Complex ecological phenotypes on phylogenetic trees: a hidden Markov model for comparative analysis of multivariate count data

Michael C. Grundler[1,*] and Daniel L. Rabosky[1]

[1]*Museum of Zoology and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI USA 48109*

[*]*Corresponding author: mgru@umich.edu*

ABSTRACT

The evolutionary dynamics of complex ecological traits – including multistate representations of diet, habitat, and behavior – remain poorly understood. Reconstructing the tempo, mode, and historical sequence of transitions involving such traits poses many challenges for comparative biologists, owing to their multidimensional nature and intraspecific variability. Continuous-time Markov chains (CTMC) are commonly used to model ecological niche evolution on phylogenetic trees but are limited by the assumption that taxa are monomorphic and that states are univariate categorical variables. Thus, a necessary first step when using standard CTMC models is to categorize species into a pre-determined number of ecological states. This approach potentially confounds interpretation of state assignments with effects of sampling variation because it does not directly incorporate empirical observations of resource use into the statistical inference model. The neglect of sampling variation, along with univariate representations of true multivariate phenotypes, potentially leads to the distortion and loss of information, with substantial implications for downstream macroevolutionary analyses. In this study, we develop a hidden Markov model using a Dirichlet-multinomial framework to model resource use evolution on phylogenetic trees. Unlike existing CTMC implementations, states are unobserved probability distributions from which observed data are sampled. Our approach is expressly designed to model ecological traits that are intra-specifically variable and to account for uncertainty in state assignments of terminal taxa arising from effects of sampling variation. The method uses multivariate count data for individual species to simultaneously infer the number of ecological states, the proportional utilization of different resources by different states, and the phylogenetic distribution of ecological states among living species and their ancestors.

The method is general and may be applied to any data expressible as a set of observational counts from different categories.

Keywords: Ecological niche evolution; intraspecific variation; hidden Markov model; macroevolution; comparative methods; Dirichlet-multinomial

Most species in the natural world make use of multiple, categorically-distinct types of ecological resources. Many butterfly species use multiple host plants, for example (Ehrlich & Raven 1964; Robinson 1999). Insectivorous warblers in temperate North America use multiple distinct microhabitats and foraging behaviors (MacArthur 1958), as do honeyeaters in mesic and arid Australia (Miller et al. 2017). The evolution of novel patterns of resource use can impact phenotypic evolution (Martin & Wainwright 2011; Davis et al. 2016), diversification (Mitter et al. 1988; Givnish et al. 2014), community assembly (Losos et al. 2003; Gillespie 2004), and ecosystem function (Harmon et al. 2009; Bassar et al. 2010). Consequently, there has been substantial interest in understanding how ecological traits related to resource use evolve and for exploring their impacts on other evolutionary and ecological phenomena (Vrba 1987; Futuyma & Moreno 1988; Forister et al. 2012; Price et al. 2012; Burin et al. 2016).

Making inferences about the evolutionary dynamics of resource use, however, first requires summarizing the complex patterns of variation observed among taxa into traits that can be modeled on phylogenetic trees. It is widely recognized that the real-world complexities of resource use are not adequately described by a set of categorical variables (Hardy & Linder 2005; Hardy 2006). Nonetheless, it is also true that major differences in resource use can sometimes be summed up in a small set of ecological states, a point made by Mitter et al. (1988)

in their study of phytophagy and insect diversification. For this reason, continuous-time Markov chain (CTMC) models, which require classifying species into a set of character states, have become commonplace in macroevolutionary studies of ecological trait evolution (Kelley & Farrell 1998; Nosil 2002; Price et al. 2012; Hardy & Otto 2014; Cantalapiedra et al. 2014; Burin et al. 2016). CTMC models describe a stochastic process for evolutionary transitions among a set of character states and are used to infer ancestral states and evolutionary rates, and to perform model-based hypothesis tests (O'Meara 2012).

The utility of continuous-time Markov chains for studying the evolutionary dynamics of resource use is limited by the modeling assumption that taxa are monomorphic for ecological states (Hardy & Linder 2005; Hardy 2006). As a practical solution, most empirical studies define one or more generalized states to accommodate species that use multiple resource types and that therefore cannot be characterized as specialists for a particular resource (Alencar et al. 2013; Price et al. 2012; Burin et al. 2016; Gajdzik et al. 2019). Another solution, rather than classifying each species as a specialist or a generalist, represents each resource category with a binary score of present or absent (Janz et al. 2001; Colston et al. 2010; Hardy 2017). In this case, the ecological state of a species is the set resources scored as present. Each of these approaches is one solution to the modeling challenge posed by intraspecific variation in resource use, but both solutions neglect variation in the relative importance of different resources for different taxa. Consequently, species classified in single state can nonetheless exhibit substantial differences in patterns of resource use, creating challenges for interpreting evolutionary transitions among character states as well as for understanding links between character state evolution and diversification.

Another limitation of continuous-time Markov chains for modeling resource use evolution emerges from the fact that species are classified into ecological states without regard for the quality and quantity of information available to perform the classification exercise. As an example, species with few ecological observations might be classified as specialists for a particular resource, when their apparent specialization is strictly a function of the small number of ecological observations available for the taxon. More generally, by failing to use a statistical model for making resource state assignments, we neglect a major source of uncertainty in our data: the uneven and incomplete knowledge of resource use across different taxa. This uncertainty, in turn, has substantial implications for how we project patterns of resource use onto a set of resource states. By failing to account for uneven and finite sample sizes characteristic of empirical data on resource use we cannot be certain if state assignments reflect true similarities or differences in resource use or are merely the expected outcome of sampling variation.

Consider the simple four-species example in Figure 1. Panels (i) and (ii) illustrate the true resource states and their phylogenetic distribution across a set of four species and their ancestors. Here, an ancestral specialist evolved a generalist diet via a single transition (panel ii), such that there are two extant species with the ancestral specialist diet (species X and Y) and two with the derived generalist diet (species P and Q). In panels (iii) and (iv) the relative importance estimates of three food resource categories in the diets of four species are used to classify each species into one of two diet states. These relative importance estimates are based on uneven, and in some cases quite small, sample sizes, consistent with many empirical datasets (Vitt & Vangilder 1983; Shine 1994; Alencar et al. 2013). In panel (v) we imagine repeating the state assignment process on independent datasets while holding the samples sizes fixed to those in panel (iii), which reveals that both the initial state assignments and the number of states from (iv) are highly

sensitive to real-world levels of sampling variation. This has obvious implications for downstream macroevolutionary analyses. There is a serious risk of conflating different (similar) state assignments with different (similar) diets when the differences (similarities) are expected, even in the absence of true differences (similarities), from sampling variation alone. In the analyses that underlie Figure 1, we find that more than 70 percent of tip state classifications do not match the true pattern of resource use.

Is this a problem in practice? This issue is difficult to assess because few studies provide information about the sample sizes that underlie state assignments. In most cases, ecological states are simply asserted as known. It is also important to emphasize that the specific problem in Figure 1 is an outcome of a more general problem: standard CTMC models have a limited ability to model complex ecological phenotypes because of the assumption that states in the model are categorical variables. While it is true that CTMC models work with a countable state space, it is not true that the states of the system must represent categorical variables. In a hidden Markov model, the observed data are assumed to be the outcome of a CTMC where the states are not directly observable. Instead, states are probability distributions from which observed data are sampled. Although hidden state CTMC models already used in macroevolution (Marazzi et al. 2012; Beaulieu et al. 2013; Beaulieu & O'Meara 2016; Caetano et al. 2018) can be interpreted like this they generally are not because in these cases the observed data are categorical variables and the hidden states are indistinguishable from the observed categories. Instead, hidden states are interpreted as unobserved factors that affect rates of change or rates of diversification. Because of this, the potential flexibility of hidden state models for modeling complex phenotypes remains poorly explored.

In this paper we use the formulation of hidden states as probability distributions to develop a CTMC model for studying the evolutionary dynamics of ecological resource use on phylogenetic trees. Our approach is explicitly designed to model resource traits that are intra-specifically variable and to account for uncertainty in ecological state assignments of terminal taxa arising from effects of sampling variation. We assume that each state is an unobserved (latent) multinomial distribution and that observed data are sampled outcomes from these latent distributions (see panels (i) through (iii) of Fig. 1). The number of states in the model and the states themselves are not directly observed and are estimated from the data. Using simulations and an empirical dataset of snake diets, we show how the method can use observational counts to simultaneously infer the number of resource states, the proportional utilization of resources by different states, and the phylogenetic distribution of ecological states among living species and their ancestors. The method is general and applicable to any data expressible as a set of observational counts from different resource categories.

MATERIALS & METHODS

Model description

We assume that the data for each species are represented by a vector of J category counts. Each category is a resource (e.g. a diet or habitat component), and each count represents the number of observations of a species utilizing a particular resource. Each node in a phylogeny is to be placed into one of K distinct resource states. States are unobserved, even at the tips of a phylogeny: the observed data consist of sampled outcomes from an underlying (latent) multinomial distribution that represents the state for each species. All count data are drawn independently from their respective states and counts from each state are also independent of one

another. We assume that the multinomial parameters for each state are drawn from a common Dirichlet distribution with parameter β. This parameterization allows us to analytically marginalize over the unknown multinomial parameters underlying each state so that the likelihood of the observed data is the product of K independent Dirichlet-multinomial distributions (Appendix). The parameter β acts as a vector of pseudo-counts. Higher values require more data for the model to discriminate two samples as having originated from different states. Letting X denote the resource state assignments for nodes in the phylogeny, the likelihood of the set of count data D_k generated from state k is,

$$p(D_k|X,\beta) = \frac{\Gamma(J\beta)}{\Gamma(n_k + J\beta)} \frac{\prod_j \Gamma\left(n_k^j + \beta\right)}{\Gamma(\beta)^J} \tag{1}$$

where n_k is the total number of observations generated from state k and n_k^j is the subset of those observations that represent utilization of resource category j. The full likelihood for the count data is just,

$$p(D|X,\beta) = \prod_k p(D_k|X,\beta) \tag{2}$$

This model for count data is closely related to topic models of word composition in a collection of text documents (Blei et al. 2003; Yin and Wang 2014) and to population genetic models of allele frequency composition in a set of populations (e.g., program STRUCTURE: Pritchard et al. 2000). The key difference here is that the state assigned to a taxon is the outcome of evolution and is not independent of the states of other lineages. Conceptually this is similar to phylogenetic threshold models, where the full likelihood combines a probability model for the evolution of an unobserved variable and a probability model for sampling the observed data conditioned on the set of unobserved variables (Felsenstein 2012; Revell 2014). We model evolution as a Poisson process where the rate of change is the same between all states (i.e. there

is no evolutionary trend in the model) but varies among lineages. We introduce two mechanisms for accommodating this rate variation.

