Syst. Biol. 64(2):307–324, 2015

DOI:10.1093/sysbio/syu098
Advance Access publication December 18, 2014

Among-Character Rate Variation Distributions in Phylogenetic Analysis of Discrete Morphological Characters

LUKE B. HARRISON[1,*] AND HANS C. E. LARSSON[2]

[1]*Redpath Museum, McGill University, 859 Sherbrooke Street West Montreal, Quebec, Canada H3A 0C4 and* [2]*Redpath Museum, McGill University, 859 Sherbrooke ST W, Montreal, Quebec, Canada H3A 0C4*

[*]*Correspondence to be sent to: Redpath Museum, McGill University, 859 Sherbrooke Street West Montreal, Quebec, Canada H3A 0C4; E-mail: luke.harrison@mail.mcgill.ca*

Received 21 June 2014; reviews returned 4 September 2014; accepted 23 November 2014
Associate Editor: Peter Foster

Abstract.—Likelihood-based methods are commonplace in phylogenetic systematics. Although much effort has been directed toward likelihood-based models for molecular data, comparatively less work has addressed models for discrete morphological character (DMC) data. Among-character rate variation (ACRV) may confound phylogenetic analysis, but there have been few analyses of the magnitude and distribution of rate heterogeneity among DMCs. Using 76 data sets covering a range of plants, invertebrate, and vertebrate animals, we used a modified version of MrBayes to test equal, gamma-distributed and lognormally distributed models of ACRV, integrating across phylogenetic uncertainty using Bayesian model selection. We found that in approximately 80% of data sets, unequal-rates models outperformed equal-rates models, especially among larger data sets. Moreover, although most data sets were equivocal, more data sets favored the lognormal rate distribution relative to the gamma rate distribution, lending some support for more complex character correlations than in molecular data. Parsimony estimation of the underlying rate distributions in several data sets suggests that the lognormal distribution is preferred when there are many slowly evolving characters and fewer quickly evolving characters. The commonly adopted four rate category discrete approximation used for molecular data was found to be sufficient to approximate a gamma rate distribution with discrete characters. However, among the two data sets tested that favored a lognormal rate distribution, the continuous distribution was better approximated with at least eight discrete rate categories. Although the effect of rate model on the estimation of topology was difficult to assess across all data sets, it appeared relatively minor between the unequal-rates models for the one data set examined carefully. As in molecular analyses, we argue that researchers should test and adopt the most appropriate model of rate variation for the data set in question. As discrete characters are increasingly used in more sophisticated likelihood-based phylogenetic analyses, it is important that these studies be built on the most appropriate and carefully selected underlying models of evolution. [Among-character rate variation; Bayesian model selection; discrete morphological characters; morphological phylogenetics; phylogenetics; rate heterogeneity.]

Likelihood-based methods for phylogenetic analysis using discrete morphological characters (DMCs) are becoming more commonly used for estimating evolutionary relationships among extinct and extant taxa (e.g., Wiens et al. 2005; Müller and Reisz 2006; Lee and Worthy 2012). This class of methods has also been used for "total evidence" studies of phylogeny, which use both molecular sequence data and DMCs to simultaneously infer phylogeny (Nylander et al. 2004; Asher and Hofreiter 2006; Wiens et al. 2010; Müller et al. 2011; Dávalos et al. 2012; O'Leary et al. 2013, among others) and more recently, divergence times and phylogeny (Pyron 2011; Ronquist et al. 2012a; Wood et al. 2013). DMCs are typically modeled using continuous-time Markov chains in a framework nearly identical to that used for molecular sequences (Lewis 2001). However, DMCs are not nucleotides. Their unique properties, which include lack of constant characters, arbitrary state labeling, and character matrix order (when $n > 1$), enforce restricted substitution models and modifications to the likelihood calculation in order to maintain statistical consistency and accurate branch length estimation (Lewis 2001). The most commonly used model of character evolution is Lewis' (2001) Mk(V) model, implemented in popular maximum likelihood and Bayesian software packages for phylogenetic inference (e.g., MrBayes: Ronquist et al. 2012b and RAxML: Stamatakis 2006). However, there have been comparatively few model selection studies in likelihood-based phylogenetic analysis of DMCs (but see especially Clarke and Middleton 2008). The comparative lack of attention given to likelihood-based models of DMC evolution may be due to lingering concerns over the sufficiency of time-homogeneous Markov models to model morphological evolution (Goloboff and Pol 2006; Spencer and Wilberg 2013; Sterli et al. 2013). The increasing size of data sets of DMCs, now exceeding thousands of characters across greater than 50 taxa (e.g., Livezey and Zusi 2006, 2007; Naish et al. 2012; O'Leary et al. 2013), suggests that an exploration of model choice and of more complex models is warranted (Clarke and Middleton 2008; Larsson et al. 2012).

As with most phylogenetic studies using molecular sequences, most likelihood-based analyses using DMCs allow rates to vary across characters to account for among-character rate variation (ACRV). Lewis (2001) originally suggested the use of the discretized gamma distribution to model among-character rate heterogeneity and this has been nearly universally adopted. The discretized gamma distribution was a reasonable choice for a number of reasons: it is commonly employed for modeling among-site rate heterogeneity in molecular sequences, is computationally tractable, and is an inherently flexible model, accommodating data sets with little or extensive rate variation (Yang 2006; Pupko and Mayrose 2010; see

also Waddell et al. 1997; Mayrose et al. 2005; Huelsenbeck and Suchard 2007; Izquierdo-Carrasco et al. 2011 for alternatives). Several Bayesian phylogenetic analyses of DMCs have tested whether the discrete gamma-distributed rates model improves fit over an equal rates model and the results are nearly universally positive, indicating that ACRV is likely a feature of DMC matrices (e.g., Wiens et al. 2005; Müller and Reisz 2006; Ayache and Near 2009; Fröbisch and Shoch 2009; Prieto-Márquez 2010).

Wagner (2012) recently argued on a theoretical basis that the lognormal distribution should be a better model to accommodate ACRV for DMCs if evolutionary rates are the product of an underlying probabilistic process (like nucleotides), selection, and hierarchical interactions between characters. Employing the first systematic analysis of ACRV in DMCs, he further demonstrated empirical evidence for his hypothesis across invertebrate groups ($n = 115$ data matrices) using a customized maximum-likelihood model. However, it remains uncertain if his hypothesis is generalizable beyond invertebrates and whether the same pattern would be observed with widely used likelihood-based software packages (e.g., MrBayes; Ronquist et al. 2012b) and models (e.g., Mk; Lewis 2001) for phylogenetic analysis in a typical analytical setting. Wagner (2012) considered both rates of state change and state derivation but here only overall rates of change of characters in matrices are considered.

Rate heterogeneity in molecular sequences is known to mislead phylogenetic analysis if unaccounted for, especially in simple models (Yang 1996; Sullivan and Swofford 2001). It is therefore important to not only understand the degree of among-character rate heterogeneity present in DMC matrices but also which discretized distribution (lognormal or gamma) is an empirically better fit to a typical matrix of characters, and what effect model misspecification has on phylogenetic analysis. Furthermore, to the authors' knowledge, the behavior of the discrete approximation to either the continuous gamma or lognormal distribution under varying numbers of rate categories has not been examined in the context of DMCs. The choice of the number of discrete rate categories is a trade-off between computational tractability and an accurate representation of the underlying continuous probability distribution (Yang 1994).

The question of whether a lognormal or gamma distribution best models ACRV in data sets of DMCs bears on important questions about the dynamics of morphological evolution and development. Do morphological characters behave as relatively independent units of evolution and development whose among-character rates would be best represented with the gamma-distributed ACRV model? Or, is the evolution of DMCs governed by a mixture of more complex probabilistic processes rooted in biology and integrated through pleiotropic, epistatic, and selective pressures (i.e., best modeled with a lognormally distributed ACRV model)? The latter has long been suggested by anatomists and evolutionary biologists who argue that morphological evolution and development are characterized by structured, hierarchical processes (Riedl 1978). These questions also have broad implications for the accurate reconstruction of phylogenies and the validity of the findings of phylogenetic comparative analyses based on these phylogenies: does ACRV model misspecification strongly affect phylogeny and maximum-likelihood parameter estimation?

The analysis presented here has four objectives: (i) a generalized empirical test of Wagner's (2012) hypothesis using an existing software package under typical Bayesian analytic conditions; (ii) an empirical quantification of the degree of among-character rate heterogeneity in published data matrices of DMCs; (iii) a quantification of the effect of distribution choice and, in particular, misspecification on phylogenetic analysis and; finally, (iv) an analysis of the optimal number of discrete categories to approximate the continuous lognormal and gamma distributions to model ACRV in DMCs.

Materials and Methods

Data Sets

A total of approximately 200 data sets of DMCs used in published, mostly peer-reviewed articles and volumes were downloaded from TreeBASE (www.treebase.org; last accessed December 11, 2014). Matrices were downloaded using an R-script subject only to the requirement that they contained at least 50 characters. Because matrices of DMCs are frequently resampled or augmented by multiple authors and republished, the original data set contained redundancies. Therefore, each matrix and its associated primary literature were inspected and pruned to create as independent a set of matrices as possible, retaining the largest data sets when similar matrices were available. This resulted in a final set of 77 matrices of DMCs that were more or less independent, although some characters and taxa will inevitably have been sampled in different matrices (Supplementary References and Supplementary Table S1; available on Dryad at http://dx.doi.org/10.5061/dryad.067qg). The final subset consisted of representatives of plant ($n = 25$), invertebrate ($n = 25$) and vertebrate ($n = 27$) data matrices (Supplementary Table S1). This is particularly important as biologists working with different taxonomic groups may exhibit cultural biases in character choice and coding (e.g., discretization of morphoclines and treatment of multistate characters). The relative complexity of groups of organisms themselves may also affect estimates of rates (Schopf et al. 1975). The number of characters per data set ranged from 50 to 444 with an average of 138 ± 94 (mean $\pm$ SD) characters whereas the number of operational taxonomic units (OTUs) per matrix ranged from 9 to 246 with an average of 57 ± 39 OTUs. The original matrices were processed to remove OTUs that consisted only

of missing data; these were typically present because the original analysis was partitioned and included taxa with molecular sequence data that were not sampled for morphology. Source articles were also reviewed to determine which OTUs were designated as outgroups and these were maintained in all analyses described below (Supplementary References and Supplementary Table S1). Throughout the following analyses, all multistate characters were considered as unordered because the use of ordering is variable and non-random among systematists.

