Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 457–465 doi: 10.1111/j.2041-210X.2011.00176.x

Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification

Rupert A. Collins[1]*, Laura M. Boykin[1], Robert H. Cruickshank[2] and Karen F. Armstrong[1]

[1] *Bio-Protection Research Centre, PO Box 84, Lincoln University 7647, Canterbury, New Zealand; and* [2] *Department of Ecology, Faculty of Agriculture and Life Sciences, Lincoln University 7647, Canterbury, New Zealand*

Summary

1. DNA barcoding studies use Kimura's two-parameter substitution model (K2P) as the *de facto* standard for constructing genetic distance matrices. Distances generated under this model then provide the basis for most downstream analyses, but uncertainty in model choice is rarely explored and could potentially affect how reliably DNA barcodes discriminate species.

2. Using information-theoretic approaches for a data set comprising 14 472 DNA barcodes from 14 published studies, we tested whether the K2P model was a good fit at the species level and whether applying a better fitting model biased error rates or changed overall identification success.

3. We report that the K2P was a poorly fitting model at the species level; it was never selected as the best model and very rarely selected as a credible alternative model. Despite the lack of support for the K2P model, differences in distance between best model and K2P model estimates were usually minimal, and importantly, identification success rates were largely unaffected by model choice even when interspecific threshold values were reassessed.

4. Although these conclusions may justify using the K2P model for specimen identification purposes, we found simpler metrics such as *p* distance performed equally well, perhaps obviating the requirement for model correction in DNA barcoding. Conversely, when incorporating genetic distance data into taxonomic studies, we advocate a more thorough examination of model uncertainty.

Key-words: Akaike information criterion, DNA barcoding, Kimura-two-parameter, model selection, pairwise distances, taxonomy

Introduction

In the context of phylogeny estimation, models play an important role in determining our interpretation of evolution. Relationships, branch lengths and rates over time are all approximated in the light of processes assumed by a model (Kelchner & Thomas 2007), and investigations using simulated and real data have shown that model selection can influence both support values and tree topologies (Cunningham *et al.* 1998; Buckley & Cunningham 2002; Lemmon & Moriarty 2004; Ripplinger & Sullivan 2008). A model selection procedure aims to identify a model, which can best represent mutational processes, while minimising the loss of predictive ability through overparameterisation (Sullivan & Joyce 2005).

In terms of choosing between models, advances in information theory have allowed for more effective discrimination between competing schemes (Posada & Buckley 2004). Implementation of information-theoretic approaches such as the Akaike Information Criterion (AIC) now allows for the assessment of model fit, as well as taking into account increases in variance by penalising overparameterisation and information loss (Posada & Buckley 2004; Bos & Posada 2005; Sullivan & Joyce 2005). We are now also able to assess relative support for a given set of substitution models using AIC weights (Posada & Buckley 2004; Posada 2008). This approach is particularly useful given that an alternative model may be an equally good estimator as the model with the lowest AIC value (Kelchner & Thomas 2007). These weights approximate probabilities for a given set of models, and evidence ratios between these weights offer a comparison of support for competing models (Anderson 2008).

For DNA barcoding, or similar specimen identification procedures using sequence data, the parameter of interest is usually nucleotide similarity. This measure of change between lineages provides the basis for data summary and

*Correspondence author. E-mail: rupertcollins@gmail.com
Correspondence site: http://www.respond2articles.com/MEE/

specimen identification (Hebert *et al.* 2003). Similarity is inferred though pairwise comparison between homologous sequences and can be expressed as a genetic distance: the number of substitutions per site in a given alignment. These distances are then used in the generation of identification success rates with, for example, nearest neighbour thresholds or neighbour-joining phylograms. Due to this reliance on distance metrics, a robust and effective estimate of these distances is a prerequisite for non-expert end-users of barcode data to have confidence in specimen identifications from public reference data bases, such as Bold (Ratnasingham & Hebert 2007).

In terms of generating genetic distances, sequence similarity can be derived directly from observed data as raw *p* distances, but unobserved substitutions at mutational hot spots such as third codon positions can lead to an underestimation of differences between lineages (Sullivan & Joyce 2005). Mathematical models used in phylogenetics correct for this saturation by applying a more realistic scenario of nucleotide substitution than observed from raw data and can vary considerably in complexity (Bos & Posada 2005). In DNA barcoding studies, Kimura's two-parameter model (Kimura 1980), hereafter referred to as the K2P model, is the *de facto* standard metric for computing these distances (Ward 2009). The K2P model provides a substitution framework with a free parameter for both transitions and transversions, accounting for the likely higher substitution rate of transitions in mitochondrial DNA (Kimura 1980; Wakeley 1996). Base frequencies are assumed to be equal under this model, but departures from this assumption are common in real data sets, and different nucleotide compositions may influence particular types of substitution rate (Tamura 1992; Galtier & Gouy 1995; Ward *et al.* 2005).

