Richard Meier[1] / Jeffrey A. Thompson[1] / Mei Chung[2] / Naisi Zhao[2] / Karl T. Kelsey[3] / Dominique S. Michaud[2] / Devin C. Koestler[1]

A Bayesian framework for identifying consistent patterns of microbial abundance between body sites

[1] Department of Biostatistics and Data Science, University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, USA, E-mail: rmeier2@kumc.edu. https://orcid.org/0000-0003-1062-1444.
[2] Department of Public Health and Community Medicine, Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, USA
[3] Department of Epidemiology, Department of Pathology and Laboratory Medicine, Brown University, 70 Ship Street, Providence, RI 02912, USA

Abstract:
Recent studies have found that the microbiome in both gut and mouth are associated with diseases of the gut, including cancer. If resident microbes could be found to exhibit consistent patterns between the mouth and gut, disease status could potentially be assessed non-invasively through profiling of oral samples. Currently, there exists no generally applicable method to test for such associations. Here we present a Bayesian framework to identify microbes that exhibit consistent patterns between body sites, with respect to a phenotypic variable. For a given operational taxonomic unit (OTU), a Bayesian regression model is used to obtain Markov-Chain Monte Carlo estimates of abundance among strata, calculate a correlation statistic, and conduct a formal test based on its posterior distribution. Extensive simulation studies demonstrate overall viability of the approach, and provide information on what factors affect its performance. Applying our method to a dataset containing oral and gut microbiome samples from 77 pancreatic cancer patients revealed several OTUs exhibiting consistent patterns between gut and mouth with respect to disease subtype. Our method is well powered for modest sample sizes and moderate strength of association and can be flexibly extended to other research settings using any currently established Bayesian analysis programs.

Keywords: association, Bayesian, consistent pattern, microbial abundance, microbiome, zero-inflated beta regression

DOI: 10.1515/sagmb-2019-0027

1 Introduction

Microbial communities inhabit virtually every part of the human body and can differ across individuals. Even within the same individual, microbial communities often change with anatomical location (Faith et al., 2013). In this context, it is not surprising that the human microbiome plays an important role in a wide range of diseases, including even life threatening conditions such as cancers. In their review, Goodman and Gardner (2018) summarize several compelling examples, such as increased *Fusobacterium* species associating with tumors in colon and *Helicobacter pylori* inducing lymphoma and gastric cancer. More recently, bacteria have been identified in pancreatic tissue in cancer patients (del Castillo et al., 2019) and have been shown to play a role in carcinogenesis in the pancreas (Pushalkar et al., 2018). Additional studies have also reported evidence that certain oral bacteria and periodontal disease associate with an increased risk in pancreatic cancer (Michaud et al., 2012; Fan et al., 2016). Finally, it has been shown that *Fusobacterium nucleatum*, a common oral bacterium, produces a protein that allows itself and other bacteria to travel through the endothelium in the mouth and into the blood stream, allowing them to migrate to other body sites (Fardini et al., 2011). Despite the empirical evidence, little is understood about how these associations originate and no confirmatory study has conclusively established their biological mechanism. This motivates the question of whether microbes exist for which changes in abundance (mean relative abundance or rate of presence) with respect to disease status in the oral cavity correspond to changes in abundance in gut samples. In other words, are there microbes for which fluctuations

Richard Meier is the corresponding author.

in their abundance are preserved across disease status between mouth and gut? Identification of such species, exhibiting pairwise stratified association (PASTA) between two body sites, may allow further insight into mechanisms and the biology behind a disease. Furthermore, it may also provide new opportunities for treatment or detection and even potentially enable a researcher or medical professional to learn about the disease in the gut by monitoring oral samples. Considering that gut samples can only be acquired through invasive surgical procedures, PASTA microbes could constitute invaluable clinical markers.

Data arising from 16S rRNA sequencing for assessing the microbiome takes the form of compositional count tables. The term operational taxonomic unit (OTU) can be understood as a group of closely related microbes on a given taxonomic level, for example: phylum, genus, or species. For a given experiment, in which abundance of microbes is quantified in a series of biological samples, each cell in row i and column j of such a table represents how often a species or OTU i was observed in sample j (Table 1). Unfortunately, these data are intricate with total column counts (sequencing depth and microbial yield) differing between samples, high frequency of zero values (i.e. sparsity), and the constant sum constraint problem that can create spurious associations when few rows dominate the majority of counts (Tsilimigras & Fodor, 2016; Gloor et al., 2017).

Table 1: Hypothetical example of a microbial abundance data table.

Genus	Sample 1 Count	Sample 2 Count	Sample 3 Count	Sample 4 Count	⋯ ⋯
Actynomices	0	0	3	5	⋯
Atopobium	0	27	10	6	⋯
Fusobacterium	0	14	0	0	⋯
⋯	⋯	⋯	⋯	⋯	⋯
Sample total	671	2390	1502	1883	⋯

Rows represent genera, which are groups of closely related microbes and an example of a type of operational taxonomic unit (OTU). For a given sample and OTU, each cell in the table counts how often said OTU was observed in said sample through the 16S rRNA sequencing technique. All counts in this type of table are expected to increase with the total number of observed OTUs in the respective sample. These column totals can be understood as the sample signal intensity and change based on experimental parameters for each sample.

Due to its complexity, many different modeling strategies have been proposed for the analysis of microbial 16S rRNA abundance data. When investigating an individual microbe (or a specific group of microbes), current strategies predominantly aim to understand the relationship between abundance and selected phenotypes. Three major parametric approaches are employed by most researchers: discrete data models such as Zero-inflated Poisson or Zero-inflated Negative Binomial regression (Zhang et al., 2017; Xia, Sun & Chen, 2018); log-ratio Aitchison models that explicitly address the constant sum constraint by treating the ratio of abundance counts between two taxonomic units as the response (Shi, Zhang & Li, 2016; Tsilimigras & Fodor, 2016; Gloor et al., 2017); and lastly, relative abundance models that transform counts into sample proportions and fit semi-continuous models to the data such as Zero-inflated Beta regression (ZIBR) (Chen & Li, 2016; Peng, Li & Liu, 2016; Xia, Sun & Chen, 2018). Each approach can present specific advantages and limitations, where the most suitable model will depend on the circumstances of the research study. While log-ratio Aitchison models are mandatory in datasets either measuring high phylogenetic levels with few taxonomic units or exhibiting low community diversity (Tsilimigras and Fodor 2016), discrete data and relative abundance models are convenient to address sparsity in high diversity settings. To date, neither of these modeling strategies has been utilized to test for PASTA relationships and there presently exists no general testing approach that is applicable regardless of the parametric modeling strategy. Alternatively, non-parametric inter-rater strategies can be employed to test for agreement or association between body-sites. These strategies assume that there are individual raters that are presented with two different scenarios or cases, each of which they have to assign to either a category or numeric value. The methods then ask the question whether individual raters tend to make assignments that agree or associate between the two scenarios. Popular examples are Cohen's kappa (Cohen, 1960; Fleiss, 1971) for categorical responses and Pearson or Spearman correlation for numeric responses (Schober, Boer & Schwarte, 2018). These methods do not necessarily require knowledge about the distribution of the response and are applicable even if there is strong disagreement or variability between individual raters. However, they do not allow accounting for confounders or other sources of variation, and they require paired samples.

Here, we present an approach to test for PASTA that is applicable regardless of the data model and regardless whether all, some, or none of the samples are paired. The question of PASTA relationships with respect to body site is translated into a question of association of population parameters (such as mean relative abundance) between the two body sites. A test is then proposed based on applying a correlation statistic to parameter estimates. Testing and adjusting for paired samples is made convenient by utilizing a Bayesian modeling framework. For the purpose of illustration, this paper will focus on modeling relative abundance via a ZIBR

model, though as stated before, the approach is not limited to any partiular data model. After establishing the data model and introducing the approach, viability and performance are evaluated via simulation studies and through the analysis of a biological dataset involving microbiome data collected from the gut and specific oral sites in patients with pancreatic cancer and other diseases of the foregut. Finally, strengths, limitations and opportunities for future methodological development are discussed.

