PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# A marginalized two-part Beta regression model for microbiome compositional data

**Haitao Chai**[1,2], **Hongmei Jiang**[3], **Lu Lin**[1], **Lei Liu**[2,4]*

**1** Institute for Financial Studies, Shandong University, Jinan, Shandong, China, **2** Department of Preventive Medicine, Northwestern University, Chicago, Illinois, United States of America, **3** Department of Statistics, Northwestern University, Evanston, Illinois, United States of America, **4** Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri, United States of America

* lei.liu@wustl.edu

## Abstract

In microbiome studies, an important goal is to detect differential abundance of microbes across clinical conditions and treatment options. However, the microbiome compositional data (quantified by relative abundance) are highly skewed, bounded in [0, 1), and often have many zeros. A two-part model is commonly used to separate zeros and positive values explicitly by two submodels: a logistic model for the probability of a specie being present in Part I, and a Beta regression model for the relative abundance conditional on the presence of the specie in Part II. However, the regression coefficients in Part II cannot provide a marginal (unconditional) interpretation of covariate effects on the microbial abundance, which is of great interest in many applications. In this paper, we propose a marginalized two-part Beta regression model which captures the zero-inflation and skewness of microbiome data and also allows investigators to examine covariate effects on the marginal (unconditional) mean. We demonstrate its practical performance using simulation studies and apply the model to a real metagenomic dataset on mouse skin microbiota. We find that under the proposed marginalized model, without loss in power, the likelihood ratio test performs better in controlling the type I error than those under conventional methods.

## Author summary

Semi-continuous compositional data are typically analyzed using two-part models which separately describe the probability of zero values and the distribution of positive values. The second part of the model provides a conditional interpretation of covariate effects on the positive response. However, it is of great interest in many applications to assess the covariate effect on the marginal mean of the response. For this purpose, we propose a marginalized two-part model by reparameterizing the marginal mean in Part II. We show that the proposed marginalized two-part model outperforms conventional methods by simulation studies in terms of controlling the Type I error and maximizing the power. We apply our method to a microbiota dataset, and find consistent results with our simulation studies.

**Citation:** Chai H, Jiang H, Lin L, Liu L (2018) A marginalized two-part Beta regression model for microbiome compositional data. PLoS Comput Biol 14(7): e1006329. https://doi.org/10.1371/journal.pcbi.1006329

**Editor:** Dan Knights, University of Minnesota, UNITED STATES

**Received:** July 13, 2017

**Accepted:** June 26, 2018

**Published:** July 23, 2018

**Data Availability Statement:** The data are publicly available at https://www.nature.com/articles/ncomms3462#supplementary-information.

**Funding:** This research was partly supported by AHRQ R01 HS 020263 and National Natural Science Foundation of China (Grant No. 11571204). HC was supported by the China Scholarship Council (201606220131) as a visiting student at Northwestern University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

In recent years, metagenomics studies have been growing rapidly due to the advances of next-generation sequencing (NGS) technologies [1]. Microbiota have been known to be associated with various diseases, e.g., obesity and diabetes [2, 3], Crohn's disease [4], bacterial vaginosis [5], and cancer [6, 7].

The microbial abundance is usually measured in read counts. However, such quantities are not directly comparable across samples due to the uneven total sequence counts of samples. Therefore, the read counts are often normalized to relative abundances which sum to 1 for all microbes in a sample [8]. Relative abundance can be characterized by a point mass at zero and a right-skewed continuous distribution with a positive support, the so-called "semi-continuous" or "zero-inflated continuous" data. The zero values indicate that certain microbes are absent in the sample, or the rare microbes are present but missed due to undersampling, while the continuous distribution with a positive support describes the levels of relative abundance among the present microbes.

The relative abundance is often described by a two-part model [9], which separates zeros and positive values explicitly by two submodels: a logistic model for the probability of the outcome being positive in Part I and a (generalized) linear regression model for the amount of the (transformed) positive value in Part II. An important issue in such two-part models is to determine the distributional form in Part II. The nonzero relative abundance data are non-normally distributed and bounded in [0, 1). Beta distribution has been used to model this outcome. A two-part Beta regression model can be thus developed [10–12]. It includes two sets of parameters, one in the logistic regression for the presence of a microbe, and the other in the Beta regression for the relative abundance conditional on the presence of the microbe. These two sets of parameters are interpreted as effects on the presence of a microbe and on the level of relative abundance given that the microbe is present, respectively. That is, there is a conditional interpretation in Part II. However, it is often of great interest to have a straightforward interpretation of covariate effects on the overall marginal (unconditional) mean. For example, [13] proposed a marginalized two-part log-normal model by parameterizing covariates effects directly in terms of the marginal mean.

As conventional two-part Beta regression models do not provide an unconditional interpretation of covariate effects, we propose a marginalized two-part Beta regression model for microbiome abundance data which parameterizes covariate effects in terms of the marginal mean. The proposed model not only accounts for the zero-inflated nature of the microbiome data but also yields more interpretable effect estimates.

Of note, an alternative to describe zero-inflated data is the Tobit model [14] where zero values are considered as left censored observations of the underlying true negative values (of Normal or other distributions accommodating negative values). However, the Tobit model is not appropriate for the Beta distribution which does not have a support of negative values. Consequently, the Tobit model cannot be applied directly to the relative abundance data.

## Models

In the following Section, we will introduce the conventional two-part Beta regression model and the proposed marginalized two-part Beta regression model. We will also describe their properties to assess the overall impact of covariates on the marginal mean, and demonstrate that the proposed model outperforms the conventional model.

## Two-part Beta regression model

We begin with the conventional two-part model with a Beta component in Part II [10–12]. For a given operational taxonomic unit (OTU), let $Y_i$ denote its semi-continuous relative abundance for subject $i$, where $0 \leq Y_i < 1$ and $i = 1, 2, \ldots, n$. Specifically, a two-part Beta regression model has the following form:

$$\begin{aligned} Y_i &\sim 0 \quad \text{with probability} \quad 1 - p_i \\ &\sim \text{Beta}(\mu_i\phi, (1-\mu_i)\phi) \quad \text{with probability} \quad p_i, \end{aligned}$$

where the density function of the Beta distribution is parameterized as

$$\frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma[(1-\mu_i)\phi]} y_i^{\mu_i\phi-1}(1-y_i)^{(1-\mu_i)\phi-1},$$

with $\mu_i$ $(0 < \mu_i < 1)$ and $\phi$ $(\phi > 0)$ being the mean and dispersion parameters of the Beta distribution, respectively, and $p_i$ is the probability that the observation $Y_i$ is from the Beta distribution. The two-part model describes the probability $p_i$ in the logistic component and the conditional mean in the Beta component as functions of covariates,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\alpha}, \tag{1}$$

$$\text{logit}(\mu_i) = \text{logit}[E(Y_i|Y_i > 0)] = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of regression coefficients, $\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ip})^T$ is the $(p + 1)$ dimensional covariate vector (including an intercept) for the $i$-th subject. We assume identical covariates for both parts of the model for simplicity of notation. One can instead allow for different sets of covariates for the two parts.

## Marginalized two-part Beta regression model

To obtain interpretable covariate effects on the marginal mean, we propose the following marginalized two-part Beta regression model. Let $\nu_i = E(Y_i)$ be the marginal mean of $Y_i$. The first part of the proposed marginalized two-part model is the same as Part I in the conventional two-part model,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\alpha}. \tag{3}$$

In Part II, the marginal (unconditional) mean $\nu_i$, instead of the conditional mean $\mu_i$, is modeled as a function of covariates:

$$\text{logit}(\nu_i) = \log\left(\frac{\nu_i}{1-\nu_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\gamma}. \tag{4}$$

As we can see, the marginalized two-part model not only captures zero-inflation and skewness as the conventional two-part model, but also allows us to examine covariate effects on the overall marginal mean.