The use of the K2P model in DNA barcoding began with Hebert *et al.* (2003), who stated: 'For the species-level analysis, nucleotide-sequence divergences were calculated using the K2P model, the best metric when distances are low (Nei & Kumar 2000) as in this study' (p. 315). Hebert *et al.* were presumably referring to the following passage in Nei & Kumar (2000): 'Even the *p* distance becomes very similar to other distance measures when $p \leq 0{\cdot}1$. Therefore when one is studying closely related sequences, there is no need to use complex distance measures. In this case, it is better to use a simpler one, because it has smaller variance'' (p. 40–41; also see p. 112). This point made by Nei & Kumar is important because at a fundamental level, and despite the widespread use of the K2P model in DNA barcoding, it remains to be demonstrated whether model-corrected distances are justified over using the uncorrected *p* distances (i.e. can the raw data serve adequately for the purpose required?). Although it has been noted that barcode variation within species is generally low (Ward 2009; Hebert *et al.* 2010), it is not clear whether simple measures could systematically bias results by underestimating change (Sullivan & Joyce 2005). In terms of specimen identification, an underestimate of genetic distance may increase the number of false-negative 'lumping' errors, while overestimating change may increase false-positive 'splitting' errors (Meyer & Paulay 2005). This is linked to the principal of the barcoding gap, which relies on individuals within a species being more similar to one another than to the closest individual of another species (Meyer & Paulay 2005; Meier *et al.* 2008). It may be that when simple measures such as *p* distances are used, this gap is decreased, hindering identification success. For an effective specimen identification system, it is important, therefore, to fully understand how measures of inferred similarity (model-corrected distances) or observed similarity (uncorrected distances) could affect results.

Two recently published studies have investigated the application of substitution models in DNA barcoding, but offer fundamentally different conclusions. Fregin *et al.* (2011), based on their analysis of 120 cytochrome *b* sequences from 61 acrocephalid bird species, recommended 'only distances based on the optimal substitution model should be used'. In contrast, Srivathsan & Meier (2011) looked at 5283 published COI sequences from 200 genera and showed that 'the use of uncorrected distances yields higher or similar identification success rates' (compared to K2P correction). These contradictory findings suggest the question of model specification deserves further attention.

Given the availability of model selection software such as jModelTest (Guindon & Gascuel 2003; Posada 2008), it seems an appropriate time to re-examine how sensitive DNA barcode analyses are to alternative models and ask whether the indiscriminate use of the K2P model is really justified. Using an explicit test of DNA barcode data under justifiable model selection criteria, we specifically aim to address the following: (i) is the K2P a well-fitting model at the species level; (ii) how different are distances generated under a better model to those generated under the K2P model; (iii) can applying different models change identification success rates and estimations of the barcoding gap; (iv) does model correction in general perform better than using no model; and (v) how did Fregin *et al.* (2011) and Srivathsan & Meier (2011) reach such conflicting conclusions?

Materials and methods

DATA ACQUISITION

Fourteen data sets were obtained in Fasta format from project pages on Bold. These data sets comprised large studies of relatively well-known taxonomic groups including butterflies (Hajibabaei *et al.* 2006; Lukhtanov *et al.* 2009; Dincă *et al.* 2011), birds (Kerr *et al.* 2009a, b, 2007; Johnsen *et al.* 2010), fishes (Ward *et al.* 2005; Hubert *et al.* 2008; Rasmussen *et al.* 2009; Wong *et al.* 2009; Steinke *et al.* 2009a,b) and bats (Francis *et al.* 2010). Well known faunas were chosen to minimise discrepancies between the molecular data and taxonomy. Bold sequence identifiers (taxon names) were trimmed using regular expressions to include only GenBank accession number and taxonomic identification (species name). Alignment was carried out by Bold, followed by visual editing using translated amino acids in Mega4 (Tamura *et al.* 2007).

SPECIES-LEVEL MODEL SELECTION

To test whether the K2P is a well-fitting model at the species level, each data set was split into species using the Ape package (Paradis *et al.* 2004) for R (R Development Core Team 2010), with species delimited by their unique binomials. The individual species data were exported in Nexus format, and species with less than five individuals were excluded to represent a data set of at least an average intraspecific sample size (Ward *et al.* 2009). Using nested Unix shell scripts, the program jModelTest was run as a batch process for each species in each data set, producing a corresponding jModelTest output file. All 11 substitution schemes were tested (Posada 2008), along with base frequency and rate variation options (total 44 models). An invariant sites parameter was not included, as species comprising a single haplotype could not be optimised under this setting in jModelTest. The model frequencies and AIC weights for the best and K2P models were extracted from the jModelTest output files using shell commands.

DIFFERENCE BETWEEN K2P AND BEST MODEL

To test how different intraspecific K2P distances are from best-model distances, we first used batch processes in Paup* (Swofford 2003) to calculate pairwise comparisons under standard K2P distance settings (distance = K2P). Next, estimations for the best model were generated as maximum likelihood (ML) distances (distance = ml), with likelihood settings derived from jModelTest's Paup* block output. Shell scripting was used to manipulate corresponding likelihood settings from the jModelTest output into the Nexus file for each species, before initiating Paup* as a concatenated batch process. K2P distances were then subtracted from best-model estimates for each pairwise comparison. For this analysis using Paup*, the pairwise deletion option for missing data was used (missdist = ignore), and undefined distances were set to 'NA' (undefined = asterisk); all other settings were default. Except for K2P, abbreviated nomenclature of models follows Posada (2008); the K2P model is referred to as the K80 model by this author.