2 Methods

2.1 Experimental design

A study suitable to answer the previously described research question can be broken down into the following steps. First, an appropriate subject population exhibiting the disease or target phenotype is identified and biological samples from the two body sites of interest are collected. Multiple samples from the same patient within and across body sites are possible, but not necessarily required. Next, sample preparation and 16S rRNA sequencing are performed. This sequencing technique aims to identify and count hypervariable DNA patterns that are specific to microbial species and OTUs, but that do not exist in human DNA; the rationale being that the DNA content of a group of microbes is approximately proportional to their abundance in the sample. So, by counting how often signatures belonging to a specific OTU are observed, we can obtain an estimate of its abundance relative to how many microbes were observed, in total. After OTUs have been counted, our proposed statistical test is performed individually for each OTU, testing the null hypothesis that there is no PASTA relationship for each specific set of considered microbes. This test is performed by fitting a statistical regression model to the row vector of abundance values corresponding to a target OTU, followed by the calculation of a test statistic T_θ based on the parameter estimates (for example, rate of absence or mean abundance) obtained from said model. This statistic will be small when H_0 is true and large when H_0 is false. An overview of the experimental design for testing the hypothesis of PASTA can be found in Figure 1.

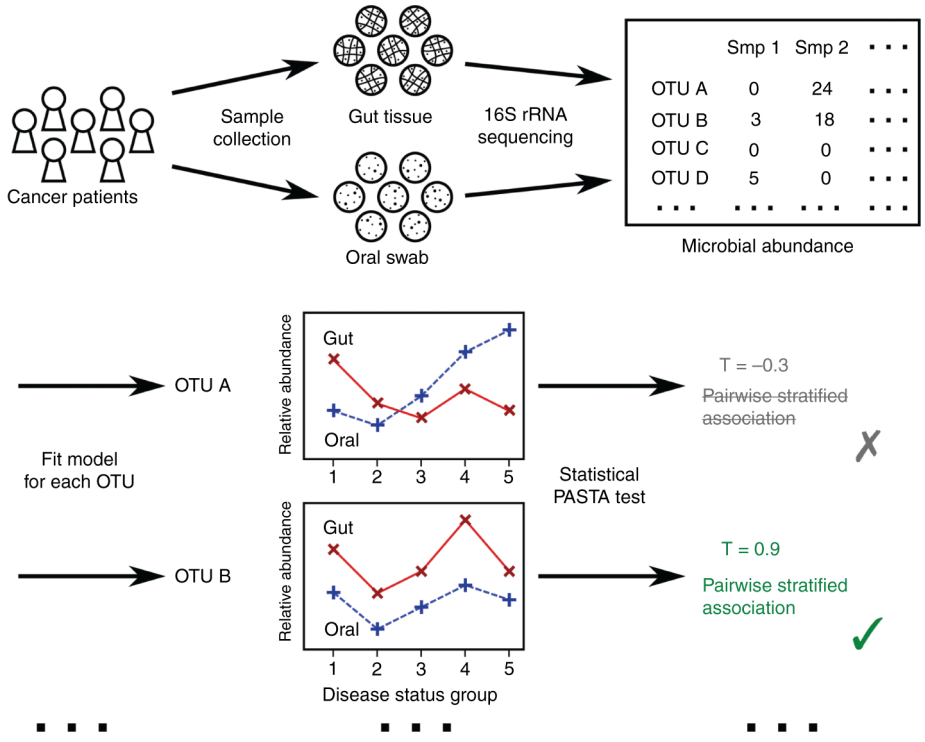

Figure 1: Overview of the experimental setup to test for pairwise stratified association (PASTA). Oral and gut samples are obtained from cancer patients and 16S rRNA sequencing is performed on each sample. The resulting microbial abundance data is used to fit a statistical regression model to each observed OTU across all samples. Finally, abundance estimates across strata are used to test whether abundance patterns in disease status are preserved between mouth and gut.

2.2 Data model

In what follows we consider abundance on two taxonomic levels: the genus and the Amplicon Sequence Variant (ASV) level, the latter representing unique biological sequences that were identified from 16S genes (Callahan, McMurdie & Holmes, 2017). In order to make abundance values comparable across samples and bring them to the same scale, raw counts are first transformed into relative abundance values. For a given sample, relative abundance of an OTU refers to the number of times that OTU was observed, scaled by the total number of observed OTUs for that sample. It represents the proportion of times an OTU was observed in a given sample.

Let Y_k denote the relative abundance of a specific OTU for sample k. This response can be modeled as a Zero-inflated Beta distribution with probability density $f_{Y_k}(y|p_k, \omega_k, \phi_k)$. This model assumes that the case $Y_k = 0$ occurs with probability p_k and that given $Y_k > 0$, the response Y_k follows a Beta distribution with mean ω_k and dispersion ϕ_k. For a given OTU and sample, the probability of absence p defines how likely it is to observe no microbe comprising that OTU within said sample. The mean non-zero relative abundance ω represents the mean relative abundance given that microbes comprising the OTU are actually observed. The mean of Y_k, the overall mean relative abundance, is then $E[Y_k] = \mu_k = \omega_k(1 - p_k)$. The probability density function of this distribution can be expressed as follows:

$$f_{Y_k}(y) = \begin{cases} p_k & \text{if } y = 0 \\ (1 - p_k) \cdot \frac{\Gamma(\phi_k)}{\Gamma(\omega_k \cdot \phi_k)\Gamma((1-\omega_k)\cdot\phi_k)} y^{\omega_k \cdot \phi_k - 1}(1 - y)^{(1-\omega_k)\cdot\phi_k - 1} & \text{if } y > 0 \end{cases} \tag{1}$$

Before the statistical PASTA test can be performed, a Bayesian ZIBR model is fit to the data Y_k utilizing the likelihood f_{Y_k} and assuming a common dispersion parameter $\phi_k = \phi$ for all samples. The estimated posterior distributions of ω and p resulting from this model are then subsequently used to conduct the test.

Let $\boldsymbol{\omega}$ denote the vector of mean relative abundances for all samples, $\mathbf{p}$ denote the vector of probabilities of absence for all samples, $\boldsymbol{\beta}, \boldsymbol{\delta}$ denote coefficient vectors, $\mathbf{b}, \mathbf{d}$ denote random effect vectors and $\mathbf{Q}, \mathbf{R}, \mathbf{W}, \mathbf{X}$ represent design matrices. The design matrices code how covariates impact the model parameters via the following link functions:

$$logit(\boldsymbol{\omega}) = \boldsymbol{\beta}\mathbf{X} + \mathbf{bR} \quad \text{and} \quad logit(\mathbf{p}) = \boldsymbol{\delta}\mathbf{W} + \mathbf{dQ} \tag{2}$$

For our application, the matrices $\mathbf{W}$ and $\mathbf{X}$ are used to model the strata of body site and disease status, but can additionally be used to adjust for other, fixed covariates (e.g. subject age, gender, smoking status, and/or other potential confounders or sources of variation). On the other hand, the optional inclusion of $\mathbf{Q}$ and $\mathbf{R}$ permits one to account for correlation structures, such as within-subject correlation when multiple samples are collected from the same patient. If, for example, the probability of absence of a given OTU in sample k from subject j_k is assumed to be impacted by disease status g_{j_k}, body site s_k, age A_{j_k} and within-subject correlation, we would formulate $\mathbf{W}$, $\mathbf{Q}$ and our model for the probability of absence as follows:

$$logit(p_k) = \delta_{1,g_{j_k},s_k} + \delta_2 A_{j_k} + d_{1,j_k} \tag{3}$$

Here, δ_{1,g_{j_k},s_k} captures the effect of disease status g_{j_k} and body site s_k on the probability of absence, δ_2 represents the effect of age on the probability of absence, and d_{1,j_k} is a subject specific random intercept. Analogously, if mean relative abundance is believed to be impacted by the same covariates in the same way, except that age was believed to have no effect, we would formulate $\mathbf{X}$, $\mathbf{R}$ and our model for non-zero relative abundance such that:

$$logit(\omega_k) = \beta_{1,g_k,s_k} + b_{1,k} \tag{4}$$

Here, β_{1,g_{j_k},s_k} captures the effect of disease status g_{j_k} and body site s_k on the mean non-zero relative abundance and b_{1,j_k} is a subject specific random intercept. The specific models used in our analysis of the pancreatic cancer dataset are presented in the Results Section are further discussed in Section 2.7.