Bayesian Analysis

MrBayes v3.2.2-r512 (Ronquist et al. 2012b) was modified following Yang (1994) to implement lognormally distributed among-character rates using median (or quantile in the terminology of Pupko and Mayrose 2010) discretization with K equiprobable rate classes. MrBayes was selected to investigate these models because it implements the Lewis (2001) MkV model for morphological characters and is frequently used to estimate phylogeny of DMCs. The MkV model applies Lewis' (2001) correction for the acquisition bias inherent in morphological data: because constant characters are not informative in parsimony-based analyses, they have traditionally been excluded from data matrices of DMCs. The MkV model considers only the variable DMCs and estimates the probability of unsampled constant characters in order to condition the model likelihood on the fact that only variable characters are included (for further details see Lewis 2001). The application of ACRV models will lead to larger corrections for acquisition bias relative to an equal-rates model, as the lower rate categories will imply that unsampled constant characters are more probable, although the effect on the analysis is unclear and to the authors' knowledge remains unexplored. The arithmetic mean of the lognormal distribution was fixed to 1.0 such that the rate category multipliers varied only with the distribution shape parameter σ^2 (probability density function: Equation (1); Johnson et al. 1994; Yang 1994).

$$f\left(x;\sigma^2,\mu=-\left(\frac{\sigma^2}{2}\right)\right)=\frac{1}{x\sigma\sqrt{2\pi}}\mathrm{e}^{\frac{-ln(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Median or quantile discretization was implemented by using (Equation (2)) and the standard normal quantile function to determine the midpoint of the K rate category classes followed by rescaling the category means so that the mean of the discrete distribution is 1.0; r_i is the rate for category $i=\{1,\ldots,K\}$ and $\Phi^{-1}(p)$ is the standard normal quantile function. The lognormal shape parameter and model were otherwise treated identically to the gamma distribution shape parameter and model in terms of implementation.

$$r_i=\frac{e^{\mu+\sigma\phi^{-1}(p)}K}{\Sigma_i e^{\mu+\sigma\phi^{-1}(p)}},p=\frac{2i-1}{2K},\mu=-\left(\frac{\sigma^2}{2}\right) \quad (2)$$

Skinner (2010) used a similar discrete approximation of the lognormal distribution to integrate over rate heterogeneity across lineages but parameterized his model such that rates varied with the standard deviation of the lognormal distribution (designated as σ) rather than its shape parameter. It is also possible to parameterize the lognormal distribution based on the non-squared shape parameter σ. Here, the squared shape parameter was chosen but all three methods should produce a nearly identical discrete approximation, although the effective parameter space and posterior distribution of the distribution parameter will differ. MrBayes implements both equal site rate and discrete gamma-distributed site rate models. By default MrBayes uses mean rather than median discretization of the gamma distribution and this method was used for all analyses described here. Initial computations with both the modified and unmodified versions of MrBayes resulted in several computations failing because of excessive operator auto-tuning on the distribution shape parameters; to work around this issue, an upper limit on the auto-tuning parameter was added. The relationships between the lognormal and gamma distribution shape parameters (σ^2 and $\alpha=\beta$, respectively) and the category rates at $K=4$ and $K=12$ are graphically depicted in Figure 1. The modified source code of MrBayes is available in the Supplementary Materials.

For each data set, MrBayes was used to determine the marginal model likelihood of the three candidate models (equal rates [EQ], gamma-distributed rates [GA], and lognormally distributed rates [LN]), integrating over uncertainty in all other parameters, including phylogenetic topology by using stepping-stone sampling (Xie et al. 2011; Ronquist et al. 2012b). For each model, MrBayes was executed using four independent runs using Metropolis-coupled Markov Chain Monte Carlo (MCMC) sampling with three hot Markov chains and one cold chain for a total of 16 MC^2 chains per model per data set. The substitution model was set to the MkV model using the variable-characters only ascertainment correction (Lewis 2001; MrBayes command: lset coding = variable). For both unequal-rates models, four discrete rate categories were used. Because stepping-stone sampling also considers the prior in calculating the marginal likelihood (Xie et al. 2011), two alternative priors on the distribution shape parameters were considered: uniform and exponentially distributed. First, uniform priors on the interval [0.0001,200] were applied to the lognormal and gamma distribution shape parameters σ^2 and α, respectively. This interval covers the range of effective variability for these parameters (see Fig. 1) and the latter prior is the default implemented in MrBayes for the gamma rates model. In addition, the analysis was repeated using exponential priors with

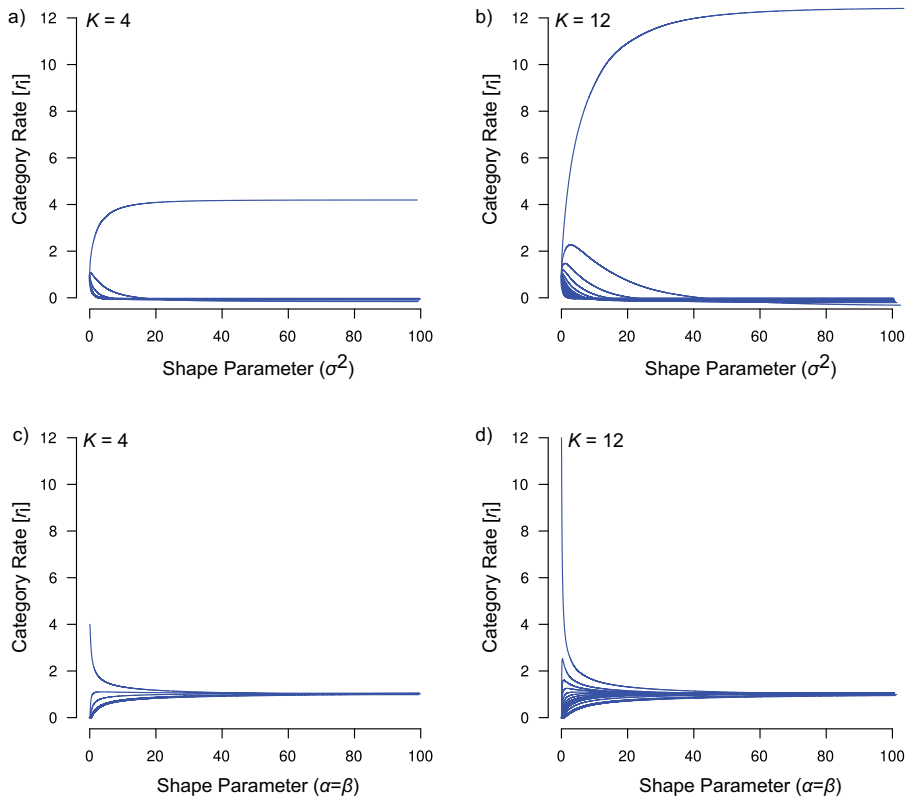

FIGURE 1. Plot of discrete category rates against distribution shape parameters for the discretized lognormal distribution at $K=4$ (a) and $K=12$ (b) and the gamma distribution at $K=4$ (c) and $K=12$ (d) discrete rate classes.

a mean of 1.0 on the respective distribution shape parameters.

Stepping-stone sampling was used to calculate marginal model likelihood for each rate model under each set of priors, sampling from the posterior to the prior for a total of at least 15 million generations up to approximately 30 million generations for the largest data set, sampling every 1000 generations (parameters for each data set are included in Supplementary Table S2). In all cases, initial sampling from the posterior was conducted for at least 5 million generations for each independent run to provide posterior estimates of the topology, branch lengths, and distribution parameters. Stepping-stone sampling toward the prior began after 5 million generations, discarding the first 25% of samples from each sampling step as a burnin. The default sampling scheme using 50 steps toward the prior, drawing δ values from a beta distribution with $\alpha=0.4$ was applied. All MrBayes analyses were conducted using the non-MPI version of MrBayes and the CLUMEQ/Colosse and CLUMEQ/Guillimin high-performance computing (HPC) facilities at McGill University and the University of Laval in Québec, Canada. Total computation time for the analyses described in this article exceeded 1 core-year.

From the initial 5 million generations sampling from the model's posterior, which were extracted from the MrBayes output with an R script using the APE R package and tools from the DendroPY and BEAST software packages (Paradis et al. 2004; Sukumaran and Holder 2010; Drummond et al. 2012; R Core Team 2012; R script available in the Supplementary Materials), a burnin fraction of 0.25 was discarded from each of the four independent runs. Convergence was assessed by using MrBayes to calculate the final average standard deviation of split frequencies (ASDSFs) and by ensuring estimated sample sizes (ESSs) for all parameters exceeded 200 using the R package CODA (Drummond et al. 2006; Plummer et al. 2006; Supplementary Table S2; available on Dryad at http://dx.doi.org/10.5061/dryad.067qg). Marginal posterior estimates for the majority-rule consensus topology, branch lengths and shape parameters were summarized and annotated using the MrBayes commands sump and sumt.

For each data set, the mean estimated marginal model likelihood from the four independent runs was retrieved from the MrBayes output for each model and this value was used as the estimated marginal model likelihood under each prior scheme. Bayes factors were computed as twice the log difference in marginal model likelihoods following Nylander et al. (2004) and were used to assess relative support for the models; the degree of rate heterogeneity was estimated by computing Bayes factors between the equal rates (EQ) and both unequal-rates (GA and LN) models under each distribution shape prior. Interpretation of Bayes factors is subjective, and guidelines from Kass and Raftery (1995; Nylander

et al. 2004) were initially followed: $0 < 2 \cdot \ln(B_{10}) < 2$ no evidence for model 1 over model 0, $2 < 2 \cdot \ln(B_{10}) < 6$ positive evidence for model 1 over 0, $6 < 2 \cdot \ln(B_{10}) < 10$ strong evidence for model 1 over 0, $2 \cdot \ln(B_{10}) > 10$ very strong evidence for model 1 over 0. However, initial calculations determined that because of variability in estimates of marginal model likelihoods, a more conservative interpretation was adopted that considered all differences of $2 \cdot \ln(B_{10}) < 6$ to be equivocal. Under each prior combination for the unequal-rates models, if rate heterogeneity was detected using a conservative threshold of $2 \cdot \ln(B_{10}) > 10$ for both the GA:EQ and LN:EQ comparisons, relative support between the lognormal and gamma models was examined by further calculating Bayes factors between these models. Bayes factors were also calculated between the model runs with the highest mean marginal model likelihood under each prior scheme. Pairwise Bayes factors between the LN and GA models were plotted using histograms in R (R Core Team 2012).

Parsimony-Based Estimation of Rate Distributions

Although parsimony- and likelihood-based methods are very different, the distribution of total changes per character inferred by parsimony offers an approximation to the distribution of character rates (e.g., Tourasse and Gouy 1997). Recovered rate distributions will be biased toward more equal rates because parsimony assumes all character changes are equiprobable and will underestimate rates in quickly evolving characters due to saturation (Felsenstein 2004). In order to estimate the distribution of among-character rates, maximum parsimony was used to determine the total number of changes for each character across the entire phylogeny for six focal data sets. Of these matrices, two were from data sets where the EQ model was an equivocal fit to the data relative to the GA and LN models under both prior schemes, two were from data sets that showed evidence ($2 \cdot \ln[B_{LG}] > 6$) for the LN model over the GA model and two that showed evidence for the GA model ($2 \cdot \ln[B_{GL}] > 6$) under both prior schemes. For each data set examined, PAUP* v4.0b10 (Swofford 2003) was used to map the characters under the ACCTRAN optimization onto each sampled phylogeny (>15,000 trees) in the marginal posterior distribution of trees derived from the equal rates (EQ) Bayesian analysis for each data set (details above). Rare constant characters were pruned from the analyses as the majority of DMC data matrices do not include constant characters. Distributions for each sampled phylogeny from the posterior were then combined for each data set and summarized using relative frequency histograms generated using R and the APE package (Paradis et al. 2004; R Core Team 2012; R script to execute PAUP* and process its output available in the Supplementary Materials).