IDENTIFICATION SUCCESS

To test the influence of model selection on identification success rate, both intraspecific and interspecific values were required, so distances were generated from the undivided data sets, which also included the previously excluded species with less than five individuals. To illustrate the effects of different substitution schemes, we used a selection of standard 'off the shelf' models in Paup*, offering a variety of parameterisations from simple to complex: JC, F81, K2P, TrN, HKY, HKY + Γ and GTR + Γ. Gamma shape values were derived from jModelTest. We measured identification success rates using the 'best close match' criterion of Meier *et al.* (2006), but also see Ross *et al.* (2008) and Austerlitz *et al.* (2009) for additional comparisons including tree-based methods. For the 'best close match', a conspecific nearest neighbour ($k = 1$) within a threshold per cent value was recorded as a 'correct' identification; a non-conspecific nearest neighbour within the threshold was an 'incorrect' identification; more than one equally close species (including the correct species) within the threshold was 'ambiguous'; and no match within the threshold was reported as a 'no identification'. The threshold was initially set at the 1% value, as used by the Bold identification engine (Ratnasingham & Hebert 2007), but because threshold values are likely to be contingent upon the models they are generated under, we also optimised new thresholds for each model and data set. This optimisation procedure minimises false-positive (no matches within x of query) and false-negative (more than one species match within x of query) errors for a range of threshold values (0·2–5·0% in 0·2% increments). To assess the effect of model selection on magnitude of the barcoding gap, both maximum intraspecific and minimum interspecific distances were calculated (Meier *et al.* 2008), with the barcoding gap expressed as minimum interspecific distance divided by maximum intraspecific distance; singletons were not considered for intraspecific variation, and intraspecific values of zero were replaced with a value of 0.001536098 (corresponding to a single nucleotide change over 651 bp). Analyses were carried out in R using the DNA barcoding package Spider (Brown *et al.* in press; Paradis *et al.* 2004).

Results

SPECIES-LEVEL MODEL SELECTION

From the 14 data sets, we extracted 1446 species with ≥5 individuals, resulting in 14 472 DNA barcodes; the mean number of barcodes per species was 10 (Table 1). For the individual species tested by jModelTest ($n = 1446$), the model most frequently selected as best (zero AIC Δ value) was the HKY ($n = 579$), followed by F81 ($n = 312$) and TrN ($n = 264$). Overall, 20 models were selected by the AIC, and the K2P model was never selected as best model (Fig. 1). Models with a gamma shape parameter were selected on 7·95% of occasions. The AIC weight (w) of the best model ranged between 0·08 and 0·64 (mean $w = 0{\cdot}21$). As an alternative model, the AIC weight for the K2P was no greater than 0·019 (mean $w = 0{\cdot}000134$). The mean evidence ratio (E) for the best model vs. K2P model weight was $E = 1{\cdot}9 \times 10^{33}$ (range = 10·0 to $2{\cdot}8{\times}10^{36}$). A representation of the relative model weights is shown in Fig. 2.

DIFFERENCE BETWEEN K2P AND BEST MODEL

In calculating distances within species, a total of 191 402 pairwise comparisons were made. When the K2P distance was subtracted from the best-model distance, 31·2% of the total comprised zero change, and 39·6% were greater than zero and less than 0·1%; 8·12% showed a difference greater than 1%, and 15·6% were negative (K2P distance larger than best-model distance). Average differences were 0·64% (mean) and 0·00012% (median); range was −0·068% to 136·7%. A density plot illustrating the differences between the K2P model and best-model distances for each data set is presented in Fig. 3.

IDENTIFICATION SUCCESS

A total of 21 514 DNA barcodes were used to measure identification success. Under the 1% Bold threshold, differences in identification success for all models varied by no greater than 0·04%; the two models with gamma shape parameters (HKY + Γ and GTR + Γ) had the lowest correct identification rates of 91·81% (Table 2). Optimised threshold values varied according to data set (range 0·2–1·2%), but not by model,

Table 1. Summary and citations for data sets used in this study, with numbers of individuals per species remaining after filtering for ≥5 individuals

Data set citation	Taxon	No. spp. ≥5 indiv.	No. indiv.	Seqs. per sp.
Dincă *et al.* (2011)	Romanian butterflies	144	1273	8·8
Francis *et al.* (2010)	Southeast Asian bats	88	1736	19·7
Hajibabaei *et al.* (2006)	Tropical Lepidoptera	65	723	11·1
Hubert *et al.* (2008)	Canadian freshwater fishes	132	1203	9·1
Johnsen *et al.* (2010)	Scandinavian birds	31	173	5·6
Kerr *et al.* (2007)	North American birds	230	2386	10·4
Kerr *et al.* (2009b)	Argentinian birds	106	687	6·5
Kerr *et al.* (2009a)	Palearctic birds	148	1063	7·2
Lukhtanov *et al.* (2009)	Central Asian butterflies	34	192	5·6
Rasmussen *et al.* (2009)	North American salmonids	8	934	116·8
Steinke *et al.* (2009b)	Ornamental marine fishes	162	1169	7·2
Steinke *et al.* (2009a)	Pacific Canadian fishes	107	1029	9·6
Ward *et al.* (2005)	Australian marine fishes	148	921	6·2
Wong *et al.* (2009)	Commercial sharks	43	983	22·9
Total		1446	14 472	10·0 (avg.)

avg., mean; indiv., individuals; spp./sp., species; seqs., sequences.

