



**HAL**  
open science

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

Johannes R Börk, Francis K C Hui, Robert B O'Hara, José Montoya

► **To cite this version:**

Johannes R Börk, Francis K C Hui, Robert B O'Hara, José Montoya. Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, 2018, 27 (12), pp.2714-2724. 10.1111/mec.14718 . hal-02404692

**HAL Id: hal-02404692**

**<https://hal.science/hal-02404692>**

Submitted on 17 Nov 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## 14 **Abstract**

15 **Background** In addition to the processes structuring free-living communities, host-associated microbial communi-  
16 ties (i.e., microbiotas) are directly or indirectly shaped by the host. Therefore, microbiota data have a hierarchical  
17 structure where samples are nested under one or several variables representing host-specific features. In addition, mi-  
18 crobiota data are often collected across multiple levels of biological organization. Current statistical methods do not  
19 accommodate this hierarchical data structure, and therefore cannot explicitly account for the effects of host-specific  
20 features on structuring the microbiota.

21 **Methods** We introduce a unifying model-based framework developed specifically for analyzing host-microbiota  
22 data spanning multiple levels of biological organization. While we chose to discern among the effects of host species  
23 identity, host phylogeny, and host traits in structuring the microbiota, the presented framework can straightforwardly  
24 accommodate any recorded data that includes host-specific features. Other key components of our modeling frame-  
25 work are the powerful yet familiar outputs: (i) model-based ordination to visualize the main patterns in the data,  
26 (ii) co-occurrence networks to visualize microbe-to-microbe associations, and (iii) variance partitioning to assess the  
27 explanatory power of the included host-specific features and how influential these are in structuring the microbiota.

28 **Results** The developed framework was applied to published data on marine sponge-microbiota. We found that a  
29 series of host traits that are likely phylogenetically conserved underpinned differences in both abundance and species  
30 richness among sites. When controlling for these differences, microbiota composition among sites was confounded by  
31 numerous site and host-specific features. At the host level, host traits always emerged as the prominent host-specific  
32 feature structuring the microbiota.

33 **Conclusions** The proposed framework can readily be applied to a wide range of microbiota systems spanning mul-  
34 tiple levels of biological organization, allowing researchers to systematically tease apart the relative importance of  
35 recorded and/or measured host-specific features in structuring the microbiota. The study of free-living species com-  
36 munities have significantly benefited from the increase in model-based approaches. We believe that it is time for  
37 research on host-microbiota to leverage the strengths of a unifying model-based framework.

## 38 Introduction

39 Ecological communities are the product of both stochastic and deterministic processes. While environmental factors  
40 may set the upper bound on carrying capacity, competitive and facilitative interactions within and among taxa deter-  
41 mine the identity of the species present in local communities. Ecologists are often interested in inferring ecological  
42 processes from patterns and determining their relative importance for the community under study ([39]). During the  
43 last few years, there has been a growing interest in developing new statistical methods aimed toward ecologists and the  
44 analysis of multivariate community data (see e.g., [17] and references within). There are many metrics for analyzing  
45 such data, however, these have a number of drawbacks, including uncertainty of selecting the most appropriate null  
46 models/randomization tests, low statistical power, and the lack of possibilities for making predictions. One framework  
47 which has become increasingly popular in ecology is joint species distribution models (JSDMs,[28, 40, 25]). JSDMs  
48 are an extension of generalized linear mixed models (GLMMs, [3]) where multiple species are analyzed simultane-  
49 ously, with or without measured environmental data, revealing community-level responses to environmental change.  
50 Because JSDMs are an extension of GLMMs, they can partition variance among fixed and random effects to assess the  
51 relative contribution of different ecological processes, such as habitat filtering, biotic interactions and environmental  
52 variability ([25]). Also, with the increase of trait-based and phylogenetic data in community ecology, together with the  
53 growing appreciation that species interactions are constrained by the “phylogenetic baggage” they inherit from their  
54 ancestors ([34]), this type of models can further accommodate information on both species traits and phylogenetic  
55 relatedness among species ([14, 15, 1, 25]). As such, JSDMs represents a rigorous statistical framework which allows  
56 ecologists to gain a more mechanistic view of the processes structuring ecological communities ([40]).

57 In parallel to recent developments in community ecology, there is the growing field of microbial ecology studying  
58 both free-living and host-associated communities (i.e., microbiotas). While microbial ecologists can adapt many of the  
59 new statistical approaches developed for traditional multivariate abundance data (see e.g., [4]), researchers studying  
60 microbiotas need to consider an additional layer of processes structuring the focal community: microbiotas are also  
61 shaped directly or indirectly by their hosts. Interactions between hosts and microbes often involve long-lasting and  
62 sometimes extremely intimate relationships where the host animal may have evolved a capacity to directly control the  
63 identity and/or abundance of its microbial symbionts ([21]). Similarly to an environmental niche, host-specific features  
64 can be viewed as a multidimensional composite of all the host-specific factors governing microbial abundances and/or  
65 occurrences within a host. These may represent everything from broad evolutionary relationships among host species  
66 ([11]) to distinct ecological processes, such as the production of specific biomolecules within a single host species

67 ([18]). Furthermore, microbiotas often encompass multiple levels of biological organization, as e.g., samples may  
68 be collected from different body sites on numerous host individuals, and/or from different host species across larger  
69 spatial scales. At each level of biological organization, a different set of processes are likely to be influencing the  
70 microbiota.

71 While a few recent JSDMs have been applied to microbiota data ([1, 5, 36, 44]), none of these models explicitly and  
72 transparently account for the aforementioned host-specific features. This extra layer of processes creates a hierarchical  
73 data structure where samples are nested under one or several nominal variables representing recorded and/or measured  
74 host-specific features. On the other hand, as JSDMs are naturally multi-levelled, they can easily account for such a  
75 hierarchical data structure, including the hierarchy implicit in data spanning multiple levels of biological organization  
76 ([24, 20]). An example of such a data set is the gut microbiota of the Amboselli baboons (see e.g., [37]), where  
77 individual baboons are raised in matriarchal family groups which are part of larger social groups. Individuals may  
78 disperse from their family groups to other social groups when reaching adulthood. Individual baboons are therefore  
79 nested within both family and social groups, and researchers may want to investigate what processes acting on which  
80 social level of organization are most likely governed the gut microbiota.

## 81 **Discerning among processes through joint distribution models**

82 How processes related to host-specific features structure the microbiota are largely unknown. At the same time, to  
83 analyze such data requires a unifying, model-based framework capable of discerning amongst various host-specific  
84 features spanning multiple levels of biological organization. To fill this gap, we propose a novel JSDM framework  
85 specifically aimed at analyzing microbiota data which explicitly accounts for host-specific features across multiple  
86 levels of biological organization. Other key components of our proposed modeling framework include: (i) model-  
87 based ordination to visualize the main patterns in the data (ii) co-occurrence networks to visualize microbe-to-microbe  
88 associations, and (iii) variance partitioning to assess the explanatory power of the included host-specific features and  
89 their influence in structuring the microbiota (Figure 2). While our models can discern among the effects of host  
90 species identity, host phylogeny and host traits, they can straightforwardly accommodate any recorded and/or measured  
91 data on host-specific features. However, information on host phylogenetic relatedness and host traits are particularly  
92 useful in order to disentangle whether the microbiota under study is non-randomly structured among the branches of  
93 a host phylogeny such that related host species harbor more similar microbes (i.e., indicating vertical transmission) or  
94 whether the microbiota is non-randomly structured among environments reflecting different host traits (i.e., indicating  
95 horizontal transmission).

96 By applying our developed modeling framework to sponge-microbiota data, we set out to investigate a set of  
97 fundamental, but non-mutually exclusive questions of interest. Broadly, we are interested in whether the sponge  
98 microbiota are governed by processes at the site and/or host species level. More specifically we ask whether the  
99 microbiota associated with: (i) the same host species and/or (ii) phylogenetically closely related host species and/or  
100 (iii) host species with similar traits, are more similar irrespective of the spatial distance between the sites where they  
101 were collected. We also investigate whether host species in closely located areas harbor more similar microbiotas  
102 than host species collected in sites farther apart. Finally, we generate microbe-to-microbe association networks using  
103 our proposed framework, but acknowledge that we do not have any *a-priori* hypotheses regarding which microbes are  
104 more or less likely to be co-occur. To our knowledge, this is the first unifying model-based framework specifically  
105 developed for analyzing host-microbiota.