Optimal Number of Discrete Rate Categories

To estimate the minimum number of discrete rate categories required to accurately approximate the continuous lognormal and gamma distributions, an analysis similar to Yang's (1994: Fig. 5) was performed. The four focal data sets included in the parsimony analysis described above showing strong evidence for the unequal-rates models were used. These data sets were then reanalyzed to determine the marginal model likelihood of both unequal-rates models using the same MrBayes analysis described above but modified to vary the number of discrete rate categories K across the following values {2, 3, 4, 5, 6, 8, 10, 14, 18} for both the GA and LN rate models under each prior scheme (total 32 model runs). Full stepping-stone sampling analyses, estimating marginal model likelihood, integrating over uncertainty in parameters including topology were applied for each K. Because these analyses were computationally intensive at high K, only two independent runs of four Markov chains each were calculated for each combination of rate-model, prior and K. As well, only 11 million generations per run were considered, discarding a sample from the posterior of 1 million generations but otherwise following the same procedure described above. Marginal model likelihood for each K, prior and model was then plotted against K to determine the minimum number of categories sufficient to represent the continuous distribution.

Focal Phylogenetic Analysis

For a focal topology, branch length, and clade support comparison, one data set (TreeBaseID S12929; Sidlauskas and Vari 2008; see below for rationale) was analyzed using MrBayes in a standard (i.e., not stepping-stone sampling) MCMC analysis. The analysis was conducted under four different ACRV models: equal rates (EQ), gamma-distributed rates (GA), and lognormally distributed rates (LN) at $K = 4$ and $K = 12$ rate categories. MC^3 sampling was conducted using four independent runs of four Markov chains running for 15 million generations, sampling every 1000 generations for each rate model using a uniform prior, as above, on the distribution shape parameters. All other parameters were identical to the model selection analysis described above. Convergence of independent runs was assessed by calculating Potential Scale Reduction Factors (PSRFs; Gelman and Rubin 1992) for all model parameters and split frequencies as well as calculating the final ASDSFs after combining runs and discarding a burnin fraction of 25% using the MrBayes sumt command. Sufficient sampling was assessed by ensuring that ESS of all parameters was greater than 200 using Tracer v1.5 (Drummond et al. 2006; Rambaut and Drummond 2007). Majority-rule consensus trees were calculated for each model using the MrBayes sumt command, after discarding the burnin fraction. Mean marginal posterior estimates of total tree length were calculated for each rate model using Tracer, combining samples from the posterior of each run after discarding the burnin fraction. Pairwise comparisons between clade posterior

probabilities were made between nodes shared by the compared majority-rule consensus trees using the R package APE (Paradis et al. 2004).

Results

Equal Rates Models and Unequal-Rates Models

Calculated marginal model likelihoods for each data set and each rate model are recorded in Supplementary Table S2 for both prior combinations. Variability in marginal model likelihoods between independent runs was low for most data sets, with average ranges of 1.2, 1.3, and 1.4 of log-likelihood units between replicates for the EQ, GA, and LN model runs (uniform prior model). In the most extreme example, data set S11535 (TreeBaseID; Sharkey et al. 2012: Hymenoptera), a range of 12 log-likelihood units was noted between independent runs for the GA model under the uniform prior. For this reason, this data set was excluded from further analysis. Furthermore, given the observed variability of runs, any Bayes factor difference less than 6 was deemed equivocal (see also Clarke and Middleton 2008). In 61/76 data sets, the Bayes factors comparing both unequal-rates models with the equal rates models exceeded the threshold of 10 for the presence of significant ACRV using uniform priors on the distribution shape parameters (Table 1). Using exponential priors on the distribution shape parameters, 64/76 data sets exceed the same thresholds. The exponential prior almost always led to higher likelihoods for both rates models relative to the uniform prior, and the results of the comparisons of the prior choice yielding the highest likelihood were essentially equivalent to the exponential comparisons and yielded identical tabulated results due to Bayes factor cutoffs (Table 1; Supplementary Table S2). Bayes factors comparing the lognormal rate model (under the prior choice of highest likelihood) with the equal rates were significantly correlated to the number of characters (Fig. 2a; Spearman's rank correlation test $r = 0.41$, $P < 0.001$) and much more strongly correlated to the number of OTUs present (Fig. 2b, $r = 0.86$, $P < 0.001$). Correlations calculated between GA:EQ comparisons and Bayes factors calculated using uniform priors or the exponential likelihood priors on distribution shape parameters were nearly identical (not shown).

Gamma and Lognormal Rate Distributions

Among the 61/76 data sets with significant rate heterogeneity under uniform shape distribution priors, 2/61 data sets showed evidence ($2 \cdot \ln(B_{GL}) > 6$) for the GA model over the LN model and a single data set showed very strong evidence ($2 \cdot \ln(B_{GL}) > 10$) (Table 1; Fig. 3a). Conversely, 18/61 data sets showed evidence for the LN model and 9/61 showed very strong evidence. The remaining 41 data sets (67%) were equivocal. Results for Bayes factors calculated from models using

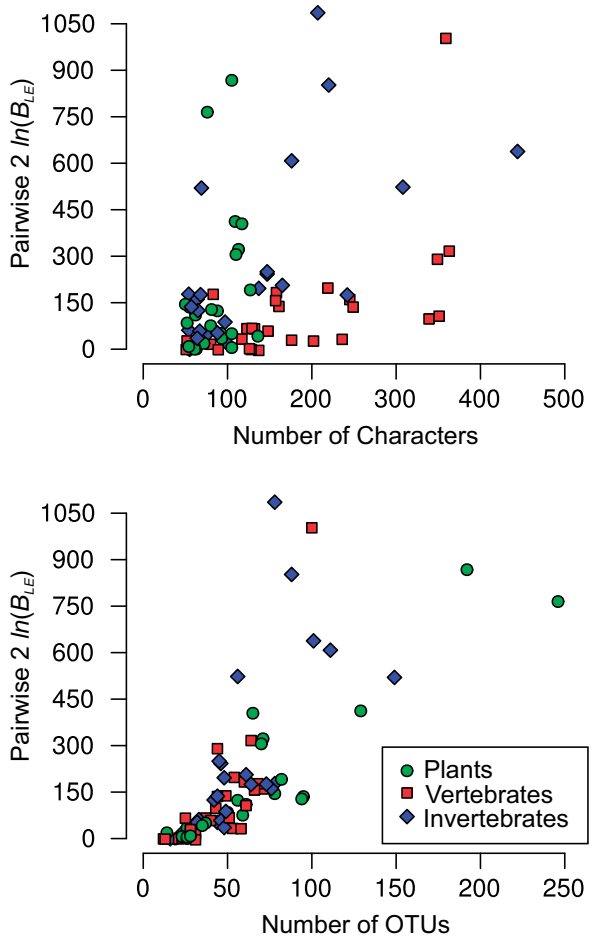

FIGURE 2. Plot of data set size in terms of a) number of DMCs and b) number of OTUs against Bayes factors calculated using twice the log difference between the mean marginal model likelihood for the equal and lognormally distributed among-character rate models. The marginal likelihood of the lognormal model was estimated using the distribution shape prior that yielded the highest model likelihood (see text). Positive Bayes factors represent support for the lognormal model over the equal rates model. Plant data sets are indicated by circles, vertebrate data sets by squares and invertebrate data sets by diamonds.

exponentially distributed priors on rate parameters were similar but more equivocal with respect to model choice and generally weakened support for the LN model relative to the GA model (Table 1; Fig. 3b). A further comparison of Bayes factors calculated between data sets using the highest marginal model likelihood recovered under either prior scheme was nearly identical to the comparison using exponential priors (Fig. 3b vs. Fig. 3c). This is because exponential priors nearly always led to the highest likelihood for each model, and where they did not, the tabulated Bayes factor results did not change because of the thresholds used and were therefore not reported in Table 1. Under exponential priors and the prior yielding the highest likelihood, 47 out of the 64 data sets with evidence for an unequal-rates model were equivocal with respect to support for the GA or LN model. However, among data sets where Bayes factors

TABLE 1. Results of the Bayesian model selection analysis of the three models of ACRV: equal, gamma-distributed, and lognormally distributed rates, using both uniform and exponential priors for the ACRV distribution shape parameter

Shape prior	Group	Number of data sets	GA/equal 2 $\ln(B_{GE})$ $2\ln(B_{GE})>10$	LN/equal 2 $\ln(B_{LE})$ $2\ln(B_{LE})>10$	Data sets with unequal rates	GA/LN 2 $\ln(B_{GL})$ $2\ln(B_{GL})>6$	GA/LN 2 $\ln(B_{GL})$ $2\ln(B_{GL})>10$	LN/GA 2 $\ln(B_{LG})$ $2\ln(B_{LG})>6$	LN/GA 2 $\ln(B_{LG})$ $2\ln(B_{LG})>10$
Uni[a]	Vertebrates	27	21 (78%)	21 (78%)	21 (78%)	0 (0%)	0 (0%)	7 (33%)	4 (19%)
Uni	Invertebrates	24	22 (92%)	22 (92%)	22 (92%)	2 (9%)	1 (5%)	8 (36%)	3 (14%)
Uni	Plants	25	18 (72%)	19 (76%)	18 (72%)	0 (0%)	0 (0%)	3 (17%)	2 (11%)
Uni	Total	76	61 (80%)	62 (82%)	61 (80%)	2 (3%)	1 (2%)	18 (30%)	9 (15%)
Exp[b]	Vertebrates	27	22 (81%)	22 (81%)	22 (81%)	1 (5%)	0 (0%)	5 (23%)	3 (14%)
Exp	Invertebrates	24	23 (96%)	23 (96%)	23 (96%)	2 (9%)	2 (9%)	5 (22%)	2 (9%)
Exp	Plants	25	19 (76%)	19 (76%)	19 (76%)	1 (5%)	0 (0%)	3 (16%)	1 (5%)
Exp	Total	76	64 (84%)	64 (84%)	64 (84%)	4 (6%)	2 (3%)	13 (20%)	6 (9%)

[a]Uniform prior for the distribution shape parameter.
[b]Exponential prior for the distribution shape parameter.

were not equivocal, the LN model was supported more often: 13/64 (20%) data sets $2\cdot\ln(B_{LG})>6$ compared with 4/64 (6%) for $2\cdot\ln(B_{GL})$. There was no obvious relationship between taxonomic group and rate model nor was standardized topological distance between consensus trees obviously related to model preference (using the prior that yielded the highest likelihood: Fig. 3d). Most consensus trees were identical between models (Fig. 3d), and inspection of individual topologies indicated that the observed differences in consensus topology were due to differences in nodes with support around the posterior probability cutoff (0.95); these nodes were included in one rate model's consensus tree but were excluded in the others. The results of the initial Bayesian analysis were used to designate six focal data sets for further analysis: two data sets where the EQ model had equivocal support relative to the unequal-rates models under all prior schemes (TreeBase ID No.: S10249; Holland et al. 2010: Phalacrocorcidae and S13029; Bourdon 2011: Odontopterygiformes), two with evidence for the LN model (S10265; Frick et al. 2010: Linyphiidae and S12929; $2\cdot\ln(B_{LG})=9.6$ and 11.4 under the uniform priors and 6.87 and 8.37 under the exponential) and two with evidence for the GA model (S2128; Liljeblad et al. 2008: Cynipidae and S12833; Lambkin and Bartlett 2011: Exoprosopini; $2\cdot\log_e(B_{LG})=$ 8.9 and 13.3 under the uniform priors and 10.4 and 13.9 under the exponential priors). Bayes factors for these data sets under the prior choice that yielded the highest likelihood were identical to the exponential prior scheme.

