MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 1944–1957 doi: 10.1111/mec.13590

Assessing the performance of DNA barcoding using posterior predictive simulations

ANTHONY J. BARLEY and ROBERT C. THOMSON
Department of Biology, University of Hawai'i at Mānoa, Honolulu, HI 96822, USA

Abstract

Accurate estimates of biodiversity are required for research in a broad array of biological subdisciplines including ecology, evolution, systematics, conservation and biodiversity science. The use of statistical models and genetic data, particularly DNA barcoding, has been suggested as an important tool for remedying the large gaps in our current understanding of biodiversity. However, the reliability of biodiversity estimates obtained using these approaches depends on how well the statistical models that are used describe the evolutionary process underlying the genetic data. In this study, we utilize data from the Barcode of Life Database and posterior predictive simulations to assess the performance of DNA barcoding under commonly used substitution models. We demonstrate that the success of DNA barcoding varies widely across DNA substitution models and that model choice has a substantial impact on the number of operational taxonomic units identified (changing results by ~4–31%). Additionally, we demonstrate that the widely followed practice of *a priori* assuming the Kimura 2-parameter model for DNA barcoding is statistically unjustified and should be avoided. Using both data-based and inference-based test statistics, we detect variation in model performance across taxonomic groups, clustering algorithms, genetic divergence thresholds and substitution models. Taken together, these results illustrate the importance of considering both model selection and model adequacy in studies quantifying biodiversity.

Keywords: biodiversity, clustering algorithms, genetic distances, model adequacy, operational taxonomic units, substitution models

Received 30 September 2015; revision received 5 January 2016; accepted 18 January 2016

Introduction

Quantifying and understanding the diversity of life on earth is one of the major challenges for the biological sciences in the 21st century (Losos *et al.* 2013). The total number of species present on earth and the taxonomic boundaries among them are fundamental pieces of information for a large number of fields ranging from evolution, development, medicine and speciation to ecology, conservation, agriculture and pest management (Cracraft 2002; Sites & Marshall 2004; Tewksbury *et al.* 2014). Despite their fundamental nature, these remain open questions on a global scale. For example, recent estimates of the number of species on earth vary by two orders of magnitude (e.g. Chapman 2009; Hamilton *et al.* 2010; Mora *et al.* 2011, 2013; Costello *et al.* 2012, 2013; Scheffers *et al.* 2012; Adams *et al.* 2014), which reflects a strong lack of information about these questions. Amidst this uncertainty, it is clear that there is a substantial gap between the number of currently described species and the actual number of species that are believed to exist, as well as that biodiversity is being lost to extinction at an increasingly rapid rate (Wilson 2004; Wiens 2007; Rodrigues *et al.* 2010; Dubois 2011; Mora *et al.* 2011; Niemiller *et al.* 2013; Régnier *et al.* 2015).

In recent years, the use of genetic data and statistical models has contributed strongly to advances in the field of evolutionary biology and these will continue to serve as essential tools in efforts aimed at addressing gaps in our understanding of biodiversity. For example, the use

Correspondence: Anthony J. Barley, Fax: (808) 956 4745; E-mail: ajbarley@hawaii.edu

© 2016 John Wiley & Sons Ltd

of genetic distances (calculated from DNA sequence data using models of DNA substitution) has become an increasingly popular approach for performing rapid biodiversity assessments and identifying operational taxonomic units (OTUs), particularly in 'hyperdiverse' clades of organisms (e.g. Smith *et al*. 2005; Meier *et al*. 2006; Saitoh *et al*. 2015). Much of this popularity can be attributed to the rise of DNA barcoding for species identification and discovery (Hebert *et al*. 2003). Genetic distances have also been used as a proxy for preliminarily delimiting species in lieu of, or in combination with, morphological characters in an increasing number of 'cryptic species complexes' that have been identified (Bradley & Baker 2001; Vences *et al*. 2005; Fregin *et al*. 2012; Nagy *et al*. 2012; Leavitt *et al*. 2015). In some cases, researchers have cited genetic distances equivalent to or greater than those observed among already described species as evidence that populations likely represent distinct species (e.g. Burbrink *et al*. 2000; Price 2010; He *et al*. 2014; Guayasamin *et al*. 2015). Alternatively, other researchers have sought to identify a 'barcoding gap' in genetic distances that represents the distinction between interspecific and intraspecific taxonomic sampling (Hebert *et al*. 2004; Derycke *et al*. 2010; Čandek & Kuntner 2015). Researchers working on basic ecological and evolutionary questions in poorly understood biological communities also require estimates of species diversity and an understanding of community composition (e.g. Pfenninger *et al*. 2007; Pompanon *et al*. 2012; Geisen *et al*. 2015). This has lead to an increase in the use of metabarcoding for quantifying biodiversity in mass-collected biological samples or environmental DNA (Coissac *et al*. 2012; Taberlet *et al*. 2012b; Yoccoz 2012).

The reliability of inferences drawn under a specified model depends on the extent to which the model adequately describes the process that generated the data (Brown 2014a,b). In the case of DNA barcoding, the extent to which calculated genetic distances reflect true genetic distances among taxa depends on how well the model of sequence evolution describes the evolutionary processes that produced the observed sequence data (Felsenstein 2004). If the simplifying assumptions of substitution models systematically bias genetic distance calculations, this may subject inferences regarding evolutionary patterns and processes to errors of accuracy and/or precision. DNA barcoding typically involves calculating pairwise genetic distances under standard Markov models of sequence evolution. Individuals in barcode gene data sets are subsequently grouped into OTUs using the genetic distance matrices in sequence clustering algorithms based on a divergence threshold (e.g. Sun *et al*. 2009; Puillandre *et al*. 2012; Ratnasingham & Hebert 2013). The field has historically relied on a single, relatively simple, model of substitution (the Kimura 2-parameter model; K2P) for calculating distances (Kimura 1980). This largely reflects historical inertia following Hebert *et al*. (2003), which suggested that, because barcoding data sets generally consist of closely related sequences, the K2P model should be sufficient for genetic distance calculation. Standard model selection approaches have widely shown that the K2P may not be an ideal model for many barcoding data sets; however, there is conflicting evidence as to whether or not these inadequacies impact species identification success (Collins *et al*. 2012; Fregin *et al*. 2012; Srivathsan & Meier 2012). Because of this, there is a considerable possibility that the continuing widespread use of the K2P model may systematically bias barcoding studies. Beyond this, the impact that the choice of substitution model has on the resulting estimate of the number of OTUs is in need of further exploration (Kekkonen & Hebert 2014).

Posterior predictive assessments of model performance provide a natural framework for addressing these issues (Gelman *et al*. 2013). These assessments generally involve (i) estimating a joint posterior distribution of model parameters by performing Bayesian inference under a model, (ii) simulating posterior predictive data sets using draws of parameter values from the posterior and (iii) comparing the observed data to the posterior predictive distribution using a test statistic. The rationale for this approach is that a model that adequately describes the observed data should be able to predict (or simulate) new data that are similar to the observed data (as measured by a test statistic). In this study, we utilize posterior predictive simulation to assess the fit of Markov models of sequence evolution to DNA barcoding data sets. We employ this approach in a comprehensive evaluation of model fit in DNA barcoding using data available from the Barcode of Life Database (BOLD) across a large set of animal and plant clades. In doing so, we address several questions: (i) Does the choice of substitution model impact the number of operational taxonomic units that are identified in DNA barcoding studies? (ii) Do standard substitution models adequately describe DNA barcoding data sets? (iii) Do model inadequacies impact the reliability of biodiversity estimates obtained using DNA barcoding? and (iv) In what circumstances do various substitution models perform poorly, and can this provide guidance for future studies? We take this approach with the ultimate goal of directly linking the adequacy of models of sequence evolution in calculating genetic distances to the number of OTUs that are identified in DNA barcoding.