Let $t, i \in \mathbb{N}$ denote placeholder indices for any potential coefficient specified in the above ZIBR model. Then all posterior distributions were estimated, utilizing the following independent prior distributions:

$$\pi(\beta_t), \pi(\delta_t) \sim N(0, 100);\ \pi(b_{t,i}) \sim N(0, \zeta_t);\ \pi(d_{t,i}) \sim N(0, \xi_t); \tag{5}$$

$$\pi(\sqrt{\phi}) \sim Unif(1,100);\ \pi(\zeta_t^{-1}), \pi(\xi_t^{-1}) \sim Gam(0.01, 0.01) \tag{6}$$

Priors were chosen to be weakly informative, with the exception of $\sqrt{\phi}$ being restricted to values larger than or equal to 1 in an attempt to stabilize estimation of means. Under these priors and for some integer vectors $\mathbf{T}, \mathbf{I_1}, \mathbf{I_2}$ the posterior distribution of parameters will then satisfy the following:

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \mathbf{d}, \boldsymbol{\zeta}, \boldsymbol{\xi}|\mathbf{Y}) \propto & \prod_{t_1}^{T_1} \pi(\beta_{t_1}) \cdot \prod_{t_2}^{T_2} \pi(\delta_{t_2}) \cdot \prod_{t_3}^{T_3} \left(\pi(\zeta_{t_3}) \prod_{i_{t_3}}^{I_{1,t_3}} \pi(b_{t_3,i_{t_3}}|\zeta_{t_3}) \right) \cdot \\ & \prod_{t_4}^{T_4} \left(\pi(\xi_{t_4}) \prod_{i_{t_4}}^{I_{2,t_4}} \pi(d_{t_4,i_{t_4}}|\xi_{t_4}) \right) \cdot f(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{b}, \mathbf{d}, \boldsymbol{\zeta}, \boldsymbol{\xi})\end{aligned} \tag{7}$$

There are generally no analytical solutions for the posterior distributions of the coefficients when random effects are present. Regardless, whether the model structure is a special case that allows for analytical calculation of posterior distributions or whether we employ a more complex model where this is not possible, posterior distributions can be estimated via Markov chain Monte Carlo (MCMC) methods. Briefly, MCMC procedures allow one to draw arbitrarily large samples from a posterior distribution that will numerically approximate the said distribution, as the number of draws increases. Models were fit via this method in the software OpenBUGS (version 3.2.3 rev 1012) via the R (version 3.4.0) package "R2OpenBUGS" (version 3.2.3.2).

2.3 Formal definition of pairwise stratified association (PASTA)

In order to understand how the Bayesian regression model can be used to conduct the desired hypothesis test, we will first provide a formal definition of PASTA. Let s denote a grouping variable for which two groups are to be compared. For our purposes, this grouping variable represents body sites: $s = 1$ denotes gut and $s = 2$ denotes mouth. Let g denote another grouping variable with three or more distinct categories. This grouping variable will represent different types of disease status, more specifically cancer-subtype. Let θ_{sg} be a population parameter of the response for a given body site s and disease status g. The population parameter represents fundamental properties of the distribution of the response. For the here considered ZIBR model, p, ω and μ are relevant candidates for θ. If PASTA holds for a given OTU, then either p, ω or μ will associate between the two body sites, because they all relate to the magnitude of abundance.

We thus define: The parameter θ exhibits PASTA with respect to s and g if there exists an increasing function $h(x)$ such that $\theta_{1g} = h(\theta_{2g})$ holds for all $g \in \{1, 2, \ldots, G\}$, where $G \geq 3$. Conceptually, this definition says that as we move from one disease status group to another, if θ increases in oral samples, it will also increase in gut samples. Analogously, if θ decreases from one disease status group to another in the mouth, it will also decrease in the gut. A visualization is provided in Figure 2.

Body site	Disease status			
	$g = 1$	$g = 2$	$g = 3$	...
$s = 1$ (gut)	θ_{11}	θ_{12}	θ_{13}	...
$s = 2$ (mouth)	$\theta_{21} = h(\theta_{11})$	$\theta_{22} = h(\theta_{12})$	$\theta_{23} = h(\theta_{13})$	...

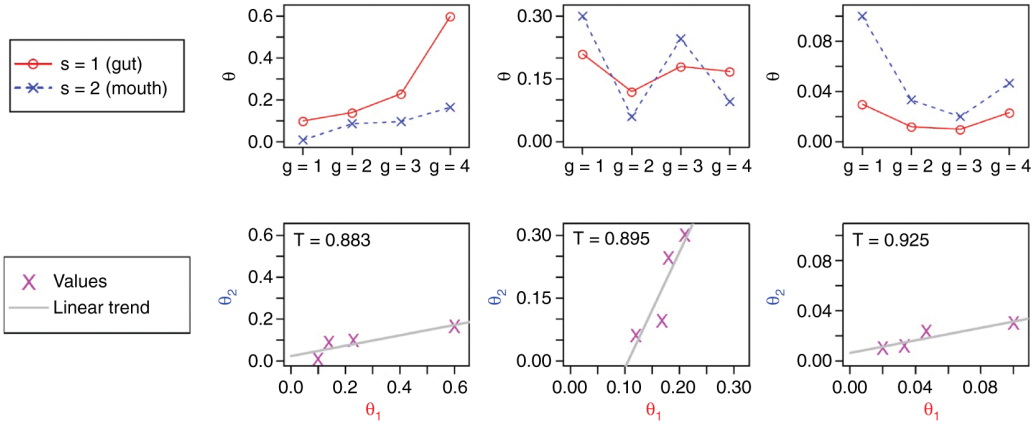

Figure 2: Visualization of pairwise stratified association (PASTA). Let θ represent a population parameter of interest, for example the mean relative abundance of a particular OTU. Each column of sub-figures below the table are examples of a PASTA relationship, i.e. of h being an increasing function. The first row plots parameter values of mouth and gut side-by-side and demonstrates that a variety of different scenarios are covered by this definition. In the second row, plotting parameter values of gut against parameter values of mouth reveals their association through a trend. T denotes Pearson correlation values between gut and mouth.

2.4 Testing for PASTA

Let $T(\mathbf{x}, \mathbf{y}) \in [-1, 1]$ denote a correlation statistic between two numerical vectors $\mathbf{x}, \mathbf{y}$; for example, the Pearson or Spearman correlation statistic. Under this definition, $T_\theta = T(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denotes the correlation statistic calculated for the two parameter vectors corresponding to $s = 1$ (e.g. parameters for disease status groups in the mouth) and $s = 2$ (e.g. parameters for disease status groups in the gut). Generally, if a PASTA relationship holds between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, this statistic should assume a larger value compared to cases where such a relationship does not hold. This means that we are able to formulate our desired test by rejecting H_0 if T_θ is larger than a specific threshold and fail to reject H_0 if it is less than said threshold.

In summary, assume that $\theta_{1g} = h(\theta_{2g})$ implies $T_\theta > t_c$ for some $-1 < t_c < 1$. The constant t_c represents a meaningful degree of association. For example, a value of $t_c = 0$ would mean that any tangible degree of association is meaningful, where a value of $t_c = 0.5$ would mean that a moderate degree of association is meaningful. This definition is useful because T_θ can score the degree of association without explicitly having to specify the shape of h. Considering the complexity of the biology underlying the samples, specifying h in advance may not only be hard to justify, but strong deviations of a chosen h from the true h could also result in missing promising associations. Instead, our regression model will allow each stratum (s, g) to have an independent effect on the response, leading to a unique, agnostic posterior distribution of each θ_{sg}. These unique posteriors are then in turn used to calculate the posterior distribution of T_θ and conduct the hypothesis test.

Based on this scoring definition of PASTA, we formulate our hypotheses in the following way:

$H_0 : T_\theta \leq t_c$, i.e. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ do NOT exhibit PASTA

$H_1 : T_\theta > t_c$, i.e. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ DO exhibit PASTA

While deriving analytical solutions of the distribution of $T_\theta|\mathbf{Y} = T(\boldsymbol{\theta}_1|\mathbf{Y}, \boldsymbol{\theta}_2|\mathbf{Y})$ will depend on the data model and may be difficult or even impossible to obtain depending on the modeling scenario, a general testing procedure can still be derived. As decribed earlier, MCMC methods allow one to conveniently obtain a large sample of posterior draws of each θ_{sg}, even when obtaining analytical solutions of posterior distributions is not possible. Furthermore, plugging the posterior draws of each MCMC iteration into T allows one to obtain posterior draws from T_θ itself. Let α denote the target credibility threshold, H_0 is then rejected if the lower bound $t_{Q\alpha}$ of the one-sided credible interval of $T_\theta|\mathbf{Y}$ exceeds t_c. This is equivalent to rejecting H_0 if the estimated probability of no association exceeds α, i.e. $Pr(T_\theta|\mathbf{Y} \leq t_c) < \alpha$. In detail, the step by step process for testing H_0 is as follows:

1. Specify a likelihood for the response data $\mathbf{Y}$ and prior distributions for the parameters $\boldsymbol{\theta}$.
2. Utilize a MCMC sampling scheme to draw a large number of samples from the posterior distributions of the parameters $\boldsymbol{\theta}|\mathbf{Y}$. One draw from the Markov Chain contains a unique draw for each θ_{sg}.
3. Calculate $T_v^* = T(\boldsymbol{\theta}_1^{*v}, \boldsymbol{\theta}_2^{*v})$ where $\boldsymbol{\theta}^{*v}$ denotes the vth MCMC draw. Then $\mathbf{T}^*$ is a large sample of the posterior distribution $T_\theta|\mathbf{Y}$.
4. Calculate the $\alpha \cdot 100\%$ sample quantile $t_{Q\alpha}$ of $\mathbf{T}^*$. If the Markov Chain is sufficiently long, the sample quantiles of $\mathbf{T}^*$ will closely approximate the quantiles of the true posterior distribution. The value $t_{Q\alpha}$ is thus the lower bound of the $(1 - \alpha) \cdot 100\%$ one-sided credible interval of $T_\theta|\mathbf{Y}$.
5. Reject H_0 if the lower bound $t_{Q\alpha}$ is larger than t_c.

