# Uncovering the drivers of host-associated microbiota with joint species distribution modeling

Johannes R. Björk[1,4*], Francis K.C. Hui[2], Robert B. O'Hara[3], and Jose M. Montoya[4]

[1]Department of Biological Sciences, University of Notre Dame, United States

[2]Mathematical Sciences Institute, The Australian National University, Canberra, Australia

[3]Department of Mathematical Sciences, NTNU, Trondheim, Norway

[3]Biodiversity and Climate Research Centre, Frankfurt, Germany

[4]Theoretical and Experimental Ecology Station, CNRS-University Paul Sabatier, Moulis, France

[1,4*]*rbjork@nd.edu (Corresponding author)*

[2]*francis.hui@anu.edu.au*

[3]*bob.ohara@ntnu.no*

[4]*josemaria.montoyateran@sete.cnrs.fr*

# Abstract

In addition to the processes structuring free-living communities, host-associated microbiota are directly or indirectly shaped by the host. Therefore, microbiota data have a hierarchical structure where samples are nested under one or several variables representing host-specific factors, often spanning multiple levels of biological organization. Current statistical methods do not accommodate this hierarchical data structure, and therefore cannot explicitly account for the effect of the host in structuring the microbiota. We introduce a novel extension of joint species distribution models (JSDMs) which can straightforwardly accommodate and discern between effects such as host phylogeny and traits, recorded covariates like diet and collection sites, among other ecological processes. Our proposed methodology includes powerful yet familiar outputs seen in community ecology overall, including: (i) model-based ordination to visualize and quantify the main patterns in the data; (ii) variance partitioning to asses how influential the included host-specific factors are in structuring the microbiota; and (iii) co-occurrence networks to visualize microbe-to-microbe associations.



**Keywords: Host-associated; Microbiota; Microbiome; Joint species distribution models; Generalized linear mixed models; Bayesian inference**


# Introduction

Ecological communities are the product of stochastic and deterministic processes; while environmental factors may set the upper bound on carrying capacity, competitive and facilitative interactions within and among taxa determine the identity of the species present in local communities. Ecologists are often interested in inferring ecological processes from patterns and determining their relative importance for the community under study [1]. During the last few years, there has been a growing interest in developing new statistical tools aimed toward ecologists and the analysis of multivariate community data (see e.g., [2]). Many of the distance-based approaches however, have a number of drawbacks, including uncertainty of selecting the most appropriate null models, low statistical power, and the lack of possibilities for making predictions [3]. One alternative, model-based framework which has become increasingly popular in community ecology is joint species distribution models (JSDMs, [4]). Such models are an extension of generalized linear mixed models (GLMMs, [5]), where multiple species are analyzed simultaneously often together with environmental variables, thereby revealing community level responses to environmental change. By incorporating both fixed and random effects, sometimes at multiple levels of biological organization, JSDMs have the capacity to

assess the relative importance of processes such as environmental and biotic filtering versus stochastic variability. Furthermore, with the increase of trait-based and phylogenetic data in community ecology together with the growing appreciation that species interactions are constrained by the "phylogenetic baggage" they inherit from their ancestors [6], JSDMs can further accommodate information on both species traits and phylogenetic relatedness among species [7, 8, 9, 10]. Finally, accounting for phylogenetic relatedness among species can greatly improve estimation accuracy and power when there is a phylogenetic signal in species traits and/or residual variation ([11]).

To model covariances between a large number of species using a standard multivariate random effect, as a standard JSDM [4, 12] does, is computationally challenging; the number of parameters that needs to be estimated when assuming a completely unstructured covariance matrix increases rapidly (quadratically) with the number of species. An increasingly popular tool for overcoming this problem, which is capable of modeling such high-dimensional data, is latent factor models [13]. In community ecology, latent factor models and JSDMs have been combined to allow for a more parsimonious yet flexible way of modeling species covariances in large communities [10, 14]. Such an approach offers a number of benefits. First, latent factors provide a method of explicitly accounting for residual correlation. This is important because missing covariates, ecological interactions and/or spatio-temporal correlation will induce residual correlation among species, which, if not accounted for, may lead to erroneous inference. Second, latent factors facilitate model-based ordination in order to visualize and quantify the main patterns in rows and/or columns of the data [15, 16]. While traditional distance-based ordination techniques may confound location (i.e., the mean abundance) and dispersion (i.e., the variability) effects [3], model-based ordination directly models the mean-variance relationship and can therefore accurately distinguish between the two effects [17, 18]. Finally, the estimated factor loadings can be conveniently interpreted as indicating whether two species co-occur more or less often than by chance as well as the direction and strength of their co-occurrence, thus allowing a latent factor approach to robustly estimate large species-to-species co-occurrence networks [19]. Note that an important decision when fitting latent factor models, is the choice of the number of latent factors. While less than five is usually sufficient for a good approximation to correlations, there is a trade-off between model complexity and the model's capacity to capture the true correlation structure ([13]). An alternative approach is to use variable selection, which automatically shrinks less-informative latent factors to zero ([20]).

In parallel to community ecology, there is a growing field of microbial ecology studying both free-living and host-associated microbiota. While microbial ecologists can adopt many of the same statistical tools developed for traditional multivariate abundance data (see e.g., [21]), researchers studying host-associated microbiota need to consider an additional layer of processes structuring the focal community, namely that host-associated microbiota are addition-

ally shaped directly or indirectly by their hosts. For example, interactions between hosts and microbes often involve long-lasting and sometimes extremely intimate relationships, where the host may have evolved the capacity to directly control the identity and/or abundance of its microbial symbionts [22, 23]. Similar to an environmental niche, the host must be viewed as a multidimensional composite of all host-specific factors driving the occurrence and/or abundance of microbes within a host–everything from broad evolutionary relationships between host species [24] to the direct production of specific biomolecules within a single host individual [25]. As a result, host-associated microbiota have a hierarchical data structure where samples are nested under one or several variables representing recorded and/or measured host-specific factors sometimes spanning multiple levels of biological organization.

In this article, we propose a novel extension of JSDMs to analyze host-associated microbiota, based around explicitly modeling its characteristic hierarchical data structure. In doing so, our proposed model can straightforwardly accommodate and discriminate among any measured host-specific factors. Over the past few years, there has been an increase of model-based approaches aimed specifically toward the analysis of host-associated microbiota (see e.g., [12, 26, 27, 28]). To our knowledge however, our proposed model is the first to explicitly and transparently account for the aforementioned hierarchical structure that is inherent in data on host-associated microbiota (Fig 1). Other key features of the proposed model, which are inherited from JSDMs and latent factor models, include: (1) parsimonious modeling of the high-dimensional correlation structures typical of host-associated microbiota; (2) model-based ordination to visualize and quantify the main patterns in the data; (3) variance partitioning to assess the explanatory power of the modeled host-specific factors and their influence in shaping the microbiota; and finally (4) co-occurrence networks to visualize OTU-to-OTU associations. Furthermore, by building our model in a probabilistic, i.e., Bayesian framework, we can straightforwardly sample from the posterior probability distribution of the correlation matrix computed by the factor loadings; this means that we can choose to look at, or further analyze the correlations that have at least e.g., 95% (or even 97% or 99%) probability.

We apply our proposed model to two published data sets. While we include the effect of host phylogenetic relatedness in both case studies, we illustrate the flexibility of our approach by adapting the proposed model to overdispersed counts and presence-absence responses, and study-specific meta data relevant to each case study. By utilizing recent progress in latent factor modeling, our proposed model can also assist in cases where meta data are scarce by finding latent "hidden" variables driving the microbiota.

