Research Article

Farhad Hatami, Emma Beamish, Albert Davies, Rachael Rigby, and Frank Dondelinger*

A Hierarchical Bayesian Approach for Detecting Global Microbiome Associations

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: The human gut microbiome has been shown to be associated with a variety of human diseases, including cancer, metabolic conditions and inflammatory bowel disease. Current approaches for detecting microbiome associations are limited by relying on specific measures of ecological distance, or only allowing for the detection of associations with individual bacterial species, rather than the whole microbiome. In this work, we develop a novel hierarchical Bayesian model for detecting global microbiome associations. Our method is not dependent on a choice of distance measure, and is able to incorporate phylogenetic information about microbial species. We perform extensive simulation studies and show that our method allows for consistent estimation of global microbiome effects.
Additionally, we investigate the performance of the model on two real-world microbiome studies: a study of microbiome-metabolome associations in inflammatory bowel disease, and a study of associations between diet and the gut microbiome in mice. We show that we can use the method to reliably detect associations in real-world datasets with varying numbers of samples and covariates.

Keywords: Microbiome, global effects, Bayesian modeling

1 Introduction

Recent years have seen an explosion in the amount of genomic sequencing data collected for a variety of organisms. One area of particular interest is the nascent field of metagenomics. Metagenomics is concerned with the study of genetic material that can be found in samples from diverse environments ranging from soil and water to body cavities (Wooley et al., 2010). The collection of bacterial genetic material identified in a sample is called the microbiome. For human health, the human gut microbiome has been shown to be associated with a variety of conditions, including colorectal cancer (Ahn et al., 2013), metabolic diseases (Le Chatelier et al., 2013) and inflammatory bowel disease (Halfvarson et al., 2017).

The recent glut of microbiome studies has led to increased interest in developing robust and efficient methods for analysing this type of data. A typical microbiome study will first sequence the bacterial genetic material contained in a sample, and will then use the genetic information to identify and quantify the organisms that the material originated from. In the case of the microbiome, the latter is most commonly done by using the 16S ribosomal RNA genes, which allow for the identification of bacterial species that colonize the human gut. Each fragment of 16S genetic material in the sample can then be classified by comparing it to known reference genomes, and species can be differentiated from each other by looking at genetic similarity. The result will be a count matrix quantifying the number of times genetic material from each species was detected.

Farhad Hatami, Frank Dondelinger, Lancaster University, Centre for Health Informatics Computation and Statistics, Lancaster, UK, e-mail address of first and forth authors: farhad.hatami@bdi.ox.ac.uk, fdondelinger.work@gmail.com
Emma Beamish, University of Liverpool, Department of Musculoskeletal Biology, Liverpool, UK, e-mail: e.beamish@liverpool.ac.uk
Albert Davies, Furness General Hospital, Barrow-In-Furness, UK, e-mail: Albert.Davies@mbht.nhs.uk
Rachael Rigby, Lancaster University, Department of Biomedicine and Life Sciences, Lancaster, UK, e-mail: rachael.rigby@lancaster.ac.uk

Previous approaches to analysing microbiome data can be classified in two main categories: distance-based methods and regression-based methods. In distance-based methods we define a distance metric between observed microbiome samples. Common measures include Jaccard (dis)similarity, Bray-Curtis dissimilarity and the UniFrac distance (Lozupone and Knight, 2005). The latter has the advantage that it takes into account the phylogeny, or genetic relatedness, of the species in the samples when calculating the distance; loosely speaking, samples with related species will be closer than samples with unrelated species. One can then use a statistical test such as the PERMANOVA to test whether the centroids and dispersion of samples within two or more groups are identical (see e.g. Chen et al., 2012).

Regression-based methods tend to treat the count data as following either a negative binomial (Zhang et al., 2017) or Dirichlet-multinomial distribution (Chen and Li, 2013). In the former, the effects of observed covariates on species are usually treated as independent, although variance parameters may be shared across species. In the latter, the covariance across species is defined by the Dirichlet-multinomial distribution, but the regression will still define individual parameters representing the association between each species and the observed covariate. Unlike distance-based methods, regression methods do not try to infer a global effect of a covariate on the microbiome. However, they are better able to deal with multiple continuous covariates, and are not sensitive to a choice of distance measure.

In our work, we combine the advantages of the regression and distance-based approaches. We developed a Bayesian hierarchical log-ratio regression model, where we model multiple species as multiple outcomes. Unlike most standard regression approaches for microbiome data, we do not model the species counts directly, but rather treat the data as proportional and apply the log-ratio transform (Aitchison, 1982) to obtain multivariate-Gaussian distributed pseudo-observations. This allows us to bring in the phylogenetic information via the covariance matrix. Global associations between the microbiome and the covariates are estimated using a shared hyper-prior on the species-specific effect estimates. Efficient gradient-based estimation of the discrete parameters in the model is achieved via the Gumbel-Softmax (Jang et al., 2016; Maddison et al., 2016). We demonstrate the robustness and efficiency of our method using an extensive simulation study, and apply it to case studies in inflammatory bowel disease and mouse gut microbiome analysis.

2 Methods

2.1 Log Ratio Model

Let us assume that we are collecting microbiome data from N subjects, and in each sample we are counting the occurrence of S species or taxons. We treat the microbiome data as compositional, i.e., even though we are observing count data, we only consider the proportion of each species. If we observe an S-dimensional count vector $\mathbf{z}_i$ for sample i, then the proportion of species s can be estimated as $\rho_{i,s} = \frac{z_{i,s}}{\sum_{s'}^{S} z_{i,s'}}$. We can now transform the species proportions for species $s \in \{1, ..., S-1\}$ by using the log-ratio transform (Aitchison, 1982):

$$y_{i,s} = \log\left(\frac{\rho_{i,s}}{\rho_{i,S}}\right) \tag{1}$$

We assume that the $\mathbf{y}_i$ are approximately multivariate Gaussian distributed, and model them as:

$$\mathbf{y}_i \sim \mathbf{MVN}(\boldsymbol{\mu}_i, c\boldsymbol{\Sigma}) \tag{2}$$

In our application, $\boldsymbol{\Sigma}$ is informed by the phylogenetic tree inferred from the sequencing data. Briefly, under the assumption of a Brownian motion evolutionary model, the covariance of two species is equivalent

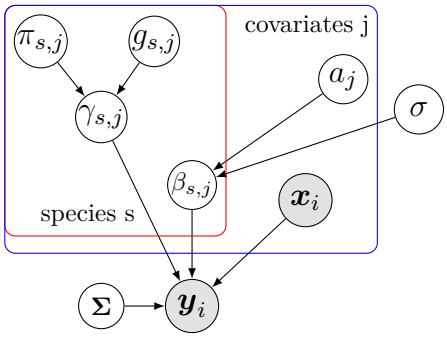


Fig. 1: Graphical representation of the hierarchical regression model for detecting microbiome associations. Shaded nodes are observed values, unshaded nodes are parameters. Here, $\mathbf{y}_i$ denotes the log-ratio transformed species abundances in sample i, $\boldsymbol{\Sigma}$ is the covariance matrix informed by the phylogenetic relationships, $\mathbf{x}_i$ denotes the observed covariates, and a_j denotes the global effect of covariate j on $\mathbf{y}_i$. Please see the text for the remaining parameters, as well as the detailed description of the model.

to the branch length from the root to the most recent common ancestor in the phyogenetic tree. If no phylogenetic tree is available, then we set $\boldsymbol{\Sigma} = \mathbf{I}$, the $(S-1) \times (S-1)$ identity matrix. The hyperparameter c, $c > 0$ is a scaling parameter controlling the desired fit to the data. For the data in this study we found that $c = 1$ led to satisfactory models that showed no evidence of over- or under-fitting.

