Comprehensive evaluation of statistical approaches for differential metaproteomics

Authors:

Tjorven Hinzke[*#1,2,3], Benoit J. Kunath[*#3,4,5], J. Alfredo Blakeley-Ruiz[3], Abigail Korenek[3], Simina Vintila[3], Paul Wilmes[4,6], Manuel Kleiner[#3]

*contributed equally

Corresponding authors: tjorven.hinzke@uni-greifswald.de, benoit.kunath@lih.lu, manuel_kleiner@ncsu.edu

[1] University of Greifswald, partner of the Greifswald Mire Centre, Greifswald, Germany
[2] Helmholtz Institute for One Health, Greifswald, Germany
[3] Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA
[4] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[5] Multiomics Data Science, Department of Cancer Research, Luxembourg Institute of Health, Strassen, Luxembourg
[6] Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Key words:
Quantitative proteomics, regression, Bayesian statistics, generalized linear mixed models, edgeR, limma, microbiome, holobiont, microbial community

Abstract

Background

Metaproteomics characterizes and compares molecular phenotypes of organisms in communities by comprehensively analyzing their protein expression profiles using statistical methods. However, not all statistical methods are suitable for determining differentially abundant protein groups in metaproteomic analyses. Statistical challenges in metaproteomics include: data sparsity, non-normality, compositionality, and large between-sample variability. These challenges can potentially be addressed with several data processing steps, including imputation, normalization, transformation, and selection of the appropriate statistical tests. The potential combinations of different processing methods create a complex matrix of analysis options and it is currently unclear how these combinations impact the results of statistical tests on metaproteomic data.

Results

To determine what data processing methods and statistical tests are best for identifying differentially abundant proteins in metaproteomics datasets, we generated a set of thirteen metaproteomic samples with known compositions, known differences, and differing levels of complexity. These defined metaproteomes address the general challenges outlined above, using various scenarios in metaproteomic data analyses. We compared over 110 different statistical analysis combination options, including regression-based tools, general statistics inference, and machine learning techniques. We found that several combinations within the frameworks of limma, edgeR, MaAslin2, custom linear and Bayesian linear models, and random forests all offer suitable evaluation options.

Conclusions

We highlight key recommendations for differential expression analysis in metaproteomics. Our work enables improved assessment of statistical methods for metaproteomics by establishing a framework for testing statistical approaches, including comprehensive raw mass spectrometry data and reproducible benchmarking code.

Background

Metaproteomics is the term used for approaches that comprehensively characterize gene expression in microbiomes at the protein level (Kleiner 2019; Van Den Bossche et al. 2025; Wilmes and Bond 2004). Metaproteomics usually involves high-resolution liquid chromatography and mass spectrometry methods to identify and quantify tens of thousands of peptides that are then used to identify and quantify thousands of proteins in each sample. Based on protein abundances that differ between samples from different environments, treatments, or conditions, we can determine how metabolic and physiological processes in the microbiome respond to situations such as changing diet in the gut (Blakeley-Ruiz et al. 2022; 2025; Levi Mortera et al. 2024), ecological succession in acid mine drainage (Mueller et al. 2011), organic matter availability at different depths in the ocean (Bergauer et al. 2018), or host-symbiont and host-pathogen interactions (Abbondio et al. 2023; Gruber-Vodicka et al. 2019).

Despite the successful application of metaproteomics to samples from many environments there is currently no consensus on how to determine which proteins are actually differentially abundant in different groups of samples. In fact, this consensus is even lacking for regular single-organism proteomics (Langley and Mayr 2015; Wolski et al. 2023; Yang et al. 2022; Zhu et al. 2020). Additionally, it remains to be shown whether statistical tools used for proteomics, or other -omics, work as expected for complex metaproteomic analyses.

Statistical data analysis is rendered challenging by several features of metaproteomics data (please refer to the Box for details). These features and consequential challenges include: different possible protein abundance measures, i.e., spectral counts (SpC), or area for chromatographic peaks of peptides from MS ion currents (AUC), data sparsity and concomitant imputation, batch effects and non-Gaussian data distribution, and compositionality. After taking these challenges and the resulting data pre-processing steps into account, there is still the open question of the statistical test or tool for inferring differentially abundant proteins. Comparative evaluations of statistical tools for (meta-)proteomics are mostly based on the use of spike-in proteins (Malinowska et al. 2012; Pursiheimo et al. 2015; Ramus et al. 2016), which, while undoubtedly useful, cannot fully capture the complexity of environmental metaproteomes. Alternatively, evaluations use experimental data without a known ground truth, and/or simulated data (Langley and Mayr 2015; Li et al. 2010; Pursiheimo et al. 2015), meaning that their direct transferability is still unknown. Likewise, approaches for evaluating statistical tools for other omics- and microbiome techniques include mainly experimental, downsampled, or simulated data (e.g., Calgaro et al. 2020; Jonsson et al. 2016; Nearing et al. 2022; Weiss et al. 2017). Following from this, the use of complex samples of controlled composition (mock communities or ground truth samples) for statistical method evaluation is largely missing.

Here, we generated complex defined metaproteomes with known compositions and measured them with a standard metaproteomics workflow to use as ground-truth data. We designed our mock communities to address various challenges of metaproteomics data analysis: (a) small, but biologically important, changes in protein abundances, (b) very large changes in total protein

abundances of specific species in the microbiome, and (c) potential misassignment of identified proteins to the incorrect species when sequences are very similar. We measured these communities with a data-dependent acquisition (DDA) metaproteomics workflow to use as ground-truth data. We subsequently used the metaproteomic data from these defined metaproteomes to compare various data preparation and statistical inference methods using both SpC and AUC quantification methods. We tested various commonly used statistical tools used in (meta)omics and microbial community differential abundance analysis (e.g., edgeR, DESEq2, limma; Chen et al. 2025; Love et al. 2014; Ritchie et al. 2015) as well as general statistical inference methods (t-test, Wilcoxon), machine learning (random forests; Breiman 2001), and Bayesian statistics-based approaches with various combinations of normalization and/or transformation approaches. This framework for testing statstical approaches can be used in the future to expand to other data acquisition approaches such as data-indepenendent mass spectrometry (DIA).

———————————————————

Box

Statistical terminology

Discrete and continuous data: Discrete data are represented as whole numbers (integers), e.g., count data. Continuous data, on the other hand are fractional and require decimals (float).

Compositional data: Compositional data are proportions or percentage data, or can be represented as such, i.e., data relative to a total. Proteomics mass spectrometry data are inherently compositional: for example, a cell is made up of 10 % protein A, 5 % protein B, etc. In addition to this "biological compositionality", the mass spectrometry data acquisition introduces compositionality, because there is an upper limit to the amount of data that can be acquired. It has been advocated that specific methods have to be used to address this feature of compositionality (Gloor et al. 2017).

Data sparsity and imputation: Especially in complex metaproteomic datasets, not all proteins present in the sample, and certainly not all peptides of specific proteins, will be detected in all conditions, especially when proteins are lowly abundant. Some statistical procedures cannot deal with missing data or zeros, e.g., because they are using a log transformation (and the log of zero is not defined). Imputation, i.e., replacement of missing values or of zero by some specific value (or values, in the case of multiple imputation), can be used to address missing values. Multiple different imputation methods exist, ranging from replacement with a small constant (e.g., the smallest number possible in the dataset or a fraction thereof), to complex statistical procedures, and their use depends on the assumption of whether the data is missing at random (MAR) or missing not at random (MNAR), or consists of a MAR/MNAR mixture (Lazar et al. 2016). Alternatively, imputation-free models, which take the pattern of missing values into account (Plancade et al. 2022) may be used, but missingness patterns are informative only for a higher number of replicates.

Normalization vs Transformation: Normalization is used to make data comparable between MS runs, batches, etc. For example, if different amounts of peptides were injected for the respective MS runs, one would expect this to cause differences in protein group abundances. Consequently, these effects have to be corrected. For that, a multitude of normalization methods exist (e.g., Välikangas et al. 2016), and the method used will impact the outcome of differential expression analysis. Transformations, on the other hand, convert the data to adhere to a certain data distribution. Statistical tests often operate under the assumptions of a specific data distribution (e.g., Gaussian, Poisson, etc.). To meet these assumptions, data can be transformed (e.g., log, square root, etc.). Different transformations can be suitable for the same data, and they will impact downstream statistical inference.

Type I error: Detection of false positives, i.e., a variable (protein group) is determined by a test as being significiantly different between groups, whereas, in fact, it is not.

Type II error: Detection of false negatives, i.e., a variable (protein group) is determined by a test as not being significantly different between groups, whereas, in fact, it is.

Power of a statistical test: Probability for a test to detect a true significant difference.

Overdispersion: Overdispersion means that the variability in the data is greater than that expected by a given model. This happens for example for Poisson models: in the Poisson distribution, the mean equals the variance. But in real-life count data, the variance is often larger (overdispersed).

Statistical tests

Parametric vs non-parametric tests: Parametric tests, such as the t-test, assume that the data stem from a specific underlying distribution, e.g., that they were sampled from a Gaussian normal distribution. The "parameters" in a normal distribution would, for example, be the mean and variance, which define the exact shape of the distribution. Non-parametric tests, such as Wilcoxon's rank-sum test, do not assume a certain distribution, i.e., they have no parameters (hence non-parametric). This generally results in lower power of non-parametric tests and thus a lower ability to actually detect existing statistical differences. The assumption of independence of the samples being compared holds true for both parametric and non-parametric tests. While samples need to be independent, correlated data (e.g., tests on the same subjects, or time series) necessitates adapted approaches explicitly taking into account these correlations (e.g., paired tests, or mixed models).

Generalized linear mixed model (GLMM): At the core, GLMMs are linear regression models. They assume a linear relationship between the predictor variables (e.g., the protein group abundance) and the response variable (e.g., the environmental or experimental condition). However, the linearity assumption cannot hold true when, for example, the response is constrained between 0 and 1. Here, the "generalized" comes into play: it refers to a link function, which is a function of the response variable that is linearly correlated with the predictors – for example, the log. The "mixed" refers to the inclusion of random effects in addition to fixed effects as predictors. Random effects can be batches of samples that are sampled at a specific time point or extracted together, or subjects in the case of repeated measurements, but also proteins. Mixed models can thus inherently incorporate non-independence between observations.

Frequentist vs Bayesian statistics: Frequentist statistics refers to "classical" null hypothesis testing, most often done by many repetitions of an experiment, i.e., many independent replicates. Bayesian statistics, on the other hand, offer a mathematical framework to update prior knowledge with data, generating posterior knowledge. For example, if the expected range of standard errors of protein abundance measurements is known from previous experiments, this can be explicitly used as a "prior". One difference between both (combinable) approaches lies in the interpretation of the uncertainty estimations: a frequentist 95% confidence interval holds the true population value 95% of the time when repeating an experiment under the same conditions. A Bayesian credibility interval, on the other hand, is the range in which the true value is expected to be at 95% probability.

Random forests: Random forests (Breiman 2001) are a supervised machine learning technique, which can be used to infer to which class a sample belongs (or, in the case of a quantitative response, be used for regression), and to determine which variables drive the distinction between classes. It uses bootstrapped sample subsets (i.e., samples are drawn with replacement), and for classification splits these subsets to generate nodes with as little mixture between pre-assigned classes as possible. For these splits, a random subset of variables (here: protein groups) are used. The full split is referred to as a decision tree. The final result is the majority vote of the trees. As random forests can be used for high-dimensional data, do not easily overfit and render variable importance measures, they are well suited for meta-omics datasets (Díaz-Uriarte and Alvarez De Andrés 2006).

General considerations for statistics in metaproteomics

Protein quantification in (meta-)proteomics: Spectral counts, also known as peptide spectrum matches (Spc or PSMs), are the number of peptide fragment spectra (MS/MS or MS2) that are matched to a peptide sequence. The alternative to spectral counting is the extraction of peptide ion intensity data from the LC-MS/MS data (extracted ion current, XIC). This data can be used for intensity-based quantification by determining the maximum intensity for specific peptide ions in the XIC, or by integration of the chromatographic peak area for a specific peptide ion (area under the curve, AUC). Depending on the quantification method used either count data (SpC) or continuous data (AUC) are generated, which can necessitate different data transformations.

Data exploration: For (meta-)proteomics, common data exploration steps include hierarchical clustering and (N)MDS plots, to examine whether the data support the expectations, and, more importantly, the experimental design. We expect these steps to be done prior to applying any statistics, see for example (Van Den Bossche et al. 2025) for details.

[End of box]

———————————————————————-

Methods

Design and generation of defined metaproteomes

Our defined metaproteomes were mixed at the peptide level and consisted of defined combinations of pure culture peptides with background peptide mixtures ("matrices" in Table 1). All defined metaproteomes were generated in biological quadruplicates. Also, each pure culture produced in biological quadruplicates for each condition (Table 1, Supplementary Table 1). The first of three background matrices consisted of peptides from stool sampled from mice with a conventional microbiota (Blakeley-Ruiz et al. 2025), simulating a sample with a high level of complexity. The second matrix was generated from corn roots that were grown under sterile conditions, and were therefore of low complexity. The last matrix consisted of stool sampled from gnotobiotic mice with a defined gut microbiome of 13 species. This last matrix provided a medium level of complexity. For the conventional and gnotobiotic mice samples, NC State's Institutional Animal Care and Use Committee approved all experimental activity (Protocol # 18-034-B and 18-165-B). Neither euthanasia nor anesthesia were used in the collection of these samples.