Parsimony Analysis

Parsimony estimation of character rate distributions of the six focal data sets was summarized using relative frequency histograms (Fig. 4). The two data sets where the EQ model had equivocal support relative to the unequal-rates models had relatively few changes per character (Fig. 4a,b). The data sets where the LN model was supported had relatively more characters with low rates and a rapidly decreasing number of characters with higher rates of change (Fig. 4c,d). Finally, the data sets where the GA model was supported had a relatively more uniform distribution of character rates (Fig. 4e,f).

Optimal Number of Discrete Rate Categories

Model likelihoods were always highest under the exponential prior for a given rates model (Table 2; Fig. 5). Variability within runs and between likelihoods under adjacent K values for the same data sets was relatively high but an overall trend is apparent: in the four focal data sets examined, regardless of whether the gamma or lognormal rate model was originally the best fit, four discrete rate categories appeared to be sufficient to represent the continuous gamma distribution (Table 2; dotted lines in Fig. 5). Beyond four rate categories, increasing K did not significantly

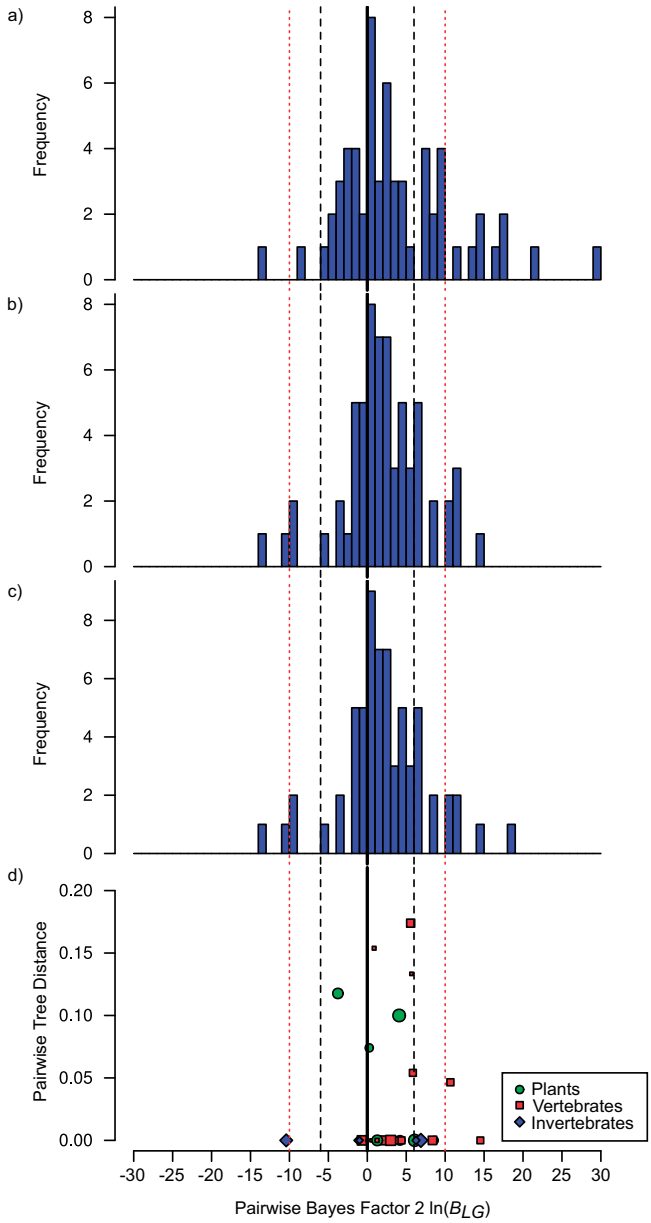

FIGURE 3. Bayesian comparison of the discrete gamma and lognormally distributed among-character rates models. Histograms summarize the Bayes factors 2 $\ln(B_{LG})$ comparing the gamma and lognormal models of ACRV for each data set under a) uniform shape distribution priors, b) exponential shape distribution priors and c) highest likelihood under either prior. Negative 2 $\ln(B_{LG})$ values indicate support for the gamma model over the lognormal model whereas positive values indicate support for the lognormal model. Vertical lines indicate 2 $\ln(B_{LG})=\{-6,6\}$, and 2 $\ln(B_{LG})=\{-10,10\}$, where there is respectively, strong evidence and very strong support for one model over the other. Standardized pairwise topology distance, calculated using the dist.topo function available in the APE R package (Paradis et al. 2004) and standardized by mean number of nodes, d) between consensus trees (posterior probabilities >0.95) estimated under the gamma and lognormal models plotted against 2 $\ln(B_{LG})$. Note that the number of pairwise comparisons is less in d) because comparisons were only made between analyses where the ASDSFs was less than 0.01. Data points are scaled to relative data set size. Plant data sets are indicated by circles, vertebrate data sets by squares and invertebrate data sets by diamonds.

increase the GA model's marginal likelihood. However, in both data sets where the lognormal model was a better fit at $K=4$ in the original analysis, the marginal model likelihood under the LN model increased with K and only begins to asymptote at $K>10$–12 (Fig. 5a,b). In data sets that preferred the gamma model, the behavior of the LN model with respect to K was similar to the GA (Fig. 5c,d) but decreased at $K>4$ for the S2128 data set (Fig. 5c) indicating that the continuous distribution would be a worse fit than the discretized distribution (see Discussion).

Effect on Phylogenetic Topology and Branch Lengths: Sidlauskas and Vari (2008)

Majority-rule consensus trees inferred by MrBayes under the equal (i), gamma, $K=4$ (ii), lognormal, $K=4$ (iii) and lognormal, $K=12$ (iv) ACRV models for data set S12929 for Sidlauskas and Vari (2008) are summarized in Figure 6. PSRFs were 1.0 and ESS was more than 200 for all model parameters and split frequencies. The final ASDSF was less than 0.01 for each model run, suggesting that each independent run for each rate model converged to the posterior distribution. This data set, which demonstrated greatest fit to the LN model (see Fig. 5b), was used in a phylogenetic analysis of relationships in the Anostomidae family of South American fishes and consisted of 158 DMCs coded for 60 OTUs (Sidlauskas and Vari 2008). The data matrix had 3% missing data and characters had an average of 2.35 states. The means of the marginal posterior distribution of total tree lengths were, respectively, 4.54, 6.65, 7.07 and 9.1 character changes per character for the EQ, GA ($K=4$), LN ($K=4$) and LN ($K=12$) ACRV models. There were several differences between the topology inferred under the EQ model and trees inferred under the GA and LN models: the EQ model resolved more nodes than the GA or LN models (Fig. 6a, arrows). There were one or two nodes that differed between the GA and LN models ($K=4$ or $K=12$; Fig. 6b–d, starred arrows). Clade posterior probabilities were highly similar between the GA and LN models (Fig. 7c) but quite different in the EQ model (Fig. 7a,b). The LN ($K=12$) had slightly higher overall clade posterior probabilities relative to the GA model (Fig. 7d) and LN ($K=4$) model (not shown).

DISCUSSION

Rate Heterogeneity in Data Sets of DMCs

Several previous Bayesian studies of DMCs have tested equal rates models against gamma-distributed rates models for individual data sets (e.g., Wiens et al. 2005; Müller and Reisz 2006; Ayache and Near 2009; Fröbisch and Shoch 2009; Prieto-Márquez 2010). The conclusions of those studies must be qualified by the use of the harmonic mean estimator (HME) for model likelihood which can be biased, especially toward more complex

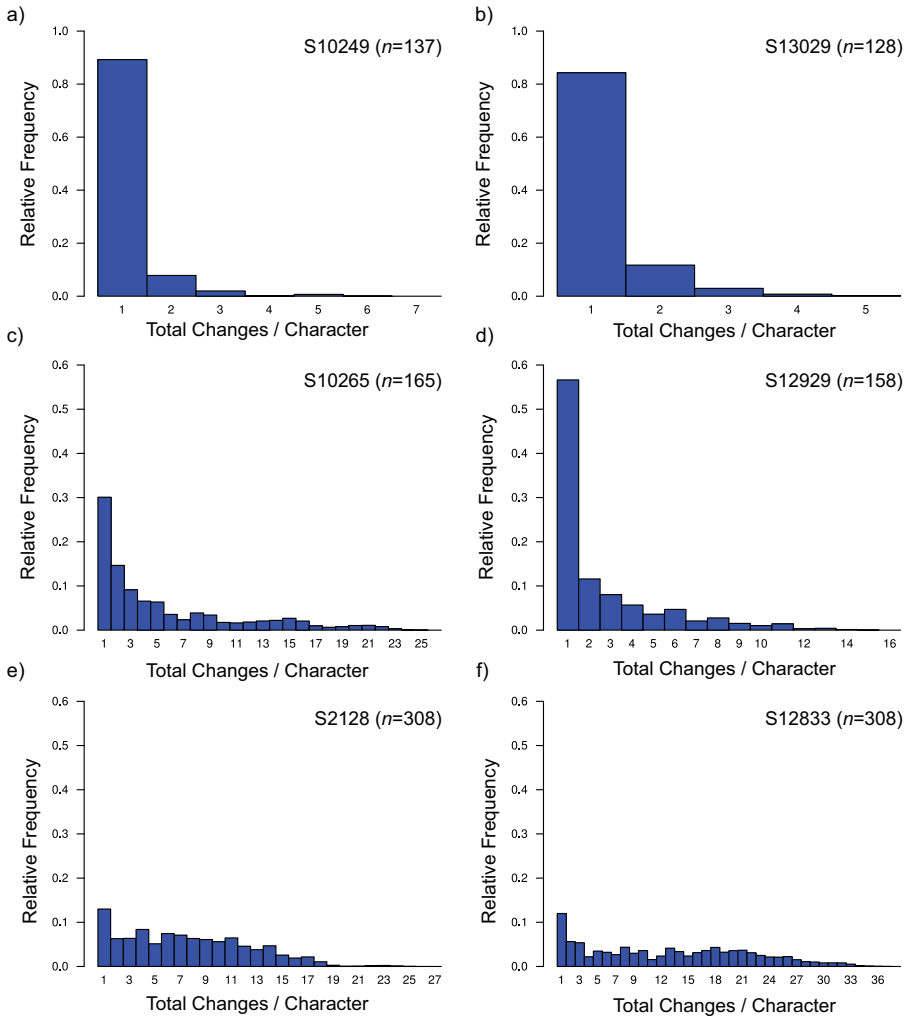

FIGURE 4. Parsimony reconstruction of the distribution of total character changes per character for six focal data sets using PAUP*. These relative frequency histograms summarize total character changes per character pooled across a sample of topologies (> 10000) from the marginal posterior of topologies from the equal-rates Bayesian analysis (see text); *n* is the number of DMCs present in each data set. The equal rates model was at least an equivocal fit to data sets a) S10249 and b) S10329 relative to the unequal-rates models; the lognormal model was preferred for data sets c) S10265 and d) S12929 and the gamma model was preferred for data sets e) S2128 and f) S12833.

models (Xie et al. 2011). Here, using the stepping-stone sampling method to estimate marginal model likelihood, which weights model likelihoods by their priors to penalize more complex models (Xie et al. 2011; Ronquist et al. 2012b), most data sets still exhibited significant rate heterogeneity by Bayes factors indicating very strong evidence for unequal-rates models over equal-rates models (Table 1 and Fig. 2). Bayes factors supporting unequal-rates models were more strongly correlated to the number of OTUs compared with the number of characters (Fig. 2), a pattern not observed in molecular sequences (e.g., Mayrose et al. 2005: Fig. 3). This suggests that increasing data set size, especially numbers of OTUs, should increase the ability to detect rate heterogeneity. This analysis corroborates those previous studies using the HME and suggests that ACRV is significant in DMC matrices.