Materials and methods

Briefly, the analysis pipeline that we use here follows three steps. We first assembled a large series of empirical

© 2016 John Wiley & Sons Ltd

data sets to use for substitution model performance assessments. For each data set, we then selected a 'best fit' model of molecular evolution and performed DNA barcoding analyses under both the chosen model and the K2P model to estimate the number of OTUs. Finally, we performed Bayesian phylogenetic inference and posterior predictive simulation to assess the fit of each substitution model to each data set (Fig. 1).

Data set assembly

We first downloaded all available data for animals and plants from BOLD (http://www.boldsystems.org; Accessed: June 18, 2014) and assembled the DNA sequences into exemplar DNA barcode data sets similar to those that are widely employed in the literature. For animals, all sequences for the CO1 gene that were >500 bp were parsed into fasta files for each genus of organisms. Identifying a universal barcoding gene in plants has been a much more contentious process (CBOL Plant Working Group *et al.* 2009), although the majority of available data on BOLD for plants are for the *rbcL* and *matK* genes. Therefore, for each phylum of plants, we utilized the barcoding gene that had the most data available when assembling data sets for each genus. We then filtered all the data sets and included only those that had at least 5 named species and 30 sequences available so that we could focus on well-sampled clades. This approach reflected our rationale that barcoding data sets typically consist of a large number of sequences for multiple, relatively closely related species.

Sequences were aligned using MUSCLE v3.8.31 (Edgar 2004) and quality filtered using TRIMAL v1.2

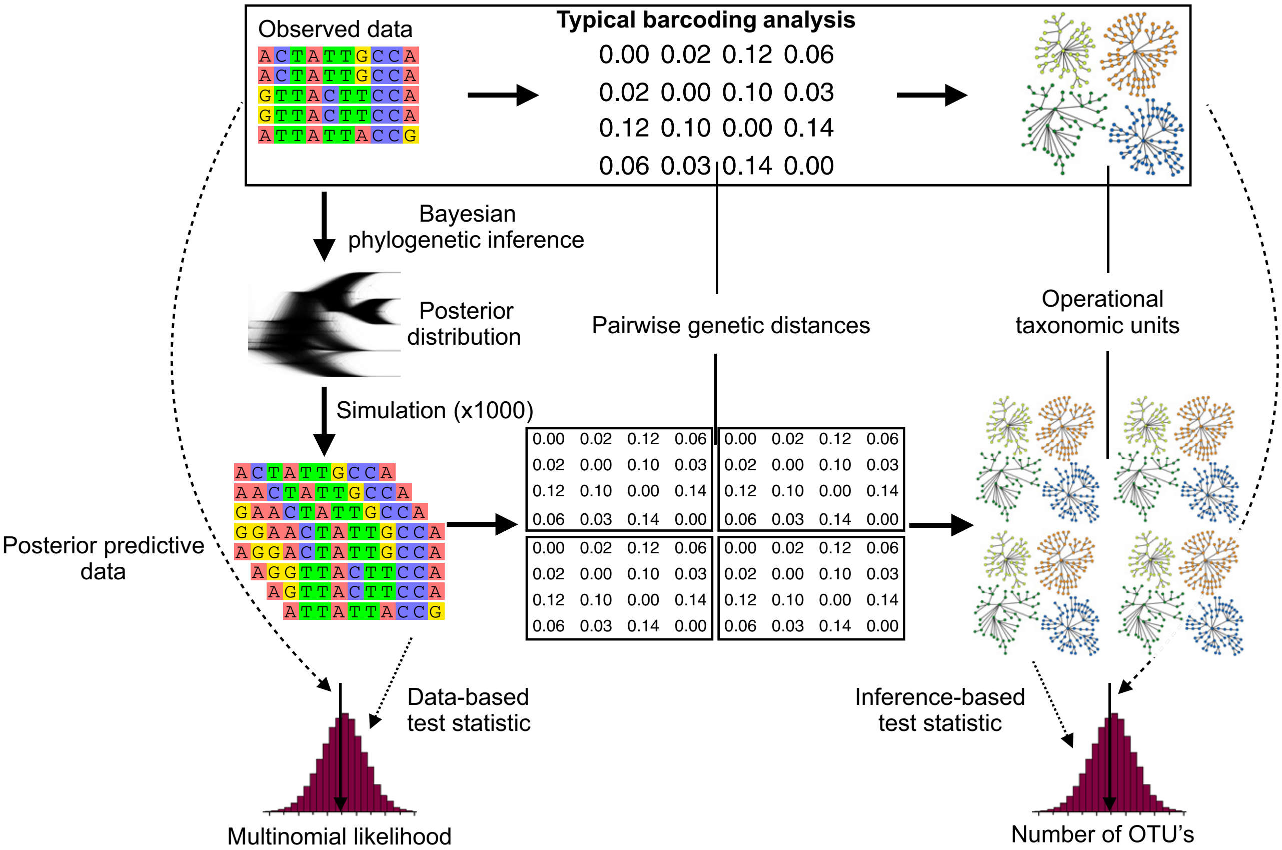

Fig. 1 Outline of the model adequacy assessments. A typical empirical barcoding approach involves calculating pairwise genetic distances under a substitution model and then using a clustering algorithm to assign sequences to OTUs. Here, we generate a posterior distribution of substitution model parameters by performing Bayesian phylogenetic inference. Posterior predictive simulation involves fitting the model to a data set by sampling parameter values from the posterior and using them to simulate new data sets. The number of OTUs in each predictive data set is also determined by calculating pairwise genetic distances and performing clustering; this effectively generates a posterior predictive distribution for the number of OTUs. Histograms illustrate the comparison of the observed value of the test statistic to the distribution of values calculated from the posterior predictive data sets. In the first, we compare the original data set to the simulated data sets using a data-based test statistic (the multinomial likelihood). In the second, we compare the number of OTUs delimited using the original data to the number delimited in each of the predictive data sets.

© 2016 John Wiley & Sons Ltd

(Capella-Gutiérrez *et al.* 2009) as follows. After an initial alignment was performed, poorly aligned sequences were removed using the 'resoverlap' and 'seqoverlap' commands (removing sequences below a 50% threshold). For animals, alignments were trimmed to the 648-bp 'barcoding region' of the CO1 gene (Hebert *et al.* 2003). Sequences that contained more than 100 bp of missing data or 'Ns' due to ambiguous base calls were then removed from the alignment. Finally, any remaining gaps were removed from the alignments, and alignments that were <500 base pairs long or contained <30 sequences were discarded. A small number of alignments contained an anomalously large number of individuals, with one or several species being substantially overrepresented. For computational efficiency, if an alignment contained more than 800 total individuals, sequences were filtered out of it so that no species was represented by more than 20 individuals in the alignment.

