llperm: A Permutation of Regressor Residuals Test for Microbiome data

Markus Viljanen (markus.viljanen@rivm.nl)
National Institute for Public Health and the Environment - RIVM
Hendriek Boshuizen
National Institute for Public Health and the Environment - RIVM

Research Article

Keywords: microbiome, bioinformatics, statistics

Posted Date: June 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1669365/v1

RESEARCH

llperm: A Permutation of Regressor Residuals Test for Microbiome data

Markus Viljanen* and Hendriek Boshuizen

*Correspondence:
markus.viljanen@rivm.nl
National Institute for Public Health and the Environment - RIVM, PO Box 1, 3720BA Bilthoven, Netherlands
Full list of author information is available at the end of the article

Abstract

Background: Differential abundance testing is an important aspect of microbiome data analysis, where each taxa is fitted with a statistical test or a regression model. However, many models do not provide a good fit to real microbiome data. This has been shown to result in high false discovery rates. Permutation tests are a good alternative, but a regression approach is desired for small data sets with many covariates, where stratification is not an option.

Results: We present an extension of the Permutation of Regressor Residuals (PRR) test suitable for microbiome data, and a new R package 'llperm' which implements popular regression models in this framework. Simulations based on real data show that the approach outperforms the likelihood based models in both Power and False Discovery Rate. The PRR-test approach is able to maintain the correct nominal false positive rate expected from the null hypothesis, while having equal or greater power to detect the true positives as models based on likelihood at a given false positive rate.

Conclusions: The PRR-test was shown to provide a useful new tool for microbiome data analysis. Likelihood models can have a shockingly high rate of false positives and it is not possible to adjust for this in real data sets where the ground truth is unknown. As standard models may not provide a good fit to data, robustness gained from this approach can be viewed as a major benefit.

Keywords: microbiome; bioinformatics; statistics

1 Introduction

Statistical tools and computational methods have an important role in analysing microbiome data. Modern microbiome data sets are created by sequencing marker genes or the entire metagenome in a sample, and mapping these sequences to operational taxonomic units (OTUs), amplicon sequence variants (ASVs), or species or other phylogenetic levels [1]. We refer to these microbiome units as taxa regardless of the aggregation level. A data set typically has hundreds to thousands of taxa and comparatively few samples. The sample is described by sampling unit (e.g. subject) and environmental characteristics. These additional variables are important because the microbiome (unlike the genome) can both modify and be modified by these factors [2].

The goal of statistical analysis is to identify associations between the microbiome and biological, environmental, genetic, clinical or experimental conditions, while taking into account possible confounding factors [3]. The research hypothesis is typically formulated as a null hypothesis, such as "There is no difference in the microbiome composition of comparison groups". Several different types of analyses

can be considered. A common statistical analysis of microbiome data is differential abundance (DA) testing, where each taxon is sequentially tested for a difference in taxon abundance given the experimental groups and covariates in the sample [4].

Classic statistical tests, such as Pearson correlation, T-test and ANOVA, are used to compare groups in microbiome data [5, 6, 7, 8] even though the distributional assumptions can be suspect. When there are covariates under consideration, standard regression approaches have become popular tools. The Negative Binomial distribution, and packages like edgeR and DESeq2 based on it, are sometimes recommended for microbiome data. [9, 10]. While simulation studies show good performance, it has been pointed out that more realistic data do not satisfy their distributional assumptions [2, 11]. This can result in many false positives, implying the methods have a poor False Discovery Rate (FDR) control [9, 12, 10, 13, 4, 2, 14, 15].

Permutation tests provide a robust approach for a comparison of experimental groups because the FDR is maintained at the nominal level [16]. With a limited number of confounding factors, stratification can be employed [17]. However, if the data set has a small sample size and multiple covariates, a regression approach with similar robustness properties as a permutation test is desired. In this paper, we present an extension of the Permutation of Regressor Residuals Test (PRR-test) [18] suitable for microbiome data. We show that this method controls the FDR within the regression approach, enabling a robust test of comparison groups or environmental gradients while taking into account the covariates.

2 Methods

2.1 Testing differential abundance

For person $i = 1, ..., n$ and taxa $j = 1, ..., m$, define the detected counts $Y_{i,j}$ as a matrix $Y \in \mathbf{N}^{n \times m}$. Our goal is to detect the differentially abundant taxa, which we denote by the binary vector $y^* \in \{0,1\}^m$ where $y_j^* = \mathbb{I}(\text{taxa } j \text{ is differentially abundant})$. The null hypothesis is that there is no difference in the counts of a taxa between the experimental groups. We test hundreds of taxon j and obtain a vector of p-values $p_j \in [0,1]^m$ from a single experiment. A good statistical hypothesis test should have the ability to 1) control the probability of a type I error (false positive result) at the nominal significance level γ, and 2) have sufficient power (i.e. true positive rate) for detecting the differentially abundant taxa. [11]. We quantify the FPR and Power of the test with:

$$\text{FPR} = \frac{\sum_{j=1}^{m} \mathbb{I}(p_i < \gamma \text{ and } y_j^* = 0)}{\sum_{j=1}^{m} \mathbb{I}(y_j^* = 0)}$$
$$\text{TPR} = \frac{\sum_{j=1}^{m} \mathbb{I}(p_i < \gamma \text{ and } y_j^* = 1)}{\sum_{j=1}^{m} \mathbb{I}(y_j^* = 1)}$$

2.2 Model definition

Given that microbiome data often contain many zero counts, we define both single distribution models as well as zero-inflated models, consisting of a part modelling the probability of a zero ('zero' component) and a part modelling the number of counts (the 'count'component). For the zero-inflated model, define the 'count' component related covariates as a matrix $X \in \mathbf{R}^{n \times (p+q)}$ and 'zero' component related covariates as a matrix $Z \in \mathbf{R}^{n \times (s+t)}$, with the corresponding coefficients $\alpha \in \mathbf{R}^{p+q}$ and $\beta \in \mathbf{R}^{s+t}$. This generalizes the simple case $X = Z$ where same covariates are considered

to influence both counts and zero-inflation, as well as the single distribution model where the 'zero' component is omitted. Define the likelihood function for taxa j:

$$L_j(Y, X, Z, \alpha, \beta) = \prod_{i=1}^{n} f(Y_{i,j}, X_{i,:}, Z_{j,:}, \alpha, \beta)$$

Denote the maximum likelihood solution $\hat{\alpha}, \hat{\beta}$:

$$\hat{\alpha}, \hat{\beta} := \text{argmax}_{\alpha,\beta} L_j(Y, X, Z, \alpha, \beta)$$

We factorize the matrices X and Z, and the corresponding coefficients α and β, into covariates of interest $X^{'} \in \mathbf{R}^{n \times p}, Z^{'} \in \mathbf{R}^{n \times s}$ and other covariates $X^{''} \in \mathbf{R}^{n \times q}, Z^{''} \in \mathbf{R}^{n \times t}$:

$$\begin{array}{ll} X = (X^{'}, X^{''}) & \alpha = (\alpha^{'}, \alpha^{''}) \\ Z = (Z^{'}, Z^{''}) & \beta = (\beta^{'}, \beta^{''}) \end{array}$$

The null hypothesis is that the regression coefficients for the covariate of interest is zero for both components is $\alpha^{'} = 0$ and $\beta^{'} = 0$. It is also possible to test only one covariate of interest while taking into account the other in model fitting.