Gamma and Lognormal Rate Distributions

Felsenstein (2001) argued that for molecular sequences, either the gamma or lognormal distribution should be effective to model among-site rates simply because both are distributions on the interval $[0, \infty]$. He further argued that it would be difficult to differentiate these distributions without large quantities of data. Felsenstein's arguments were directed toward molecular sequences and did not discuss the added effect of discretization. Here, in matrices of DMCs, it was possible to discriminate between the discretized gamma and lognormal ACRV models among some data sets of modest size. This analysis strongly concurs with Felsenstein (2001) in that most data sets show equivocal support for either the GA or LN models, at least when using a 2 ln(BF) < 6 criterion under two different

TABLE 2. Results of the Bayesian analysis of the discrete approximation of the gamma and lognormal distributions of ACRV with variable numbers of discrete rate categories

						Mean model marginal model likelihood								
Data set	Number of OTUs	Number Chars.	EQ model ln(L)	Rate model	Shape prior	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=8$	$K=10$	$K=14$	$K=18$
S10265	111	176	−5582	LN	Uni[a]	−5299	−5283	−5281	−5280	−5276	−5271	−5268	−5267	−5267
				GA	Uni	−5299	−5286	−5286	−5286	−5287	−5284	−5283	−5282	−5282
S12929	60	158	−2660	LN	Uni	−2576	−2573	−2572	−2570	−2568	−2566	−2565	−2562	−2561
				GA	Uni	−2576	−2578	−2577	−2577	−2578	−2578	−2576	−2579	−2577
S12833	78	207	−11714	LN	Uni	−11227	−11181	−11176	−11177	−11175	−11177	−11176	−11179	−11178
				GA	Uni	−11226	−11174	−11169	−11170	−11169	−11169	−11167	−11166	−11167
S2128	56	308	−9478	LN	Uni	−9231	−9220	−9220	−9222	−9224	−9223	−9225	−9224	−9226
				GA	Uni	−9229	−9216	−9216	−9216	−9215	−9216	−9216	−9215	−9216
						Mean model marginal model likelihood								
Data set	Number of OTUs	Number Chars.	EQ model ln(L)	Rate model	Shape prior	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=8$	$K=10$	$K=14$	$K=18$
S10265	111	176	−5582	LN	Exp[b]	−5293	−5282	−5278	−5276	−5273	−5270	−5267	−5266	−5264
				GA	Exp	−5295	−5280	−5281	−5282	−5284	−5281	−5279	−5279	−5277
S12929	60	158	−2660	LN	Exp	−2574	−2570	−2569	−2567	−2565	−2563	−2562	−2561	−2560
				GA	Exp	−2572	−2574	−2573	−2573	−2573	−2573	−2572	−2572	−2573
S12833	78	207	−11714	LN	Exp	−11222	−11177	−11173	−11175	−11174	−11175	−11173	−11176	−11177
				GA	Exp	−11224	−11168	−11165	−11164	−11163	−11165	−11165	−11164	−11163
S2128	56	308	−9478	LN	Exp	−9226	−9217	−9216	−9218	−9218	−9219	−9222	−9220	−9221
				GA	Exp	−9227	−9212	−9212	−9213	−9212	−9213	−9214	−9212	−9212

[a]Uniform prior for the distribution shape parameter.
[b]Exponential prior for the distribution shape parameter.

prior specifications. However, among the subset of the data matrices that are not equivocal, evidence for the lognormally distributed ACRV model was detected more often than evidence for the gamma rates model; this offers some, albeit qualified, support to Wagner (2012)'s hypothesis. However, this conclusion must be strongly qualified as clearly data set dependent: in some data sets there was strong evidence for gamma-distributed rates model, especially under an exponential prior on the rate distribution shape parameter (Table 1 and Fig. 3a,c). Wagner (2012) also noted support for gamma-distributed models in some data sets, especially considering character change rates and particularly among trilobites where both models were equally supported. Here, preference for either rate model was not obviously related to major taxonomic division or data set size (see Fig. 3d)

However, it is possible that other data set-specific qualities may influence ACRV model preference. Atomistic character coding (enumerating phenotypes into a set of finest discernible variations) might lead to data sets better modeled with lognormal distributions. Some authors deliberately follow this mode of characterization with the justification that it may more objectively divide phenotypes into phylogenetic characters (e.g., Livezey and Zusi 2006). This mode of characterization is also more likely to record autapomorphies, which are considered phylogenetically uninformative in parsimony-based analyses, but are informative in likelihood-based methods. A caveat of atomistic characterization, though, is that these characters are more likely to interact among themselves, through developmental and/or functional associations, and the outcome of these probabilistic processes may be best modeled with a lognormal distribution (Wagner 2012).

Although many DMC matrices still retain some degree of atomistic characterization, most authors also strive to incorporate some degree of biologically relevant dependencies. Phenotypic qualities are often not completely independent from each other and complex interactions arise from developmental and/or functional relationships. High degrees of interactions define relatively discrete evolutionary modules (Wagner and Altenberg 1996) and it is these modular units of phenotypes that more complex characters attempt to describe. The construction of module-informed characters should maximize the independence of characters (Wagner 1995; Houle 2001; Kim and Kim 2001). Complex characters that are coded explicitly to be maximally independent may be better modeled using gamma-distributed rate models. Indeed, in reality, DMC matrices are composed of a continuum of independent and non-independent characters and may warrant the use of a model-averaging approach employing both gamma and lognormal distributions (see below). Additionally, DMC matrices could be explored using these methods to discover degrees of linkages and potential novel interaction partners among characters. Comprehensive testing of these questions is beyond the scope of the analysis presented here.

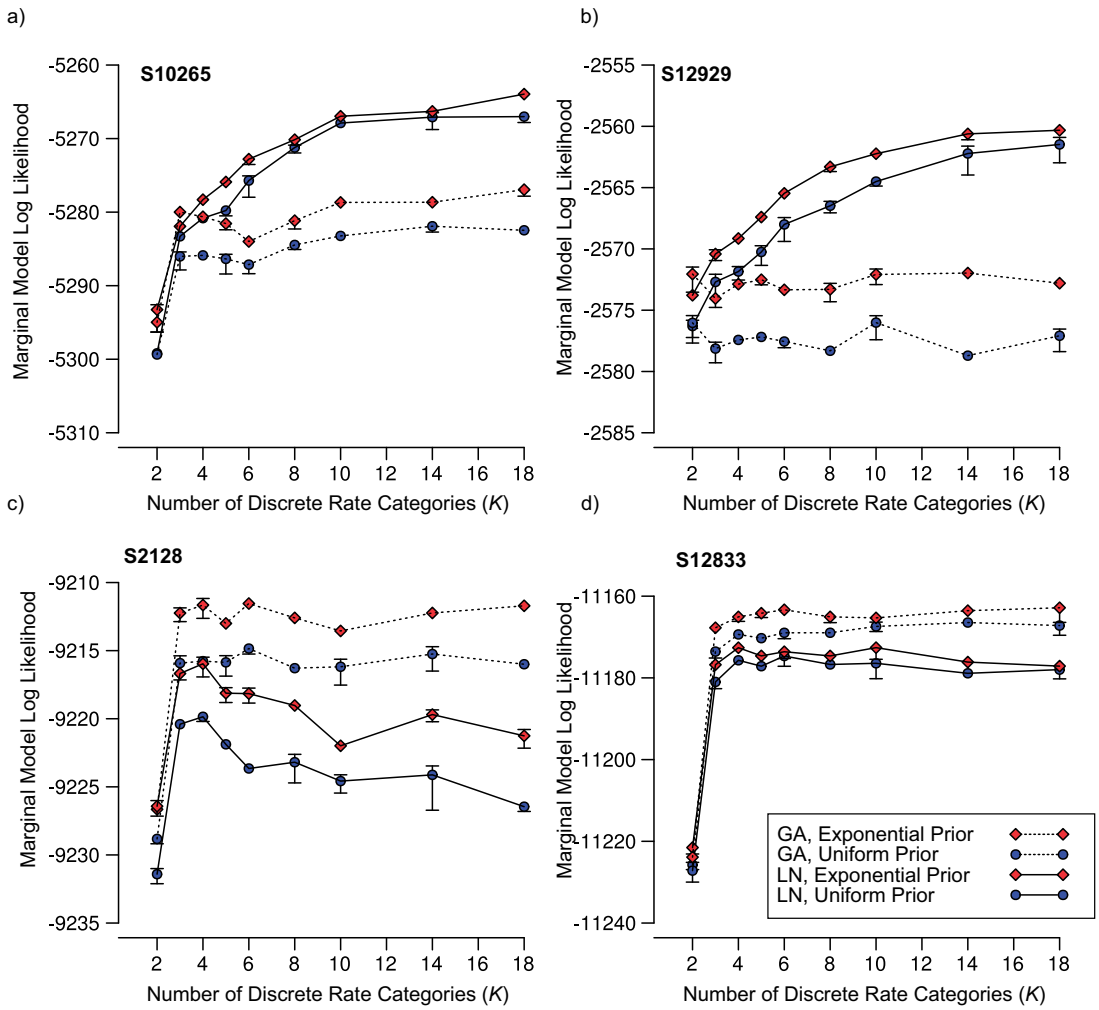

FIGURE 5. Relationship between mean marginal model log likelihood for the discrete gamma (GA) and lognormal (LN) models of ACRV and the number of discrete rate classes (K) under uniform and exponential shape distribution prior choices. LN model likelihoods are solid lines whereas the GA model likelihoods are dotted lines. Marginal model likelihoods calculated under the exponential model are diamonds, whereas those calculated under the uniform prior model are circles. The lognormal rates model was a better fit to data sets S10265 and S12929 (a,b) whereas the gamma rate model was a better fit to S2128 and S12833 (c,d). Although the gamma distribution is accurately represented using $K = 4$ rate classes, the continuous lognormal distribution is better represented with greater number of discrete rate categories ($K > 8$) but only data sets where it has the best fit (a,b). Error bars represent the range of marginal model likelihoods from the two independent stepping-stone sampling runs (see text).