2.2 Hierarchical regression model

Having established the log ratio model, we now want to formulate $\boldsymbol{\mu}_i$ as a linear expression with information sharing across the species (or response variables). Let:

$$\boldsymbol{\mu}_i = (\boldsymbol{\Gamma} \odot \mathbf{B})\mathbf{x}_i^T \tag{3}$$

where $\mathbf{x}_i$ is a vector of p observed covariates for subject i, $\mathbf{B}$ is an $(S-1) \times p$ matrix of positive linear coefficients and $\boldsymbol{\Gamma}$ is an $(S-1) \times p$ matrix with entries in $\{-1, 1\}$. Estimation of global effects is achieved via a prior on the elements of $\boldsymbol{\beta}_j$, the $S-1$-dimensional vectors corresponding to the columns in $\mathbf{B}$:

$$\beta_{s,j} \sim \mathcal{N}^+(a_j, \sigma^2) \quad \forall s \in \{1, ..., S-1\} \tag{4}$$

where a_j denotes the global effect of covariate j on all species. A graphical representation of the resulting hierarchical model can be seen in Figure 1. In eq. (4), $\mathcal{N}^+$ denotes the positive part of the Gaussian distribution truncated at zero. Note that the global effect only defines the prior for the strictly positive magnitude $\beta_{s,j}$ of the species-specific effects. Placing a prior on the signed effect size would not have the desired effect; since the size of each biological sample is fixed, an increase in some species must lead to a decrease in other species, and so the global effect size under such an alternative model would not be reflective of the total change across species.

We also need to define a distribution for the elements $\gamma_{s,j}$ of $\boldsymbol{\Gamma}$. The natural approach would be to treat them as binary variables following a Bernoulli distribution. However, this requires us to perform inference over discrete variables, which can be onerous, and does not lend itself well to gradient-based inference techniques. Instead, we opt for a Gumbel-Softmax approximation to the Bernoulli distribution (Jang et al., 2016; Maddison et al., 2016). The Gumbel-Softmax is a continuous relaxation of the categorical distribution. In general, for a categorical random variable h with class probabilities π_1, π_2, ..., π_K, we can represent h as:

$$\mathbf{h} = \text{one-hot}(\text{argmax}_k[g_k + \log \pi_k]) \tag{5}$$

where one-hot denotes the one-hot encoding of a categorical variable as a K-dimensional vector (hence the change from h to $\mathbf{h}$). The g_k are sampled from a Gumbel(0,1) distribution. Note that this representation is

exactly equivalent to a categorical distribution; to obtain the continuous relaxation we need to replace the argmax by a softmax function, which gives:

$$\delta_k = \frac{\exp\left((\log(\pi_k) + g_k)/\tau\right)}{\sum_{k'=1}^{K} \exp\left((\log(\pi_{k'}) + g_{k'})/\tau\right)} \tag{6}$$

for $k \in \{1, ..., K\}$, where $\boldsymbol{\delta} \in \Delta^{K-1}$, the $(K-1)$-dimensional simplex. Here τ is a tuning parameter that controls the smoothness, with $\tau = 0$ reverting to the discrete case. In our case, $K = 2$, which means that eq. (6) simplifies to:

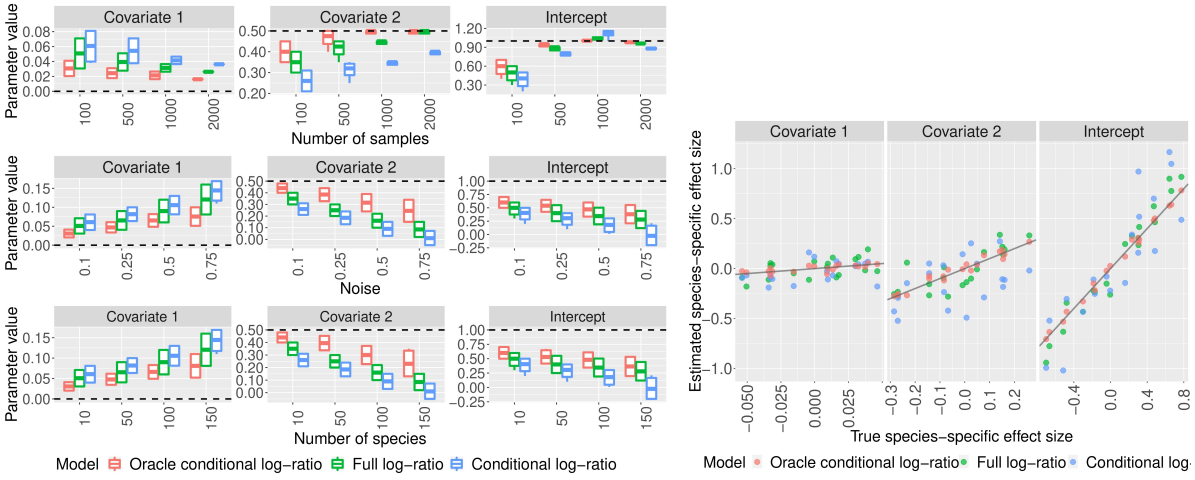
$$\delta_{s,j} = \frac{\exp\left((\log(\pi_1) + g_1)/\tau\right)}{\exp\left((\log(\pi_1) + g_1)/\tau\right) + \exp\left((\log(\pi_{-1}) + g_{-1})/\tau\right)} \tag{7}$$

where $\pi_{-1} = 1 - \pi_1$. We drop the subscript k as one value of δ fully defines the two-dimensional vector on the simplex, but introduce the subscript (s, j) to denote the value of δ for species s and covariate j. We use the subscripts 1 and -1 to denote Gumbel draws for $\gamma_{s,j} = 1$ and $\gamma_{s,j} = -1$ respectively. Since $\delta_{s,j}$ is either 0 or 1, we set:

$$\gamma_{s,j} = 2\delta_{s,j} - 1 \tag{8}$$

This has the desired property that if $\delta_{s,j} \approx 1$, $\gamma_{s,j} \approx 1$ and if $\delta_{s,j} \approx 0$, $\gamma_{s,j} \approx -1$.