In the S1 Text, we can see that the likelihood of the conventional two-part model can be reparameterized to that of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\phi$ in the marginalized model. However, the interpretation of covariate effects are different in the two frameworks, which will be elaborated in the next subsection.

The estimation of the marginalized two-part model can be carried out in SAS Proc NLMIXED (The main code is shown in S1 Code). To obtain starting values of the estimation, a logistic model and a Beta regression model are fitted for the binary part and the positive part, respectively. Then the estimates of these two models are used as starting values for the two-part marginalized model. The convergence of the estimation is determined by a threshold value $1 \times 10^{-8}$ for the relative gradient, a common convergence criterion in SAS Proc NLMIXED. This criterion is satisfied in our simulations for all replicates, and in the real data analysis for all 131 OTUs.

## Interpretation of covariate effects

**For the conventional model.** Using the conventional two-part model shown in Eqs (1) and (2), $\beta_j$ is interpreted as the effect of a unit increase in the $j$th covariate on the logit of the conditional mean of $Y_i$ given $Y_i$ is positive. In many applications, however, the primary interest is to examine the impact of covariates on the overall marginal mean $E(Y_i)$. For the conventional two-part model, we have

$$E(Y_i) = p_i E(Y_i | Y_i > 0) = \frac{\exp(\boldsymbol{X}_i^T \boldsymbol{\alpha})}{1 + \exp(\boldsymbol{X}_i^T \boldsymbol{\alpha})} \cdot \frac{\exp(\boldsymbol{X}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}_i^T \boldsymbol{\beta})}. \tag{5}$$

Along the lines of [15], we can assess the effect of the $j$-th continuous covariate $x_{ij}$ on the unconditional mean as

$$\frac{\partial}{\partial x_{ij}} (\text{logit}[E(Y_i)]) = \frac{\partial}{\partial x_{ij}} (\text{logit}[p(x_{ij})\mu(x_{ij})]), \tag{6}$$

where

$$\begin{aligned} p(x_{ij}) &= \frac{\exp[x_{ij}\alpha_j + \boldsymbol{X}_{i(-j)}^T \boldsymbol{\alpha}_{(-j)}]}{1 + \exp[x_{ij}\alpha_j + \boldsymbol{X}_{i(-j)}^T \boldsymbol{\alpha}_{(-j)}]}, \\ \mu(x_{ij}) &= \frac{\exp[x_{ij}\beta_j + \boldsymbol{X}_{i(-j)}^T \boldsymbol{\beta}_{(-j)}]}{1 + \exp[x_{ij}\beta_j + \boldsymbol{X}_{i(-j)}^T \boldsymbol{\beta}_{(-j)}]}, \end{aligned}$$

with $\alpha_j$ and $\beta_j$ being the coefficients corresponding to $x_{ij}$ in the conventional two-part model and $\boldsymbol{X}_{i(-j)}$, $\boldsymbol{\alpha}_{(-j)}$, and $\boldsymbol{\beta}_{(-j)}$ be the corresponding vectors with the $j$-th covariate removed.

A straightforward calculation shows that (6) can be equivalently written as

$$\frac{\partial}{\partial x_{ij}} (\text{logit}[E(Y_i)]) = c_1(\alpha_j, \beta_j)\alpha_j + c_2(\alpha_j, \beta_j)\beta_j, \tag{7}$$

where

$$\begin{aligned} c_1(\alpha_j, \beta_j) &= \frac{1 - p(x_{ij})}{1 - p(x_{ij})\mu(x_{ij})}, \\ c_2(\alpha_j, \beta_j) &= \frac{1 - \mu(x_{ij})}{1 - p(x_{ij})\mu(x_{ij})}. \end{aligned}$$

As the logit transformation is a monotonically increasing function in the interval (0, 1), the hypothesis test of the covariate effects on the marginal mean is equivalent to that on its logit transformation. In Eq (7), the logit transformation of the marginal mean abundance is independent of covariate $x_{ij}$ if both $\alpha_j$ and $\beta_j$ are zero. However, if $\alpha_j$ and $\beta_j$ have opposite signs, even when they are not zero, the logit transformation of the marginal mean abundance may be still independent of covariate $x_{ij}$. Furthermore, the coefficients $c_1(\alpha_j, \beta_j)$ and $c_2(\alpha_j, \beta_j)$ in Eq (7) are functions of $\alpha_j$ and $\beta_j$. Thus, the independence between the marginal mean and covariate $x_{ij}$ cannot be tested simply as the hypothesis of $\alpha_j = 0$ and $\beta_j = 0$, e.g., by the likelihood ratio test. Instead, the Delta method has to be used on the hypothesis test of Eq (7), which depends on $\boldsymbol{X}_{i(-j)}$ in a complicated way.

When the interest is to assess the effect of a discrete variable on response, e.g., placebo vs. treatment, Eq (7) no longer applies. Without loss of generality, consider a binary covariate $x_{ik}$ taking value 0 or 1. Similar to [15], the difference in the logit transformation of the marginal mean with $x_{ik} = 1$ vs. $x_{ik} = 0$ is used to evaluate the impact on the expected marginal mean response.

Under the conventional two-part model, the difference between the logit transformations with $x_{ik} = 1$ and $x_{ik} = 0$ is

$$\begin{aligned} & \operatorname{logit}[E(Y_i|x_{ik}=1)] - \operatorname{logit}[E(Y_i|x_{ik}=0)] \\ = \; & \operatorname{logit}[p(x_{ik}=1)\mu(x_{ik}=1)] - \operatorname{logit}[p(x_{ik}=0)\mu(x_{ik}=0)] \\ = \; & \alpha_k + \beta_k + b_1(\alpha_k) + b_2(\beta_k) + b_3(\alpha_k, \beta_k), \end{aligned} \tag{8}$$

where

$$\begin{aligned} b_1(\alpha_k) &= \ln\left(\frac{1+\exp[\boldsymbol{X}_{i(-k)}^T\boldsymbol{\alpha}_{(-k)}]}{1+\exp[\alpha_k+\boldsymbol{X}_{i(-k)}^T\boldsymbol{\alpha}_{(-k)}]}\right), \\ b_2(\beta_k) &= \ln\left(\frac{1+\exp[\boldsymbol{X}_{i(-k)}^T\boldsymbol{\beta}_{(-k)}]}{1+\exp[\beta_k+\boldsymbol{X}_{i(-k)}^T\boldsymbol{\beta}_{(-k)}]}\right), \\ b_3(\alpha_k,\beta_k) &= \ln\left(\frac{1-p(x_{ik}=0)\mu(x_{ik}=0)}{1-p(x_{ik}=1)\mu(x_{ik}=1)}\right). \end{aligned}$$

It is worth noting that $b_1(\alpha_k)$, $b_2(\beta_k)$, and $b_3(\alpha_k, \beta_k)$ all equal to 0 if $\alpha_k$ and $\beta_k$ are 0. Similar to the continuous covariate, the logit transformation of the marginal mean abundance does not depend on the binary covariate $x_{ik}$ if both $\alpha_k$ and $\beta_k$ are zero. However, even though neither of the coefficients is zero, the transformed mean abundance may still be independent of the binary covariate $x_{ik}$ when $\alpha_k$ and $\beta_k$ have opposite signs. Eq (8) indicates that the independence between the response and the binary covariate $x_{ik}$ cannot be ascertained by directly testing $\alpha_k = 0$ and $\beta_k = 0$ by e.g., the likelihood ratio test, as shown in the simulation studies and the real data analysis.