2.3 Permutation scheme

We explain how to calculate a likelihood and a permutation of regression residuals test [18] based p-value in three stages:

2.3.1 Calculate residuals for the covariate of interest from a least squares problem

The basic idea of the PRR test is that we replace the covariate of interest by their residual given by a linear regression on the remaining covariates. We first predict the covariate of interest $X^{'}$ from the other covariates $X^{''}$ by solving the least squares problem $\hat{\Sigma} := \text{argmin}_{\Sigma} \|X^{'} - X^{''}\Sigma\|^2$, and then we calculate the residuals $\tilde{X} := X^{'} - X^{''}\hat{\Sigma}$. While $X^{'}$ may be correlated with $X^{''}$, replacing it by the residual $\tilde{X}$ ensures that it is not correlated. The same is done with Z to obtain $\tilde{Z}$. The maximum value of the likelihood is the same with the residuals as it is with the covariates of interest. We then permute the residuals to estimate the null distribution and therefore the p-value. In case $X^{'}$ (and $Z^{'}$ when present) is a categorical variable with m categories, it is represented in the model matrix as a set of m-1 dummy variables, and the least squares problem consists of a system of m-1 regression equations, delivering m-1 residuals, which are used in place of the dummy variables.

2.3.2 For each resampling iteration, calculate p-values using the permuted residuals

For every resampling iteration $b = 1, ..., B$, use $\mathcal{I}_b(n)$ to denote a random permutation of row indexes $\{1, ..., n\}$. We then substitute the factorized matrices X and Z by matrices without/with the permuted residuals:

$$\begin{array}{ll} X^0 = (\tilde{X}, X^{''}) & X^b = (\tilde{X}_{\mathcal{I}_b(n),:}, X^{''}) \\ Z^0 = (\tilde{Z}, Z^{''}) & Z^b = (\tilde{Z}_{\mathcal{I}_b(n),:}, Z^{''}) \end{array}$$

The likelihood ratio pivotal has an asymptotic Chi-squared distribution, from which a p-value can be calculated:

$$p_{j,b} = \mathcal{X}^2_{p+q}\left(\frac{L_j(Y, X^b, Z^b, \hat{\alpha}^b, \hat{\beta}^b)}{L_j(Y, X^0, Z^0, \hat{\alpha}^0, \hat{\beta}^0)}\right)$$

where p and q are the number of columns in $X^{'}$ and $Z^{'}$ respectively, that is 1 in case of a continuous variable, and m-1 in case of an variable with m categories.

2.3.3 Calculate a p-value

First, a p-value based on likelihood can be calculated directly

$$\hat{p}_j = \mathcal{X}^2_{p+q}\left(\frac{L_j(Y, X^{''}, Z^{''}, \hat{\alpha}^{''}, \hat{\beta}^{''})}{L_j(Y, X^0, Z^0, \hat{\alpha}^0, \hat{\beta}^0)}\right)$$

Second, a p-value based on permutation of regression residuals can be calculated based on the resampling iterations:

$$p_j = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(p_{b,j} < \hat{p}_j)$$

2.4 Model regression specification

The likelihood of a single observation $f(Y_{i,j}, X_{i,:}, Z_{j,:}, \alpha, \beta)$ can have an arbitary specification in our model. We consider the following eight regression models in the experiments, which can be classified as Poisson or Binomial type models with zero-inflation and/or overdispersion (compound Gamma or Beta-distribution) illustrated in Table 1 and Table 2.

Table 1 Poisson family of models

		Overdispersion	
		no	yes
Zero-Inflation	no	Poisson	Negative Binomial
	yes	ZI Poisson	ZI Negative Binomial

Table 2 Binomial family of models

		Overdispersion	
		no	yes
Zero-Inflation	no	Binomial	Beta Binomial
	yes	ZI Binomial	ZI Beta Binomial

The raw abundance counts are not directly comparable across samples in real data sets. These counts do not directly reflect the true amount of DNA, but also sample DNA quality, concentration, amplification, barcoding and sequencing act as factors which make the taxa counts in a sample larger or smaller in an unpredictable way [19, 20]. Therefore the taxon abundance can only be analysed relative to the library sizes $s_i = \sum_{j=1}^{m} Y_{i,j}$ [4, 11]. This is directly incorporated in the Binomial distributions because the counts are drawn from the library size. Poisson distributions can include an offset($\log(s_i)$) term in the regression equation to include the library size.

2.5 Model implementation (llperm)

We propose an R package called 'llperm' that implements the model. Our package extends the 'glmperm' R package implemented by Werft (2010) [18], which in turns is an extension of 'logregperm' R package proposed by Potter (2005) [21]. The original package implemented the novel permutation test procedure for inference in logistic regression models, whereas the glmperm extended this into Generalized Linear Models (GLM) where more than one covariate can be involved together with the covariate of interest. Our package 'llperm' in turn extends this implementation in three ways to better fit microbiome data:

1. The covariate of interest can also occur as a categorical covariate with multiple levels.
2. We generalized the implementation to any likelihood based model, which enables additional distributions with zero-inflation and overdispersion (Poisson, ZIPoisson, NegBin, ZINegBin, Bin, ZIBin, BetaBin, ZIBetaBin, ...).
3. In case of zero-inflated models, the regression coefficients related to the count- and the zero-component can be simultaneously tested.