This process is generally applicable regardless of the data model or the parameter being tested, as long as each θ_{sg} can be estimated without constraining them to a parameter space that implies PASTA.

2.5 Pancreatic cancer patient dataset

In order to evaluate validity of the approach in the context of microbiome data, analyses were performed based on a biological 16S rRNA sequencing dataset first published in del Castillo et al. (2019). This dataset contained

samples of various gut and oral sites from 77 patients with pancreatic cancer with age range 31–86 years. Sequencing was performed utilizing the Illumina MiSeq System and alignments were performed using BLASTN against a reference library combining sequences from HOMD (version 14.5), Greengenes Gold and the NCBI 16S rRNA reference sequence set. OTU counts were obtained utilizing the QIIME (Quantitative Insights Into Microbial Ecology 22) software package version 1.9.1, while the unique Amplicon Sequence Variant (ASV) counts were calculated using the QIIME2 software package release 2018.4. The former was used to obtain taxonomic genus level counts, whereas the latter was used to obtain rarefied ASV level information, calculated based on sequencing data rarefied at a sampling depth of 1200. Both genus level and ASV level counts were considered for analysis. Before fitting statistical models to the data, relative abundance values of less than 0.01 were treated as noise and set to 0. To ensure inference was based on sufficient signal, OTUs and ASVs were only tested if more than 5of all samples exhibited non-zero values.

The dataset was used to both guide simulation studies (described in the next section) and to deploy models to identify potential microbes that may exhibit a PASTA pattern.

2.6 Simulation studies

Before simulations were performed, an empirical approach was pursued in order to obtain sampling distributions of the parameters p, ω, ϕ that would be representative of biological microbiome data. First, a marginal, unstratified ZIBR model was fit to the pancreatic cancer dataset that assumed all samples of relative abundance for a given OTU originated from the same distribution. These model fits yielded a single estimate of p, ω and ϕ for each OTU. These estimates were then assumed to be representative of or approximate the true distribution of parameters in biological data. In the next step, the estimates were used to obtain smooth probability distributions that parameters could be sampled from during the simulation studies. For both p and ω, individual Beta distribution models were fit to the marginal estimates in order to obtain their smooth sampling distributions. On the other hand, $\log \phi$ was sampled via a Normal distribution through an observed linear relationship between $\log \phi$ and $\log \omega$ that was present on the ASV level and the genus level. More specifically, since our models assumed fixed dispersion among all groups, dispersion was sampled from $(\log \phi | \min_{sg}\{\log \omega_{sg}\}) \sim N(a \min_{sg}\{\log \omega_{sg}\} + b, \sigma^2)$, where a, b, σ^2 differed between genus and ASV level.

After the smooth sampling distributions were obtained, the performance of PASTA tests was evaluated via simulations. Let t denote a target, fixed degree of association, n denote the number of observations in each stratum (s, g) and $t_c = 0$ denote the tested degree of association. A single simulation run was carried out by first randomly drawing all θ_{sg} parameters from the representative sampling distributions, until $|T_\theta - t| < 0.001$ was satisfied. This process yields parameters that are both representative and that also exhibit a target degree of association (within a small error margin). Next, the drawn parameters satisfying this condition were plugged into the likelihood of the ZIBR data model, which was in turn used to draw a random sample of relative abundance values. This simulated pseudo-data was then used to fit the Bayesian ZIBR model and conduct our hypothesis test. Each considered scenario was simulated 1000 times and statistical power for given t, n and t_c was then estimated as the proportion of times H_0 (i.e. $T_\theta \leq 0$) was rejected. We specifically considered Pearson correlation as choice for $T(x, y)$ in this simulation.

An additional restriction was put in place for sampling pseudo-data in order to prevent rare cases of sparse datasets with insufficient signal to perform the analysis. If a generated pseudo-dataset contained more than three sub-strata (s, g) in which all observations exhibit a response value of either all 0 or all 1, then it was rejected and a new pseudo-dataset was sampled.

2.7 Model fitting

Let j_k denote the unique identifier index for the subject and s_k denote the body site that sample k originated from. Also, let and g_{j_k} denote the disease status for subject j_k, let X_k be the log of total sample abundance for sample k and let b_{j_k} denote the random intercept for subject j_k. The three different models that were utilized in this study are shown below:

Model A: $logit(\omega_k) = \beta_{s_k, g_{j_k}}$ & $logit(p_k) = \beta_{s_k, g_{j_k}}$

Model B: $logit(\omega_k) = \beta_{s_k, g_{j_k}} + b_{j_k}$ & $logit(p_k) = \beta_{s_k, g_{j_k}} + b_{j_k}$

Model C: $logit(\omega_k) = \beta_{s_k, g_{j_k}} + b_{j_k}$ & $logit(p_k) = \beta_{1, s_k, g_{j_k}} + X_k \beta_2 + b_{j_k}$

Model A was utilized in the simulation studies. Model B was utilized for fitting ASV level data, while Model C was utilized for fitting genus level data. This choice was made because scaling OTU counts to relative abundance

will only make non-zero relative abundance comparable between samples, but not the rate of absence. This is due to the fact that, even if the true probability of absence p for a specific OTU is very high, if more microbes are overall observed in sample 1 than in sample 2, then the probability of observing none of the microbes belonging to the target OTU in sample 1 is much lower than in sample 2. For example, if a total of 1,000,000 microbes live in a body site and 100 of them belong to the genus *Prevotella*, then if we randomly extract 1000 microbes from this body site with our sample, we would expect to only rarely find one of these 100 microbes in our sample. However, if our sample randomly extracts 100,000 microbes from the body site, it would be rare to find none of the 100 microbes in it that belong to the genus *Prevotella*. So since the genus level data was not rarefied, the total sample abundance differed between samples and an adjustment was necessary, whereas the ASV level data was rarefied and did not require adjustment for total sample abundance.

In order to achieve potentially better convergence behaviour and to simplify and speed up the model fitting, the logistic regression component of the model was fit independently of the Beta regression component, in all cases. The resulting posterior chains of p and ω were then used to calculate the posterior chain of ω. This approach is justified under the assumption that p and ω are independent after adjusting for covariates, but may be inadequate when there are confounders affecting both parameters not accounted for in the model.

3 Results

3.1 Simulation studies

Performance of our proposed approach was first evaluated using series of simulation studies. In an attempt to obtain sampling distributions of parameters that would approximate biological distributions, unstratified ZIBR models were fit to each OTU in the pancreatic cancer dataset (see Methods for details of this dataset). Unstratified parameter estimates were then used to obtain smooth sampling distributions of ω, p, ϕ. Finally, these sampling distributions were used to generate many pseudo-datasets satisfying H_1 and performance was evaluated when applying the previously described testing approach to the simulated dataset.

Sampling distributions for parameters were similar for both genus and ASV level. However, for ω, the mean non-zero relative abundance, distributions tended to be slightly further concentrated toward 0.0 on the ASV level as compared to the genus level. Further, distributions of p tended to be slightly more concentrated toward 1.0 on the ASV level as compared to the genus level. In both cases a linear relationship was observed between $\log \omega$ and $\log \phi$ which was ultimately used to sample ϕ conditionally on ω (Figure 3).