Barcoding models

Our goal was to assess the adequacy of both the widely used K2P model and a 'best fit' model of sequence evolution that was chosen using a standard model selection approach. Genetic distances calculated under each model would subsequently be used in sequence-divergence-based clustering algorithms to identify OTUs. We examined two commonly employed clustering algorithms for these assessments: the Automatic Barcode Gap Discovery method (ABGD; Puillandre *et al.* (2012)) and the hierarchical clustering algorithm implemented in hcluster (Sun *et al.* 2009). hcluster implements a complete-link hierarchical clustering algorithm based on a specified divergence threshold. The ABGD algorithm seeks to infer a 'barcode gap' between intraspecific and interspecific genetic distances. It ranks pairwise distances by increasing values and uses a local slope function to identify the first statistically significant peak in the slope values that corresponds to a gap in the distribution of pairwise distances. Finally, it partitions the data into OTUs based on the inferred barcode gap. For each data set, we selected the best fitting model of molecular evolution from among 12 commonly used substitution models that range from very simple to more complex and that can be implemented in MrBayes (Ronquist *et al.* 2012). This included the JC, F81, K2P, HKY85, SYM and GTR substitution models with or without allowing for rate variation across sites using a gamma distribution (Felsenstein 2004). jModelTest2 was employed to select among the substitution models using the AICc criteria (Darriba *et al.* 2012). For a small number of data sets, the simple JC, F81 and K2P models were selected as the best fit model. Because we were primarily interested in assessing model performance in circumstances where overly simplistic models are arbitrarily used, these data sets were not included in subsequent analyses.

We calculated pairwise genetic distance matrices for each data set under both the K2P and selected models using PAUP* v4.0b10 (Swofford 2003), utilizing the ML estimate for the shape parameter of the gamma distribution (α) from jModelTest2 when appropriate. Preliminary analyses suggested that accurate estimates for α were difficult to obtain for some data sets that contained a large amount of rate variation across sites and we therefore conservatively excluded data sets from further analysis if the estimate for α was <0.03. The genetic distance matrices were then used with ABGD and hcluster to calculate the empirical estimates for the number of OTUs for each data set using divergence threshold values ranging from 1% to 10%. For ABGD, we used the default proxy for minimum gap width of 1.5 and focused on examining model performance in analyses where the prior limit to intraspecific diversity was set at 0.01, 0.02 and 0.03, as these are recommended as the optimal parameter values by the authors of the method. For hcluster, we focused on divergence threshold values of 0.03, 0.05 and 0.10 in animals, and 0.01, 0.03 and 0.05 for plants. Pairwise sequence divergence values of 3–5% are frequently cited in the barcoding literature as the threshold separating inter- and intraspecific sampling in animals. Genetic distance values >10% are frequently cited as evidence for distinct populations potentially representing undescribed species, and since genetic distances calculated under increasingly complex models tend to be larger, we felt this represented a reasonable upper limit at which to assess the effects of model performance. Lower divergence threshold values were used for plants because the barcoding genes used generally exhibit much lower levels of sequence divergence than CO1 does in animals (CBOL Plant Working Group *et al.* 2009).

Posterior predictive checks

We utilized Bayesian phylogenetic inference to estimate the posterior distribution of substitution model parameters for an empirical data set, which we subsequently drew samples from to simulate predictive data sets. Test statistics used in posterior predictive checks can be broadly viewed as 'data-based' (which assess whether the observed and posterior predictive data sets exhibit similar characteristics) or 'inference-based' (where inferences drawn from the observed and posterior predictive data sets are compared; Brown 2014a,b). Inference-based test statistics are used to link deficiencies of the model to their impact on resulting inferences, as

© 2016 John Wiley & Sons Ltd

researchers are often primarily concerned whether inadequacies of the model actually influence particular aspects of inference they are interested in (as opposed to merely affecting the similarity of the data sets themselves). We included both types of test statistics in our model adequacy assessments for comparison. For each data set, we performed Bayesian phylogenetic inference under the K2P and the selected models of molecular evolution using MrBayes 3.2.2 (Ronquist *et al.* 2012). Each data set was analysed as a single partition (although we explore the impact of partitioning below) and, with the exception of the prior on branch lengths, all other priors were left at their default values.

Branch length estimation in Bayesian phylogenetics is currently an active area of research, and the most appropriate way to set priors on branch lengths is a subject of considerable focus (Brown *et al.* 2010; Marshall 2010; Rannala *et al.* 2012; Zhang *et al.* 2012; Nelson *et al.* 2015). The use of independent and identically distributed exponential priors has been shown to cause overly long branch lengths to be inferred for certain data sets, particularly 'intraspecific' data sets where true branch lengths are likely short (as is often the case for barcoding data sets). In practice, two approaches have been implemented to circumvent this issue. The first is to choose a different parameterization for the prior on branch lengths that induces less bias, while the second is to inform prior distributions based on empirical data. The use of the compound Dirichlet tree-length prior has helped improve branch length estimation in Bayesian phylogenetics for many data sets (Rannala *et al.* 2012; Zhang *et al.* 2012), although it is also clear that this prior has not entirely alleviated the problem (Nelson *et al.* 2015). Informing priors based on ML parameter estimates from the data set to be analysed has been criticized for being non-Bayesian and for artificially reducing uncertainty in the posterior distribution (Zhang *et al.* 2012). Informing priors based on other 'similar' data sets has also been proposed (Nelson *et al.* 2015). However, even when data sets are available for comparison, assessing data set 'similarity' can be difficult and does not guarantee that estimates obtained using those data sets will be appropriate for the focal data set being analysed (Nelson *et al.* 2015). Given that we were interested purely in assessing adequacy of the substitution models, limiting potential biases from other aspects of the phylogenetic model was important. Therefore, we chose to utilize the compound Dirichlet priors on branch lengths and inform the parameters of these distributions based on the ML parameter estimates for each data set. Given the lack of consensus on this subject, as well as the importance of having accurate branch length parameter estimates for simulating posterior predictive data sets, we felt that this was the most effective approach for achieving our model assessment goals. We also felt this approach was the most conservative in terms of giving the substitution model the best 'opportunity' to perform well (i.e. when good information is available for informing priors) and avoiding bias from other aspects of the model that would influence our results. The compound Dirichlet prior on branch lengths in MrBayes has four parameters: the tree length shape (α_T), rate (λ_T), Dirichlet concentration (α) and internal to external branch length ratio (c). For each data set, we performed maximum-likelihood phylogenetic analysis under the K2P and the best fit models of molecular evolution using GARLI v2.0 (Zwickl 2006). Five replicates were run for each analysis, choosing the tree from the run with the highest likelihood as the optimal tree. For our analyses, we set $\alpha_T = 1$ to represent a diffuse prior on tree length and performed ML parameter estimation to obtain parameter estimates for λ_T, α and c. We used the Nelder–Mead method as implemented in the 'OPTIM' function in R 3.0 and the R package 'BBMLE' v1.0.17 for ML parameter estimation (Nelson *et al.* 2015). Under the compound Dirichlet tree-length prior, the tree length is considered drawn from a gamma distribution with mean α_T/λ_T. The variance of the branch lengths in the tree is specified by α, while c specifies the ratio of the prior means for the internal and external branch lengths.

Each analysis was run initially using two replicates with four chains each for 10 million generations, sampling every thousand generations. To ensure sufficient mixing, the temperature parameter was given a value between 0.01 and 0.05, and adjusted to accommodate data sets that exhibited poor mixing (<20%) among chains in initial analyses. Convergence was assessed using a Python script that checked three diagnostics for each parameter in the analyses: the potential scale reduction factor (PSRF), the minimum effective sample size (ESS) and the average ESS. If the PSRF was >1.001, the minimum ESS was <1000, or the difference between the minimum ESS and the average ESS was >2000, and analyses were manually checked for convergence using TRACER v1.6 (Rambaut *et al.* 2014). In most cases, 10 million generations was a sufficient run length; however, some larger data sets were run for 20 million generations (sampling every 2000 generations) after preliminary analyses failed to achieve convergence and sufficient ESSs of all parameters. In a small number of MrBayes analyses, the estimated marginal distribution for one or more parameters did not stabilize over the course of the runs, which usually indicates that the data are not sufficiently informative to estimate those parameters. These data sets were ultimately dropped from the study.