We specify truncated Gaussian $\mathcal{N}^+(0, 1)$ priors for the global effects a_j, which also serves to regularize the model when the sample size is smaller than the number of parameters, which is the case for almost all the models studied here. We further place a Beta prior on the parameter π_1, where the hyperparameters can be used to encode any prior beliefs about the prevalence of positive or negative associations; for our study we set both equal to 1.



(a) (b)

Fig. 2: Performance comparison between the models. (a) Comparison of the performance of the full, conditional and oracle conditional log-ratio models on simulated data when varying the sample size (top row), noise variance of the species-specific effects (middle row) and number of species or taxonomic units (bottom row). The linear simulation model consists of the intercept term and two covariates with global effect sizes 1, 0.5 and 0 respectively. The dashed horizontal lines show the size of the true effect, and the boxplots represent the posterior mean estimates obtained from 20 independent simulated datasets. (b) Comparison of the performance of the full, conditional and oracle conditional log-ratio models on simulated data when estimating the species-specific effect size ($\beta_{s,j}$). Each point corresponds to the estimated effect size for one species in one model. The linear simulation model consists of the intercept term and two covariates with global effect sizes 1, 0.5 and 0 respectively. The solid lines represent $y = x$ (note that the x-axes are on different scales).

2.3 Inference

We are interested in performing fully Bayesian posterior inference over the parameters of this model. A convenient framework is Hamiltonian Monte Carlo (HMC), which allows us to take advantage of the differentiability of the Gumbel-Softmax to estimate gradients and speed up the Bayesian inference. We employ the Stan sampler to perform HMC sampling; for more details on the the Stan sampler and modelling language see (Carpenter et al., 2017).

In microbiome research, it is common to encounter datasets consisting of large numbers of operational taxonomic units (OTUs). Despite the efficiency of HMC sampling, sampling hundreds or thousands of parameters $\beta_{s,j}$ and $\gamma_{s,j}$ is not currently feasible on a personal computer. We therefore propose an approximate estimation, as an alternative to the full log-ratio model, where we only estimate $\beta_{s,j}$ as part of the sampling, and the $\gamma_{s,j}$ parameters are estimated in a separate inference step. The approximate inference procedure is as follows:

1. Fit independent linear models $y_{i,s} = \mathbf{x}_i^T \boldsymbol{\beta}_s^* + \epsilon_{i,s}$, $\epsilon_{i,s} \sim \mathcal{N}(0, \sigma_s^2)$ to the log-transformed microbiome counts for all species s. Note that $\boldsymbol{\beta}_s^*$ is distinct from the rows of $\mathbf{B}$ in the full model.
2. Set $\gamma_{s,j} = \text{sign}(\beta_{s,j}^*)$ where the sign() operation returns -1 if $\beta_{s,j}^*$ is negative and 1 otherwise.

We refer to the model with $\boldsymbol{\Gamma}$ inferred from the data as the full log-ratio model, and to the model with $\boldsymbol{\Gamma}$ fixed according to the two-step procedure above as the conditional model, since inference is conditional on the estimate of $\boldsymbol{\Gamma}$. We can think of the conditional model as an approximation to the full joint model $P(\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}) = P(\mathbf{B}, \boldsymbol{\theta} | \boldsymbol{\Gamma}, \mathbf{Y}, \mathbf{X}) P(\boldsymbol{\Gamma} | \mathbf{Y}, \mathbf{X})$, where $\boldsymbol{\theta}$ collects all parameters not in $\mathbf{B}$ or $\boldsymbol{\Gamma}$, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the log-ratio response and the design matrix with the observed covariates, respectively. Instead of estimating $P(\boldsymbol{\Gamma} | \mathbf{Y}, \mathbf{X})$, which would require marginalising over all other parameters, we use $\hat{\boldsymbol{\Gamma}} = \text{sign}(\hat{\boldsymbol{\beta}}^*)$ where $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_S^*)$ is the maximum likelihood estimate of the effect sizes in independent linear models of $P(\mathbf{y}_s | \mathbf{x}_s, \boldsymbol{\beta}_s^*)$ for each s. Inference is performed on the conditional model $P(\mathbf{B}, \boldsymbol{\theta} | \hat{\boldsymbol{\Gamma}}, \mathbf{Y}, \mathbf{X})$, which is less computationally expensive.

3 Simulation Study

3.1 Setup

We perform a simulation study under different scenarios to test the performance of the full and conditional log-ratio models. In particular, we study how well the full model can predict the global effect (a_j) and species-specific effect ($\beta_{s,j}$) parameters under scenarios with different sample size n, variance parameter σ^2, and number of species (or taxonomic units) S and compare the performance to that of the conditional model. In each experiment the variables that are not changing are set to $n = 100$, $\sigma^2 = 0.1$ and $S = 10$. The global effect sizes in the simulation model are set fixed with $a_j \in \{0, 0.5, 1\}$.

For each unique setting of the simulation parameters, we generate 20 random datasets according to the following procedure. First, a random phylogenetic tree is generated using the `ape` R package (Paradis and Schliep, 2018). Based on this tree, the covariance matrix $\boldsymbol{\Sigma}$ under the Brownian motion evolutionary model is calculated. Next, we generate an $N \times p$ matrix of covariates $\mathbf{X}$ with $x_{i,j} \sim \mathcal{N}(0, 1)$. For each covariate j and species s, $\gamma_{s,j}$ are generated from a binomial distribution with $p = 0.5$. Species-specific coefficients $\beta_{s,j}$ are generated according to eq. (4). Then we simply calculate $\boldsymbol{\mu}_i$ according to eq. (3), and sample $\boldsymbol{y}_i$ according to eq. (2). The $\rho_{i,s}$ must sum to 1, and, for $s \in \{1, ..., S-1\}$, must be equal to $\rho_{i,S} \exp(y_{i,s})$ by eq. (1). To achieve this we set $\rho_{i,S} = 1 / \sum_s^{S-1} \exp(y_{i,s})$. We can then calculate the $\rho_{i,s}$ and simulate count data $\mathbf{z}_i$ from a multinomial distribution with parameters $\boldsymbol{\rho}_i$, and number of trials equal to $100 * S$.

The bench-marking platform used for this study ran R-3.6.1 to generate datasets and invoke `rstan`. Models were fitted using the Stan software (Carpenter et al., 2017) version 2.19.1. A comparison between popular methods such as the PERMANOVA (generalized uniFrac method using the `GUniFrac` R package,

Chen and Chen, 2018), ANOSIM (non-parametric multivariate analysis using the `vegan` R package, Clarke, 1993) and MiRKAT (microbiome regression-based kernel association test using the `MiRKAT` R package, Zhao et al., 2015) has been done. The sampling was executed in parallel using 4 different chains each with 8000 iteration with a stopping criteria at convergence on 4 Intel Ivy Bridge cores each running at 1.15GHz speed and 12GB of RAM memory in total.