The first mechanism takes advantage of the random local clocks model introduced by Drummond and Suchard (2010). In this framework, there is an overall rate of evolution Λ and an unknown number of lineages that deviate from this rate by a set of multiplicative constants. Specifically, the root node is defined to have a relative rate of 1 while the relative rates of all other nodes are equal to the relative rate of their ancestor multiplied by a branch-specific positive rate multiplier. Complexity is controlled through the use of a prior that makes it unlikely for many of these multipliers to differ from unity. Under the fully symmetric Poisson model with random local clocks the probability of change across an ancestral-descendant branch is,

$$p\left(X_i \middle| X_{pa(i)}, \Lambda, r_i\right) = \left(1 - e^{-K\Lambda r_i t_i}\right)\frac{1}{K} + \delta_{X_i X_{pa(i)}} e^{-K\Lambda r_i t_i} \tag{3}$$

where X_i is the state of node i, $X_{pa(i)}$ is the state of i's ancestor, r_i is the branch-specific normalized relative rate of evolution, t_i is the length of the branch in units of time, and $\delta_{X_i X_{pa(i)}}$ is the Kronecker delta. For simplicity, we may occasionally notate transition probabilities by p_{ii} and p_{ij} to indicate the probability that a descendant's state is the same as, or different than, the state of its ancestor. The likelihood of the node states is just,

$$p(X|\Lambda, r) = \prod_i p\left(X_i \middle| X_{pa(i)}, \Lambda, r_i\right) \tag{4}$$

where the product is taken over all nodes and $p\left(X_i \middle| X_{pa(i)}, \Lambda, r_i\right) \equiv \frac{1}{K}$ if node i is the root. The normalized rates effectively expand and contract the temporal durations of the branches. They are derived by scaling the relative rates in such a way that the total absolute time in which evolution has had to occur is held constant even as the effective time is allowed to vary over phylogeny (that is, $\sum_i t_i = \sum_i r_i t_i$).

The second mechanism for accommodating rate heterogeneity is essentially a saturated version of the random local clocks model where each branch has a unique rate of evolution. Following Huelsenbeck et al. (2008), this allows us to model branch-specific rates as nuisance parameters drawn independently from a Gamma distribution with parameter vector $(\alpha, 1)$. This model induces the same distribution of node states as a model where the number of expected character state changes along a branch is the same for all branches (Appendix). This has elsewhere been termed the ultra-common mechanism model (Steel 2011) to mark its contrast with the no-common mechanism model (Tuffley and Steel 1997) from which it derives. In this case the probability of change across an ancestral-descendant branch is,

$$p\left(X_i \middle| X_{pa(i)}, \alpha\right) = \left(1 - \frac{1}{(K/(K-1)+1)^{\alpha}}\right)\frac{1}{K} + \delta_{X_i X_{pa(i)}} \frac{1}{(K/(K-1)+1)^{\alpha}} \tag{5}$$

Phylogenetic signal is controlled by the parameter α, which is equal to the expected number of state changes that occur from ancestor to descendant. As $\alpha \to 0$, phylogenetic signal approaches 1 because descendants almost surely resemble their ancestors. As $\alpha \to \infty$, phylogenetic signal approaches 0 because a descendant's state becomes independent of its ancestor's state and resembles a random draw from a discrete uniform distribution. The likelihood of the node states is just,

$$p(X|\alpha) = \frac{1}{K} {p_{ii}}^{n} {p_{ij}}^{m} \tag{6}$$

where *n* is the number of nodes with the same state as their ancestor, *m* is the number of nodes with a different state than their ancestor, and the factor $\frac{1}{K}$ accounts for the root state probability.

Bayesian inference

We simulated the posterior distribution of node states and model parameters using the Metropolis-Hastings algorithm (Hastings 1970). The different proposal mechanisms are described below.

Updating node states.— A Gibbs update mechanism is used for proposing changes to node states. The full conditional distribution for the state of a node can be written as,

$$p(X_i|X_{\neg i}, D, \beta, \theta) = \frac{p(D, X|\beta, \theta)}{p(D, X_{\neg i}|\beta, \theta)}$$

$$\propto \frac{p(D, X|\beta, \theta)}{p(D_{\neg i}, X_{\neg i}|\beta, \theta)}$$

$$= \frac{p(D|X, \beta)}{p(D_{\neg i}|X_{\neg i}, \beta)} \frac{p(X|\theta)}{p(X_{\neg i}|\theta)} \tag{7}$$

where the symbol $\neg i$ denotes the exclusion of node i. Here, depending on whether (4) or (6) is used as the likelihood model for node states, θ is equal to (Λ, r) or α, respectively. Because changing a node state only affects the branches incident to the affected node, all terms in the ratio $\frac{p(X|\theta)}{p(X_{\neg i}|\theta)}$ not involving those branches cancel and it simplifies to,

$$\frac{p(X|\theta)}{p(X_{\neg i}|\theta)} = p\left(X_i|X_{pa(i)}, \theta\right) \prod_k p\left(X_{ch(i,k)}|X_i, \theta\right)$$

where $X_{ch(i,k)}$ is the state of the k-th immediate descendant of node i. Similarly, altering the state of a node only affects the likelihood of count data associated with a single state so that factors in the ratio $\frac{p(D|\beta)}{p(D_{\neg i}|\beta)}$ not involving, say, state k cancel, and it simplifies to,

$$\frac{p(D|X, \beta)}{p(D_{\neg i}|X_{\neg i}, \beta)} = \frac{p(D_k|X, \beta)}{p(D_{k\neg i}|X_{\neg i}, \beta)}$$

$$= \frac{\prod_j \Gamma\left(n_{k\neg i}^j + N_i^j + \beta\right)}{\prod_j \Gamma\left(n_{k\neg i}^j + \beta\right)} \frac{\Gamma(n_{k\neg i} + J\beta)}{\Gamma(n_{k\neg i} + N_i + J\beta)}$$

where N_i is the total number of observations for node i and N_i^j is the subset of those observations that represent utilization of resource category j. Note that for nodes with no associated count data (which includes all internal nodes), this factor is equal to 1. To perform the Gibbs update in practice, we calculate the conditional likelihood for each diet state according to equation (7) and choose a state with probability proportional to its conditional likelihood, using the sum of all conditional likelihoods as a normalizing constant. The marginal posterior probability that a node is in a given diet state is simply the fraction of posterior samples where it appears in that state.

Once a state is sampled for a node any count data associated with that node are added to the set of count data generated from the sampled state. Because the Dirichlet distribution is conjugate to the multinomial distribution, the posterior distribution of the multinomial distribution underlying each state is also Dirichlet distributed with parameter $\left(n_k^1 + \beta, \dots, n_k^J + \beta\right)$. During the course of updating node states we keep track of the average expected proportional utilization of each resource by each state. The expected proportional utilization of resources is simply the mean of the posterior distribution which is $\left(\frac{n_k^1+\beta}{\sum_j n_k^j + J\beta}, \dots, \frac{n_k^J+\beta}{\sum_j n_k^j + J\beta}\right)$.

Updating β.—The symmetric hyperparameter β controls the shape of the Dirichlet prior distribution on the latent multinomial distributions underlying each resource state. When $\beta = 1$ the distribution is uniform over the J-dimensional simplex of resources. When $\beta < 1$ the distribution concentrates toward the corners of the simplex, and when $\beta > 1$ the distribution concentrates toward the center. Because empirical datasets are typically sparse with many zeros, we assume that β is uniformly distributed on the interval (0, 1) and update its value using a sliding window proposal mechanism. The prior and proposal ratios are 1.

Updating α.—When equation (6) is used to compute the likelihood the hyperparameter α controls phylogenetic signal. Although it can take on any positive value the likelihood surface plateaus relatively quickly as its magnitude increases and phylogenetic signal decays. By solving the logarithm of (6) for the maximum likelihood estimate of α we find that

$$\hat{\alpha} = -\frac{\log\left(\frac{fK-1}{K-1}\right)}{\log(K/(K-1)+1)}$$

where f is the fraction of nodes that have the same state as their ancestor. Values of $f \leq \frac{1}{K}$ are consistent with infinite values of α. We therefore bound α above by the value $-\log\left(\frac{\frac{\lfloor N/K \rfloor+1}{N}K-1}{K-1}\right)/\log(K/(K-1)+1)$, where N is the number of nodes (not including the root) in the phylogeny. We assume that α is uniformly distributed between zero and this upper value and update its value using a sliding window proposal mechanism. The prior and proposal ratios are 1.

Updating Λ.—When equation (4) is used to compute the likelihood the parameter Λ controls phylogenetic signal. As for α, we want a reasonable upper bound for Λ because the likelihood plateaus relatively quickly as the rate increases. By replacing branch lengths in equation (4) with the average branch length and solving for the maximum likelihood estimate of Λ under the assumption of no rate variation we find that

$$\hat{\Lambda} = -\frac{\log\left(\frac{fK-1}{K-1}\right)}{K\bar{t}}$$

where $\bar{t}$ is the average branch length. As before, values of $f \leq \frac{1}{K}$ are consistent with infinite values of Λ. We therefore bound Λ above by the value $-\log\left(\frac{\frac{\lfloor N/K \rfloor+1}{N}K-1}{K-1}\right)/K\bar{t}$. We

assume that Λ is uniformly distributed between zero and this upper value and update its value using a sliding window proposal mechanism. The prior and proposal ratios are 1. We use the same prior distributions and proposal mechanisms detailed by Drummond and Suchard (2010) for updating the number of local random clocks and the rate multipliers associated with local clocks.

Implementation

Functions for fitting the model to data are provided as an R package available from github.com/blueraleigh/phyr. The package includes two R functions that call compiled C programs implementing the random local clocks and ultra-common mechanism models.

Simulation study

To illustrate application of the method we designed a simulation study using an empirical dataset on pseudoboine snake diets (Alencar et al. 2013). Our rationale for basing simulations on an empirical dataset is to ensure that properties of the data used to evaluate performance of the method are consistent with real studies, especially the distribution of observations per taxon and the distribution of resource specialization. Pseudoboine snakes are common members of the squamate communities found in lowland rainforests of South America. Predominantly terrestrial or semi-arboreal, these snakes mainly eat small mammals, lizards, and other snakes. The dataset includes 606 observations of prey items from 8 prey categories for 32 species of pseudoboine snakes. Per species sample sizes range from 1 to 56 observations (or 0.125-fold to seven-fold coverage, where coverage is the number of observations divided by the number of resource categories). We reanalyzed these data using a 33-species pseudoboine phylogeny extracted from

the posterior distribution of trees in Tonini et al. (2016). The dataset is illustrated in Figure 2. The original publication coded each species with at least 8 diet observations into a set of 5 specialist diet states and 1 generalist diet state. Species were considered specialists if the resource represented at least 70 percent of recorded prey items (as in our Figure 1). When we applied the resampling procedure illustrated in Figure 1 to this empirical dataset under the assumption that the original state assignments represented the "truth", we found that in approximately 20 percent of resampled datasets at least one original state (not always the same state) was not present and that in about 84 percent of cases at least one species was coded incorrectly (although overall coding accuracy was high, ranging from 0.77 to 1). Thus, this dataset illustrates some of the concerns raised in our introduction but is also well-sampled enough and shows enough variation to facilitate the estimation of separate multinomial distributions.

