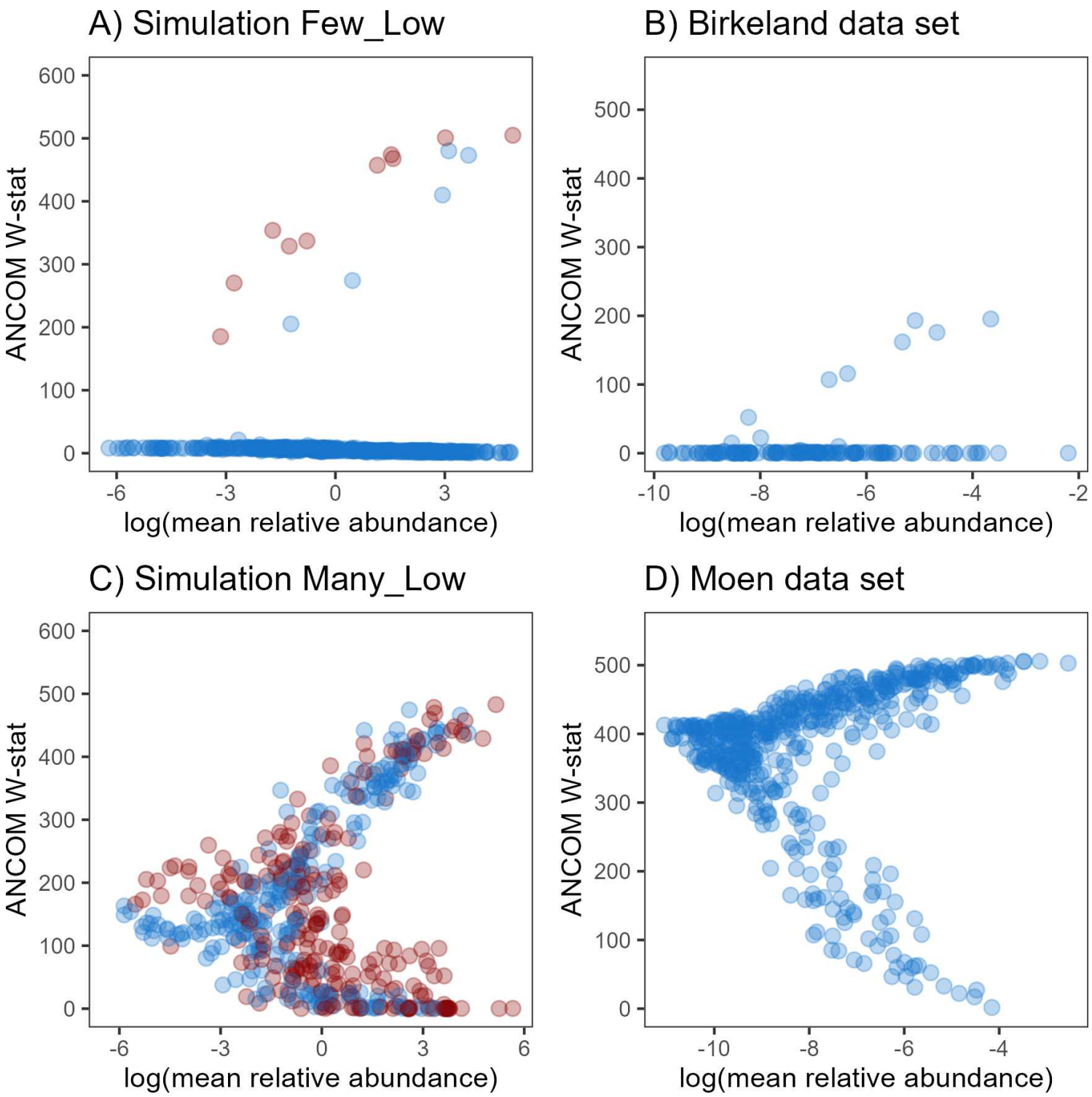

Figure 5

Mean relative abundance (log-scale) plotted versus ANCOM W-stat. (A) The “Few-Low” simulation scenario and (B) Birkeland data set, (C) “Many-Low” simulation scenario and (D) Moen data set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- KHOMICHetal2021preprintSupportingInformation.pdf
- S1Fig.pdf
- S1File.xlsx
- S1Table.xlsx
- S2Fig.pdf
- S2File.xlsx
- S2Table.xlsx
- S3File.rdata
- S4File.r
- S5File.r