The model is easy to apply using the R formula syntax. We extend the original glmperm function (prr.test) and implemented one new function for GLMs (prr.test.glm) and another for likelihood models (prr.test.ll). Example:

```
# Fit 'glmperm' to Diet = Vegan (0/1)
fit <- prr.test(ASV159 ~ Diet01 + Age + Urbanization + Gender +
     Education + offset(log(library_size)), var = "Diet01",
     data=otu.counts, nrep=1000, family=poisson())

# Fit 'llperm' to Diet (4 groups)
fit <- prr.test.glm(ASV159 ~ Diet + Age + Urbanization + Gender +
     Education + offset(log(library_size)), test.var = "Diet",
     data=otu.counts, nrep=1000, family=Poisson())

# Fit 'llperm': Negative Binomial
fit <- prr.test.ll(ASV159 ~ Diet + Age +  Urbanization + Gender +
     Education + offset(log(library_size)), test.var = "Diet",
     data=otu.counts, nrep=1000, family=NegBin())

# Fit 'llperm': ZI Negative Binomial, test both count and zero
fit <- prr.test.ll(ASV159 ~ Diet + Age + Urbanization + Gender +
     Education + offset(log(library_size)) | Diet + Age +
     Urbanization + Gender + offset(log(library_size)),
     test.var = "Diet", which="both", data=otu.counts,
     nrep=1000, family=ZINegBin())

# Likelihood and permutation p-values
summary(fit)
```

3 Simulations

We performed simulations in order to validate our method. When validating a method, it is important that the simulations resemble real life situations, and not an artificial situation in which the assumptions of the method are met. In our case this means that we do not want to generate data according to a specific statistical distribution, but want to use a distribution as present in real data. The ground truth is unknown in a real data set, so it is not possible to compare model predictions for taxa that are truly differentially abundant between the groups. Therefore we use the real dataset as the foundation for generating simulated data where the 'signal', i.e. truly differentially abundant taxa, is known.

3.1 Real data underlying the simulation

The VEGA data set [22] studied the extent to which antibiotic resistant bacteria occur in vegetarians and non-vegetarians. Faecal samples were collected from volunteers and used to detect the Extended-Spectrum Beta-Lactamases (ESBL) producing bacteria, while 16s rRNA sequencing was used to see what microbiota were present. These data can also be used to study the relation between microbiota abundance and diet (vegan, meateater, fisheater, vegetarian), taking into account confounders such as sex, age, urbanization, pets at home, medication and travel history. The data set has 149 persons and 531 ASVs that occur in at least 10% of persons. The microbiome is therefore represented by a 149 x 531 table of counts. For example, the counts for 'ASV305' in Figure 1 could indicate some difference in diet groups.

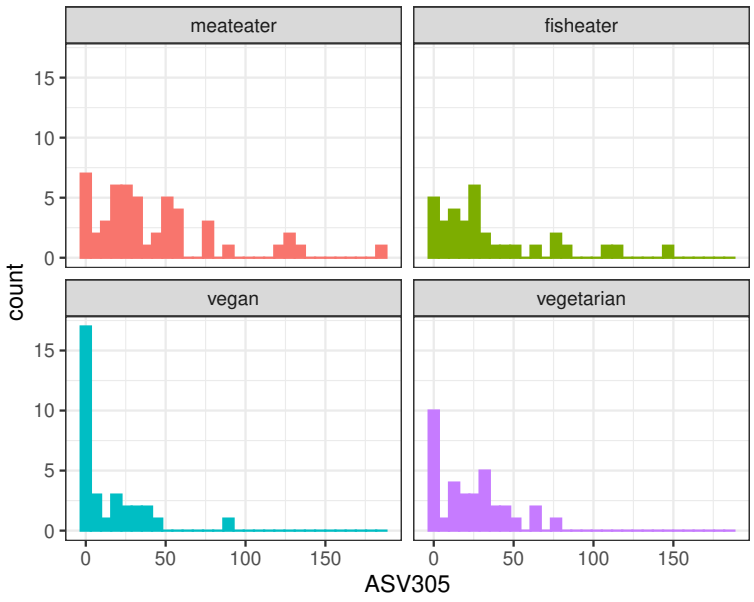

Figure 1 Example of raw data: 16S rRNA sequence counts for a single taxa 'ASV305', which appear overdispersed (mean 30, variance 1133) and zero-inflated, possibly influenced by Diet.

3.2 Simulated data

3.2.1 Adding signal to the real data

For each simulated dataset, we assigned each person in our data to one of 4 groups (meateater, fisheater, vegetarian, vegan) with equal 25% probability, irrespective of his/her real status. In each group, 10% of the taxa are randomly chosen to be differentially abundant. If a taxa is differentially abundant in a person, the counts are multiplied by an effect size (+25%, +50%, +100%, +200%, +400%). However, note that this only modifies non-zero counts.

We additionally introduced signal in the zero counts by decreasing their probability. For every taxon, we first calculated the baseline odds of the counts being non-zero, and assigned this to every individual. If the taxon is differentially abundant in a given person, this odds was multiplied by the effect size, and the probability of a non-zero sample was calculated from this increased odds. For the entire sample we then used this probability to draw whether or not the particular sample was non-zero, and if so we sampled without replacement a non-zero counts from the existing data. At some point the number of non-zero counts available for sampling are depleted (as we increased the probability of non-zero samples) and the remaining samples are assigned zero's. This implies that the counts remain the same but get shuffled so that the non-zero counts are more likely to occur in a sample where this taxon is differentially abundant. Each sample in this group then has an increased probability of a non-zero count, that is further multiplied by the effect size used.

3.2.2 Adding covariates

We made a similar simulated data set containing confounding factors. In addition to the diet, we included two additional simulated covariates for every subject: Urbanization (low/high) and Age (20-69). The effect of urbanization was simulated like that of diet: subjects were allocated to low/high urbanization and 10% of the taxa were made differentially abundant in both groups with an effect size +200%. Ages of 20, 21, ..., 69 were allocated to each subject and a differential effect was added for 10% of taxa with the effect depending linearly on age from 0% to 400%. These effects increase both the counts and the odds of non-zero counts. So there are three sources of signal to disentangle: different 10% of taxa are differentially abundant for each diet group, urbanization, and affected by age.

In order to act as confounders, urbanization and age need to be correlated to the diet group. Table 3 shows the probability of being assigned to a joint Diet and Urbanization group used to produce such a correlation, and Table 4 shows the probability of being assigned into a particular age range given diet group. We uniformly assigned age within this age range. Some taxa might now be detected as differentially abundant, not because the diet really influenced them, but because they also tended to have a different degree of urbanization and age.