Simulated datasets were generated from $K = 2, 3, 4$, and 5 diet states using the empirical sample size distribution with the original 8 food resource categories. For each K we first performed Bayesian inference under the ultra-common mechanism model to estimate the unobserved multinomial distributions. The estimated multinomial distributions were subsequently used to simulate diet observations. For each K we simulated 20 datasets at each of 7 different levels of phylogenetic signal (0.1, 0.3, 0.5, 0.6, 0.7, 0.8, and 0.9) using the transition probabilities in both equations (3) and (5), resulting in 560 datasets for each model and 1,120 datasets altogether. We defined phylogenetic signal as $p_{ii} - p_{ji}$, which ranges from 0 to 1 and quantifies how much information a descendant's state provides about the state of its ancestor (Royer-Carenzi et al. 2013). Using equation (5) for transition probabilities results in phylogenetic signal equal to $\left(\frac{1}{K/(K-1)+1}\right)^{\alpha}$. We used this result to calculate the value of α for each simulation.

When equation (3) is used for transition probabilities each branch has a unique phylogenetic signal. Because phylogenetic signal is a convex function of branch length, the average phylogenetic signal of all branches is greater than or equal to the phylogenetic signal of the average branch, which is $e^{-K\Lambda\bar{t}}$. We used the phylogenetic signal of the average branch to calculate the value of Λ for each simulation, which we applied to all branches (i.e., datasets did not include random local clock variation). Interestingly, for a given branch length (measured as expected number of state changes) phylogenetic signal with equation (5) is always greater than phylogenetic signal with equation (3), suggesting that estimating the rate of evolution trades off with estimating ancestral node states (Gascuel and Steel 2018). For each simulated dataset we ran a set of Markov chains with 1, 2, …, up to K+3 diet states. Each chain was run for 160,000 iterations after a burnin of 30,000 iterations, sampling every 128 iterations to yield 1,250 posterior samples.

Determining the number of resource states

Because the model does not include a process for generating the number of states, we must perform analyses across multiple values of K and apply an *a posteriori* inference procedure to choose between them. A similar problem is encountered when trying to infer the number of demes from multi-locus genotype data with the program STRUCTURE (Pritchard et al. 2000). Our approach is to choose the largest value of K for which all states are unambiguously assigned to at least one terminal taxon. Specifically, by examining the terminal nodes of the phylogeny we can determine the maximum marginal posterior probability assigned to each state across all terminal nodes. Call the smallest of these q_K. A low value of q_K implies that at least one state is not assigned to any individuals (terminal taxa) with high probability. As K varies from low to

high there will come a point when additional states become redundant to previous states. When that occurs, q_K will drop well below 1. In other words, a point will be reached when at least one state is never unambiguously assigned to at least one terminal node. To choose the best value for K we first identify the K with the steepest drop in q_K with respect to q_{K-1}. $K^* = K - 1$ then becomes a candidate best choice. If $q_{K^*} = 1$ (or nearly so) we treat K^* as the best choice. Otherwise, we keep setting $K^* = K^* - 1$ until $q_{K^*} = 1$. This procedure, which we call the q_K rule, is illustrated in Figure 3 on the empirical dataset.

Assessing model adequacy

In practice, the number of states identified by the q_K rule will depend on the underlying data and may change as data are updated. This is because the model may be unable to distinguish truly different samples as having arisen from separate distributions if sample sizes are small. Therefore, empirical applications of the method require a way of assessing how well the inferred multinomial distributions explain the empirical resource observations. Our approach is to compute a per-species adequacy score that measures the similarity of each taxon, with respect to sampled observations, to other taxa assigned to the same state. The procedure we describe effectively measures the compositional heterogeneity of the model-inferred states with respect to sampled diet observations. If the model is fully adequate and accurately describes patterns of resource use, then all species assigned to a given state will have similar sampled diets (e.g., the observed data). However, some species may be assigned to states even where they have dissimilar sampled observations from other species in the same state, reflecting overdispersed diet distributions for the state. Such overdispersion might arise if species are assigned to the "wrong" state. These incorrect state assignments might be preferred under the model if a species

does not have enough observations to provide information about the existence of a distinct state, and so the species is conservatively assigned to the immediately ancestral diet state. To compute the adequacy score, we first draw a state for each terminal node from its array of marginal posterior probabilities computed using equation (7) (which utilizes count data and evolutionary history). Then we visit each terminal node in turn and perform the following exercise. Given the configuration of states for all terminal nodes, we identify the set of count data generated by the state of the current node and compute the likelihood of these data using equation (1) (which only utilizes count data). Next, we compute the likelihood of these same data, using equation (1) together with equation (2), but assuming that they were generated from two states by placing the current node in its own unique state. Finally, we take the negative log likelihood ratio of these two values. We repeat this procedure for a thousand independent configurations and record the average negative log likelihood ratio (adequacy score) for each terminal node. Large negative adequacy scores highlight terminal nodes that were assigned to a state whose other members have a strongly dissimilar pattern of resource utilization. This procedure is illustrated in Figure 4 on the empirical dataset (see also Figure S4).

RESULTS

Overall, the q_K rule correctly identified the number of resource states in 492 of 560 simulations from the ultra-common mechanism model (Fig. 5). In the 68 cases where the method incorrectly identified the number of states, it underestimated the number of states by one (61 instances), two (4 instances), and three states (2 instance) and overestimated the number of states by one state in one instance. When the q_K rule was used with the random local clocks model it correctly identified the number of states in 475 of 560 simulations (Fig. S1). In the 85 cases

where the method incorrectly identified the number of states, it underestimated the number of states by one (77 instances) and two states (8 instances). Failure to correctly identify the number of states commonly occurs when the number of observations generated by a state is small relative to the number of observations from other states. This happens when the terminal nodes representing a state have poorly sampled diets causing the state to be subsumed into the state of near relatives.

Estimation of latent multinomial probabilities underlying resource states was highly accurate (Figs. 5, S1, S5). In simulations from both the ultra-common mechanism model and the random local clocks model, the correlation between true and estimated probabilities ranged from 0.92 to 0.999 with a mean of 0.99. Such high values occur because even when the number of states is underestimated, the number of sampled observations from the missing states are few enough in number that they do not appreciably alter the estimated proportions of the other states (Fig S6).

Across all levels of phylogenetic signal, and with simulations from both the ultra-common mechanism model and the random local clocks model, the method consistently classified greater than ninety-percent of terminal nodes in the correct state. For internal nodes, reconstruction accuracy was comparable to reconstruction accuracy for terminal nodes at the highest level of phylogenetic signal but, in accordance with expectation, decayed toward the expected accuracy of a random guess as phylogenetic signal declined toward zero (Figs. 6, S2).

This behavior was mirrored in the posterior distributions of α and Λ (Figs. 7, S3). When phylogenetic signal was low, posterior distributions of rate estimates were diffuse and resembled their uniform prior distributions except that they were shifted away from the lower bound. As

phylogenetic signal increased, posterior distributions of rate estimates concentrated around the true values used to generate simulated datasets.

DISCUSSION

We developed a comparative method for macroevolutionary analysis of multivariate count data. The method is general and may be applied to any data expressible as a set of observational counts from different categories. Such datatypes arise frequently in community ecology and behavior. Potential applications include the comparative analysis of diet, foraging behavior, activity patterns, and habitat preferences. The method is similar to standard continuous-time Markov chain models of phenotypic evolution but differs in several important respects. First, the number of states in the model and the states themselves are unobserved and must be estimated from empirical data on resource use. Second, each state is an unobserved multinomial distribution rather than a categorical variable. This latter property enables researchers to model ecological traits that show intraspecific variation and to account for uncertainty in the state assignments of terminal taxa that arises from the effects of sampling variation.

Simulations revealed that the new method is generally able to determine the correct number of states and that it provides accurate estimates of the underlying (unobserved) multinomial distributions, both for terminal taxa as well as internal nodes. We designed simulations around empirical patterns of resource use in a dataset on snake diets (Alencar et al. 2013). Therefore, caution is warranted in generalizing the good performance observed in the current study to other datasets. In particular, performance of the model will depend on the idiosyncrasies of individual datasets, including the distribution of sample sizes and the

distribution of overlaps in resource use among species. We expect that states represented by few observations will be difficult to infer, especially if those states show appreciable overlap with other states.

Our empirical analysis identified at least two feeding modalities among the set of species Alencar et al. (2013) recognize as "generalists": species that feed predominantly on snakes but that regularly eat lizards and mammals, and species that feed predominantly on mammals and lizards. Ancestral state estimates strongly suggest that each of these feeding modalities arose from a more specialized diet comprised almost entirely of lizards. This is in contrast to the results of Alencar et al. (2013), which imply that nearly all origins of specialized feeding modalities occurred from a generalist ancestor, although direct comparison of results is made difficult by the use of different phylogenies in the two studies.

As currently implemented, the approach described here does not directly model gains and losses or substitutions of different resources. Indeed, no resource is ever truly absent from the reconstructed states (although its proportional representation may approach zero as β becomes small). This contrasts with biogeographic-type models that explicitly model resource use expansions, contractions, and substitutions (e.g. see Hardy (2017) for application of Ree and Smith's (2008) dispersal-extinction-cladogenesis model to binary encoded diet data). Although these types of changes are implicit in the sequence of reconstructed states derived from the model, future studies might want to explore how to combine more complex evolutionary models with the current model for count data. Nonetheless, the advantage of a simple evolutionary model is that it has broad scope. It would be possible, for example, to apply our method to continuous characters by keeping the same evolutionary model but changing the model for observations from a Dirichlet-multinomial to a multivariate-normal distribution, which could

then be applied to other data types used for quantifying resource use such as stable isotope ratios of carbon and nitrogen.

One challenge for comparative methods is their limited ability to model ecological phenotypes that cannot be neatly summarized by a single value (Hardy & Linder 2005; Hardy 2006). Recent years have seen progress in this direction for continuous traits, including models that accommodate intraspecific variation, function-valued traits, and other non-gaussian data (Ives et al. 2007; Felsenstein 2008; Evans et al. 2009; Jones & Moriarty 2013; Goolsby 2015; Quintero et al. 2015). The general approach developed here, where each state is a multinomial distribution rather than a categorical variable, extends this progress to traits like diet and habitat that are typically treated as categorical variables. By placing an emphasis on individual natural history observations, the method draws attention to the central role such observations play in evolutionary biology (Greene 1986) and to the many remaining opportunities for developing comprehensive ecological databases that advance our understanding of biodiversity (Hortal et al. 2015).