3.2 Results

In this section we investigate and compare the performance of:

- the full log-ratio model,
- the conditional log-ratio model which uses a linear marginal model for each species to determine the species-specific effect signs $\gamma_{s,j}$ and then infers the size of the species-specific effects $\beta_{s,j}$ and the global effect a_j conditional on the $\gamma_{s,j}$,
- the "oracle conditional model", where the $\gamma_{s,j}$ parameters are set to the true values used for simulating the data.

Figure 2 shows the performance of the models for estimating the global effect size in three different experiments: changing the sample size, noise variance on the $\beta_{s,j}$ parameters, and number of species. Dashed lines represent the true values of the global effect (a_j) in the simulation model. It can be seen that increasing the number of samples n improves the accuracy of prediction and reduces the variance. On the other hand, increasing the variance parameter for the effect values and increasing the number of species (or taxonomic units) both make estimation of the global effect parameters harder. We see that there is not much difference between the performance of the full model and that of the oracle conditional model, but as could be expected, the full log-ratio model produces less biased estimates compared to the conditional log-ratio model, with the oracle log-ratio model outperforming both. The bias in the conditional log-ratio model effect size estimates can be explained by the inherent model misspecification due to the estimation of $\gamma_{s,j}$ via independent linear regressions.

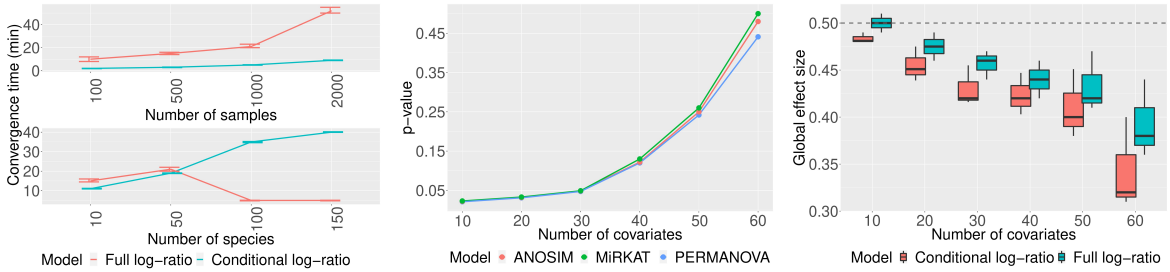
Figure 2 also reveals that even for large sample sizes, a small bias remains when estimating the global effect size for covariate 1, whose simulated value was set to zero. We observe that the simulated species-specific effect sizes tend to be small but non-zero. In the absence of a strong regularizing prior on the global effect, this leads to a small non-zero global effect. We chose not to impose such a prior to avoid over-regularising.

Figure 2b shows a comparison between the performance of the three models when predicting the species-specific effect size $\beta_{s,j}$. We note that species-specific effects for covariate 1 and the intercept are predicted well across all three models. For covariate 2, we again see a significant bias in the estimation for the conditional log-ratio model. Looking at both Figure 2a and Figure 2b we can see that in most cases the full log-ratio model outperforms the conditional model.

We have also checked the sensitivity of our model to differences in the variance parameters for the species-specific effect sizes $\beta_{s,j}$ (see supplementary), and note that despite the assumption of a common variance parameter, our model is robust to varying noise levels across species. Furthermore, in figure 6c (see supplementary), we can see the distribution of the estimated spices-specific effects when changing the number of spices. It can be seen that both models converge to a small non-zero value, however, in the full model sampler converged in a few iterations but the conditional model took longer to explore.

Figure 3b shows the p-values obtained using the PERMANOVA (generalized UniFrac), ANOSIM and MiRKAT methods applied to simulated data with $n = 100$ samples, $\sigma^2 = 0.1$ and $S = 10$. The number of covariates increases from $p = 10$ to $p = 60$. We note that empirically, for a large number of covariates (in this case, when the number of covariates exceeds 30), the statistical power becomes insufficient to reject the null hypothesis of no global effect.

The PERMANOVA, ANOSIM and MiRKAT models quantify the effect of a covariate via the variability in microbial compositions, and do not try to estimate a global effect size. Hence we cannot directly compare



(a) (b) (c)

Fig. 3: (a) Efficiency and estimation performance of the proposed models. Comparison of the computational time to convergence of the full and conditional log-ratio models, on the simulated data when increasing sample size n and number of species (or taxonomic units) S. The error bars show the 95% confidence interval calculated over 20 independent simulated dataset. (b) Comparison of the performance of the ANOSIM, MiRKAT and PERMANOVA models on simulated data when increasing the number of covariates. (c) Comparison of the performance of the full and conditional models on simulated data when increasing the number of covariates. The y axis represents the estimated global effects with the dashed line representing the mean value.

with our log-ratio models. Figure 3c shows the performance of the full and conditional log-ratio models when increasing the number of covariates. We see that increasing the number of covariates results in a slight negative bias in the estimated effect size. However, even for the maximum of $p = 60$ covariates, the estimated global effect size remains far from zero under our models, indicating a lower false negative rate compared to the methods in Figure 3b.

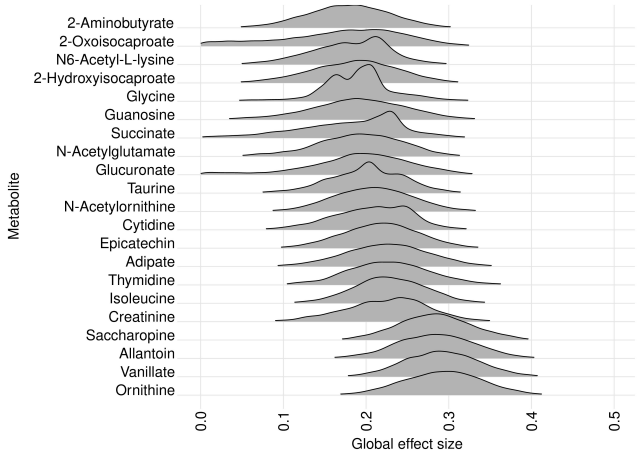
We also investigate the computational efficacy of the full and conditional log-ratio models. Figure 3a shows the run time to convergence under each of the models. We see that the sampler for the conditional model takes longer for large S; this is a consequence of setting the $\gamma_{s,j}$ parameters fixed which results in less efficient exploration of the sample space and longer convergence times. The sudden drop in time for $S \geq 100$ is explained by the fast convergence of the chains in the HMC sampler to a small non-zero value (when the true global effect value is set to zero). A comparison is shown in figure 7, where we can see the estimated spices-specific effects when changing the number of spices, across all different chains of HCM. It can be seen that when $S \geq 100$ the sampler in the full model converged in a few number of iterations but the conditional model took longer to explore.