**For the marginalized model.** In the marginalized two-part model Eqs (3) and (4), the effect of a continuous covariate $x_{ij}$ on the marginal mean $E(Y_i)$ can be characterized by

$$\frac{\partial}{\partial x_{ij}}(\operatorname{logit}[E(Y_i)]) = \gamma_j, \tag{9}$$

where $\gamma_j$ is the coefficient corresponding to $x_{ij}$ in Eq (4). Thus, the effect of the covariate $x_{ij}$ on the marginal mean abundance is determined by its coefficient in the marginalized model. With the marginalized two-part model, we can estimate the coefficient $\gamma_j$ as well as test the effect on the marginal mean.

As for a binary covariate in the marginalized two-part model, the difference in logit transformation of the marginal mean with $x_{ik} = 1$ vs. $x_{ik} = 0$ can be expressed as

$$\operatorname{logit}[E(Y_i|x_{ik}=1)] - \operatorname{logit}[E(Y_i|x_{ik}=0)] = \gamma_k \tag{10}$$

One can see that the effect of a binary covariate $x_{ik}$ on the marginal mean abundance is determined by its coefficient $\gamma_k$ in the marginalized two-part model. The logit transformation of the marginal mean abundance with $x_{ik} = 1$ is bigger than that with $x_{ik} = 0$ when $\gamma_k$ is positive, and the reverse is true when $\gamma_k$ is negative.

## Results

In this section, simulation studies and real data analysis are presented to assess the performance of the proposed marginalized and the conventional two-part models. Results show that the proposed model outperforms the conventional model, which is consistent with the theoretical results.

### Simulation studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed marginalized two-part model. To test the effect of the covariate on the overall marginal mean $E(Y_i)$, likelihood ratio tests (LRT) are performed and compared under the marginalized two-part (MTP) model and the conventional two-part (CTP) model. In addition, the two sample T-test and the Wilcoxon rank sum test are also compared.

We assume that, in both parts, there is only one binary covariate $x_1$, which is generated from the Bernoulli distribution with $p = 0.5$. However, according to the interpretation of the covariate effects in the preceding section, the proposed model can be applied to multiple covariates. The response $y_i$ is generated below:

$$\begin{aligned} \text{logit}(p_i) &= \alpha_0 + \alpha_1 x_{i1}, \\ \text{logit}(v_i) &= \gamma_0 + \gamma_1 x_{i1}, \\ f(y_i) &= (1 - p_i)^{1_{(y_i=0)}} \times [p_i \text{Beta}(\mu_i \phi, (1 - \mu_i)\phi)]^{1_{(y_i>0)}}, \end{aligned}$$

where $\mu_i = \frac{1+\exp(-\boldsymbol{x}_i^T \boldsymbol{\alpha})}{1+\exp(-\boldsymbol{x}_i^T \boldsymbol{\gamma})}$ is the conditional mean given that $y_i$ is positive and $\phi$ is the dispersion parameter of the Beta distribution.

In the simulation studies, 1000 samples of sizes 200 and 400 are generated. We set the parameters as $\alpha_0 = 1.5$, $\gamma_0 = -2.5$, and $\phi = 1$, while $\alpha_1$ and $\gamma_1$ may have different values according to which of the two criteria are under study: the type I error or the power.

First, we evaluate the type I error for testing the null hypothesis $H_0$: *the binary covariate $x_1$ has no effect on the overall marginal mean of $y_i$*. In the MTP model, this is equivalent to testing $H_0^M : \gamma_1 = 0$ as shown in Eq (10). However, testing $H_0^C : \alpha_1 = \beta_1 = 0$ in the CTP model is not equivalent to testing $H_0$. Specifically, according to Eq (8), even though neither of the coefficients is zero, the binary covariate $x_1$ may still have no effect on the marginal mean. This means that the conventional model cannot control the type I error for testing $H_0$ when both $\alpha_1$ and $\beta_1$ are non-zero.

The results are shown in Fig 1. Type I errors are calculated under two settings: $\alpha_1 = 0$, $\gamma_1 = 0$ and $\alpha_1 = 1$, $\gamma_1 = 0$. For each setting, two $\alpha$-levels are considered: 0.01 and 0.05. As we can see from Fig 1, under the first setting ($\alpha_1 = 0$, $\gamma_1 = 0$), all the methods control the type I error reasonably well. Under the setting $\alpha_1 = 1$, $\gamma_1 = 0$, the LRT under the MTP and the T-test control the type I error well, while the LRT under the CTP and the Wilcoxon test cannot control the type I error, especially the LRT under the CTP model. Because in this setting, testing $H_0^C$ in the CTP model is not equivalent to testing the null hypothesis $H_0$.

The powers under two different settings, $\alpha_1 = 0$, $\gamma_1 = 1$ and $\alpha_1 = 1$, $\gamma_1 = 1$, are shown in Fig 2. As we can see, the LRT under the CTP and the MTP are the most powerful methods with

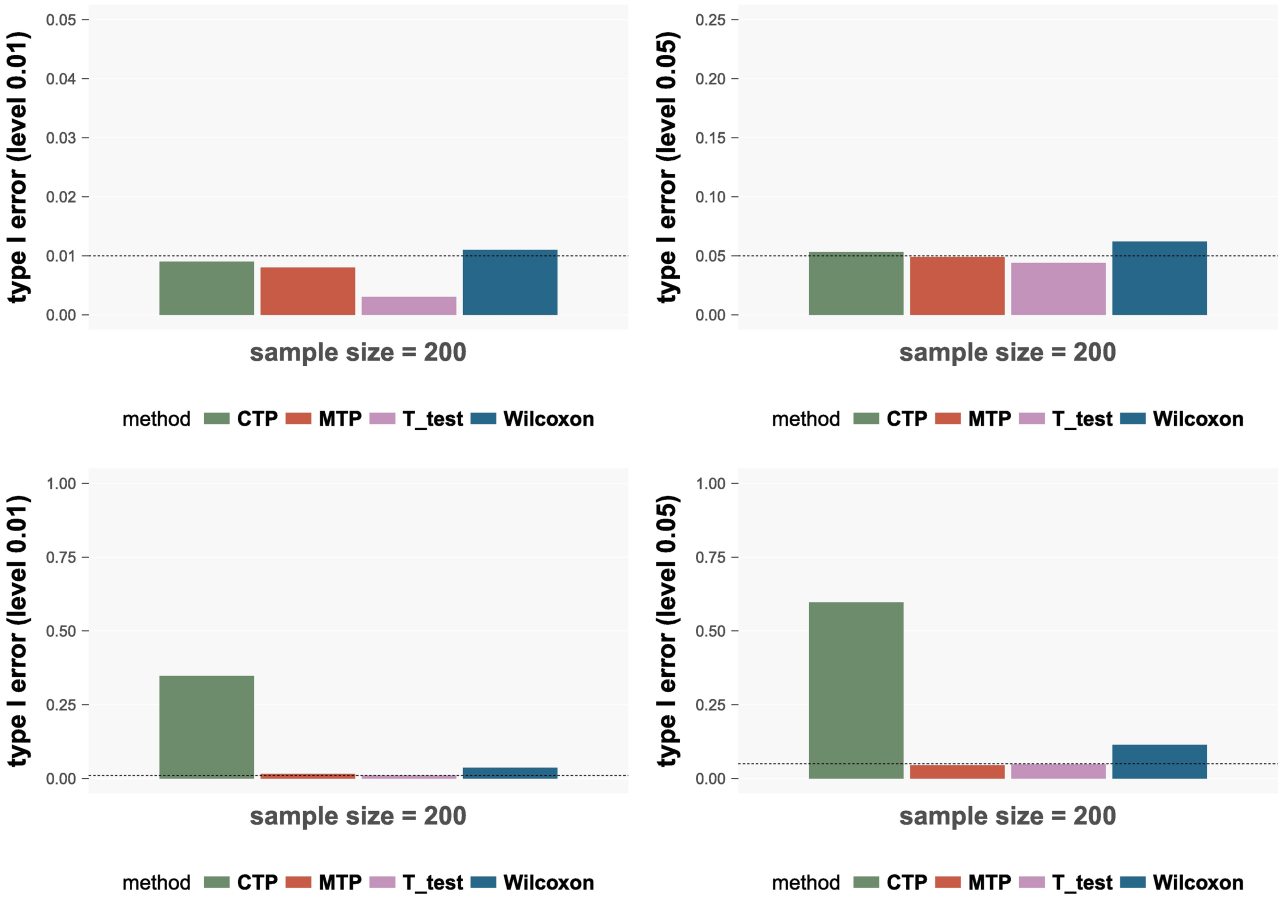

**Fig 1. Type I errors of the four methods.** The results in the upper panels correspond to the setting $\alpha_1 = 0, \gamma_1 = 0$ and the lower panels correspond to setting $\alpha_1 = 1, \gamma_1 = 0$. In each setting, the left panel shows the results for significance level 0.01 and the right panel shows the results for level 0.05. The dashed horizontal line in each panel represents the correct level. The results for sample size 400 can be found in S1 Fig in Supporting information.

