BIBS Seminar

Viljanen and Boshuizen *BMC Bioinformatics* (2022) 23:540
https://doi.org/10.1186/s12859-022-05088-w

BMC Bioinformatics

RESEARCH Open Access

llperm: a permutation of regressor residuals test for microbiome data

Markus Viljanen* and Hendriek Boshuizen

Presenter: Mozaffar Hosain

14 September 2023

Contents

Introduction

- To analyze microbiome data, statistical tools and computational methods are necessary.
- Microbiome data often contain hundreds to thousands of taxa and only a small number of samples.
- The samples are characterized by various variables including the environmental factors, as the microbiome can both influence and be influenced by these factors.
- Statistical analysis of microbiome data identifies the associations between the microbiome and biological, environmental, genetic, clinical or experimental factors while accounting for potential confounding factors.

Introduction

- This analysis typically begins with the formulation of a null hypothesis, such as “There is no difference in the microbiome composition between comparison groups.”
- A common approach is differential abundance (DA) testing, which sequentially tests each taxon to determine if there are differences in taxon abundance between experimental groups.
- Pearson correlation, *t*-test, and ANOVA, are used to compare groups even though the distributional assumptions can be suspect.

Introduction

- Regression approaches are widely used when taking covariates into account.
- Two popular methods edgeR, and DESeq2 based on negative binomial distribution perform well for simulation studies but more realistic data do not satisfy their distributional assumptions.
- This may lead to many false positives, indicating that the methods do not effectively control the type 1 error rate.

Introduction

- Permutation tests offer a robust non-parametric method to compare experimental groups as the type 1 error rate is retained at the nominal level.
- In this article, a permutation-based method called Permutation of Regressor Residuals (PRR) is proposed which controls the type 1 error rate within the regression framework.
- Previously, others developed an R package called 'glmperm' for the Generalized Linear Model (GLM) family, but it lacks support for count regression with zero-inflation.
- The authors of this paper provided with an extended R package ‘llperm’ (Log-Likelihood) suitable for microbiome data, which implemented over dispersed and zero-inflated count regression models.

Methods

Testing differential abundance

- Let Y_{ij} be the count for subject i $(i = 1, 2, \ldots, n)$ and taxon j $(j = 1, 2, \ldots, m)$.
- The objective is to detect the differentially abundant taxa.
- The null hypothesis of interest is that there is no difference in the counts of a taxa between the experimental groups.
- The test is performed for all m taxa and obtain a vector of p-values $p_j \in [0, 1]^m$.

Methods

Model definition

- The approach in this article also takes into account 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).
- Define the ‘count’ component related covariates as a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times(p+q)}$ with p columns related to the covariate of interest and q other columns.
- The ‘zero’ component related covariates is also defined as a matrix $\mathbf{Z} \in \mathbb{R}^{n\times(s+t)}$ with s columns related to the covariate of interest and t other columns.
- The corresponding vectors of coefficients are $\boldsymbol{\alpha} \in \mathbb{R}^{(p+q)}$ and $\boldsymbol{\beta} \in \mathbb{R}^{(s+t)}$.

Methods

Model definition

- Let Y_i be the count for subject i $(i = 1, 2, \ldots, n)$ from any taxon.
  - $Y_i \sim f(y_i)$ with $E(Y_i) = \mu_i$, and link function for GLM: $g(\mu_i)$
  - Model without zero-inflation: $g(\mu_i) = \boldsymbol{X\alpha}$
  - Model with zero-inflation: Suppose p_i is the probability that Y_i is from the excess zero state.

    model for 'count' part: $g(\mu_i) = \boldsymbol{X\alpha}$

    model for 'zero' part: $\text{logit}(p_i) = \boldsymbol{Z\beta}$

Model definition

- For the simple case, $\boldsymbol{X} = \boldsymbol{Z}$, that is same covariates are considered to influence both counts and zero-inflation, as well as the model without zero component.
- The likelihood function for taxa j:

$$L(\boldsymbol{Y}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f(Y_{ij}, \boldsymbol{\alpha}, \boldsymbol{\beta})\ ;\ \boldsymbol{Y}_j \text{ is the vector of } n \text{ observations in taxon } j$$

- Maximum likelihood estimates $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\boldsymbol{\beta}}$ are such that

$$\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}} := \text{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}} L(\boldsymbol{Y}_j, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Model definition