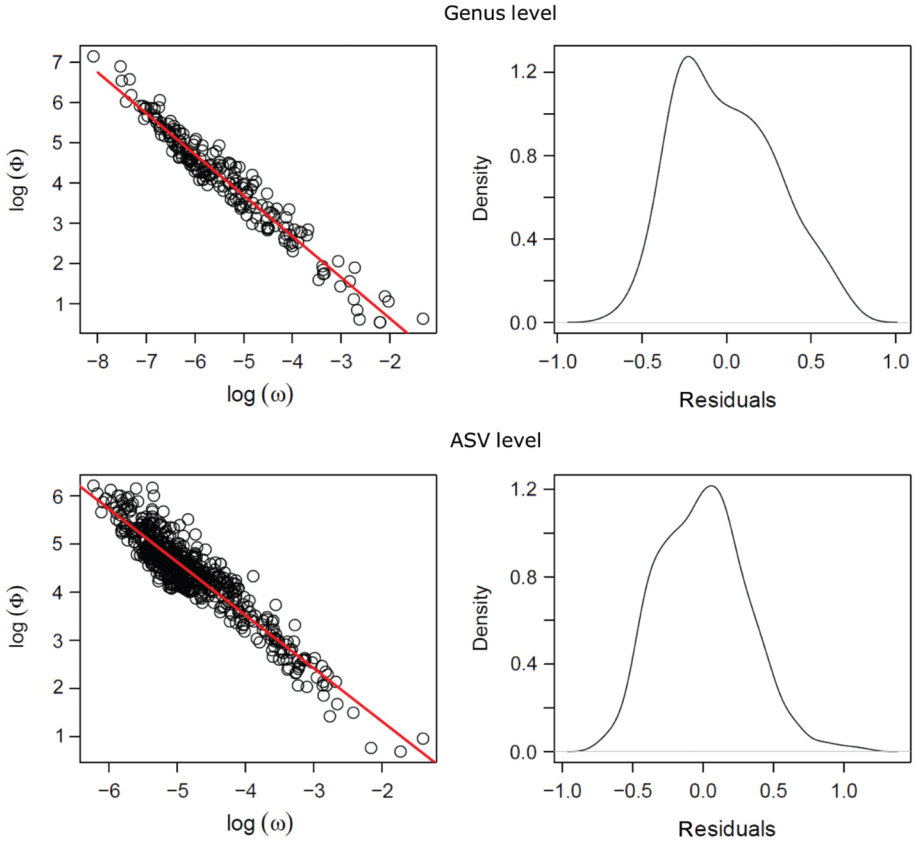

Figure 3: Observed relationships between marginal distributions of ω and ϕ estimated from the pancreatic cancer dataset. For both the genus and the ASV level, parameters were estimated marginally for each OTU across all observations without any stratification. When plotting marginal parameter estimates of ω and ϕ a linear relationship can be observed on the log scale. This relationship was utilized to sample ϕ conditionally on ω in the simulation studies.

In summary, the following sampling distributions were obtained:

$$\begin{aligned}
\text{Genus: } & p \sim Beta(1.67, 0.4); \quad \omega \sim Beta(0.63, 53.27); \\
& \log \phi | \min_{sg}\{\log \omega_{sg}\}) \sim N(-1.02 \min_{sg}\{\log \omega_{sg}\} - 1.41, 0.3^2) \\
\text{ASV: } & p \sim Beta(7.35, 0.49); \quad \omega \sim Beta(1.46, 121.12); \\
& \log \phi | \min_{sg}\{\log \omega_{sg}\}) \sim N(-1.10 \min_{sg}\{\log \omega_{sg}\} - 0.89, 0.31^2)
\end{aligned}$$

As expected, simulations of biological data revealed that analyses on the genus level were overall more powerful than on the ASV level, regardless of which population parameter was investigated (Figure 4). Assuming $t_c = 0$, four disease status groups, 95% credible intervals and utilizing Pearson correlation, the highest power was achieved when testing PASTA of ω. Under a moderate degree of association of $T_\theta = 0.537$ a target power of 0.8 was reached for 5 samples per stratum on the genus level and 15 samples per stratum on the ASV level. Type 1 error rates appeared adequately calibrated to the 5% significance level ranging from 0.03 to 0.056 on the genus level and from 0.032 to 0.06 on the ASV level. Despite the relatively modest within group sample size needed to detect a moderate degree of association for ω with adequate statistical power, there appeared to be considerably less power for tests of p. Under a high degree of association of $T_\theta = 0.834$ a target power of 0.8 was reached for 40 samples per stratum on the genus level. On the ASV level, utilizing as many as 80 samples per stratum resulted in a power of only 0.59 for the same T_θ. Type 1 error rates also appeared mostly calibrated in this scenario, but showed deflatation for smaller sample sizes, assuming a value of 0.027 on the genus level and 0.009 on the ASV level.

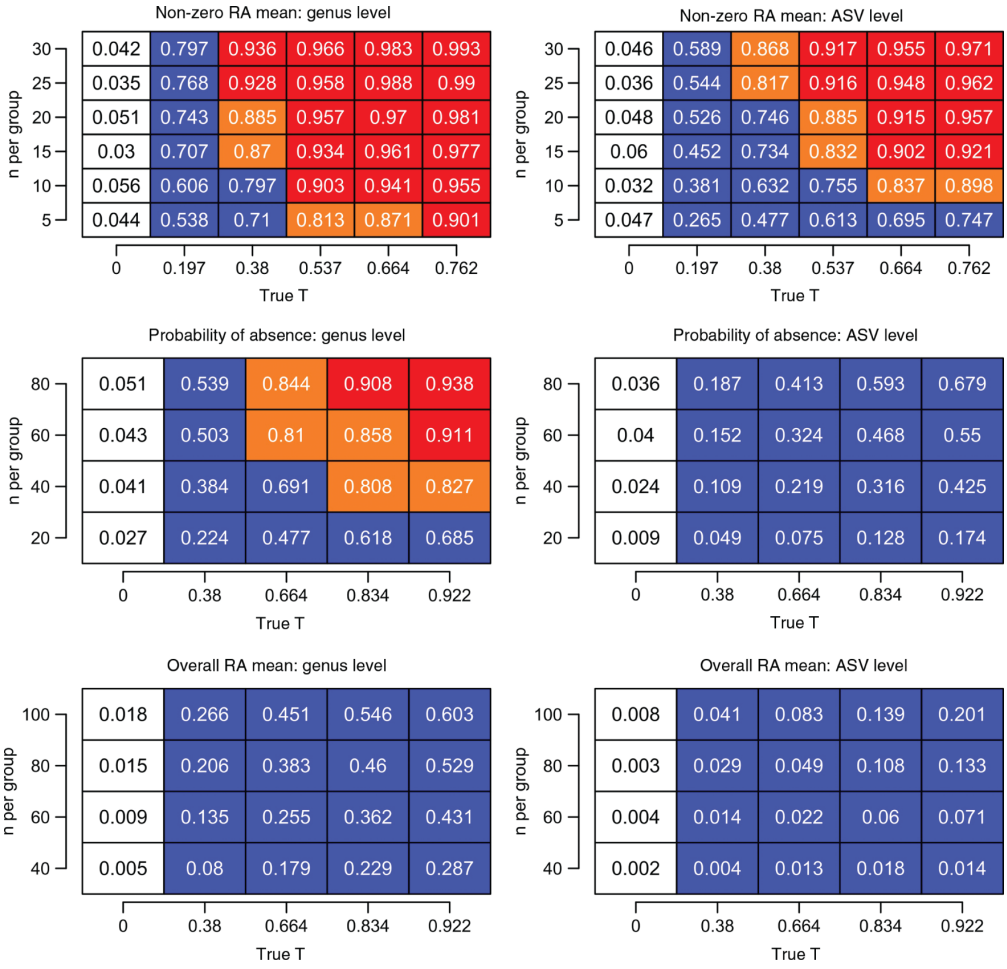

Figure 4: Results of the simulation studies. Power plots are displayed for testing PASTA of various population parameters with $t_c = 0$ at both ASV and genus level. The term "n per group" refers to the number of samples available in each of the eight sub-group combinations resulting from two body sites and four different levels of disease status. H_0 was rejected if $Pr(T_\theta|\mathbf{Y} \leq 0) < 0.05$. Type 1 error rates are displayed in white colored boxes with black fonts. Power values less than 0.8 are colored blue, values larger than 0.9 are colored red and values between 0.8 and 0.9 are colored orange. Genus level pseudo data generally has higher statistical power than the ASV level. High performance is achieved by the non-zero mean ω, while an increased sample size is required for the probability of absence p. Tests of the overall mean μ result in low performance, when only mildly constraining sparsity.

Testing PASTA of the overall mean $\mu = \omega(1-p)$ was also investigated. While improving with increasing size of effect and sample size, the power for this parameter was lower than when considering ω, p individually. Even when considering the large degree of association $T_\theta = 0.834$ and using 100 samples per stratum, the genus level scenario achieved a power of only 0.546. Notably, type 1 error rates were consistently deflated, ranging from 0.005 to 0.018 on the genus level and 0.002–0.008 on the ASV level. Type 1 error rates were deflated across all simulated scenarios, reaching values of less than or equal 0.018 or less.