## 106 **Materials and methods**

### 107 **Sponge microbiota as a case study**

108 To illustrate our modeling framework, we acquired data on marine sponge-microbiota from different host species  
109 collected at different geographic sites across the globe (Figure 1, Table S1). As marine sponges are commonly divided  
110 into two groups reflecting a suite of morphological and physiological traits—coined *High* and *Low Microbial Abundance*  
111 (HMA/LMA) sponges—collection sites are nested within host species which are further nested within one of the two  
112 traits. While the HMA-LMA division in a strict sense refers to the abundance of microbes harbored by the host, HMA  
113 sponges have a denser interior, including narrower aquiferous canals and smaller choanocytes compared to LMA  
114 sponges whose architecture are more fitted for pumping large volumes of water ([38]). As a consequence, HMA and  
115 LMA sponges tend to harbor different microbiotas, with the latter often showing a higher similarity to the free-living  
116 microbial community present in the surrounding sea water ([2, 33]).

### 117 **Data compilation**

118 To assess variation in microbial abundances and co-occurrences across different sponges species collected at different  
119 sites, we compiled a data set of sponge-associated bacterial 16S rRNA gene clone-library sequences published in  
120 NCBI GenBank (<http://www.ncbi.nlm.nih.gov>) between September 2007 and August 2014. All sponge species in  
121 the data set were required to be present in at least two different collection sites and be associated with at least 10

122 different sequences per site. The final data set contained a total of 3874 nearly full-length 16S rRNA gene sequences  
123 from 9 HMA and 10 LMA sponge species collected at 48 different sites ( $n_{HMA}=28$ ,  $n_{LMA}=20$ ) across the Atlantic,  
124 Pacific Ocean, Mediterranean and Red Seas (Figure 1, Table S1). The 16S rRNA gene sequences were aligned and  
125 clustered into operational taxonomic units (OTUs) representing family-level (at 90% nucleotide similarity, [42, 32])  
126 using mothur v.1.32.1 ([31]). At higher and lower sequence similarities, OTU clusters tended to become either too  
127 narrow or too broad, generating too sparse data for our models. Finally, as clone-libraries do not circumvent the need  
128 for cultivation, the OTUs modelled here correspond to the most common members of the sponge-microbiota.

## 129 **Phylogenetic reconstructions**

130 We retrieved nearly full-length sponge 18S rRNA gene sequences published in NCBI GenBank (<http://www.ncbi.nlm.nih.gov>)  
131 (see e.g., [10]). Sequences were aligned using the default options in ClustalW (1.83) ([16]). The phylogenetic relation-  
132 ship between the sponge species were reconstructed by implementing a HKY +  $\Gamma_4$  substitution model using BEAST  
133 (1.7.4) ([6]). For a few host species (*I. oros*, *H. simulans*, *M. methanophila* and *X. testudinaria*), the 18S rRNA gene  
134 sequence was unavailable. In these cases, we constrained the sponge species to the clade containing its genera.

135 A posterior distribution of phylogenies were sampled using Markov Chain Monte Carlo (MCMC) simulations as  
136 implemented in BEAST. We ran 4 independent chains each for 20 million generations saving every 4000<sup>th</sup> sample and  
137 discarding the first 25% as burn-in. This resulted in 20,000 generations from the posterior distribution. Convergence  
138 was evaluated using Tracer (v1.5) ([30]). We summarized the output of the four chains as a consensus phylogeny.  
139 Assumeing Brownian motion so that each covariance between host species  $i$  and host species  $j$  is proportional to their  
140 shared branch length from the most recent common ancestor ([7]), we used the variance-covariance matrix of the  
141 consensus phylogeny  $\Sigma(\text{phylo})$  as prior information in Equation 3, such that  $\mu(\text{phylo})_s \sim \mathcal{MVN}(\mathbf{0}, \Sigma(\text{phylo}))$ . Note  
142 that as the host species-specific variance i.e., the diagonal elements of the variance-covariance matrix is scaled to one  
143 by the construction of  $\Sigma(\text{phylo})$ , we multiplied it with a scaling factor  $\tau$  as seen in the formulation in (3).

## 144 **Joint species distribution models**

145 We developed a Bayesian joint species distribution modeling framework to jointly model the abundance and co-  
146 occurrence of OTUs across multiple sites, while also accounting for host species identity, host phylogenetic related-  
147 ness, and host traits (HMA and LMA, hereafter termed *ecotype*). Another important feature of the models we propose  
148 is the inclusion of latent factors, serving three main purposes. First, they allow for a parsimonious yet flexible way

149 of modeling correlations between a large number of taxa. That is, given the number of taxa recorded often has the  
150 same order or exceeds the number of sites, as is characteristic of most multivariate abundance data including the one  
151 analyzed here, modeling the covariation between all taxa using an unstructured correlation matrix is often unreliable  
152 due to the large number of elements in the matrix that need to be estimated ([40]). Using latent factors instead of-  
153 fers a more practical solution, via rank reduction, to model correlations in such high dimensional settings. Second,  
154 latent factors allow for performing model-based unconstrained and residual ordination in order to visualize the main  
155 patterns in the data ([12, 13]). While traditional distance-based ordination techniques easily confound location and  
156 dispersion effects ([41]), model-based ordination properly models the mean-variance relationship, and can therefore  
157 accurately detect differences between the two. Third, latent factors allow for inferring associative networks identified  
158 by correlations and partial correlations ([24]).

159 We considered two response types commonly encountered in ecology and biogeography; negative binomial re-  
160 gression for overdispersed counts and probit regression for presence-absence. As such, the response matrix being  
161 modelled consisted of either counts or presence-absence of  $n$  OTUs observed at  $m$  sites. The rows of the response  
162 matrix have a hierarchical structure typical for many microbiota data. Specifically, the  $m = 48$  sites are nested within  
163 the  $s = 19$  host species, with the 19 host species nested within one of  $r = 2$  ecotypes (Figure 2). Due to their lack of  
164 information, OTUs with less than 5 presences across sites and with a total abundance of less than 5 were removed,  
165 resulting in 65 modelled OTUs.

166 **NB model:** Due to the presence of overdispersion in the counts, a negative binomial distribution with a quadratic  
167 mean-variance relationship was assumed for the response matrix  $y_{ij}$ , such that  $\text{Var}(y_{ij}) = v_{ij} + \phi_j v_{ij}^2$  where  $\phi_j$  is the  
168 OTU-specific overdispersion parameter. The mean abundance was related to the covariates using a log link function.  
169 We denote the response and mean abundance of OTU  $j$  at site  $i$  by  $y_{ij}$  and  $v_{ij}$ , respectively.

170 **Probit model:** Presence ( $y_{ij} = 1$ ) or absence ( $y_{ij} = 0$ ) of OTU  $j$  at site  $i$  was modelled by a probit regression,  
171 implemented as  $y_{ij} = 1_{z_{ij} > 0}$  where the latent liability  $z_{ij}$  is a linear function of the covariates, including the probit link  
172 function.

173 Below, we present specifications for the negative binomial (NB) model only, as the probit model description is  
174 similar except the distribution assumed at the response level of the model (Equation S1).

Let  $\mathcal{N}(\mu, \sigma^2)$  denote a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and analogously, let  $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  denote  
a multivariate normal distribution with mean vector and covariance matrix  $\boldsymbol{\Sigma}$ . Then, we have the model formulation



as follows

$$y_{ij} \sim \text{Negative-Binomial}(v_{ij}, \phi_j); \quad i = 1, \dots, 48; \quad j = 1, \dots, 65 \quad (1)$$

$$\log(v_{ij}|z_i) = \alpha_i + \beta_j + \sum_{q=1}^2 Z_{iq}^S \lambda_{qj}^S + \sum_{q=1}^2 Z_{s[i]q}^H \lambda_{qj}^H; \quad q = 1, \dots, 2 \quad (2)$$

$$\beta_j \sim \text{Cauchy}(0, 2.5)$$

$$\alpha_i \sim \mathcal{N}(\mu_i, \sigma^2(\text{host}))$$

$$\mu_i = \mu(\text{host})_{s[r]} + \tau * \mu(\text{phylo})_s; \quad r = 1, 2; \quad s = 1, \dots, 19 \quad (3)$$

$$\mu(\text{host})_{s[r]} \sim \mathcal{N}(\mu(\text{ecotype})_r, \sigma^2(\text{ecotype}))$$