SUMMARY

We described a novel methodological framework for studying the evolutionary dynamics of complex ecological traits on phylogenetic trees. Previous approaches to this problem have assumed that ecological states are categorical variables and that species are monomorphic for particular states. We relaxed this assumption through the use of a hidden Markov model that treats ecological states as unobserved probability distributions from which observed data are sampled. Although our method is designed for the analysis of multivariate count data, we suggest

that the approach of treating states as hidden probability distributions has wide applicability and will greatly facilitate the comparative analysis of novel sources of ecological data.

ACKNOWLEDGEMENTS

This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship and by the University of Michigan Department of Ecology and Evolutionary Biology Block Grant.

LITERATURE CITED

Alencar L.R.V., Gaiarsa M.P., Martins M. 2013. The evolution of diet and microhabitat use in pseudoboine snakes. South American Journal of Herpetology. 8:60–66.

Bassar R.D., Marshall M.C., López-Sepulcre A., Zandona E., Auer S.K., Travis J., Pringle C.M., Flecker A.S., Thomas S.A., Fraser D.F., Reznick D.N. 2010. Local adaptation in Trinidadian guppies alters ecosystem processes. PNAS. 107:3616–3621.

Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology. 65:583–601.

Beaulieu J.M., O'Meara B.C., Donoghue M.J. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in Campanulid angiosperms. Systematic Biology. 62:725–737.

Blei D.M., Ng A.Y., Jordan M.I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3:993–1022.

Burin G., Kissling W.D., Guimarães P.R., Sekercioglu C.H., Quental T.B. Omnivory in birds is a macroevolutionary sink. Nature Communications. 7:11250.

Caetano D.S., Beaulieu J.M., O'Meara B.C. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution. 72:2308–2324.

Cantalapiedra J.L., FitzJohn R.G., Kuhn T.S., Hernández Fernández M., DeMiguel D., Azanza B., Morales J., Mooers A.Ø. 2014. Dietary innovations spurred the diversification of ruminants during the Caenozoic. Proceedings of the Royal Society B: Biological Sciences. 281:20132746.

Colston T.J., Costa G.C., Vitt L.J. 2010. Snake diets and the deep history hypothesis. Biological Journal of the Linnean Society. 101:476–486.

Davis A.M., Unmack P.J., Vari R.P., Betancur-R R. 2016. Herbivory promotes dental disparification and macroevolutionary dynamics in grunters (Teleostei: Terapontidae), a freshwater adaptive radiation. The American Naturalist. 187:320–333.

Drummond A.J., Suchard M.A. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology. 8:114.

Ehrlich P.R., Raven P.H. 1964. Butterflies and plants: a study in coevolution. Evolution. 18:586–608.

Evans M.E.K., Smith S.A., Flynn R.S., Donoghue M.J. 2009. Climate, niche evolution, and diversification of the "bird-cage" evening primroses (*Oenothera*, Sections *Anogra* and *Kleinia*). The American Naturalist. 173:225–240.

Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. The American Naturalist. 171:713–725.

Felsenstein J. 2012. A comparative method for both discrete and continuous characters using the threshold model. The American Naturalist. 179:145–156.

Forister M.L., Dyer L.A., Singer M.S., Stireman III J.O., Lill J.T. 2012. Revisiting the evolution of ecological specialization, with emphasis on insect-plant interactions. Ecology. 93:981–991.

Futuyma D.J., Moreno G. 1988. The evolution of ecological specialization. Annual Review of Ecology and Systematics. 19:207–233.

Gajdzik L., Aguilar-Medrano R., Frédérich B. 2019. Diversification and functional evolution of reef fish feeding guilds. Ecology Letters. 22:572–582.

Gascuel O., Steel M. 2018. A Darwinian uncertainty principle. bioRxiv. https://doi.org/10.1101/506535

Gillespie R. 2004. Community assembly through adaptive radiation in Hawaiian spiders. Science. 303:356–359.

Givnish T.J., Barfuss M.H.J., Van Ee B., Riina R., Schulte K., Horres R., Gonsiska P.A., Jabaily R.S., Crayn D.M., Smith J.A., Winter K., Brown G.K., Evans T.M., Holst B.K., Luther H., Till W., Zizka G., Berry P.E., Sytsma K.J. 2014. Adaptive radiation, correlated and contingent evolution, and net diversification in Bromeliaceae. Molecular Phylogenetics and Evolution. 71:55–78.

Goolsby E.W. 2015. Phylogenetic Comparative Methods for Evaluating the Evolutionary History of Function-Valued Traits. Systematic Biology. 64:568–578.

Greene H.W. 1986. Natural history and evolutionary biology. In: Feder M.E., Lauder G.V., editors. Predator–prey relationships: perspectives and approaches from the study of lower vertebrates. University of Chicago Press. p. 198.

Hardy C.R. 2006. Reconstructing ancestral ecologies: challenges and possible solutions. Diversity and Distributions. 12:7–19.

Hardy C.R., Linder H.P. 2005. Intraspecific variability and timing in ancestral ecology reconstruction: a test case from the cape flora. Systematic Biology. 54:299–316.

Hardy N.B. 2017. Do plant-eating insect lineages pass through phases of host-use generalism during speciation and host switching? Phylogenetic evidence. Evolution. 71:2100–2109.

Hardy N.B., Otto S.P. 2014. Specialization and generalization in the diversification of phytophagous insects: tests of the musical chairs and oscillation hypotheses. Proceedings of the Royal Society B: Biological Sciences. 281:20132960.

Harmon, L. J., Matthews B., Des Roches S., Chase J.M., Shurin J.B., Schluter D. 2009. Evolutionary diversification in stickleback affects ecosystem functioning. Nature. 458:1167–1170.

Hastings W.K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. 57:97–109.

Hortal J., de Bello F., Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M., Ladle R.J. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics. 46:523–549.

Huelsenbeck J.P., Ané C., Larget B., Ronquist F. 2008. A Bayesian perspective on a non-parsimonious parsimony model. Systematic Biology. 57:406–419.

Ives A.R., Midford P.E., Garland Jr T. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Systematic Biology. 56:252–270.

Janz N., Nyblom K., Nylin S. 2001. Evolutionary dynamics of host-plant specialization: a case study of the tribe Nymphalini. Evolution. 55:783–796.

Jones N.S., Moriarty J. 2013. Evolutionary inference or function-valued traits: Gaussian process regression on phylogenies. Journal of the Royal Society Interface. 10:20120616.

Kelley S.T., Farrell B.D. 1998. Is specialization a dead end? The phylogeny of host use in *Dendroctonus* bark beetles (Scolytidae). Evolution. 52:1731–1743.

Losos J.B., Leal M., Glor R.E., de Queiroz K., Hertz P.E., Rodríguez Schettino L., Chamizo Lara A., Jackman T.R., Larson A. 2003. Niche lability in the evolution of a Caribbean lizard community. Nature. 424:542–545.

MacArthur R.H. 1958. Population ecology of some warblers of northeastern coniferous forests. Ecology. 39:599–619.

Marazzi B., Ané C., Simon M.F., Delgado-Salinas A., Luckow M., Sanderson M.J. 2012. Locating evolutionary precursors on a phylogenetic tree. Evolution. 66:3918–3930.

Martin C.H., Wainwright P.C. 2011. Trophic novelty is linked to exceptional rates of morphological diversification in two adaptive radiations of *Cyprinodon* pupfish. Evolution. 65:2197–2212.

Miller E.T., Wagner S.K., Harmon L.J., Ricklefs R.E. 2017. Radiating despite a lack of character: ecological divergence among closely related, morphologically similar honeyeaters (Aves: Meliphagidae) co-occurring in arid Australian environments. The American Naturalist. 189:E14–E30.

Mitter C., Farrell B., Wiegmann B. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? The American Naturalist. 132:107–128.

Nosil P. 2002. Transition rates between specialization and generalization in phytophagous insects. Evolution. 56:1701–1706.

O'Meara B.C. 2012. Evolutionary inferences from phylogenies: a review of methods. Annual Review of Ecology and Systematics. 43:267–285.

Price S.A., Hopkins S.S.B., Smith K. K., Roth V.L. 2012. Tempo of trophic evolution and its impact on mammalian diversification. PNAS. 109:7008–7012.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 155:945–959.

Quintero I., Keil P.K., Jetz W., Crawford F.W. 2015. Historical biogeography using species geographical ranges. Systematic Biology. 64:1059–1073.

Ree R.H., Smith S.A. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Systematic Biology. 57:4–14.

Revell L.J. 2014. Ancestral character estimation under the threshold model from quantitative genetics. Evolution. 68:743–759.

Robinson G.S. 1999. HOSTS - a database of the hostplants of the world's Lepidoptera. Nota Lepidopterologica. 22:35–47.

Royer-Carenzi M., Pontarotti P., Didier G. 2013. Choosing the best ancestral character state reconstruction method. Mathematical Biosciences. 242:95–109.

Shine R. 1994. Allometric patterns in the ecology of Australian snakes. Copeia. 1994:851–867.

Steel M. 2011. Can we avoid “SIN” in the house of “no common mechanism”? Systematic Biology. 60:96–109.

Tonini J., Beard K.H., Barbosa Ferreira R., Jetz W., Pyron R.A. 2016. Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. Biological Conservation. 204:23–31.

Tuffley C., Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of Mathematical Biology. 59:581–607.

Vitt L.J., Vangilder L.D. 1983. Ecology of a snake community in northeastern Brazil. Amphibia-Reptilia. 4:273–296.

Vrba E.S. 1987. Ecology in relation to speciation rates: some case histories of Miocene-Recent mammal clades. Evolutionary Ecology. 1:283–300.

Yin J., Wang J. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’14.:233–242.

