MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2017) **17**, 1168–1182 doi: 10.1111/1755-0998.12658

Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: *Phrynosoma*)

CHRISTOPHER BLAIR*† and ROBERT W. BRYSON JR.‡§

**Department of Biological Sciences, New York City College of Technology, The City University of New York, 300 Jay Street, Brooklyn, NY 11201, USA, †Biology PhD Program, CUNY Graduate Center, 365 5th Avenue, New York, NY 10016, USA, ‡Department of Biology and Burke Museum of Natural History and Culture, University of Washington, 4331 Memorial Way Northeast, Seattle, WA 98195, USA, §Moore Laboratory of Zoology, Occidental College, Los Angeles, CA 90041, USA*

Abstract

Biodiversity reduction and loss continues to progress at an alarming rate, and thus, there is widespread interest in utilizing rapid and efficient methods for quantifying and delimiting taxonomic diversity. Single-locus species delimitation methods have become popular, in part due to the adoption of the DNA barcoding paradigm. These techniques can be broadly classified into tree-based and distance-based methods depending on whether species are delimited based on a constructed genealogy. Although the relative performance of these methods has been tested repeatedly with simulations, additional studies are needed to assess congruence with empirical data. We compiled a large data set of mitochondrial ND4 sequences from horned lizards (*Phrynosoma*) to elucidate congruence using four tree-based (single-threshold GMYC, multiple-threshold GMYC, bPTP, mPTP) and one distance-based (ABGD) species delimitation models. We were particularly interested in cases with highly uneven sampling and/or large differences in intraspecific diversity. Results showed a high degree of discordance among methods, with multiple-threshold GMYC and bPTP suggesting an unrealistically high number of species (29 and 26 species within the *P. douglasii* complex alone). The single-threshold GMYC model was the most conservative, likely a result of difficulty in locating the inflection point in the genealogies. mPTP and ABGD appeared to be the most stable across sampling regimes and suggested the presence of additional cryptic species that warrant further investigation. These results suggest that the mPTP model may be preferable in empirical data sets with highly uneven sampling or large differences in effective population sizes of species.

Keywords: ABGD, GMYC, ND4, Phrynosoma, PTP, speciation

Received 16 October 2016; revision received 25 January 2017; accepted 26 January 2017

Introduction

Proper identification and delimitation of species is of utmost importance for most fields of biology (de Queiroz 2007). Many research programmes address fundamental questions through a comparative framework, necessitating the use of both a robust phylogeny and accurate species assignments for hypothesis testing. Molecular-based species delimitation methods can generally be classified into single- or multilocus, and discovery and validation-based techniques (Carstens *et al.* 2013). DNA barcoding threshold methods (Hebert *et al.* 2003, 2004; Hebert & Gregory 2005; Edgar 2010; Puillandre *et al.* 2012a) comprise one common example of a single-locus technique, where threshold or cut-off values are used to differentiate inter- from intraspecific divergences. The refined single linkage (RESL) method, for example, is a popular clustering algorithm implemented within the Barcode of Life Data Systems (Ratnasingham & Hebert 2013) to delineate operational taxonomic units (OTUs) based on animal COI barcode data. Although these threshold-type methods continue to be a quick and effective way to document and describe diversity, they do not take into account tree structure and often rely on arbitrarily defined thresholds (e.g. 2%–3% pairwise sequence divergence) to delimit species (Blaxter 2004; Hebert & Gregory 2005; Hamilton *et al.* 2011).

More recently, coalescent-based methods of species delimitation have become common, in part due to the continual ease in which researchers can generate vast quantities of molecular data (Leaché & Fujita 2010; Fujita

Correspondence: Dr. Christopher Blair, Fax: 718-260-5342; E-mails: cblair@citytech.cuny.edu; cblair@gc.cuny.edu

et al. 2012). Bayesian multilocus coalescent methods, for example, can explicitly account for gene tree/species tree incongruence when delimiting species and estimating a species tree (Jones 2014; Yang & Rannala 2014). Although an attractive alternative to threshold-type methods, the utility of many Bayesian multilocus coalescent methods for large data sets remains uncertain due to the relatively large computational demand of the algorithms (Yang & Rannala 2010, 2014; Satler *et al.* 2013; Leaché *et al.* 2014).