After proteins were extracted and digested, and peptide concentrations were determined as detailed in the Supplementary Methods, we mixed the resulting peptides of pure cultures in different known quantities with peptides extracted from one of the three matrices to create the final defined metaproteomes with known composition (see Results and Discussion Figure 1).

Table 1: Pure cultures and matrices used for generating defined metaproteomes.

Pure cultures		
Name	**Description**	**Culturing conditions**
Thermus thermophilus	Gram-negative bacterium; not present in mouse microbiome	High and low temperatures
Chlamydomonas reinhardtii	Eukaryote; not present in mouse microbiome	High and low light
Rhizobium leguminosarum bv. *viciae* 3841	Gram-negative bacterium; not present in mouse microbiome, plant symbiont	Strain comparison with *R. leguminosarum* VF39
Rhizobium leguminosarum bv. *viciae* VF39	Gram-negative bacterium; not present in mouse microbiome, plant symbiont	Strain comparison with *R. leguminosarum* 3841
Escherichia coli	Gram-negative bacterium; potentially present in mouse microbiome	Complex (LB) and minimal media (M9)
Bacteroides thetaiotaomicron	Gram-negative bacterium; present in mouse microbiome	Single condition
Matrices		
Name	**Type**	**Rationale**
Mouse faecal pellets	High complexity	Most similar to real experimental conditions and challenges
Sterile corn roots	Low complexity	Low protein diversity; similar to situations observed, e.g., in symbioses
Gnotobiotic mouse faecal pellets	Medium complexity	Analysis of background complexity effects by comparison with complex mouse stool

Metaproteomic database generation

The metaproteomic databases were assembled as a combination of "building blocks," which included proteins expected to be found in the pure cultures, predicted mouse or maize proteomes, the matrix microbial translated metagenome, mouse dietary proteins (when applicable), and common laboratory contaminants (Supplementary Table 2). For the generic mouse stool matrix, DNA extraction and metagenomic analysis were performed (see Supplementary Methods) in order to generate a sample-specific database, as required for the metaproteomic analysis (Blakeley-Ruiz and Kleiner 2022). Each building block was clustered at 95% sequence identity using CD-HIT and then combined for the different mixes. Please refer to the Supplementary Methods for details.

LC-MS/MS measurements and analyses

Peptides from the pure cultures and the different mixes were analyzed using a nanoLC-MS/MS system consisting of a Dionex UltiMate 3000 RSLCnano (Thermo Scientific) connected to an

Orbitrap Exploris 480 (Thermo Scientific) equipped with an EASY-Spray source (Thermo). Samples were run in randomized block design. For each run, we loaded 1 µg of peptides per sample onto a trap column (Acclaim PepMap100, C18, 5 µm, 100 Å, 0.3 mm i.d. ×5 mm, ThermoScientific) and backflushed onto a 75 cm analytical column (EASY-Spray C18, 2 µm, 100 Å, 75 µm i.d., Thermo Scientific). Peptides were separated on the analytical column using a 140 min gradient of 95% eluent A [0.1% (v/v) formic acid], 5% eluent B [80% (v/v) acetonitrile, 0.1% (v/v) formic acid] to 31% (v/v) B in 102 min, 31 to 50% (v/v) B in 18 min, and finally to 99% B for 20 min at a flow rate of 300 nl/min. The mass spectrometer was operated in data-dependent mode and the 15 most intense peptide precursor ions were selected for fragmentation and MS/MS acquisition. The selected precursor ions were then excluded from repeated fragmentation for 25 s. The resolution was set to R = 60,000 and R = 15,000 for MS and MS/MS, respectively. As lock mass we used the ambient ion 445.12 m/z.

Detection of differential abundances

We used R 4.3.1 (R Core Team 2023). For details on package versions, please refer to Supplementary Table 3.

Data preprocessing, filtering, and imputation

We analyzed differential protein abundance in samples by comparing one defined metaproteome vs. another, within the same matrix. These defined metaproteomes simulate independent experimental conditions, environmental conditions, spatial sampling points, or other groups of samples to be compared. We used master proteins for our analysis. Master proteins are representative sequences chosen for protein groups, where protein groups are one or multiple protein sequences which share a set of peptide-spectrum matches (Nesvizhskii and Aebersold 2005; Van Den Bossche et al. 2025). Selection of master proteins for specific protein groups can differ in pure cultures vs. defined metaproteomes, leading to empty matches and thereby inaccurate comparisons. We addressed this issue by matching on a per-protein level: only one protein of the pure culture protein group needed to be present in a defined metaproteome protein group for matching. Sparsity in the dataset was reduced by filtering the SpC data for most implementations to retain only proteins that had at least 5 PSMs in at least 2 out of 8 samples. AUC data were initially filtered to arrive at approximately the same number of proteins as for SpC, to make the two quantification approaches comparable. The effect of filtering was subsequently assessed with selected statistical methods (see Supplementary Results and Supplementary Figure 1). We imputed missing values with 0 or, where necessary, with a small constant (1/5th of the smallest value in the dataset). For random forests that use clr transformed data, we also tested the zcompositions count imputation (Palarea-Albaladejo and Martín-Fernández 2015).

Statistical analyses

Regression-based tests

(Bayesian) (Generalized) linear mixed model

For (generalized) linear regression, we used a random effects only-model:

$$\text{value} \sim (1 + \text{condition} \mid \text{protein})$$

We implemented Gaussian regression with transformed data, as well as Poisson and negative binomial regression in lme4 v. 1.1.34 (Bates et al. 2015). Significant differences were defined as differences where the interval of the slope of the random effect ± 1.96 the standard deviation did not include 0 (i.e., where the 95% confidence interval of the random slope was strictly positive or strictly negative). Few linear model evaluations produced boundary fits.

For the Bayesian approach, we used brms (Bürkner 2018) with the same model formula as above. We used 3 chains, 6000 iterations with 2000 of these assigned to warmup, and chain thinning of 10. We assigned significance if the 2.5 to 97.5% credibility interval of the conditional random effect excluded zero. We tested Gaussian, as well as exponentially modified Gaussian, students, and skew-normal distributions, with different normalizations. We mostly used the specified default priors, aside from a model with Gaussian distribution and chiP (lambda=0.9) normalization and transformation, where we also adapted priors. In this case, we modelled, with a Student's t-distribution, the intercept (3,0,0.1), residual standard deviation sigma (3,0,1), and standard deviation of the group-level (random) effect (3,0,2). For SpC data, brms evaluations in matrix 3 produced low effective sample sizes (ESS), with the exception of brms_gauss_chiP09_prior, i.e., adaptation of priors was necessary to achieve sufficiently fast convergence.

Corncob

The R package corncob (Martin et al. 2020) uses a beta-binomial regression model, which was developed for differential abundance analysis of taxa in microbial communities. It explicitly takes overdispersion into account, thus allowing for testing differential relative abundance together with differential variance. For that, it uses a logit link, and directly uses absolute abundances and total counts in the model, instead of normalizing beforehand. FDR correction is implemented in the algorithm. We used corncob v. 0.3.1 (Martin et al. 2022).

DESeq2 (negative binomial)

DESeq2 (Love et al. 2014), originally developed for RNA-Seq data, models data with a negative binomial distribution, which can be formulated as a mixture of gamma and Poisson distributions. It normalizes for different sequencing depths and uses a logarithmic link in its generalized linear model. A shrinkage of dispersion estimates within groups is also employed, with stronger shrinkage for lower information-containing variables. DESeq2 has FDR calculation implemented.

edgeR 4.0

EdgeR (Chen et al. 2025) uses a negative binomial distribution for counts, and can accommodate different experimental designs via generalized linear models. It employs empirical Bayes moderation for outlier treatments, i.e., the dispersions are estimated from the data, and variations moderated towards the common trend (Chen et al. 2025). The newest version edgeR v4.0 uses unbiased quasi-likelihood dispersion estimates (QLE) also for small counts. We used the FDR calculation as implemented.

Limma

Limma (Ritchie et al. 2015), originally developed for microarray data (= continuous intensity data) and extended to accommodate RNAseq data (= count data), uses a variable-wise linear model with global parameters to share information across genes and samples. An empirical Bayes moderation is used to moderate the residual variance. Limma has FDR calculation implemented.

MaAsLin2

MaAsLin2 (Mallick et al. 2021) uses generalized linear mixed models, with the option to use different model specifications, as well as normalizations and transformations. It accommodates different experimental designs and the inclusion of metadata. We used the implemented FDR value calculation.

Ensemble machine learning

Random forest

For random forests (Breiman 2001), we used the variable importance measure of Janitza et al. (Janitza et al. 2018). Additionally, we aggregated results over forests per comparison to increase stability. We used 500 forests and 75,000 trees per forest when not normalizing at the species level, and 500 forests with 50,000 trees per forest when normalizing at the species level in the defined metaproteomes, as well as in the pure cultures. We assigned protein groups as significantly differentially abundant when the mode of the p-value density distribution of forests was below 0.05, or when the median of p-values of all forests for one comparison was below 0.05. Monotonic transformations like a log transfomation will not impact the random forest results, as the splits are the same. Some random forest p-value estimations were inaccurate due to too few non-negative values (the alternative Altmann approach for p-value estimation was prohibitively slow and therefore not used). Each random forest result was aggregated over forests as well as over trees to ameliorate p-value issues.

Null hypothesis testing

Welch's t-test

While the Student's t-test assumes normality and homogeneity of variances, Welch's t-test approximation does not require homogeneity of variances (Welch 1947). As Welch's t-test performs better type I error control with heterogenous variances and looses little power (if any)

as compared to the Student's t-test, it is recommended to be used as default instead of the Student's t-test (Delacre et al. 2017). We calculated Benjamini-Hochberg corrected FDR (Benjamini and Hochberg 2000; Benjamini and Yekutieli 2001) values from p-values using adjust_pvalue in rstatix (Kassambara 2023).

Wilcoxon rank-sum test

The Wilcoxon rank-sum test is based on ranking observations of both groups by their abundance and then summing these ranks per group. The Wilcoxon rank-sum test does not make assumptions about the data distribution and is thus a non-parametric test to compare two conditions (Wilcoxon 1945). FDR values were calculated from p-values using adjust_pvalue in rstatix (Kassambara 2023).

Approaches not included

Some approaches were initially tested, but not included in the final analyses. XGBoost (Chen and Guestrin 2016) needed prohibitively complex hyperparameter tuning in order to give meaningful results. ALDEx2 (Fernandes et al. 2014) is known to have low power (Calgaro et al. 2020), an observation we also made during testing. An explicit scale model, as very recently published (Nixon et al. 2025), was not tested, and is not expected to ameliorate low power issues in our few-replicate setting, especially as it was noted that scale uncertainty inclusion decreases sensitivity (Gloor et al. 2025). Likewise, the experimental layout we describe here, i.e., few replicates per condition, does not lend itself to neural network-based machine learning approaches.

Results and Discussion

Design of defined metaproteomes and statistics evaluation framework

We designed defined metaproteomes with varying degrees of complexity to evaluate differential expression analysis statistical tools and methods. We built these defined metaproteomes by mixing pure culture peptides with one of three complex matrix peptides in defined amounts. For the matrices, we chose mouse stool (high complexity), sterile corn roots (low complexity) and gnotobiotic mouse stool (medium complexity). For the pure cultures, we grew five microbial species, namely *Bacteroides thetaiotaomicron*, *Chlamydomonas reinhardtii*, *Escherichia coli*, *Rhizobium leguminosarum*, and *Thermus thermophilus* in one or two different culturing conditions, or used different strains of the same species (see Results and Discussion Figure 1). For the species grown in two conditions we chose conditions that induce distinct protein abundance profiles. The resulting defined metaproteomes corresponded well between the actual measured species abundance and the theoretical (pre-defined) input abundances for each species. This shows that the sample generation process was successful and that these samples can be used to determine the performances of statistical tests considering various metaproteomic data challenges (Figure 1).