- The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, and the corresponding vectors of parameters are factorized as:

$$\boldsymbol{X} = \left(\boldsymbol{X}^{*},\ \boldsymbol{X}^{\dagger}\right), \boldsymbol{\alpha} = \left(\boldsymbol{\alpha}^{*}, \boldsymbol{\alpha}^{\dagger}\right)$$

$$\boldsymbol{Z} = \left(\boldsymbol{Z}^{*},\ \boldsymbol{Z}^{\dagger}\right), \boldsymbol{\beta} = \left(\boldsymbol{\beta}^{*}, \boldsymbol{\beta}^{\dagger}\right),$$

where covariates of interest: $\boldsymbol{X}^{*} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Z}^{*} \in \mathbb{R}^{n\times s}$

other covariates: $\boldsymbol{X}^{\dagger} \in \mathbb{R}^{n\times q}$, $\boldsymbol{Z}^{\dagger} \in \mathbb{R}^{n\times t}$

- The null hypothesis is $\boldsymbol{\alpha}^{*} = \mathbf{0}$ and $\boldsymbol{\beta}^{*} = \mathbf{0}$.

Methods

Permutation scheme

- A *p*-value is calculated using both
  - ➢ a likelihood ratio test (LRT)
  - ➢ a permutation of regressor residuals (PRR)
- These are calculated in three different steps.

Permutation scheme

❑ Step 1: Calculate residuals for the covariate of interest from a least squares problem

- First predict the covariate of interest X^* by a least squares regression on the other covariates $\boldsymbol{X}^{\dagger}$.
- The residuals are calculated as: $\tilde{X} = X^* - \boldsymbol{X}^{\dagger}\hat{\Sigma}$, where Σ is the vector of parameters of the regression of X^* on $\boldsymbol{X}^{\dagger}$.
- X^* and $\boldsymbol{X}^{\dagger}$ may be correlated but X^* and $\tilde{X}$ are not.
- The residuals $\tilde{Z}$ can be obtained similarly.

Permutation scheme

❑ Step 1: Calculate residuals for the covariate of interest from a least squares problem

- The residuals are then permuted to obtain the *p*-values.
- When X^* (and Z^* if present) is a categorical variable with c categories, it is represented in the model matrix as a set of $(c-1)$ dummy variables.
- The least squares regression then consists of a system of $(c-1)$ regression equations.
- Thus $(c-1)$ residuals can be obtained, which are used in place of the dummy variables.

Permutation scheme

❑ Step 2: For each resampling iteration, calculate *p*-values using the permuted residuals

- The factorized matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ then can be expressed as:

without permutation: $\boldsymbol{X^0} = \left(\tilde{X},\ \boldsymbol{X^\dagger}\right), \boldsymbol{Z^0} = \left(\tilde{Z},\ \boldsymbol{Z^\dagger}\right)$

with the permuted residuals: $\boldsymbol{X^b} = \left(\tilde{X}_{I_b(n)},\ \boldsymbol{X^\dagger}\right), \boldsymbol{Z^b} = \left(\tilde{Z}_{I_b(n)},\ \boldsymbol{Z^\dagger}\right)$;

where $I_b(n)$ denotes a random permutation of row indices $\{1, 2, \ldots, n\}$ for the iteration b and $b = 1, 2, \ldots B$.

Permutation scheme

❑ Step 2: For each resampling iteration, calculate *p* values using the permuted residuals

- The likelihood ratio statistic has an asymptotic chi-square distribution from which the *p*-values can be obtained

  ➢ $(\hat{\boldsymbol{\alpha}}^{\mathbf{0}}, \hat{\boldsymbol{\beta}}^{\mathbf{0}})$ and $(\hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}, \hat{\boldsymbol{\beta}}^{\boldsymbol{b}})$ denote the OLS estimates based on unpermuted and permuted residuals respectively.

  ➢ $p_{j,\,b} = \chi^2_{p+s}\left(-2ln\left(\frac{L(\boldsymbol{Y_j}, \boldsymbol{X^b}, \boldsymbol{Z^b}, \hat{\boldsymbol{\alpha}}^{\boldsymbol{b}}, \hat{\boldsymbol{\beta}}^{\boldsymbol{b}})}{L(\boldsymbol{Y_j}, \boldsymbol{X^0}, \boldsymbol{Z^0}, \hat{\boldsymbol{\alpha}}^{\mathbf{0}}, \hat{\boldsymbol{\beta}}^{\mathbf{0}})}\right)\right)$,

  where $p =$ # of columns in X^*, $s =$ # of columns in Z^*,

  for a continuous variable, $p = s = 1$, and $(c-1)$ for a categorical variable with c categories.