Taken together, Figures 2, 2b and 3a indicate that for small sample sizes ($n < 100$) one could use either the full or conditional log-ratio models, as long as the number of covariates is not too large. For large sample sizes the conditional model could be deployed to improve computational efficiency, but there is a trade-off in terms of estimation accuracy. For large number of species S ($S > 50$), the full log-ratio model seems more efficient than the conditional model, perhaps because it can take advantage of Hamiltonian paths in the posterior landscape that are closed off when conditioning on $\mathbf{\Gamma}$. In general, we would recommend to use the full log-ratio model where possible, and only resort to the conditional log-ratio model when resource constraints and large sample sizes make use of the full model impractical.

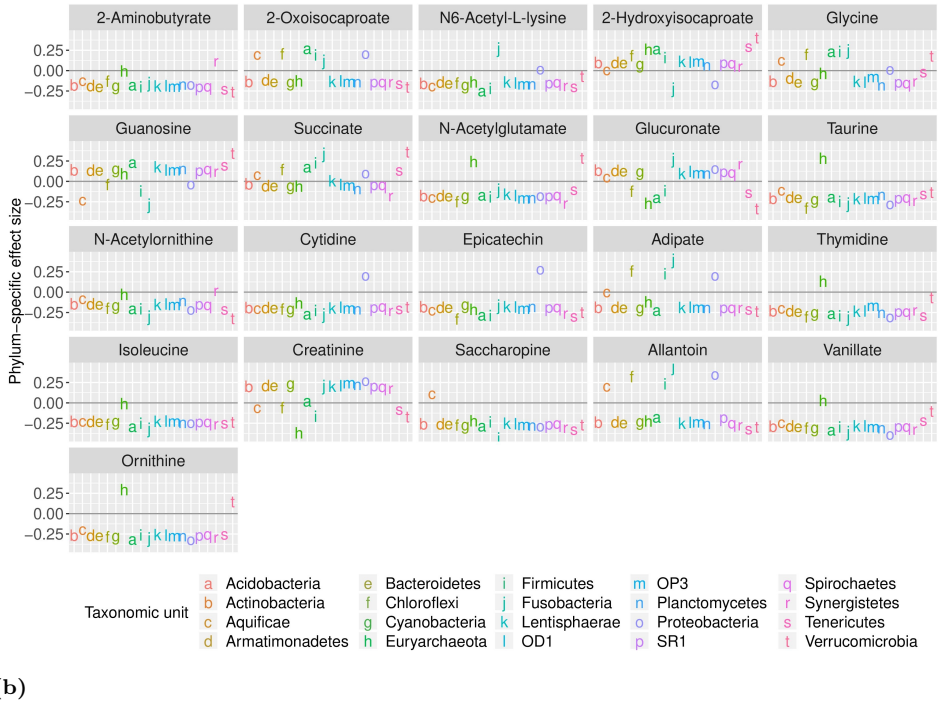
4 Real-World Applications

4.1 Detecting Associations

Using metabolomics as a measure of health, or to predict response to pharmacologics is a very attractive prospect, as measuring metabolites in urine, for example, may provide a non-invasive and potentially inexpensive screening method that has potential for application to a plethora of pathologies.



(a)



(b)

Fig. 4: Application of the log-ratio model to a microbiome-metabolome association study (Beamish, 2017). (a) Density plots showing the posterior distribution of the global effect size associating each metabolite to changes in the microbiome. (b) Posterior mean estimates of the phylum-specific effect size for each metabolite across the different bacterial phyla.

We apply the full log-ratio model to a dataset collected as part of a pilot study into inflammatory bowel disease (IBD) (Beamish, 2017). IBD is an umbrella term used to describe two conditions (Crohn's disease and ulcerative colitis) that are characterized by chronic inflammation of the gastrointestinal tract. 42 gut bacteria samples were obtained from patients undergoing colonoscopy at Lancashire Teaching Hospitals Trust (LTHTR) and University Hospitals of Morecambe Bay Trust (UHMBT), and metagenomic profiling was performed to obtain the bacterial species counts. The patients were additionally asked to provide a urine sample, and metabolite concentrations found in those samples were recorded. One objective of the study was to investigate associations between the gut microbiome and the urinary metabolite profiles.

After quality control, our cohort consists of $n = 40$ people (IBD cases and controls), for which we have measured a set of 3 clinical covariates: age, sex, and the hospital where the sample was collected. Note that

we do not investigate any associations with IBD status, as number of cases was too small to make any reasonable inferences. Instead, we focus on detecting associations between the microbiome and individual metabolites. We consider the set of the $S = 21$ most abundant operational taxonomic units (OTUs or species) at the phylum level. As a result, our design matrix $\mathbf{X}$ is a 40×5 matrix consisting of the intercept, the metabolite level, age, sex and sample collection location. As the data is limited, we do not try to infer the variance parameter σ^2, but keep it constant and equal to 0.1. Finally, we have a total of 261 metabolites.

Figure 4a shows the posterior distribution of estimated global effect size a_j for the metabolites with the largest posterior mean effect estimate. We see that out of these 21 metabolites, saccharopine, allantoin, vanillate and ornithine have the strongest association with the microbiome. Some metabolites show interesting bimodal patterns in their posterior effect distribution; this is particularly obvious for glycine. Such patterns could potentially be indicative of more than one mechanism that links the microbiome to metabolite levels.

Figure 4b shows the posterior mean estimate of the species-specific effect sizes $\beta_{s,j}$, specifically the association of each phylum with each metabolite. We note that for the metabolites with the largest posterior effects, the majority of species-specific effects are strongly negative, indicating that these metabolites are associated with a decrease in most phylum counts, compensated by an increase in just a couple of phyla.

We note that our analysis identified several phyla-associated metabolites that have previously been shown to be linked to markers of intestinal health or nutrition, demonstrating potential impacts on health that are reflective of microbiota changes. First, our study suggests that adipate is strongly associated with changes in the microbiome. Adipate, the salts and esters of adipic acid, are increased in the urine of breast-fed vs. formula-fed neonates (Dessì et al., 2016), a now well-known essential factor in the establishment of a diverse, healthy microbiome. We observe effect sizes that are indicative of an association between adipate and an increase in Firmicutes, Proteobacteria and Bacteroidetes, the major components of a healthy adult intestinal flora. Therefore, clinical measurement of adipate may have potential as biomarkers for intestinal health.

Epicatechin is a polyphenolic flavonoid abundantly present in plants, cocoa and that offers protection against diabetes by its ability to mimic the effects of insulin and has strong antioxidant properties (Kanehisa and Goto, 2000; Samarghandian et al., 2017). The association we observed between epicatechin and proteobacteria abundance may prove useful in the prediction of pathologies associated with proteobacteria expansion, such as IBD or nonalcoholic fatty liver disease (Rizzatti et al., 2017). Of course, any potential mechanisms that underpin these associations will require further investigation but collectively this demonstrates the broad and significant potential for clinical application of our model. Several metabolites, including isoleucine, thymidine, ornithine and vanillate, were associated with an increase in cyanobacteria (Figure 4b). Cyanobacteria are unusual intestinal microflora (Ley et al., 2005), in that they perform oxygenetic photosynthesis and cyanobacteria-based supplements, such as Spirulina, regarded as functional foods, have been shown to support immunological function (Finamore et al., 2017). These associations may allow us to get a clearer picture of the functionality of the gut microbiome, by assessing the functions of these metabolites.