https://doi.org/10.1371/journal.pcbi.1006329.g001

the power close to 1 in all settings. The Wilcoxon test performs a little worse than the LRT while the T-test has the lowest power.

We also estimate the coefficients in the MTP model under the setting $\alpha_1 = 1, \gamma_1 = 1$. The results in Table 1 demonstrate that the biases are negligible and the coverage probabilities are acceptably close to the nominal level 0.95 for all the model parameters. In addition, we observe small differences between the empirical standard errors and our estimates. The mean squared errors for sample size 400 are smaller than those for sample size 200.

According to the simulation results, the LRT under the MTP model has the best performance: it controls the type I error reasonably well and also achieves the best power. The T-test has the similar performance in the error control while it is not as powerful as the LRT under the MTP model. The LRT under the CTP model is powerful, however, it fails to control the type I error. The Wilcoxon test has poor performances in both the error control and power than the LRT under the MTP model.

To assess the robustness of the proposed method, we consider a setting where positive responses are generated from another distribution. First of all, the only covariate $x_i$ is generated from the Uniform distribution on (0, 1), while the response $y_i$ has the following

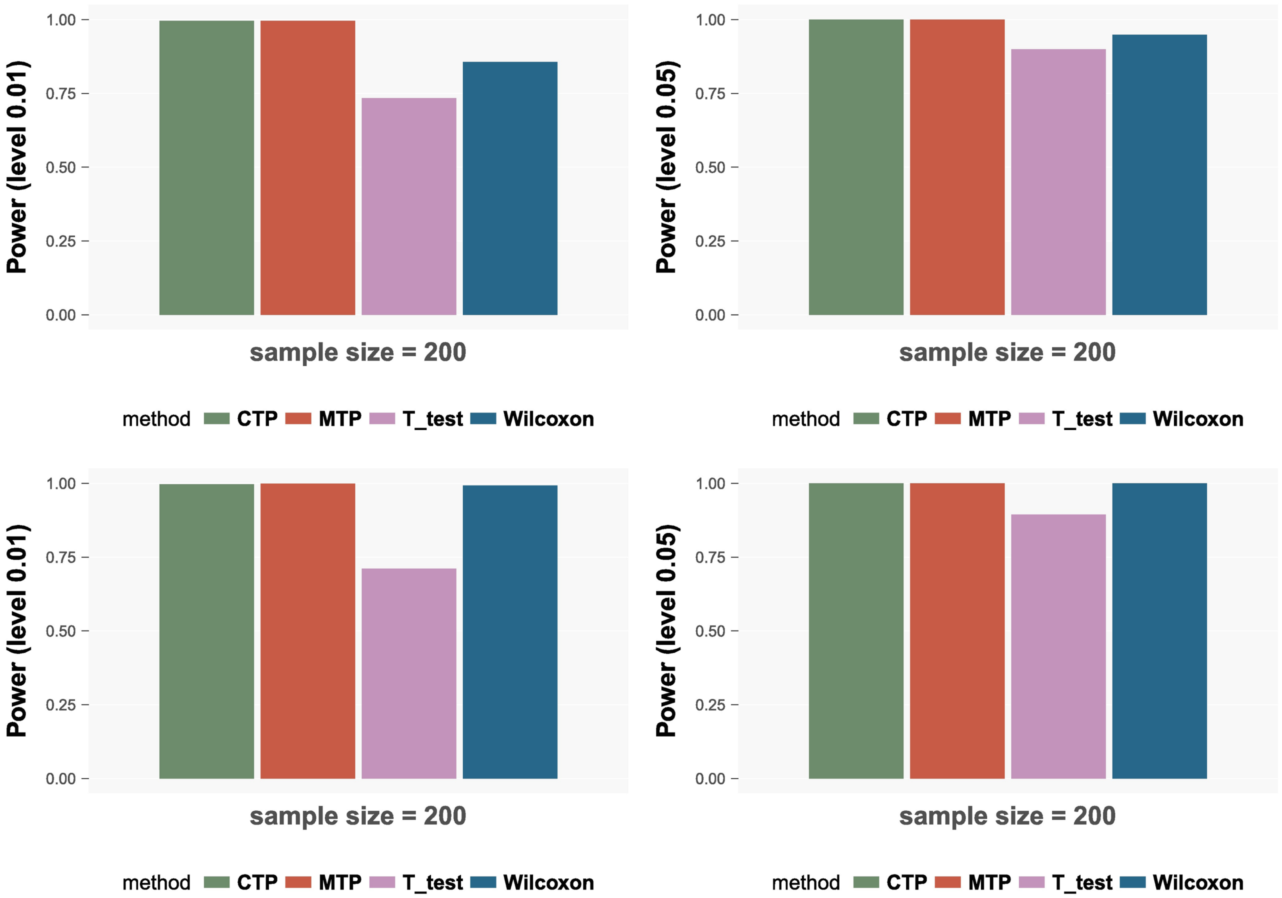

**Fig 2. Powers of the four methods.** The upper panels show the powers corresponding to the setting $\alpha_1 = 0$, $\gamma_1 = 1$ and the lower panels show the powers corresponding to the setting $\alpha_1 = 1$, $\gamma_1 = 1$. In each setting, the left panel shows the results for significance level 0.01 and the right panel shows the results for level 0.05. The powers for sample size 400 are shown in S2 Fig in Supporting information.

https://doi.org/10.1371/journal.pcbi.1006329.g002

**Table 1. Estimates of the coefficients in the marginalized two-part model under the setting $\alpha_1 = 1$, $\gamma_1 = 1$.**

| Parameter | sample size = 200 | | | | | sample size = 400 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est | SE | SEM | CP | MSE | Est | SE | SEM | CP | MSE |
| $\alpha_0 = 1.5$ | 1.5321 | 0.2695 | 0.2646 | 0.955 | 0.0736 | 1.5153 | 0.1921 | 0.1854 | 0.943 | 0.0375 |
| $\alpha_1 = 1$ | 1.0345 | 0.4876 | 0.4803 | 0.960 | 0.2387 | 1.0078 | 0.3457 | 0.3320 | 0.947 | 0.1195 |
| $\gamma_0 = -2.5$ | -2.5104 | 0.1673 | 0.1727 | 0.956 | 0.0281 | -2.5074 | 0.1210 | 0.1219 | 0.949 | 0.0147 |
| $\gamma_1 = 1$ | 0.9962 | 0.1758 | 0.1803 | 0.949 | 0.0309 | 1.0002 | 0.1253 | 0.1273 | 0.957 | 0.0157 |
| $\phi = 1$ | 1.0323 | 0.1374 | 0.1331 | 0.956 | 0.0199 | 1.0157 | 0.0929 | 0.0923 | 0.954 | 0.0089 |

Est: mean of the parameter estimates;

SE: standard error of the parameter estimates;

SEM: sample mean of the standard error estimates;

CP: coverage probability of the corresponding 95% confidence interval.

https://doi.org/10.1371/journal.pcbi.1006329.t001

distribution:

$$y_i \quad \sim \quad 0 \quad \text{with probability} \quad 1 - p_i,$$

where

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 x_i;$$

and the overall marginal mean $\nu_i$ of the response is

$$\text{logit}(\nu_i) = \gamma_0 + \gamma_1 x_i.$$

Instead of the Beta distribution, positive responses are generated from the Binomial distribution Bin(100, $\mu_i$) and then divided by 100 to make them bounded in (0, 1). As in the previous simulation, we set $\mu_i = \frac{1+\exp(-\boldsymbol{x}_i^T \boldsymbol{\alpha})}{1+\exp(-\boldsymbol{x}_i^T \boldsymbol{\gamma})}$. The probability of having exactly 0 success in 100 trials is $(1 - \mu_i)^{100}$, which is negligible with the proper choice of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Thus almost all the zero values in this zero-inflated Binomial data are structural zeros.