Permutation scheme

❑ Step 3: Calculate a *p*-value of the permutation test

- For *j*th taxon, a *p*-value based on the LRT is calculated first

$$\hat{p}_j = \chi^2_{p+s}\left(-2ln\left(\frac{L\left(\boldsymbol{Y_j}, \boldsymbol{X}^{\dagger}, \boldsymbol{Z}^{\dagger}, \hat{\boldsymbol{\alpha}}^{\dagger}, \hat{\boldsymbol{\beta}}^{\dagger}\right)}{L\left(\boldsymbol{Y_j}, \boldsymbol{X^0}, \boldsymbol{Z^0}, \hat{\boldsymbol{\alpha}}^{\boldsymbol{0}}, \hat{\boldsymbol{\beta}}^{\boldsymbol{0}}\right)}\right)\right)$$

- Finally, the *p*-value based on the permutation of regressor residuals is

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

Model regression specification

- Following regression models were considered

Table 1 Poisson family of models

Zero-inflation	Overdispersion	
	No	Yes
No	Poisson	Negative binomial
Yes	ZI Poisson	ZI negative binomial

Table 2 Binomial family of models

Zero-inflation	Overdispersion	
	No	Yes
No	Binomial	Beta binomial
Yes	ZI binomial	ZI beta binomial

Model implementation (llperm)

- To implement the model, the authors proposed a R package called ‘llperm’.
- This package extends the ‘glmperm’ R package implemented by Werft et al (2010), which in turns is an extension of ‘logregperm’ R package proposed by Potter (2005).
- The package ‘logregperm’ implemented the novel permutation test procedure for inference in logistic regression models.
- On the other hand, the ‘glmperm’ extended that into Generalized Linear Models (GLM) where more than one variables can be involved together with the covariates of interest.

Model implementation (llperm)

- The package ‘llperm’ extended the following considerations to better fit microbiome data.

1. The covariate of interest can occur as a category with multiple levels.
2. The implementation is generalized to any likelihood based model, which enables additional distributions with zero-inflation and overdispersion.
3. In case of zero-inflated models, the regression coefficients related to the count- and the zero-component can be simultaneously tested.

Simulations

Real data underlying the simulation

- To generate simulated data, the real dataset was used as the foundation where the ‘signal’ (truly differentially abundant taxa) is known.
- The VEGA data set (Dierikx et al, 2017) studied the extent to which antibiotic resistant bacteria occur in vegetarians and non-vegetarians.
- 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.

Real data underlying the simulation

- The data set has 149 persons and 531 ASVs. The microbiome data can be represented by a 149 × 531 table of counts.
- For example, the counts for ‘ASV305’ in Fig. 1 could indicate some difference in diet groups.