# Methods

We applied the proposed methodology to two published data sets on host-associated microbiota. Both datasets possess two main features which characterize many host-associated microbiota data, namely high dimensionality i.e., the number of OTUs is a non-negligable proportion of the number of samples, and sparsity i.e., most OTUs are rarely observed. The first data set comprise 90 samples from 20 sponge species collected in four closely located sites in the Bocas del Toro archipelago (Fig S1, for original study see [29]). The meta data contain apart from collection site, a classification of hosts into either High Microbial Abundance (HMA) or Low Microbial Abundance (LMA) sponges (hereafter termed *ecotype*). This classification is based on the abundance of microbes harbored by the host and determined by transmission electron microscopy [30]. The authors constructed a host phylogeny from 18S rRNA gene sequences (downloaded from GenBank) by implementing a relaxed-clock model in MrBayes. The data have a hierarchical structure with $n = 90$ samples nested within $S = 20$ host species and $L = 4$ collection sites. Host species are then further nested under one of $R = 2$ ecotypes. The response matrix had already been filtered to only include OTUs (defined at 97% similarity) with at least 500 reads, but we further removed OTUs with less than 20 presences across samples, resulting in $m = 187$ modeled OTUs.

The second data set consists of 59 neotropical bird species with a total of 116 samples from the large intestine. Host species were collected from 12 lowland forests sites across Costa Rica and Peru (Fig S2, for original study see [31]). The meta data include bird taxonomy and several covariates–including dietary specialization, stomach contents and host habitat. The authors sequenced and used the mitochondrial locus ND2 to reconstruct the host phylogeny by implementing a partitioned GTR + $\Gamma$ model in BEAST. Similarly to the sponge data set, this data set has a hierarchical structure with $n = 116$ samples nested within $S = 59$ host species and $L = 12$ collection sites. We filtered the response matrix to include OTUs (defined at 97% similarity) with at least 50 reads and 40 presences across samples, resulting in $m = 151$ modeled OTUs. Of the full list of covariates available, we included diet, stomach content, sex, elevation and collection site as explanatory predictor variables in our model. While diet and geography have been shown to influence the human gut microbiota (see e.g., [32, 33]), the effect of sex and elevation is less known.

## Joint species distribution models

We considered two response types commonly encountered in host-associated microbiota data: counts and presence-absence. Formally, let the response matrix being modeled consist of either counts or presence-absence records of $m$ OTUs from $n$ samples, and let $y_{ij}$ denote the response of the $j$-th OTU in the $i$-th sample. Also, let $\mathcal{N}(\mu, \sigma^2)$ denote

a univariate normal distribution with mean $\mu$ and variance $\sigma^2$, and analogously, let $\mathcal{MVN}(\boldsymbol{\mu}, \Sigma)$ denote a multivariate normal distribution with mean vector and covariance matrix $\Sigma$. We now split our model formulation up into the two case studies/response types.

**Case Study 1 (Counts):** Due to the presence of overdispersion that was quadratic in nature, as confirmed by a mean-variance plot of the OTU counts (not shown), we assumed a negative binomial distribution for the responses. Specifically, we considered a negative binomial distribution with a quadratic mean-variance relationship for the element $y_{ij}$, such that $\text{Var}(y_{ij}) = \psi_{ij} + \phi_j \psi_{ij}^2$ where $\phi_j$ is the OTU-specific overdispersion parameter. The mean abundance was related to the covariates using a log-link function. Denoting the mean abundance of OTU $j$ in sample $i$ by $\psi_{ij}$, then we have

**Model 1**

$$y_{ij} \sim \text{Negative-Binomial}(\psi_{ij}, \phi_j), \qquad i = 1, \ldots, n = 90, \quad j = 1, \ldots, m = 187 \tag{1}$$

$$\log(\psi_{ij}) = \alpha_i + \gamma_j + \sum_{q=1}^{5} Z_{iq}\lambda_{qj} + \sum_{q=1}^{5} Z^H_{s[i]q}\lambda^H_{qj}, \qquad q = 1, \ldots, 5 \tag{2}$$

$$\begin{aligned}
\alpha_i &\sim \mathcal{N}(\mu(\text{host})_{s[i]}, \sigma^2(\text{sample})) \\
\mu(\text{host})_s &= \mu(\text{ecotype})_s + \mu(\text{site})_s + \mu(\text{phylo})_s \times \theta_{\text{phylo}}, \qquad s = 1, \ldots, S = 20 \\
\mu(\text{ecotype})_s &\sim \mathcal{N}(\mu_{r[s]}, \sigma^2(\text{ecotype})) \\
\mu(\text{site})_s &\sim \mathcal{N}(\mu_{l[s]}, \sigma^2(\text{site})) \\
\boldsymbol{\mu}(\text{phylo})_s &\sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{C}(\text{phylo})) \\
\mu_r &\sim Cauchy(0, 2.5), \qquad r = 1, \ldots, R = 2 \\
\mu_l &\sim Cauchy(0, 2.5), \qquad l = 1, \ldots, L = 4 \\
\gamma_j &\sim Cauchy(0, 2.5) \\
\theta_{phylo} &\sim Exp(0.1)
\end{aligned} \tag{3}$$

To clarify the above formulation, $s$, $r$ and $l$ index effects that are attributed to the $S = 20$ host species, $R = 2$ ecotypes and $L = 4$ sites respectively. For instance, "$s[i]$" and "$r[s]$" denote "sample $i$ nested within host species $s$" and "host species $s$ nested within ecotype $r$", respectively (Fig 1). In equation (2), the quantities $\alpha_i$ and $\gamma_j$ represent sample and OTU-specific effects, respectively. The former adjusts for differences in sequencing depth among samples, while the latter controls for differences in OTU total abundance. The inclusion of $\alpha_i$ serves two main purposes. First

and foremost, including $\alpha_i$ allows us to account for the hierarchical data structure and its effect on sample total abundance specifically. In particular, to account for sample $i$ being nested within host species $s$ (which are further nested within ecotype $r$) and site $l$, the sample effects $\alpha_i$ are drawn from a normal distribution with a mean that is a linear function of three host-specific effects: host ecotype $\mu$(ecotype); host collection site $\mu$(site); and host phylogeny $\mu$(phylo). Furthermore, the host ecotype $\mu$(ecotype) and host collection site $\mu$(site) effects are themselves drawn from a normal distribution with an ecotype and site-specific mean, respectively. Second, the inclusion of $\alpha_i$ means that the resulting ordinations constructed by the latent factors on the sample $Z_{iq}$ and host species $Z^H_{s[i]q}$ level are in terms of species composition only, as opposed to a composite of abundance and composition if the site effects were not included in the formulation. We included five latent factors at both the sample and host species level, and both $Z_{iq}$ and $Z^H_{s[i]q}$ were assigned standard normal priors $\mathcal{N}(0, 1)$ with the assumption of zero mean and unit variance to fix the location and scale (see Chapter 5, [34]). Furthermore, to address rotational variance, the upper triangular component of both loading matrices (i.e., sample $\boldsymbol{\lambda}$ and host species $\boldsymbol{\lambda}^H$ level) are fixed to zero with the diagonals constrained to be positive [35]. As recommended by Polson and Scott [36], and analogous to the prior distributions we use for the mean $\mu_r$ and $\mu_l$, we used a weakly informative prior in the form of a half-Cauchy distribution with a center and scale equal to 0 and 2.5 for the overdispersion parameter $\psi$. Moreover, following Gelman et al. [37], we used the same distribution with location and scale equal to 0 and 1 as prior information on the variance parameters: $\sigma^2$(sample); $\sigma^2$(ecotype); and $\sigma^2$(site). Based on our empirical investigation, we found that the use of such priors stabilized the MCMC sampling substantially without introducing too much prior information, compared to using more uninformative prior distributions. Lastly, the quantity $\boldsymbol{C}$(phylo) corresponds to a phylogenetic correlation matrix constructed from the host phylogeny by assuming Brownian motion evolution such that the covariances between host species are proportional to their shared branch length from the most recent common ancestor [38]. The phylogenetic parameter $\theta_{\text{phylo}}$ quantifies variance that can be attributed to the phylogenetic effect, and is drawn from an exponential distribution with a rate parameter of 0.1. Similar to the half-Cauchy priors, this prior distribution provides a weak level of regularization–a rate parameter of 0.1 gives a prior mean of 10, thus preventing the estimated variance of getting implausibly large.