In this simulation study, 1000 samples of sizes 200 and 400 are generated. The parameters are set as $\alpha_0 = 2$, $\gamma_0 = -0.5$, while $\alpha_1$ and $\gamma_1$ may have different values in order to calculate the type I errors and the powers.

The type I errors are calculated under two settings: $\alpha_1 = 0$, $\gamma_1 = 0$ and $\alpha_1 = 1$, $\gamma_1 = 0$. For each setting, two $\alpha$-levels are considered: 0.01 and 0.05. As we can see from Fig 3, under both settings, the proposed marginalized model controls the type I error reasonably well. The conventional model controls the type I error under the setting $\alpha_1 = 0$, $\gamma_1 = 0$ while it fails under the setting $\alpha_1 = 1$, $\gamma_1 = 0$, similar to Fig 1.

As shown in Fig 4, both the marginalized model and the conventional model have power equal to 1 under all settings.

From the simulation studies we can conclude that the proposed marginalized two-part Beta regression model is powerful and control the type I error well. Also, it is robust against model misspecification.

## Real data analysis

In this section, the proposed marginalized two-part model and the conventional two-part model are applied to a real metagenomic dataset on mouse skin microbiota to investigate the effects of immunization on the relative abundances of 131 core OTUs [16, 17]. The data are publicly available at https://www.nature.com/articles/ncomms3462#supplementary-information. In addition to the likelihood ratio tests under CTP and MTP, the T test and the Wilcoxon rank sum test are also included for comparison. All the tests are carried out with Bonferroni's correction.

The skin dataset contains the relative abundances of the most common 131 OTUs for 261 mouse skin samples, including 78 non-immunized and 183 immunized individuals. There is a presence of a large portion of zero abundances in the skin data, ranging from 0 to 68.97% with average 33.03% and median 33.72% (see S3 and S4 Figs). The positive values are highly right skewed and the logit transformations in the MTP model and the CTP model capture the skewness (See S5 Fig).

Fig 5 shows the results for these four methods. As we can see, the LRT under the marginalized two-part model results in significant effects of immunization on 45 (namely, 31 + 14) OTUs. The LRT under the conventional two-part model has significant results for all these 45 OTUs, and 14 (namely, 8 + 4 + 2) additional OTUs. The T test identifies 31 of these 45 OTUs and another 7 (namely, 4 + 3) OTUs. Similar to the LRT under conventional two-part model,

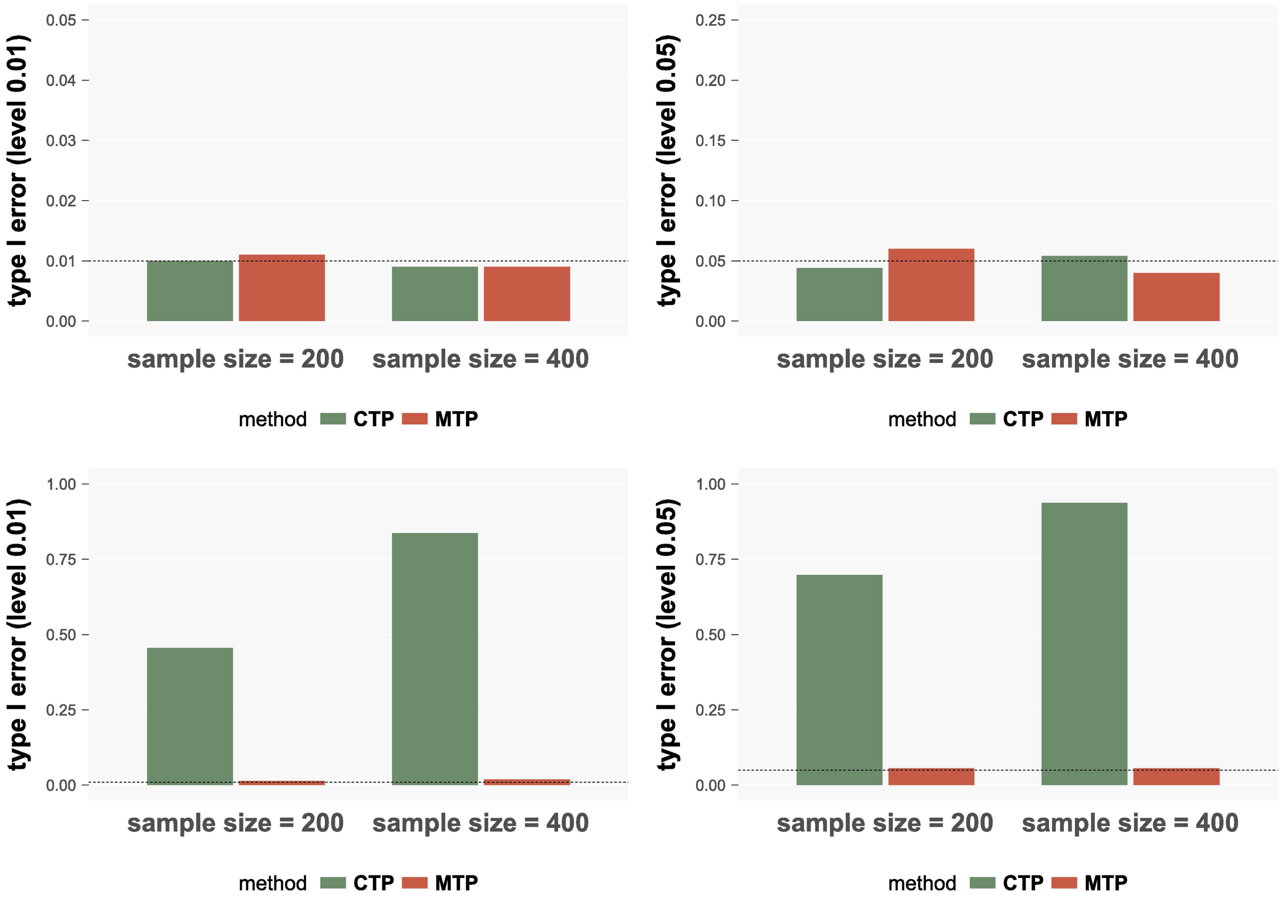

**Fig 3. Type I errors for the CTP model and the MTP model.** The results in the upper panels correspond to the setting $\alpha_1 = 0, \gamma_1 = 0$ and the lower panels correspond to the setting $\alpha_1 = 1, \gamma_1 = 0$. In each setting, the left panel shows the results for significance level 0.01 and the right panel shows the results for level 0.05. The dashed horizontal line in each panel represents the correct $\alpha$-level.

https://doi.org/10.1371/journal.pcbi.1006329.g003

the Wilcoxon test identifies all these 45 OTUs and 21 (namely, 9 + 8 + 4) additional OTUs. Finally, 60 OTUs are not identified by any methods.