4.2 Estimating the Global Effect of Diet on the Gut Microbiome

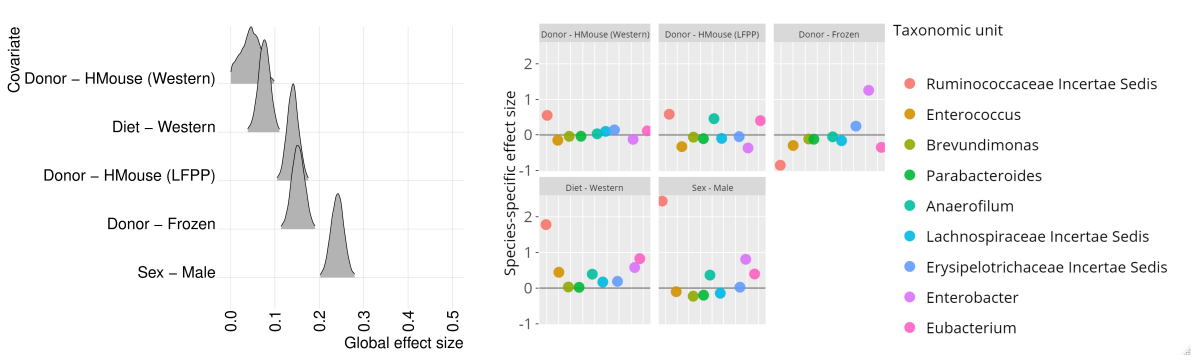
It is well-known that diet has a profound effect on the composition of the gut microbiome (David et al., 2014; Singh et al., 2017). However it is difficult to quantify this effect using traditional microbiome analysis methods, as we either need to look at individual species, or rely on a pre-specified distance metric. In (Turnbaugh et al., 2009), the authors took the latter approach to characterise the differences in microbiota in mice whose gut tracts were colonised by donor bacteria of human or mouse origin, and who received either a Western or low-fat plant-based (LF/PP) diet. The dataset is described in Table 1. Donor bacteria either originated from human microbiome samples (fresh or frozen), or from mice who had previously been implanted with (fresh) human samples (HMouse) and had received either the LF/PP or Western diet. The dataset consists of $n = 444$ fecal samples, obtained at various time points from 15 mice, and we use the

Tab. 1: Dataset Description for The Gut Microbiome Study in (Turnbaugh et al., 2009).

Number of observations	Variable
394	Sex - Male
50	Sex - Female
137	Donor - Fresh
92	Donor - Frozen
175	Donor - HMouse (LF/PP Diet)
40	Donor - HMouse (Western Diet)
300	Diet - LF/PP
144	Diet - Western

bacterial abundance data for $S = 160$ taxonomic units at the genus level. Note that we did not have access to the information on sample times or sample source, and so were unable to take this aspect into account in our model.

Using the UniFrac distance measure, (Turnbaugh et al., 2009) found that recipient mice consuming a Western diet showed a similar microbiome composition that was distinct from mice on the LF/PP diet, regardless of donor. We applied the full log-ratio model to the dataset to determine whether this was reflected in the global microbiome effect of diet on the microbiome, and to study the individual (genus-specific) effects. We include sex, donor type and diet as covariates. The reference taxonomic unit for the log-ratio transform was the Dorea genus, although investigation of alternative reference taxonomic units showed little change in the results (see supplementary information).



(a) (b)

Fig. 5: Application of the log-ratio model to a gut microbiome study (Turnbaugh et al., 2009). The reference taxonomic unit for the log-ratio transformation is Dorea. (a) Density plots showing the posterior distribution of the global effect size associating each covariate to changes in the microbiome. Note that the reference category for diet is LF/PP, meaning that the global effect is the change with respect to that diet. Similarly the reference category for bacterial donor is fresh (human), and the reference category for sex is female. (b) Posterior mean estimates of the genus-specific effect size for each metabolite across the different bacterial genera. We display here the 10 OTUs with the highest effect size for the association with diet.

The results of our study are shown in Figure 5. We observe that overall there is a significant (non-zero) effect of diet, and a strong effect of using the frozen samples. This is concordant with the findings in (Turnbaugh et al., 2009). We also note that receiving a bacterial implant from a donor mouse on the LF/PP diet seems to have a stronger effect than receiving an implant from a donor mouse on the Western diet, compared to having a human donor. Ruminococcaceae Incerta Sedis is strongly associated with the change in diet, a finding which is backed up by the literature (Clarke et al., 2013). Similarly, Eubacteria colonization of the gut has been shown to be affected by diet (Simmering et al., 2002). Finally, we see that

Enterobacteria is associated with the change in diet; previous studies in rats have shown that a high-fat diet results in considerably more propionate and acetate producing species, including Enterobacteria (Singh et al., 2017). Interestingly, this bacterial genus also shows a strong association with implantation by frozen donor bacterial samples.

5 Conclusion

We have developed a hierarchical Bayesian regression model for detecting global microbiome associations, and have shown that the model produces consistent estimates in a simulation study. We then applied the model to investigate microbiome-metabolome associations in a pilot study of inflammatory bowel disease patients and global associations with diet in a study of mouse gut microbiomes. The model is superior to existing approaches in that it allows for integrating phylogenetic information, it estimates global as well as species-specific effects, and it does not rely on a specific choice of ecological distance measure.

There are some limitations to our study. The IBD dataset used in this paper only consists of a small number of samples, which reduced our ability to extend the inference to a larger number of species. For a larger sample size, the full log-ratio model could also be applied to taxonomic units at the 'order' or 'genus' level, as our simulation study showed that inference for up to 150 species is feasible without major modifications. However, we have seen that for large number of samples ($n \geq 500$), the conditional log-ratio model can be deployed as it showed comparable prediction accuracy to the full model at greater computational efficiency.

In our model, information about the phylogenetic relationships between species is incorporated via a simple Brownian motion model of evolution (Felsenstein, 1973). This model was originally developed for the evolution of traits, and assumes a priori that closely related species have traits that are positively correlated, which is a reasonable assumption. Here we are using it to model species abundances, where this assumption is likely an oversimplification, as it does not incorporate negative correlation due to inter-species competition. Further modelling work is needed to derive more realistic models of the impact of phylogenetic relatedness on abundance.