Stepping-stone sampling considers the prior in the estimation of the marginal model likelihood (Xie et al. 2011). In the case of the identical uniform priors used here, because of the nature of the underlying distributions, more prior weight was given to equal rates among sites using the gamma-model while the lognormal was the converse (see Fig. 1). Under the exponentially distributed prior, the opposite is true because the direction of variation (i.e., rates becoming more equal) is reversed among these two rate distributions (Fig. 1). Here, both priors were tested to determine the effect of the prior on model choice, which was appreciable (see Fig. 3a vs. Fig. 3b). Even considering the prior choice that produced the highest marginal model likelihood for each model, there is still a weak preference for the lognormal model among the data sets that are not equivocal (Fig. 3c). However, there were also some data sets that conclusively favored the gamma rates model under the same comparison. The influence of the prior on the marginal model likelihood was relatively strong among these data sets and may reflect a relatively large weight of the prior compared with the information content of the data. Future studies using larger, potentially more informative data sets of DMCs may also offer sufficient information to better discriminate these two models. Finally, an alternative "generalized" stepping-stone approach to calculate marginal model likelihoods that does not require sampling close to the actual prior might offer more efficient and precise model comparison (e.g., Fan et al. 2011) but this does not yet accommodate variable topologies (Baele et al. 2012).

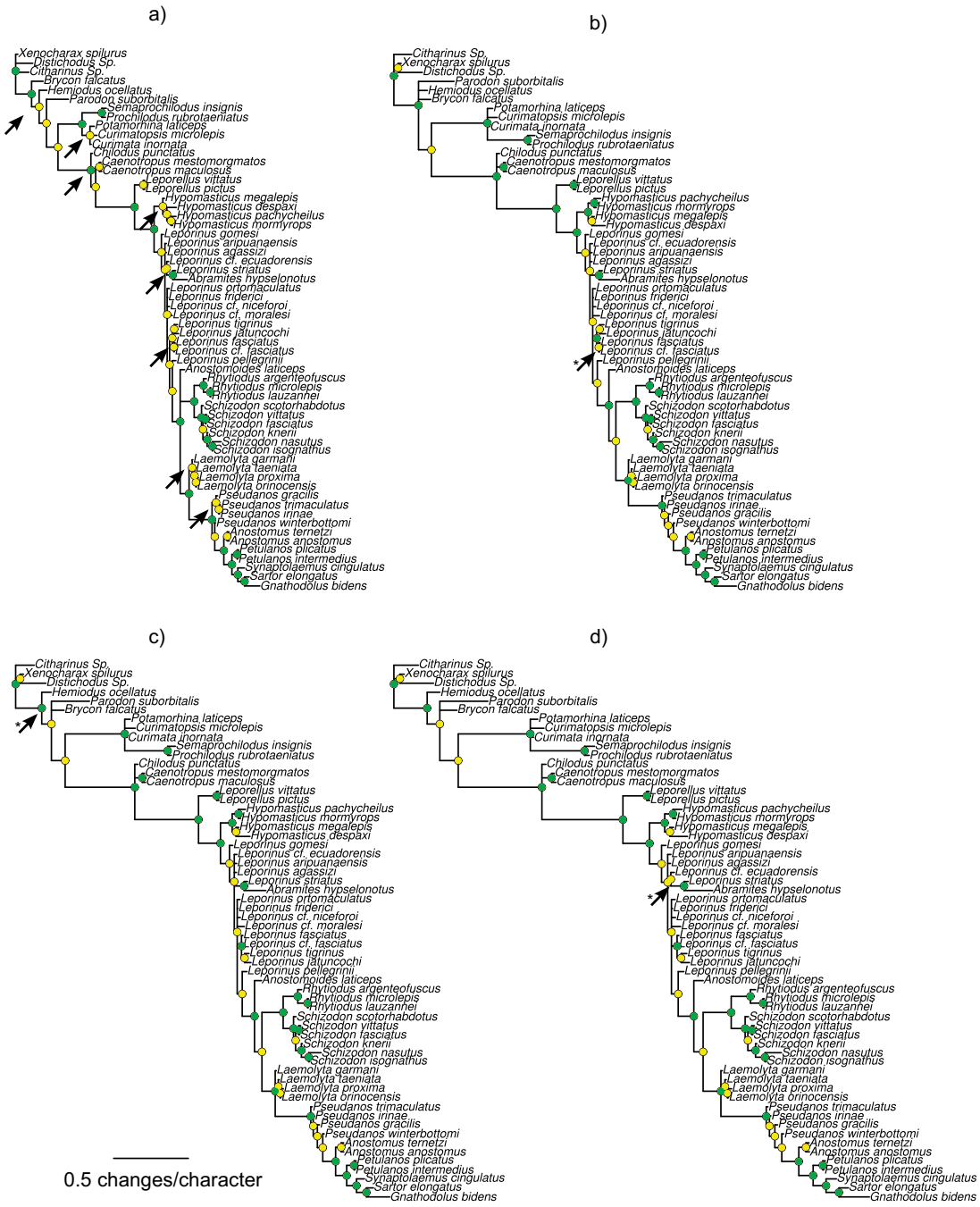

FIGURE 6. Majority-rule consensus summarizing the posterior distribution of phylogenies inferred under the a) equal-rates model, the b) gamma-rates model ($K=4$), and the lognormal rates model ($K=4$, c; $K=12$, d). Clade posterior probabilities are indicated by dots: lighter gray (yellow) >0.5 and darker gray (green) >0.95. Branch lengths were annotated using the mean of the marginal posterior distribution of the respective branch lengths using MrBayes. Means of the marginal posterior distribution of total tree lengths were, respectively, 4.54, 6.65, 7.07 and 9.1 character changes per character for the EQ, GA ($K=4$), LN ($K=4$) and LN ($K=12$) models. Arrows indicate topological differences (see text for details). Color version of this figure is available at *Systematic Biology* online.

Parsimony-based reconstruction of six focal character rate distributions (Fig. 4) corroborated the Bayesian analysis and hinted why the lognormal distribution was a better fit to some data sets. In both data sets where the equal-rates model was an equally good choice relative to unequal-rates models (S10249 and S13029; Fig. 4a,b), parsimony reconstruction of character changes confirmed a largely equal distribution of rates. The two data sets where the evidence for the LN was recovered (Fig. 4c,d) had proportionally more characters with slow

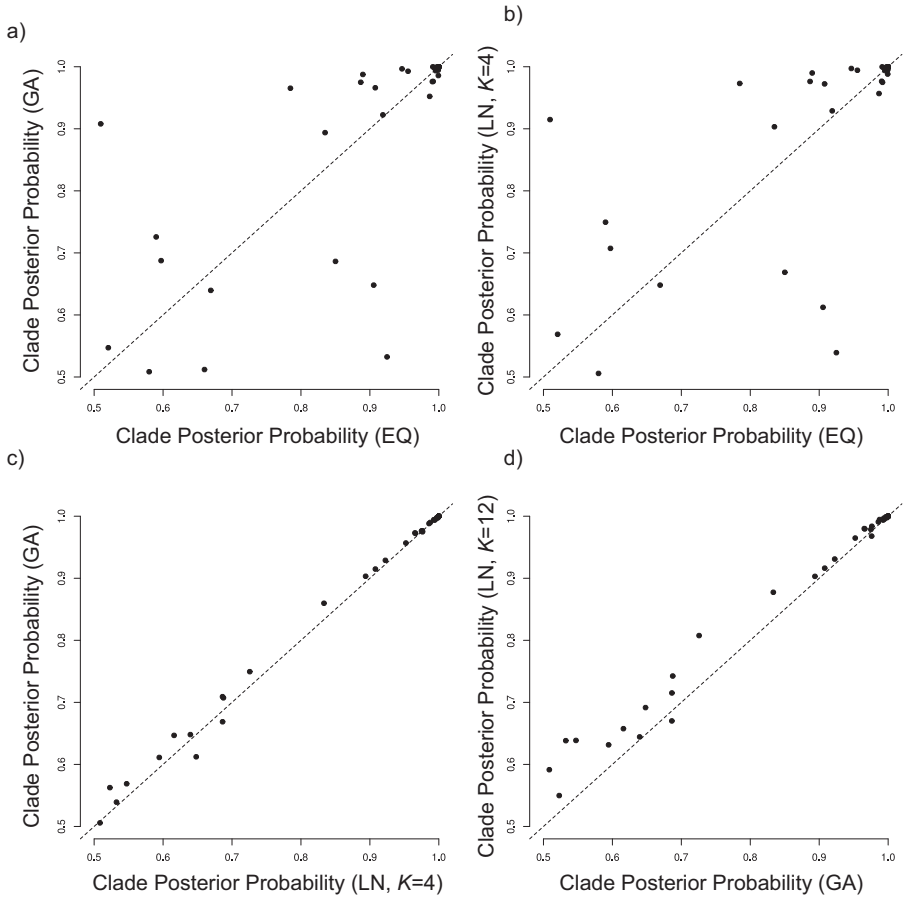

FIGURE 7. Comparison of clade posterior probabilities for the S12929 data set under different among-character rate heterogeneity models. a) Equal-rates model versus gamma-distributed rates ($K=4$) model, b) equal-rates versus lognormal-distributed ($K=4$) rates model, c) gamma-distributed rates ($K=4$) versus lognormal-distributed rates model ($K=4$) and d) gamma-distributed rates ($K=4$) versus lognormal distributed rate model ($K=12$).

rates and a smaller number of characters with much higher rates. In contrast, the data sets where evidence for the GA model was recovered had much more even distribution of character change rates, that is, more unequal rates (Fig. 4e,f). Category rates when $K=4$ for the LN and GA models under varying distribution shape parameters (Fig. 1) hint at why the lognormal distribution was a better fit to data sets containing many characters with low, but not equally low rates. Compared with the gamma distribution (Fig. 1c,d), the lognormal distribution (Fig. 1a,b) has several slow categories and one fast rate category while the gamma distribution's rate category means are generally more evenly spaced on the $\{1,\ldots,K\}$ interval except for the small interval near 0. The parsimony analysis supports the hypothesis that the choice of ACRV models in select data sets is driven by underlying patterns of character evolution. However, the complex interaction between subjective factors (character construction and coding) and objective factors (phylogenetic signal, number of OTUs, etc.) make thorough understanding of the observed patterns complex.

Optimal Number of Discrete Rate Categories

Computation time increases linearly with the number of discrete rate categories used to approximate the continuous lognormal or gamma distributions as the overall character transition probabilities are integrated over all rate categories (Yang 1994). Because of this significant computational cost, it is important to minimize the number of rate categories while maintaining an accurate approximation of the continuous distribution. However, it is theoretically possible that an increased number of rate categories will lead to more rapid convergence of the MCMC algorithm (Wagner P., personal communication). For molecular sequences, Yang (1994) determined that four rate categories were sufficient to represent the gamma distribution when using mean discretization of rate categories. Mayrose et al. (2005) argued that larger data sets may require a greater number of categories or a gamma-mixture model. The choice of four categories is common in evolutionary studies and is usually the default setting in software packages (e.g., RAxML

v7.2.8; Stamatakis 2006). Because DMCs are neither nucleotides nor amino acids, the choice of four may not be appropriate. Here, given the high computational cost at high numbers of rate categories (K), only four focal data sets were assessed: two matrices where the LN model was supported (S12929 and S10265) and two matrices where GA model was supported (S2128 and S12833).