**Case Study 2 (Presence-absence):** We modelled the presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$) of OTU $j$ in sample $i$ using probit regression, implemented via the indicator function $1_{z_{ij}>0}$ where the latent score is normally distributed with the mean equal to a linear function of the covariates and latent factors, and variance set equal to one. The hierarchical model was set up as follows:

**Model 2**

$$z_{ij} \sim \alpha_i + L_{ij} + \sum_{q=1}^{5} Z_{iq}\lambda_{qj}, \qquad i = 1,\ldots,n = 116, \quad j = 1,\ldots,m = 151, \quad q = 1,\ldots,5 \tag{4}$$

$$L_{ij} = \gamma_j + \sum_{k=1}^{5} X_{ik}\beta_{kj}, \qquad k = 1,\ldots,5 \tag{5}$$

$$\alpha_i \sim \mathcal{N}(\mu(\text{host})_{s[i]}, \sigma^2(\text{sample}))$$

$$\mu(\text{host})_s = \mu(\text{non-phylo})_s + \mu(\text{phylo})_s \times \theta_{\text{phylo}}, \qquad s = 1,\ldots,S = 59 \tag{6}$$

$$\mu(\text{non-phylo})_s \sim \mathcal{N}(\mu_s, \sigma^2(\text{host}))$$

$$\boldsymbol{\mu}(\text{phylo})_s \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{C}(\text{phylo}))$$

$$\mu_s \sim Cauchy(0, 2.5)$$

$$\gamma_j \sim Cauchy(0, 2.5)$$

$$\psi_{ij} \sim \textit{half-Cauchy}(0, 2.5)$$

$$\sigma^2(\text{sample}) \sim \textit{half-Cauchy}(0, 1)$$

$$\theta_{phylo} \sim Exp(0.1)$$

While the above description is largely the same as that of *Model 1*, we also included here a linear predictor $L_{ij}$ to model the effects of five available covariates (represented by the model matrix $X_{ik}; k = 1,\ldots,5$) on species composition (equation (5)). The linear predictor $L_{ij}$ thus acts to explain covariation between OTUs due to the measured explanatory predictor variables, while the latent factors account for the remaining, residual covariation. Similarly to *Model 1*, including $\alpha_i$ means that the covariation between OTUs is in terms of species composition only. By drawing the sample effects $\alpha_i$ from a normal distribution with a mean that is a linear function of both non-phylogentic $\mu(non-phylo)$ and phylogenetic $\mu(\text{phylo})$ host effects (equation (6)), we account for the hierarchical structure present in the data. Furthermore, from the loading matrix $\boldsymbol{\lambda}$, we computed a covariance matrix as $\Omega = \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, which we subsequently convert to a correlation matrix for studying the OTU-to-OTU co-occurrence network.

For both case studies, we used Markov Chain Monte Carlo (MCMC) to estimate the models via `JAGS` [39] and the `runjags` package [40] in `R` [41]. For each model, we ran one chain with dispersed initial values for 300,000 iterations saving every $10^{th}$ sample and discarding the first 25% of samples as burn-in. We evaluated convergence of model parameters by visually inspecting trace and density plots using the R packages *coda* [42] and *mcmcplots* [43], as well as using the Geweke diagnostic [44].

## Variance partitioning

To discriminate among the relative contributions of the various factors driving covariation in the JSDMs, we partition the explained variance by the row effects ($\alpha_i$), the linear predictor ($L_{ij}$), and the loadings ($\lambda_{qj}$ and $\lambda^H_{qj}$) into components reflecting sample and host level effects. Such a variance decomposition is analogous to the sum-of-squares and variance decompositions seen in Analysis of Variance (ANOVA) and linear mixed models ([45]). Depending on the response type, the row effects capture variance in relative abundance (*Model 1*) or species richness (*Model 2*), while the linear predictor and the loadings capture variance in species composition. As mention above, when the linear predictor is included in (*Model 2*), the loadings capture residual variation not accounted for by the modeled covariates. Variance partitioning therefore allows us to asses the explanatory power of the hierarchical data structure, and measured covariates including "hidden" factors, and how influential each of them are in structuring the host-associated microbiota ([10]).

We now discuss in more detail how we partition the explained variance into components attributed to the row effects ($\alpha_i$) for *Model 1*, and the loadings ($\lambda_{qj}$) together with the linear predictor ($L_{ij}$) for *Model 2*. Let $V_{\text{total}}$ denote the total variance of the $\alpha_i$, while $V_{\text{sample}}$, $V_{\text{ecotype}}$, $V_{\text{site}}$ and $V_{\text{phylo}}$ denote the variances for the sample, host ecotype, host collection site and host phylogeny, respectively. Then for Case Study 1 we have,

$$
\begin{aligned}
V_{\text{total}} &= V_{\text{sample}} + V_{\text{ecotype}} + V_{\text{sites}} + V_{\text{phylo}}, \qquad \text{where} \\
V_{\text{sample}} &= \sigma^2(\text{sample}) \\
V_{\text{ecotype}} &= \sigma^2(\text{ecotype}) \\
V_{\text{site}} &= \sigma^2(\text{site}) \\
V_{\text{phylo}} &= \theta^2_{\text{phylo}},
\end{aligned}
$$

and for Case Study 2 we have,

$$
\begin{aligned}
V_{\text{total}} &= V_{\text{linpred}} + V_{\text{residual}} + V_{\text{sample}} + V_{\text{non-phylo}} + V_{\text{phylo}}, \text{where,} \\
V_{\text{linpred}_j} &= \text{var}(\text{Diet}_i \times \beta_{j1}) + \text{var}(\text{StomachContents}_i \times \beta_{j2}) + \text{var}(\text{Sex}_i \times \beta_{j3}) + \text{var}(\text{Elevation}_i \times \beta_{j4}) + \text{var}(\text{Site}_i \times \beta_{j5}) \\
V_{\text{residual}} &= diag(\mathbf{\Omega}) \\
V_{\text{sample}} &= \sigma^2(\text{sample}) \\
V_{\text{non-phylo}} &= \sigma^2(\text{non-phylo}) \\
V_{\text{phylo}} &= \sigma^2(\text{phylo})
\end{aligned}
$$

In the second partitioning, the quantity $V_{\text{linpred}}$ represents the variance explained by the linear predictor $L_{ij}$, the $V_{\text{residual}}$ represents the residual variance not accounted for by the modeled predictor variables i.e., as explained by the diagonal elements of the residual covariance matrix $\Omega$, and finally the $V_{\text{sample}}$, $V_{\text{non-phylo}}$ and $V_{\text{phylo}}$ to variance attributed to the hierarchy present on the row effects $\alpha_{ij}$.

# Results

Below we present the main results for each case study. We used the 95% highest density interval (HDI) as a measure of statistical significance. That is, if a parameter or a pairwise parameter comparison excludes zero, then we conclude that the posterior probability of the difference being significantly different from zero exceeds 95%.