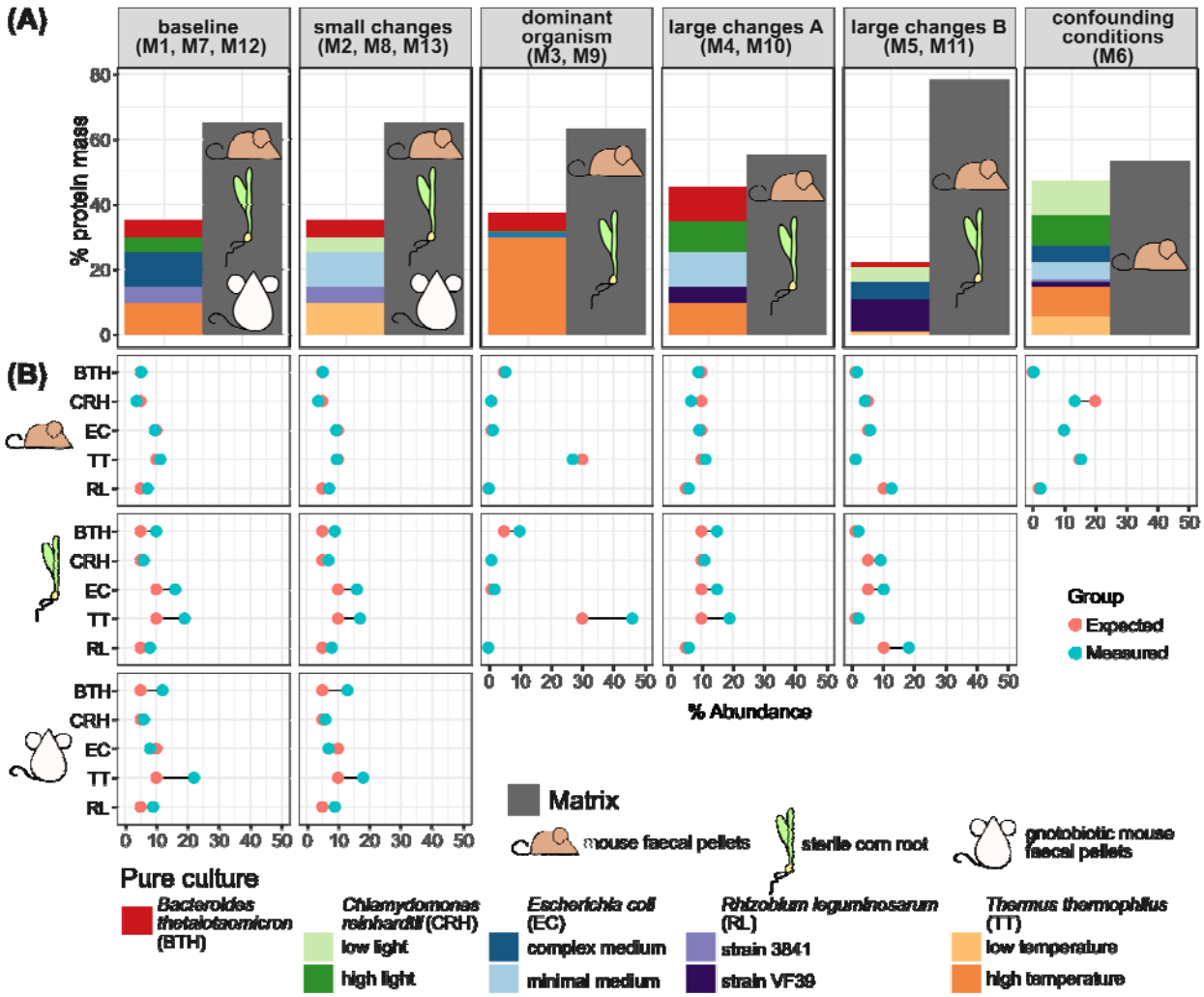


Figure 1: Design of defined metaproteomes and comparison of measured with pre-defined species abundances. (A) Defined metaproteome compositions (see Supplementary Table 4 for percentage input per species). Grey bar corresponds to the peptide abundance of the respective matrix. We used one of three matrixes; some mixes were generated in the same composition with different matrices. (B) Comparison of the measured species abundances (proportion of peptide signal summed per species to total signal) with their pre-defined input abundance (mass-% of peptides). Color-blindness accessibility note: the order of pure cultures in the legend corresponds to the order in the barplots (left to right columns = top to bottom).

For testing statistical data analysis methods, we used regression-based approaches, machine-learning based techniques, and null hypothesis testing (Table 2). Different tests are compatible with different normalization and/or transformation techniques (Table 2 and Methods). When applicable, we set the nominal false discovery rate (FDR) for detection of significant differences to 5 %. The FDR measures how many of the detected significantly different variables are, in fact, not significantly different between conditions.

Table 2: Combinations of statistical tests and normalization (+transformation) methods used in this study. Note that only certain normalization/transformation and test combinations are compatible, as given below. The method corncob does not normalize beforehand. S: used for spectral count data, A: used for AUC data. Abbreviations correspond to abbreviations used throughout the study, please see below table. For details and references, please refer to the Methods section.

	normalization and/or transformation															
Statistical test	chiPower (chiP)	clr	CSS_CPLM	log	logNSAF	NSAF	RLE	TSS	astTSS	logTSS	logitTSS	sqrtTSS	TMM	TMMwsp	voom	vsn
bayesian linear regression (brms)	S A			S						S A						S A
corncob																
DESeq2							S									
edgeR							S						S	S		
limma															S A	S A
(generalized) linear regression (lm)	S A				S					S A		S A	S A			S A
MaAsLin2		S A	S A						S A	S A	S A		S			
Random forest (rf)	S A	S A		A	S	S		S A		S A		S	S A			S A
Welch t-test										S A						
Wilcoxon										S A						

Abbreviations used throughout this study

ast	arc sine square root transformation
clr	centered log-ratio transformation
CPLM	compound Poisson linear model
CSS	cumulative sum scaling
log	log2 transformation
logit	logit transformation
NSAF	normalized spectral abundance factors
TSS	total sum scaling
TMM	trimmed mean of M values
TMMwsp	TMM with singleton pairing
RLE	relative log expression
sqrt	square root transformation
vsn	variance stabilization normalization

Additional abbreviations used for statistical methods

brms	**Bayesian linear regression**
exgauss	exponentially modified gaussian distribution
gauss	gaussian distribution
prior	adaptation of the prior (instead of using the default)
sn	skew-normal distribution
Student	Student's t-distribution
edgeR 4.0	**empirical analysis of DGE in R**
ET	exact test
QLE	quasi-likeliood estimation

lm	**generalized linear regression**
betainf	zero-inflated beta distribution
poisson	Poisson distribution
limma	**linear models for microarray data**
voom	voom transformation
deqms	proteomics-specific variance estimation for limma
MaAsLin2	**Microbiome Multivariable Associations with Linear Models**
rf	**random forest**
impconst	0 imputation with small constant after transformation
med	filtering for median of random forest p values <0.05
mtry20	random forest mtry=20 (i.e., 20 variables taken per split)
mtry200	random forest mtry=200 (i.e., 200 variables taken per split)
relaxed	relaxed filter for random forest
sf0.6	random forest with sample fraction 0.6
zcomp	imputation via zcompositions package
brms, lm, rf	
chip09	chiPower with lambda=0.9 (other numbers represent other lambda values accordingly)
MaAsLin2, lm	
negbin	negative binomial distribution

We based evaluation of statistical tests for the defined metaproteomes on metrics derived from true positive (TP), true negative (TN), false positive (TP), and false negative (FN) identifications (Figure 2). The ground truth used for these calculations depended on the type of comparison:

(i) If normalizing for species abundance and comparing between different culturing conditions of an organism, we defined as ground-truth those proteins that were statistically significantly different between the culturing conditions when measuring the pure cultures. Proteins that were only detected in pure cultures, but not in the defined metaproteomes, were not taken into account. Few protein groups were only detected in defined metaproteomes, but not in pure cultures. We removed these from our analyses, as we were not able to assess based on our definition whether these are true or false positives/negatives. This comparison applies, for example, for *T. thermophilus* between M2 and M3 (Figure 1).

(ii) If not normalizing for species abundance, but comparing proteins of one species between conditions, with the same total species abundance in both mixes (i.e., no shifts in relative protein abundances which are solely caused by the different relative abundances of the organism), the ground-truth was comprised of the same proteins as in (i). One example is the comparison of *T. thermophilus* between M1 and M2 (Figure 1).

(iii) If normalizing for species abundance and comparing proteins of one species within the same condition, or not normalizing and comparing between the same condition and same abundance, there are by definition no TP, and all proteins that are detected as significantly

different are by definition FP. See, for example *T. thermophilus* M1 vs. M3, or *T. thermophilus* M1 vs. M4 (Figure 1).

(iv) If not normalizing for species abundance for a specific species grown under one condition, and the abundance of that species was different between mixes, then by definition all proteins from that species are differentially abundant and can be considered TP if detected as statistically significantly different. In this comparison, there are no FP or TN, but only FN, i.e., those proteins of the species which were not detected as being significantly differentially abundant. This type of comparison occurs, for example, for *T. thermophilus* M1 vs. M3 (Figure 1).

(v) The comparison of no normalization for species abundance, but changes in condition and species abundance cannot be evaluated, because it is not known whether the changes are due to species abundance changes, condition changes, or both (i.e., changes in abundance of the organism are confounded with changes abundance of individual proteins due to different culturing conditions). We assigned "not available" (NA) to these values. Please note that in a real-life dataset, where the abundances of organisms are usually not known beforehand, it is thus necessary to normalize on organism (or sub-community) level, if biologically meaningful conclusions on the organism level should be drawn (see also Kleiner, 2017). This pertains, e.g., to *T. thermophilus* M2 vs. M3 (Figure 1).

(vi) We also assigned "NA" to TP and TN where the organism was not present in both conditions, as all detected proteins are already wrongly detected; there is no ground truth for this type of comparison. See, for example, *R. leguminosarum* strain VF39 in M1 vs. M2 (Figure 1).

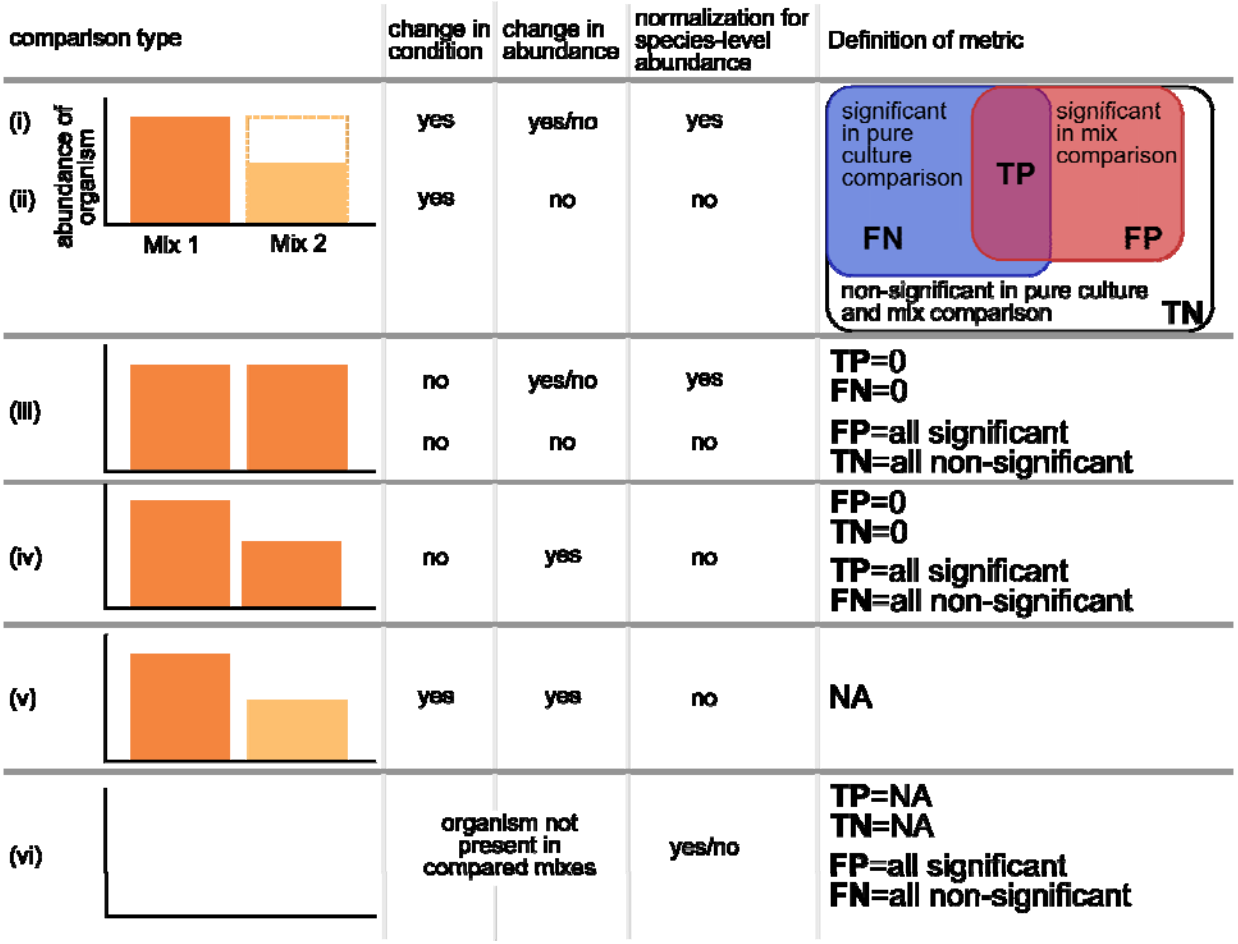


Figure 2: Definition of outcomes for each metric used for performance assessment of statistical tests in this study. For reference to scenarios, please refer to the text. Dark vs. light orange: Culturing condition 1 vs. culturing condition 2.