Discrepancies in performance were found to be directly related to precision of parameter estimates. When plotting the posterior means of T_θ against their true simulated values across various simulation runs, the variation around the identity line consistently increased from ω to p, aswell as from genus to ASV level (Figure 5). Analogously, posterior distributions of T_θ were found to on average become more diffuse and more biased towards 0, when moving from ω to p or from genus to ASV level. When performing simulation runs of a scenario with low relative precision, in which p was sampled from a $Uniform(0.85, 0.95)$ distribution, the posterior distribution of T_θ was on average almost perfectly centered at zero and highly diffuse.

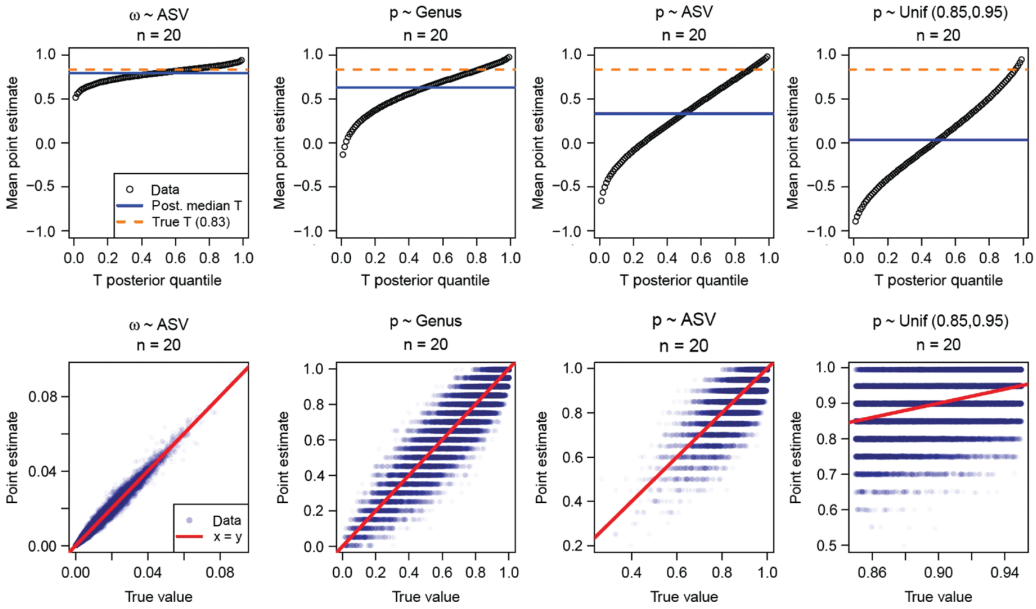

Figure 5: Effects of the relative precision of parameter estimates on the posterior distribution of T_θ. The first row shows the average point estimate of posterior quantiles of T_θ across simulation runs for various simulation scenarios. The second row shows the associated plots of the parameters' posterior means versus their true values across simulation runs. As the relative precision of parameter estimates decreases, the posterior distribution of T_θ becomes more diffuse and more biased towards 0.

Poor power when testing μ was also found to be related to two additional factors. Detailed results of simulations accounting for these factors are displayed in Additional File 1. The deflated type 1 error rates when utilizing 95% credible intervals, lead to overly conservative tests that negatively affected power. Calibrating type 1 errors to 5% by adjusting lower bounds of the credible intervals of T_θ for each considered sample size, lead to a consistent improvement in power, reaching a value of 0.73 for $T_\theta = 0.922$ and 80 samples per stratum on the genus level. The second factor that affected performance was the employed liberal three sub-strata rule, which allowed up to three strata to exhibit exclusively zeroes. Since high rates of absence were simulated, this case often naturally occurred leading to the three respective posterior estimates being imputed with the vague prior distribution, which is very imprecise. A follow-up simulation restricting all strata to have at least one non-zero observation, lead to a consistent increase in power, reaching a value of 0.82 for $T_\theta = 0.922$ and 80 samples per stratum on the genus level. In both settings, overall performance for testing μ was consistently lower than for testing p regardless of taxonomic levels. When both calibrating type 1 error and restricting non-zero observations at the same time, power increased further but was not consistently better than for testing p.

Additionally, three sets of simplified supplementary simulations were also performed to showcase how the PASTA testing approach can be analogously utilized in other data models. A simple Poisson regression model and a log ratio Aitchison model both achieved performance metrics slightly less performant but overall comparable to testing PASTA of the mean non-zero relative abundance via Beta regression. In particular, the Beta regression model appeared to achieve higher power for small sample sizes than the other two approaches and the log-ratio Aitchison model appeared to perform slightly worse than the Poisson regression model. On the other hand, testing PASTA of the overall mean in a zero-inflated Poisson model, utilizing the same smooth sampling distribution of zero-inflation rate p as for the ZIBR model on the genus level, achieved performance metrics comparable to testing PASTA of the overall mean in the ZIBR model. Even though minor differences with respect to calibration of type 1 error and statistical power were observed across the different models, the testing approach was overall viable regardless of the scenario. A detailed summary of these simulations is provided in Additional File 2.

3.2 Applying the approach to biological data

The 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), is a systematic classification of medical conditions provided by the World Health organization. Disease status information in this dataset was available via ICD-10 codes for each subject. The clinical pancreatic cancer dataset contained four predominant cancer-types: C24.x, C25.x, K86.2 and other, where ".x" denotes a further sub-type that could differ by subject and "other" refers to pancreatic cancer in various other categories or other diseases of the foregut. OTUs exhibiting significant PASTA with respect to disease status were successfully identified for both genus and ASV level in this dataset. For analysis we considered coding cancer sub-type in two ways: four group coding as described above and three group coding, which collapsed "K86.2" and "other" into one group. On the genus level, when coding disease status into four groups three genera exhibiting PASTA between mouth and gut were identified: *Fusobacterium*, *Haemophilus* and *Veillonella* (Table 2). After substratifying oral sites into saliva, tongue, buccal and gum, these association were found to be preserved for some of the site pairs: *Fusobacterium* also exhibited PASTA between gut an saliva sites; *Haemophilus* also exhibited PASTA between gut and gum, as well as gut and tongue; *Veillonella* also exhibited PASTA between gut and gum. Several genera also exhibited PASTA between individual mouth sites (Table 3). Two genera exhibited PASTA between four pairs of mouth sites: *Fusobacterium* and *Actinomyces*. Three genera exhibited PASTA between two pairs of mouth sites: *Atopobium*, *Haemophilus* and *Prevotella*. Six genera exhibited PASTA in only one pair of mouth sites.

Table 2: Genus level OTUs showing evidence of PASTA between gut and mouth sites when dividing ICD10 code into four groups.

Genus	Gut & mouth (all)	Gut & buccal	Gut & gum	Gut & saliva	Gut & tongue
Fusobacterium	$\mu^{\cdot\cdot}, p^{\cdot\cdot}$	–	–	$\mu^{\cdot\cdot}$	–
Haemophilus	p^{*}	–	μ^{*}, p^{*}	–	$p^{\cdot\cdot}$
TM7-G1	–	–	–	–	$p^{\cdot\cdot}$
Veillonella	$p^{\cdot\cdot}$	–	$p^{\cdot\cdot}$	–	–

For a given genus, a parameter is included in this table if it was marginally significant, or when significance is achieved when T is either Pearson or Spearman correlation. For a given population parameter θ, marginal significance ($Pr(T|\mathrm{Y} \leq 0) < 0.1$) is denoted by $\theta^{\cdot\cdot}$ and significance ($Pr(T|\mathrm{Y} \leq 0) < 0.05$) is denoted by θ^{*}. Three parameters were investigated: μ, ω, p. Due to low power in this exploratory setting multiple testing was not adjusted for.

Table 3: Genus level OTUs showing evidence of PASTA between mouth sites when dividing ICD10 code into four groups.

Genus	Buccal & gum	Buccal & saliva	Buccal & tongue	Gum & saliva	Gum & tongue	Gum & tongue
Actinomyces	–	$\omega^{\cdot\cdot}$	ω^{*}	$\mu^{\cdot\cdot}, p^{*}$	–	μ^{*}
Atopobium	–	–	–	–	p^{*}	$p^{\cdot\cdot}$
Fusobacterium	p^{*}	$\omega^{\cdot\cdot}, p^{\cdot\cdot}$	$\mu^{\cdot\cdot}$	p^{*}	–	–
Haemophilus	–	$\omega^{\cdot\cdot}$	–	–	–	$\mu^{\cdot\cdot}, \omega^{\cdot\cdot}$
Prevotella	–	$\mu^{\cdot\cdot}, \omega^{\cdot\cdot}$	–	–	–	$\mu^{\cdot\cdot}$

For a given genus, a parameter is included in this table if it was marginally significant, or when significance is achieved when T is either Pearson or Spearman correlation. For a given population parameter θ, marginal significance ($Pr(T|\mathrm{Y} \leq 0) < 0.1$) is denoted by $\theta^{\cdot\cdot}$ and significance ($Pr(T|\mathrm{Y} \leq 0) < 0.05$) is denoted by θ^{*}. Three parameters were investigated: μ, ω, p. Due to low power in this exploratory setting multiple testing was not adjusted for. Six OTUs showing association for only one pair of mouth sites are not shown in this Table.