$$\mu(\text{ecotype})_r \sim \text{Cauchy}(0, 2.5)$$

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

175 To clarify the above formulation, the subscript  $r$  indexes ecotype,  $s$  indexes host species and  $i$  indexes sites, such that  
 176 “ $s[i]$ ” and “ $s[r]$ ” means “site  $i$  nested within host species  $s$ ” and “host species  $s$  nested within ecotype  $r$ ”, respectively.  
 177 In Equation (2), the quantities  $\alpha_i$  and  $\beta_j$  represent site and OTU-specific effects, respectively. The former adjusts for  
 178 differences in site total abundance (species richness in the probit case), whereas the latter controls for differences in  
 179 OTU total abundance (OTU prevalence across sites in the probit case). From a purely statistical point of view, this  
 180 can be thought of as a model-based analog of studying alpha and beta diversity, respectively. The inclusion of  $\alpha_i$   
 181 serves two main purposes. First and foremost, including  $\alpha_i$  allows us to account for the hierarchical structure of the  
 182 data and its effect on site total abundance (species richness in the probit case) specifically. In particular, to account  
 183 for site  $i$  being nested within host species  $s$  which in turn is nested within ecotype  $r$ , the site effects  $\alpha_i$ ’s are drawn  
 184 from a normal distribution with a mean that is a linear function of both a host-specific mean  $\mu(\text{host})_{s[r]}$  and a host-  
 185 specific phylogenetic effect  $\mu(\text{phylo})_s$  (Equation 3). Furthermore, the host effects themselves are drawn from a normal  
 186 distribution with a ecotype-specific mean  $\mu(\text{ecotype})_r$ . Second, it means the resulting ordinations constructed by the  
 187 latent factors at the site  $Z_{iq}^S$  and host species  $Z_{s[i]q}^H$  level are in terms of composition only, as opposed to a composite  
 188 of site total abundance (species richness in the probit case) and composition (i.e. microbiota structure) when site  
 189 effects are not included ([12]). In other words, by accounting for the hierarchical structure present in the data, the  
 190 model-based ordinations are able to distinguish between microbiota composition and structure. It also means that the  
 191 corresponding factor loadings  $\lambda_{qj}^S$  and  $\lambda_{qj}^H$  which quantify each OTU’s response to the latent factors and subsequently

192 the correlations among OTUs at the two different levels of biological organization are driven by OTU-specific effects  
193 only, as opposed to correlations additionally induced by site and host-specific features.

194 Note that, in contrast to the means  $\mu$ 's, the variance parameters  $\sigma^2(\text{host})$  and  $\sigma^2(\text{ecotype})$  are common across all  
195 hosts and ecotypes. This implies that, *a-priori*, hosts and ecotypes can differentiate in location (mean) but not in  
196 dispersion (variance). However, as we will see later in the Results section, the ordinations for hosts and ecotypes can  
197 still, *a-posteriori*, vary substantially in terms of location and dispersion. We fitted each model with and without site  
198 effects  $\alpha_i$  included, so that two types of ordinations and association networks were constructed. When site effects were  
199 included, the ordinations on both levels of biological organization are in terms of microbiota composition, whereas  
200 when site effects are not included, the ordinations represent microbiota structure. The inclusion of  $\alpha_i$  also allows us to  
201 discern among OTU-to-OTU correlations induced by OTU-specific effects from those induced by site and host-specific  
202 features. For the model without site effects  $\alpha_i$  included, its associated nested structure were removed from Equation  
203 (2), such that  $\log(v_{ij}|z_i) = \beta_j + \sum_{q=1}^2 Z_{iq}^S \lambda_{qj}^S + \sum_{q=1}^2 Z_{s[i]q}^H \lambda_{qj}^H$ . As is conventional with ordination, we set  $q = 2$  so that  
204 once fitted, the latent factors  $Z_{i,q} = (Z_{i1}, Z_{i2})$  were plotted on a scatter plot to visualize the main patterns in the data  
205 ([12]). From the corresponding factor loadings  $\lambda_{qj}$ , a variance-covariance matrix was computed as  $\Omega = \lambda_{1j}(\lambda_{2j})^T$ ,  
206 and subsequently converted to a correlation matrix and plotted as a OTU-to-OTU association network ([24]).

207 To complete the above formulation, we assigned priors to the appropriate hyperparameters. For the OTU-specific  
208 overdispersion parameters  $\phi_j$  (Equation 1), we chose to assign a weakly-informative Gamma prior,  $\text{Gamma}(0.1, 0.1)$ .  
209 The standard deviations for host  $\sigma(\text{host})$  and ecotype  $\sigma(\text{ecotype})$  in Equations (2)-(3) were assigned uniform priors  
210  $\text{Unif}(0, 30)$ . The latent factors in Equation (2) on the site  $Z_{iq}^S$  and host species  $Z_{iq}^H$  level were assigned normal priors  
211  $\mathcal{N}(0, 1)$ . The corresponding OTU-specific coefficients, i.e., the  $\lambda_{qj}^S$ 's and the  $\lambda_{qj}^H$ 's in Equation (2) were assigned  
212 Cauchy priors with center and scale parameters of 0 and 2.5, respectively, while taking to account the appropriate  
213 constraints for parameter identifiability (see citeHui2015, for details). The Cauchy distribution was used because it is  
214 good example of a weakly-informative normal prior ([9]). Finally, the phylogenetic scale parameter  $\tau$  was drawn from  
215 a weakly-informative exponential prior with a rate parameter of 0.1.

## 216 Variance partitioning

217 One of the main advantages of the differing levels in the hierarchy in Equations (1)-(3) is that we can calculate the  
218 total variance of the  $\mu_i$ 's and partition this variance into components reflecting variation in site total abundance (species  
219 richness in the probit case) attributable to differences in host species identity  $\mu(\text{host})_s$ , host phylogenetic relatedness  
220  $\mu(\text{phylo})_s$  and host traits  $\mu(\text{ecotype})_r$ . This means that we can assess the explanatory power of the host-specific features

221 and how influential each of them are in structuring the microbiota. Such a variance decomposition is analogous to  
222 sum-of-squares and variance decompositions seen in Analysis of Variance (ANOVA) and linear mixed models ([23]).

Let  $V_{\text{total}}$  denote the total variance of the  $\mu_i$ 's, while  $V_{\text{host}}$ ,  $V_{\text{phylo}}$  and  $V_{\text{ecotype}}$  denote the variances due to host species identity, host phylogeny and host ecotype, respectively. Then we have,

$$V_{\text{total}} = V_{\text{host}} + V_{\text{phylo}} + V_{\text{ecotype}} + (\mu(\text{ecotype})_{LMA} - \mu(\text{ecotype})_{HMA})^2, \quad \text{where} \quad (4)$$

$$V_{\text{ecotype}} = \sigma^2(\text{ecotype}) \quad (5)$$

$$V_{\text{host}} = \sigma^2(\text{host}) \quad (6)$$

$$V_{\text{phylo}} = \tau^2 \quad (7)$$

223 Where  $\sigma^2(\text{host})$  reflects the intraspecific variation among sites nested within host species with small values of  $V_{\text{host}}/V_{\text{total}}$   
224 implying that sites nested within the same host species are more similar within than between host species.  $\tau^2$  corre-  
225 sponds the intraspecific variation among sites nested within host species that can be attributed to hosts' phylogenetic  
226 relatedness, meaning that small values of  $V_{\text{phylo}}/V_{\text{total}}$  provide evidence that the host phylogeny has little influence  
227 on variation in site total abundance (species richness in probit case).  $\sigma^2(\text{ecotype})$  accounts for intraspecific variation  
228 among host species nested within the two ecotypes, whereas  $(\mu(\text{ecotype})_{LMA} - \mu(\text{ecotype})_{HMA})^2$  is the difference in  
229 variation between the two ecotypes. Therefore,  $(\mu(\text{ecotype})_{LMA} - \mu(\text{ecotype})_{HMA})^2/V_{\text{total}}$  represents the proportion of  
230 total variation in site total abundance (species richness in the probit case) driven by ecotype. That is, if the proportion  
231  $V_{\text{ecotype}}/V_{\text{total}}$  is small compared to  $(\mu(\text{ecotype})_{LMA} - \mu(\text{ecotype})_{HMA})^2/V_{\text{total}}$ , then host species' microbiota are more  
232 similar within rather than between ecotypes.