Table 3 Joint probability of Diet and Urbanization

		Diet			
		meateater	fisheater	vegetarian	vegan
Urbanization	low	20%	15%	10%	5%
	high	5%	10%	15%	20%

Table 4 Conditional probability of Age given Diet

		Diet			
		meateater	fisheater	vegetarian	vegan
Age	[20,30)	0%	10%	30%	40%
	[30,40)	10%	15%	25%	30%
	[40,50)	20%	20%	20%	20%
	[50,60)	30%	25%	15%	10%
	[60,70)	40%	30%	10%	0%

4 Results

We first compare the 4 diet groups (meateater, fisheater, vegetarian, vegan) in the simulations without confounding factors, and then we introduce Urbanization (low/high) and Age (20-60) as additional covariates which introduce confounding, as described in the section Simulations. To both data sets we either introduce signal only in the counts, or in both counts & zeros. This results in four experiments that we run 50 times each to eliminate sampling variation. All experiments were run in parallel on a high-performance RedHat 7.9 LSF Linux cluster with R version 4.0.5.

In each experiment, we compare the likelihood model and the PRR-test by presenting the following four metrics:

1 True Positive Rate (TPR) at a p-value = 0.05 threshold.
2 False Positive Rate (FPR) at a p-value = 0.05 threshold.
3 Power when the p-value is chosen such that true FPR = 0.05 (Power@0.05),
4 Area Under the ROC Curve up to the FPR = 0.10 (AUC@0.10), normalized by the maximum area attainable.

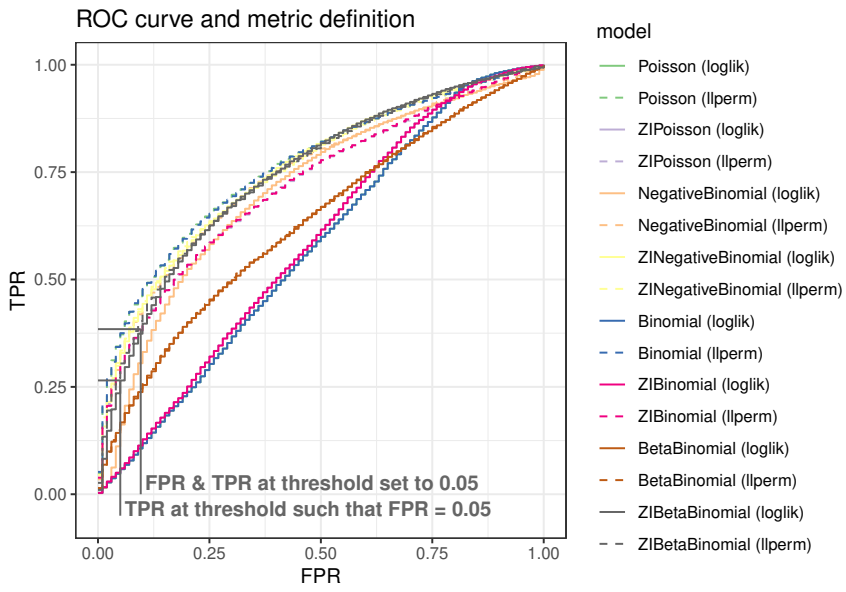

Figure 2 ROC curve on diet groups with confounding variables, signal in both counts & zeros. ZIBetaBinomial (loglik) has a TPR 0.38 and 0.10 FPR at a 0.05 p-value threshold. If we set the threshold such that FPR equals 0.05, the TPR@0.05 is 0.27. We can similarly calculate the AUC@0.10 as area under TPR over FPR values 0.00-0.10.

These are illustrated by the ROC curve in Figure 2. Note that Power@0.05 and AUC@0.10 can not be calculated in real data, because we cannot set the threshold

at a given FPR rate without knowing the truly differentially abundant taxa, but can be calculated from simulations.

4.1 Group comparison without confounding

For the first experiment, we try to detect taxa that are differentially abundant in a simple comparison of Diet groups without confounding variables. The likelihood and PRR-test based model statistics are shown in Table 5 for signal in counts and Table 6 for signal in counts & zeros given an effect size +100% .

Table 5 Model comparison on diet groups (signal in counts, effect size +100%)

Family	Type	Power	FPR	Power@0.05	AUC@0.10
Poisson	(loglik)	0.99 (±0.00)	0.98 (±0.00)	0.09 (±0.00)	0.52 (±0.01)
Poisson	(llperm)	0.29 (±0.01)	0.05 (±0.00)	0.28 (±0.01)	0.72 (±0.01)
ZIPoisson	(loglik)	0.99 (±0.00)	0.93 (±0.00)	0.11 (±0.00)	0.52 (±0.01)
ZIPoisson	(llperm)	0.40 (±0.00)	0.05 (±0.00)	0.40 (±0.01)	0.77 (±0.01)
NegativeBinomial	(loglik)	0.11 (±0.00)	0.02 (±0.00)	0.21 (±0.01)	0.66 (±0.01)
NegativeBinomial	(llperm)	0.26 (±0.00)	0.05 (±0.00)	0.26 (±0.01)	0.70 (±0.01)
ZINegativeBinomial	(loglik)	0.49 (±0.01)	0.15 (±0.00)	0.28 (±0.01)	0.63 (±0.01)
ZINegativeBinomial	(llperm)	0.34 (±0.00)	0.05 (±0.00)	0.34 (±0.01)	0.75 (±0.01)
Binomial	(loglik)	0.99 (±0.00)	0.98 (±0.00)	0.09 (±0.00)	0.52 (±0.01)
Binomial	(llperm)	0.29 (±0.01)	0.05 (±0.00)	0.28 (±0.01)	0.72 (±0.01)
ZIBinomial	(loglik)	0.99 (±0.00)	0.93 (±0.00)	0.11 (±0.00)	0.52 (±0.01)
ZIBinomial	(llperm)	0.41 (±0.00)	0.05 (±0.00)	0.40 (±0.01)	0.76 (±0.01)
BetaBinomial	(loglik)	0.14 (±0.00)	0.09 (±0.00)	0.09 (±0.00)	0.59 (±0.01)
BetaBinomial	(llperm)	0.08 (±0.00)	0.05 (±0.00)	0.08 (±0.00)	0.59 (±0.01)
ZIBetaBinomial	(loglik)	0.33 (±0.01)	0.05 (±0.00)	0.33 (±0.01)	0.71 (±0.01)
ZIBetaBinomial	(llperm)	0.34 (±0.01)	0.05 (±0.00)	0.35 (±0.01)	0.74 (±0.01)

Table 6 Model comparison on diet groups (signal in counts & zeros, effect size +100%)