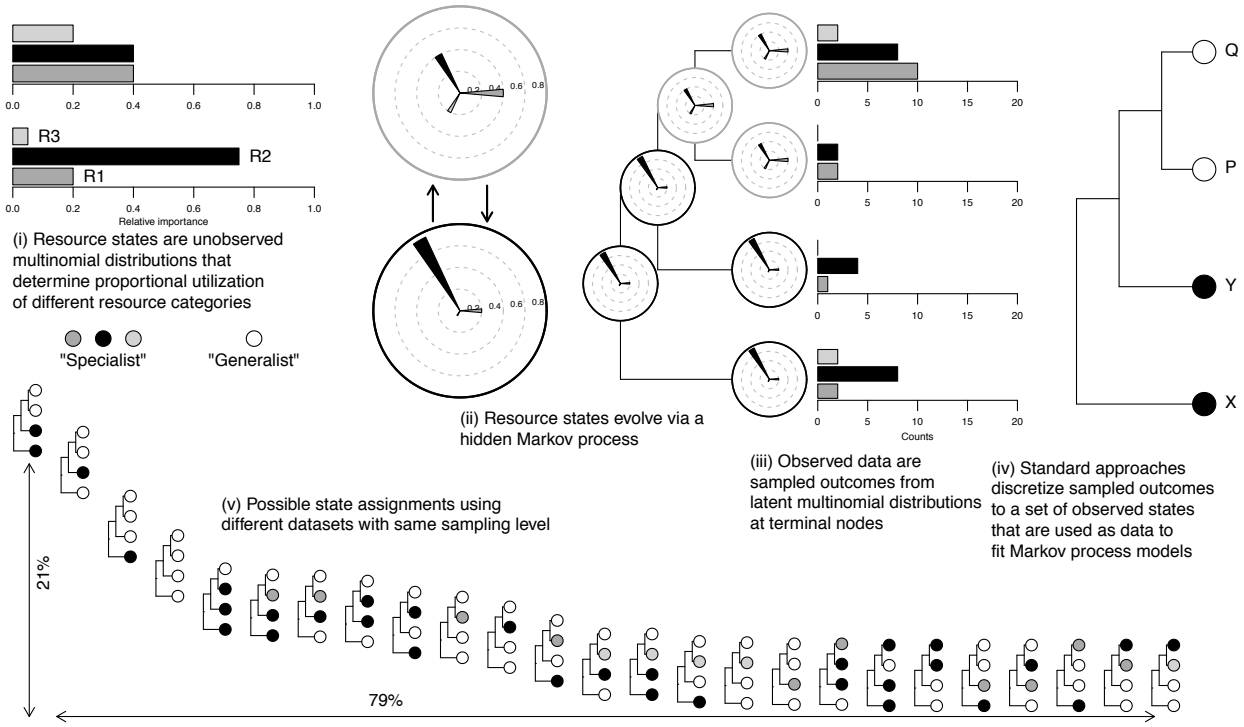
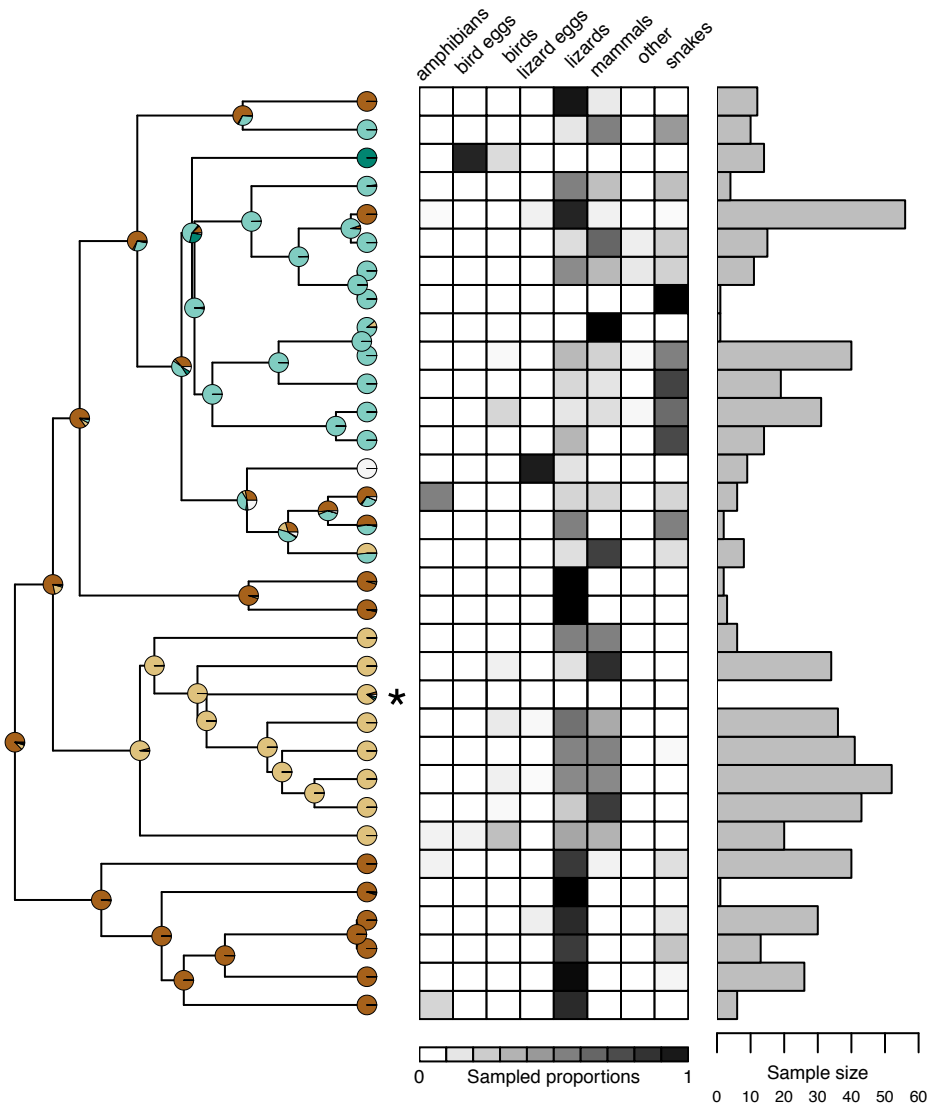


Figure 1. Distribution and representation of multivariate ecological phenotypes (i, ii, iii), data as sampled by researchers (iii), and sampled states as typically represented by univariate categorical traits (iv, v). Loss and distortion of information associated with complex phenotypes motivates the development of the Dirichlet-multinomial model described in this article. (i) True resource states are unobserved multinomial distributions that determine the proportional utilization of three dietary resource categories by four species. (ii) The resource state of a species is the outcome of evolution via a hidden Markov process where the hidden states are the unobserved multinomial distributions. Here, the multinomial distributions from (i) are represented as rose plots: the direction of a spoke identifies the resource category and the length of a spoke is equal to the proportional utilization of that category. The phylogeny depicts the true evolutionary history of change. (iii) Observed data are sampled outcomes from these latent multinomial distributions. (iv) Sampled outcomes are projected onto a set of resource states. Here, a species is considered a “specialist” on a particular category if the sampled proportion of the category

exceeds 0.7. Otherwise, it is considered a "generalist". In this case, the dataset and cutoff value align to match each species with its correct modal resource category. (v) The same state assignment process is repeated with different datasets while keeping the sample sizes for each species identical to (iii). State assignments are sorted along the *x*-axis according to their frequency of occurrence in 1,000 independent datasets. Datasets were generated by sampling from the two multinomial distributions in (i). Note that the procedure correctly matches all species with their modal food resource in a minority of cases and results in a variable number of states across datasets. The implication for macroevolutionary studies using this state assignment procedure is that we cannot be certain whether state assignments are reflective of true patterns of resource use or are merely the expected outcome of sampling variation.



683

684 **Figure 2.** Summary attributes of the snake dietary dataset used to parameterize the simulation
685 study, including phylogeny of pseudoboine snakes (left), relative prey frequencies (middle), and
686 total numbers of food observations per snake species (right). Dark colors in the prey frequency
687 matrix indicate higher sampled proportions of a particular prey item in a given diet. The
688 phylogeny and sampled diet observations were used to infer the "true" (unobserved) diet states
689 and their evolutionary history on the phylogeny (colored circles). Each color in a pie chart is a
690 specific model-inferred state and the size of a slice represents the marginal posterior probability

that a node belongs to that state. Note that the terminal node marked with an asterisk is missing data. It is treated like an internal node and information about its probable diet state is drawn only from what the model has learned about the states of its neighbors and the likelihood of evolutionary change. Here, the Dirichlet-multinomial model inferred 5 states, corresponding to 3 specialist (> 70% specificity for a single prey group) and 2 generalist diets. Note that the diet states are not observed directly, even at the tips of the tree; rather, all observed data are assumed to be sampled from a set of unknown multinomial distributions.

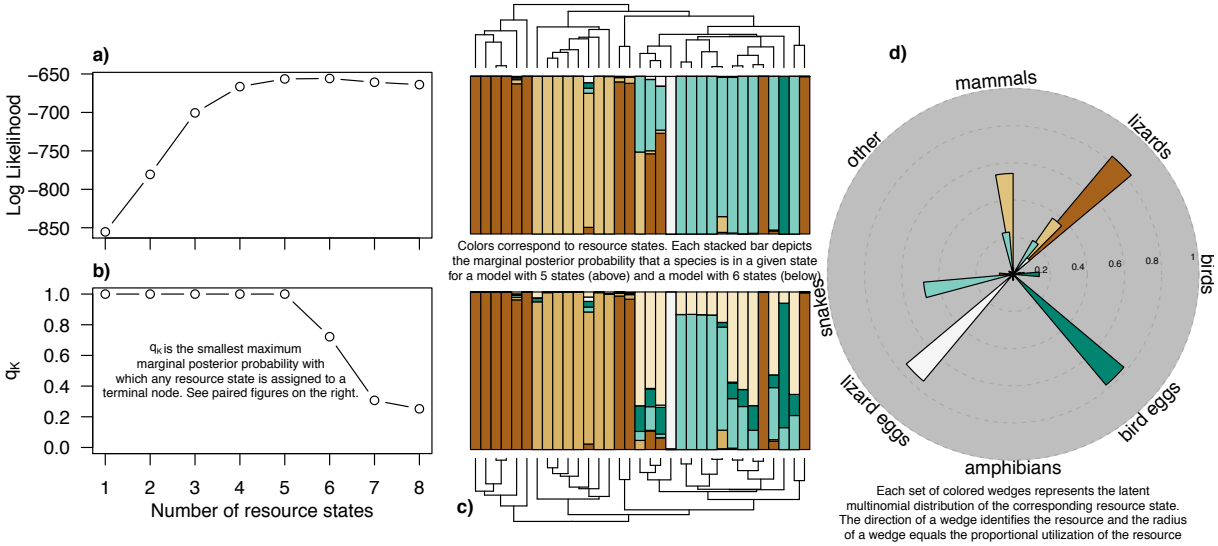


Figure 3. Illustration of the *a posteriori* criterion for determining the number of states in the model. Panel (a) shows the average log likelihood of the empirical data as a function of the number of diet states. Panel (b) depicts how q_K, the smallest maximum marginal posterior probability with which a state is assigned to terminal taxa, changes as a function of the number of states. Inspection of the marginal posterior probabilities reveals that the sixth state is never unambiguously assigned to a terminal node (panels b and c). For this reason, a model with five resource states is considered optimal. The proportional utilization of different food resources by those five states is illustrated by the rose plot in panel (d).