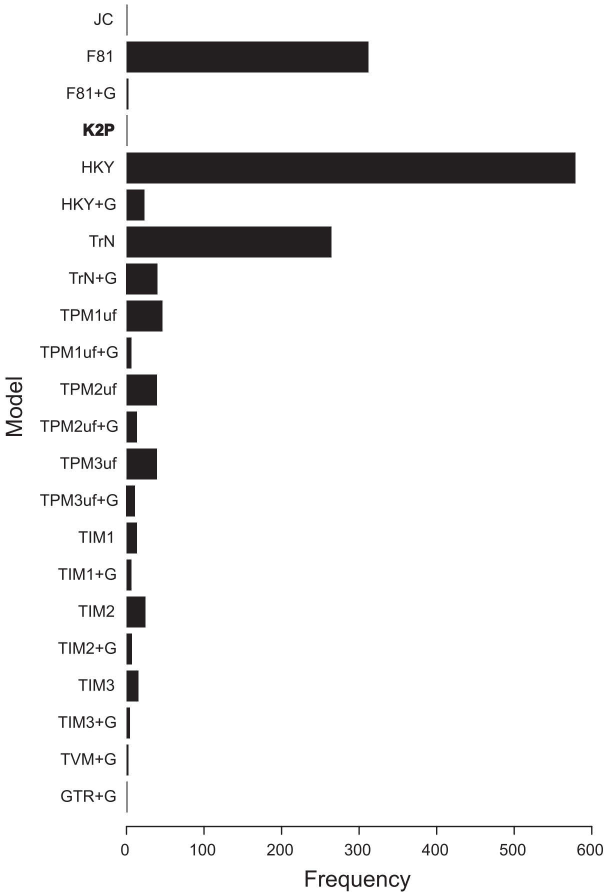

Fig. 1. Frequency of per-species models selected by jModelTest under the Akaike Information Criterion. The Kimura two-parameter (K2P) model is highlighted in bold (frequency = 0). Except for the K2P model, abbreviated nomenclature of models follows Posada (2008). Summary of the properties of these models can also be found in Posada (2008).

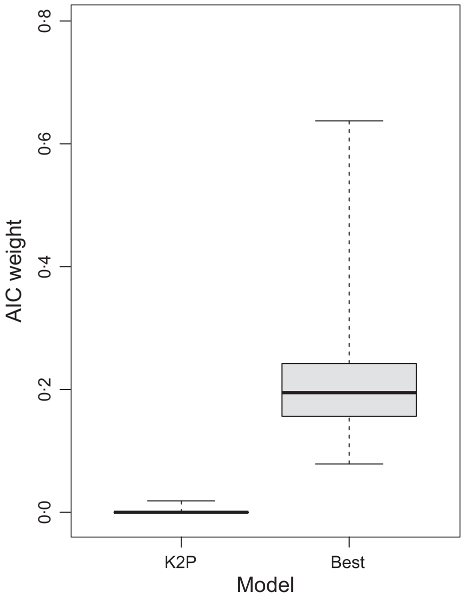

Fig. 2. Distribution of Akaike Information Criterion (AIC) weights for best and Kimura two-parameter (K2P) models. Whiskers extend to full range of data; boxes represent quartiles; black lines show median values.

except for the GTR + Γ threshold for Dincă *et al.* (2011) (Table 3). Identification success varied by up to 0·28% under optimised thresholds, with *p* distance having the highest value and the GTR + Γ model with the lowest (Table 2). Ambiguous identification tended to decrease with model complexity, along with an increase in incorrect and unidentifiable individuals (Table 2). In terms of the distribution of the barcoding gap under different models, for schemes without a gamma parameter, median values remained generally similar with smallest interspecific distances between 12·33× and 13·17× maximum intraspecific distances; the models with a gamma parameter had higher median (16·02× to 16·59×) and also higher maximum values (Fig. 4). No barcode gap was found for between 8·72% (*p* distance) and 8·50% (HKY + Γ) of individuals. Overall, the effect of model selection on all distances (both intraspecific and interspecific) is represented in Fig. 5.

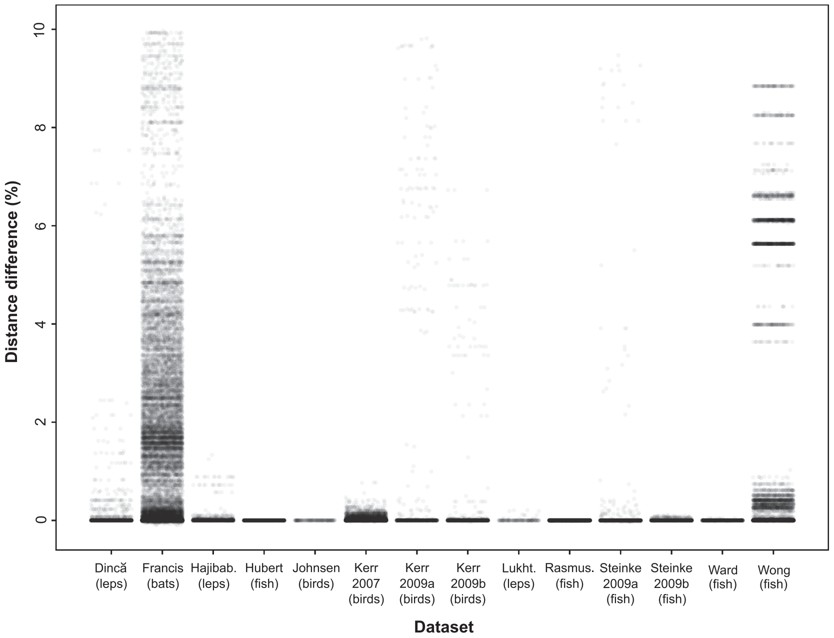

Fig. 3. Jittered density plot showing per cent difference between best Akaike Information Criterion model and Kimura two-parameter model distances for each of 14 data sets. The *y*-axis limit was set to 10% to assist presentation. The plot was created in R using ggplot2 (Wickham 2009).

