Luke O. Frishkoff, Perry de Valpine, Leithen K. M'Gonigle

Phylogenetic occupancy models integrate imperfect detection and phylogenetic signal to analyze community structure

Ecology

Appendix S1

Influence of number of sites surveyed on inference of phylogenetic signal
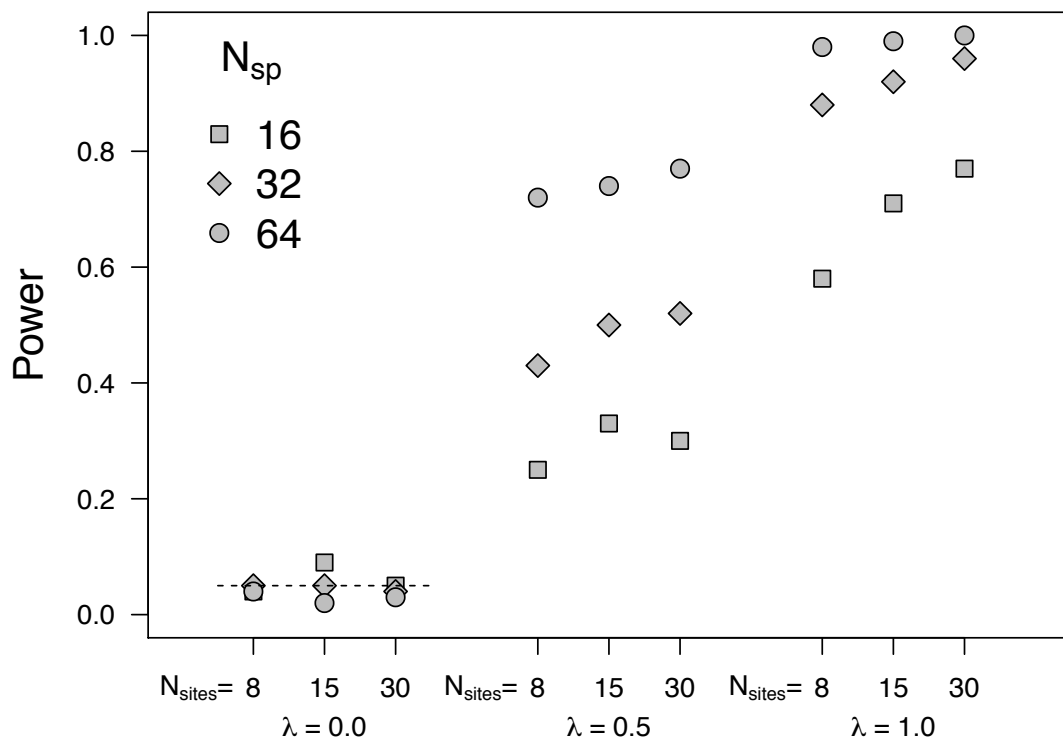

Figure S1: Statistical power to detect phylogenetic signal when it exists increases primarily as a function of number of species sampled, rather than as the number of sites sampled. Points designate the proportion out of 100 simulated datasets for which 95% HPDs for λ exclude 0, varying the number of species, number of sites, and the generating value for λ. The dotted dashed line indicates the 0.05 type I error rate expected when λ actually is equal to 0.

Bayes factors as an alternative to HPD

As an alternative to using the 95% highest posterior density we tested the performance of Bayes factors for inferring whether phylogenetic signal existed in the data (i.e. $\lambda > 0$). To do so we ran a POM, this time adding an inclusion parameter W for λ:

$\lambda = \lambda_0 * W$

Specifically we model these parameters according to the priors:

$W \sim Bernoulli(0.5)$

$\lambda_0 \sim Uniform(0, 1)$

As a result, at each step of the MCMC, W can equal either 0 or 1. When the inclusion parameter equals 1, λ is freely estimated in the model via the parameter λ_0. When W is equal to 0, then λ is also set to 0, and $\mathbf{C}_{phylo}$ is set equal to the identity matrix. The relative frequency that the inclusion parameter spends equal to 1 versus 0 is an approximation of the Bayes Factor (as the prior probability is equal for each state), giving the weight of evidence for the POM versus the non-phylogenetic MSOM model. Since λ is restricted between 0 and 1, we did not encounter some of the MCMC mixing problems that can arise from indicator variables (Hooten and Hobbs 2015).

Different Bayes factor cut offs have been proposed as providing sufficient evidence for considering one model categorically more likely than another (*e.g.* Kass and Raftery 1995). Here we consider a Bayes factor of 3 (equivalent to model weight of 75% in favor of POM) to indicate support for the POM over the MSOM. We evaluate the type I error rate by quantifying the proportion of 100 datasets simulated with $\lambda = 0$ for which

the Bayes factor on λ is >3. We evaluate the POM's type II error rate when λ is simulated as 0.5 and 1 for which the Bayes factor for λ is < 3.

Performance of Bayes factors on simulated data

We found that type I and type II error rate from Bayes factors generally conform to those provided by the 95% HPP (Figure S2). Bayes factors seem less likely to demonstrate inflated type I error when numbers of species are low, though they also display systematically slightly less power than using 95% HPP to determine whether $\lambda=0$.

We found that Bayes factors also seem to be effective for selecting between a POM and a MSOM when traits also influence response to an environmental gradient (Table S1). In conditions in which there is phylogenetic correlation in the residuals after accounting for the effect of the trait the POM is generally favored. In contrast, when there is not phylogenetic correlation in the residuals the MSOM is favored. This allows inferior models to be avoided in cases were either the POM or MSOM has inflated root mean square error (Figure 4a).

