APPENDIX 3

Description of the hierarchical Bayesian generalized linear model with latent variables

We fitted a statistical joint species distribution model (Warton *et al.* 2015), which combines information on environmental covariates, species traits and phylogenetic constraints, as well as the sampling study design. The modelling framework used here is described in Abrego, Norberg & Ovaskainen (2016b), as well as Ovaskainen et al. (2016a, b) and Abrego et al. (2016), where the application of the framework with other research questions and organismal groups is also demonstrated. As an alteration to Abrego *et al.* (2016b), the phylogenetic relationships of the taxa were accounted for slightly differently. Below we describe the structure of the modelling framework (including the deviations from Abrego *et al.* 2016b), as well as the specific details of each of the individual models.

We denote the sampling unit by the index $i = 1, \dots, n_y$; the focal species (parasite species or bacterial orders; referred to as species in this section) by the index $j = 1, \dots, n_s$; and the species-specific traits by the index $l = 1, \dots, n_t$; where n_y is the total number of sampling units (consisting of several samples from individual lemurs, obtained at different times points); n_s is the number of parasite species or bacterial orders, and n_t is the number of traits. We denote the presence-absence data by y_{ij}, so that $y_{ij} = 1$ if the species j was found in sampling unit i and otherwise $y_{ij} = 0$. We model the presence (and absence) of the species with a probit regression model, where $y_{ij} = 1_{z_{ij}>0}$, and the latent occurrence score is modelled as $z_{ij} = L_{ij} + \varepsilon^H_{H(i)j} + \varepsilon^P_{P(i)j} + \varepsilon^A_{A(i)j} + \epsilon_{ij}$. The residual term $\epsilon_{ij} \sim N(0,1)$ corresponds to the probit link function, as it is fixed to $\sigma_j^2 = 1$. The fixed effects L_{ij} are modelled as regression parameters $L_{ij} = \sum_k x_{i⍰}\beta_{jk}$, where terms x_{ik} denote the covariate measured for sampling unit i for condition k, and the regression parameters β_{jk} denote the response of species j to k.

We assumed that the species-specific regression coefficients β_{jk} follow the multivariate normal distribution $\boldsymbol{\beta}_{j\cdot} \sim N(\boldsymbol{\mu}_{j\cdot}, \mathbf{V})$, centred around the expectation $\boldsymbol{\mu}_{j\cdot}$. As the dot singles out a row/column in a matrix, $\boldsymbol{\beta}_{j\cdot}$ denotes the vector of regression coefficients for species j, and $\boldsymbol{\mu}_{j\cdot}$ is a vector of mean responses, hence describing the expected environmental niche of species j. We used the $\boldsymbol{\mu}_{j\cdot}$ parameters to model the influence of species-specific traits on the species' responses to their habitat characteristics: We modelled the expected response as a linear combination of the traits: $\mu_{jk} = \sum_l t_{jl}\gamma_{lk}$, where t_{jl} is the value of trait l for species j, and the parameter γ_{lk} measures the effect of trait l on covariate k. To account for phylogenetic relationships, the covariance structure of the multivariate normal distribution has been modified so, that $\boldsymbol{\beta}_{\cdot\cdot} \sim N\left(\boldsymbol{\mu}_{\cdot\cdot}, V \otimes \left[|\rho|\left(1_{\rho>0}\mathbf{C} + 1_{\rho<0}\mathbf{C}^{-1}\right) + (1-|\rho|)\mathbf{I}\right]\right)$, where $\mathbf{C}$ is the phylogenetic correlation matrix, and $-1 \leq \rho \leq 1$ measures the strength of the phylogenetic signal. If $\rho = 0$, then the residual variance is independent among the species, but when ρ

approaches $\rho = -1$ or $\rho = 1$, species' environmental niches become fully structured by their phylogeny. A positive parameter value $\rho > 0$ indicates that related species have more similar niches than by random, whereas $\rho < 0$ suggest that closely related species have less similar niches than expected. This part of our modelling framework differs from Abrego *et al.* (2016b), where the parameter ρ could only vary between 0 and 1, i.e. $0 \leq \rho \leq 1$, hence not separating for covariance due to niche similarity or dissimilarity.