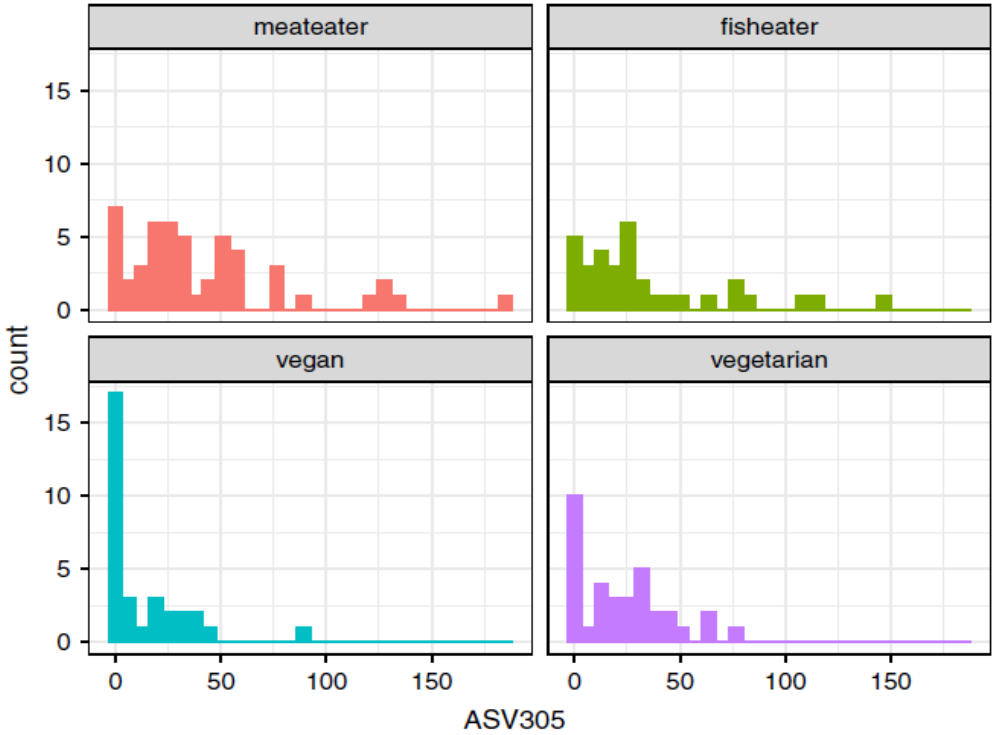

Fig. 1 Example of raw data: 16S rRNA sequence counts for a single taxa ‘ASV305’, which appear overdispersed (mean 30, variance 1133) and zero-inflated, possibly influenced by Diet

Simulations

Simulated data

❑ Adding signal to the real data

- For each simulated dataset, each person in real data is assigned to one of 4 groups (meateater, fisheater, vegetarian, vegan) with equal 25% probability, irrespective of the real status.
- 10% of the taxa are randomly chosen to be differentially abundant in each group.
- If a taxa is differentially abundant in a person, the counts are multiplied by an effect size (+25%, +50%, +100%, +200%, +400%).
- Signal in the zero counts were additionally introduced by decreasing their probability.
- For every taxon, the baseline odds was calculated first of the counts being non-zero, and assigned this to every individual.

Simulations

Simulated data

❑ Adding signal to the real data

- 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, this probability was used to draw whether or not the particular sample was non-zero, and if so, a non-zero count was sampled without replacement from the existing data.
- At some point the number of non-zero counts available for sampling are depleted, and the remaining samples are assigned zero's.

Simulations

Simulated data

❑ Adding signal to the real data

- 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.

Simulations

Simulated data

❑ Adding covariates

- A similar simulated data set was formed that contains confounding factors.
- For each subject, two additional covariates Urbanization (low/high) and Age (20–69) were included.
- 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%.

Simulations

Simulated data

- There are three sources of signal:

  different 10% of taxa are differentially abundant for

  - each diet group
  - urbanization
  - affected by age

Simulated data

- Table 3 shows the probability of being assigned to a joint Diet and Urbanization group

Table 3 Joint probability of diet and urbanization

Urbanization	Diet			
	Meateater (%)	Fisheater (%)	Vegetarian (%)	Vegan (%)
Low	20	15	10	5
High	5	10	15	20

Table 4 Conditional probability of age given diet

Age	Diet			
	Meateater (%)	Fisheater (%)	Vegetarian (%)	Vegan (%)
[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

- Table 4 shows the conditional probability of being assigned into a particular age range given a diet group.

Simulations

Results

- At first, the 4 diet groups (meateater, fisheater, vegetarian, vegan) in the simulations are compared without confounding factors.
- Then Urbanization (low/high) and Age (20–60) are added as confounding factors.
- To both data sets the signal is introduced only in the counts, or in both counts & zeros.

Simulations

Results

- The LRT and the PRR-test are compared by presenting

    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
    4. Area Under the ROC curve up to the FPR = 0.10

- These are illustrated by the ROC curve in Fig. 2.