Family	Type	Power	FPR	Power@0.05	AUC@0.10
Poisson	(loglik)	1.00 (±0.00)	0.98 (±0.00)	0.12 (±0.00)	0.54 (±0.01)
Poisson	(llperm)	0.49 (±0.01)	0.05 (±0.00)	0.48 (±0.01)	0.78 (±0.01)
ZIPoisson	(loglik)	0.99 (±0.00)	0.93 (±0.00)	0.11 (±0.00)	0.53 (±0.01)
ZIPoisson	(llperm)	0.39 (±0.01)	0.05 (±0.00)	0.39 (±0.01)	0.75 (±0.01)
NegativeBinomial	(loglik)	0.17 (±0.00)	0.02 (±0.00)	0.32 (±0.01)	0.66 (±0.01)
NegativeBinomial	(llperm)	0.44 (±0.01)	0.05 (±0.00)	0.43 (±0.01)	0.73 (±0.01)
ZINegativeBinomial	(loglik)	0.59 (±0.00)	0.15 (±0.00)	0.37 (±0.01)	0.69 (±0.01)
ZINegativeBinomial	(llperm)	0.42 (±0.00)	0.05 (±0.00)	0.42 (±0.01)	0.74 (±0.01)
Binomial	(loglik)	1.00 (±0.00)	0.98 (±0.00)	0.12 (±0.00)	0.54 (±0.01)
Binomial	(llperm)	0.49 (±0.01)	0.05 (±0.00)	0.48 (±0.01)	0.78 (±0.01)
ZIBinomial	(loglik)	0.99 (±0.00)	0.93 (±0.00)	0.11 (±0.00)	0.53 (±0.01)
ZIBinomial	(llperm)	0.39 (±0.00)	0.05 (±0.00)	0.39 (±0.01)	0.75 (±0.01)
BetaBinomial	(loglik)	0.28 (±0.01)	0.10 (±0.00)	0.19 (±0.01)	0.64 (±0.01)
BetaBinomial	(llperm)	0.20 (±0.01)	0.05 (±0.00)	0.20 (±0.01)	0.64 (±0.01)
ZIBetaBinomial	(loglik)	0.39 (±0.00)	0.05 (±0.00)	0.39 (±0.01)	0.72 (±0.01)
ZIBetaBinomial	(llperm)	0.41 (±0.00)	0.05 (±0.00)	0.41 (±0.01)	0.73 (±0.01)

Most likelihood based models without overdispersion have shockingly high false positive rates: over 90% of non-differentially abundant taxa are detected as false positives for (ZI)Binomial and (ZI) Poisson distributions. Overdispersed models do better, but still have too high false positive rates. Only ZIBetaBinomial produced the correct nominal 5% FPR, while having the power to detect 33% (counts) or 39% (counts&zeros) of the differentially abundant taxa. The PRR-test based models all had the correct nominal 5% FPR rate, and the zero-inflated models all had power of 34-41% (counts) or 39%-42% (counts&zeros) to detect the taxa. In a more realistic setting where signal occurs in both counts and zeros, the standard Binomial and Poisson models based on likelihood are useless but become surprisingly effective with the PRR-test, achieving 49% power and 5% FPR.

In Figure 4 in the appendix, these statistics are plotted as a function of the effect size. These findings are consistent with different effect sizes; a models' power increases as the effect size increases. With different effect sizes PRR-test based models maintain the correct nominal FPR, while likelihood based models maintain the high rate of false positives.

4.2 Group comparison with confounding

For the second experiment, we try to detect taxa that are differentially abundant between Diet groups with confounding variables. The likelihood and PRR-test based model statistics are shown in Table 7 for signal in counts and Table 8 for signal in counts & zeros given an effect size +100% .

Table 7 Model comparison on diet groups with confounding variables (signal in counts, effect size +100%)

Family	**Type**	**Power**	**FPR**	**Power@0.05**	**AUC@0.10**
Family	Type	Power	FPR	Power@0.05	AUC@0.10
Poisson	(loglik)	0.99 (±0.00)	0.98 (±0.00)	0.09 (±0.00)	0.53 (±0.01)
Poisson	(llperm)	0.22 (±0.01)	0.05 (±0.00)	0.22 (±0.01)	0.69 (±0.01)
ZIPoisson	(loglik)	0.99 (±0.00)	0.94 (±0.00)	0.10 (±0.01)	0.53 (±0.01)
ZIPoisson	(llperm)	0.32 (±0.01)	0.05 (±0.00)	0.32 (±0.01)	0.72 (±0.01)
NegativeBinomial	(loglik)	0.11 (±0.00)	0.04 (±0.00)	0.15 (±0.01)	0.56 (±0.01)
NegativeBinomial	(llperm)	0.22 (±0.01)	0.06 (±0.00)	0.21 (±0.01)	0.66 (±0.01)
ZINegativeBinomial	(loglik)	0.46 (±0.01)	0.16 (±0.00)	0.27 (±0.01)	0.69 (±0.01)
ZINegativeBinomial	(llperm)	0.29 (±0.01)	0.05 (±0.00)	0.29 (±0.01)	0.72 (±0.01)
Binomial	(loglik)	0.99 (±0.00)	0.98 (±0.00)	0.09 (±0.00)	0.53 (±0.01)
Binomial	(llperm)	0.22 (±0.01)	0.05 (±0.00)	0.22 (±0.01)	0.69 (±0.01)
ZIBinomial	(loglik)	0.99 (±0.00)	0.94 (±0.00)	0.10 (±0.01)	0.53 (±0.01)
ZIBinomial	(llperm)	0.32 (±0.01)	0.05 (±0.00)	0.32 (±0.01)	0.73 (±0.01)
BetaBinomial	(loglik)	0.14 (±0.01)	0.11 (±0.00)	0.08 (±0.00)	0.59 (±0.01)
BetaBinomial	(llperm)	0.08 (±0.00)	0.06 (±0.00)	0.07 (±0.00)	0.58 (±0.01)
ZIBetaBinomial	(loglik)	0.34 (±0.01)	0.10 (±0.00)	0.23 (±0.01)	0.64 (±0.01)
ZIBetaBinomial	(llperm)	0.26 (±0.00)	0.05 (±0.00)	0.26 (±0.01)	0.68 (±0.01)

Table 8 Model comparison on diet groups with confounding variables (signal in counts & zeros, effect size +100%)