The LRT under the CTP model and the Wilcoxon test identify more OTUs than the LRT under the MTP due to their failure to control the type I error as shown in Simulation studies (Fig 1). Actually, for those 14 OTUs identified by the CTP but not by the MTP, all of them have significant coefficients in Part I of the two-part model. Out of the 21 OTUs that are identified by the Wilcoxon test but not by the MTP, 17 have significant coefficients in Part I of the two-part model. This corresponds to the setting $\alpha_1 = 1, \gamma_1 = 0$ where both the CTP and the Wilcoxon test have much higher type I errors than the MTP (See the lower panel of Fig 1). Because it is less powerful than the MTP (Fig 2), the T test identifies less OTUs than the MTP.

Table 2 shows 10 most significant OTUs from the MTP model. As in [17], for OTUs which cannot be classified at the species level, the next highest classifiable taxonomic level (denoted by 'o', 'f' and 'g' for order, family, and genus, respectively) is displayed. We use a number in the superscript to distinguish among different OTUs with the same classification name. The detailed results of all the 45 OTUs identified by the proposed MTP model are shown in S1 Table.

Moreover, for most of the 131 OTUs, the proposed marginalized two-part model fits the observed data better than the conventional two-part model. Fig 6 shows the density curves of

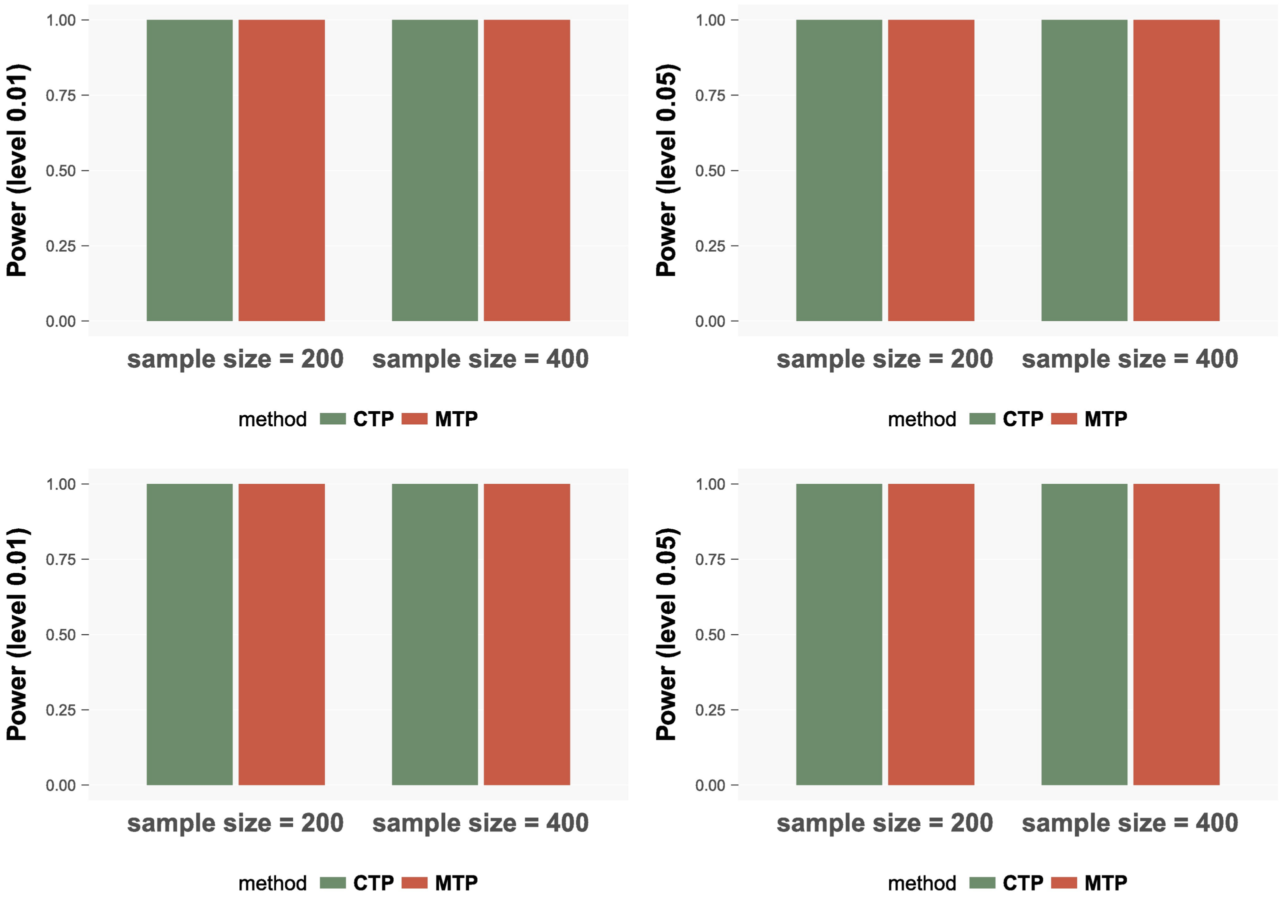

**Fig 4. Powers for the CTP model and the MTP model.** The powers in the upper panels correspond to the setting $\alpha_1 = 0, \gamma_1 = 1$ and the lower panels correspond to setting $\alpha_1 = 1, \gamma_1 = 1$. In each setting, the left panel shows the results for significance level 0.01 and the right panel shows the results for level 0.05.

https://doi.org/10.1371/journal.pcbi.1006329.g004

the observed relative abundances, the predicted relative abundances using the MTP model, and the predicted relative abundances using the CTP model for two OTUs. As we can see, the MTP model fits the observed data much better than the CTP model.

## Discussion

In this paper, we propose a marginalized two-part Beta regression model for semi-continuous microbiome compositional data. This model allows investigators to obtain covariate effects on the marginal mean of the outcome. It takes into account the compositional and zero-inflation nature of the microbiome relative abundance data. It also has an unconditional interpretation of the covariate effect on the marginal mean. Our proposed marginalized two-part model has satisfactory performance in both simulation studies and real data analysis.

For count outcomes exhibiting many zeros, a zero-inflated Poisson (ZIP) regression model or a zero-inflated negative binomial (ZINB) model, is often employed to examine the relation between covariates and the response. To model the overall population mean count directly, the marginalized ZIP model and the marginalized ZINB model were proposed by [18] and [19], respectively. However, in the case of bounded count data, the ZIP is questionable while the zero-inflated binomial (ZIB) model and its extension for over-dispersion: the zero-inflated