233 We used Markov Chain Monte Carlo (MCMC) simulation method by running JAGS ([26]) in R ([29]) through  
234 the *rjags* ([19]) package to sample from the joint posterior distribution of the model parameters. We ran 1 chain with  
235 dispersed initial values for 100,000 iterations saving every  $10^4$ h sample and discarding the first 50% of samples as  
236 burn-in. We evaluated convergence of model parameters by visually inspecting trace and density plots using the R  
237 packages *coda* ([27]) and *mcmcplots* ([22]).

## 238 Results

239 We did not observe any large qualitative differences between the negative binomial (NB) and probit models of our  
240 framework. As noted above, an interesting difference between the two models is the interpretation of the row and

241 column totals. Modeling counts means that row and column totals correspond to site and OTU total abundance,  
242 respectively, rather than species richness and OTU prevalence across sites as in the case of presence-absences. Even if  
243 the two are very similar, the latter has a more straightforward interpretation as alpha and beta diversity. We present the  
244 main results for both models below, but relegate figures associated to the probit model to the supplementary material.

245 At the site level, without adjusting for differences among sites (i.e. not including  $\alpha_i$ ), host ecotype appeared as the  
246 major host-specific feature driving differences in microbiota structure (Figure 3A-B, S2A-B). After adjusting for site  
247 effects, while simultaneously accounting for host species identity, host phylogenetic relatedness and host ecotype, sites  
248 clustered, i.e., they harbored similar microbiota composition, to a lesser extent by host ecotype (Figure 3C-D, S2C-  
249 D). The variance partitioning showed that differences among sites in terms of abundance and richness were largely  
250 driven by host phylogenetic relatedness (Figure 4, S3), suggesting that ecotype is phylogenetically conserved within  
251 *Porifera*. It also indicates that composition among sites, similarly to abundance and richness, is confounded by site  
252 and host-specific features, such as geographic distance, host species identity, host phylogenetic relatedness and host  
253 ecotype. For example, a few sites clustered by host species (e.g., HMA hosts *Aplysina cualiformis*, *Aplysina fluva*,  
254 *Ircinia felix*, and *Ircinia oros*), but at closer inspection, the geographic distance between several of these sites were  
255 low (Figure 3C, Figure S2C). At the host-species level, hosts always clustered according to ecotype, indicating that  
256 the set of traits encompassing HMA and LMA hosts are indeed important for structuring the microbiota (Figure 5A-B,  
257 S4A-B).

258 A closer look at  $\alpha_i$ , the parameter adjusting for site effects, showed that sites belonging to the same host species and  
259 sites belonging to either of the two host ecotypes often had similar posterior means, with HMA hosts typically having  
260 narrower credible intervals (Figure 6A, S5A). However, these differences were not present in the mean parameter  
261 of  $\alpha_i$ , i.e., the  $\mu(\text{host})_{s[r]}$  (Figure 6B, S5B), further indicating that microbiota composition, more than differences in  
262 abundance and richness, is driving the observed HMA-LMA dichotomy.

263 We did not find any distance-decay relationship where microbiota similarity among sites decrease with increasing  
264 geographic distance. However weak, we observed that HMA and LMA hosts had opposite slopes in the model not  
265 controlling for site effects, indicating that LMA microbiota may be more influenced by local environmental conditions  
266 (Figure S1, S6). Interestingly, for the NB model, the slope of LMA hosts switched sign in the model adjusted for site  
267 effects. (Figure S1).

268 We generated OTU-to-OTU association networks where links between OTUs represented either positive and nega-  
269 tive abundance correlations and co-occurrences with at least 95% posterior probability. On one hand, by not adjusting  
270 for site effects, correlations between OTUs are induced by not only OTU-specific effects, but also by site and host-

271 specific features. We found many more correlations in the model not controlling for site effects (Figure 7A-B, S7A-B)  
272 compared to the model that did (Figure 7C-D, S7C-D). The site level (Figure 7A-C, S7A-C) generally had more cor-  
273 relations compared to the host species level (Figure 7B-D, S7B-D). On the site level, correlations were likely induced  
274 by, in addition to host-specific features, site effects such as geographic distance, coexisting host species and/or sim-  
275 ilar environmental preferences among OTUs, whereas on the host species level, correlations were only induced by  
276 host-specific features. The probit model detected more correlations on both levels compared to the NB model (Fig-  
277 ure S7). This is likely due to the difference in the nature of the correlations, i.e., co-occurrences (probit model) versus  
278 abundance correlations (NB model).

## 279 Discussion

280 Discerning amongst the many processes structuring microbiotas is one of the big new challenges facing ecology  
281 and evolution. However, the complexity of these communities often preclude their understanding, and we currently  
282 lack a mechanistic view of the processes structuring these systems. Motivated by these challenges, we developed a  
283 joint species distribution modeling (JSDM) framework to enhance our understanding of how host-specific features  
284 influence and structure the microbiota, both in terms of the abundance/species richness and composition of microbes.  
285 The presented framework builds upon and extends existing JSDMs by specifically targeting the hierarchical structure  
286 typically characterizing microbiota data. For example, our framework can be seen as microbiota adapted phylogenetic  
287 generalized linear mixed models where we model host species traits and phylogenetic relatedness on the rows of the  
288 response matrix, as opposed to on the columns as seen in the typical specification of these models ([14]).

289 Whether host phylogeny and/or host traits structure the microbiota reveal important information about the under-  
290 lying processes. We found a strong phylogenetic signal on microbial abundance and species richness among hosts,  
291 but at the same time, we did not observe a clear clustering by host phylogeny. Instead, the sponge-microbiota always  
292 showed a strong clustering by host traits (i.e. HMA/LMA), indicating (1) that host traits may be phylogenetically  
293 conserved within *Porifera* and/or (2) that the microbiota may be adapted to the different host environments associated  
294 with the two traits. Traditional ordination methods, such as principal coordinate analysis (PCoA) and non-metric mul-  
295 tidimensional scaling (NMDS) does not allow for such a systematic dissection of the patterns and the likely processes  
296 structuring host-microbiotas.

297 Other advantages compared to traditional ordination methods are that model-based ordination is implemented and  
298 developed by directly accommodating the statistical properties of the data at hand ([12]). Failure to account for, e.g.,

299 the mean–variance relationship can lead to misleading results (see [41] for details and discussion). Another advantage  
300 of our modeling framework is that the constructed ordinations are able to distinguish between microbiota composition  
301 and structure. For instance, we found that on the host species level, ecotype (HMA/LMA) emerged as the major host-  
302 specific feature driving microbiota structure and composition, whereas on the site level, structure and composition  
303 was confounded by numerous factors. Furthermore, calculating total variance and partitioning this into components  
304 reflecting variation attributable to different host-specific features, such as host traits and phylogenetic relatedness,  
305 allows researchers to assess the relative importance of possible ecological processes.

306 It has become increasingly popular in microbial ecology to visualize OTU-to-OTU association networks from  
307 correlations (e.g. [8, 43]). A key feature of the presented framework is the use of latent factors as a parsimonious  
308 approach for modeling correlations between a large number of taxa. Beyond OTU-specific effects, such as e.g.,  
309 interspecific interactions, correlations amongst OTUs may be induced by site and/or host-specific features. Therefore,  
310 by modeling the microbiota on multiple levels of biological organization, while simultaneously controlling for site  
311 effects and its hierarchical structure (i.e. the host-specific features), it is possible to gain a better understanding of the  
312 possible interaction structures. However, as these associations are of correlative nature, they should not be regarded as  
313 ecological interactions, but merely as hypotheses of such ([24, 35]).

314 Finally, the presented framework can readily be applied to a wide range of microbiota systems spanning multiple  
315 levels of biological organization, where the main interest lies in teasing apart the relative importance among host-  
316 specific features in structuring the microbiota. It can further be adapted to accommodate additional information, such  
317 as e.g., phylogenetic relatedness among microbes, spatial distance between sites, and/or environmental covariates  
318 directly acting on the hosts. Such a flexible modeling framework offers many exciting avenues for methodological  
319 advancements that will help to enhance our understanding of the numerous processes structuring host-microbiotas.

## 320 **References**

- 321 [1] Tuomas Aivelo and Anna Norberg. Parasite-microbiota interactions potentially affect intestinal communities in  
322 wild mammals. *bioRxiv*, 2016.
- 323 [2] Johannes R. Björk, C. Díez-Vives, Rafel Coma, Marta Ribes, and José M. Montoya. Specificity and temporal  
324 dynamics of complex bacteria-sponge symbiotic interactions. *Ecology*, 94(12):2781–2791, 2013.