Family	**Type**	**Power**	**FPR**	**Power@0.05**	**AUC@0.10**
Family	Type	Power	FPR	Power@0.05	AUC@0.10
Poisson	(loglik)	1.00 (±0.00)	0.99 (±0.00)	0.11 (±0.00)	0.53 (±0.01)
Poisson	(llperm)	0.37 (±0.01)	0.05 (±0.00)	0.37 (±0.01)	0.74 (±0.01)
ZIPoisson	(loglik)	0.99 (±0.00)	0.94 (±0.00)	0.10 (±0.01)	0.54 (±0.01)
ZIPoisson	(llperm)	0.30 (±0.01)	0.05 (±0.00)	0.30 (±0.01)	0.72 (±0.01)
NegativeBinomial	(loglik)	0.17 (±0.00)	0.05 (±0.00)	0.16 (±0.01)	0.48 (±0.01)
NegativeBinomial	(llperm)	0.34 (±0.01)	0.05 (±0.00)	0.33 (±0.01)	0.71 (±0.01)
ZINegativeBinomial	(loglik)	0.52 (±0.01)	0.15 (±0.00)	0.33 (±0.01)	0.70 (±0.01)
ZINegativeBinomial	(llperm)	0.34 (±0.01)	0.05 (±0.00)	0.34 (±0.01)	0.71 (±0.01)
Binomial	(loglik)	1.00 (±0.00)	0.99 (±0.00)	0.11 (±0.00)	0.53 (±0.01)
Binomial	(llperm)	0.37 (±0.01)	0.05 (±0.00)	0.38 (±0.01)	0.74 (±0.01)
ZIBinomial	(loglik)	0.99 (±0.00)	0.94 (±0.00)	0.10 (±0.01)	0.54 (±0.01)
ZIBinomial	(llperm)	0.30 (±0.01)	0.05 (±0.00)	0.30 (±0.01)	0.72 (±0.01)
BetaBinomial	(loglik)	0.26 (±0.00)	0.10 (±0.00)	0.17 (±0.00)	0.63 (±0.01)
BetaBinomial	(llperm)	0.17 (±0.00)	0.05 (±0.00)	0.17 (±0.00)	0.63 (±0.01)
ZIBetaBinomial	(loglik)	0.38 (±0.01)	0.10 (±0.00)	0.27 (±0.01)	0.63 (±0.01)
ZIBetaBinomial	(llperm)	0.30 (±0.01)	0.05 (±0.00)	0.30 (±0.01)	0.67 (±0.01)

When confounding factors are added to the regression model, the situation is similar. As is to be expected, the models lose some Power when additional covariates are introduced, and have more Power when there is signal in both counts and zeros. Of the likelihood based models, only the Negative Binomial had a correct nominal 5% FPR rate with Power of 11% (counts) or 17% (counts&zeros). The PRR-test based models all had the correct nominal 5% FPR rate, and the zero-inflated models

had Power of 26-32% (counts) or 30-34% (counts&zeros). When there is signal in both counts and zeros, again the standard Binomial and Poisson models based on likelihood are useless but become surprisingly effective with the PRR-test, achieving 37% power and 5% FPR. In Figure 5 in the appendix, these statistics are again plotted as a function of the effect size.

Discussion

Models with overdispersion and zero-inflation are generally better in both likelihood and PRR-test based approaches. Findings are consistent with different effect sizes. The PRR-test not only controls the FDR but also seems to improve the Power in a regression setting. Surprisingly, in a more realistic setting where the signal in counts co-occurs with signal in zeros -both in the same direction -, the PRR-test makes even the standard Poisson and Binomial models perform well. It seems that zero-inflated models are most needed if a signal has been introduced only to counts, because the random variation in the occurrence of non-zero counts tends to obfuscate the signal, making the models without zero-inflation lose power.

The results generally align with previous literature, except for two findings. First, the standard Negative Binomial seems to have too low FPR (0.02) in the comparison of groups without confounding. We investigated that this seemed to be caused by some taxas having very high zero-inflation in our data. With a better fitting zero-inflated Negative Binomial we did observe the expected too high FPR. Also when we simulated data from Negative Binomial (instead of simulations based on real data) we observed a too high FPR. Second, the standard Beta Binomial has a very low Power and results differ - due to different assumptions on overdispersion - from that of using the Negative Binomial. We found that this model also has significant problems with excess zeros which regularly cause numerical problems. Sometimes the likelihood cannot even be evaluated outside a narrow neighbourhood of the solution, necessitating very accurate starting values for the optimization process.

One surprise in doing this work was that the glm.nb function from the MASS package converged to a different solution compared to our likelihood based implementation in some of the datasets (supplementary materials S1). The divergent glm.nb solutions had either very small or very large p-value, and was caused by a lack of convergence of the estimate of the overdispersion parameter, which went unnoticed as the function returned a converged status. This made the FPR even larger than with our implementation. With the exception of this issue with MASS, our results tended to be identical to those delivered by other packages.

We argued that simulating data by resampling a real data set provides more realistic results than simulating data from a known statistical distribution. However, our simulations are based on a single dataset. This might not fully reflect all possible data in microbiome studies. Also we assumed the original data set did not contain signal, so the data used for simulation might be more overdispersed than data that are truly without signal. Also, adding signal by multiplying the counts will increase the variance in simulated data. Nevertheless we believe our simulation gives a good indication of the relative merits of the different methods. We publish the data set, a reproducible R Markdown source code for the simulation experiments, and a simple implementation of the method in the Appendix.

Conclusion

The PRR-test was shown to provide useful new tools for microbiome data analysis. Standard regression models based on it are able to maintain the correct nominal false positive rate expected from the null hypothesis, while having equal or greater power to detect the true positives as models based on likelihood at a given false positive rate. Likelihood models can have a shockingly high rate of false positives and it is not possible to adjust for this in real data sets where the ground truth is unknown. This method therefore provides a new approach which is competitive in power, but also offer insurance against model misspecification. As standard models may not provide a good fit to data, so such robustness can be viewed as a major benefit.

Declarations

Availability of data and materials

The dataset and the source code for the experiments of this article are available in the repository 'llperm': https://github.com/majuvi/llperm

Acknowledgements

We gratefully acknowledge the researchers involved with the collection of the VEGA data set.

Author's contributions

MV wrote the draft of the manuscript, with contributions from HB. Both authors edited together and approved the final version. MV was responsible for data processing and experiments, with experimental design done jointly by both authors. HB proposed using the PRR-test for microbiome data, with a software implementation from MV.

Funding

RIVM Strategic Programme (SPR)

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Comparing packages

We verified our implementation by comparing it with other R-packages. Our results were virtual identical to likelihood based methods, and other methods tended to give the same results for almost all taxa. Exception was glm.nb from the MASS package, which produced many p-values close to 0 or 1 but otherwise agreed with our method. This is illustrated in Figure 3.