We note that our method does not currently explicitly model the common problem of zero-inflation. Zeros in the count data are dealt with by adding a small constant before applying the log-ratio transform. Despite the simplicity of this approach, we have not observed significantly biased estimates as a result in any of our simulations.

In this paper, we focused on comparisons with methods for detecting global effects on the microbial community. Numerous other methods exist that are limited to detecting species-specific effects. Methods such as FZINBMM (Zhang and Yi, 2020) and ALDEx2 (Gloor, 2015), as well as more general count models like edgeR (Robinson et al., 2010) and DESeq2 (Anders and Huber, 2010) do not allow for the estimation of global effects. Additionally, the latter two do not properly model the compositional nature of metagenomic data. As a consequence, comparisons with these methods would have been of limited value. We did investigate the potential of these methods for initialising the $\gamma_{s,j}$ parameters in the conditional log-ratio model, but found the performance not significantly better than the method described in our paper.

An interesting avenue for future research is whether our model can scale to larger numbers of species S and sample sizes n. We note that time to convergence seems to increase polynomially with n; we would therefore not expect to be able to scale to tens of thousands of samples, and a different inference method would need to be employed. One option is variational inference, which has been successfully used as an alternative to full Bayesian inference in other settings (Hensman et al., 2013).

More crucially, we would like the method to scale to larger numbers of species. Currently, our results indicate that HMC inference in the full model converges faster as S increases. Our hypothesis is that for small S, the gradient estimate for a_j is unreliable as it depends on the estimates of $\beta_{s,j}$. For large S, we have more information to estimate the gradients, and thus experience fewer rejections during HMC sampling. However, estimation of the parameters will be poor unless n also increases. In future work, we

will investigate regularization approaches to estimate sparse $\beta_{s,j}$ parameters in situations with small sample sizes.

Funding: This work has been supported by the University Hospitals of Morecambe Bay Trust SIFT funding and the Academy of Medical Sciences.

Appendix A - Sensitivity Analysis

We have performed a sensitivity analysis using simulated data to test the impact of our modelling assumption that the microbial species share a common noise parameter. The simulation model is the same as described in Section 3 in the main manuscript, but instead of drawing the $\beta_{s,j}$ from a truncated Gaussian with common parameter σ^2, we instead sample σ_j^2 from a truncated Gaussian with standard deviation 0.1. The results are shown in supplementary Figure 6. We can compare this to Figure 2 to see that the estimated species-specific effects and the global effects stay consistent when using different values σ_j^2 for each species j at the simulation stage.

Appendix B - Hamiltonian Monte Carlo Traces

Figure 3a indicates that for larger number of species S, the convergence time of the Hamiltonian Monte Carlo algorithm decreases. In Supplementary Figure 7, we show additional example trace plots for the species-specific effect parameters $beta_{s,j}$ under the default simulation setup described in the main paper. We observe that for $S \geq 100$, convergence under the full log-ratio model is reached earlier, but the posterior distribution is wider than for $S < 100$. The conditional log-ratio model reaches a similar equilibrium state, but the width of the distribution of samples parameters reduces more gradually, possibly as a result of the restrictions imposed by the two-stage approach. It is reasonable to assume that increasing S also increases the information available to estimate the gradients of the log likelihood function with respect to the global effect size parameters, and thus the chains experience fewer rejections during HMC sampling. However, as the number of samples does not increase, and the number of species-specific parameters increases, the estimate overall becomes more uncertain, resulting in a wider posterior distribution. This can also be observed in Figure 2a.

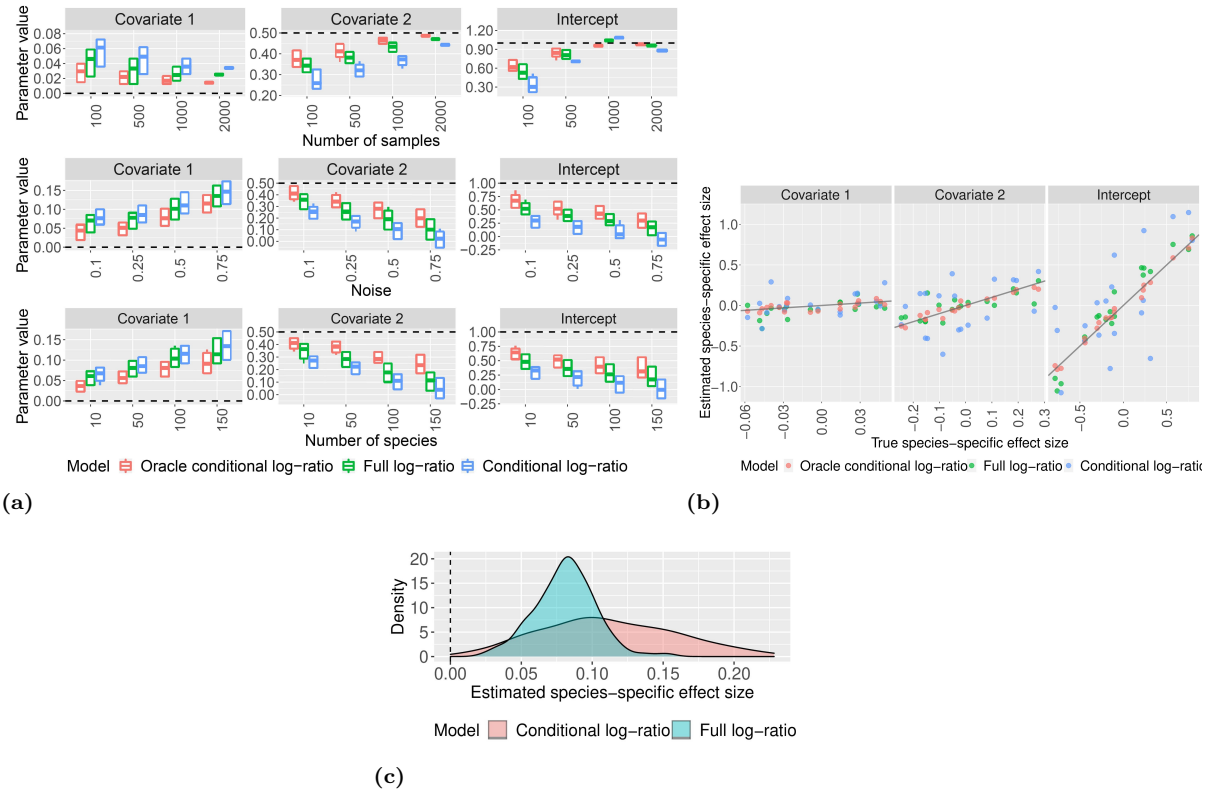


Fig. 6: Model performance when simulating data under species-specific variance in the effect sizes. (a) Comparison of the performance of the full, conditional and oracle conditional log-ratio models on simulated data when varying the sample size (top row), noise variance of the species-specific effects (middle row) and number of species or taxonomic units (bottom row). Notice that x axis is now the mean value of p randomly drawn values centered on the relevant noise and with standard deviation 0.1. This is to have different noise parameters for different species when simulating the data. The linear simulation model consists of the intercept term and two covariates with global effect sizes 1, 0.5 and 0 respectively. The dashed horizontal lines show the size of the true effect, and the boxplots represent the posterior mean estimates obtained from 20 independent simulated datasets. (b) Comparison of the performance of the full, conditional and oracle conditional log-ratio models on simulated data when estimating the species-specific effect size ($\beta_{s,j}$). Each point corresponds to the estimated effect size for one species in one model. The linear simulation model consists of the intercept term and two covariates with global effect sizes 1, 0.5 and 0 respectively. The solid lines represent $y = x$ (note that the x-axes are on different scales). (c) Comparison of the distribution of the estimated species-specific effect size resulted from the full and conditional models for a relatively large number of species $S = 100$. The dashed vertical line shows the true global effect.