For comparison types (i) and (ii), the final evaluation was based on the following metrics (all ranging between 0 and 1, and chosen so that they will be 1 for ideal conditions):

(1) Positive predictive value (PPV) = Precision = complement of the false discovery rate (FDR; (Benjamini and Hochberg 2000; Benjamini and Yekutieli 2001)):

(2) Negative predictive value (NPV) = complement of the false omission rate (FOR):

(3) True positive rate (TPR) = sensitivity = recall = complement of the false negative rate (FNR):

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P} = 1 - FNR$$

(P: all protein groups which are actually positive, i.e., statistically significantly different)

(4) True negative rate (TNR) = specificity = complement of the false positive rate (FPR):

$$TNR = \frac{TN}{TN + FP} = \frac{TN}{N} = 1 - FPR$$

(N: all protein groups which are actually negative, i.e., not significantly differentially abundant)

(5) Balanced accuracy (BAcc; (Brodersen et al. 2010)):

$$BAcc = \frac{TPR + TNR}{2}$$

(6) F1 score = harmonic mean of precision and recall (Sasaki 2007):

$$F1 = \frac{2 * PPV * TPR}{PPV + TPR} = \frac{2 * TP}{2 * TP + FP + FN}$$

(7) P4 score (Sitarz 2022), which balances the PPV, TPR, TNR, and NPV (i.e., if the P4 score is close to one, all of the metrices used are close to one):

$$P4 = \frac{4}{\frac{1}{PPV} + \frac{1}{TPR} + \frac{1}{TNR} + \frac{1}{NPV}} = \frac{4 * TP * TN}{4 * TP * TN + (TP + TN) * (FP + FN)}$$

For the comparisons (iii) to (vi), calculation of these metrics does not make sense, as some (or all) of the base metrics (TP, FP, TN, FN) are 0 or NA.

Determining statistical tests for pure culture ground truth differentially abundant proteins

We used pure culture proteome data to identify protein groups whose differential abundances between mixes were sufficiently reliable to use for downstream analyses of statistical test performance. For that, we determined the ground truth of statistically significantly different protein groups between conditions using the proteomes of the pure cultures that were the inputs

for the defined metaproteomes. To do this, we first determined which statistical test to use for pure culture proteome comparisons. To determine this test, we used the defined metaproteomes with the mouse fecal pellet matrix (M1 to M5) that had EC, TT, and CRH at the same condition, but at different abundances, meaning that for each of these three species all proteins should be statistically significantly different in the respective tests, when not normalizing on the species level. We then identified the tests yielding the best balance between true positives and true negatives. For SpC, we defined a threshold of 70% TP and 95% TN, resulting in us using brms with family=gaussian and chipower (lambda=0.9) normalized and transformed data (Figure 3) for pure culture proteome comparisons to identify differentially abundant protein groups to be used as ground truth (Supplementary Tables 5a, 5b). For AUC, we used thresholds of 70% TP and 85% TN, giving lm_vsn as pure culture ground-truth test (Supplementary Tables 6a, 6b).

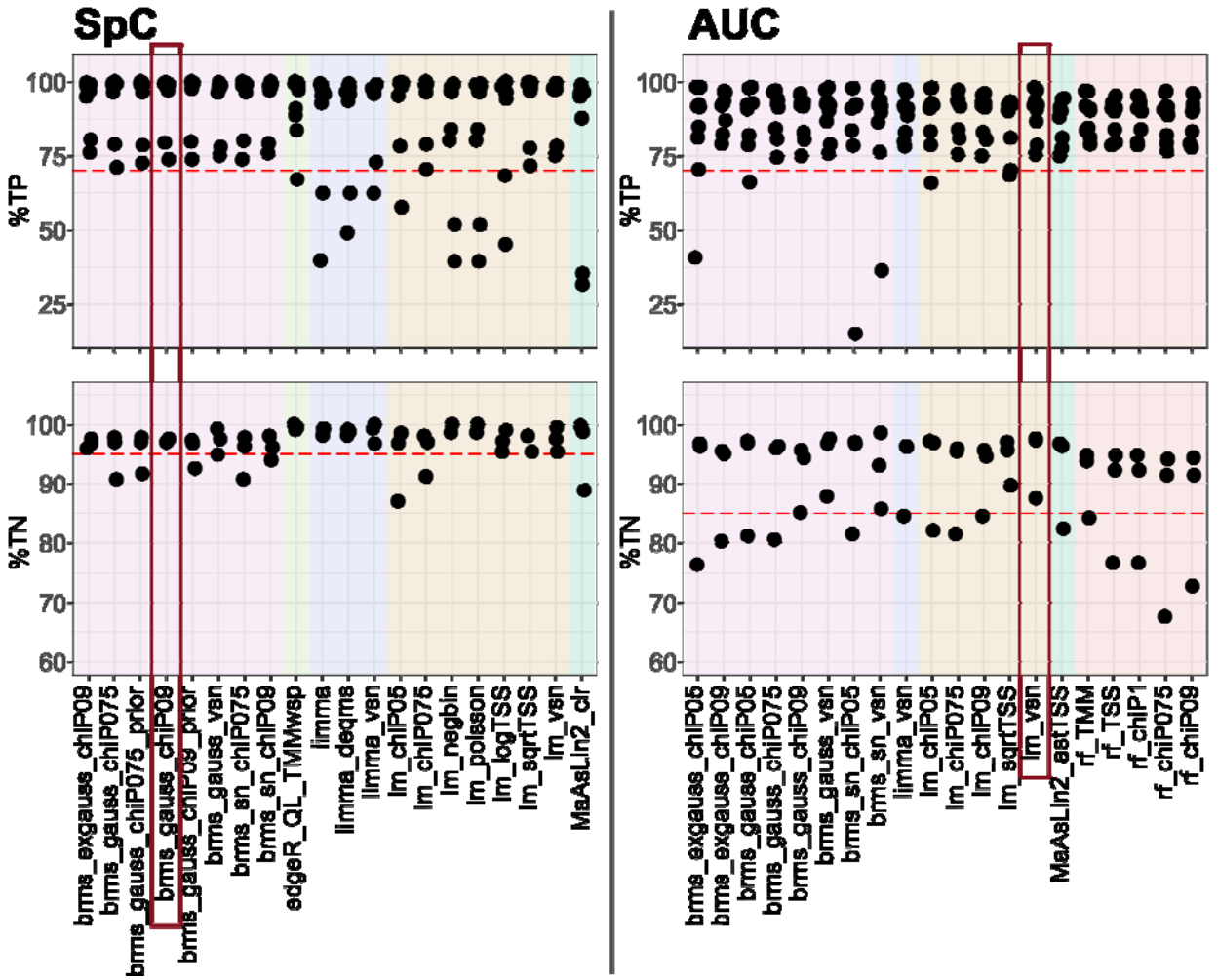


*Figure 3: Determination of the ground-truth test for evaluating statisticallly significantly different proteins in pure cultures. We pre-selected tests with a median %TP and %TN of 95 for SpC, and of 90 for AUC data. %TP and %TN: Percentage of true positives/ true negatives of all proteins of the respective organism (*T. thermophilus, E. coli, *or* C. reinhardtii*) in comparisons where the expectation was that all (100%) proteins of an organism would be differentially abundant, because the organism was mixed into the compared samples at different abundances (but same condition). Red boxes indicate the tests used to identify differentially abundant proteins in the pure cultures, which we define as the statistical ground-truths. Red dashed lines: 70% TP and 95% TN for SpC, 70% TP and 85%TN for AUC, respectively.*

Many statistical approaches perform similarly well, and some perform poorly

With the determined ground truth of statistically significant proteins, we next determined how each statistical test was able to identify differentially abundant proteins in the mouse faecal pellet mixes 1 to 5, using the results of comparison types (i) and (ii) as outlined in Figure 2. No single statistical approach clearly out-performed the others (Figure 4). Please note that the scores we chose can range from 0 to 1 and trend towards 1 for desirable performance outcomes. Some statistics clearly performed worse than the majority of tests (e.g., Wilcoxons rank-sum test, DESeq2) as well as some normalization methods tested for random forest implementations. Based on these results we narrowed the list of tests to use in further evaluations to well-performing tests, which included tests with a median TNR and TPR in the top 25% of all tests. Since some tests with high median performance suffered from a large spread in TNR and TPR, we added the requirement that the the TNR and TPR lower quartiles for each test had to be greater than those of 25% of all tests (Supplementary Tables 5c, 5d, 6c, 6d). This resulted in us retaining 30 tests for SpC-based approaches (42% of all 72 tests), and 17 of 47 (36%) for AUC-based tests (Figure 4). Due to the fixed cutoff, some tests were removed, which performed only slightly less good than these selected tests. Generally, we noticed that the NPV performed worse than all other metrics, meaning that all tests miss a large share of actually significant differences (Figure 4).

Our results partially corroborate comparisons of statistical approaches that were done on other (meta)-omics data types. We found that compositional analyses methods generally did not perform better than methods not explicitly taking compositionality into account, which is in line with previous results from a study comparing statistical approaches fo 16S rRNA gene amplicon and WGS data (Calgaro et al. 2020). However, in that study, DESeq2, limma-voom, and corncob were recommended as best-performing, while in our study DESeq2 and corncob were among the worst-performing approaches. Another study on use of statistical approaches for RNA-Seq data found that DESEq2, limma-voom and edgeR (version 3) produced too-high false positive rates and the Wilcoxon rank-sum test was recommended (Li et al. 2022). In contrast, we found that limma-voom performed well on our data, and Wilcoxon rank-sum test performed poorly. These contrasting results between studies comparing statistical approaches for omics data are likely due to several causes, including 1) differences in underlying data structures of different (meta)-omics data types despite some similarities (e.g., compositionality and sparsity), and 2) different frameworks for evaluating performance of statistical approaches. While some past studies focused heavily on FPRs or FDRs (=1-PPV) (e.g., Calgaro et al. 2020; Li et al. 2022), we used a balanced framework: we took into account not only true and false positives, but also true and false negatives, i.e., PPV and NPV (as complements of the FDR and FOR, chosen so that they will approach 1 for a desired outcome), as well as TPR and TNR. For final scoring, we compared the F1 score to the more balanced P4 score, and focused on the latter. For example, if we focused only on TNRs, the Wilcoxon test would be a clear winner, however, when also considering TPRs, it becomes immediately clear that this test is unsuitable. The same principle holds true for DESeq2, which fares well with regard to the TPR, but had the lowest TNR. The

tests we chose for more in-depth analyses show a more balanced performance, and thus decrease the risk for a biased data analysis.