We next performed posterior predictive simulations for each data set under both the K2P and selected

© 2016 John Wiley & Sons Ltd

models. Posterior predictive sequence data sets were simulated using PuMA v1.0 (Brown & ElDabaje 2009). One thousand parameter values and trees for simulation were randomly sampled from the post burn-in posterior distributions of each MrBayes analysis (sampling 500 from each independent run after the first 50% of samples had been discarded as burn-in). We first utilized the multinomial likelihood to assess model adequacy (Bollback 2002), which is a data-based test statistic used to compare the frequency of site patterns between observed and simulated data sets. In this case, the value of the test statistic is calculated for all simulated data sets, and this distribution of values is then compared to the value of test statistic in the empirical data set. For a given data set analysed under a plausible model, the expectation for the multinomial likelihood posterior predictive *P*-value is 0.5 (Brown & ElDabaje 2009), and values close to 0 or 1 indicate poor model fit. We also compared these results to a model adequacy assessment using an inference-based test statistic that would allow us to directly assess adequacy from the perspective of the focal inference in DNA barcoding studies: the number of OTUs identified. We calculated the number of OTUs in each of the simulated posterior predictive data sets under a range of sequence divergence thresholds using both ABGD and hcluster as we did for each of the empirical data sets. This resulted in a posterior predictive distribution for the number of OTUs in each data set for both the K2P and selected models of sequence evolution. Model adequacy was assessed by determining whether the empirical estimate for the number of OTUs in a data set fell within the 95% highest posterior density of the posterior predictive distribution (using the 2.5% and 97.5% quantiles).

Model improvement

We took several approaches to identify circumstances in which the K2P and selected models performed poorly and directions for model improvement in DNA barcoding. First, we quantified model adequacy using two continuous variables: the frequency of the empirical estimate for the number of OTUs within the posterior predictive distribution (PPD) and a standardized measure of error (calculated as:

$$\frac{|x - \bar{x}|}{\sigma}$$

where x is the empirical estimate of the number of OTUs for a data set, and $\bar{x}$ and σ are the mean and standard deviation of the PPD, respectively). These variables were plotted against summary statistics of the data sets that were hypothesized to be associated with variation in the adequacy of the K2P model (e.g. data set size, average genetic distance among sequences, variance in base frequency composition, tree length, number of described species, mean substitution rate estimates).

Partitioning across sites

Partitioning has been demonstrated to be an important tool for accommodating process heterogeneity across sites in molecular data sets (Bull *et al.* 1993; Nylander *et al.* 2004) and for accurate model-based phylogenetic inference (Brandley *et al.* 2005; Brown & Lemmon 2007; Fan *et al.* 2011; Kainer & Lanfear 2015). To assess whether data partitioning could improve model performance in DNA barcoding, we chose a random subset of 80 data sets exhibiting poor model performance under the selected model (data sets where the observed number of OTUs failed to fall within the posterior predictive distribution at any divergence threshold for a given clustering algorithm). We used AICc implemented in PartitionFinder v1.1.1 (Lanfear *et al.* 2012) to select a partitioning scheme and corresponding models of molecular evolution. We allowed for different codon positions to be treated as separate partitions and for substitution models to be chosen among the same set as used above. PartitionFinder suggested the use of multiple partitions for 60 of the 80 data sets for which we performed phylogenetic inference and posterior predictive simulation under the selected partitioned models as discussed above. Rather than *a priori* assigning sites in an alignment to different partitions, site-specific profiles and rates can also be treated as random variables and estimated using a Dirichlet process (Lartillot & Philippe 2004). We utilized the CAT-GTR model implemented in Phylobayes-MPI v1.5a to perform phylogenetic inference and posterior predictive simulation on these same 80 data sets. The analyses were run for 10 000 cycles, sampling every 10 cycles, and 1000 posterior predictive data sets were simulated using draws from the posterior. In both cases, pairwise genetic distances for the posterior predictive data sets were calculated as above using the selected best fit model, and sequences were clustered into OTUs under the two clustering algorithms.

Process heterogeneity across taxa

All of the most widely used substitution models assume that the substitution process is homogeneous across taxa within a data set. We were therefore further concerned that variation in the substitution process across taxa within a data set could represent another potential source of bias (Yang & Roberts 1995; Nielsen 2002; Foster 2004; Dutheil *et al.* 2012; Jayaswal *et al.* 2014). To examine whether compositional heterogeneity could be a source of model bias in our data sets, we selected a random subset of 40 data sets that exhibited poor model performance (including the 20 data sets for which PartitionFinder suggested the use of a single partition) and performed a chi-squared test of homogeneity using the P4 v1.0 python package (Foster 2004). The posterior predictive data sets simulated using PuMA were used to generate the appropriate null distribution of chi-squared test statistic values for comparison to the observed data.

Results

The data set assembly and filtering strategy resulted in a total of 2083 data sets containing a total of 269 919 sequences that we used for posterior predictive model assessments. As expected given the data availability, the majority of the data sets that we assembled were for arthropods (1245 or ~60%), followed by chordates (442), plants (217; 169 for matK which included the angiosperms and 48 for rbcL), molluscs (127) and 52 from other groups of animals. The most frequently selected model for the barcode data sets was GTR + Γ, followed by HKY + Γ (Table 1). Our comprehensive analysis of barcoding data indicated that the model of sequence evolution that is used to calculate genetic distances frequently (and sometimes substantially) impacts the number of OTUs that are identified using sequence divergence clustering algorithms (Fig. 2). Depending on the method and threshold employed, the total number of OTUs that were identified across all data sets ranged from 15 687 to 35 478. The difference in the total number of OTUs identified between using the K2P and selected model ranged from 1087 to 8103 or 4.5–34.1% of the total number of OTUs identified.

The data-based model adequacy assessments using the multinomial likelihood indicated that the K2P model was unlikely to be a plausible model for nearly any of the data sets examined in our study (the test statistic for the observed data set was only found to be within the 95% highest posterior density (HPD) of the PPD for 3% of all data sets). Conversely, the selected model generally performed well in the multinomial likelihood model adequacy assessments (with the test statistic for the observed data being contained in the 95% HPD for 83% of all data sets; Fig. 3). This stark contrast suggests that the K2P substitution model, in general, poorly captures the process of sequence evolution in DNA barcoding data sets. Model adequacy assessments using the number of OTUs identified as an inference-based test statistic also generally indicated that using the selected model of sequence evolution for genetic distance calculation was superior to using the K2P model (Table 2). However, the difference in model adequacy between the selected and K2P models did not appear as large using this metric, indicating that deficiencies of the K2P model in calculating genetic distances do not always bias the number of OTUs that are identified. Model performance varied between the two clustering algorithms and across divergence thresholds, models and taxonomic groups. All models generally performed well for plant data sets, exhibiting little variation in model adequacy. This was likely a result of the fact that the two plant barcoding genes exhibit low rates of substitution and little genetic divergence was observed among taxa. Conversely, all models performed most poorly in the arthropod data sets, likely a result of the fact that these data sets include the largest number of individuals and species, and largest genetic distances among taxa (arthropod data sets also comprised the majority of data sets in our study). The GTR + Γ model generally exhibited better performance in data sets where it was selected as the best fit model than the HKY + Γ model did where it was selected.

All models exhibited better performance when using the ABGD algorithm to delimit OTUs than when using the hcluster algorithm. The selected model also consistently performed better than the K2P model at all ABGD divergence thresholds (by ~10%; Table 2). Performance across divergence thresholds was similar under the ABGD algorithm, although the lowest threshold had the highest number of data sets for which the observed number of OTUs was included in the 95% HPD of the PPD. This is consistent with the observation that both the K2P and selected models generally 'underpredicted' the number of OTUs when model performance problems occurred.