## Case study 1

We applied *Model 1* to data on sponge host-associated microbiota [29]. The fitted model revealed that more than 86% of the variation in relative abundance among samples could be attributed to processes operating on the host-species level (Table 1; Fig 2). More specifically, 57% of this variation was explained by host phylogenetic relatedness, even though the 95% HDI for the phylogenetic effects did not exclude zero for any of the host species. While this suggests the presence of a phylogenetic signal in one or more host traits affecting microbial abundance and/or occurrence, it also indicates that no particular host species or host species clade have a stronger signal than the rest. Easson and Thacker [29] used the Bloomberg's K statistic and found a significant signal of the host phylogeny on the inverse Simpson's index. This index measures the diversity of a community, but is strongly influenced by the relative abundance of its most common species ([46]). The authors specifically noted that host species *Aiolochroia crassa, Aplysina cauliformis*

and *Aplysina fulva* from the order Verongida, along with host *Erylus formosus* from the order Astrophorida had higher values of this index compared to the rest of the host species. Similarly, we found that the same four hosts harbored more abundant (Fig 2) and distinctively different microbiotas than the other host species (Fig 3). Pairwise comparisons of these four hosts showed that *A. crassa* harbored markedly different microbial composition compared to its two closest relatives *A. cauliformis* and *A. fulva* (Table S1; Table S2). These three hosts were nonetheless collected at the same site. The two species from the genus *Aplysina* on the other hand, harbored very similar microbiota composition to that of host *E. formosus* even if they were collected some 17,000 km apart.

Host ecotype and collection site roughly explained two thirds of the remaining variation in relative abundance (Table 1). Furthermore, the host species level explained 39% of the variation beyond differences in relative abundance, with the remaining variation explained by the latent factors on the sample level. While samples did not cluster based on ecotype or sites, samples belonging to HMA hosts generally formed tighter clusters compared to samples from LMA hosts (Fig S3). Note however that because the sampling scheme in the original study confounded host ecotype and collection site, it is impossible to fully disentangle the two.

## Case study 2

Fitting *Model 2* to the data on neotropical bird gut-associated microbiota [31] revealed that only 9% of the variation in species richness among samples could be explained by processes acting on the host species level, including processes related to the host phylogeny. The remaining 91% of this variation was captured by processes operating on the sample level (Table 2). Of the total variance in species occurrence, variation in species richness only accounted for, on average, about 17%. The modeled predictor variables explained 69% of the total variance, and varied from a minimum of less than 0.01% to a maximum of 99.7% across all OTUs (Fig 5). The predictor variable that had the largest average effect on microbiota composition was collection site (21.33%, Table 2). None of the estimated regression coefficients for the predictor variables excluded zero (Fig S4). Furthermore, the ordination plots constructed from the the first two latent factors did not reveal any obvious clustering by e.g., host taxonomy (at the order level), collection site, or diet (broad dietary specialization) (Fig 6; Fig S5; Fig S6).

We ran an edge betweeness community detection algorithm [47] on the correlation matrix computed from the loading matrix $\lambda$ where links represent positive and negative co-occurrences with at least 95% posterior probability. We colored nodes by their bacterial taxonomic affiliation at the phylum level. This revealed a large tightly knit cluster with well connected nodes in the centre and less connected nodes in the periphery of the cluster. The network displayed equal proportion of positive and negative co-occurrences, and with no apparent clustering of OTUs belonging to certain

phyla (Fig S7). Caution should, however, be taken when interpreting statistical interactions: these are residual species-to-species co-occurrences that can only be considered as hypotheses for ecological interactions, and without additional biological information it is impossible to definitively confirm or assess their nature ([19, 48, 49]).

# Discussion

In this paper, we have developed a joint species distribution model (JSDM) aimed towards analyzing host-associated microbiota data. The present work builds upon and extends existing JSDMs by specifically targeting the hierarchical structure implicit in host-associated microbiota studies, while also including several other features that are attractive for analyzing such data. First, we have shown how overdispersed counts and presence-absence data, two common features of host-microbiota data can be modeled under a single framework by implementing a negative binomial and a probit distribution with the appropriate link function. Furthermore, we have utilized recent progress in latent factor modeling in order to represent the high-dimensional nature of host-microbiota data as a rank-reduced covariance matrix, thus making the estimation of large OTU-to-OTU covariance matrices computationally tractable. By doing so, we have also demonstrated how latent factors, both alone or together with measured covariates, can be used for variance partitioning and further visualized as ordinations and co-occurrence networks. Lastly, depending on the modelled response function, we have illustrated that the variance partitioning of the hierarchy present on the rows can be represented in terms of either relative abundance or species richness.

We adapted our proposed model to make use of two published data sets on host-associated microbiota. Although our goal was not to compare the results from these two case studies, such a systematic comparison can be done using a model-based approach like ours. Broadly, the data analyzed here suggest that markedly different processes are shaping the microbiota harbored by these different host organisms. Individually, the main results from each of our two models were generally in agreement with the results reported in their respective original study; for example, *Model 1* identified the same four host species reported by Easson and Thacker [29] to have more abundant and distinctively different microbiotas compared to the other analyzed hosts. Similarly to Hird et al. [31], the ordinations produced by *Model 2* did not cluster by host diet, host taxonomy nor collection site. By partitioning variance among fixed and random effects, *Model 2* further showed that there was substantial variation across OTUs in terms of which predictor variables explained the most variance.

While distance-based methods such as PERMANOVA still remains one of the most widely used non-parametric methods to analyze host-associated microbiota data, model-based approaches are increasingly recognized to outper-

form such analyses (see e.g., [3, 17, 27]), and we see our proposed model as making a strong case for further empirical comparisons between distanced-based and model-based approaches to analyzing microbiota data.

There are a number of extensions one could make to the proposed model. Perhaps the most important of these stems from the growing recognition that high-throughput DNA sequencing produces compositional data, i.e., non-negative counts with an arbitrary sum imposed by the sequencing platform, which can produce spurious correlations if not properly accounted for (see e.g., [50, 51, 52]). Because of the log-link function used in *Model 1*, it is possible to parameterize this model and regard it in terms of compositional effects (see [53] and also noting the fact that the negative binomial distribution can itself be parameterized as a hierarchical Poisson model with Gamma distributed random effects), although for ease of estimation and interpretation we chose to adopt the standard negative binomial parameterization. This topic remains an area of active research, and there are currently several model-based methods (see e.g., [54, 55, 56, 57]) to infer co-occurrence networks, each with its own set of assumptions–it is not yet conclusive that any one of these methods outperforms the rest. Other model extensions and modifications can also be made in order to answer specific ecological questions of interest. For example, whether closely related host species harbor closely related microbes (i.e., host-microbiota phylogenetic congruence), or whether similarity among host-associated microbiota decreases as a function of increasing geographical distance or social connectance between hosts. Such questions may be answered for instance, by incorporating a phylogenetic effect acting on the columns of the response matrix, and by implementing a Gaussian process model that quantifies the degree of spatial and/or social autocorrelation between hosts, respectively. These two “flavors" of JSDMs and mixed models more generally have previously been considered in community ecology, both separately [58, 59, 60] and combined [8], although both computation and successful estimation and inference of all the model parameters remain a major issue especially with the high-dimensional nature of host-associated microbiota data. In summary, while substantial methodological advances have been made over the past few years in developing an extensive model framework for community ecological data, to date there exists no similar unifying framework for modeling host-associated microbiota which is directly tailored to the hierarchical and correlation structures present as well as questions of interest specific to such data. Our proposed model, which explicitly accounts for the host’s effect in structuring its microbiota, takes us closer to that goal.

## Acknowledgements

We thank Dr. Robert W. Thacker and Dr. Sarah Hird for sharing their data sets on the microbiota associated to marine sponges and neotropical bird species, respectively. We further thank three anonymous reviewers for providing constructive feedback and comments. J.R.B was supported by an FPI Fellowship from the Spanish Government (BES-

2011-049043). J.M.M. was supported by the French LabEx TULIP (ANR-10-LABX-41; ANR-11-IDEX-002-02), by the Region Midi-Pyrenees project (CNRS 121090) and by the FRAGCLIM Consolidator Grant, funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement number 726176)

## Code and Data Availability

All code and data are available on Open Science Framework (DOI:10.17605/OSF.IO/T9NXH) with a tutorial on how to fit the model and analyze the output. While we used JAGS to fit the model in this study, we have translated the model into Greta ([61]) which is an R style probabilistic language that scales better to large data sets.