- 325 [3] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H.  
326 Stevens, and Jada-Simone S. White. Generalized linear mixed models: a practical guide for ecology and evolu-  
327 tion. *Trends in Ecology & Evolution*, 24(3):127–135, 2009.
- 328 [4] Miklós Bálint, Mohammad Bahram, A. Murat Eren, Karoline Faust, Jed A. Fuhrman, Björn Lindahl, Robert B.  
329 O’Hara, Maarja Öpik, Mitchell L. Sogin, Martin Unterseher, and Leho Tedersoo. Millions of reads, thousands of  
330 taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews*,  
331 40(5):686, 2016.
- 332 [5] James S. Clark, Diana Nemergut, Bijan Seyednasrollah, Phillip J. Turner, and Stacy Zhang. Generalized joint at-  
333 tribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecological Monographs*,  
334 87(1):34–56, 2017.
- 335 [6] A.J. Drummond, M.A. Suchard, D. Xie, and Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST  
336 1.7. *Molecular Biology And Evolution*, 12:1969–1973, 2012.
- 337 [7] Joseph Felsenstein. Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15, 1985.
- 338 [8] Jonathan Friedman and Eric J. Alm. Inferring Correlation Networks from Genomic Survey Data. *PLOS Compu-*  
339 *tational Biology*, 8(9):1–11, 09 2012.
- 340 [9] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A Weakly Informative Default Prior  
341 Distribution for Logistic and Other Regression Models. 2(4):1360–1383, 2008.
- 342 [10] Volker Gloeckner, Markus Wehrl, Lucas Moitinho-Silva, Christine Gernert, Peter Schupp, Joseph R. Pawlik,  
343 Niels L. Lindquist, Dirk Erpenbeck, Gert Wörheide, and Ute Hentschel. The HMA-LMA Dichotomy Revisited:  
344 an Electron Microscopical Survey of 56 Sponge Species. *The Biological bulletin*, 227(1):78–88, 2014.
- 345 [11] Mathieu Groussin, Florent Mazel, Jon G. Sanders, Chris S. Smillie, Sébastien Lavergne, Wilfried Thuiller, and  
346 Eric J. Alm. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nature*  
347 *Communications*, 8:14319EP, Feb 2017.
- 348 [12] Francis K C Hui, Sara Taskinen, Shirley Pledger, Scott D. Foster, and David I. Warton. Model-based approaches  
349 to unconstrained ordination. *Methods in Ecology and Evolution*, 6(4):399–411, 2015.
- 350 [13] Francis K.C. Hui. boral–Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R.  
351 *Methods in Ecology and Evolution*, 2016.

- 352 [14] Anthony R. Ives and Matthew R. Helmus. Phylogenetic metrics of community similarity. *The American natu-*  
353 *ralist*, 176(5):E128–E142, 2010.
- 354 [15] Arne Kaldhusdal, Roland Brandl, Jörg Müller, Lisa Möst, and Torsten Hothorn. Spatio-phylogenetic multispecies  
355 distribution models. *Methods in Ecology and Evolution*, 6(2):187–197, 2015.
- 356 [16] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wal-  
357 lace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal W and Clustal X version 2.0.  
358 *Bioinformatics*, 23(21):2947, 2007.
- 359 [17] Pierre Legendre and Olivier Gauthier. Statistical methods for temporal and space-time analysis of community  
360 composition data. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1778), 2014.
- 361 [18] Shirong Liu, Andre Pires da Cunha, Rafael M. Rezende, Ron Cialic, Zhiyun Wei, Lynn Bry, Laurie E. Comstock,  
362 Roopali Gandhi, and Howard L. Weiner. The Host Shapes the Gut Microbiota via Fecal MicroRNA. *Cell Host*  
363 *& Microbe*, 19(1):32–43, 2016.
- 364 [19] Plummer Martyn, Alexey Stukalov, and Matt Denwood. Bayesian Graphical Models using MCMC. *R News*,  
365 6(1):7–11, 2006.
- 366 [20] Joseph B. Maxwell, William E. Stutz, and Pieter T. J. Johnson. Multilevel Models for the Distribution of Hosts  
367 and Symbionts. *PLOS ONE*, 11(11):1–15, 11 2016.
- 368 [21] Margaret McFall-Ngai, Michael G. Hadfield, Thomas C.G. Bosch, Hannah V. Carey, Tomislav Domazet-Lošo,  
369 Angela E. Douglas, Nicole Dubilier, Gerard Eberl, Tadashi Fukami, Scott F. Gilbert, Ute Hentschel, Nicole King,  
370 Staffan Kjelleberg, Andrew H. Knoll, Natacha Kremer, Sarkis K Mazmanian, Jessica L. Metcalf, Kenneth Neal-  
371 son, Naomi E Pierce, John F. Rawls, Ann Reid, Edward G. Ruby, Mary Rumpho, Jon G. Sanders, Diethard Tautz,  
372 and Jennifer J. Wernegreen. Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of*  
373 *the National Academy of Sciences*, 110(9):3229–3236, 2013.
- 374 [22] Curtis S. McKay. Create Plots from MCMC Output. *R News*, 2015.
- 375 [23] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining R<sup>2</sup> from generalized  
376 linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013.



- 377 [24] Otso Ovaskainen, Nerea Abrego, Panu Halme, and David Dunson. Using latent variable models to identify  
378 large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*,  
379 7(5):549–555, 2016.
- 380 [25] Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas  
381 Roslin, and Nerea Abrego. How to make more out of community data? A conceptual framework and its imple-  
382 mentation as models and software. *Ecology Letters*, 2017.
- 383 [26] Martyn Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*, 2003.
- 384 [27] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence Diagnosis and Output  
385 Analysis for MCMC. *R News*, 6(1):7–11, 2006.
- 386 [28] Laura J. Pollock, Reid Tingley, William K. Morris, Nick Golding, Robert B. O’Hara, Kirsten M. Parris, Peter A.  
387 Vesk, and Michael A. Mccarthy. Understanding co-occurrence by modelling species simultaneously with a Joint  
388 Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406, 2014.
- 389 [29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Comput-  
390 ing, Vienna, Austria, 2016.
- 391 [30] A. Rambaut, M.A. Suchard, D. Xie, and A.J. Drummond. Tracer v1.6. 2013.
- 392 [31] Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister,  
393 Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres,  
394 Gerhard G. Thallinger, David J. Van Horn, and Carolyn F. Weber. Introducing mothur: Open-source, platform-  
395 independent, community-supported software for describing and comparing microbial communities. *Applied and*  
396 *Environmental Microbiology*, 75(23):7537–7541, 2009.
- 397 [32] Susanne Schmitt, Peter Tsai, James Bell, Jane Fromont, Micha Ilan, Niels Lindquist, Thierry Perez, Allen Ro-  
398 drigo, Peter J. Schupp, Jean Vacelet, Nicole Webster, Ute Hentschel, and Michael W. Taylor. Assessing the  
399 complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *The*  
400 *ISME Journal*, 6(3):564–576, 2012.
- 401 [33] Torsten Thomas, Lucas Moitinho-Silva, Miguel Lurgi, Johannes R Björk, Cole Easson, Carmen Astudillo-  
402 García, Julie B Olson, Patrick M Erwin, Susanna López-Legentil, Heidi Luter, et al. Diversity, structure and  
403 convergent evolution of the global sponge microbiome. *Nature Communications*, 7(11870), 2016.