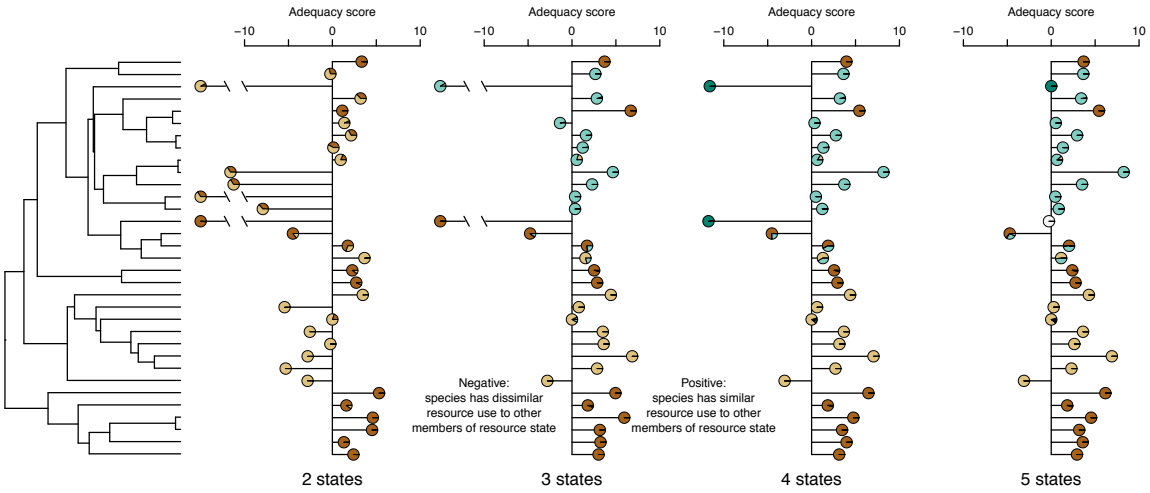


Figure 4. Goodness of fit of inferred multinomial distributions to sampled observations in the empirical snake diet dataset when the inference model assumes $K = 2, 3, 4$, and 5 states. Adequacy scores are computed on a per-species basis; negative scores imply that sampled observations for a species are different from other taxa assigned to the same state. Such negative scores might arise when a species has sampled diet that is distinct from that of close relatives but where insufficient data (few observations) are available to inform the model as to the existence of a new and distinct dietary state. Each line segment is a species in the empirical dataset. Each pie chart depicts the marginal posterior probabilities that a species is in a given diet state. Thus, fitting a model with only 2 states results in a third of the species sharing a diet state with ecologically dissimilar species. In a model with 5 states nearly all species in the empirical dataset share a similar pattern of resource utilization with the other species assigned to the same diet state. The two species with negative scores have a relatively high sampled percentage of amphibian prey items or a relatively broad diet compared to other species in the dataset but only modest sample sizes (compare to Fig. 2).

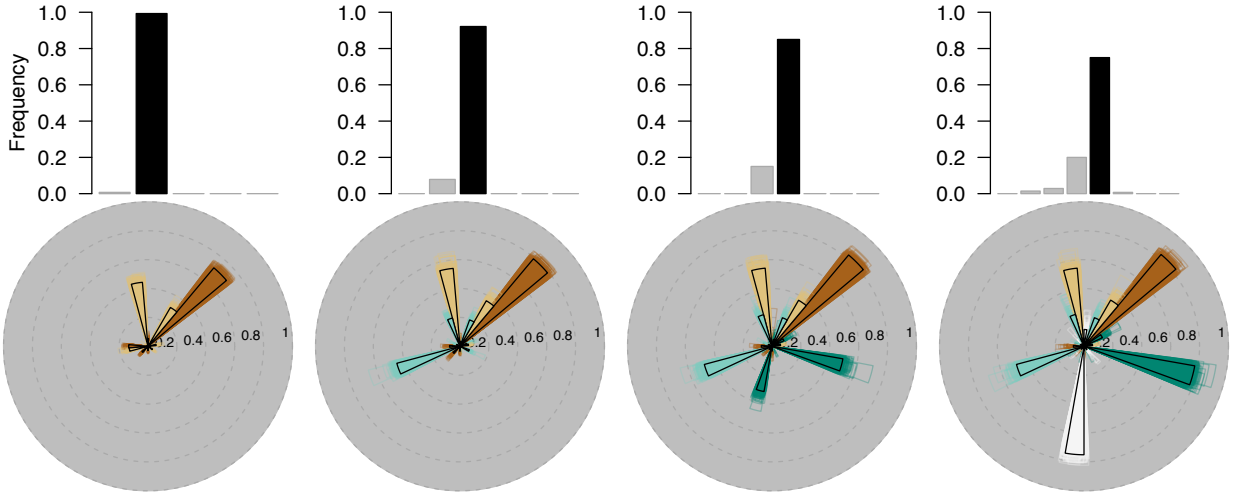


Figure 5. Performance of the method at identifying the true number of states (top) and the latent multinomial distributions underlying states (bottom). Each column summarizes the results of a set of 140 simulated datasets where the number of resource states varied from 2 (leftmost column) to 5 (rightmost column). *Top:* each bar chart shows the frequency of the number of resource states estimated using the q_K rule discussed in the main text; black bars are the number of states in the generating model. The method correctly identifies the number of generating states in most cases. *Bottom*: each set of solid colored spokes represents the latent multinomial distribution underlying a resource state in the generating model. The direction of a spoke identifies its corresponding resource and its length equals the proportional utilization of that resource. Directions are slightly offset between diets states so that spokes with contacting edges represent the same resource. Spokes with colored outlines but unfilled centers are the model estimated multinomial distributions for the corresponding state. The method correctly identifies the major and minor resources of each state in most cases.

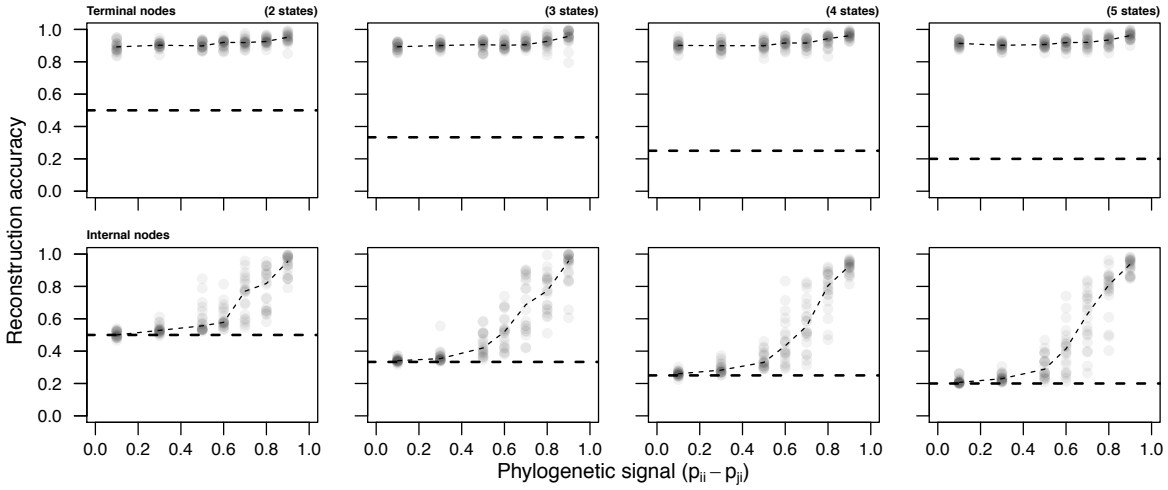


Figure 6. Accuracy of model-inferred resource state assignments for terminal and internal nodes across varying levels of phylogenetic signal. Each point is the reconstruction accuracy for a single simulation. Reconstruction accuracy is the probability that a randomly selected node was classified in the correct resource state. The horizontal dashed lines correspond to the expected reconstruction accuracy of a random guess. The reconstruction accuracy of terminal nodes is relatively constant across levels of phylogenetic signal. The reconstruction accuracy of internal nodes is comparable to the accuracy of terminal nodes when phylogenetic signal is high and, as expected, decays toward the random expectation as phylogenetic signal declines. This decline is expected because, with low signal, the tip data provide little information about states at internal nodes. Light weight dashed lines connect the median accuracy at each level of phylogenetic signal.

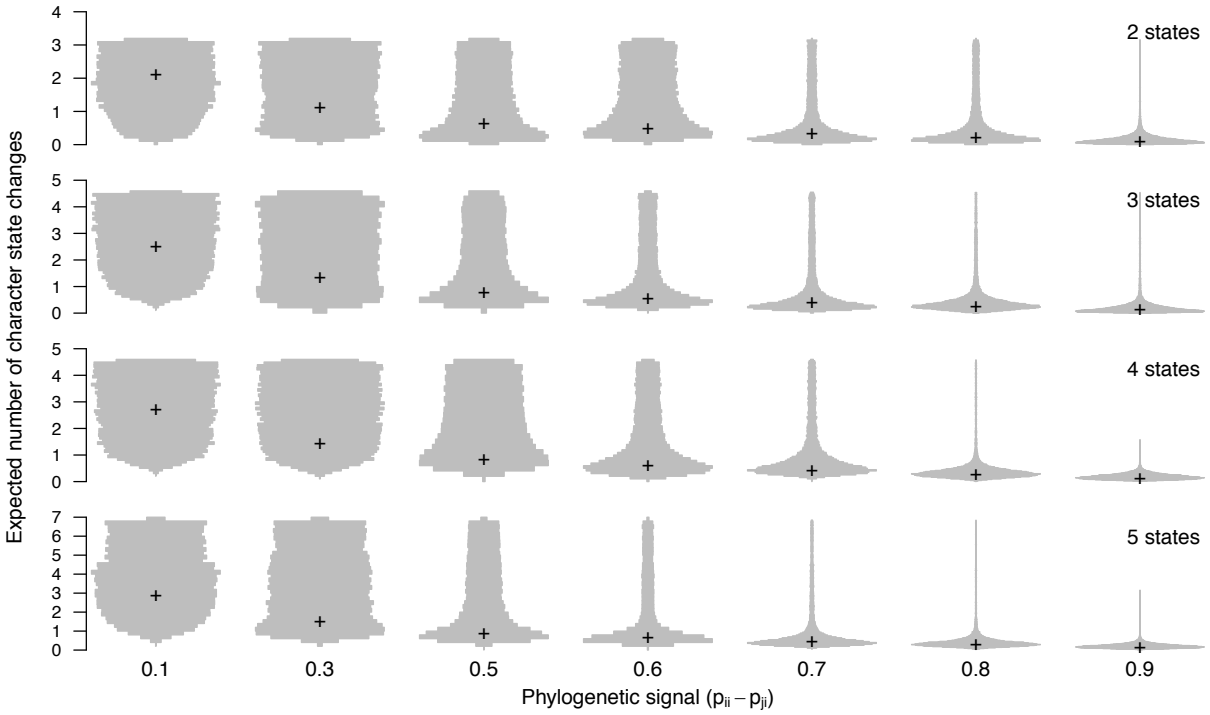


Figure 7. Posterior distributions of the expected number of character state changes per branch for simulated datasets with $K = 2, 3, 4$, and 5 states. Each violin plot summarizes the posterior distributions of 20 simulations at each of 7 different levels of phylogenetic signal. The black hatch mark within each violin is the expected number of changes per branch that was used to generate simulated datasets. The width of a violin measures how frequently estimated values appear in the posterior distributions. Phylogenetic signal increases as the expected number of changes decreases. At low phylogenetic signals posterior distributions largely recapitulate their uniform prior distributions but as phylogenetic signal increases they become concentrated around the true values.