# References

[1] Mark Vellend. "Conceptual Synthesis in Community Ecology". In: *The Quarterly Review of Biology* 85.2 (2010), pp. 183–206. DOI: 10.1086/652373.

[2] P. Legendre and L. Legendre. *Numerical Ecology*. Developments in environmental modelling. Elsevier, 2012. ISBN: 9780444538680.

[3] David I. Warton, Stephen T. Wright, and Yi Wang. "Distance-based multivariate analyses confound location and dispersion effects". In: *Methods in Ecology and Evolution* 3.1 (2012), pp. 89–101. DOI: 10.1111/j.2041-210X.2011.00127.x.

[4] Laura J. Pollock et al. "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)". In: *Methods in Ecology and Evolution* 5.5 (2014), pp. 397–406. DOI: doi.org/10.1111/2041-210X.12180.

[5] Benjamin M. Bolker et al. "Generalized linear mixed models: a practical guide for ecology and evolution". In: *Trends in Ecology & Evolution* 24.3 (2009), pp. 127–135. ISSN: 0169-5347. DOI: 10.1016/j.tree.2008.10.008.

[6] John N. Thompson. *The coevolutionary process*. University of Chicago Press, 1994.

[7] Anthony R. Ives and Matthew R. Helmus. "Phylogenetic metrics of community similarity." In: *The American naturalist* 176.5 (2010), E128–E142. DOI: 10.1086/656486.

[8] Arne Kaldhusdal et al. "Spatio-phylogenetic multispecies distribution models". In: *Methods in Ecology and Evolution* 6.2 (2015), pp. 187–197. DOI: 10.1111/2041-210X.12318.

[9] Tuomas Aivelo, Anna Norberg, and Andy Fenton. "Parasite–microbiota interactions potentially affect intestinal communities in wild mammals". In: *Journal of Animal Ecology* 87.2 (), pp. 438–447. DOI: 10.1111/1365-2656.12708.

[10] Otso Ovaskainen et al. "How to make more out of community data? A conceptual framework and its implementation as models and software". In: *Ecology Letters* 20.5 (2017), pp. 561–576. DOI: 10.1111/ele.12757.

[11] Daijiang Li and Anthony R. Ives. "The statistical need to include phylogeny in traitbased analyses of community composition". In: *Methods in Ecology and Evolution* 8.10 (2017), pp. 1192–1199. DOI: 10.1111/2041-210X.12767.

[12] Fan Xia et al. "A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis". In: *Biometrics* 69.4 (2013), pp. 1053–1063. DOI: 10.1111/biom.12079.

[13] David I. Warton et al. "So Many Variables: Joint Modeling in Community Ecology". In: *Trends in Ecology and Evolution* 30 (2015), pp. 1–14. DOI: 10.1016/j.tree.2015.09.007.

[14] Andrew D. Letten et al. "Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models". In: *Journal of Ecology* 103.5 (2015), pp. 1264–1275. DOI: 10.1111/1365-2745.12428.

[15] Francis K.C. Hui. "boral–Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R". In: *Methods in Ecology and Evolution* 7.6 (2016), pp. 744–750. DOI: https://doi.org/10.1111/2041-210X.12514.

[16] Francis K.C. Hui. "Model-based simultaneous clustering and ordination of multivariate abundance data in ecology". In: *Computational Statistics & Data Analysis* 105 (2017), pp. 1–10. DOI: https://doi.org/10.1016/j.csda.2016.07.008.

[17] Francis K.C. Hui et al. "Model-based approaches to unconstrained ordination". In: *Methods in Ecology and Evolution* 6.4 (2015), pp. 399–411. DOI: https://doi.org/10.1111/2041-210X.12236.

[18] Michael B. Sohn and Hongzhe Li. "A GLM-based latent variable ordination method for microbiome samples". In: *Biometrics* (2017). DOI: 10.1111/biom.12775.

[19] Otso Ovaskainen et al. "Using latent variable models to identify large networks of species-to-species associations at different spatial scales". In: *Methods in Ecology and Evolution* 7.5 (2016), pp. 549–555. DOI: 10.1111/2041-210X.12501.

[20] A. Bhattacharya and D. B. Dunson. "Sparse Bayesian infinite factor models". In: *Biometrika* (2011), pp. 291–306. DOI: 10.1093/biomet/asr013.

[21] Miklós Bálint et al. "Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes". In: *FEMS Microbiology Reviews* 40.5 (2016), p. 686. DOI: 10.1093/femsre/fuw017.

[22] Roeland L. Berendsen, Corné M.J. Pieterse, and Peter A.H.M. Bakker. "The rhizosphere microbiome and plant health". In: *Trends in Plant Science* 17.8 (2012), pp. 478 –486. DOI: 10.1016/j.tplants.2012.04.001f.

[23] Margaret McFall-Ngai et al. "Animals in a bacterial world, a new imperative for the life sciences". In: *Proceedings of the National Academy of Sciences* 110.9 (2013), pp. 3229–3236. DOI: doi.org/10.1073/pnas.1218525110.

[24] Mathieu Groussin et al. "Unraveling the processes shaping mammalian gut microbiomes over evolutionary time". In: *Nature Communications* 8 (2017). DOI: 10.1038/ncomms14319.

[25] Shirong Liu et al. "The Host Shapes the Gut Microbiota via Fecal MicroRNA". In: *Cell Host & Microbe* 19.1 (2016), pp. 32–43. DOI: 10.1016/j.chom.2015.12.005.

[26] Lizhen Xu, Andrew D. Paterson, and Wei Xu. "Bayesian latent variable models for hierarchical clustered count outcomes with repeated measures in microbiome studies". In: *Genetic Epidemiology* 41.3 (2017), pp. 221–232. DOI: 10.1002/gepi.22031.

[27] Neal S. Grantham et al. *MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments*. 2017. eprint: arXiv:1703.07747.

[28] Xinyan Zhang et al. "Negative binomial mixed models for analyzing microbiome count data". In: *BMC Bioinformatics* 18.1 (2017), p. 4. DOI: 10.1186/s12859-016-1441-7.

[29] Cole G. Easson and Robert W. Thacker. "Phylogenetic signal in the community structure of host-specific microbiomes of tropical marine sponges". In: *Frontiers in Microbiology* 5 (2014), p. 532. DOI: 10.3389/fmicb.2014.00532.

[30] Volker Gloeckner et al. “The HMA-LMA Dichotomy Revisited: an Electron Microscopical Survey of 56 Sponge Species.” In: *The Biological bulletin* 227.1 (2014), pp. 78–88. DOI: 10.1086/BBLv227n1p78.

[31] Sarah M. Hird et al. “Comparative Gut Microbiota of 59 Neotropical Bird Species”. In: *Frontiers in Microbiology* 6 (2015), p. 1403. DOI: 10.3389/fmicb.2015.01403.

[32] Brian D. Muegge et al. “Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans”. In: *Science* 332.6032 (2011), pp. 970–974. DOI: 10.1126/science.1198719.

[33] Tanya Yatsunenko et al. “Human gut microbiome viewed across age and geography”. In: *Nature* 486.7402 (2012), pp. 222–227. DOI: 10.1038/nature11053.

[34] Anders Skrondal and Sophia Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, 2004. ISBN: 9780203489437. URL: https://books.google.com/books?id=YUpDqCzb-WMC.

[35] John Geweke and Guofu Zhou. “Measuring the price of the Arbitrage Pricing Theory”. In: 9.2 (1996), pp. 557–587. URL: http://www.jstor.org/stable/2962214.

[36] Nicholas G. Polson and James G. Scott. “On the Half-Cauchy Prior for a Global Scale Parameter”. In: *Bayesian Analysis* 7.4 (Dec. 2012), pp. 887–902. DOI: 10.1214/12-BA730.

[37] Andrew Gelman et al. “A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models”. In: *The Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383. DOI: 10.1214/08-AOAS191.

[38] Joseph Felsenstein. “Phylogenies and the Comparative Method”. In: *The American Naturalist* 125.1 (1985), pp. 1–15.