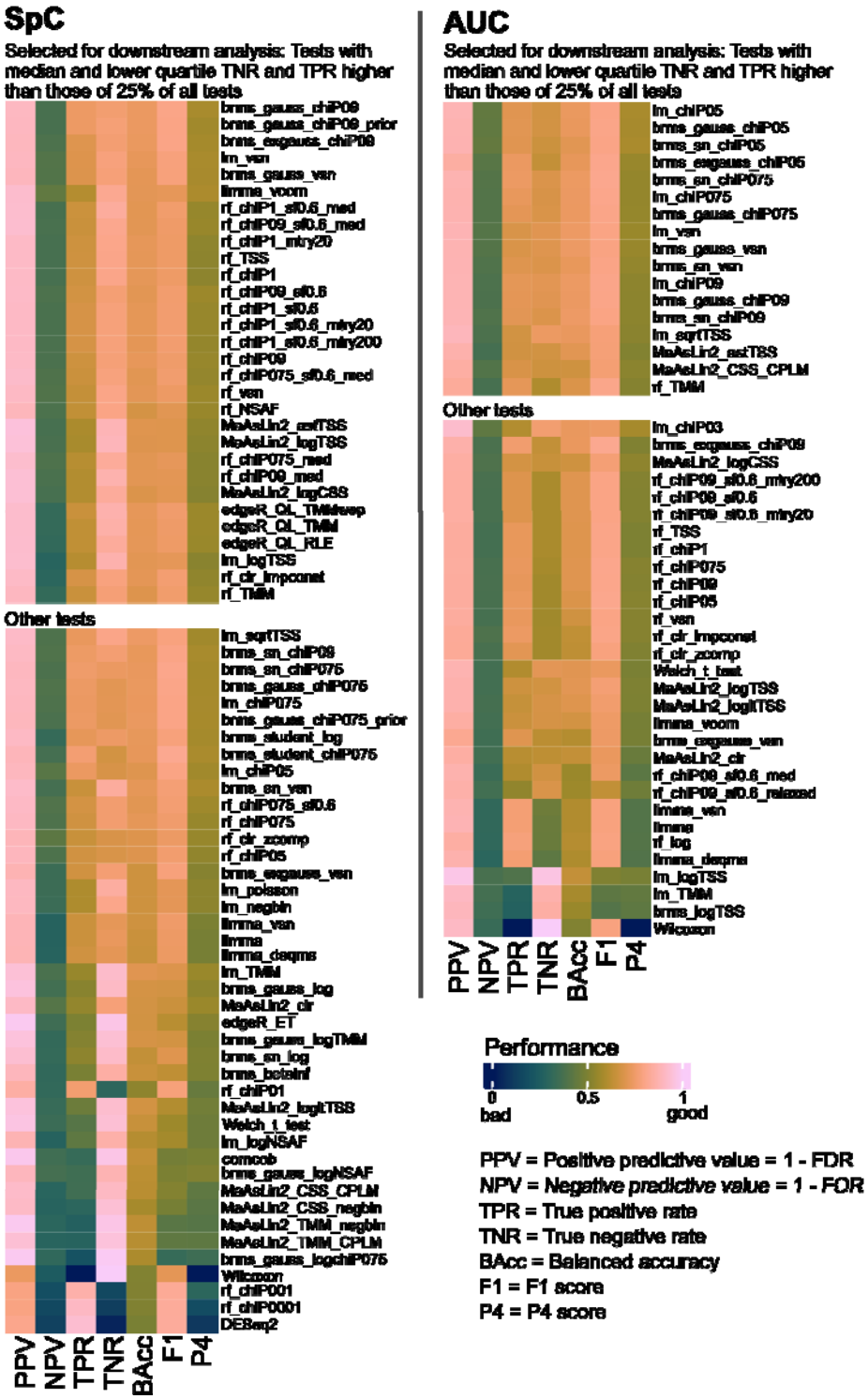


Figure 4: Median test evaluation metrics for mouse faecal pellet matrix comparisons (M1 to M5). Upper part: tests with a median and lower quartile TNR and TPR higher than those of 25% of all tests.The tests in the upper part were used for further in-depth analysis. Data was clustered row-wise based on Canberra distances.

Type of comparison and data pre-treatment impact test performance

For the 30 SpC tests and 17 AUC tests for which we carried out in-depth analyses, we found that the type of comparison, in terms of organisms, as well as in terms of organism abundance compared, had a major impact on performance, as demonstrated by the large spread of P4 scores for individual tests (Figure 5, Supplementary Figure 2 for all matrices). The type of mix impacted the outcome more than the statistical test chosen: the intraclass correlation (ICC), reflecting the similarity of test results, was (for SpC) 0.89 (F-test, $p < 0.01$) between tests, but only 0.02 (F-test, $p < 0.01$) between mixes. For AUC, the ICC between tests was even higher at 0.95 (F-test, $p < 0.01$) and lower between mixes (0.006, F-test, $p < 0.01$).

Specific comparison types that led to low performance, indicated by low P4 scores across tests, involved species that were present in the defined metaproteome samples at 1%. We observed this low performance for low-abundance species for both SpC and AUC-based analyses. This low performance of tests on samples with low-abundance species is likely due to more inaccurate quantification during LC-MS/MS analysis. Additionally, all comparisons that involved *C. reinhardtii* had low P4 scores, especially for AUC analysis, as compared to comparisons including *T. thermophilus* and *E. coli*, even when considering comparisons with the same species abundance ratios (see Supplementary Figure 2 and Supplementary Tables 5 and 6 for all matrices). Notably, *C. reinhardtii* showed very low numbers of proteins constantly identified across all mixes in the SpC data compared to its overall high number of proteins identified. This lack of overlap likely explains the low P4, caused by a low NPV (Supplementary Tables 5c, 5d, 6c, 6d and Supplementary Figure 3). For the AUC data, on the other hand, overlap between identifications was very high (Supplementary Figure 4), indicating protein mis-identifications and thus AUC quantification distortions, which might have been caused by the MS1-based transfer of peptide identifications between datasets (i.e., match between runs), as performed in ProteomeDiscoverer (see manual), but also in other software such as MaxQuant (Tyanova et al. 2016).

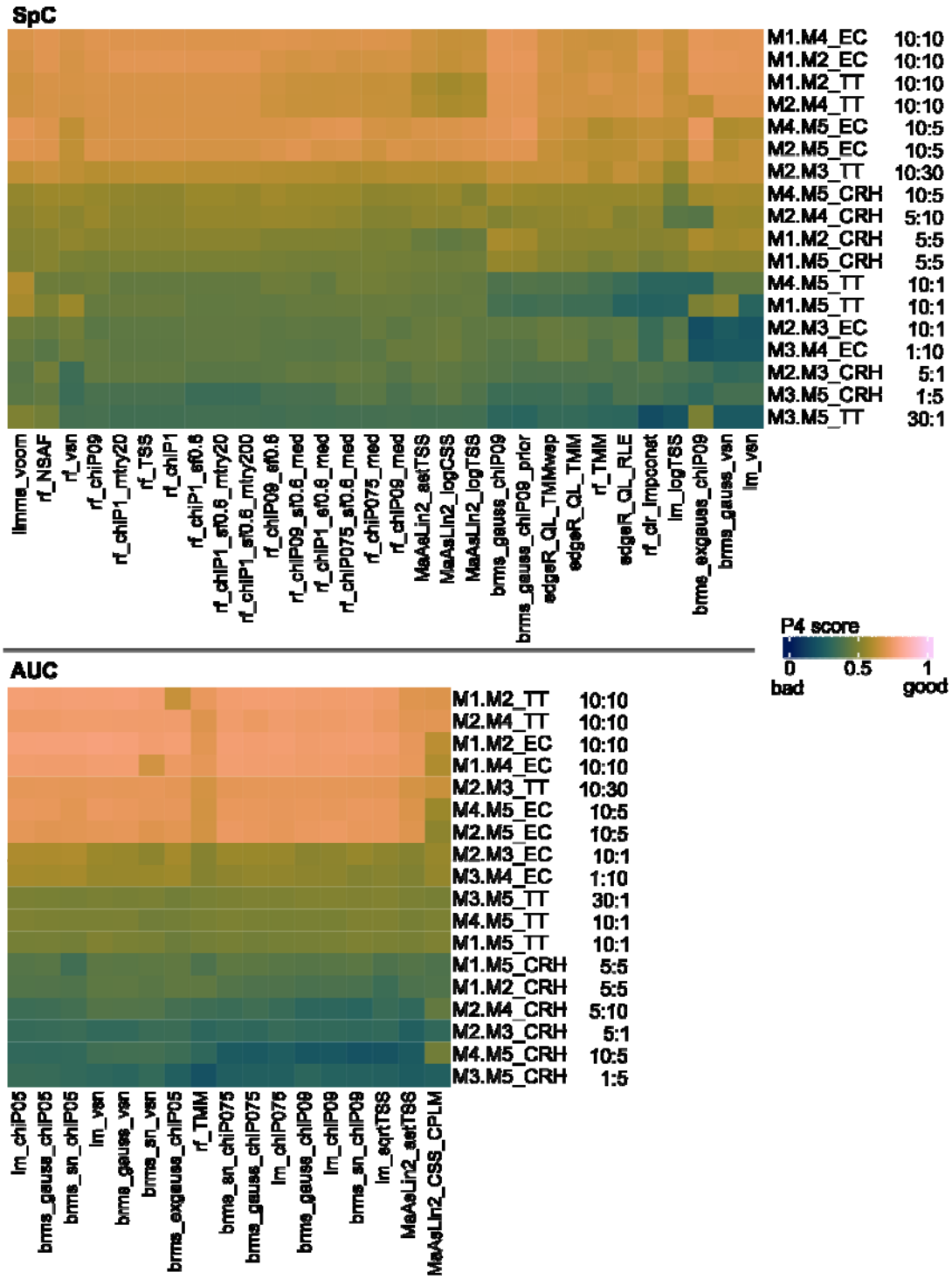


Figure 5: Heatmaps of P4 scores for the best-performing statistical approaches in mouse faecal pellet matrix (comparisons M1 to M5). Tests shown here were chosen as the best-performing ones based on data shown in Figure 4. The P4 score ranges from 0 to 1 and trends towards 1 for desirable performance outcomes. Data was clustered row- and column-wise based on Canberra distances. CRH: Chlamydomonas reinhardtii, EC: Escherichia coli, TT: Thermus thermophilus. Numbers to the right give the abundance of the respective organism in percent in the compared mixes.

While median test performances and P4 scores mainly clustered by the statistical test itself (i.e., most random forest implementations clustered together, as did most limma, etc.), normalization and transformation did in some cases have a noticable impact as well (Figure 4, Figure 5): the voom normalization greatly increased performance of limma, while chiPower pretreatment with a too-low lambda worsened the random forest performances. On the other hand, random forest performance was at most slightly impacted by across-forest summarization via kernel density vs. median statistics or hyperparameter tuning. For edgeR, a quasi-likelihood estimation outperformed the exact test statistics (Figure 4).

Explicitly accounting for compositionality, e.g., via a clr transformation, has been advocated as mandatory for microbiome datasets (Gloor et al. 2017). According to our analysis, normalization and transformation impacted test performance, but accounting for compositionality was not necessary and in fact use of clr transformed data (but not of chiP transformed data) led to worse performance (Figure 4). Similar effects were shown by a study on amplicon sequencing data, where clr, isometric log-ratio, and additive log-ratio transformations performed slightly worse as compared to normalizations which do not take compositionality into account (Yerke et al. 2024).

For random forests, imputation with a small constant performed slightly better than imputation based on compositionality (zcomp), indicating that more intricate methods are not always necessary (Figure 4). While more complex or data-dependent imputation methods exist (see, e.g., Lazar et al. 2016; Webb-Robertson et al. 2015), as well as methods explicitly taking into account the pattern of missingness, we found these to be unsuitable for the low number of replicates often available in metaproteomic studies and we are therefore not addressing these here. More stringent pre-filtering led to more true positives, but had inconsistent effects on the number of true negatives (see Supplementary Results and Supplementary Figure 1).

Best tests per test category highlight test-specific differences

To determine the set of best-performing tests for more detailed analyses, we subsetted the tests to only include the test with the highest median PPV (=lowest FDR) per test category (i.e., brms, edgeR, limma, lm, MaAslin2, and random forests; edgeR and limma are only applicable to SpC), in the well-performing tests for all comparisons involving M1 to M5. This resulted, for SpC, in the six tests brms_gauss_chiP09, edgeR_QL_RLE, limma_voom, lm_logTSS, MaAsLin2_astTSS, and rf_chiP09_med. For AUC, we included the four tests brms_sn_chiP09, lm_sqrtTSS, MaAsLin2_astTSS, and rf_TMM (Figure 6, Supplementary Tables 5e, 6e). The spread of results per test is still wide, underlining that the specific comparisons performed impact the results. Additionally, the F1 score showed more differences between tests as compared to the P4 score, highlighting that the F1 score is only impacted by the PPV and TPR, whereas the P4 score is a balanced score of PPV, NPV, TPR, and TNR. The general result of high PPV but rather low NPV, as discussed above, holds true here as well. For SpC-based tests,

brms_gauss_chiP09 clearly outperformed the others in terms of the TPR, but had a slightly lower TNR. On the other hand, limma_voom had the highest median NPV, and the lowest performance spread of the TNR. The AUC-based tests generally had a lower median PPV and larger spread of PPV (and most other metrics) compared to the SpC-based tests. Within the AUC-based test, differences between medians were mostly lower than for SpC-based tests. The performance of all tests decreased for comparisons with confounded data (i.e., with M6) as expected (Supplementary Results and Discussion, Supplementary Figure 5, Supplementary Tables 5f, 6f).