- 404 [34] John N Thompson. *The coevolutionary process*. University of Chicago Press, 1994.
- 405 [35] Gleb Tikhonov, Nerea Abrego, David Dunson, and Otso Ovaskainen. Using joint species distribution models for  
406 evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and*  
407 *Evolution*, 8(4):443–452, 2017.
- 408 [36] Hirokazu Toju, Masato Yamamichi, Paulo R. Guimarães Jr, Jens M. Olesen, Akihiko Mougi, Takehito Yoshida,  
409 and John N. Thompson. Species-rich networks and eco-evolutionary synthesis at the metacommunity level.  
410 *Nature Ecology & Evolution*, 1:0024EP, Jan 2017.
- 411 [37] Jenny Tung, Luis B. Barreiro, Michael B. Burns, Jean-Christophe Grenier, Josh Lynch, Laura E. Grieneisen,  
412 Jeanne Altmann, Susan C. Alberts, Ran Blekman, and Elizabeth A. Archie. Social networks predict gut micro-  
413 biome composition in wild baboons. *eLife*, 4:e05224, 2015.
- 414 [38] Jean Vacelet and Claude Donadey. Electron microscope study of the association between some sponges and  
415 bacteria. *Journal of Experimental Marine Biology and Ecology*, 30(3):301–314, 1977.
- 416 [39] Mark Vellend. Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology*, 85(2):183–206,  
417 2010.
- 418 [40] David I. Warton, F. Guillaume Blanchet, Robert B. O’Hara, Otso Ovaskainen, Sara Taskinen, Steven C. Walker,  
419 and Francis K. C. Hui. So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology and*  
420 *Evolution*, 30:1–14, 2015.
- 421 [41] David I. Warton, Stephen T. Wright, and Yi Wang. Distance-based multivariate analyses confound location and  
422 dispersion effects. *Methods in Ecology and Evolution*, 3(1):89–101, 2012.
- 423 [42] Nicole S. Webster, Michael W. Taylor, Faris Behnam, Sebastian Lucker, Thomas Rattei, Stephen Whalan,  
424 Matthias Horn, and Michael Wagner. Deep sequencing reveals exceptional diversity and modes of transmis-  
425 sion for bacterial sponge symbionts. *Environmental Microbiology*, 12(8):2070–2082, 2010.
- 426 [43] Li C. Xia, Joshua A. Steele, Jacob A. Cram, Zoe G. Cardon, Sheri L. Simmons, Joseph J. Vallino, Jed A.  
427 Fuhrman, and Fengzhu Sun. Extended local similarity analysis (eLSA) of microbial community and other time  
428 series data with replicates. *BMC Systems Biology*, 5(Suppl 2):S15, 2011.
- 429 [44] Lizhen Xu, Andrew D. Paterson, and Wei Xu. Bayesian latent variable models for hierarchical clustered count  
430 outcomes with repeated measures in microbiome studies. *Genetic Epidemiology*, 41(3):221–232, 2017.

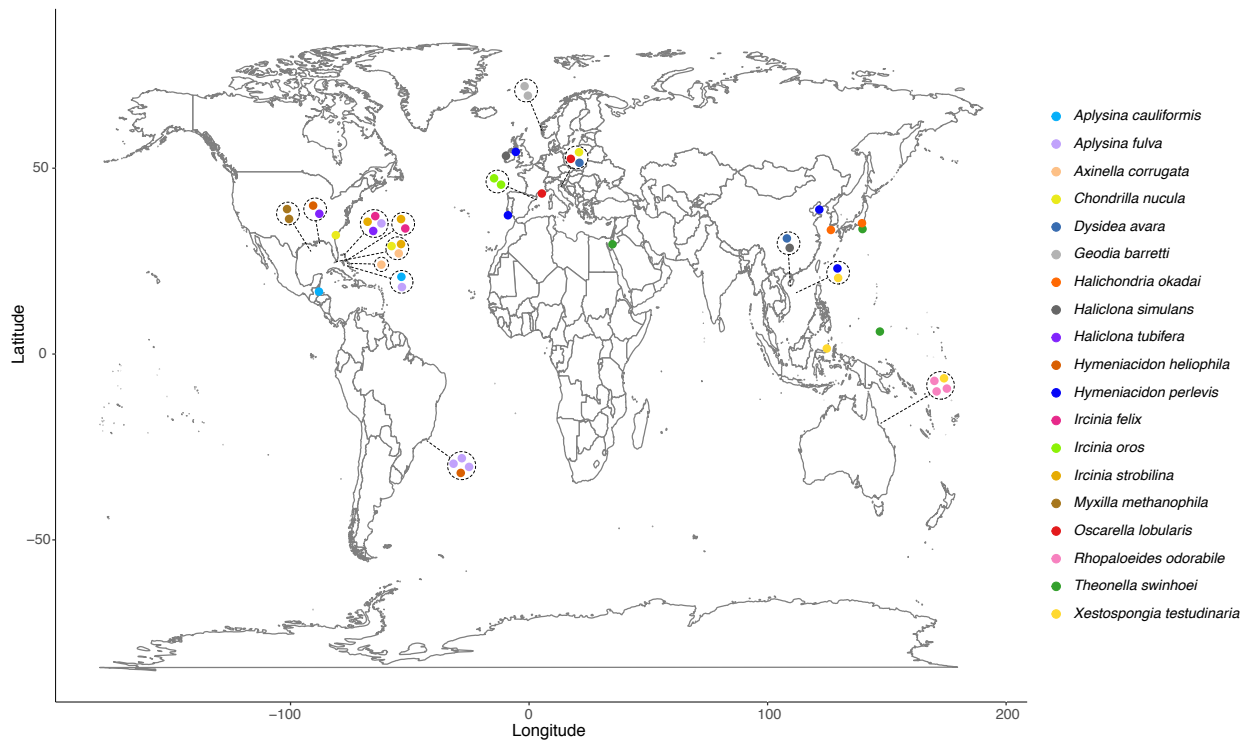


Figure 1: Overview of the broad spatial scale for which the data is distributed. Each point represents a collection site and each color represents a host species. Note that some host species coexist within the same site. See Table S1 for more detailed information.

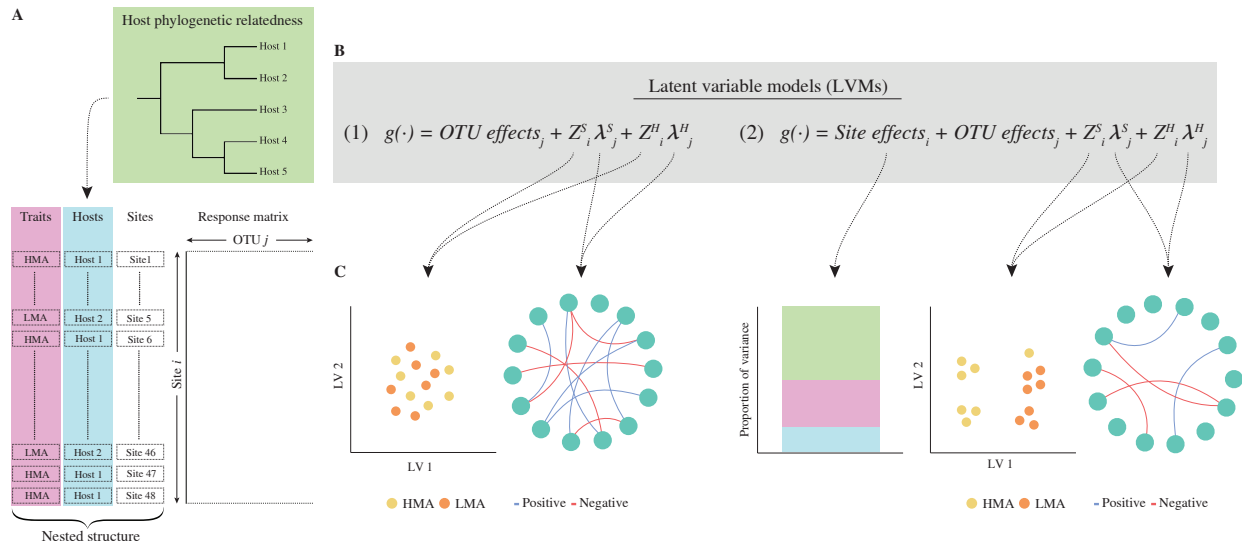


Figure 2: Conceptual figure of the modeling framework. Panel A shows a schematic figure of the response matrix. While columns correspond to OTUs, rows have a hierarchical structure where sites are nested within host species which are further nested within host traits (High Microbial Abundance (HMA) and Low Microbial Abundance (LMA)). At the host species level, the framework also accounts for phylogenetic relatedness. Panel B shows the two different joint species distribution models (JSDMs) with latent factors for site ( $S$ ) and host species ( $H$ ) level, each representing a different level of biological organization. The  $g(\cdot)$  represents the different link function associated to the different response types. Panel C shows the corresponding output; because model (1) does not include site effects, its resulting ordination constructed from the latent factors are in terms of microbiota structure (i.e., a composite of abundance and composition), and because model (2) includes site effects, its resulting ordination constructed from the latent factors are in terms of microbiota composition only. The OTU-to-OTU association networks constructed from the corresponding factor loadings also differ for the two JSDM models. Note that ordinations and association networks are produced both on the site and host species level, respectively. Finally, as the site effects are nested within the host-specific features, model (2) partition variance in microbiota abundance or species richness into components directly reflecting the included host-specific features.