[39] Martyn Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. 2003.

[40] Matt Denwood. “runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS”. In: *Journal of Statistical Software* (2016). DOI: 10.18637/jss.v071.i09.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2016.

[42] Martyn Plummer et al. *CODA: Convergence Diagnosis and Output Analysis for MCMC*. 2016. URL: https://cran.r-project.org/web/packages/coda/.

[43] Curtis S. McKay. *Create Plots from MCMC Output*. 2015. URL: https://cran.r-project.org/web/packages/mcmcplots/.

[44] John F. Geweke. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Clarendon Press, Oxford, UK, 1991.

[45] Shinichi Nakagawa and Holger Schielzeth. "A general and simple method for obtaining R2 from generalized linear mixed-effects models". In: *Methods in Ecology and Evolution* 4.2 (2013), pp. 133–142. DOI: 10.1111/j.2041-210x.2012.00261.x.

[46] Bart Haegeman et al. "Only Simpson Diversity can be Estimated Accurately from Microbial Community Fingerprints". In: *Microbial Ecology* 68.2 (2014), pp. 169–172. DOI: 10.1007/s00248-014-0394-5.

[47] Gabor Csardi and Tamas Nepusz. "The igraph software package for complex network research". In: *InterJournal* Complex Systems (2006), p. 1695. URL: http://igraph.org.

[48] Gleb Tikhonov et al. "Using joint species distribution models for evaluating how speciestospecies associations depend on the environmental context". In: *Methods in Ecology and Evolution* 8.4 (2017), pp. 443–452. DOI: 10.1111/2041-210X.12723.

[49] Pollock L. J. Zurell D. and W. Thuiller. "Do joint species distribution models reliably detect interspecific interactions from co occurrence data in homogenous environments?" In: *Ecography* (). DOI: 10.1111/ecog.03315.

[50] Hongzhe Li. "Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis". In: *Annual Review of Statistics and Its Application* 2.1 (2015), pp. 73–94. DOI: 10.1146/annurev-statistics-010814-020351.

[51] Matthew C.B. Tsilimigras and Anthony A. Fodor. "Compositional data analysis of the microbiome: fundamentals, tools, and challenges". In: *Annals of Epidemiology* 26.5 (2016), pp. 330–335. DOI: 10.1016/j.annepidem.2016.03.002.

[52] Gregory B. Gloor et al. "Microbiome Datasets Are Compositional: And This Is Not Optional". In: *Frontiers in Microbiology* 8 (2017), p. 2224. DOI: 10.3389/fmicb.2017.02224.

[53] David I. Warton and Peter Guttorp. "Compositional analysis of overdispersed counts using generalized estimating equations". In: *Environmental and Ecological Statistics* 18.3 (2011), pp. 427–446. DOI: 10.1007/s10651-010-0145-9.

[54] Jonathan Friedman and Eric J. Alm. "Inferring Correlation Networks from Genomic Survey Data". In: *PLOS Computational Biology* 8.9 (Sept. 2012), pp. 1–11. DOI: 10.1371/journal.pcbi.1002687.

[55] Huaying Fang et al. "CCLasso: correlation inference for compositional data through Lasso". In: *Bioinformatics* 31.19 (2015), pp. 3172–3180. DOI: 10.1093/bioinformatics/btv349.

[56] Zachary D. Kurtz et al. "Sparse and Compositionally Robust Inference of Microbial Ecological Networks". In: *PLOS Computational Biology* 11.5 (May 2015), pp. 1–25. DOI: 10.1371/journal.pcbi.1004226.

[57] Emma Schwager et al. "A Bayesian method for detecting pairwise associations in compositional data". In: *PLOS Computational Biology* 13.11 (Nov. 2017), pp. 1–21. DOI: 10.1371/journal.pcbi.1005852.

[58] Anthony R. Ives and Matthew R. Helmus. "Generalized linear mixed models for phylogenetic analyses of community structure". In: *Ecological Monographs* 81.3 (2011), pp. 511–525. DOI: 10.1890/10-1264.1.

[59] Otso Ovaskainen et al. "Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models". In: *Methods in Ecology and Evolution* (2015). DOI: https://doi.org/10.1111/2041-210X.12502.

[60] James T. Thorson et al. "Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range". In: *Methods in Ecology and Evolution* (2015). DOI: https://doi.org/10.1111/2041-210X.12359.

[61] Nick Golding. *greta: Simple and Scalable Statistical Modelling in R*. 2018. URL: https://cran.r-project.org/web/packages/greta/index.html.

Table 1: Variation explained by the hierarchy present on $alpha_i$, i.e., the host effects $\mu(\text{host})_s$.

| | |
|---|---|
| Phylogeny | 57.09% |
| Ecotype | 14.58% |
| Site | 14.51% |
| Sample | 13.82% |

Table 2: Variation attributed to the linear predictor $L_{ij}$, the residual variation captured the diagonal elements of the residual covariance matrix $\mathbf{\Omega}$, and by the hierarchy present on the row effects $\alpha_{ij}$, i.e., the host effects $\mu(\text{host})_s$.

| | |
|---|---|
| Collection site | 21.33% |
| Stomach contents | 16.13% |
| Elevation | 15.97% |
| Diet | 13.59% |
| Sex | 2.12% |
| Residuals | 13.89% |
| Sample | 15.5% |
| Non-Phylogeny | 0.65% |
| Phylogeny | 0.82% |

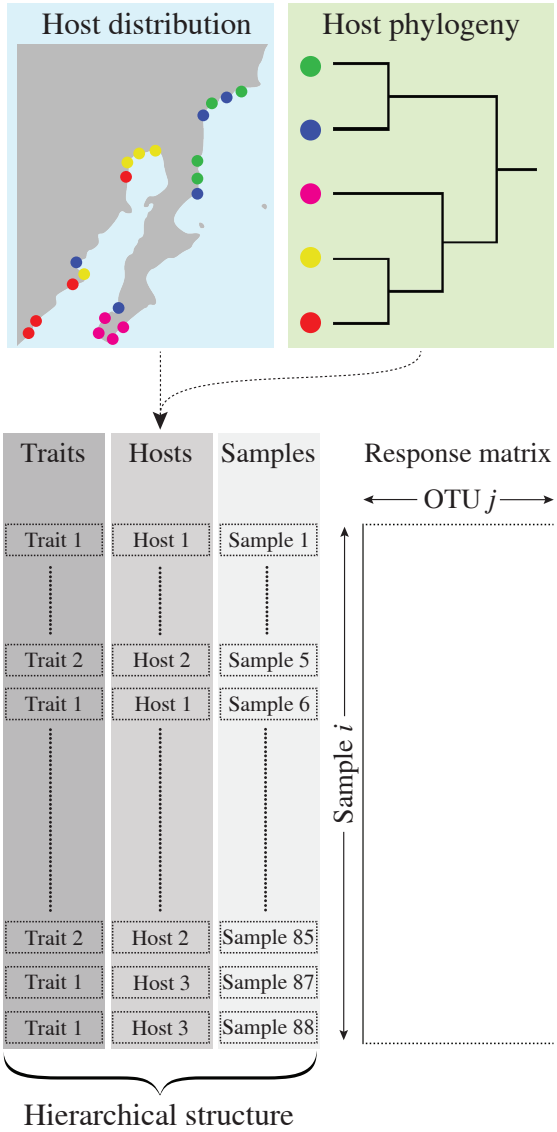


Figure 1: Host-associated microbiota data have a hierarchical data structure. In this example, samples are nested within host species which in turn are nested under species traits. As there are also data on the host's geographical distribution, host species can be further nested within observation/collection sites. Additional data that are often available is the host species phylogeny. The proposed model extension can straightforwardly accommodate for this hierarchical data structure and discriminate their importance in structuring the microbiota.