Table 2. Identification success rates using the 'best close match' criterion of Meier *et al.* (2006) across a selection of models for $n = 21\,514$ individuals

Dist. measure	Ambig. %	Correct %	Incorrect %	No ident. %
p distance	2·35 (2·31)	91·84 (90·81)	0·91 (0·75)	4·90 (6·13)
JC	2·34 (2·31)	91·84 (90·77)	0·91 (0·75)	4·91 (6·17)
F81	2·33 (2·31)	91·85 (90·77)	0·92 (0·75)	4·91 (6·17)
K2P	2·34 (2·31)	91·84 (90·76)	0·91 (0·75)	4·91 (6·18)
TrN	2·30 (2·29)	91·85 (90·76)	0·94 (0·78)	4·91 (6·18)
HKY	2·32 (2·31)	91·85 (90·76)	0·92 (0·76)	4·91 (6·18)
HKY + Γ	2·31 (2·29)	91·81 (90·75)	0·93 (0·77)	4·95 (6·20)
GTR + Γ	2·30 (2·29)	91·81 (90·53)	0·94 (0·77)	4·95 (6·41)

ambig., ambiguous; dist., distance; ident., identification.
Threshold values were determined from Bold's 1% (open values) or were optimised according to error minimisation (values in parentheses); refer to Table 3 for optimised threshold values.

Discussion

Although the species-level analyses show that the K2P was never selected as the best model, picking a model with the lowest AIC value may ignore credible alternative models that are also good approximators (Alfaro & Huelsenbeck 2006; Kelchner & Thomas 2007; Anderson 2008). Therefore, it could have been possible that the K2P model was a reasonable alternative model, but when we considered AIC weights and evidence ratios between models to assess support, we found that the K2P was without exception a poorly approximating model at the species level; the lowest evidence ratio was 10 : 1 against the K2P. It is likely that the assumption of equal base frequencies led to the rejection of the K2P model in most cases, thus favouring the otherwise similar F81 and HKY models with unequal frequencies (Fig. 1). In general, substitution schemes tended to be relatively simple at the species level, with either equal rates (F81), or separate transition/transversion rates (HKY) selected. In terms of the suitability of the AIC for answering these questions, other model selection criteria such as likelihood ratio tests or the Bayesian Information Criterion (BIC) could have been considered here, but these measures are considered to be based on weak philosophical foundations, and the latter has a tendency to give high weights to poorly fitting models (Anderson 2008; Posada & Buckley 2004).

Overall, there was little difference between intraspecific distances optimised under best model or K2P model parameters. The majority (86·3%) of the difference was either zero or minor less than (± 0·1%). The Francis *et al.* (2010) bat data set had the largest differences (Fig. 3). When this data set was excluded, 93·9% of differences in distance were less than ± 0·1%. At least a third of the bat species analysed in this study had multiple divergences of over 2% K2P distance (Francis *et al.* 2010). This study group reflects a high proportion of underestimated diversity, and this discrepancy between current taxonomy and DNA data indicates that the species-level units

Table 3. Optimised distance thresholds for each data set under a selection of models

Data set	*p* dist. %	JC %	F81 %	K2P %	TrN %	HKY %	HKY + Γ%	GTR + Γ%
Dincă *et al.* (2011)	0·4	0·4	0·4	0·4	0·4	0·4	0·4	**0·2**
Francis *et al.* (2010)	1·2	1·2	1·2	1·2	1·2	1·2	1·2	1·2
Hajibabaei *et al.* (2006)	0·2	0·2	0·2	0·2	0·2	0·2	0·2	0·2
Hubert *et al.* (2008)	0·4	0·4	0·4	0·4	0·4	0·4	0·4	0·4
Johnsen *et al.* (2010)	0·6	0·6	0·6	0·6	0·6	0·6	0·6	0·6
Kerr *et al.* (2007)	0·6	0·6	0·6	0·6	0·6	0·6	0·6	0·6
Kerr *et al.* (2009b)	0·6	0·6	0·6	0·6	0·6	0·6	0·6	0·6
Kerr *et al.* (2009a)	0·8	0·8	0·8	0·8	0·8	0·8	0·8	0·8
Lukhtanov *et al.* (2009)	0·4	0·4	0·4	0·4	0·4	0·4	0·4	0·4
Rasmussen *et al.* (2009)	0·4	0·4	0·4	0·4	0·4	0·4	0·4	0·4
Steinke *et al.* (2009b)	1·2	1·2	1·2	1·2	1·2	1·2	1·2	1·2
Steinke *et al.* (2009b)	0·8	0·8	0·8	0·8	0·8	0·8	0·8	0·8
Ward *et al.* (2005)	0·6	0·6	0·6	0·6	0·6	0·6	0·6	0·6
Wong *et al.* (2009)	0·2	0·2	0·2	0·2	0·2	0·2	0·2	0·2
Mean	0·6	0·6	0·6	0·6	0·6	0·6	0·6	0·6

dist., distance.
Thresholds were optimised for a range of values (0·2–5·0%) under a procedure that minimises false-positive and false-negative error rates (Meyer & Paulay 2005). The threshold varying by model is highlighted in bold.