On the ASV level, two ASVs exhibited PASTA with respect to p between mouth and gut when disease status was coded into four groups. When coding disease status into three groups, the same two ASVs as before and three additional ASVs exhibited PASTA with respect to p between mouth and gut. Notably, among these additional ASVs was a candidate belonging to the *Fusobacterium* genus. Further details of the ASV level analysis are discussed in Chung et al. (2019).

4 Discussion

The methodology presented in this publication successfully establishes a general framework to test for pairwise stratified association (PASTA) in microbial abundance or relative abundance. The approach first estimates posterior distributions of population parameters $\boldsymbol{\theta}|\mathbf{Y}$ within the strata of body site and disease status and subsequently calculates a correlation statistic T_θ between body sites, which scores their degree of association. This allows researchers to identify individual microbes or groups of microbial species that show consistent abundance patterns between different body sites with respect to the disease status of patients or any other relevant categorical grouping variable.

While this work focuses on identifying preserved patterns between body sites, anti-correlated relationships, where an increase in one body site corresponds to a decrease in another body site, may also be also of biological interest. Such associations are represented by a decreasing functional relationship between the two body sites. Our approach can also be used to identify these relationships by flipping the inequalities in H_0 and H_1 and rejecting the null when $Pr(T_\theta|\mathbf{Y} \geq t_c) < \alpha$. If either correlated or anti-correlated relationships are to be identified a two-sided test can be analogously formularized testing $H_0 : -t_c \leq T_\theta \leq t_c$.

It has to be noted that while many possible T may be adequate to detect a wide variety of increasing relationships h, the choice of T can favour certain shapes of h. If the Pearson correlation is employed, linear relationships will achieve higher scores than rapid, exponential growth relationships, since it measures the degree of linear association. In this case, overly large values of t_c (for example $t_c = 0.8$) should be avoided as they may lead to falsely rejecting non-linear, increasing relationships. While rank-correlation measures such as the Spearman correlation may be more generally applicable, they may also be less powerful, especially when few groups are considered ($g < 5$). In cases with a small number of groups, the discrete nature of the rank-correlation statistic is more pronounced. When utilizing Spearman correlation it is helpful to keep in mind that T_θ can only assume 4 discrete values when $g = 3$, 11 discrete values when $g = 4$ and 21 possible values when $g = 5$.

Care should also be exercised when interpreting significant associations. The test for PASTA is concerned with trend, agreement or association between $s = 1$ and $s = 2$ after stratification according to g, but does not at all provide information on whether the effect of site s or disease status g is biologically or clinically significant. To the contrary, it assumes that both grouping variables are inherently meaningful objects of the research hypothesis. For example, if there is no significant effect of body site (i.e. abundance is the same between mouth and gut), but abundance differs by disease status, the test statistic will likely score a high degree of association, because what is going on in one site is still associated with what is going on in the other site and this is an inherently meaningful relationship to us. However, the contrary where effect of body site is significant (i.e. abundance is different between mouth and gut) but effect of disease status is not, will not necessarily lead to a significant score of association. Scenarios are possible in which there are small effects of body site and disease status, where none are strong enough to reach statistical significance, yet the test for association may still be overall significant, as long as the trend across strata is pronounced enough. To understand the specific nature of an identified PASTA relationship, it can be useful to plot credible intervals of parameter estimates θ_{sg} side-by-side (Additional File 3) or to perform statistical follow-up tests investigating the effects of s and g. In order to reduce the burden of multiple testing and to increase the likelihood of screening for impactful associations, a researcher may also choose to first perform marginal tests confirming whether each microbe exhibits significant (or marginally significant) differences with the phenotype of interest within the gut. The restricted set of microbes exhibiting such significant differences could then be used to test for PASTA.

It should be noted that in cases where more information about h is known in advance, more powerful tests of association could be designed that leverage this information. If, for example, h was known to be linear, then the following model could be fit: $\theta_{2g} = \alpha + \theta_{1g} \cdot \beta$, which drastically reduces the number of parameters. In this setting, a PASTA test would be reduced to significance of the parameters α, β. Whilst being more powerful, such a model would also allow one to learn the relationship between mouth and gut, which could be leveraged for predicting

gut samples via mouth samples of newly observed subjects. Knowledge about the correlation structure among strata and between OTUs could also potentially be incorporated by utilizing Bayesian hierarchical prediction models with shared hyperpriors. Such sophisticated models may further provide the opportunity to increase power and more adequately reflect knowledge about the data. However, the benefit of our current approach is its general applicability and lack of assumptions about h or correlation structure in the data. Little is currently known about the form of relationships between microbes in different organs or tissues. It is therefore more important to be able to identify cases in which a relationship is present as opposed to fully characterizing the relationship. Without prior knowledge the choice of h is arbitrary and researchers run the risk of potentially missing associations that do not conform with this choice. A researcher can first use our approach to identify microbes exhibiting promising associations, then look at point estimates and credible intervals of parameters across strata to learn about the shape of h. This may then motivate building a prediction that is grounded in empirical evidence. Another benefit of an adequately chosen multivariate Bayesian hierarchical model is that it allows one to test whether OTU A in mouth associates with OTU B in gut. While such a model has the potential to provide a more powerful test, the here proposed approach does allow one to identify this type of association by including the response values from OTU A in oral samples and the response values of OTU B in gut samples into the model and conducting the test analogously. However, if such a strategy was employed, ϕ may have to be estimated individually per body site, as the assumption of constant dispersion is likely to not hold between different OTUs.

Results of the simulation studies reveal that the testing procedure is able to successfully identify PASTA patterns. The decreased performance on the ASV level can be attributed to decreased signal intensities and the overall increase in sparsity of non-zero observations. The substantial drop in performance when investigating PASTA of p was demonstrated to be a result of overall lower precision in estimation, compared to ω. Since probabilities of absence are generally high and concentrated towards 1.0 across OTUs and strata, the differences between them are often small. In this scenario, to be able to reliably quantify differences and assess trends with adequate precision, larger sample sizes are required. This problem is thus a limitation of the zero-inflated data and not the testing approach itself.

Investigating the overall mean μ, may not always be viable when utilizing the ZIBR model. Since its estimation is based on estimates of both p and ω, its estimates are subject to more sources of variation, resulting in poorer precision and lower power. Our simulation suggests that if the properties of the population that is to be analyzed are well known, adjusting the quantile $t_{Q\alpha}$ to calibrate type 1 error rates is a viable strategy to improve performance. If this was not the case and a researcher was convinced that inference based on μ was more biologically meaningful than considering the individual components p and ω, alternative models may be considered. For relative abundance data an adequate choice may be the marginalized ZIBR model as proposed by Chai et al. (2018) which directly estimates μ as a function of covariates. These estimates could then be used analogously to test for PASTA relationships using the here proposed approach.

The supplementary simulations provided in Additional File 2 should also be interpreted with caution. While their results do provide information about general viability of our testing approach in the respective scenario, they may not be suitable to infer superiority of either modeling approach. Direct comparison of the models based on the simulation scenarios could be biased, since in each scenario pseudo-data was generated differently and the number of estimated parameters also differed between models.

The fact that in the pancreatic cancer patient dataset OTUs can be identified that show associations between mouth and gut, as well as between individual oral sites suggests that they may be promising candidates for potential biomarkers. Among these were *Fusobacterium* and *Haemophilus*, both oral bacteria recently found to distinguish pancreatic head carcinoma patients from healthy subjects (Lu et al., 2019). Also, species belonging to the genera *Fusobacterium* and *Prevotella* (even though the latter was only found to show association between mouth and gut) have been shown to associate with periodontal disease (Chiranjeevi et al., 2014; Chen et al., 2018). These results lend further credence to the disease related connection between microbial abundance in mouth and gut and suggests that our method leads to conclusions consistent with the literature. More future research will be needed to validate these findings.

The simulation studies also confirmed that tests of PASTA applied to the pancreatic cancer patient dataset are likely underpowered due to the limited sample size. It should be noted that many OTUs could not be tested due to too high zero-inflation and thus insufficient signal. These two factors likely explain why relatively few candidates were identified when conducting the tests. Future studies may consider larger sample sizes or aim to improve the yield of observed counts in each sample to alleviate this issue. Our results suggest that differences in the extent of zero-inflation between groups may be generally hard to detect for small to medium sized studies when more granular phylogenetic levels are targeted.