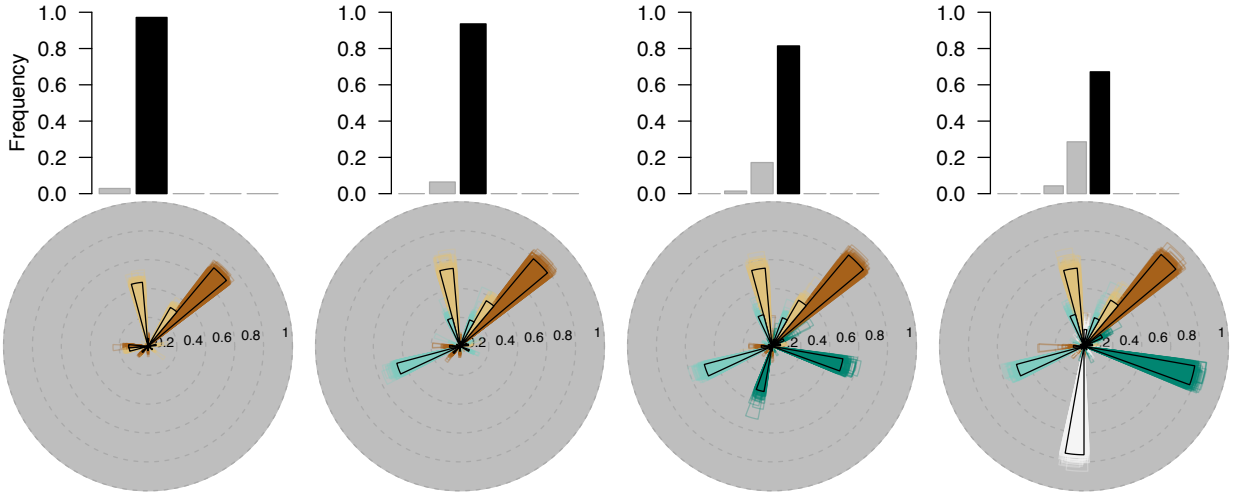


Figure S1. As for Figure 5 in the main text except that simulations were made using transition probabilities from equation (3) rather than from equation (5).

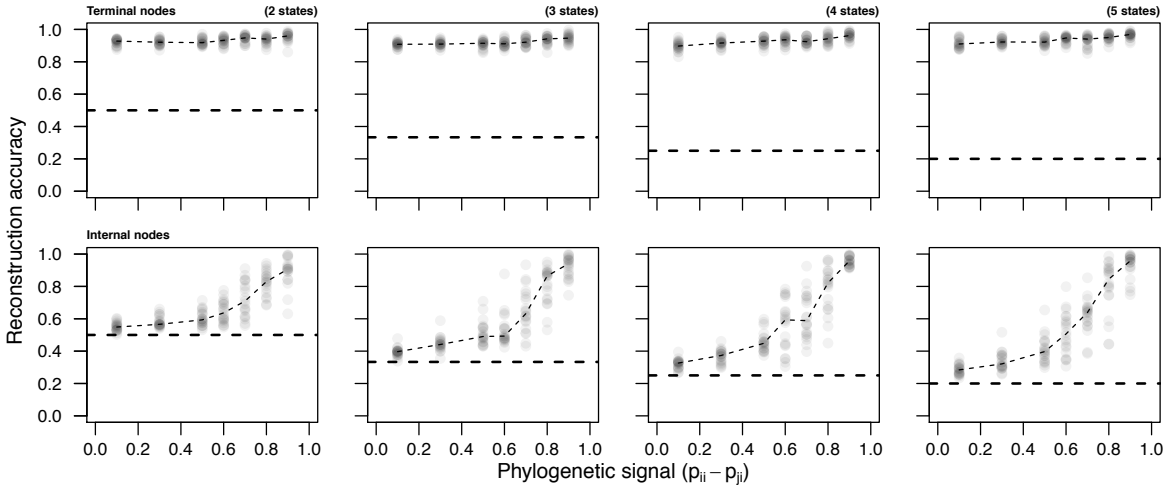


Figure S2. As for Figure 6 in the main text except that simulations were made using transition probabilities from equation (3) rather than from equation (5).

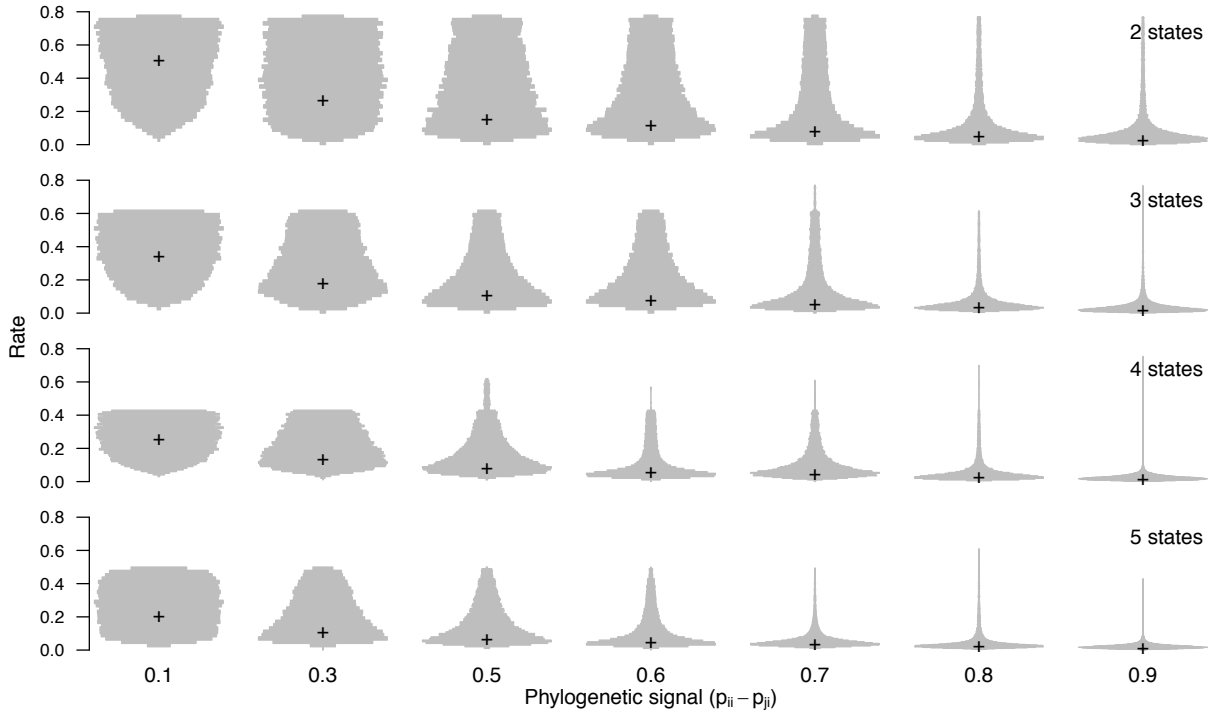


Figure S3. As for Figure 7 in the main text except that simulations were made using transition probabilities from equation (3) rather than from equation (5) and phylogenetic signal refers to the phylogenetic signal of the average branch length.

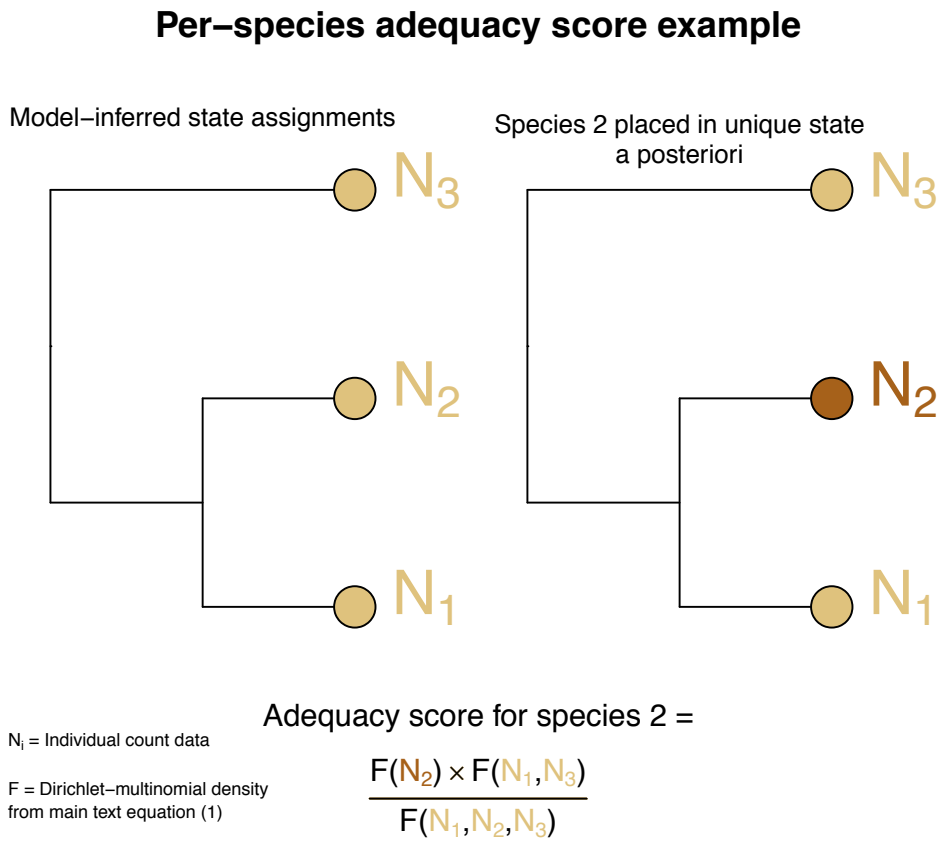


Figure S4. The main text defines a per-species adequacy score to assess how well the estimated multinomial distributions explain sampled observations for terminal taxa. The toy example in this figure depicts how such a score is calculated for species 2. In the main text, we use the negative logarithm of the likelihood ratio calculation depicted in the figure. Thus, positive values for the log likelihood ratio imply that the estimated multinomial is a good fit to sampled observations; negative values imply the opposite.