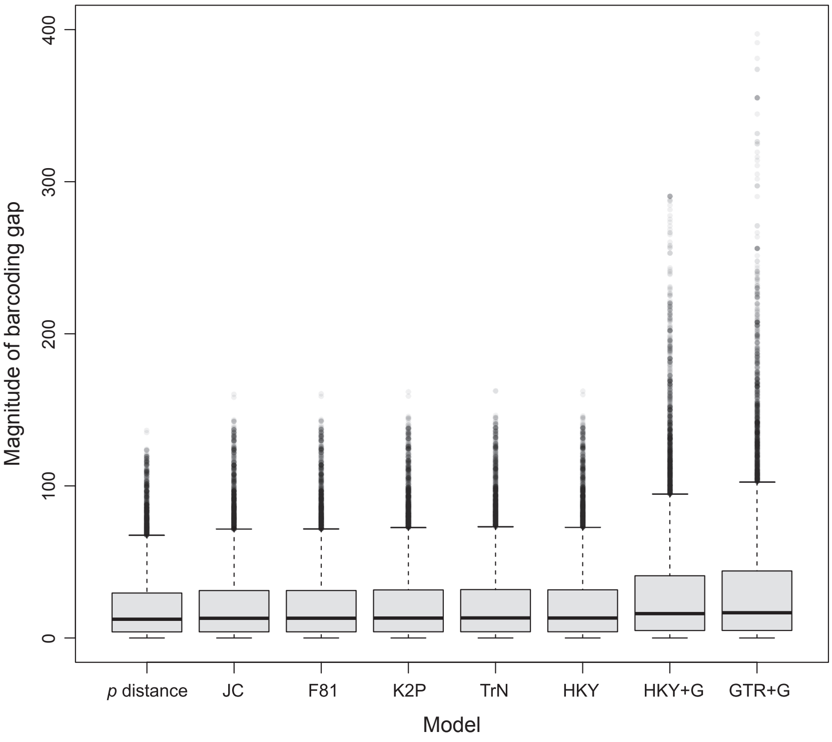

Fig. 4. Distribution of variation in the magnitude of the barcoding gap according to model for $n = 20\ 643$ individuals. The barcoding gap is expressed as interspecific divergence as a multiple of intraspecific divergence and was calculated by dividing each minimum interspecific value by the corresponding maximum intraspecific value. Singletons were not considered for intraspecific variation. Whiskers extend to 1·5× interquartile range, black lines show median values, and points represent outlying data.

from this study were probably not comparable with the other data sets we used. Conversely for the other data sets, species-level diversity may have been artificially reduced, as it was not clear from the methods sections of the publications cited (Table 1) whether code numbers or designations such as cf. were appended to species names during the morphological identification process or were post hoc assignments based on barcode divergences. As these would be considered different species in our analysis, an indication of how this may have affected results is necessary; of all 14 472 individuals, only 7% failed to satisfy a regular expression conforming to a correctly constructed binomial. However, regardless to the degree of

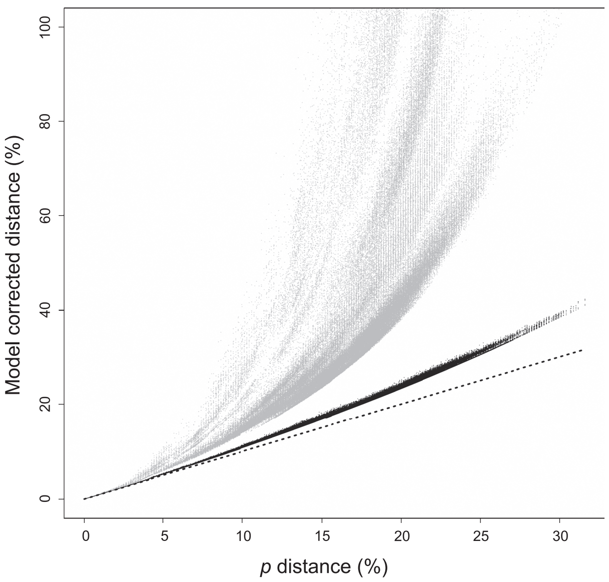

Fig. 5. Scatter plot of a representative random sample (n = 100 000) of intraspecific and interspecific distances as a function of increasing p distance. Models with a gamma shape parameter (HKY + Γ and GTR + Γ) are shown by grey points, p distance by the dotted line, and distances derived under the JC, F81, K2P, TrN and HKY models by black points.

match between barcodes and taxonomic names, optimising intraspecific distances under a more statistically justifiable model than the K2P did not substantially change them in the majority of cases (Fig. 3).