Model adequacy varied substantially across taxonomic groups and divergence thresholds when assessed using the hcluster algorithm. In contrast to the ABGD algorithm, all models generally 'overpredicted' the number of OTUs when the observed estimate did not

Table 1 Results of model selection analyses showing the number of data sets for which each model was selected as the best fit substitution model. Models were selected using the AICc criteria and the total number of alignments included in the study was 2083

Model	CO1	matK	rbcL
GTR + Γ	986	90	26
GTR	3	23	0
HKY + Γ	767	21	9
HKY	6	34	0
SYM + Γ	3	0	8
K80 + Γ	101	0	5

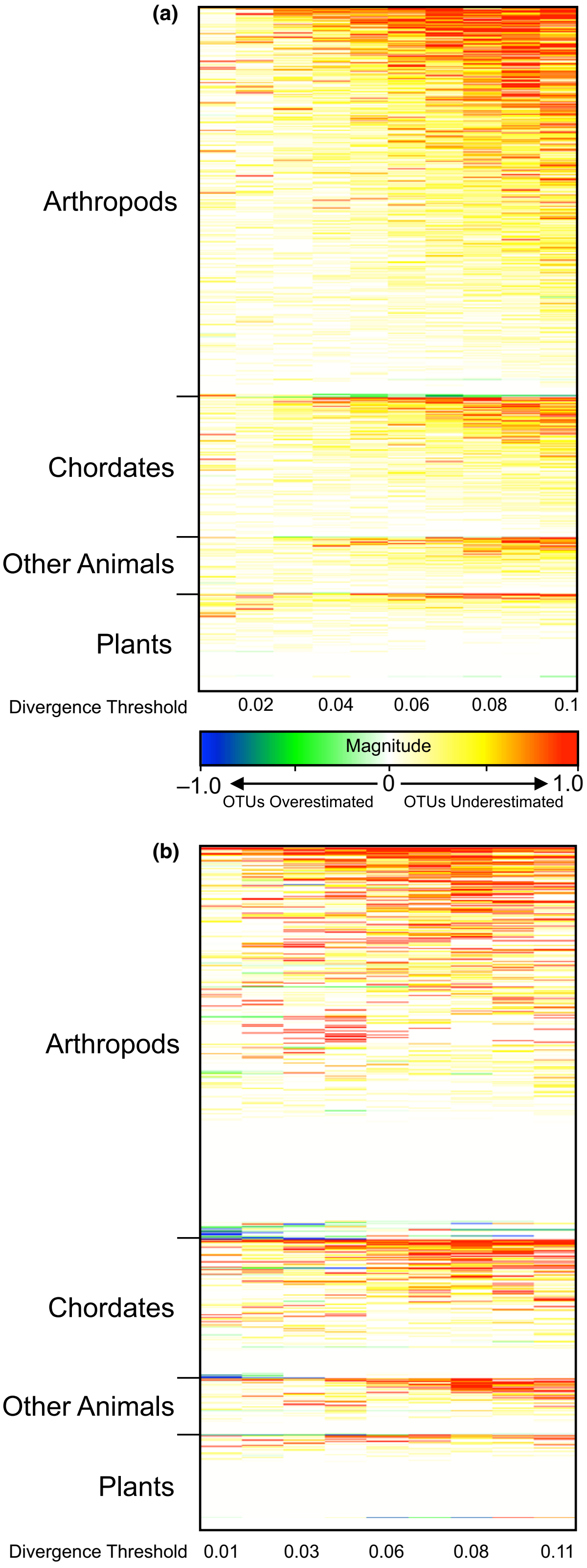

Fig. 2 Heat map showing the difference in the empirical estimate for the number of OTUs between the best fit and K2P models across a range of divergence thresholds for a) the hcluster algorithm and b) the Automatic Barcode Gap Discovery algorithm. Differences are expressed as a proportion relative to the number of species in each data set based on taxonomy and are truncated at a maximum value of 1.0 (or −1.0) for ease of visualization.

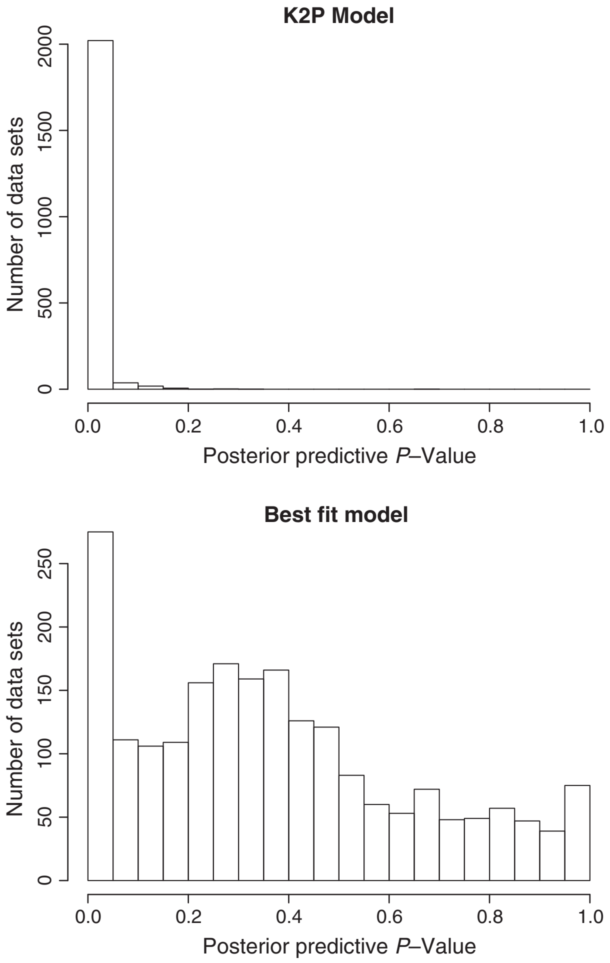

Fig. 3 Comparison of multinomial likelihood test statistic values between K2P and best fit models of molecular evolution for all data sets. For a plausible model, the expectation for the posterior predictive *P*-value is 0.5, whereas values close to 0 or 1 indicate poor model fit. For the K2P model, only 3% of all data sets had test statistic values that fell within the 95% HPD (compared to 83% for the best fit model).

fall within the 95% HPD of the PPD for the hcluster algorithm. Consequently, model performance was better at the higher divergence thresholds, where the selected model again outperformed the K2P. However, the K2P model actually performed better than more complex models in predicting the number of OTUs for arthropods using the hcluster algorithm at the two lower divergence thresholds (although this was not the case for other taxonomic groups). This may result from a combination of the fact that this group included the most complex data sets in the study and the consistent

© 2016 John Wiley & Sons Ltd

Table 2 Results of DNA barcoding model adequacy assessments across taxonomic groups with the number of data sets given in parentheses. Numbers for the three distance thresholds for each method indicate the proportion of data sets for which the empirical estimate of the number of OTUs falls within the 95% HPD of the posterior predictive distribution at that threshold (see text for threshold information)

	Arthropods (1245)		Chordates (442)		Other Animals (179)		Plants (matK: 169)		Plants (rbcL: 49)		Plants Combined		Animals Combined	
	K2P	Best Fit	K2P	Best Fit	K2P	Best Fit	K2P	Best Fit	K2P	Best Fit	K2P	Best Fit	K2P	Best Fit
ABGD Threshold 1	0.74	0.83	0.80	0.86	0.92	0.94	0.89	0.91	0.87	0.94	0.89	0.91	0.77	0.85
ABGD Threshold 2	0.70	0.81	0.75	0.83	0.85	0.91	0.93	0.92	0.92	0.90	0.93	0.92	0.73	0.83
ABGD Threshold 3	0.72	0.82	0.78	0.84	0.87	0.91	0.98	0.96	0.96	0.94	0.98	0.96	0.75	0.84
hcluster threshold 1	0.49	0.33	0.49	0.57	0.37	0.43	0.67	0.71	0.75	0.63	0.69	0.69	0.48	0.40
hcluster threshold 2	0.72	0.58	0.69	0.74	0.44	0.53	0.65	0.66	0.71	0.69	0.66	0.66	0.69	0.61
hcluster threshold 3	0.66	0.70	0.73	0.84	0.65	0.68	0.83	0.86	0.60	0.71	0.78	0.82	0.68	0.73

overprediction of the number of OTUs using the hcluster algorithm.