Single-locus, coalescent-based methods like the general mixed Yule coalescent model (GMYC; Pons *et al.* 2006; Fujisawa & Barraclough 2013) have become a popular tree-based species delimitation technique often applied to barcoding data (e.g. animal mitochondrial DNA). The GMYC model uses maximum likelihood and an ultrametric gene tree to model the transition between inter- and intraspecific branching patterns. Branching patterns older than the inferred threshold represent speciation events (Yule process), whereas younger branching indicates neutral coalescence within species. GMYC has been used in numerous empirical studies (e.g. Monaghan *et al.* 2009; e.g. Barraclough *et al.* 2009; Hamilton *et al.* 2011; Gebiola *et al.* 2012; Esselstyn *et al.* 2012; Blair *et al.* 2015), and recent simulations and empirical data suggest that the method is fairly robust to different assumptions (Esselstyn *et al.* 2012; Reid & Carstens 2012; Fujisawa & Barraclough 2013; Talavera *et al.* 2013; Tang *et al.* 2014). The Poisson tree processes (PTP/bPTP) model is similar in that it seeks to model the transition in branch lengths between vs. within species (Zhang *et al.* 2013). However, PTP estimates branching processes using the expected number of substitutions (vs. time in GMYC) and thus utilizes a nonultrametric phylogenetic tree as input. One limitation of the original PTP model is that it assumes only two independent distributions to model branch lengths (one exponential distribution for speciation and one exponential distribution for coalescence). This generally ignores the stochastic variation among species due to different population sizes and demographic histories. Conversely, the recently developed multirate Poisson tree processes model (mPTP) fits multiple independent exponential distributions to each delimited species to explicitly account for differences in sampling intensity and/or population history (Kapli *et al.* 2016). Although the mPTP model may potentially lead to more accurate delimitations vs. other single-locus methods, testing and comparison using empirical data characterized by highly heterogeneous sampling intensity and/or large differences in genetic diversity among species is lacking.

Horned lizards (*Phrynosoma*) are a genus of phrynosomatid lizards consisting of 17–21 species distributed from Canada to Guatemala (Leaché & Linkem 2015; Montanucci 2015). The unique morphology and behaviour of these lizards, including ocular blood squirting (Sherbrooke 2003), have made them the target of numerous systematic investigations. Early studies based on mitochondrial DNA (mtDNA) sequences yielded conflicting phylogenetic relationships with both nuclear and morphological data (e.g. Reeder & Montanucci 2001; Hodges & Zamudio 2004; Leaché & McGuire 2006), presumably because of mtDNA introgression (Leaché & McGuire 2006; Mulcahy *et al.* 2006). More recently, next-generation sequencing has been used to estimate a robust phylogeny for the genus, with results suggesting that cladogenesis initiated in the Miocene around 22 million years ago (Leaché & Linkem 2015). Phylogeographic studies have also been conducted on several species including *P. douglasii* (Zamudio *et al.* 1997), *P. mcallii* (Mulcahy *et al.* 2006), *P. platyrhinos* (Jezkova *et al.* 2016), *P. coronatum* (Leaché *et al.* 2009) and *P. orbiculare* (Bryson *et al.* 2012). Many of these studies have indicated the presence of undocumented cryptic diversity, but whether any of this diversity may warrant species status has not been evaluated in detail for most groups (but see Leaché *et al.* 2009). These studies have also revealed substantially different levels of intraspecific diversity, ranging from deep lineages within *P. orbiculare* to low levels of genetic diversity in *P. mcallii*. More recent taxonomic work on the genus has suggested that levels of diversity may be underestimated at a clade-wide level (Nieto-Montes de Oca *et al.* 2014; Montanucci 2015).

Given the need for additional empirical studies to compare and contrast single-locus 'discovery-based' species delimitation methods, particularly in cases with large differences in sampling intensity and levels of intraspecific diversity, in this study, we assess congruence among four tree-based and one distance-based methods of delimiting horned lizard species. We particularly focus on the utility of one method (mPTP) with the ability to accommodate highly heterogeneous data sets comprised of species with dramatically different levels of molecular diversity. Based on our results, we provide further quantitative evidence for undescribed species in the genus.

Materials and methods

Data collection

We obtained from GenBank a total of 368 orthologs of the mitochondrial ND4 gene from multiple representatives of *Phrynosoma*. Although many molecular phylogenetic studies have been performed on the genus, there have been relatively few phylogeographic investigations of species using the same marker. We extracted from GenBank multiple sequences from *P. mcallii* (Mulcahy *et al.* 2006), *P. platyrhinos* (Jezkova *et al.* 2016), *P.*

orbiculare (Bryson *et al.* 2012) and the *P. douglasii* complex (Zamudio *et al.* 1997). We included singletons of several other species, including *P. cornutum*, *P. coronatum*, *P. solare*, *P. asio* and *P. taurus*. Although the COI gene is the generally accepted standard for most animal barcoding studies (Hebert *et al.* 2003; Ratnasingham & Hebert 2013), ND4 likely serves as a good proxy due to the linked inheritance of mtDNA. Full sampling information can be found in Table S1 (Supporting information). All gene trees and alignments used in this study can be found on Dryad (doi:10.5061/dryad.r7989).