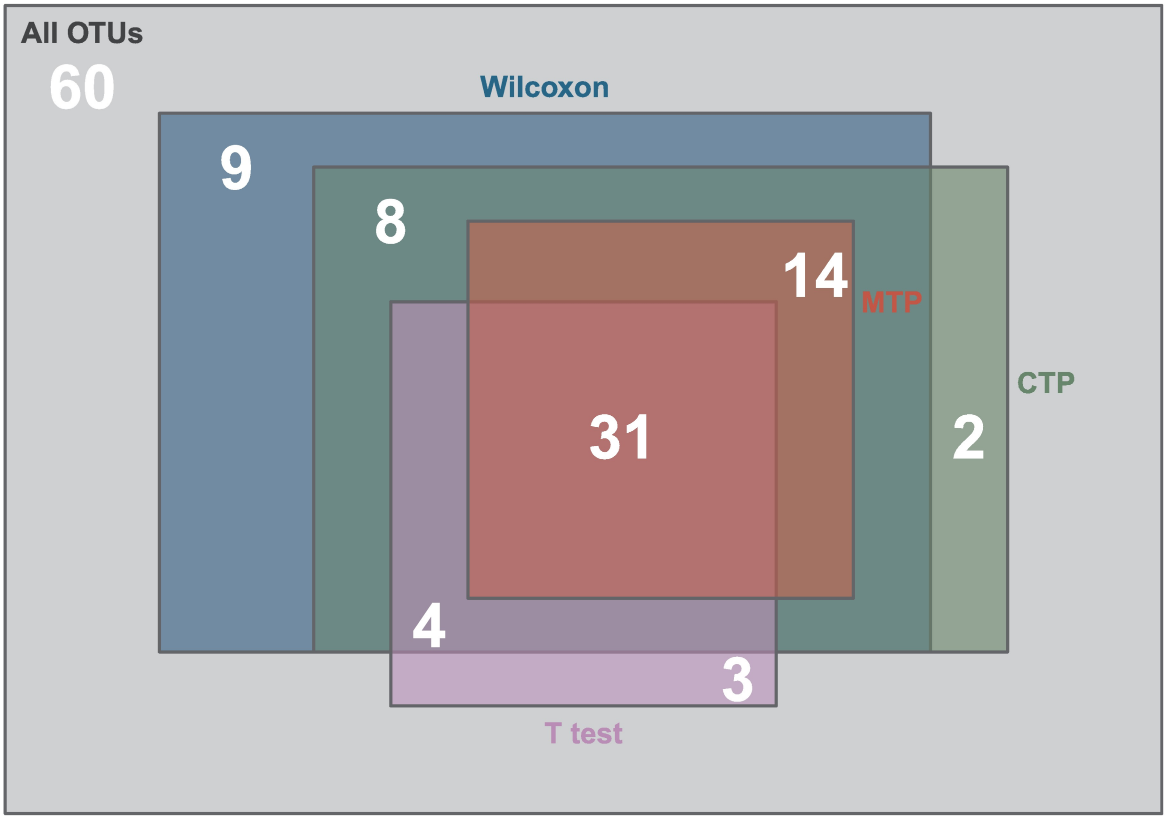

**Fig 5. Venn diagram for the OTUs.** Among all the 131 OTUs, 60 OTUs are not identified by any methods and the other 71 OTUs are identified by at least one method. For example, "31" in the intersection of all sets indicates that 31 OTUs are identified by all methods; while "4" located in the intersection of three sets, indicates that 4 OTUs are identified by three methods, namely, the T test, the CTP model, and the Wilcoxon test.

https://doi.org/10.1371/journal.pcbi.1006329.g005

beta-binomial (ZIBB) model, are available in [20–22]. It is of interest to develop a marginalized modeling approach for ZIB or ZIBB.

More recently, there has been increasing interest in analyzing correlated zero-inflated semi-continuous data. The correlation may stem from the structure of clustered data or from

**Table 2. Top 10 OTUs identified by the MTP model.**

| Rank | ID | Species | Est | SE | p Value |
|---|---|---|---|---|---|
| 1 | 237040 | g_Alicyclobacillus | 2.3262 | 0.3148 | < 1E-16 |
| 2 | 101810 | g_Helicobacter[1] | -1.3934 | 0.1460 | < 1E-16 |
| 3 | 52884 | g_Helicobacter[2] | 1.9091 | 0.2875 | 4.88E-15 |
| 4 | N10167 | o_Bacteroidales[1] | 2.5761 | 0.4234 | 8.33E-15 |
| 5 | 381715 | f_Ruminococcaceae | 2.7674 | 0.4777 | 6.82E-14 |
| 6 | 269548 | g_Helicobacter[3] | 1.2357 | 0.1833 | 3.57E-13 |
| 7 | N26397 | Acetobacter aceti | 1.8841 | 0.3055 | 4.86E-13 |
| 8 | 294146 | Acetobacterorleanensis | 2.0662 | 0.3387 | 5.66E-13 |
| 9 | N2007 | Acinetobacterlwoffii | 2.5549 | 0.4498 | 1.19E-12 |
| 10 | N8891 | g_Mucispirillum | 1.5868 | 0.2679 | 1.53E-11 |

Est: estimation of the coefficient of treatment in the second submodel;

SE: standard error of the coefficient of treatment in the second submodel.

https://doi.org/10.1371/journal.pcbi.1006329.t002

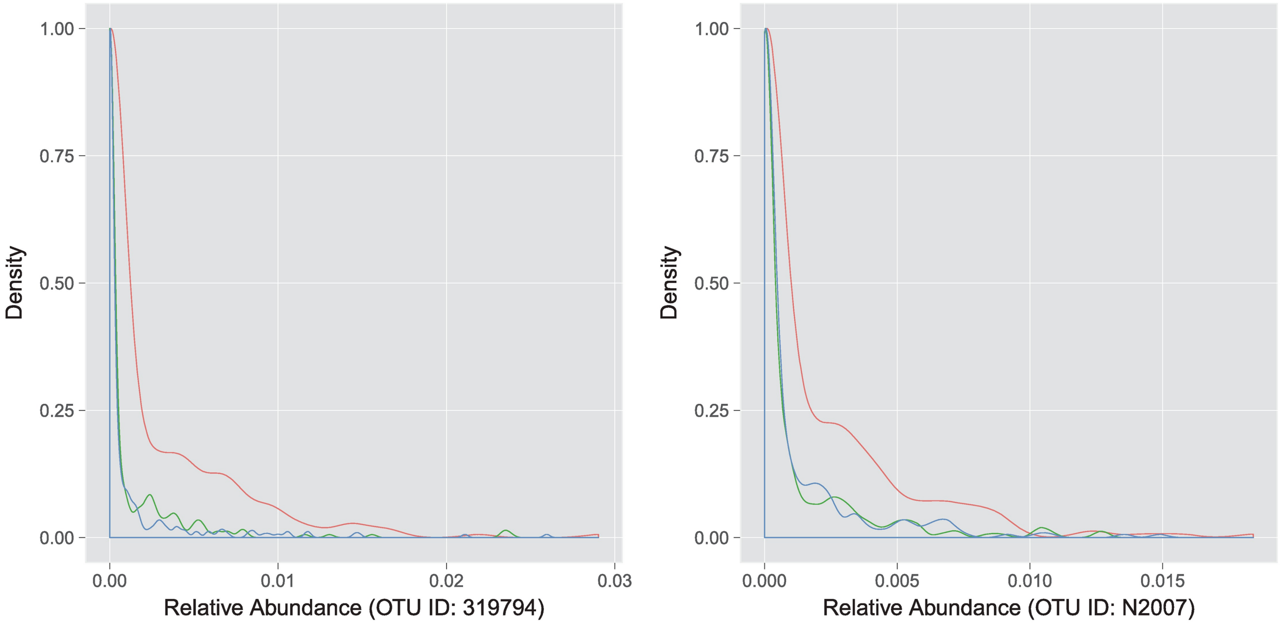

**Fig 6. Density curves for two OTUs.** The blue curve shows the density of the observed data. The green curve shows the density of predictions from the MTP model while the red curve represents the density of predictions from the CTP model.

https://doi.org/10.1371/journal.pcbi.1006329.g006

longitudinal data where repeated measures are correlated for the same subject. Typically, random effects are included to account for the correlations between observations [10, 15, 23–25]. However, similar limitation exists in these two-part random effects models, as they cannot account for covariate effects on the marginal mean. Recently, Smith et al. [26] proposed a marginalized two-part model for longitudinal semicontinuous data based on the log-skew normal distribution for positive values. In future studies, we will extend our marginalized two-part model to correlated semi-continuous data bounded by 0 and 1.

Finally, it is of interest to consider different microbiomes together, taking into account the constraint that the relative abundances of all OTUs sum to 1. Scealy and Welsh [27, 28] considered Kent models for such compositional data. It merits further consideration to incorporate zero values in the Kent model framework.