Although the relationships between the data set characteristics we examined and performance of the K2P model appeared to be complex, the best predictor of poor performance we examined was genetic distance among individuals in a data set, followed to a lesser extent by the number of sequences in an alignment, although both explained a relatively small amount of variance in the results (Fig. 4). Thus, as might be expected, the K2P model appears to perform more poorly in complex data sets. For data sets exhibiting poor performance under an unpartitioned best fit substitution model, model performance (assessed using the hcluster algorithm) improved in 38/40 data sets when a partitioned model (selected using PartitionFinder) was used (i.e. the estimated number of OTUs in the observed data now fell within the posterior predictive distribution for at least one divergence threshold, although frequently performance improved at multiple thresholds). Partitioning improved model performance in 11/20 data sets examined for the ABGD algorithm. Therefore, data set partitioning across codon positions and the use of multiple models of substitution appear to be an important potential source of model inadequacy in DNA barcoding. Use of the CAT-GTR model also improved model performance in many of these data sets. Model performance assessed using the hcluster algorithm improved in 50/53 data sets that we were able to analyse under the CAT-GTR model. The CAT-GTR model improved performance in 14/16 data sets when assessed using the ABGD algorithm. In contrast, chi-squared tests for compositional heterogeneity indicated that variation in base composition across taxa was unlikely to be a large source of model bias in DNA barcoding data sets, as the test statistic for the observed data set only fell outside the 99% credible interval of the posterior predictive distribution for 3 of 40 poorly performing data sets (9 of 40 fell outside the 95% HPD of the posterior predictive distribution).

Discussion

Although previous research has demonstrated the usefulness of DNA barcoding in certain circumstances, its utility on a variety of fronts remains understandably controversial, particularly with respect to delimiting species (Dasmahapatra & Mallet 2006; Elias *et al.* 2007; Collins & Cruickshank 2013; Ji *et al.* 2013; Percy *et al.* 2014). A more basic question regarding DNA barcoding is whether the models of sequence evolution that are actually employed adequately describe these data sets. In other words, even if we accept the biological utility of DNA barcoding, are the models of sequence evolution that are used accurately measuring the number of OTUs present in a data set? We addressed this question by examining the performance of standard substitution models in DNA barcoding data sets across the tree of life using posterior predictive simulation. We focus on the number of OTUs identified in a data set as a test statistic in these model adequacy assessments, as this is usually the focal inference in DNA barcoding studies. Our results demonstrate the importance of considering both model choice and model fit when calculating pairwise genetic distances for use in DNA barcoding.

We find that standard Markov models of sequence evolution frequently perform well in calculating pairwise genetic distances in DNA barcoding data sets. However, our results also show that more complex models consistently outperform the widely employed K2P substitution model. We also demonstrate that the use of different substitution models can substantially impact the number of OTUs that are identified in barcoding data sets. As expected, our analyses indicate that the K2P model

© 2016 John Wiley & Sons Ltd

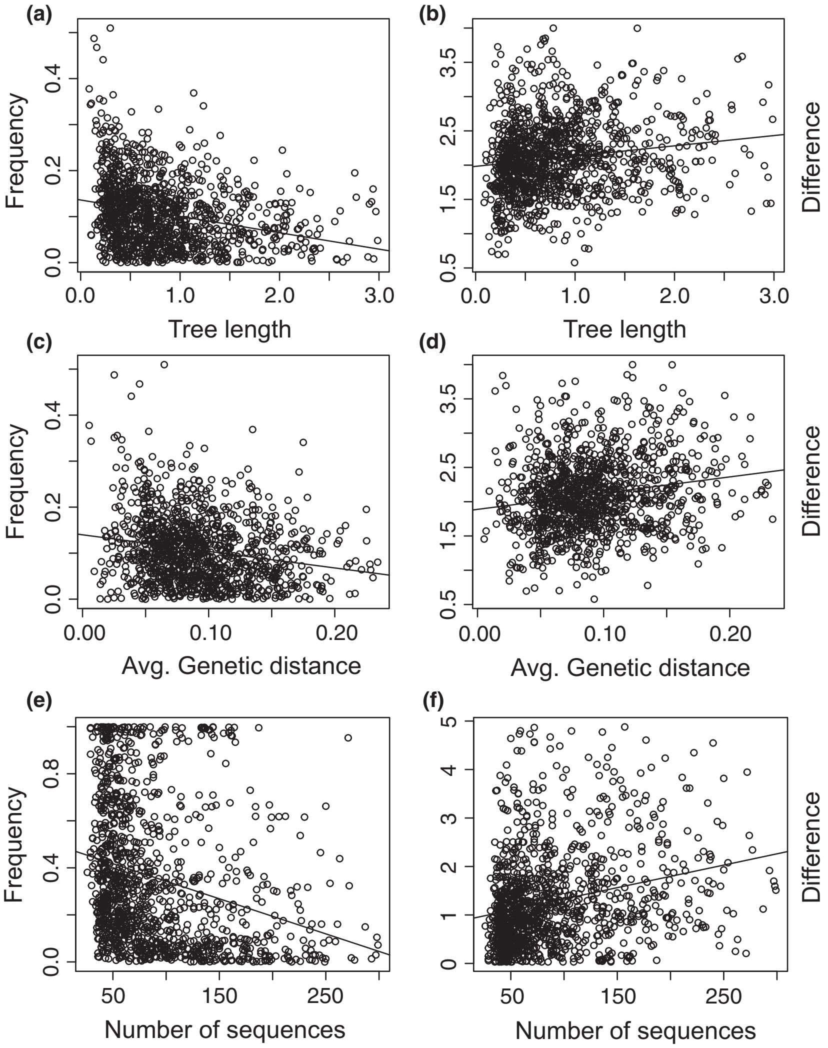

Fig. 4 Scatter plots with regression line showing summary statistics of data sets plotted against two measures of model adequacy for the K2P model: the frequency of the empirical estimate for the number of OTUs within the posterior predictive distribution (a, c, e) and the difference between the empirical number of OTUs and the mean of the posterior predictive distribution (b, d, f). Results are for the hcluster algorithm.

causes increasingly inaccurate estimates of the number of OTUs in a data set as the data grow more complex. In addition to demonstrating the importance of using a sufficiently complex model of sequence evolution, our results suggest that accommodating variation in the substitution process across sites through partitioning can substantially improve accuracy. While our analyses suggest that even the use of relatively simple partitioning strategies can drastically improve model fit, more parameter-rich approaches such as codon models have been developed that could also be employed (e.g. Halpern & Bruno 1998; Rodrigue *et al.* 2010). Although the importance of sufficiently complex models and partitioning molecular data has received a large amount of attention in the phylogenetics community, the barcoding community has been considerably slower to adopt these practices. Our results emphasize the utility of these approaches and suggest that they should be widely employed for DNA barcoding.