The analysis of marginal model likelihood under each rate model for the four focal data sets revealed divergent behavior between the GA and LN models. For matrices where the LN model was supported, marginal model likelihood under the LN model increased with K and only began to asymptote at $K > 10$–12 (Fig. 5a,b). Interestingly this pattern is similar to that observed by Skinner (2010)'s Fig. 1a when he discretized the lognormal distribution to model among-lineage rate heterogeneity. This is in contrast to the GA model, where marginal model likelihood reached an asymptote at $K \approx 4$ in data sets for which it was and was not supported (Fig. 5). It is possible that discretization of the lognormal distribution using the mean rather than the median of the respective rate categories might improve the model's performance with respect to K, as this pattern is observed in molecular data (e.g., see Yang 1994). On the other hand, the quantile median may be preferable because the median of the highest rate class will be closer to the global median than the corresponding mean, which may take on much higher values (Wagner P., personal communication). It is not immediately clear why the marginal model likelihood for the LN model in the S2128 data set actually decreased at $K > 4$; a possible explanation is that the continuous lognormal distribution would have been a poorer fit to the data and that discretization, which at lower K is less representative of the continuous distribution, buffered the effect of the model misspecification. Although it is possible that this analysis underestimated variability in marginal likelihoods by using only two runs per K per data set, these general trends appear robust. Further interpretation of smaller-scale trends would require a greater number of independent runs for each K. Because the main analysis presented above used $K = 4$ rate categories, it may underestimate the support for the LN over the GA model in some data sets or overestimate it in others. Investigators are therefore encouraged to use as many rate categories as feasible to test LN and GA models.

Increasing the number of discrete rate categories may impose non-trivial computational cost. However, phylogenetic analyses under the MkV model (Lewis 2001) have a fixed instantaneous rate matrix and few variable parameters: typically only the among-character distribution shape parameter, branch lengths and topology. Comparatively few free parameters and the small size of most character matrices compared with molecular analyses mean that the increase in computational resources required to at least double the number of rate categories to eight might not be onerous except for large data sets. This is particularly true given that multicore personal computers are now commonplace and that HPC resources coupled with parallelized phylogenetic packages are increasingly available for phylogenetic analysis of DMCs (e.g., BEAGLE; Ayres et al. 2012).

The larger number of rate categories required to sufficiently approximate the lognormal distribution (Fig. 5a,b) and the computational cost associated argue that the discretization method may not be the optimal approach to approximate the underlying continuous lognormal distribution for DMCs. Alternatives to approximate the distribution without the need for discretization (*sensu* Yang 1994) may significantly increase model fit without excessive computational cost, particularly if compared with discrete models using high values for K. The Laplace quadrature (Felsenstein 2001; Mayrose et al. 2005) has been used with success to approximate the continuous gamma distribution and similar alternatives may exist for the lognormal distribution.

Effects on Estimations of Phylogenetic Topology

The focal analysis of the S12929 data set recovered the greatest topological differences between the EQ model and both unequal-rates models (Fig. 6: arrows). Clade posterior probabilities were very different between the EQ model and both unequal-rates models (Fig. 7a,b), a fact observed in other studies comparing equal rates to gamma-distributed rates but possibly also related to coding of autapomorphies (Müller and Reisz 2006). Interestingly, the EQ model resolved more nodes in the majority-rule tree than did the GA or LN models, which is likely symptomatic of the overestimation of Bayesian clade posterior probabilities when the model is overly simplified (e.g., Alfaro and Holder 2006). Topologies inferred under the unequal-rates models were very similar: both LN models resolved an additional node at the base of the tree (Fig. 6c, starred arrow) whereas the GA resolved an additional node in the *Lepornius* genus (Fig. 6b, starred arrow). Finally, the LN ($K = 12$) resolved a different node among *Lepornius* relative to the LN ($K = 4$) or GA models (Fig. 6d, starred arrow). Functionally, the topologies recovered under the GA and LN models were nearly identical and clade posterior probabilities were also very similar for clades with PP > 0.5, although slightly higher for the LN ($K = 12$) model (Fig. 7d). Overall tree length was higher for the LN ($K = 12$) model and indeed, tree lengths appeared to be high for all rate models. This may be related to the recent observation that MrBayes' default independent exponential priors on branch lengths can lead to overestimates of total tree length in some data sets, especially if a gamma-distributed rate model is used (Brown et al. 2010; Marshall 2010; Rannala et al. 2012; Zhang et al. 2012). It would have been interesting to examine the relationship between discretization, rate model choice, and overall tree length using the version of MrBayes created by Zhang et al. (2012) that resolves

this issue by using compound Dirichlet priors on branch lengths (now incorporated into the main distribution of MrBayes). However, when this was attempted, the likelihoods estimated by this version of the MrBayes program using stepping-stone sampling were highly variable (occasionally >200 log likelihood units between independent runs of the same model) for both equal-rates and gamma-rates models. It was not possible to determine the source of this variability. Implementation issues aside, the relationship between the use of DMCs and interactions between ACRV models, overall branch lengths and the Lewis (2001) acquisition correction clearly requires further consideration. In summary, although topological differences were observed between unequal-rate models, these appear to be generally minor relative to not accounting for rate heterogeneity among characters.

Several recent studies have urged caution in the use of likelihood-based phylogenetics for DMCs (e.g., Spencer and Wilberg 2013; Xu and Pol 2014). For example, Spencer and Wilberg (2013: 665) argued, based on Wagner's (2012) analysis, that because lognormal ACRV models may be preferred over gamma ACRV models, phylogenetic analyses conducting using gamma ACRV models may lead to "potentially misrepresenting the phylogenetic signal". They further suggested the use of maximum parsimony, which treats all character changes equally across characters. The results of this analysis and other recent model-based analyses (e.g., Wagner 2012) suggest that the greater risk of misrepresenting phylogenetic signal comes from the choice to not adopt a model explicitly accommodating rate variation among characters, rather than the specific ACRV model used. However, applying the best supported model is ideal and could be important in some data sets.

Alternative Approaches to Model ACRV

The results presented here must be accompanied by a strong caveat: as with any model selection study, the results are conditioned on the choice of candidate models (Posada and Buckley 2004). Here, only three models were evaluated and this analysis does not preclude alternative models that may demonstrate significantly better fit to the data. Approaches to ACRV can be divided by analogy into "random-effects" models and "fixed-effects"-type statistical models (Yang 2006; Yang and Rannala 2012). The approaches here are random-effect models because they integrate across rate heterogeneity without expressly assigning characters to *a priori* categories. Alternative random-effects approaches include the Dirichlet-process-prior-based models, where the number of rate categories and assignments are estimated from the data (e.g., Huelsenbeck and Suchard 2007). In this context, fixed-effect type models include *a priori* partitioning of sets of characters and estimating rates within each partition. Partitioned Bayesian analysis of molecular sequences has been demonstrated to significantly increase model fit (e.g., Brandley et al. 2005; Brown and Lemmon 2007) and is an alternative and complimentary approach to dealing with heterogeneity across sites/characters, especially when biologically informed (e.g., partitioning by codon position: Shapiro et al. 2006). These approaches are now commonplace, especially for phylogenomic analyses, and extensive tools are now available to facilitate partition selection (e.g., PartitionFinder; Lanfear et al. 2012).

Clarke and Middleton (2008) found that partitioning DMCs anatomically and allowing branch lengths to vary across partitions led to higher marginal model likelihoods over an unpartitioned gamma-distributed rate model in a fossil data set. However, the results from that study are qualified by Clarke and Middleton's use of the HME for marginal model likelihood, which may overestimate likelihoods of complex models (e.g., Xie et al. 2011; Baele et al. 2012; this is not meant as a criticism as stepping-stone sampling has only recently become available in MrBayes and Clarke and Middleton expressly qualified their study on the appropriateness of the HME). Partitioned models may be particularly appropriate for DMCs because character integration (i.e., modularity) and selection affecting multiple characters may lead to correlated evolutionary changes. In the absence of easily available implementations of more complicated strategies for partitioning DMCs (e.g., Lanfear et al. 2012), investigators are encouraged to perform model testing using biologically informative *a priori* partitions and both among-character rate distributions using as large a number of discrete rate categories as possible.

A further alternative for analyses where the ACRV model is a nuisance parameter rather than the focus of the analysis is to treat the model itself as a random variable. Uncertainty in the ACRV model could then be integrated across using, for example, reversible-jump MCMC or another model-averaging approach; this technique has been used to treat the relaxed molecular clock model as a nuisance variable in divergence time analysis (Li and Drummond 2012). This approach could also be used to assess relative support for ACRV models by examining their posterior probabilities, which, in conjunction with gamma, lognormal, or other models, could be used to explicitly test hypotheses about the evolution of DMCs. Computationally, model-averaging approaches may only impose a slightly higher overall computational burden to parameterize because they remove the need to separately test multiple ACRV models.

Conclusions

Rate heterogeneity was widespread in this diverse data set of DMC matrices as indicated by strong preference for unequal-rates models. Where significant rate heterogeneity was present, Bayesian model selection suggested that most data sets were equivocal in support for the gamma-distributed and lognormally

distributed rates models. However, where model testing was not equivocal, the lognormal model was more frequently supported relative to a gamma-distributed rates model. However, the reverse was also true for several data sets. Overall, the results suggest weak evidence for Wagner's (2012) hypothesis with strong qualifications. Prior choice was found to affect marginal model likelihood estimation but the observed patterns hold after comparison of model likelihoods under the prior choice yielding the highest marginal model likelihood for each model were compared. Parsimony analysis of estimated character distributions in four focal data sets suggests that the underlying character rate distribution may drive observed patterns in some data sets. Although rate distribution choice sometimes affected topological reconstruction, this pattern was difficult to characterize and did not appear to be related to the degree of model preference as measured by pairwise Bayes factor. Lognormally distributed rate models may require a larger number ($K > 8$) of discrete rate categories to approximate the continuous distribution, in contrast to gamma-distributed rate models where $K = 4$ rate categories appeared sufficient. Researchers should therefore use as many discrete rate categories as computationally feasible to ensure an accurate representation of the underlying continuous distribution when using the lognormal distribution. Although the data were not conclusive, the rate distribution choice was observed to have a minor effect on the posterior distribution of topologies and branch lengths. Researchers are urged to test which among-character rate model is best supported by their data and explore alternative character partitioning strategies using Bayesian marginal model likelihood estimation and Bayes factors.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.067qg.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (CGS-D Doctoral Fellowship and André Hamer Postgraduate Prize to L.B.H. and Discovery Grant No. 261796-2011 to H.C.E.L.); and by a Canada Research Chair to H.C.E.L. High performance computation resources were provided by CalculQuébec/CLUMEQ.