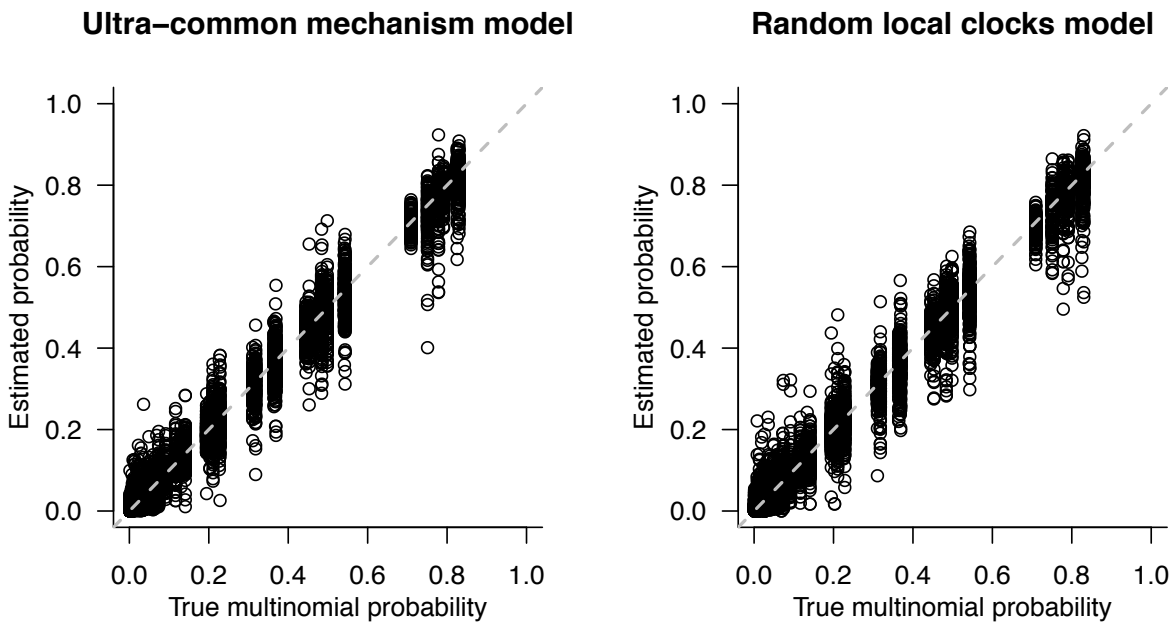


Figure S5. Correlation between true and estimated multinomial probabilities for datasets simulated under the ultra-common mechanism model and random local clocks model.

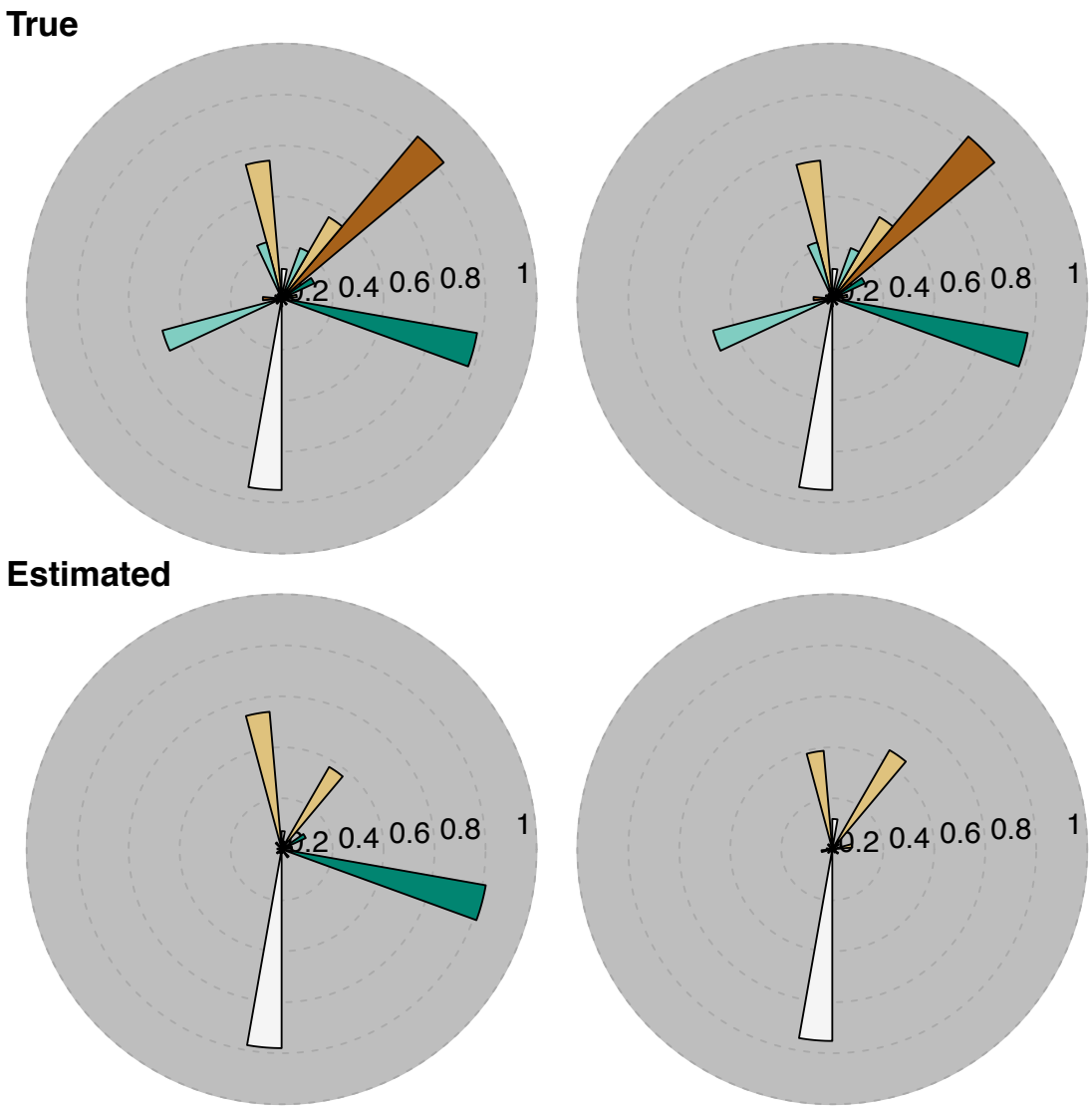


Figure S6. Two examples illustrating how high accuracy of estimated multinomial proportions is maintained even when the number of states is underestimated. In the left plot, the blue state only generated 1 sampled observation and the brown state only generated 5 sampled observations. In the right plot, the brown state only generated 8 observations; the blue state, 12; and the green state, 2. In both cases, these observations are few enough in number that they do not substantially

alter the estimated multinomial proportions when they are incorporated into the count data generated from the other states.

APPENDIX

Derivation of $p(D|X,\beta)$

We assume that conditional on the resource state the data for each species are sampled independently from the latent multinomial distribution underlying the state. That is,

$$p(D_k|X,\theta) \propto \prod_i p(d_i|X_i = k,\theta)$$

$$= \prod_i \prod_j \theta_j^{N_i^j}$$

where we have omitted terms involving multinomial coefficients as well as the dependence of θ on k for ease of notation. We also assume that for each k the multinomial parameter $\theta \sim \text{Dirichlet}(\beta_1, \dots, \beta_J)$ so that,

$$p(D_k|X,\beta) \propto \int_\theta p(D_k|X,\theta)\, p(\theta|\beta) d\theta$$

$$= \int_\theta \left(\prod_i \prod_j \theta_j^{N_i^j}\right) \frac{\Gamma\left(\sum_j \beta_j\right)}{\prod_j \Gamma(\beta_j)} \prod_j \theta_j^{\beta_j - 1}\, d\theta$$

$$= \int_\theta \left(\prod_j \theta_j^{n_k^j}\right) \frac{\Gamma\left(\sum_j \beta_j\right)}{\prod_j \Gamma(\beta_j)} \prod_j \theta_j^{\beta_j - 1}\, d\theta$$

$$= \int_\theta \frac{\Gamma\left(\sum_j \beta_j\right)}{\prod_j \Gamma(\beta_j)} \prod_j \theta_j^{n_k^j + \beta_j - 1}\, d\theta$$

$$= \frac{\Gamma\left(\sum_j \beta_j\right)}{\Gamma\left(n_k + \sum_j \beta_j\right)} \frac{\prod_j \Gamma\left(n_k^j + \beta_j\right)}{\prod_j \Gamma(\beta_j)} \int_\theta \frac{\Gamma\left(n_k + \sum_j \beta_j\right)}{\prod_j \Gamma\left(n_k^j + \beta_j\right)} \prod_j \theta_j^{n_k^j + \beta_j - 1}\, d\theta$$

$$= \frac{\Gamma\left(\sum_j \beta_j\right)}{\Gamma\left(n_k + \sum_j \beta_j\right)} \frac{\prod_j \Gamma\left(n_k^j + \beta_j\right)}{\prod_j \Gamma(\beta_j)}$$

where the last equality follows because the integrand is the density function of a Dirichlet distribution with parameter $\left(n_k^1 + \beta_1, \dots, n_k^J + \beta_J\right)$. By assuming that $\beta_j = \beta$ for all *j* this reduces to equation (1) in the main text.

Derivation of $p(X|\alpha)$

Under the fully symmetric Poisson model the probability of change across an ancestral-descendant branch is,

$$p\left(X_i \middle| X_{pa(i)}, \lambda_i, t_i\right) = \left(1 - e^{-K\lambda_i t_i}\right)\frac{1}{K} + \delta_{X_i X_{pa(i)}} e^{-K\lambda_i t_i}$$

where λ_i is the branch-specific rate of evolution, t_i is the length of the branch in units of time, and $\delta_{X_i X_{pa(i)}}$ is the Kronecker delta. Because only the product $\lambda_i t_i$ matters, we can set λ_i equal to $\frac{1}{K-1}$ (meaning that the average rate is normalized to 1) so that that $t_i = v_i$ now measures time in units of expected number of changes. We assume that each $v_i \sim \mathrm{Gamma}(\alpha, 1)$. Then,

$$p\left(X_i \middle| X_{pa(i)} \neq X_i, \alpha\right)$$

$$= \int_0^\infty p\left(X_i \middle| X_{pa(i)}, v\right) p(v|\alpha) dv$$

$$= \int_0^\infty \left(1 - e^{-\frac{K}{K-1}v}\right)\frac{1}{K}\frac{1}{\Gamma(\alpha)} v^{\alpha-1} e^{-v} dv$$

$$= \frac{1}{K} - \frac{1}{K}\int_0^\infty \frac{1}{\Gamma(\alpha)} v^{\alpha-1} e^{-v\left(\frac{K}{K-1}+1\right)} dv$$

$$= \frac{1}{K} - \frac{1}{K}\left(\frac{1}{\frac{K}{K-1}+1}\right)^\alpha \int_0^\infty \frac{\left(\frac{K}{K-1}+1\right)^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-v\left(\frac{K}{K-1}+1\right)} dv$$

$$= \frac{1}{K} - \frac{1}{K}\left(\frac{1}{\frac{K}{K-1}+1}\right)^\alpha$$

where the last equality follows because the integrand is the density function of a Gamma distribution with parameters $(\alpha, \frac{K}{K-1} + 1)$. A similar calculation shows that

$$p\left(X_i \middle| X_{pa(i)} = X_i, \alpha\right) = \frac{1}{K} + \frac{K-1}{K}\left(\frac{1}{\frac{K}{K-1}+1}\right)^{\alpha}.$$