Given that DNA barcoding data sets typically comprise sequences from closely related taxa, homogeneity in the process of molecular evolution across taxa might be a reasonable *a priori* assumption, and across the taxa we examined, we observed the best model performance in data sets exhibiting low genetic diversity (e.g. plants). Poor model fit was most common in data sets exhibiting high genetic diversity (i.e. genera likely containing more distantly related taxa such as arthropods, due to the fact that they are not as well understood taxonomically). However, we still found little evidence for compositional heterogeneity as a significant cause of bias in these data sets. If evidence of process heterogeneity is found in empirical data sets, alternative measures of genetic distance have also been proposed that could be used. For example, LogDet or paralinear distances have been proposed (Felsenstein 2004) which correct for variation in base composition across a phylogeny. A general nonstationary Markov model has also

© 2016 John Wiley & Sons Ltd

been proposed that accounts for changes in the substitution process across a tree when calculating pairwise genetic distances among taxa (Kaehler *et al.* 2015). The most flexible option for dealing with process heterogeneity across sites or taxa may be to use patristic distances calculated from branch lengths in a phylogeny for DNA barcoding studies. In doing so, distances could reflect increasingly complex phylogenetic processes and could account for gene tree heterogeneity if multiple loci are available for analysis (Dowton *et al.* 2014). Of course, DNA barcoding is generally regarded as being most useful in large data sets for performing rapid biodiversity assessments in hyperdiverse taxonomic groups. If the time and resources are available to analyse data sets using complex phylogenetic models, OTUs might be more accurately delimited using a tree-based or coalescent-based approach (e.g. Fujisawa & Barraclough 2013; Yang & Rannala 2014). Therefore, DNA barcoding studies should explicitly consider how simplifying assumptions of different models are likely to impact their results based on the individual data set being analysed.

Our study has several broader implications. First, although we focus on model adequacy from the perspective of species discovery, our results are also applicable to DNA barcoding studies that focus on species identification. This may be particularly important in taxonomic groups containing multiple, closely related species where slight inaccuracies in genetic distance calculations could lead to incorrect species identification. For data sets where inference under a particular model appears to be reliable, this approach could also be used as a parametric bootstrapping procedure for generating confidence intervals for the maximum-likelihood estimate of the number of OTUs in a barcoding data set. Finally, we found substantial differences in model fit depending on which clustering algorithm was used to assess substitution model adequacy. These results compliment another recent study demonstrating that the choice of clustering algorithm can impact the number of OTUs identified (Flynn *et al.* 2015).

Given that our results are based on taxa across the tree of life that are relatively well understood and sampled (evidenced by the fact that genetic data is available), the difference we see in the number of OTUs identified between the K2P and best fit models (~4–31%) likely represents a conservative estimate. Extrapolated across a reasonably conservative estimate of 10 million species on earth, that difference in the number of OTUs identified would equate to between 450 000 and more than 3 million species. Clearly the adequacy of statistical models used for quantifying biodiversity will have a substantial impact on our understanding of life on earth. DNA barcoding has been championed as an important step towards closing the gap between the current number of described species and total number that exist on earth (Hebert *et al.* 2003; Schindel & Miller 2005; Bik *et al.* 2012; Collins & Cruickshank 2013). Additionally, DNA barcoding is becoming a popular tool for quantifying biodiversity in studies addressing basic ecological and evolutionary questions (Taberlet *et al.* 2012a; Yu *et al.* 2012; Leray & Knowlton 2015). As DNA barcoding becomes more broadly applied across the tree of life and in increasingly diverse contexts, the composition of barcoding data sets will continue to become more complex and variable. Thus, it is unlikely that a single model of molecular evolution will adequately describe all barcoding data sets. If genetic data are to be at all useful in species identification and discovery, it is essential that the statistical models that are used can provide reliable inference in empirical data sets. Substantial biases in biodiversity estimates would have broad ramifications across studies in the fields of ecology, evolution, conservation, ecosystem structure and function, and biodiversity science.

Acknowledgements

We thank Jeremy Brown, Emilie Richards and Amber Wright for helpful discussions that improved this manuscript. We also thank Hervé Philippe and two anonymous reviewers whose suggestions helped improve this manuscript. We utilized HPC resources provided by the University of Hawaii and Louisiana State University in conducting many analyses presented here. Funding for this work was provided by an Arnold O. Beckman Postdoctoral Fellowship to AJB and by NSF awards DBI 1356796 and DEB 1354506 to RCT.

References

Adams M, Raadik TA, Burridge CP, Georges A (2014) Global biodiversity assessment and hyper-cryptic species complexes: more than one species of elephant in the room? *Systematic Biology*, **63**, 518–533.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Bollback JP (2002) Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, **19**, 1171–1180.

Bradley RD, Baker RJ (2001) A test of the genetic species concept: cytochrome-b sequences and mammals. *Journal of Mammalogy*, **82**, 960–973.

Brandley MC, Schmitz A, Reeder TW (2005) Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic Biology*, **54**, 373–390.

Brown JM (2014a) Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology*, **63**, 334–348.

© 2016 John Wiley & Sons Ltd

Brown JM (2014b) Predictive approaches to assessing the fit of evolutionary models. *Systematic Biology*, **63**, 289–292.

Brown JM, ElDabaje R (2009) PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. *Bioinformatics (Oxford, UK)*, **25**, 537–538.

Brown JM, Lemmon AR (2007) The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic Biology*, **56**, 643–655.

Brown JM, Hedtke SM, Lemmon AR, Lemmon EM (2010) When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates. *Systematic Biology*, **59**, 145–161.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ (1993) Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, **42**, 384–397.

Burbrink FT, Lawson R, Slowinski J (2000) Mitochondrial DNA phylogeography of the polytypic North American rat snake (*Elaphe obsoleta*): a critique of the subspecies concept. *Evolution*, **54**, 2107–2118.

Čandek K, Kuntner M (2015) DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources*, **15**, 268–277.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, UK)*, **25**, 1972–1973.

CBOL Plant Working Group, Hollingsworth PM, Forrest LL *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 12794–12797.

Chapman AD (2009) Numbers of living species in Australia and the world, 80. http://155.187.2.69/biodiversity/abrs/publications/other/species-numbers/2009/pubs/nlsaw-2nd-complete.pdf.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, **13**, 969–975.

Collins RA, Boykin LM, Cruickshank RH, Armstrong KF (2012) Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution*, **3**, 457–465.

Costello MJ, Wilson S, Houlding B (2012) Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology*, **61**, 871–883.

Costello MJ, May RM, Stork NE (2013) Can we name Earth's species before they go extinct? *Science (New York, N.Y.)*, **339**, 413–416.

Cracraft J (2002) The seven great questions of systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanical Garden*, **89**, 127–144.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772.

Dasmahapatra KK, Mallet J (2006) Taxonomy: DNA barcodes: recent successes and future prospects. *Heredity*, **97**, 254–255.

Derycke S, Vanaverbeke J, Rigaux A, Backeljau T, Moens T (2010) Exploring the use of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine nematodes. *PLoS ONE*, **5**, e13716.

Dowton M, Meiklejohn K, Cameron SL, Wallman J (2014) A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. *Systematic Biology*, **63**, 639–644.

Dubois A (2011) Taxonomy in the century of extinctions: taxonomic gap, taxonomic impediment, taxonomic urgency. *TAPROBANICA: The Journal of Asian Biodiversity*, **2**, 1–5.

Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, Boussau B (2012) Efficient selection of branch-specific models of sequence evolution. *Molecular Biology and Evolution*, **29**, 1861–1874.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Elias M, Hill RI, Willmott KR *et al.* (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings Biological Sciences/The Royal Society*, **274**, 2881–2889.

Fan Y, Wu R, Chen M-H, Kuo L, Lewis PO (2011) Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution*, **28**, 523–532.

Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates Inc, Sunderland.

Flynn JM, Brown EA, Chain FJJ, MacIsaac HJ, Cristescu ME (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

Foster P (2004) Modeling compositional heterogeneity. *Systematic Biology*, **53**, 485–495.

Fregin S, Haase M, Olsson U, Alström P (2012) Pitfalls in comparisons of genetic distances: a case study of the avian family Acrocephalidae. *Molecular Phylogenetics and Evolution*, **62**, 319–328.

Fujisawa T, Barraclough TG (2013) Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. *Systematic Biology*, **62**, 707–724.

Geisen S, Laros I, Vizcaíno A, Bonkowski M, de Groot GA (2015) Not all are free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoa. *Molecular Ecology*, **24**, 4556–4569.

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2013) *Bayesian Data Analysis*, 3rd edn. Chapman Hall, London.

Guayasamin JM, Krynak T, Krynak K, Culebras J, Hutter CR (2015) Phenotypic plasticity raises questions for taxonomically important traits: a remarkable new Andean rainfrog (*Pristimantis*) with the ability to change skin texture. *Zoological Journal of the Linnean Society*, **173**, 913–928.

Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, **15**, 910–917.

Hamilton AJ, Basset Y, Benke KK *et al.* (2010) Quantifying uncertainty in estimation of tropical arthropod species richness. *The American Naturalist*, **176**, 90–95.

He K, Shinohara A, Jiang X-L, Campbell KL (2014) Multilocus phylogeny of talpine moles (Talpini, Talpidae, Eulipotyphla) and its implications for systematics. *Molecular Phylogenetics and Evolution*, **70**, 513–521.

© 2016 John Wiley & Sons Ltd

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings. Biological Sciences/The Royal Society*, **270**, 313–321.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, e312.

Jayaswal V, Wong TKF, Robinson J, Poladian L, Jermiin LS (2014) Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic Biology*, **63**, 726–742.

Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Kaehler BD, Yap VB, Zhang R, Huttley GA (2015) Genetic distance for a general non-stationary markov substitution process. *Systematic Biology*, **64**, 281–293.

Kainer D, Lanfear R (2015) The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution*, **32**, 1611–1627.

Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Molecular Ecology Resources*, **14**, 706–715.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, **29**, 1695–1701.

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, **21**, 1095–1109.

Leavitt DH, Starrett J, Westphal MF, Hedin M (2015) Multilocus sequence data reveal dozens of putative cryptic species in a radiation of endemic Californian mygalomorph spiders (Araneae, Mygalomorphae, Nemesiidae). *Molecular Phylogenetics and Evolution*, **91**, 56–67.

Leray M, Knowlton N (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 2076–2081.

Losos JB, Arnold SJ, Bejerano G *et al.* (2013) Evolutionary biology for the 21st century. *PLoS Biology*, **11**, e1001466.

Marshall DC (2010) Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Systematic Biology*, **59**, 108–117.

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean? *PLoS Biology*, **9**, e1001127.

Mora C, Rollo A, Tittensor DP (2013) Comment on ‘Can we name Earth’s species before they go extinct? *Science (New York, N.Y.)*, **341**, 237.

Nagy ZT, Sonet G, Glaw F, Vences M (2012) First large-scale DNA barcoding assessment of reptiles in the biodiversity hotspot of Madagascar, based on newly designed COI primers. *PLoS ONE*, **7**, e34506.

Nelson BJ, Andersen JJ, Brown JM (2015) Deflating trees: improving Bayesian branch-length estimates using informed priors. *Systematic Biology*, **64**, 441–447.

Nielsen R (2002) Mapping mutations on phylogenies. *Systematic Biology*, **51**, 729–739.

Niemiller ML, Graening GO, Fenolio DB *et al.* (2013) Doomed before they are described? The need for conservation assessments of cryptic species complexes using an amblyopsid cavefish (Amblyopsidae: Typhlichthys) as a case study. *Biodiversity and Conservation*, **22**, 1799–1820.

Nylander J, Ronquist F, Huelsenbeck J, Nieves-Aldrey J (2004) Bayesian phylogenetic analysis of combined data. *Systematic Biology*, **53**, 47–67.

Percy DM, Argus GW, Cronk QC *et al.* (2014) Understanding the spectacular failure of DNA barcoding in willows (Salix): does this result from a trans-specific selective sweep? *Molecular Ecology*, **23**, 4737–4756.

Pfenninger M, Nowak C, Kley C, Steinke D, Streit B (2007) Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic *Chironomus* (Diptera) species. *Molecular Ecology*, **16**, 1957–1968.

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931–1950.

Price TD (2010) The roles of time and ecology in the continental radiation of the Old World leaf warblers (*Phylloscopus* and *Seicercus*). *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 1749–1762.

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.

Rambaut A, Suchard MA, Xie D, Drummond AJ (2014) Tracer.

Rannala B, Zhu T, Yang Z (2012) Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular Biology and Evolution*, **29**, 325–335.

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS ONE*, **8**, e66213.

Régnier C, Achaz G, Lambert A, Cowie RH, Bouchet P, Fontaine B (2015) Mass extinction in poorly known taxa. *Proceedings of the National Academy of Sciences*, **112**, 201502350.

Rodrigue N, Philippe H, Lartillot N (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 4629–4634.

Rodrigues ASL, Gray CL, Crowter BJ *et al.* (2010) A global assessment of amphibian taxonomic effort and expertise. *BioScience*, **60**, 798–806.

Ronquist F, Teslenko M, van der Mark P *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539–542.

Saitoh T, Sugita N, Someya S *et al.* (2015) DNA barcoding reveals 24 distinct lineages as cryptic bird species candidates in and around the Japanese Archipelago. *Molecular Ecology Resources*, **15**, 177–186.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF (2012) What we know and don’t know about Earth’s missing biodiversity. *Trends in Ecology & Evolution*, **27**, 501–510.

Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature*, **435**, 17.

Sites JW, Marshall JC (2004) Operational criteria for delimiting species. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 199–227.

Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1825–1834.

Srivathsan A, Meier R (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics*, **28**, 190–194.

Sun Y, Cai Y, Liu L *et al.* (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research*, **37**, e76.

Swofford DL (2003) *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Tewksbury JJ, Anderson JGT, Bakker JD *et al.* (2014) Natural history's place in science and society. *BioScience*, **64**, 300–310.

Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1859–1868.

Wiens JJ (2007) Species delimitation: new approaches for discovering diversity. *Systematic Biology*, **56**, 875–878.

Wilson EO (2004) Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **359**, 739.

Yang Z, Rannala B (2014) Unguided species delimitation using DNA sequence data from multiple Loci. *Molecular Biology and Evolution*, **31**, 3125–3135.

Yang Z, Roberts D (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, **12**, 451–458.

Yoccoz NG (2012) The future of environmental DNA in ecology. *Molecular Ecology*, **21**, 2031–2038.

Yu DW, Ji Y, Emerson BC *et al.* (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhang C, Rannala B, Yang Z (2012) Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. *Systematic Biology*, **61**, 779–784.

Zwickl D (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence data sets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin.

Both authors designed the research and wrote the manuscript. A.J.B. analysed the data.

Data accessibility

All DNA alignments and Python scripts used in this study are archived in the Dryad digital repository (doi:10.5061/dryad.pg470).

© 2016 John Wiley & Sons Ltd