Simulations

Results

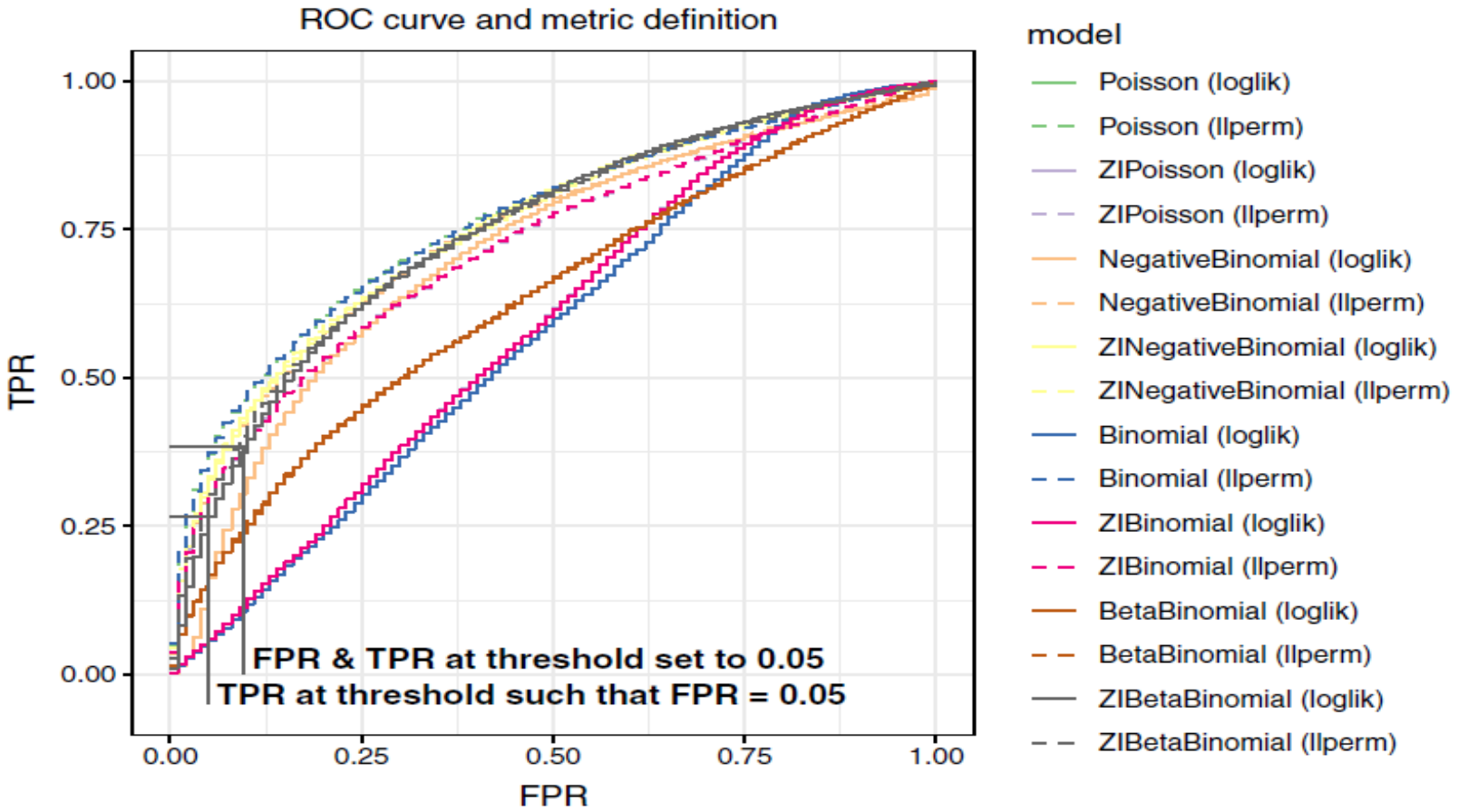

Fig. 2 ROC curve on diet groups with confounding variables, signal in both counts & zeros. ZIBetaBinomial (loglik) has a TPR 0.38 and 0.10 FPR at a 0.05 *p* value threshold. If we set the threshold such that FPR equals 0.05, the TPR@0.05 is 0.27. We can similarly calculate the AUC@0.10 as area under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

Simulations

Results

❑ Group comparison without confounding

- The LRT and PRR test based model results are shown in Table 5 for signal in counts and Table 6 for signal in counts & zeros.

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)

Simulations

Results

❑ Group comparison without confounding

- Most likelihood based models without overdispersion have high false positive rate.
- More than 90% of non-differentially abundant taxa are detected as false positives for (ZI)Binomial and (ZI) Poisson models.
- Only the ZIBetaBinomial model 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.

Results

❑ Group comparison with confounding

- Table 7 shows the results with signal in counts and Table 8 those for signal in counts & zeros both.

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
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
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)

Simulations

Results

❑ Group comparison with confounding

- Only the Negative Binomial model among the likelihood based methods had a correct nominal 5% FPR rate with a 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).