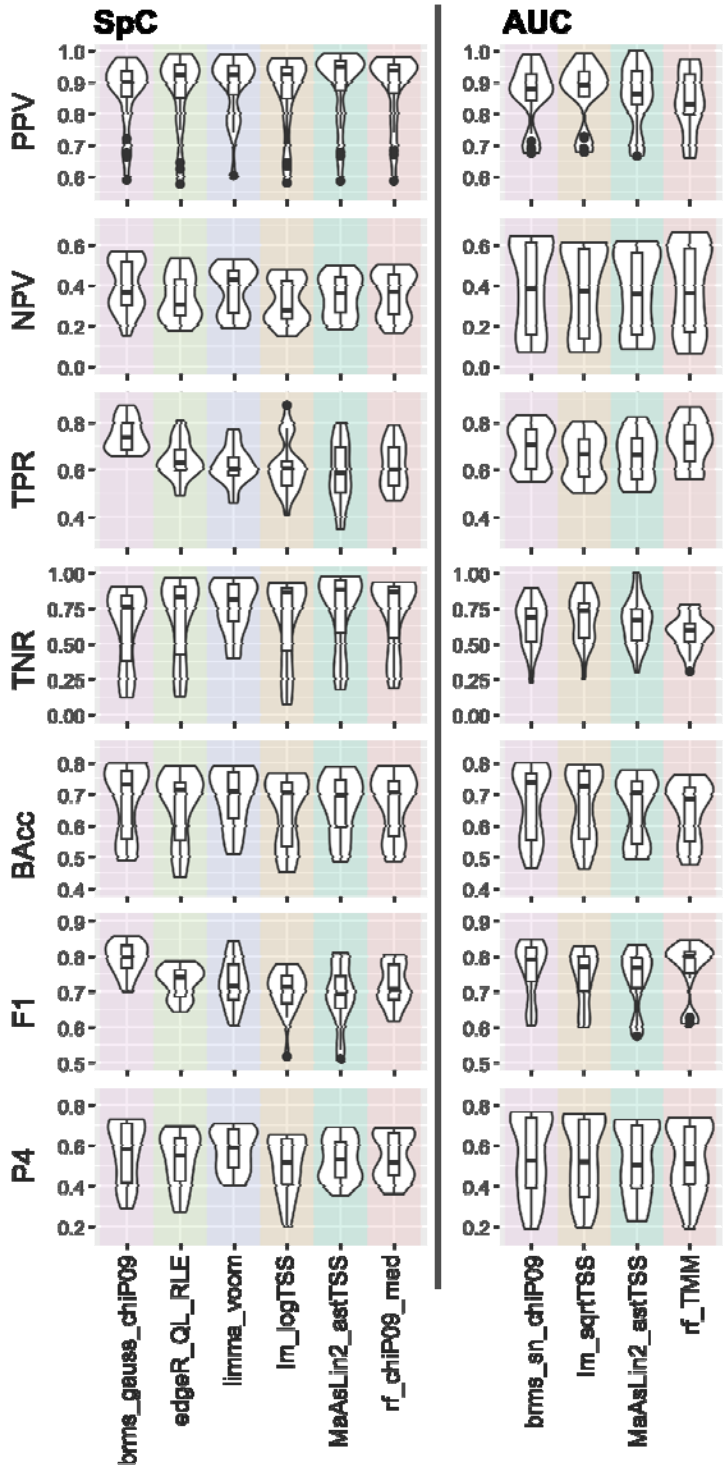


Figure 6: Test metrics for the best-performing tests in the mouse faecal pellet matrix defined by metaproteomes (M1 to M5). Shown are median values (boxplots) and value distributions (violins).

Protein (mis)identification effects

One challenge in metaproteomics is that proteins from closely related strains/species are hard to differentiate as large portions of their protein sequences are identical, which makes it less likely that differentiating peptides with differences in amino acid sequence are identified (Van Den Bossche et al. 2025). This can lead to misidentifications where the correct protein is identified, but it is assigned to the wrong species/strain, which can confound statistical tests. To address misidentifications between different bacterial strains, as well as between pure culture and matrix microbiome, we used the two *Rhizobium* strains RLVF39 and RL3841, and *B. thetaiotaomicron*, a microbial species that is also present in the mouse stool naturally (Figure 7). These analysis included comparison types (iii) to (vi) as outlined in Figure 2. Comparisons involving a higher protein abundance of one organism increased test performance for RLVF39 and RL3841 comparisons. Considering the TN, organism-level normalization generally (slightly) increased test performance (as expected). Performance decreased visibly for SpC data using the lm_logTSS considering RL3841 in the M1 vs M2 comparison (i.e., comparing 5% abundance each), likely because the model fit was comparatively low (a test of taking all *Rhizobium* proteins in this comparison together, i.e., including potential strain misidentification, produced a boundary fit, i.e., the random effect was estimated as zero).

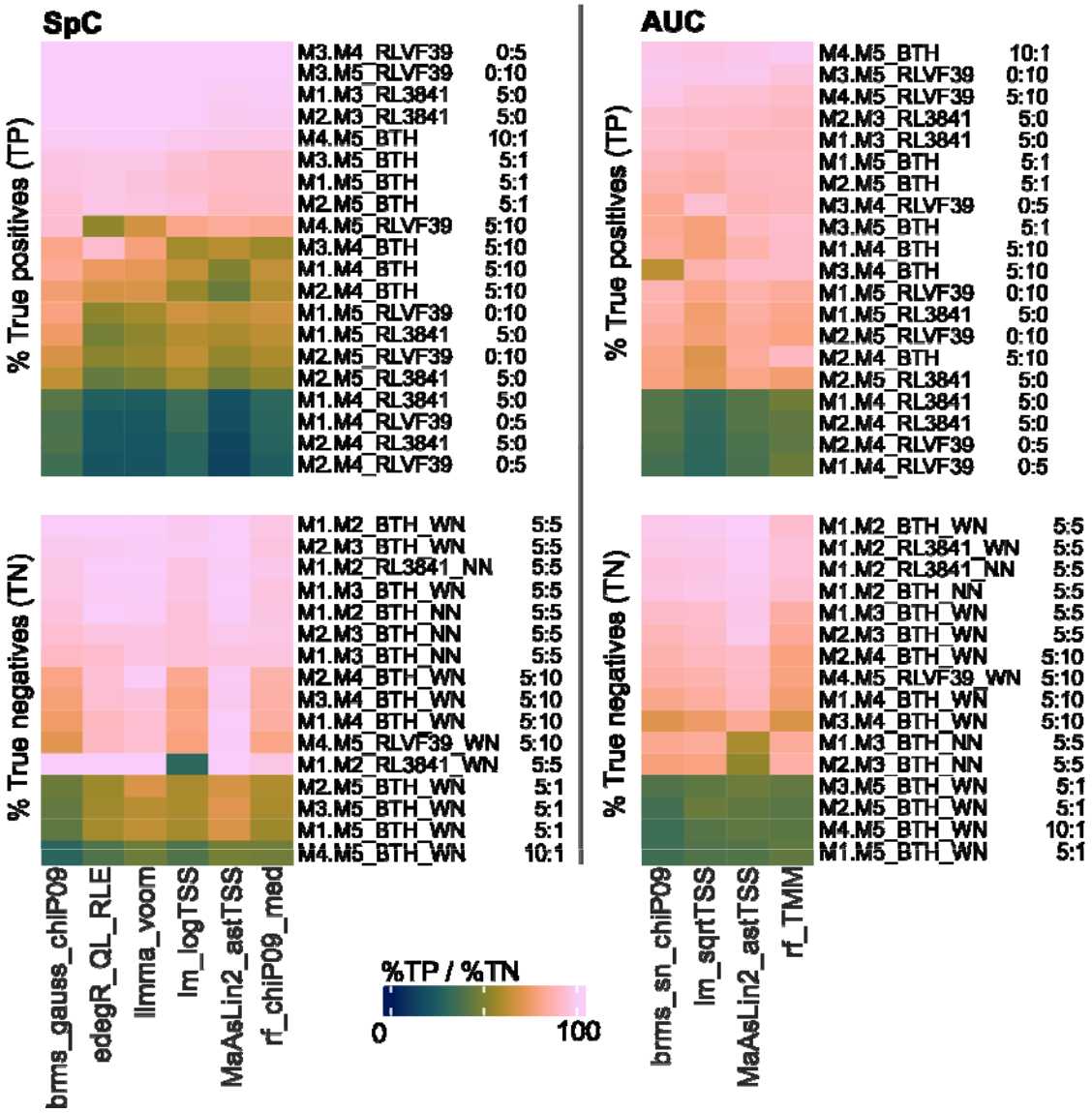
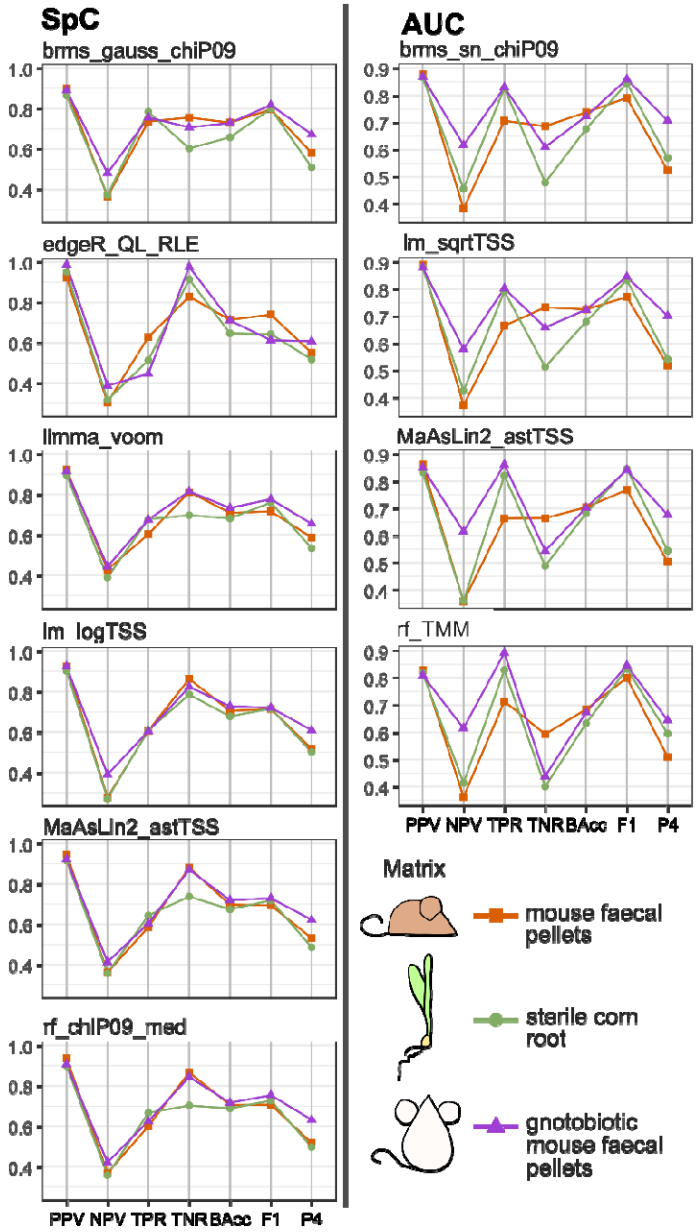


*Figure 7: Heatmaps of percentages of true positives (TP) and true negatives (TN) for pure cultures in mouse faecal pellet matrix (M1 to M5) without different culturing conditions (*B. thetaiotaomicron *and the two* Rhizobium *strains RL3841 and RLVF39). Numbers give the abundances of the respective organism in percent in the compared mixes. Data was clustered row- and column-wise based on Canberra distances. WN: with organism-level normalization, NN: no normalization on organism level. Note that no normalization is only possible for evaluation of TN; for TP to be present, there has to be an actual abundance change, and when normalizing, this change in abundance is by definition eliminated.*

Matrix complexity effects

The three matrices we used for the defined metaproteomes mainly impacted the NPV and, concomitantly, the TNR (Figure 8). While performances of the best tests were similar for SpC across all three matrices, performances of the best AUC tests varied widely between matrices. The gnotobiotic mouse matrix generally provided best performance - but it has to be noted that

735 this included only one type of comparison, i.e., between M12 and M13, thus excluding many of the challenges present for comparisons within the other two matrices. Based on the P4 values per comparison and test (Supplementary Figure 2, Supplementary Tables 5e, 5h, 5i, 6e, 6h, 6i), comparisons involving CRH and very low-abundant organisms (i.e., 1%) were especially challenging for AUC based tests, independent of the matrix.



740 *Figure 8: Evaluation metrics for best tests for all three matrices used. Shown are parallel coordinates plots with median values (please note that, while the mouse faecal pellet matrix and sterile corn root matrix include the same number and kind of comparisons, in the gnotobiotic mouse faecal pellet matrix only M12 and M13 are present limiting the number of comparisons to one). Detailed P4 score comparisons for data from corn root and gnotobiotic*
745 *mouse matrix are shown in Supplementary Figure 2.*

Conclusion

In this study, we developed a complex and controlled sample set and performance test framework to evaluate statistical approaches for differential abundance analysis in metaproteomic studies. Our known ground truth enabled a "gold standard"-based evaluation of statistical approaches for data-dependent metaproteomics data with a low number (n=4) of replicates.

Pertaining to the general challenges for metaproteomics data analysis we designed our samples for, tests generally fared worse at (a) detecting small changes, than (b) detecting large changes without increasing false discoveries. The challenge (c) of misassignments becomes relevant especially for related strains, and, while thus pertaining to all tests, was apparently somewhat better controlled by brms_gauss_chiP09 for SpC data, whereas for AUC, there was no clear difference between the best-performing tests.

Regarding differences between tests, SpC-based tests fared better than AUC-based tests at controlling the PPV (=1-FDR), i.e., detecting proteins as significantly differentially abundant which are truly differentially abundant. The NPV (=1-FOR) was low for all tests which passed initial quality assessment criteria, which means that many protein groups that were in fact significantly differentially abundant were not detected as significantly different. While in many studies significant differences between conditions are the focus of analysis, and will be interpreted in greater detail, this low NPV (=high FOR) means that we are missing a lot of true differences as they do not show up as significant in the tests. This certainly needs to be taken into account when drawing conclusions from metaproteomics data. A low NPV is, however, not a metaproteomics-specific issue, as it is likely low in many (meta-)omics approaches. NPV, however, is often not evaluated for omics-focused statistics due to the absence of a known ground truth for true positives. Our results highlight that evaluations of statistical approaches for omics data should consider a framework that enables the assessment of NPV/FOR.