References

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., Hayes, R. B., and Yang, L. (2013). Human gut microbiome and risk for colorectal cancer. *J. Natl. Cancer Inst.*, 105(24):1907–1911.

Aitchison, J. (1982). The statistical analysis of compositional data. *J. Royal Stat. Soc. Ser. B (Methodological)*, 44(2):139–160.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Preced.*, pages 1–1.

Beamish, E. (2017). *Investigating Dysbiosis as a Cause and Predictor of Intestinal Pathology*. PhD thesis, Lancaster University.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.*, 76(1).

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113.

Chen, J. and Chen, M. J. (2018). Package GUniFrac. *The Compr. R Arch. Netw. (CRAN)*.

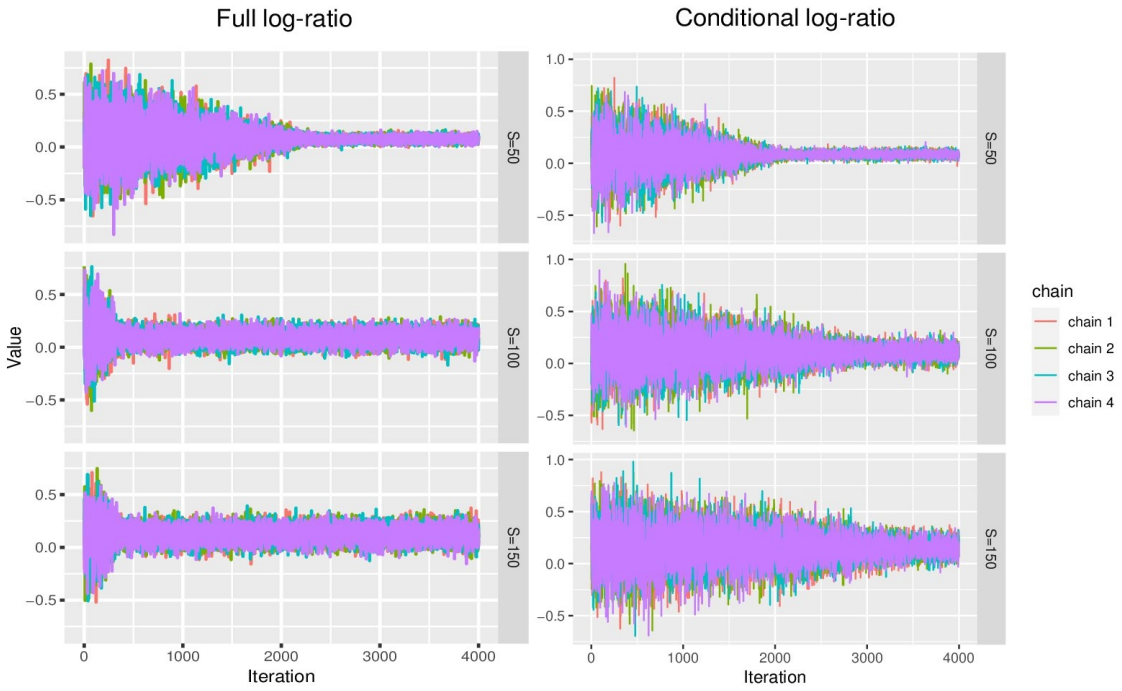


Fig. 7: Comparison of the trace plots of the HMC path for $\beta_{s,j}$ in the full and conditional models. The y-axis shows value of species-specific parameters $\beta_{s,j}$ in each chain for increasing number of species.

Chen, J. and Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals Appl. Stat.*, 7(1).

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.*, 18(1):117–143.

Clarke, S. F., Murphy, E. F., O'Sullivan, O., Ross, R. P., O'Toole, P. W., Shanahan, F., and Cotter, P. D. (2013). Targeting the microbiota to address diet-induced obesity: a time dependent challenge. *PLoS One*, 8(6):e65790.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563.

Dessì, A., Murgia, A., Agostino, R., Pattumelli, M. G., Schirru, A., Scano, P., Fanos, V., and Caboni, P. (2016). Exploring the role of different neonatal nutrition regimens during the first week of life by urinary GC-MS metabolomics. *Int. J. Mol. Sci.*, 17(2):265.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, 25(5):471.

Finamore, A., Palmery, M., Bensehaila, S., and Peluso, I. (2017). Antioxidant, immunomodulating, and microbial-modulating activities of the sustainable and ecofriendly spirulina. *Oxid. Med. Cell. Longev.*, 2017:3247528.

Gloor, G. (2015). Aldex2: Anova-like differential expression tool for compositional data. *ALDEX manual modular*, 20:1–11.

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, 2(5):17004.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *Uncertain. Artif. Intell.*, pages 282–290.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci.*, 102(31):11070–11075.

Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–8235.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The CONCRETE distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C., and Gasbarrini, A. (2017). Proteobacteria: A common factor in human diseases. *Biomed Res. Int.*, 2017:9351507.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Samarghandian, S., Azimi-Nezhad, M., and Farkhondeh, T. (2017). Catechin treatment ameliorates diabetes and its complications in streptozotocin-induced diabetic rats. *Dose Response*, 15(1):1559325817691158.

Simmering, R., Pforte, H., Jacobasch, G., and Blaut, M. (2002). The growth of the flavonoid-degrading intestinal bacterium, eubacterium ramulus, is stimulated by dietary flavonoids in vivo. *FEMS Microbiol. Ecol.*, 40(3):243–248.

Singh, R. K., Chang, H.-W., Yan, D., Lee, K. M., Ucmak, D., Wong, K., Abrouk, M., Farahnik, B., Nakamura, M., Zhu, T. H., Bhutani, T., and Liao, W. (2017). Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.*, 15(1):73.

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.*, 1(6):6ra14.

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput. Biol.*, 6(2):e1000667.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1):4.

Zhang, X. and Yi, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics*, 36(8):2345–2351.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H., and Wu, M. C. (2015). Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The Am. J. Hum. Genet.*, 96(5):797–807.