5 Conclusions

In conclusion, the performed simulation studies demonstrate the viability of the approach in the context of ZIBR models and suggest that for tests of association of mean non-zero relative abundance modest sample sizes can achieve adequate power for moderate degree of association. The simulations also highlight potential lack of power for low-level phylogeny data (e.g. species, ASV) or when more complex functions of population parameters are considered. When analyzing a biological dataset consisting of pancreatic cancer patients the approach is able to identify microbes that exhibit PASTA patterns and are consistent with independent findings of current research studies. The generality of this approach allows it to be extended to other data models and research settings, ensuring that it can be useful for researchers interested in stratified associations in the microbiome world and beyond.

Acknowledgement

We would like to extend our gratitude to Dr. Dong Pei, Lisa Neums, Stefan Graw, Qing Xia, Bo Zhang, Rosalyn Henn and Duncan Rotich of the Department of Biostatistics & Data Science at the University of Kansas Medical Center for their constructive feedback on the methodology.

Funding

Research reported in this publication was supported by NIH/National Cancer Institute grants R01 CA166150 and funder id: http://dx.doi.org/10.13039/100000054, P30 CA168524 as well as the the Kansas IDeA Network of Biomedical Research Excellence Bioinformatics Core, supported in part by the National Institute of General Medical Science, funder id: http://dx.doi.org/10.13039/100000057 award P20GM103428.

Declarations

Ethics approval and consent to participate: The study was approved by Lifespan's Research Protection Office for recruitment at RIH, as well as the Institutional Review Boards for Human Subjects Research at Brown University, Tufts University, and the Forsyth Institute.

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. The datasets analyzed for this manuscript will also be available through the NCBI under the BioProject accession no.: PRJNA421501 accompanying the publication of Chung et al. (2019).

R scripts utilized to perform the simulation studies and to analyze the pancreatic cancer dataset are available via the GitHub directory provided in the reference section (Meier 2019). An example script showing the analysis of a simple pseudo-dataset is also available in the same directory.

Competing interests: K.T. Kelsey is a consultant/advisory board member at Celintec. No potential conflicts of interest were disclosed by the other authors.

References

Callahan, B. J., P. J. McMurdie and S. P. Holmes (2017): "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," ISME J., 11, 2639–2643.

Chai, H., H. Jiang, L. Lin and L. Liu (2018): "A marginalized two-part beta regression model for microbiome compositional data," PLoS Comput. Biol., 14, e1006329.

Chen, E. Z. and H. Li (2016): "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data," Bioinformatics, 32, 2611–2617.

Chen, C., C. Hemme, J. Beleno, Z. J. Shi, D. Ning, Y. Qin, Q. Tu, M. Jorgensen, Z. He, L. Wu and J. Zhou (2018): "Oral microbiota of periodontal health and disease and their changes after nonsurgical periodontal therapy," ISME J., 12, 1210–1224.

Chiranjeevi, T., O. H. Prasad, U. Prasad, A. K. Kumar, V. Chakravarthi, P. B. Rao, P. Sarma, N. Reddy and M. Bhaskar (2014): "Identification of microbial pathogens in periodontal disease and diabetic patients of south indian population," Bioinformation, 10, 241–244.

Chung, M., N. Zhao, R. Meier, D. C. Koestler, G. Wu, E. D. Castillo, B. J. Paster, K. T. Kelsey and D. S. Michaud (2019): "Oral, gut, and pancreatic microbiome are correlated and exhibit consist co-abundance in patients with pancreatic diseases and cancer," [Manuscript in Progress].

Cohen, J. (1960): "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., 20, 37–46.
del Castillo, E., R. Meier, M. Chung, D. C. Koestler, T. Chen, B. J. Paster, K. P. Charpentier, K. T. Kelsey, J. Izard and D. S. Michaud (2019): "The microbiomes of pancreatic and duodenum tissue overlap and are highly subject specific but differ between pancreatic cancer and non-cancer subjects," Cancer Epidemiol. Biomark. Prev., 28, 370–383.
Faith, J. J., J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, M. Rosenbaum and J. I. Gordon (2013): "The long-term stability of the human gut microbiota," Science, 341, 1237439.
Fan, X., A. V. Alekseyenko, J. Wu, B. A. Peters, E. J. Jacobs, S. M. Gapstur, M. P. Purdue, C. C. Abnet, R. Stolzenberg-Solomon, G. Miller, J. Ravel, R. B. Hayes and J. Ahn (2016): "Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study," Gut, 67, 120–127.
Fardini, Y., X. Wang, S. Témoin, S. Nithianantham, D. Lee, M. Shoham and Y. W. Han (2011): "Fusobacterium nucleatum adhesin FadA binds vascular endothelial cadherin and alters endothelial integrity," Mol. Microbiol., 82, 1468–1480.
Fleiss, J. L. (1971): "Measuring nominal scale agreement among many raters," Psychol. Bull., 76, 378–382.
Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn and J. J. Egozcue (2017): "Microbiome datasets are compositional: and this is not optional," Front. Microbiol., 8, 2224.
Goodman, B. and H. Gardner (2018): "The microbiome and cancer," J. Pathol., 244, 667–676.
Lu, H., Z. Ren, A. Li, J. Li, S. Xu, H. Zhang, J. Jiang, J. Yang, Q. Luo, K. Zhou, S. Zheng and L. Li (2019): "Tongue coating microbiome data distinguish patients with pancreatic head cancer from healthy controls," J. Oral Microbiol., 11, 1563409.
Meier, R. (2019): "R scripts for simulation studies, analyses and examples related to pasta," https://github.com/richard-meier/PASTA_scripts. [Online; accessed 08-April-2019].
Michaud, D. S., J. Izard, C. S. Wilhelm-Benartzi, D.-H. You, V. A. Grote, A. Tjønneland, C. C. Dahm, K. Overvad, M. Jenab, V. Fedirko, M. C. Boutron-Ruault, F. Clavel-Chapelon, A. Racine, R. Kaaks, H. Boeing, J. Foerster, A. Trichopoulou, P. Lagiou, D. Trichopoulos, C. Sacerdote, S. Sieri, D. Palli, R. Tumino, S. Panico, P. D. Siersema, P. H. Peeters, E. Lund, A. Barricarte, J.-M. Huerta, E. Molina-Montes, M. Dorronsoro, J. R. Quirós, E. J. Duell, W. Ye, M. Sund, B. Lindkvist, D. Johansen, K.-T. Khaw, N. Wareham, R. C. Travis, P. Vineis, H. B. B. de Mesquita and E. Riboli (2012): "Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large european prospective cohort study," Gut, 62, 1764–1770.
Peng, X., G. Li and Z. Liu (2016): "Zero-inflated beta regression for differential abundance analysis with metagenomics data," J. Comput. Biol., 23, 102–110.
Pushalkar, S., M. Hundeyin, D. Daley, C. P. Zambirinis, E. Kurz, A. Mishra, N. Mohan, B. Aykut, M. Usyk, L. E. Torres, G. Werba, K. Zhang, Y. Guo, Q. Li, N. Akkad, S. Lall, B. Wadowski, J. Gutierrez, J. A. K. Rossi, J. W. Herzog, B. Diskin, A. Torres-Hernandez, J. Leinwand, W. Wang, P. S. Taunk, S. Savadkar, M. Janal, A. Saxena, X. Li, D. Cohen, R. B. Sartor, D. Saxena and G. Miller (2018): "The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression," Cancer Discov., 8, 403–416.
Schober, P., C. Boer and L. A. Schwarte (2018): "Correlation coefficients," Anesth. Analg., 126, 1763–1768.
Shi, P., A. Zhang and H. Li (2016): "Regression analysis for microbiome compositional data," Ann. Appl. Stat., 10, 1019–1040.
Tsilimigras, M. C. and A. A. Fodor (2016): "Compositional data analysis of the microbiome: fundamentals, tools, and challenges," Ann. Epidemiol., 26, 330–335.
Xia, Y., J. Sun and D.-G. Chen (2018): "Modeling zero-inflated microbiome data." In: Jiahuan Chen, (ed.), Statistical Analysis of Microbiome Data with R. Springer, Singapore. pp. 453–496.
Zhang, X., H. Mallick, Z. Tang, L. Zhang, X. Cui, A. K. Benson and N. Yi (2017): "Negative binomial mixed models for analyzing microbiome count data," BMC Bioinform., 18:4.

Supplementary Material: The online version of this article offers supplementary material (DOI: https://doi.org/10.1515/sagmb-2019-0027).