The random effects $\varepsilon^{H}_{H(i)j} + \varepsilon^{P}_{P(i)j} + \varepsilon^{A}_{A(i)j}$ model variation in species occurrences and co-occurrences at the levels of individual lemurs (H), transects (P) and years (A). The indices $H(i)$, $P(i)$ and $A(i)$ denote the lemur from which, the transect in which, and the year during which the sample i was obtained. As in Ovaskainen et al. (2016a, b) and Abrego *et al.* (2016a; b), we assume that the random effects are distributed according to the multivariate normal distributions $\varepsilon^{H}_{H\cdot} \sim N(0, \mathbf{\Omega}^{H})$, $\varepsilon^{P}_{P\cdot} \sim N(0, \mathbf{\Omega}^{P})$ and $\varepsilon^{A}_{A\cdot} \sim N(0, \mathbf{\Omega}^{A})$ where $\mathbf{\Omega}^{H}$, $\mathbf{\Omega}^{P}$ and $\mathbf{\Omega}^{A}$ are taxon-to-taxon variance-covariance matrices to be estimated. Here, the diagonal element Ω_{jj} describes the amount of variation that species j shows at the level of individual lemurs, whereas the off-diagonal element $\Omega_{j_1 j_2}$ describes the amount of covariation among the species j_1 and j_2. The variance-covariance matrix $\mathbf{\Omega}$ can be translated into a correlation matrix $\mathbf{R}$ by $R_{j_1 j_2} = \Omega_{j_1 j_2} / \sqrt{\Omega_{j_1 j_1} \Omega_{j_2 j_2}}$, which is the matrix that we will use here to represent species-to-species associations. The correlation $R_{j_1 j_2}$ measures to what extent species j_1 and species j_2 are found together more or less often than expected by random, after controlling for the environmental covariates.

We first modelled parasite associations by using a longer data set from years 2011-2012. We modelled the presence and absence of the parasites found in and on the lemurs as a function of the sex, age, aggressiveness and general condition of the lemurs, and with males we also accounted for the size of their testis (with the assumption that females can be considered as individuals with extremely small testis size). We also included the time of sampling (week) and its quadratic form (week^2) to account for the effect of seasonality. As traits we included whether the parasite has a direct or non-direct life cycle and whether it is an endo- or ectoparasite. We constructed the phylogenetic relationships with five levels: domain, kingdom, superphylum, phylum and species, assuming equal branch lengths.

We consequently modelled parasite-to-microbiota association using the data set from 2012, which included both microbiota and parasites. Here we modelled the occurrences of both the parasite species and the microbial orders as a function of the same characteristics of the lemurs as with the parasite model. We transformed the OTU abundance data into presences and absences at the level of orders. To avoid overrepresentation of very rare OTUs, we considered OTUs with >9 amplicons as presences, and ≤9 as absences. Then to avoid sequencing and OTU picking errors, we considered the OTUs present, if there were in total >99 amplicons in at least two lemur individuals. After this, the final

community matrix was constructed as presence and absence at the level of orders. This results in that the occurrence of a microbial order represents a strong presence of this particular order. We included latent random effects at the levels of individual lemurs ($\varepsilon^{H}_{H(i)j}$) and transects ($\varepsilon^{P}_{P(i)j}$), as in this model we did not have samples from multiple years. In addition to the traits also used in the parasite model, here we included whether the taxon is a parasite or part of the microbiota and microbiota was considered as having neither direct nor indirect life cycle. We constructed the phylogenetic relationships with five levels: domain, kingdom, phylum, class and order (the level of observations). We assumed equal branch lengths, but since the occurrences were modelled at the level of orders for the microbiota, but at the level of species for the parasites, we adjusted the phylogenetic correlation matrix C so, that the phylogenetic distance between the two hymenolepidid species was set to 0.99.

As a point of comparison, for both data sets, we fitted unconstrained models, where we included only sampling unit random effect ε^{S}_{ij}, which models the variation in species occurrences and co-occurrences at the level of individual samples, obtained from individual lemurs (as there could be multiple samples from one individual), and no environmental covariates, phylogenetic constrains, nor traits. Hence, we again model the presence (and absence) of the species with a probit regression model, but in the unconstrained version the fixed effects L_{ij} are modelled simply with an intercept, so that $L_{ij} = x_i\beta_j$, and the term x_i is a vector of ones of length i, and the regression parameters β_j model the overall prevalence of the species j. Thus, the variance across sampling units in the species responses is explained with the latent variables. By comparing the results for the constrained and unconstrained models, we can separate the associations that are solely due to the (dis)similar habitat requirements (e.g. when two species share the same habitat preferences, and hence co-occur more often than expected by random) or hidden by the (dis)similar habitat requirements (e.g. when two species share the same habitat preferences, but even after accounting for this, they still co-occur more often than expected by random) from the associations immune to the effects of the explanatory variables (i.e. we see the same association patterns regardless of the inclusion of the explanatory variables). This approach is analogous to comparing a constrained and an unconstrained ordination, with the difference of our approach being model-based (see e.g. Hui et al. 2015, Warton et al. 2015).