Although most changes in distance we observed among models were small, when strict thresholds are used as identification criteria (e.g. by Bold), in theory even relatively minor differences in distance could change the assignment of an unknown specimen. However, we found only a negligible decrease in identification success rate when more complex models were employed (Table 2), and although the Bold threshold value of 1% was generated from data under the K2P model, when we provided revised thresholds optimised under different models, the identification success rates continued to remain robust to model selection. This is likely due to the observation that distance values pertinent to specimen identification (i.e., largest intraspecific and smallest interspecific) were generally low enough not to be significantly affected by model correction (Figs 3 and 5). Overall, genetic distances generated under models without a gamma shape parameter scarcely deviated from estimations made by the K2P model at p distances of less than around 10%, although when a gamma shape parameter was introduced, distances had an increased proportion of correction at this level (Fig. 5). As an indication of how correction may influence a typical data set, (Ward 2009) reported mean interspecific K2P distances of 5·5% for congeneric bird species. Our results for a wider variety of taxa (Table 1) report a mean K2P distance of 6·9% for all nearest non-conspecific values and a mean maximum intraspecific value of 1·0%.

Regarding the discrepancy between conclusions presented by Fregin *et al.* (2011) and Srivathsan & Meier (2011), we find our results entirely congruent with those of Srivathsan & Meier in that substitution models have little effect on specimen identification. Our study found a slight degree of systematic bias, with more complex models having marginally lower ambiguous identification error rates (interspecific distances underestimated), but this was countered by a larger proportion of incorrect and unidentifiable specimens (intraspecific distances overestimated). When taking this bias into account, our results demonstrate that for identification purposes, p distances performed as well or marginally better (optimised thresholds) than more complex models due to the higher false-positive error rates of the latter (Table 2). Similarly, increasing model complexity produced an increase in the magnitude of the barcoding gap (Fig. 4), but this was not translated into an increase in the number of individuals for which a gap was present. We also report that increasing parameterisation further, with the inclusion of an invariant sites model (GTR + I + Γ), resulted in another increase in the magnitude of the barcoding gap and again generated a reduction in identification success (data not shown). Given the assertion of (Nei & Kumar (2000) that 'when one is studying closely related sequences, there is no need to use complex distance measures', we must ask again why models are used in DNA barcoding? Thus, it appears that observed similarity is an acceptable way to identify specimens, unless a user is particularly interested in minimising one error rate over another for a specific application.

Despite their call for better fitting models to be used in studies using genetic distances, a reanalysis of the data presented by Fregin *et al.* (2011) showed no differences according to model in either identification success rate or proportion of specimens lacking a barcode gap (TrN + Γ and p distances; their Supplementary Table 1). It is not clear to whom their advice is aimed, because their conclusions appear to blur the distinctions between DNA barcoding and DNA taxonomy, i.e. assigning unknowns to a pre-identified reference library vs. species delimitation and description (Vogler & Monaghan 2007; Padial *et al.* 2010). Although the same data can be used for both purposes, the objectives remain fundamentally different and each require distinct experimental procedures (Padial *et al.* 2010). There appears to be no standard practice regarding model correction for taxonomic questions, and different substitution frameworks are often employed among studies, frequently without a model selection procedure or justification (for references see Fregin *et al.*). When making taxonomic decisions, understanding evolutionary process is arguably more important than for DNA barcoding and may be especially critical in circumstances such as supporting a new species status for a divergent taxon. When framed in this context, we must recommend greater emphasis on model choice and therefore can agree here with the conclusions of Fregin *et al.*

We conclude by stating that model selection remains an important consideration in many disciplines, and DNA barcoding should be no different. Practitioners of DNA barcoding may feel reassured that identification rates were not significantly affected by model selection, but should also be aware that a model selection process can increasingly influence conclusions when larger distances are being considered. In taxonomic studies where these conclusions are important, statistical uncertainty in distance estimation could certainly be better explored with information-theoretic techniques such as multimodel inference and model averaging.

Acknowledgements

This work was supported by a Ministry of Agriculture and Forestry Biosecurity New Zealand scholarship for RAC, and funding from the New Zealand Tertiary Education Commission to the Bio-Protection Research Centre for LB. The authors thank David Posada for assistance with jModelTest, Samuel D. J. Brown for valuable help with R, and three anonymous reviewers who suggested improvements to the manuscript.

References

Alfaro, M.E. & Huelsenbeck, J.P. (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Systematic Biology*, **55**, 89–96.

Anderson, D.R. (2008) *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer, New York.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M. & Laredo, C. (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, **10**(Suppl. 14), S10.

Bos, D.H. & Posada, D. (2005) Using models of nucleotide evolution to build phylogenetic trees. *Developmental and Comparative Immunology*, **29**, 211–227.

Brown, S.D.J., Collins, R.A., Boyer, S., Malumbres-Olarte, J., Lefort, M-C., Vink, C. & Cruickshank, R.H. (in press) Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, doi: 10.1111/j.1755-0998.2011.03108.x.

Buckley, T.R. & Cunningham, C.W. (2002) The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution*, **19**, 394–405.

Cunningham, C.W., Zhu, H. & Hillis, D.M. (1998) Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, **52**, 978–987.

Dincă, V., Zakharov, E.V., Hebert, P.D.N. & Vila, R. (2011) Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B, Biological Sciences*, **278**, 347–355.