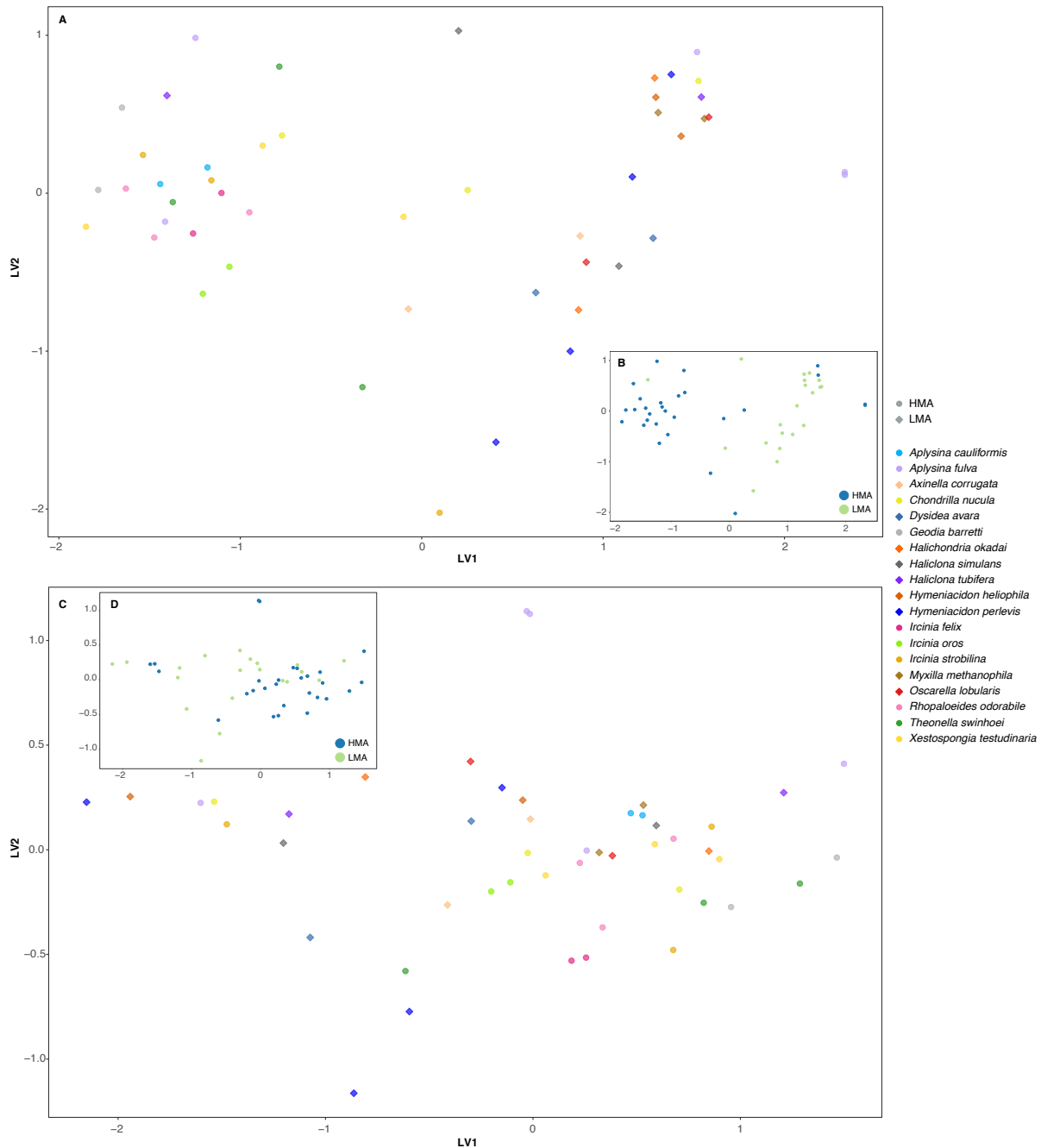


Figure 3: Model-based ordinations on the site level. Panel A and B show the model-based unconstrained ordination without site effects included. In panel A, sites are colored by host species and ecotype are depicted by different shapes (HMA=circles, LMA=diamonds), while in panel B sites are colored by ecotype only (HMA=blue, LMA=green). Panel C and D show the model-based unconstrained ordination with site effects included. In panel C, sites are colored by host species and ecotype is depicted by different shapes (HMA=circles, LMA=diamonds), while in panel D sites are colored by ecotype only (HMA=blue, LMA=green).

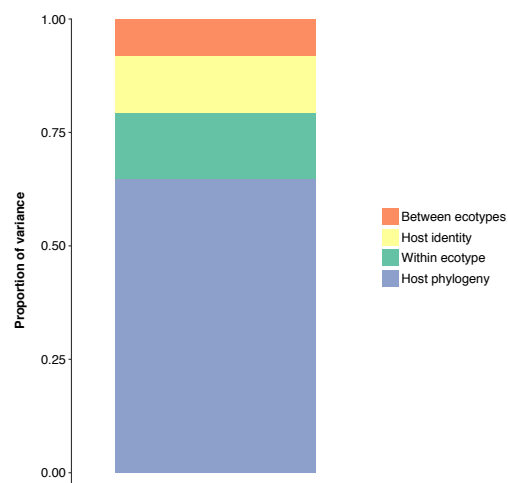


Figure 4: Explanatory power of the included host-specific features. The proportion of variance in terms of total abundance among sites explained by the included host-specific features. Yellow corresponds to variance explained by host species identity, blue to host phylogenetic relatedness, green to variance within ecotypes, and finally red corresponds to variance explained by differences among the two ecotypes.