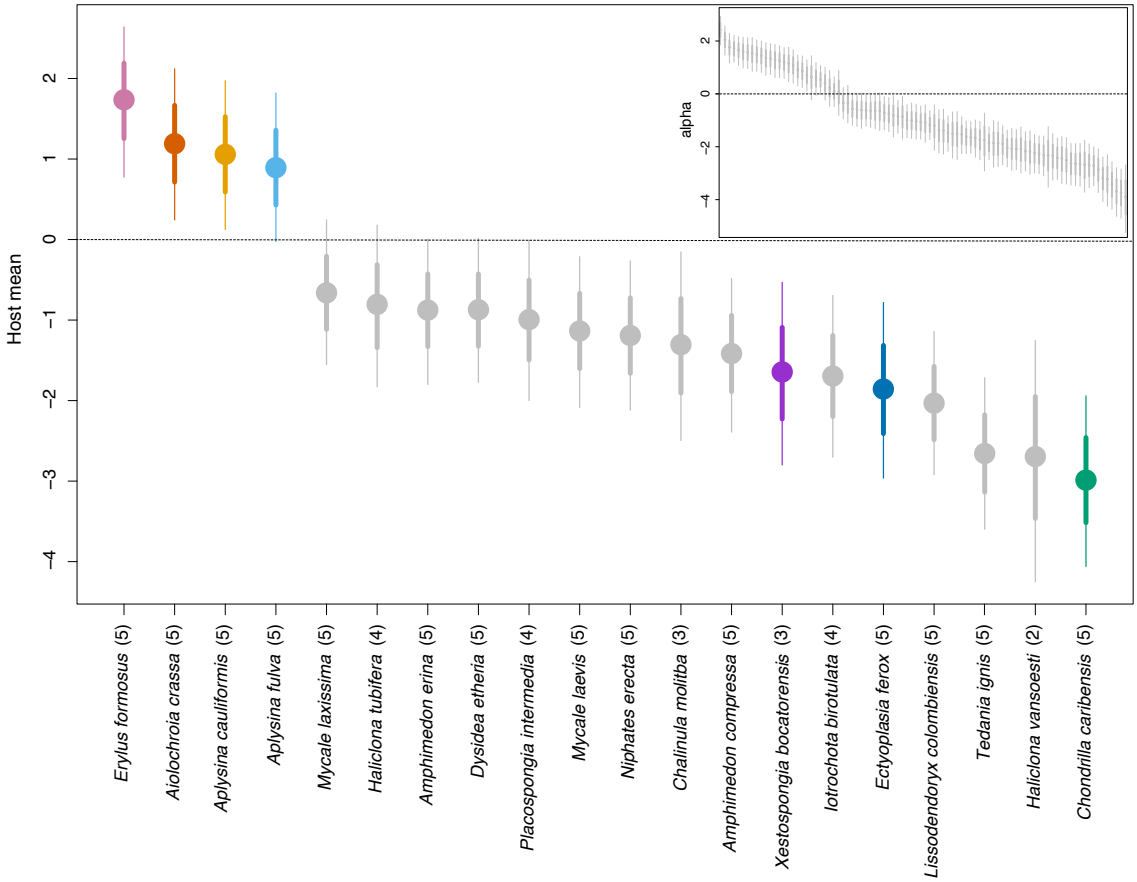


Figure 2: The main plot shows a caterpillar for the host means $\mu(\text{host})_s$, with the colors representing the 7 HMA hosts. The subplot shows a caterpillar plot for the row effects $alpha_i$. The quantiles corresponds to the 95% (thin lines) and 68% (thick lines) credible intervals. The number within the parentheses indicates how many individuals per host species were used to draw inference on.

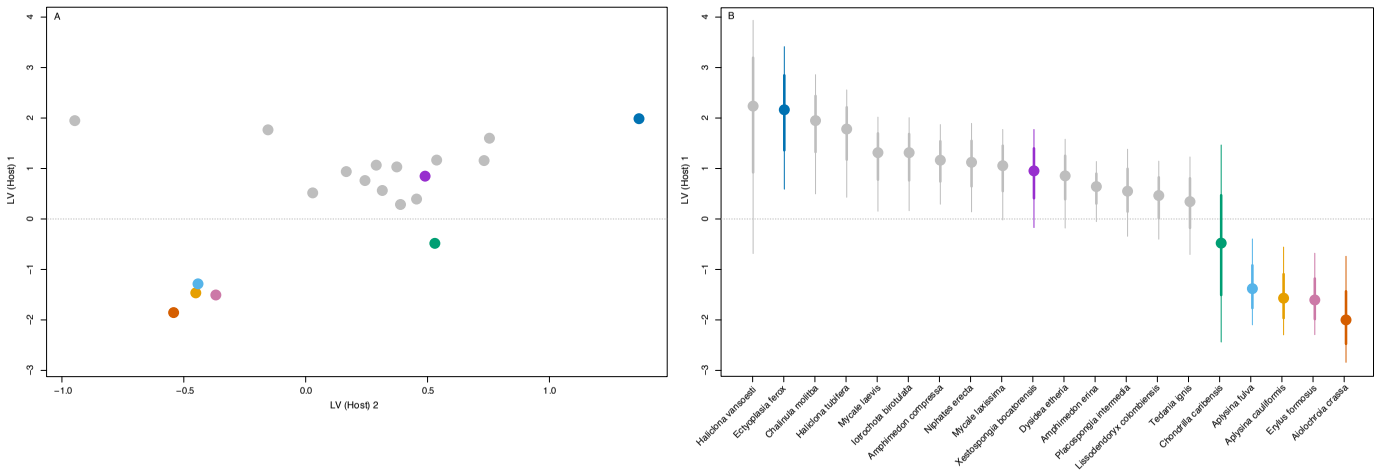


Figure 3: The left plot (A) shows the ordination constructed by the latent factors on the host species level $\boldsymbol{Z}^{\boldsymbol{H}}$, and the right plot (B) shows the corresponding caterpillar for first latent factor $Z_{i1}^{H}$. The quantiles corresponds to the 95% (thin lines) and 68% (thick lines) credible intervals.