Francis, C.M., Borisenko, A.V., Ivanova, N.V., Eger, J.L., Lim, B.K., Guillén-Servent, A., Kruskop, S.V., Mackie, I. & Hebert, P.D.N. (2010) The role of DNA barcodes in understanding and conservation of mammal diversity in Southeast Asia. *PLoS ONE*, **5**, e12575.

Fregin, S., Haase, M., Olsson, U. & Alström, P. (2011) Pitfalls in comparisons of genetic distances: a case study of the avian family Acrocephalidae. *Molecular Phylogenetics and Evolution*, doi: 10.1016/j.ympev.2011.10.003.

Galtier, N. & Gouy, M. (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 11317–11321.

Guindon, S. & Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. & Hebert, P.D.N. (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 968–971.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & deWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B, Biological Sciences*, **270**, 313–321.

Hebert, P.D.N., deWaard, J.R. & Landry, J.F. (2010) DNA barcodes for 1/1000 of the animal kingdom. *Biology Letters*, **6**, 359–362.

Hubert, N., Hanner, R., Holm, E., Mandrak, N.E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J. & Bernatchez, L. (2008) Identifying Canadian freshwater fishes through DNA barcodes. *PLoS ONE*, **3**, e2490.

Johnsen, A., Rindal, E., Ericson, P.G.P., Zuccon, D., Kerr, K.C.R., Stoeckle, M.Y. & Lifjeld, J.T. (2010) DNA barcoding of Scandinavian birds reveals divergent lineages in trans-Atlantic species. *Journal of Ornithology*, **151**, 565–578.

Kelchner, S.A. & Thomas, M.A. (2007) Model use in phylogenetics: nine key questions. *Trends in Ecology & Evolution*, **22**, 87–94.

Kerr, K.C.R., Birks, S.M., Kalyakin, M.V., Red'kin, Y.A., Koblik, E.A. & Hebert, P.D.N. (2009a) Filling the gap – COI barcode resolution in eastern Palearctic birds. *Frontiers in Zoology*, **6**, 1–13.

Kerr, K.C.R., Lijtmaer, D.A., Barreira, A.S., Hebert, P.D.N. & Tubaro, P.L. (2009b) Probing evolutionary patterns in neotropical birds through DNA barcodes. *PLoS ONE*, **4**, e4379.

Kerr, K.C.R., Stoeckle, M.Y., Dove, C.J., Weigt, L.A., Francis, C.M. & Hebert, P.D.N. (2007) Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes*, **7**, 535–543.

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.

Lemmon, A.R. & Moriarty, E.C. (2004) The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology*, **53**, 265–277.

Lukhtanov, V.A., Sourakov, A., Zakharov, E.V. & Hebert, P.D.N. (2009) DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources*, **9**, 1302–1310.

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

Meier, R., Zhang, G. & Ali, F. (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the 'barcoding gap' and leads to misidentification. *Systematic Biology*, **57**, 809–813.

Meyer, C.P. & Paulay, G. (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, 2229–2238.

Nei, M. & Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Padial, J.M., Miralles, A., De la Riva, I. & Vences, M. (2010) The integrative future of taxonomy. *Frontiers in Zoology*, **7**, 1–14.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Posada, D. (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

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. *Systematic Biology*, **53**, 793–808.

R Development Core Team (2010) R: A language and environment for statistical computing. R Development Core Team.

Rasmussen, R.S., Morrissey, M.T. & Hebert, P.D.N. (2009) DNA barcoding of commercially important salmon and trout species *Oncorhynchus* and *Salmo* from North America. *Journal of Agricultural and Food Chemistry*, **57**, 8379–8385.

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Ripplinger, J. & Sullivan, J. (2008) Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, **57**, 76–85.

Ross, H.A., Murugan, S. & Li, W.L.S. (2008) Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, **57**, 216–230.

Srivathsan, A. & Meier, R. (2011) On the inappropriate use of Kimura-2-parameter K2P divergences in the DNA-barcoding literature. *Cladistics*, doi: 10.1111/j.1096-0031.2011.00370.x

Steinke, D., Zemlak, T.S., Boutillier, J.A. & Hebert, P.D.N. (2009a) DNA barcoding of Pacific Canada's fishes. *Marine Biology*, **156**, 2641–2647.

Steinke, D., Zemlak, T.S. & Hebert, P.D.N. (2009b) Barcoding Nemo: DNA-based identifications for the ornamental fish trade. *PLoS ONE*, **4**, e3600.

Sullivan, J. & Joyce, P. (2005) Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 445–466.

Swofford, D. (2003) PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.

Tamura, K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. *Molecular Biology and Evolution*, **9**, 678–687.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Vogler, A.P. & Monaghan, M.T. (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.

Wakeley, J. (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution*, **11**, 158–162.

Ward, R.D. (2009) DNA barcode divergence among species and genera of birds and fishes. *Molecular Ecology Resources*, **9**, 1077–1085.

Ward, R.D., Hanner, R. & Hebert, P.D.N. (2009) The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, **74**, 329–356.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. & Hebert, P.D.N. (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B, Biological Sciences*, **360**, 1847–1857.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

Wong, E.H.K., Shivji, M.S. & Hanner, R.H. (2009) Identifying sharks with DNA barcodes: assessing the utility of a nucleotide diagnostic approach. *Molecular Ecology Resources*, **9**(Suppl. 1), 243–256.

Received 2 September 2011; accepted 15 November 2011
Handling Editor: Robert Freckleton