Upon investigating the reason, we found that MASS estimates the Negative Binomial distribution parameters in a two-stage process whereby first the parameters are estimated for a fixed overdispersion parameter and then the overdispersion is estimated given these parameters. This process did not converge for all taxa and sometimes indicated significant underdispersion, where overdispersion is 1/exp(theta), as illustrated by the very high thetas in Figure 3. Increasing the number of iterations would eventually crash the estimation procedure. Otherwise the package gave identical p-values and estimates of theta as indicated by the blue reference line. Our implementation tends to have equal or better Power/AUC than functions from other packages, as seen from the similar or lower p-values for taxa with signal.

An example of data for a taxon that causes this problem is given in Table 9 and the associated simple code listing below. Although the problem tends to occur in taxa with many zero-counts, it can also occur with non-zero counts if most counts are low but some are very high.

Table 9 Example: table of counts where MASS diverges

count	0	6	10	11	12	13	16	20	21	23	26	27	28	32	33	39	65
n	125	2	3	3	1	2	1	2	1	2	1	1	1	1	1	1	1

```
data = data.frame(
     N = sample(rep(count, n)),
     X = sample(c("A", "B"), sum(n), replace=T))
glm.nb(N~X, data, control=glm.control(maxit=50))
```

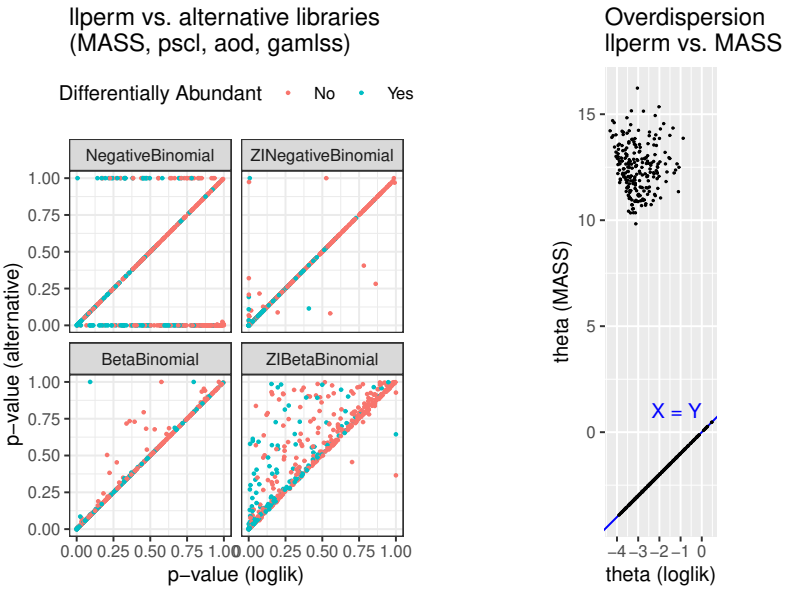

Figure 3 P-values from our implementation compared to those of other packages for taxa with and without signal (left); theta estimates from our implementation compared to those of glm.nb from MASS (right). Points in the cloud at the top indicating non-convergence of theta are the points delivering different p-values.

References

1. Hawinkel, S., Kerckhof, F.-M., Bijnens, L., Thas, O.: A unified framework for unconstrained and constrained ordination of microbiome read count data. PLoS One **14**(2), 0205474 (2019)
2. Mallick, H., Ma, S., Franzosa, E.A., Vatanen, T., Morgan, X.C., Huttenhower, C.: Experimental design and quantitative analysis of microbial community multiomics. Genome biology **18**(1), 1–16 (2017)
3. Xia, Y., Sun, J.: Hypothesis testing and statistical analysis of microbiome. Genes & diseases **4**(3), 138–148 (2017)
4. Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M.A., Stokholm, J., Al-Soud, W.A., Sørensen, S., Bisgaard, H., Waage, J.: Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rrna gene amplicon data analysis methods used in microbiome studies. Microbiome **4**(1), 1–14 (2016)
5. Chen, W., Liu, F., Ling, Z., Tong, X., Xiang, C.: Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PloS one **7**(6), 39743 (2012)
6. Kim, K.-A., Jung, I.-H., Park, S.-H., Ahn, Y.-T., Huh, C.-S., Kim, D.-H.: Comparative analysis of the gut microbiota in people with different levels of ginsenoside rb1 degradation to compound k. PloS one **8**(4), 62409 (2013)
7. Iwai, S., Fei, M., Huang, D., Fong, S., Subramanian, A., Grieco, K., Lynch, S.V., Huang, L.: Oral and airway microbiota in hiv-infected pneumonia patients. Journal of clinical microbiology **50**(9), 2995–3002 (2012)
8. Hsiao, E.Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F., *et al.*: The microbiota modulates gut physiology and behavioral abnormalities associated with autism. Cell **155**(7), 1451 (2013)
9. McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology **10**(4), 1003531 (2014)
10. Jonsson, V., Österlund, T., Nerman, O., Kristiansson, E.: Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC genomics **17**(1), 1–14 (2016)
11. Hawinkel, S., Rayner, J., Bijnens, L., Thas, O.: Sequence count data are poorly fit by the negative binomial distribution. PloS one **15**(4), 0224909 (2020)
12. Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D.: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health and disease **26**(1), 27663 (2015)
13. Jonsson, V., Österlund, T., Nerman, O., Kristiansson, E.: Variability in metagenomic count data and its influence on the identification of differentially abundant genes. Journal of Computational Biology **24**(4), 311–326 (2017)
14. Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y., Birmingham, A., *et al.*: Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome **5**(1), 1–18 (2017)
15. Hawinkel, S., Mattiello, F., Bijnens, L., Thas, O.: A broken promise: microbiome differential abundance methods do not control the false discovery rate. Briefings in bioinformatics **20**(1), 210–221 (2019)

16. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., Gloor, G.B.: 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), 1–13 (2014)
17. Ferreira, J.A.: Some models and methods for the analysis of observational data. Statistics Surveys **9**, 106–208 (2015)
18. Werft, W., Benner, A.: glmperm: A permutation of regressor residuals test for inference in generalized linear models. R J. **2**(1), 39 (2010)
19. Paliy, O., Shankar, V.: Application of multivariate statistical techniques in microbial ecology. Molecular ecology **25**(5), 1032–1057 (2016)
20. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J.: Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology **8**, 2224 (2017)
21. Potter, D.M.: A permutation test for inference in logistic regression with small-and moderate-sized data sets. Statistics in medicine **24**(5), 693–708 (2005)
22. Dierikx, C., van Duijkeren, E., Gijsbers, E., van Hoek, A., Hengeveld, P., de Greeff, S., Meijs, A., et al.: Onderzoek naar esbl-producerende bacterien onder vegetariers en niet-vegetariers: de vegastudie. RIVM Rapport (0150) (2017)