One way to increase the NPV is to use a higher number of replicates, which is often difficult in metaproteomics studies that use limited environmental samples or samples from small cohorts (e.g., prospective clinical studies). For example, RNA-Seq needed at least 20 replicates per condition to identify 85% of the truly differentially expressed genes, and it was recommended to use at least 6 replicates per condition (Schurch et al. 2016). Increasing biological replication should thus always be considered when feasible. Additionally, the impact of number of replicates on test performance is a factor worth studying further for metaproteomics.

The performance of AUC-based tests appeared to be much more dependent on sample complexity as compared to SpC analysis. We speculate that increasing complexity of MS1 spectra in our metaproteomes leads to mis-assignment of AUCs for specific mass peaks due to high peak density in the MS1 spectra.

Based on our evaluation, we recommend the analysis strategies for metaproteomics data in Table 3. Additionally, well-performing data pretreatment-test combinations are given in the upper parts of Figure 4 and in Figure 5. We provide the reproducible code, including the input files, and a readme file on how to execute the code with this study.

Table 3: Recommendation for metaproteomics statistical approaches based on our analyses. Please see the supplement for reproducible code, input files, and a readme detailing code execution.

	Metaproteomics quantification data type	
	SpC	**AUC**
General tests we recommend		
	brms_gauss_chip09 edgeR_QL_RLE limma_voom lm_logTSS MaAsLin2_astTSS rf_chiP09_med	brms_sn_chiP09 lm_sqrtTSS MaAsLin2_astTSS rf_TMM
Which computational resources are available?		
Desktop computer	lm_logTSS MaAsLin2_astTSS edgeR_QL_RLE limma_voom	lm_sqrtTSS MaAsLin2_astTSS
Server/Cluster capacities	brms_gauss_chip09 rf_chiP09_med	brms_sn_chiP09 rf_TMM
What is the expected change in the (microbial) community analyzed?		
Low	All general test recommendations	All general test recommendations
High/Unknown	limma_voom	brms_sn_chiP09
Is data pre-treatment that accounts for compositionality explicity needed?		
yes	brms_gauss_chip09 rf_chiP09_med	brms_sn_chiP09
no	All general test recommendations	All general test recommendations

One decisive factor in test choice might be the amount of computational resources needed or available to perform statistical testing. Computational speed varies widely, with limma (for SpC) and linear models being especially fast and feasible on desktop computers and laptops. EdgeR (for SpC) and MaAslin2 are suitable for desktop computers, but more intensive in terms of computational resources required. Using Bayesian regression, or random forests, requires more computational power and can usually only be executed feasibly on a high-performance desktop computer or server.

It is of note that all of the statistical inferences in our metaproteomics dataset find relative (and by no means absolute) differences between conditions and that the underlying metaproteomics quantification itself is relative, not absolute. This means that if we find a protein to be

significantly more abundant in condition 1 as compared to condition 2, that same protein might still be more abundant on an absolute scale in condition 2, if overall more protein biomass is present in condition 2. It will be interesting to see whether and how recent approaches to estimate absolute microbial abundances from relative abundances via machine learning for large-scale datasets (Nishijima et al. 2025), or to model uncertainties/prior knowledge about the underlying total microbial community size explicitly (Gloor et al. 2025; Nixon et al. 2025), can be adapted for metaproteomics and for environmental -omics studies.

In addition to the taxon-specific analyses of protein group abundances considered here, additional approaches are worth considering in a metaproteomics data analysis depending on the biological question addressed. For example, summing of protein group abundances for homologous proteins with the same functional annotation (isoforms), with and without considering taxonomic origin, can be useful for comparisons of overall (microbial) community function between different conditions. Along the same vein, protein abundances can also be summed for specific taxonomic groups or for complete metabolic pathways prior to statistical analysis.

In our study we focused on tests to compare two conditions. Such pairwise comparisons are also often at the core of more complex experimental setups. We did not evaluate approaches for longitudinal or paired/grouped designs (for the importance of taking grouped data into account, see, e.g., Vorland et al. 2025), but linear mixed models and approaches based on them, as well as random forests (Capitaine et al. 2021) can be adapted to accommodate grouped data.

Additionally, we want to stress that any statistic relies on an appropriate experimental design, and clearly formulated hypotheses (Wagner and Kleiner 2025). Such a design needs to take into account, e.g., a sufficient number of replicates, appropriate controls, and mass spectrometry measurements in randomized blocks (Oberg and Vitek 2009). It is fundamental to think about how to analyze the data from the beginning, and to take into account whether to normalize at the organism level (see Methods and Kleiner 2017).

Our statistical performance evaluation framework is directly extendable to other metaproteomics data types such as DIA data, and other meta-omics approaches. DIA has been reported to give more reproducible protein identification results for metaproteomics (Rajczewski et al. 2025), and might thus alleviate issues with data sparsity. At the same time, DIA, as well as feature mapping/match between runs (Yu et al. 2021) can also introduce additional issues, e.g., potential mismatches in the case of match between runs (Lim et al. 2019), and database search issues (Rajczewski et al. 2025), which still need to be addressed.

The evaluations of over 70 SpC and over 40 AUC statistical analysis methods, resulting recommendations, corresponding R code, and metaproteomics data we presented here will help microbiome scientists using metaproteomics to make informed choices about experimental design and statistics. The approaches we tested and recommend here are adaptable to different experimental designs. Using our results and our performance testing framework, researchers can move forward with analyzing their own meta(proteomics) data, using our provided code as a

basis (see Supplementary Files). Moreover, researchers interested in developing and testing their own statistical approaches can use our publicly available data, and our evaluation code, for direct comparisons with the test strategies presented here.

840

Supplementary Files

Supplementary Text incl. Supplementary Table 3 and Supplementary Figures 1-5
Supplementary Tables 1,2, 4-6
Reproducible code, protein group quantification files, and README on how to use the code are provided via git (https://git.uni-greifswald.de/hinzket/Stats_Metaprot), as well as a fixed record via zenodo (10.5281/zenodo.17880379)

Declarations

Ethics approval and consent to participate

NC State's Institutional Animal Care and Use Committee approved all experimental activity involving conventional and gnotobiotic mice (Protocol # 18-034-B and 18-165-B). No human data or tissue was used in this study.

Consent for publication

Not applicable.

Availability of data and material

Metagenomic sequencing data can be accessed via the ENA accession number PRJEB105044; fastq files can be directly accessed via http://ftp.sra.ebi.ac.uk/vol1/run/ERR159/ERR15997335/. All mass spectrometry proteomics data and results were deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE repository with the data set identifier PXD045390. Reviewers can currently access the repository using the following details: Username: reviewer_pxd045390@ebi.ac.uk; Password: FwNk89lz.

We provide reproducible code, protein group quantification files, and an explanation of how to use the code via git (https://git.uni-greifswald.de/hinzket/Stats_Metaprot), as well as a fixed record via zenodo (10.5281/zenodo.17880379).

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 531801029 – TRR 410 "WETSCAPES2.0", the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM138362, the US National Science Foundation (NSF, IOS #2421771), the Novo Nordisk Foundation (INTERACT, Grant number: NNF19SA0059360), the U.S. Department of Agriculture National Institute of Food and Agriculture under award No. 2022-67013-36672, the US Department of Energy (DE-SC0022996), and a National Research Fund Luxembourg (FNR) grant (number INTER/Mobility/2022/BM/16965254).

The Gnotobiotic Core at the College of Veterinary Medicine, North Carolina State University is supported by the National Institutes of Health funded Center for Gastrointestinal Biology and Disease, NIH-NIDDK P30 DK034987.

Author's contributions

T.H. and M.K. conceived study and designed defined metaproteomes; T.H. performed all statistical analyses, generated figures, and wrote initial manuscript. B.J.K. generated defined metaproteome samples together with S.V., A.K., J.A.B.-R., performed metaproteomic sample preparation, measurement, and searches, generated figures. M.K. and B.J.K. gave input on statistical analyses. M.K. and P.W. mentored and supervised involved trainees. T.H., M.K., P.W., and B.J.K. acquired funding. All authors contributed to the manuscript.

Acknowledgements

We thank Michael Greenacre for help with the chiPower transformation, and Philipp Adämmer for input on XGBoost. Elisa Kasbohm and Volkmar Liebscher provided valuable discussions on statistics. We thank Heather Maughan for valuable comments on the manuscript.

All LC-MS/MS measurements were made in the Molecular Education, Technology, and Research Innovation Center (METRIC) at North Carolina State University.

References

Abbondio, Marcello, Alessandro Tanca, Laura De Diego, et al. 2023. 'Metaproteomic Assessment of Gut Microbial and Host Functional Perturbations in *Helicobacter Pylori* - Infected Patients Subjected to an Antimicrobial Protocol'. *Gut Microbes* 15 (2): 2291170. https://doi.org/10.1080/19490976.2023.2291170.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. 'Fitting Linear Mixed-Effects Models Using **Lme4**'. *Journal of Statistical Software* 67 (1). https://doi.org/10.18637/jss.v067.i01.

Benjamini, Yoav, and Yosef Hochberg. 2000. 'On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics'. *Journal of Educational and Behavioral Statistics* 25 (1): 60–83.

Benjamini, Yoav, and Daniel Yekutieli. 2001. 'The Control of the False Discovery Rate in Multiple Testing under Dependency'. *The Annals of Statistics* 29 (4): 1165–88.

Bergauer, Kristin, Antonio Fernandez-Guerra, Juan A. L. Garcia, et al. 2018. 'Organic Matter Processing by Microbial Communities throughout the Atlantic Water Column as Revealed by Metaproteomics'. *Proceedings of the National Academy of Sciences* 115 (3). https://doi.org/10.1073/pnas.1708779115.

Blakeley-Ruiz, J Alfredo, Alexandria Bartlett, Arthur S McMillan, et al. 2025. 'Dietary Protein Source Alters Gut Microbiota Composition and Function'. *The ISME Journal* 19 (1): wraf048. https://doi.org/10.1093/ismejo/wraf048.

Blakeley-Ruiz, J. Alfredo, and Manuel Kleiner. 2022. 'Considerations for Constructing a Protein Sequence Database for Metaproteomics'. *Computational and Structural Biotechnology Journal* 20: 937–52. https://doi.org/10.1016/j.csbj.2022.01.018.

Blakeley-Ruiz, J. Alfredo, Carlee S. McClintock, Him K. Shrestha, et al. 2022. 'Morphine and High-Fat Diet Differentially Alter the Gut Microbiota Composition and Metabolic Function in Lean versus Obese Mice'. *ISME Communications* 2 (1): 66. https://doi.org/10.1038/s43705-022-00131-6.

Breiman, Leo. 2001. 'Random Forests'. *Machine Learning* 45: 5–32.

Brodersen, Kay H., Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. 2010. 'The Balanced Accuracy and Its Posterior Distribution'. *2010 International Conference on Pattern Recognition*, 3121–24.

Bürkner, Paul-Christian. 2018. 'Advanced Bayesian Multilevel Modeling with the R Package Brms'. *The R Journal* 10 (1): 395. https://doi.org/10.32614/RJ-2018-017.

Calgaro, Matteo, Chiara Romualdi, Levi Waldron, Davide Risso, and Nicola Vitulo. 2020. 'Assessment of Statistical Methods from Single Cell, Bulk RNA-Seq, and Metagenomics Applied to Microbiome Data'. *Genome Biology* 21 (1): 191. https://doi.org/10.1186/s13059-020-02104-1.

Capitaine, Louis, Robin Genuer, and Rodolphe Thiébaut. 2021. 'Random Forests for High-Dimensional Longitudinal Data'. *Statistical Methods in Medical Research* 30 (1): 166–84. https://doi.org/10.1177/0962280220946080.

Chen, Tianqi, and Carlos Guestrin. 2016. 'XGBoost: A Scalable Tree Boosting System'. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 13, 785–94. https://doi.org/10.1145/2939672.2939785.

Chen, Yunshun, Lizhong Chen, Aaron T L Lun, Pedro L Baldoni, and Gordon K Smyth. 2025. 'edgeR v4: Powerful Differential Analysis of Sequencing Data with Expanded Functionality and Improved Support for Small Counts and Larger Datasets'. *Nucleic Acids Research* 53 (2): gkaf018. https://doi.org/10.1093/nar/gkaf018.

Delacre, Marie, Daniël Lakens, and Christophe Leys. 2017. 'Why Psychologists Should by Default Use Welch's t-Test Instead of Student's t-Test'. *International Review of Social Psychology* 30 (1): 92–101. https://doi.org/10.5334/irsp.82.

Díaz-Uriarte, Ramón, and Sara Alvarez De Andrés. 2006. 'Gene Selection and Classification of Microarray Data Using Random Forest'. *BMC Bioinformatics* 7 (1): 3. https://doi.org/10.1186/1471-2105-7-3.