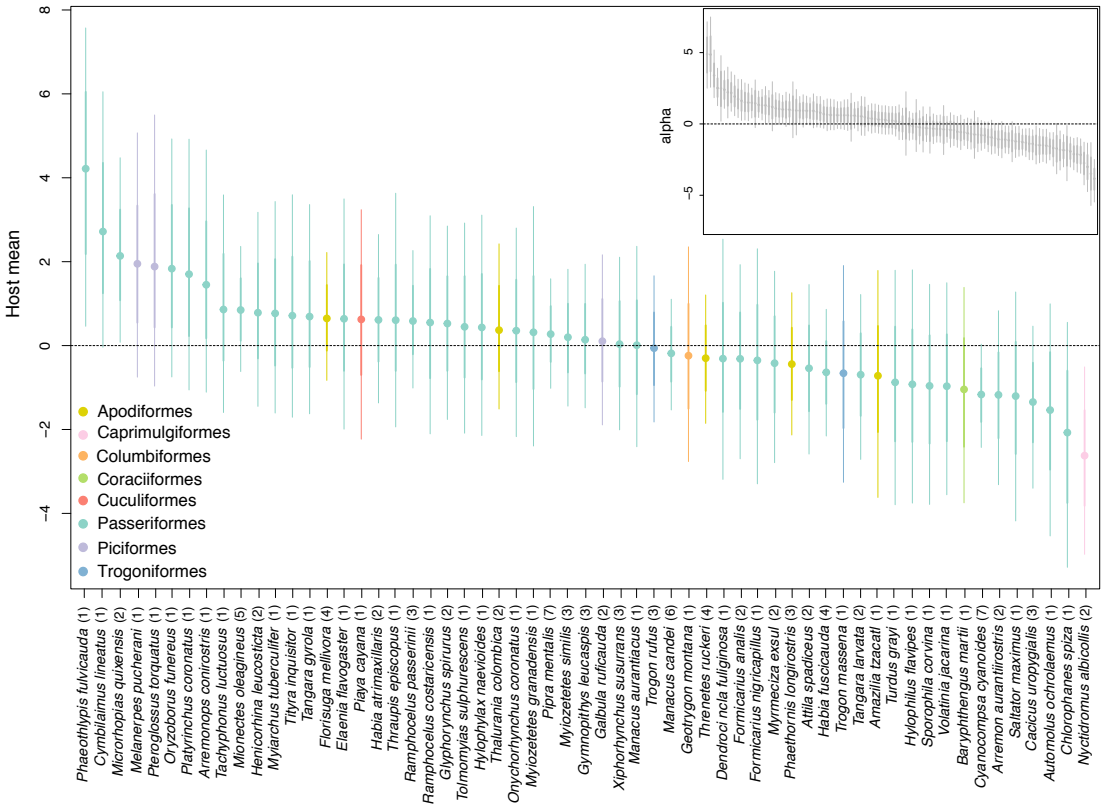


Figure 4: The main plot shows a caterpillar for the host means $\mu(\text{host})_s$ colored by host taxonomy at the order level, while the subplot shows a caterpillar plot for the row effects $alpha_i$. The quantiles corresponds to the 95% (thin lines) and 68% (thick lines) credible intervals. The number within the parentheses indicates how many individuals per host species were used to draw inference on.

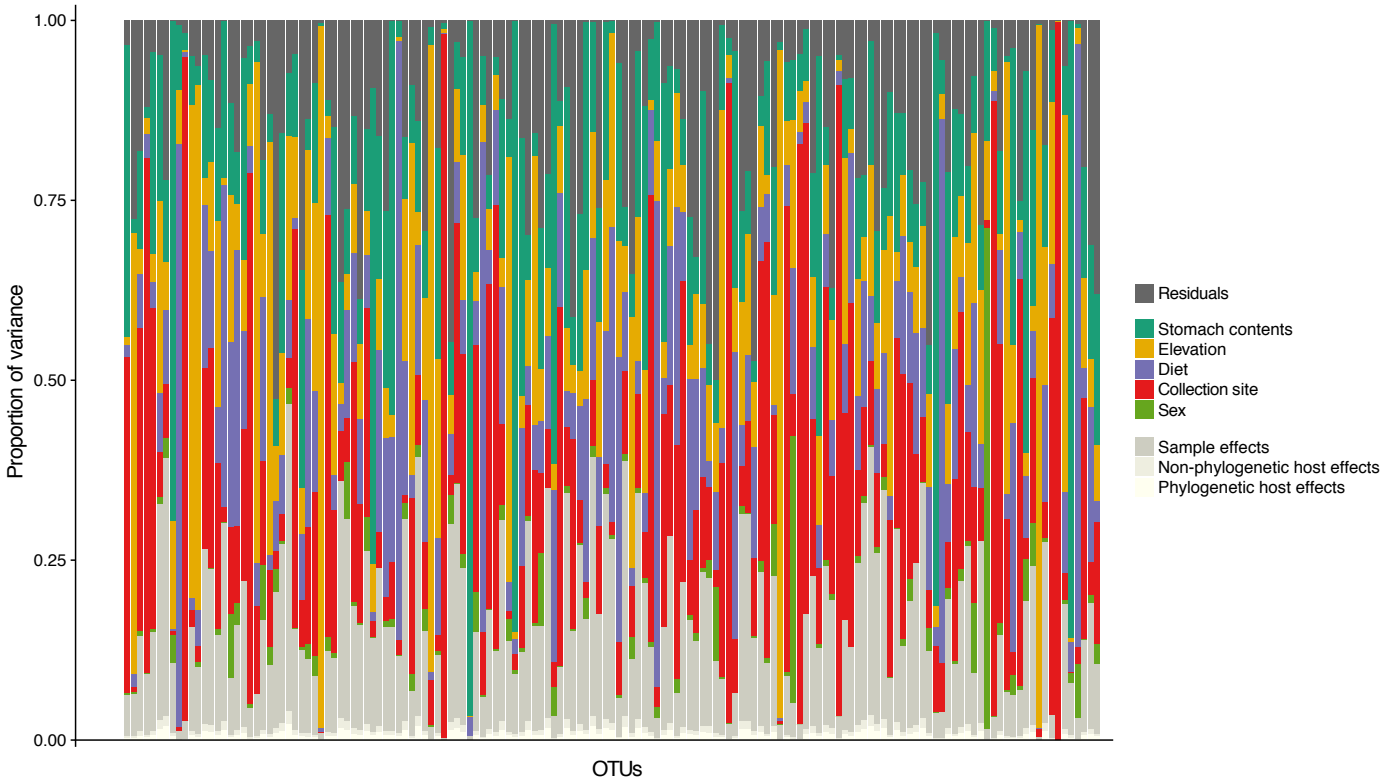


Figure 5: The y-axis shows the relative proportion of variance in species occurrences explained by the hierarchy present on $alpha_i$, the covariates included on the linear predictor $L_{ij}$, and the residual variance not accounted for by the modeled effects i.e., the diagonal elements of the residual covariance matrix $\mathbf{\Omega}$, for each OTU (x-axis).

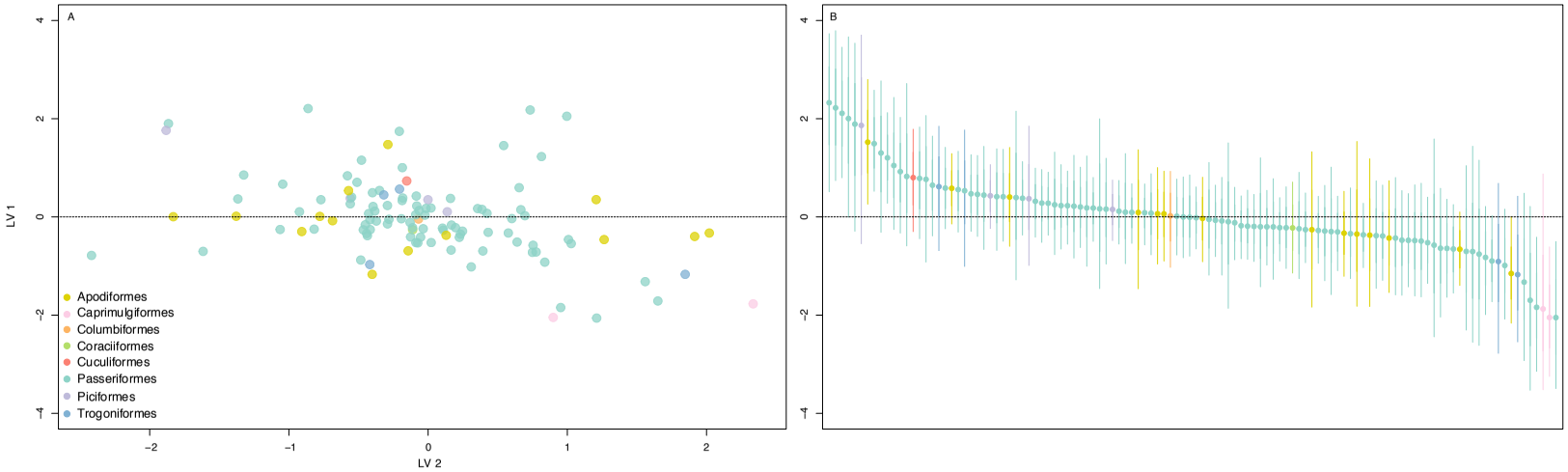


Figure 6: The left plot (A) shows the ordination constructed by the latent factors $\boldsymbol{Z}$ colored by host taxonomy (at the order level), and the right plot (B) shows the corresponding caterpillar for first latent factor $Z_{i1}$. The quantiles corresponds to the 95% (thin lines) and 68% (thick lines) credible intervals.