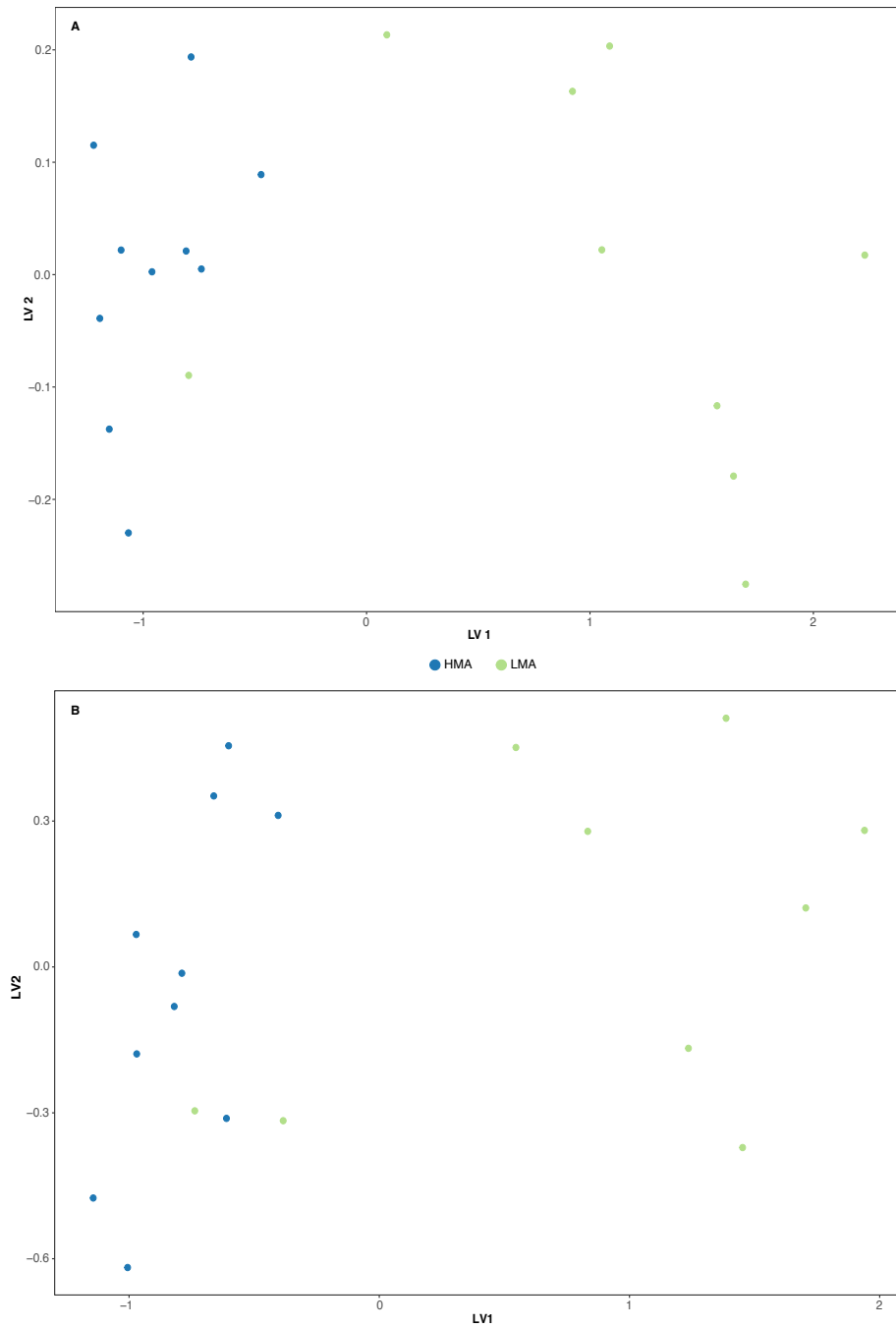


Figure 5: Model-based ordinations on the host species level. Panel A shows the model-based unconstrained ordination without site effects included, while panel B shows the model-based unconstrained ordination with site effects included. In both panels, host species are colored by ecotype (HMA=blue, LMA=green).

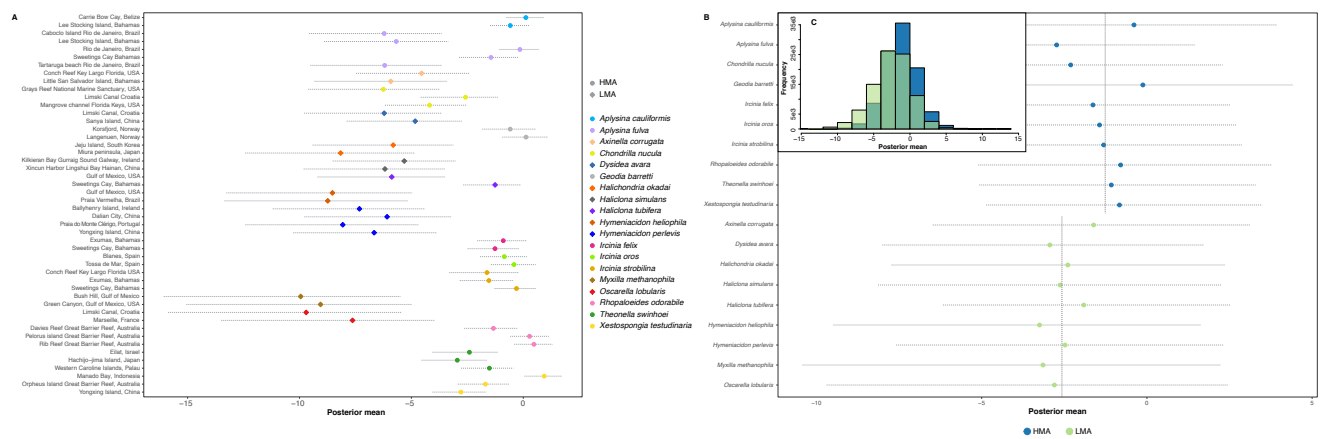


Figure 6: Variation in OTU abundance across sites and host species. Panel A shows a caterpillar plot for the parameter controlling the site effects, i.e.,  $\alpha_i$ . Each row correspond to a sites, colored by host species. The colored shape represent the posterior mean ( $\pm$  SD). The two ecotype are depicted by different shapes (HMA=circles, LMA=diamonds). Panel B shows a caterpillar plot for  $\alpha_i$ 's mean parameter, i.e., the  $\mu(\text{host})_{s[r]}$ . Rows correspond to host species colored by ecotype (HMA=blue, LMA=green). The vertical dashed lines correspond to the grand mean of each ecotype. Panel C shows the posterior probability distribution of  $\mu(\text{host})_{s[r]}$  for HMA (blue) and LMA (green), respectively.



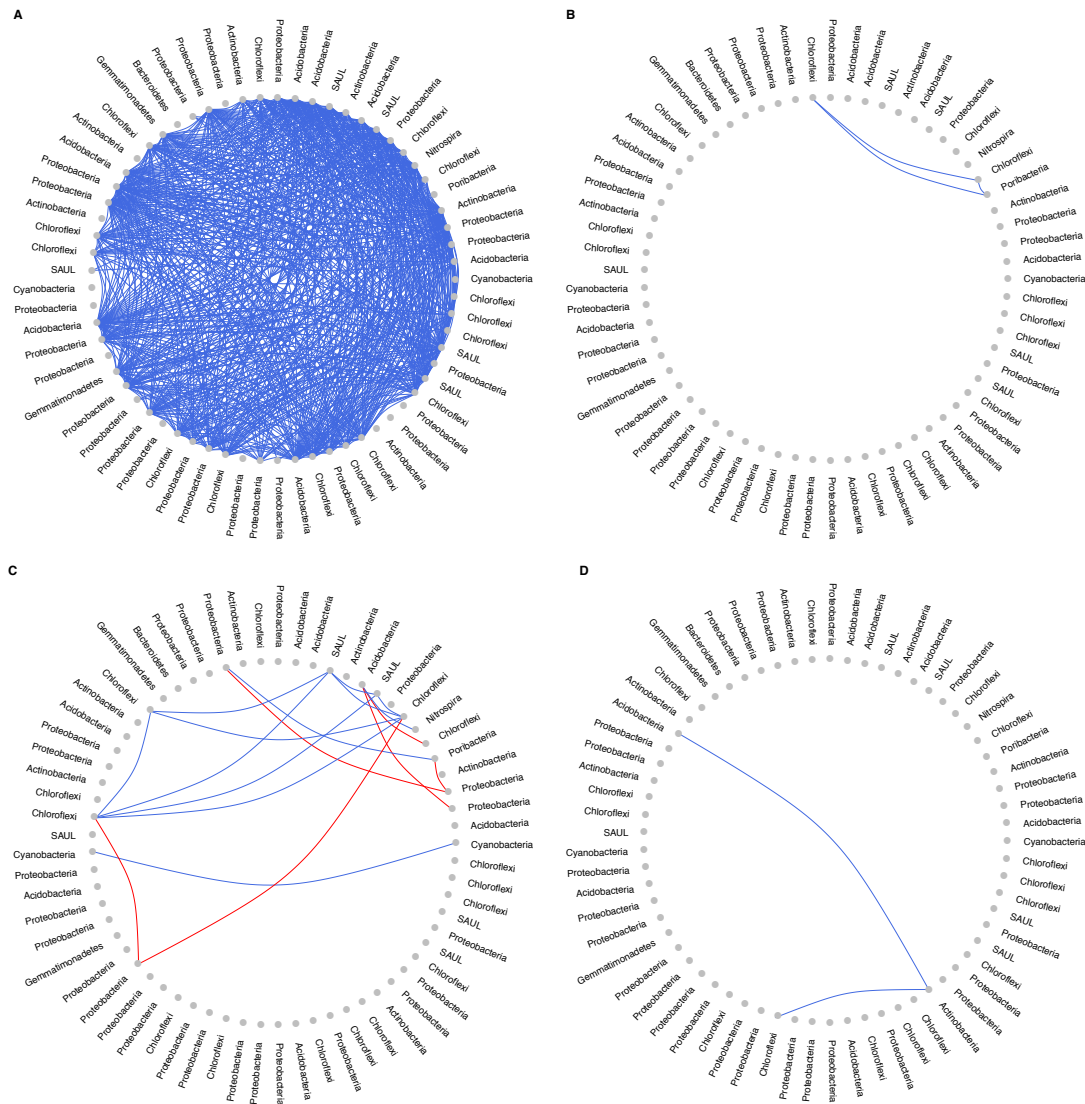


Figure 7: OTU-to-OTU association networks. Nodes represent OTUs with assigned taxonomy at the phylum-level, and links correspond to abundance correlations with at least 95% posterior probability. The top panel (A & B) shows networks generated from the model without site effects, thus correlations between OTUs are induced by both site and host-specific features as well as OTU-specific effects. The bottom panel (C & D) shows networks generated from the model with site effects included, thus correlations between OTUs are only OTU-specific effects. Panel A & C shows the association network for the site level and panel B & D shows the network for the host species level.