We chose a Bayes factor cut off of 3 because it is commonly used in the literature (Kass and Raftery 1995). Whether Bayes factors as implemented here are truly invariant to sample size, or tree topology, is beyond the scope of these analyses. When working with real (rather than simulated) data, we reiterate that users test the significance of λ using parametric bootstrapping or similar methods (Boettiger et al. 2012).
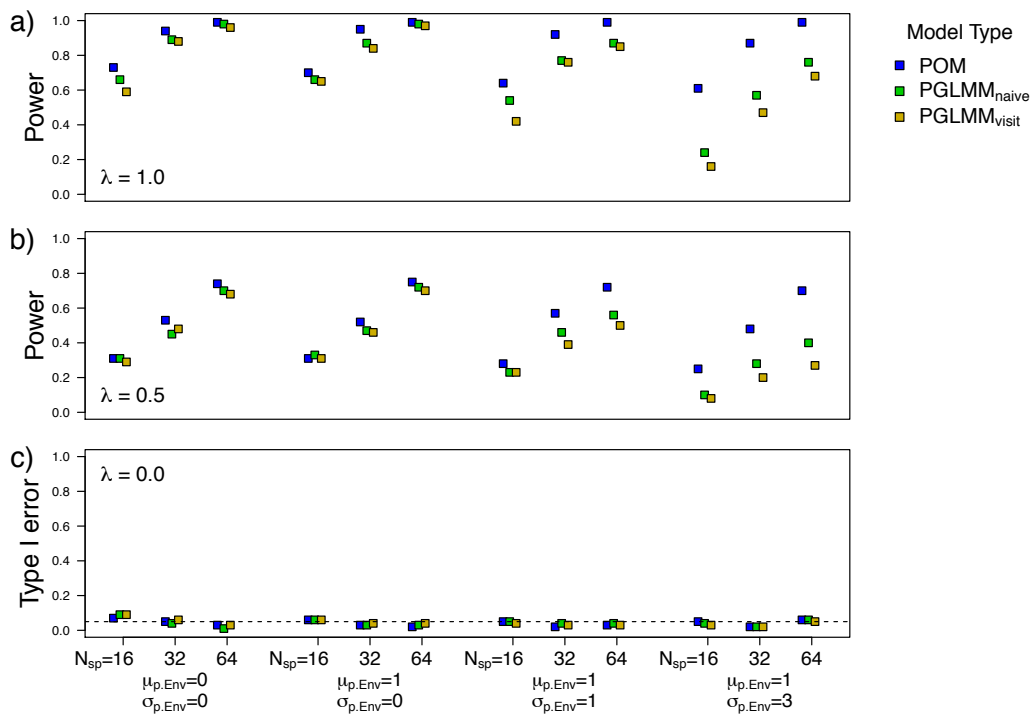
Figure S2: Points represent the proportion data simulations and model runs (out of 100) for which the indicator variable on λ yielded a Bayes factor greater than 3, indicating statistical support for λ > 0. Panels show power for different simulated values of λ: a) λ = 1.0, b) λ = 0.5, c) λ = 0. Compare to figure 3 in maintext.

Table S1: If a phylogenetic signal in responses to an environmental gradient exists, accounting for this correlation improves estimation of trait effects. For each case below, we fit both the POM and a MSOM to 100 simulated datasets and quantified estimation quality of β_{trait}. RMSE indicates the root mean square error of the β_{trait} estimate. BCI refers to the 95% Bayesian credibility interval around β_{trait}. 'Spp' indicates number of species, width of BCI reports the average length of the 2.5% to 97.5% BCI across all 100 simulations, whereas 'ratio' refers to the average ratio between all 100 MSOM:POM comparisons—values above 1 indicate that the POM has greater certainty in its estimation of β_{trait}. Width of the BCI is a metric of type II error, complementary to the power analysis presented in the main text (BCIs exclude 0). 'BCI contains true value' indicates the proportion of the 100 simulations for which the credible interval contained the true value of β_{trait} (inverse of type I error). Finally, 'POM favored' indicates the proportion of simulations for which the POM was favored over the MSOM. Here we used standard Bayes factor cut offs of 3 to indicate support.

Case	Spp	RMSE		Width of BCI			BCI contains true value		POM favored
		MSOM	POM	MSOM	POM	ratio	MSOM	POM	(BF>3)
No signal in trait, no signal in residuals	16	0.66	0.66	4.10	4.08	1.01	0.98	0.98	0.03
	32	0.44	0.46	2.50	2.46	1.02	0.97	0.94	0.05
	64	0.29	0.29	1.64	1.62	1.01	0.97	0.96	0.03
Signal in trait, No signal in residuals	16	0.80	0.86	3.98	5.54	0.75	0.97	0.97	0.05
	32	0.48	0.59	2.46	3.75	0.69	0.96	0.97	0.07
	64	0.34	0.42	1.64	2.56	0.67	0.93	0.99	0.01
No signal in trait, signal in residuals	16	0.65	0.37	3.33	2.37	1.40	0.93	0.97	0.77
	32	0.39	0.21	2.06	1.11	1.84	0.96	0.98	0.96
	64	0.27	0.12	1.33	0.63	2.12	0.96	0.95	0.99
Signal in trait, signal in residuals	16	1.05	0.68	3.11	3.54	0.90	0.73	0.94	0.71
	32	1.04	0.50	1.89	2.36	0.85	0.50	0.96	0.94
	64	0.98	0.40	1.21	1.73	0.76	0.36	0.92	0.99