Fernandes, Andrew D, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. 2014. 'Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis'. *Microbiome* 2 (1): 15. https://doi.org/10.1186/2049-2618-2-15.

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. 'Microbiome Datasets Are Compositional: And This Is Not Optional'. *Frontiers in Microbiology* 8 (November): 2224. https://doi.org/10.3389/fmicb.2017.02224.

Gloor, Gregory B, Michelle Pistner Nixon, and Justin D Silverman. 2025. 'Explicit Scale Simulation for Analysis of RNA-Sequencing Count Data with ALDEx2'. *NAR Genomics and Bioinformatics* 7 (3): lqaf108. https://doi.org/10.1093/nargab/lqaf108.

Gruber-Vodicka, Harald R., Nikolaus Leisch, Manuel Kleiner, et al. 2019. 'Two Intracellular and Cell Type-Specific Bacterial Symbionts in the Placozoan *Trichoplax* H2'. *Nature Microbiology* 4 (9): 1465–74. https://doi.org/10.1038/s41564-019-0475-9.

Janitza, Silke, Ender Celik, and Anne-Laure Boulesteix. 2018. 'A Computationally Fast Variable Importance Test for Random Forests for High-Dimensional Data'. *Advances in Data Analysis and Classification* 12 (4): 885–915. https://doi.org/10.1007/s11634-016-0276-4.

Jonsson, Viktor, Tobias Österlund, Olle Nerman, and Erik Kristiansson. 2016. 'Statistical Evaluation of Methods for Identification of Differentially Abundant Genes in Comparative Metagenomics'. *BMC Genomics* 17 (1): 78. https://doi.org/10.1186/s12864-016-2386-y.

Kassambara, A. 2023. *Rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R Package Version 0.7.2*. Released. https://CRAN.R-project.org/package=rstatix.

Kleiner, Manuel. 2017. 'Normalization of Metatranscriptomic and Metaproteomic Data for Differential Gene Expression Analyses: The Importance of Accounting for Organism Abundance'. Preprint, March 2. https://doi.org/10.7287/peerj.preprints.2846v1.

Kleiner, Manuel. 2019. 'Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities'. *mSystems* 4 (3): e00115-19. https://doi.org/10.1128/mSystems.00115-19.

Langley, Sarah R., and Manuel Mayr. 2015. 'Comparative Analysis of Statistical Methods Used for Detecting Differential Expression in Label-Free Mass Spectrometry Proteomics'. *Journal of Proteomics* 129 (November): 83–92. https://doi.org/10.1016/j.jprot.2015.07.012.

Lazar, Cosmin, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. 2016. 'Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies'. *Journal of Proteome Research* 15 (4): 1116–25. https://doi.org/10.1021/acs.jproteome.5b00981.

Levi Mortera, Stefano, Valeria Marzano, Federica Rapisarda, et al. 2024. 'Metaproteomics Reveals Diet-Induced Changes in Gut Microbiome Function According to Crohn's Disease Location'. *Microbiome* 12 (1): 217. https://doi.org/10.1186/s40168-024-01927-5.

Li, Ming, William Gray, Haixia Zhang, et al. 2010. 'Comparative Shotgun Proteomics Using Spectral Count Data and Quasi-Likelihood Modeling'. *Journal of Proteome Research* 9 (8): 4295–305. https://doi.org/10.1021/pr100527g.

Li, Yumei, Xinzhou Ge, Fanglue Peng, Wei Li, and Jingyi Jessica Li. 2022. 'Exaggerated False Positives by Popular Differential Expression Methods When Analyzing Human Population Samples'. *Genome Biology* 23 (1): 79. https://doi.org/10.1186/s13059-022-02648-4.

Lim, Matthew Y., João A. Paulo, and Steven P. Gygi. 2019. 'Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model'. *Journal of Proteome Research* 18 (11): 4020–26. https://doi.org/10.1021/acs.jproteome.9b00492.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. 'Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2'. *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Malinowska, Agata, Michał Kistowski, Magda Bakun, et al. 2012. 'Diffprot — Software for Non-Parametric Statistical Analysis of Differential Proteomics Data'. *Journal of Proteomics* 75 (13): 4062–73. https://doi.org/10.1016/j.jprot.2012.05.030.

Mallick, Himel, Ali Rahnavard, Lauren J. McIver, et al. 2021. 'Multivariable Association Discovery in Population-Scale Meta-Omics Studies'. *PLOS Computational Biology* 17 (11): e1009442. https://doi.org/10.1371/journal.pcbi.1009442.

Martin, B, D Witten, and A Willis. 2022. *Corncob: Count Regression for Correlated Observations with the Beta-Binomial. R Package Version 0.3.1.* Released.

Martin, Bryan D., Daniela Witten, and Amy D. Willis. 2020. 'Modeling Microbial Abundances and Dysbiosis with Beta-Binomial Regression'. *The Annals of Applied Statistics* 14 (1). https://doi.org/10.1214/19-AOAS1283.

Mueller, Ryan S., Brian D. Dill, Chongle Pan, et al. 2011. 'Proteome Changes in the Initial Bacterial Colonist during Ecological Succession in an Acid Mine Drainage Biofilm Community'. *Environmental Microbiology* 13 (8): 2279–92. https://doi.org/10.1111/j.1462-2920.2011.02486.x.

Nearing, Jacob T., Gavin M. Douglas, Molly G. Hayes, et al. 2022. 'Microbiome Differential Abundance Methods Produce Different Results across 38 Datasets'. *Nature Communications* 13 (1): 342. https://doi.org/10.1038/s41467-022-28034-z.

Nesvizhskii, Alexey I., and Ruedi Aebersold. 2005. 'Interpretation of Shotgun Proteomic Data'. *Molecular & Cellular Proteomics* 4 (10): 1419–40. https://doi.org/10.1074/mcp.R500012-MCP200.

Nishijima, Suguru, Evelina Stankevic, Oliver Aasmets, et al. 2025. 'Fecal Microbial Load Is a Major Determinant of Gut Microbiome Variation and a Confounder for Disease Associations'. *Cell* 188 (1): 222-236.e15. https://doi.org/10.1016/j.cell.2024.10.022.

Nixon, Michelle Pistner, Gregory B. Gloor, and Justin D. Silverman. 2025. 'Incorporating Scale Uncertainty in Microbiome and Gene Expression Analysis as an Extension of Normalization'. *Genome Biology* 26 (1): 139. https://doi.org/10.1186/s13059-025-03609-3.

Oberg, Ann L., and Olga Vitek. 2009. 'Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments'. *Journal of Proteome Research* 8 (5): 2144–56. https://doi.org/10.1021/pr8010099.

Palarea-Albaladejo, Javier, and Josep Antoni Martín-Fernández. 2015. 'zCompositions — R Package for Multivariate Imputation of Left-Censored Data under a Compositional Approach'. *Chemometrics and Intelligent Laboratory Systems* 143 (April): 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019.

Plancade, Sandra, Magali Berland, Mélisande Blein-Nicolas, Olivier Langella, Ariane Bassignani, and Catherine Juste. 2022. 'A Combined Test for Feature Selection on Sparse Metaproteomics Data—an Alternative to Missing Value Imputation'. *PeerJ* 10 (June): e13525. https://doi.org/10.7717/peerj.13525.

Pursiheimo, Anna, Anni P. Vehmas, Saira Afzal, et al. 2015. 'Optimization of Statistical Methods Impact on Quantitative Proteomics Data'. *Journal of Proteome Research* 14 (10): 4118–26. https://doi.org/10.1021/acs.jproteome.5b00183.

R Core Team. 2023. 'R: A Language and Environment for Statistical Computing'. *R Foundation for Statistical Computing, Vienna, Austria*.

Rajczewski, Andrew T., J. Alfredo Blakeley□Ruiz, Annaliese Meyer, et al. 2025. 'Data□Independent Acquisition Mass Spectrometry as a Tool for Metaproteomics: Interlaboratory Comparison Using a Model Microbiome'. *PROTEOMICS* 25 (9–10): e202400187. https://doi.org/10.1002/pmic.202400187.

Ramus, Claire, Agnès Hovasse, Marlène Marcellin, et al. 2016. 'Benchmarking Quantitative Label-Free LC–MS Data Processing Workflows Using a Complex Spiked Proteomic Standard Dataset'. *Journal of Proteomics* 132 (January): 51–62. https://doi.org/10.1016/j.jprot.2015.11.011.

Ritchie, Matthew E., Belinda Phipson, Di Wu, et al. 2015. 'Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies'. *Nucleic Acids Research* 43 (7): e47–e47. https://doi.org/10.1093/nar/gkv007.

Sasaki, Y. 2007. *The Truth of the F-Measure*. Teach Tutor Mater.

Schurch, Nicholas J., Pietá Schofield, Marek Gierliński, et al. 2016. 'How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?' *RNA* 22 (6): 839–51. https://doi.org/10.1261/rna.053959.115.

Sitarz, Mikolaj. 2022. 'Extending F1 Metric, Probabilistic Approach'. arXiv:2210.11997. Preprint, arXiv, October 26. https://doi.org/10.48550/arXiv.2210.11997.

Tyanova, Stefka, Tikira Temu, and Juergen Cox. 2016. 'The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics'. *Nature Protocols* 11 (12): 2301–19. https://doi.org/10.1038/nprot.2016.136.

Välikangas, Tommi, Tomi Suomi, and Laura L. Elo. 2016. 'A Systematic Evaluation of Normalization Methods in Quantitative Label-Free Proteomics'. *Briefings in Bioinformatics*, October 2, bbw095. https://doi.org/10.1093/bib/bbw095.

Van Den Bossche, Tim, Jean Armengaud, Dirk Benndorf, et al. 2025. 'The Microbiologist's Guide to Metaproteomics'. *iMeta* 4 (3): e70031. https://doi.org/10.1002/imt2.70031.

Vorland, Colby J., Lilian Golzarri-Arroyo, and David B. Allison. 2025. 'A Brief Guide to Statistical Analysis of Grouped Data in Preclinical Research'. *Nature Metabolism* 7 (7): 1301–4. https://doi.org/10.1038/s42255-025-01323-9.

Wagner, Maggie R., and Manuel Kleiner. 2025. 'How Thoughtful Experimental Design Can Empower Biologists in the Omics Era'. *Nature Communications* 16 (1): 7263. https://doi.org/10.1038/s41467-025-62616-x.

Webb-Robertson, Bobbie-Jo M., Holli K. Wiberg, Melissa M. Matzke, et al. 2015. 'Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics'. *Journal of Proteome Research* 14 (5): 1993–2001. https://doi.org/10.1021/pr501138h.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, et al. 2017. 'Normalization and Microbial Differential Abundance Strategies Depend upon Data Characteristics'. *Microbiome* 5 (1): 27. https://doi.org/10.1186/s40168-017-0237-y.

Welch, B. L. 1947. 'The Generalization of `Student's' Problem When Several Different Population Variances Are Involved'. *Biometrika* 34 (1/2): 28. https://doi.org/10.2307/2332510.

Wilcoxon, Frank. 1945. 'Individual Comparisons by Ranking Methods'. *Biometrics Bulletin* 1 (6): 80. https://doi.org/10.2307/3001968.

Wilmes, Paul, and Philip L. Bond. 2004. 'The Application of Two□dimensional Polyacrylamide Gel Electrophoresis and Downstream Analyses to a Mixed Community of Prokaryotic Microorganisms'. *Environmental Microbiology* 6 (9): 911–20. https://doi.org/10.1111/j.1462-2920.2004.00687.x.

Wolski, Witold E., Paolo Nanni, Jonas Grossmann, Maria d'Errico, Ralph Schlapbach, and Christian Panse. 2023. '*Prolfqua*□: A Comprehensive *R* -Package for Proteomics Differential Expression Analysis'. *Journal of Proteome Research* 22 (4): 1092–104. https://doi.org/10.1021/acs.jproteome.2c00441.

Yang, Yin, Jingqiu Cheng, Shisheng Wang, and Hao Yang. 2022. 'StatsPro: Systematic Integration and Evaluation of Statistical Approaches for Detecting Differential Expression in Label-Free Quantitative Proteomics'. *Journal of Proteomics* 250 (January): 104386. https://doi.org/10.1016/j.jprot.2021.104386.

Yerke, Aaron, Daisy Fry Brumit, and Anthony A. Fodor. 2024. 'Proportion-Based Normalizations Outperform Compositional Data Transformations in Machine Learning Applications'. *Microbiome* 12 (1): 45. https://doi.org/10.1186/s40168-023-01747-z.

Yu, Fengchao, Sarah E. Haynes, and Alexey I. Nesvizhskii. 2021. 'IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs'. *Molecular & Cellular Proteomics* 20: 100077. https://doi.org/10.1016/j.mcpro.2021.100077.

Zhu, Yafeng, Lukas M. Orre, Yan Zhou Tran, et al. 2020. 'DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis'. *Molecular & Cellular Proteomics* 19 (6): 1047–57. https://doi.org/10.1074/mcp.TIR119.001646.