Comparing alternative approaches

- The proposed methods was compared to alternative approaches.
- The counts in VEGA dataset were then modified with an otherwise identical simulation.
- The permutation based Poisson family count regression models (Poisson, ZIPoisson, NegativeBinomial, ZINegativeBinomial) were compared to alternative approaches.
- There are several other widely applicable methods that have also indicated correct false positive rate control: ALDEx2, ANCOM-BC, LinDA, and Maaslin2.

Comparing alternative approaches

- The alternative approaches indicate lower than expected false positive rates and have a considerably lower power.
- DESeq2 and EdgeR deliver very similar results to likelihood based negative binomial count regression.

Table 10 Model comparison and alternative approaches (signal in counts, effect size +100%))

Family	Type	Power	FPR	Power@0.05	AUC@0.10
DESeq2	(baseline)	0.11 (±0.00)	0.02 (±0.00)	0.17 (±0.01)	0.62 (±0.01)
EdgeR	(baseline)	0.09 (±0.00)	0.01 (±0.00)	0.19 (±0.01)	0.66 (±0.01)
ALDEx2	(baseline)	0.04 (±0.00)	0.01 (±0.00)	0.09 (±0.00)	0.59 (±0.01)
ANCOMBC	(baseline)	0.11 (±0.00)	0.06 (±0.00)	0.10 (±0.00)	0.59 (±0.01)
LinDA	(baseline)	0.10 (±0.00)	0.05 (±0.00)	0.10 (±0.00)	0.59 (±0.01)
Maaslin2 (lm)	(baseline)	0.10 (±0.00)	0.05 (±0.00)	0.10 (±0.00)	0.60 (±0.01)
kruskal_test (strata)	(baseline)	0.12 (±0.01)	0.05 (±0.00)	0.12 (±0.00)	0.62 (±0.01)
oneway_test (strata)	(baseline)	0.18 (±0.01)	0.05 (±0.00)	0.18 (±0.01)	0.65 (±0.01)
Poisson	(llperm)	0.22 (±0.01)	0.05 (±0.00)	0.21 (±0.01)	0.66 (±0.01)
Poisson	(loglik)	0.92 (±0.00)	0.85 (±0.00)	0.14 (±0.01)	0.54 (±0.01)
ZIPoisson	(llperm)	0.29 (±0.01)	0.05 (±0.00)	0.29 (±0.01)	0.70 (±0.01)
ZIPoisson	(loglik)	0.92 (±0.00)	0.78 (±0.00)	0.16 (±0.01)	0.57 (±0.01)
NegativeBinomial	(llperm)	0.23 (±0.01)	0.06 (±0.00)	0.22 (±0.01)	0.69 (±0.01)
NegativeBinomial	(loglik)	0.13 (±0.00)	0.03 (±0.00)	0.18 (±0.01)	0.60 (±0.01)
ZINegativeBinomial	(llperm)	0.29 (±0.01)	0.05 (±0.00)	0.29 (±0.01)	0.70 (±0.01)
ZINegativeBinomial	(loglik)	0.39 (±0.01)	0.10 (±0.00)	0.27 (±0.01)	0.68 (±0.01)

Discussion

- Results from simulation indicates that the PRR-test effectively controls the FPR. Additionally, it shows improved statistical power.
- While models accommodating overdispersion and zero-inflation tend to outperform in likelihood-based approaches, the discrepancies become less evident when employing the PRR-test.
- The results are mostly consistent with previous research, but there are two remarkable exceptions.

    1. In the comparison of groups without confounding, the negative binomial model exhibits an unusually low FPR of 0.02.
    2. 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.

Discussion

- The ‘MASS’ package converged to a different solution compared to this proposed ‘llperm’ in some of the datasets.
- With the exception of this issue with ‘MASS’, the results from ‘llperm’ tended to be identical to those delivered by other packages.
- Compared to simulating data from a known statistical distribution, simulating data by resampling a real data set yields more accurate findings.
- The simulations are based on a single dataset, this might not fully reflect all possible data in microbiome studies.

Conclusion

- The PRR-test is able to maintain the actual type 1 error rate. It provides the equal or greater power than the likelihood based approach at a given significance level.
- Likelihood based models can have a high rate of type 1 error, and it is not possible to adjust for this in real data sets where the actual truth is unknown.
- The proposed method therefore provides a new approach which is competitive in power.

Thank you!