Prior distributions used for the hierarchical Bayesian joint species distribution model

We used the default priors of the Matlab package 'HMSC' by Ovaskainen et al. *(submitted)*. For each component of the trait parameter γ, we assumed a normal prior with zero mean and unit variance. For the matrix **V**, which measures the amount of variation among the species-specific regression coefficients (the diagonal elements of **V**), as well as the amount of covariation among the responses to

the different covariates (the off-diagonal elements of **V**), we assumed an Inverse-Wishart prior with $n_c + 1$ degrees of freedom, and the variance-covariance matrix set to the identity matrix.

For the latent factors, the priors are as in Bhattacharya and Dunson (2011), except that $a_1 = a_2 = 50$, increasing the level of shrinkage. High amount of shrinkage may result in underestimation the influences of the random factors, and in particular miss some of the species-to-species associations, but assuming only little shrinkage would increase the risk of overfitting. For the phylogenetic signal parameter ρ, we assumed a discrete prior, which assigns a probability of $1/3$ for $\rho = 0$ (corresponding to independence among species), and the remaining probability of $2/3$ uniformly to [-1,1], discretized to 200 values to enable the use of a discrete grid sampler.

We sampled the posterior distribution using the the Gibbs sampler developed by Bhattacharya & Dunson (2011) and Ovaskainen *et al.* (2016a; b), and implemented in Matlab by Ovaskainen et al. *(submitted)*. We run the MCMC chains of all the models to 100 000 iterations, out of all of which the first half was discarded.

Assessing the model fit

We assessed the model fit and predictive power by calculating the Tjur R^2 coefficients of discrimination (Tjur 2009) for each species. Furthermore, we calculated the Spearman rank correlations between the predicted and true occurrences. We calculated the correlations at all spatial and temporal scales, (individual lemurs, transects and years).

For the parasite model, there was a good match between the predicted and true occurrences of species, as the mean Tjur R^2 coefficient of discrimination over the species was 0.13 at the level of sampling units, and the Spearman rank correlations between the predicted occurrences and the true occurrences was 0.53 at the level of sampling units, 0.71 at the level of individual lemurs, 0.98 at the transect level and 0.99 at the level of years.

For the combined model, there was a good match between the predicted and true occurrences of species, as the mean Tjur R^2 coefficient of discrimination over the species was 0.12 at the level of lemurs, and the Spearman rank correlation between the predicted occurrences and the true occurrences was 0.73 at the level of sampling units, 0.79 at the level of individual lemurs, and 0.98 at the transect level.

Estimated phylogenetic signal in the species responses to their habitat

For the parasite data, the posterior mean of the parameter ρ measuring the amount of phylogenetic signal in the data was 0.92, indicating a strong phylogenetic signal in the species responses to their environment. As there was a strong phylogenetic signal in the community response to the environment, we tested whether there is some correlation between the phylogenetic relationships between taxa and the covariance structure of the latent factors, represented by the matrix C and the residual covariance matrix $\boldsymbol{\Omega}$ by using the Mantel test with Pearson correlation. There were no significant (based on the empirical significance level from permutations ≤ 0.05) correlations at the level of lemurs ($\boldsymbol{\Omega}^H$), transects ($\boldsymbol{\Omega}^P$), nor years ($\boldsymbol{\Omega}^A$).

For the combined data, the posterior mean of the parameter ρ measuring the amount of phylogenetic signal in the data was 0.94, indicating a strong phylogenetic signal in the species responses to their environment. At the transect level there was no significant correlation between the phylogenetic relationships of taxa, but at the level of individual lemurs we found a significant ($p<0.05$) correlation of $r = 0.14$.

References:

Abrego, N., Dunson, D., Halme, P., Salcedo, I. & Ovaskainen, O. (2016a) Wood-inhabiting fungi with tight associations with other species have declined as a response to forest management. *Oikos*, **(accepted)**.

Abrego, N., Norberg, A. & Ovaskainen, O. (2016b) Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi. *Journal of Ecology*, **(accepted)**.

Bhattacharya, A. & Dunson, D.B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015) Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, **6**, 399–411.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016a) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, **7**, 549–555.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016b) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, **7**, 428–436.

Tjur, T. (2009) Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *American Statistician*, **63**, 366–372.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K.C. (2015) So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, **30**, 766–779.