## Supporting information

**S1 Text. Likelihood derivation.**
(PDF)

**S1 Code. SAS code.** The main SAS codes for the conventional two-part model and the proposed marginalized two-part model are shown in this section.
(PDF)

**S1 Fig. Type I errors for the sample size 400.** This figure shows the type I errors of the four methods for sample size 400. The results in the upper panels correspond to the setting $\alpha_1 = 0$, $\gamma_1 = 0$ and the lower panels correspond to setting $\alpha_1 = 1$, $\gamma_1 = 0$. In each setting, the left panel shows the results for significance level 0.01 and the right panel shows the results for significance level 0.05. The dashed horizontal line in each panel represents the significance level.
(TIF)

**S2 Fig. Powers for the sample size 400.** This figure shows the powers of the four methods for sample size 400. The upper panel contains the power corresponding to the setting $\alpha_1 = 0$, $\gamma_1 = 1$ and the lower panel shows the power corresponding to the setting $\alpha_1 = 1$, $\gamma_1 = 1$. In each setting, the left figure shows the results for significance level and the right panel shows the results for significance level 0.05.
(TIF)

**S3 Fig. Zero-inflation of the skin data.** The figure shows the distributions of relative abundances of 6 OTUs. From the upper panel to the lower panel and from the left to the right, the proportions of zero values for these 6 OTUs are 0.77%, 3.45%, 4.97%, 14.18%, 29.89%, and 48.28%, respectively.
(TIF)

**S4 Fig. The figure shows the percentages of zero abundance in the 261 mouse skin samples for all 131 core OTUs.** The lower quartile and the upper quartile of the percentages are 20.11% and 48.28%, respectively.
(TIF)

**S5 Fig. Skewness of the skin data.** The figure shows the histogram of the relative abundance for 6 OTUs. The first one in every panel is the histogram of the OTU in the original scale, while the second one in every panel shows the histogram after logit transformation.
(TIF)

**S1 Table. The detailed results of the MTP model.** The table shows the detailed results of all the 45 OTUs that are identified by the proposed MTP model.
(PDF)

## Author Contributions

**Formal analysis:** Haitao Chai, Lei Liu.

**Funding acquisition:** Lei Liu.

**Investigation:** Haitao Chai, Hongmei Jiang.

**Methodology:** Haitao Chai, Hongmei Jiang, Lu Lin, Lei Liu.

**Project administration:** Lei Liu.

**Software:** Haitao Chai.

**Supervision:** Hongmei Jiang, Lu Lin, Lei Liu.

**Writing – original draft:** Haitao Chai.

**Writing – review & editing:** Hongmei Jiang, Lu Lin, Lei Liu.

## References

1. Gilbert JA, Meyer F, Bailey MJ. The Future of microbial metagenomics (or is ignorance bliss?). Isme Journal. 2011; 5(5):777–779. https://doi.org/10.1038/ismej.2010.178 PMID: 21107444
2. Everard A, Cani PD. Diabetes, obesity and gut microbiota. Best Practice & Research Clinical Gastroenterology. 2013; 27(1):73–83. https://doi.org/10.1016/j.bpg.2013.03.007
3. Musso G, Gambino R, Cassader M. Obesity, diabetes, and gut microbiota: the hygiene hypothesis expanded? Diabetes Care. 2010; 33(10):2277–2284. https://doi.org/10.2337/dc10-0556 PMID: 20876708

4. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, Antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. Cell Host & Microbe. 2015; 18(4):489–500. https://doi.org/10.1016/j.chom.2015.09.008

5. Srinivasan S, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. Plos One. 2012; 7(6):e37818. https://doi.org/10.1371/journal.pone.0037818 PMID: 22719852

6. Garrett WS. Cancer and the microbiota. Science. 2015; 348(6230):80–86. https://doi.org/10.1126/science.aaa4972 PMID: 25838377

7. Schwabe RF, Jobin C. The microbiome and cancer. Nature Reviews Cancer. 2013; 13(11):800–812. https://doi.org/10.1038/nrc3610 PMID: 24132111

8. Tyler AD, Smith MI, Silverberg MS. Analyzing the human microbiome: a "how to" guide for physicians. American Journal of Gastroenterology. 2014; 109(7):983–93. https://doi.org/10.1038/ajg.2014.73 PMID: 24751579

9. Manning WG. A two-part model of the demand for medical care: preliminary results from the health insurance study. Health, Economics, and Health Economics. 1981; p. 103–123.

10. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics. 2016; 32(17):2611–2617. https://doi.org/10.1093/bioinformatics/btw308 PMID: 27187200

11. Ospina R, Ferrari SLP. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis. 2012; 56(6):1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

12. Peng X, Li G, Liu Z. Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. Journal of Computational Biology. 2015; 23(2):102–110.

13. Smith VA, Preisser JS, Neelon B, Maciejewski ML. A marginalized two-part model for semicontinuous data. Statistics in Medicine. 2014; 33(28):4891–4903. https://doi.org/10.1002/sim.6263 PMID: 25043491

14. Tobin J. Estimation of Relationships for Limited Dependent Variables. Econometrica. 1958; 26(1):24–36. https://doi.org/10.2307/1907382

15. Liu L, Strawderman RL, Cowen ME, Shih YC. A flexible two-part random effects model for correlated medical costs. Journal of Health Economics. 2010; 29(1):110–123. https://doi.org/10.1016/j.jhealeco.2009.11.010 PMID: 20015560

16. Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinformatics. 2015; 31(20):3322–3329. https://doi.org/10.1093/bioinformatics/btv364 PMID: 26079350

17. Srinivas G, Möller S, Wang J, Künzel S, Zillikens D, Baines JF, et al. Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. Nature Communications. 2013; 4(9):2462. https://doi.org/10.1038/ncomms3462 PMID: 24042968

18. Long DL, Preisser JS, Herring AH, Golin CE. A marginalized zero-inflated Poisson regression model with overall exposure effects. Statistics in Medicine. 2014; 33(29):5151–5165. https://doi.org/10.1002/sim.6293 PMID: 25220537

19. Preisser JS, Das K, Long DL, Divaris K. Marginalized zero-inflated negative binomial regression with application to dental caries. Statistics in Medicine. 2016; 35(10):1722–1735. https://doi.org/10.1002/sim.6804 PMID: 26568034

20. Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. London: Chapman & Hall; 2004.

21. Albert JM, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. Statistical Methods in Medical Research. 2014; 23(3):257–278. https://doi.org/10.1177/0962280211407800 PMID: 21908419

22. Gilthorpe MS, Frydenberg M, Cheng Y, Baelum V. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. Statistics in Medicine. 2009; 28(28):3539–3553. https://doi.org/10.1002/sim.3699 PMID: 19902494

23. Olsen MK, Schafer JL. A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. Journal of the American Statistical Association. 2001; 96(454):730–745. https://doi.org/10.1198/016214501753168389

24. Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. Statistical Methods in Medical Research. 2002; 11(4):341–355. https://doi.org/10.1191/0962280202sm291ra PMID: 12197301

**25.** Liu L, Strawderman RL, Johnson BA, O'Quigley JM. Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. Statistical Methods in Medical Research. 2016; 25(1):133–152. https://doi.org/10.1177/0962280212443324 PMID: 22474003

**26.** Smith VA, Neelon B, Preisser JS, Maciejewski ML. A marginalized two-part model for longitudinal semi-continuous data. Statistical Methods in Medical Research. 2017; 26(4):1949–1968. https://doi.org/10.1177/0962280215592908 PMID: 26156962

**27.** Scealy JL, Welsh AH. Regression for compositional data by using distributions defined on the hypersphere. Journal of the Royal Statistical Society. 2011; 73(3):351–375. https://doi.org/10.1111/j.1467-9868.2010.00766.x

**28.** Scealy JL, Welsh AH. Fitting Kent models to compositional data with small concentration. Statistics & Computing. 2014; 24(2):165–179. https://doi.org/10.1007/s11222-012-9361-5