Acknowledgments

The authors thank EIC F. Anderson, Associate Editor P. Foster, and external reviewers C. Cox and P. Wagner for their constructive comments and whose reviews have greatly improved this manuscript. They appreciate their openness in the review process. They thank the members of the Larsson laboratory at the Redpath Museum for feedback and discussion. They also thank S. B. and W. C. Chun for insightful discussions. They also thank the efforts of the individual researchers whose careful study and quantification of morphology form the raw material for our analysis and particularly their decisions to make their data sets openly available.

References

Alfaro M.E., Holder M.T. 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37:19–42.

Asher R.J., Hofreiter M. 2006. Tenrec phylogeny and the noninvasive extraction of nuclear DNA. Syst. Biol. 55:181–194.

Ayache N.C., Near T.J. 2009. The utility of morphological data in resolving phylogenetic relationships of darters as exemplified with *Etheostoma* (Teleostei:Percidae). Bull. Peabody Mus. Nat. Hist. 50:327–346.

Ayres D.L., Darling A., Zwickl D.J., Beerli P., Holder M.T., Lewis P.O., Huelsenbeck J.P., Ronquist F., Swofford D.L., Cummings M.P., Rambaut A., Suchard M.A. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61:170–173.

Baele G., Lemey P., Bedford T., Rambaut A., Suchard M.A., Alekseyenko A.V. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.

Bourdon E. 2011. The pseudo-toothed birds (Aves, Odontopterygiformes) and their bearing on the early evolution of modern birds. In: Dyke G.J., Kaiser G., editors. Living dinosaurs: the evolutionary history of modern birds. Oxford, UK: John Wiley & Sons, Ltd. p. 209–234.

Brandley M.C., Schmitz A., Reeder T.W. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst. Biol. 54:373–390.

Brown J.M., Hedtke S.M., Lemmon A.R., Lemmon E.M. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. 59:145–161.

Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

Clarke J.A., Middleton K.M. 2008. Mosaicism, modules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. Syst. Biol. 57:185–201.

Dávalos L.M., Cirranello A.L., Geisler J.H., Simmons N.B. 2012. Understanding phylogenetic incongruence: lessons from phyllostomid bats. Biol. Rev. 87:991–1024.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P.O. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28: 523–532.

Felsenstein J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. 53:447–455.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates, Inc.

Frick H., Nentwig W., Kropf C. 2010. Progress in erigonine spider phylogeny—the *Savignia*-group is not monophyletic (Araneae: Linyphiidae). Org. Divers. Evol. 10:297–310.

Fröbisch N.B., Schoch R.R. 2009. Testing the impact of miniaturization on phylogeny: paleozoic dissorophoid amphibians. Syst. Biol. 58:312–327.

Gelman A., Rubin D.B. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472.

Goloboff P.A., Pol D. 2006. Parsimony and Bayesian phylogenetics. In: Albert V.A., editor. Parsimony, phylogeny, and genomics. Oxford, UK: Oxford University Press. p. 148–159.

Holland B.R., Spencer H.G., Worthy T.H., Kennedy M. 2010. Identifying cliques of convergent characters: concerted evolution in the cormorants and shags. Syst. Biol. 59:433–445.
Houle D. 2001. Characters as the units of evolutionary change. In: Wagner G.P., editor. The character concept in evolutionary biology. San Diego: Academic Press. p. 109–140.
Huelsenbeck J.P., Suchard M.A. 2007. A nonparametric method for accommodating and testing across-site rate variation. Syst. Biol. 56:975–987.
Izquierdo-Carrasco F., Smith S.A., Stamatakis A. 2011. Algorithms, data structures, and numerics for likelihood-based phylogenetic inference of huge trees. BMC Bioinformatics 12:470.
Johnson N.L., Kotz S., Balakrishnan N. 1994. Continuous univariate distributions. Vol. 1. New York (NY): John Wiley & Sons, Ltd.
Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90: 773–795.
Kim J., Kim M. 2001. The mathematical structure of characters and modularity. In: Wagner G.P., editor. The character concept in evolutionary biology. San Diego: Academic Press. p. 215–236.
Lambkin C.L., Bartlett J.S. 2011. Bush Blitz aids description of three new species and a new genus of Australian beeflies (Diptera, Bombyliidae, Exoprosopini). Zookeys 150:231–280.
Lanfear R., Calcott B., Ho S.Y., Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.
Larsson H.C., Dececchi T.A., Harrison L.B. 2012. Morphological largess: can morphology offer more and be modeled as a stochastic evolutionary process? In: Asher R.J., Müller J., editors. From clone to bone: the synergy of morphological and molecular tools in paleobiology. Cambridge, UK: Cambridge University Press. p. 83–116.
Lee M.S., Worthy T.H. 2012. Likelihood reinstates *Archaeopteryx* as a primitive bird. Biol. Lett. 8:299–303.
Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.
Li W.L.S., Drummond A.J. 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 29:751–761.
Liljeblad J., Ronquist F., Nieves-Aldrey J.L., Fontal-Cazalla F., Ros-Farre P., Gaitros D., Pujade-Villar J. 2008. A fully web-illustrated morphological phylogenetic study of relationships among oak gall wasps and their closest relatives (Hymenoptera: Cynipidae). Zootaxa 1796:1–73.
Livezey B.C., Zusi R.L. 2006. Phylogeny of Neornithes. Bull. Carnegie Mus. Nat. Hist. 37:1–544.
Livezey B.C., Zusi R.L. 2007. Higher-order phylogeny of modern birds (Theropoda, Aves: Neornithes) based on comparative anatomy. II. Analysis and discussion. Zool. J. Linn. Soc. Linn. Soc. 149:1–95.
Marshall D.C. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. 59:108–117.
Mayrose I., Friedman N., Pupko T. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21(Suppl. 2):ii151–ii158.
Müller J., Hipsley C.A., Head J.J., Kardjilov N., Hilger A., Wuttke M., Reisz R.R. 2011. Eocene lizard from Germany reveals amphisbaenian origins. Nature 473:364–367.
Müller J., Reisz R.R. 2006. The phylogeny of early eureptiles: comparing parsimony and Bayesian approaches in the investigation of a basal fossil clade. Syst. Biol. 55:503–511.
Naish D., Dyke G., Cau A., Escuillié F., Godefroit P. 2012. A gigantic bird from the Upper Cretaceous of Central Asia. Biol. Lett. 8:97–100.
Nylander J.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J.L. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53: 47–67.
O'Leary M.A., Bloch J.I., Flynn J.J., Gaudin T.J., Giallombardo A.G., Giannini N.P., Goldberg S.L., Kraatz B.P., Luo Z.-X., Meng J., Ni X., Novacek M.J., Perini F.A., Randall Z.S., Rougier G.W., Sargis E.J., Silcox M.T., Simmons N.B., Spaulding M., Velazco P.M., Weksler M., Wible J.R., Cirranello A.L. 2013. The Placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339: 662–667.
Paradis E., Claude J., Strimmer K. 2004. APE: analysis of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.
Plummer M., Best N., Cowles K., Vines K. 2006. CODA: convergence diagnostics and output analysis for MCMC. R-News 6:7–11.
Posada D., Buckley T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53: 793–808.
Prieto-Márquez A. 2010. Global phylogeny of Hadrosauridae (Dinosauria: Ornithopoda) using parsimony and Bayesian methods. Zool. J. Linn. Soc. 159:435–502.
Pupko T., Mayrose I. 2010. Probabilistic methods and rate heterogeneity. In: Lodhi H.M., Muggleton S.H., editors. Elements of computational systems biology Vol 8. Hoboken (NJ): John Wiley & Sons, Ltd. p. 257–280.
Pyron R.A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst. Biol. 60:466–481.
R Core Team. 2012. R: a language and environment for statistical computing. Available from: URL http://www.r-project.org/. Last accessed on December 11, 2014.
Rambaut A., Drummond A.J. 2007. Tracer, version 1.5. Oxford UK: Oxford University. Available from http://tree.bio.ed.ac.uk/software/tracer/. Last accessed on December 11, 2014.
Rannala B., Zhu T., Yang Z. 2012. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29:325–335.
Riedl R. 1978. Order in living organisms: a system analysis of evolution. New York: Wiley.
Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61:973–999.
Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.
Schopf T.J., Raup D.M., Gould S.J., Simberloff D.S. 1975. Genomic versus morphologic rates of evolution: influence of morphologic complexity. Paleobiology 1:63–70.
Shapiro B., Rambaut A., Drummond A.J. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7–9.
Sharkey M.J., Carpenter J.M., Vilhelmsen L., Heraty J., Liljeblad J., Dowling A.P., Schulmeister S., Murray D., Deans A.R., Ronquist F., Krogmann L., Wheeler W.C. 2012. Phylogenetic relationships among superfamilies of Hymenoptera. Cladistics 28:80–112.
Sidlauskas B.L., Vari R.P. 2008. Phylogenetic relationships within the South American fish family Anostomidae (Teleostei, Ostariophysi, Characiformes). Zool. J. Linn. Soc. 154:70–210.
Skinner A. 2010. Rate heterogeneity, ancestral character state reconstruction, and the evolution of limb morphology in *Lerista* (Scincidae, Squamata). Syst. Biol. 59:723–740.
Spencer M.R., Wilberg E.W. 2013. Efficacy or convenience? Model-based approaches to phylogeny estimation using morphological data. Cladistics 29:663–671.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Sterli J., Pol D., Laurin M. 2013. Incorporating phylogenetic uncertainty on phylogeny-based palaeontological dating and the timing of turtle diversification. Cladistics 29:233–246.
Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Sullivan J., Swofford D.L. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.
Swofford D. 2003. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Tourasse N.J., Gouy M. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol. Biol. Evol. 14: 287–298.
Waddell P.J., Penny D., Moore T. 1997. Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol. Phylogenet. Evol. 8:33–50.
Wagner G.P. 1995. The biological role of homologues: a building block hypothesis. N. Jb. Geol. Paläont. Abh. 195:279–288.
Wagner G.P., Altenberg L. 1996. Perspectives: complex adaptations and the evolution of evolvability. Evolution 50:967–976.
Wagner P.J. 2012. Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. Biol. Lett. 8:143–146.
Wiens J.J., Bonett R.M., Chippindale P.T. 2005. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander relationships. Syst. Biol. 54:91–110.
Wiens J.J., Kuczynski C.A., Townsend T., Reeder T.W., Mulcahy D.G., Sites J.W.J. 2010. Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: molecular data change the placement of fossil taxa. Syst. Biol. 59:674–688.
Wood H.M., Matzke N.J., Gillespie R.G., Griswold C.E. 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst. Biol. 62: 264–284.
Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.
Xu X., Pol D. 2014. *Archaeopteryx* , paravian phylogenetic analyses, and the use of probability-based methods for palaeontological datasets. J. Syst. Palaeontol. 12:323–334.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.
Yang Z. 2006. Computational molecular evolution. Oxford, UK: Oxford University Press.
Yang Z., Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13:303–314.
Zhang C., Rannala B., Yang Z. 2012. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. Syst. Biol. 61: 779–784.