Figures

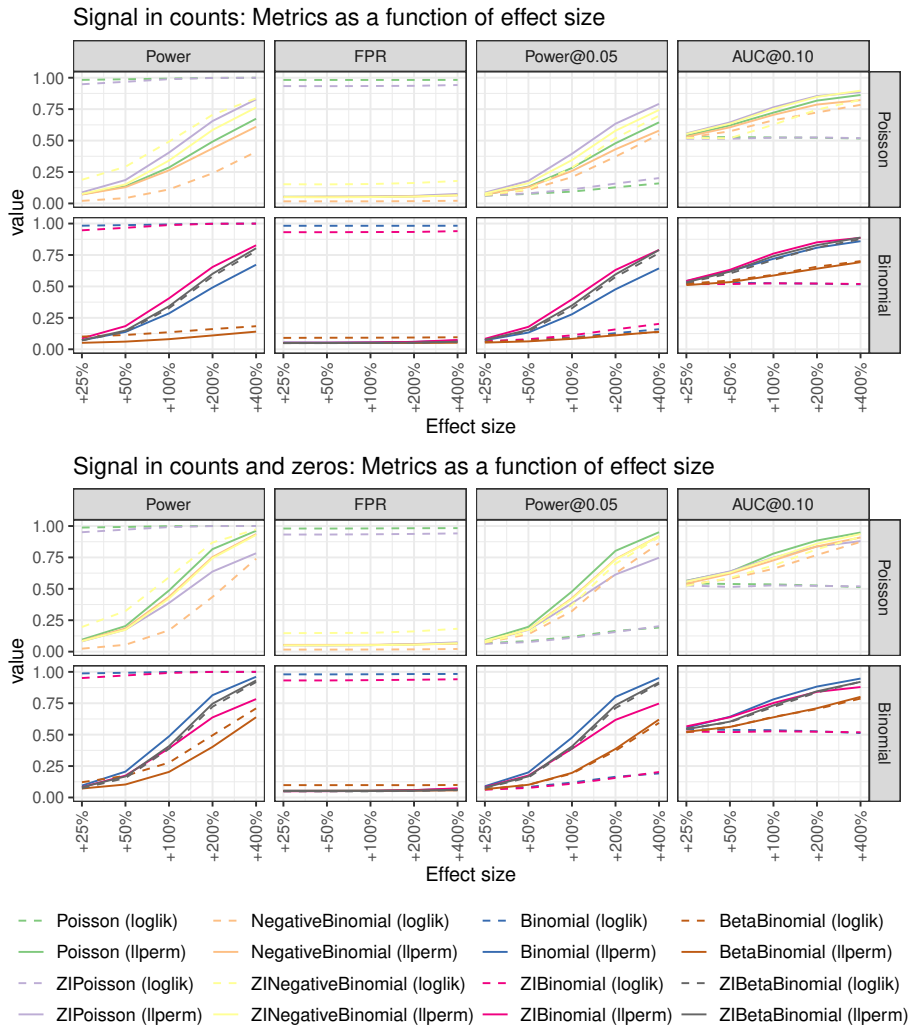

Figure 4 Model comparison on diet groups for different effect sizes (+25%, +50%, +100%, +200%, +400%)

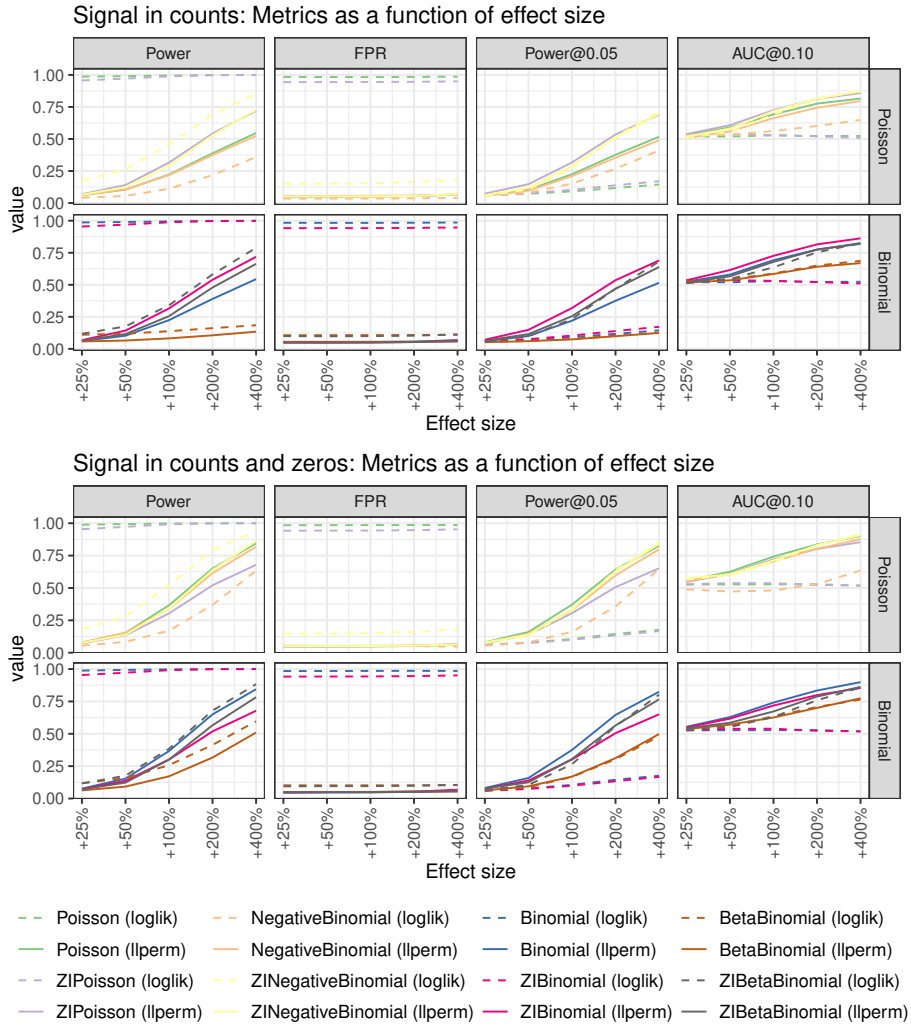

Figure 5 Model comparison on diet groups with confounding variables for different effect sizes (+25%, +50%, +100%, +200%, +400%)