Phylogenetic analysis

The primary goal of our study was to test multiple methods of single-locus species delimitation of horned lizards based on ND4 GenBank data. We were particularly interested in testing for similarities and differences in those methods requiring an ultrametric tree (GMYC-type methods) vs. those that do not rely on temporal calibration (PTP-type methods). We were also interested in the performance of both types of methods under scenarios with divergent rates of coalescence across species. We tested a total of four tree-based methods on the data, including the single-threshold GMYC (sGMYC; Pons *et al.* 2006; Fujisawa & Barraclough 2013), multiple-threshold GMYC (mGMYC; Monaghan *et al.* 2009), bPTP (Zhang *et al.* 2013) and mPTP (Kapli *et al.* 2016). All data were aligned using MUSCLE v.3.8.31 (Edgar 2004) implemented in ALIVIEW v.1.17.1 (Larsson 2014). The total alignment consisted of 871 bp, although sequence lengths for the *P. douglasii* complex were shorter (alignments available on Dryad). Prior to species delimitation analyses, we used RAXML v.8.0.0 (Stamatakis 2014) to remove duplicate haplotypes from a matrix of 368 sequences. This left a total of 220 haplotypes for species delimitation. Although identical sequences should generally be removed prior to tree-based methods of species delimitation (J. Zhang, pers. comm.), we also performed a series of duplicate analyses using all 368 sequences for comparison. We used MEGA7 (Kumar *et al.* 2016) to calculate average within-species genetic distances using the Tamura–Nei model with gamma-distributed rate heterogeneity to account for multiple substitutions. In addition, we used the R-packages APE (Paradis *et al.* 2004) and PEGAS (Paradis 2010) to calculate Watterson's estimator of theta ($\theta = 4N_e\mu$) as an indicator of effective population sizes.

We used BEAST v.2.4.3 (Bouckaert *et al.* 2014) to generate ultrametric gene trees under a strict clock and constant-size coalescent tree prior following the relative performance of clock models in previous studies (Monaghan *et al.* 2009; Satler *et al.* 2013; Talavera *et al.* 2013). A GTR + I + Γ model of substitution was used as estimated using BIC in JMODELTEST2 (Darriba *et al.* 2012). We calibrated the rate of mtDNA substitution by specifying a normal prior with a mean of 0.00805 substitutions/site/million years and sigma of 0.001 (Bryson *et al.* 2012), a rate initially estimated using vicariant scenarios in geckos (Macey *et al.* 1999). Similar substitution rates have been applied in numerous studies to estimate divergence times in several vertebrate groups (see Macey *et al.* 1999; Bryson *et al.* 2012 for examples), and our specified prior distribution accommodated uncertainty in the estimate. Analyses were run for 20 million generations, sampling every 2000. Convergence and mixing were monitored in TRACER v.1.6 (Rambaut *et al.* 2014), and ESS values >200 indicated adequate sampling of the posterior. TREEANNOTATOR v.2.4.3 (Bouckaert *et al.* 2014) was used to create a maximum clade credibility (MCC) tree using mean heights for node annotation. Maximum-likelihood (ML) phylogenetic analyses were implemented in RAXML under a GTRGAMMA model by first implementing a rapid bootstrap search (Stamatakis *et al.* 2008) with autoMRE bootstopping followed by a full ML search (-f a option). Trees were rooted using *P. asio* (Leaché & Linkem 2015).

Species delimitation analyses

Among the multiple methods of single-locus species delimitation currently available, the most popular is the GMYC model (Pons *et al.* 2006; Fujisawa & Barraclough 2013). Previous studies indicate that the GMYC model is fairly robust (Fujisawa & Barraclough 2013), especially if applied to an ultrametric tree constructed using BEAST (Tang *et al.* 2014), and that the choice of clock model and tree prior has a relatively low impact on the results (Talavera *et al.* 2013). We used the R package SPLITS (Ezard *et al.* 2009) to fit both the single- and multiple-threshold models to the data. We initially included haplotypes only for species containing multiple sequences (*P. mcallii*, *P. platyrhinos*, *P. orbiculare*, *P. hernandesi*, *P. douglasii* complex). However, sGMYC results were not significantly different from the null model of coalescence. Thus, we added singletons of *P. cornutum*, *P. coronatum*, *P. solare*, *P. asio* and *P. taurus* to increase the Yule portion of the tree and better fit the model to the data (Talavera *et al.* 2013).

bPTP analyses were performed using the online server (http://species.h-its.org/) and the ML trees from RAXML. We ran the analyses for 500 000 generations with a thinning of 500 and burn-in of 0.1. Convergence was assessed by visualizing plots of MCMC iteration vs. log-likelihood. We ran analyses both with and without the outgroup taxon (*P. asio*). As results were qualitatively similar (not shown), all subsequent comparisons were made with the outgroup to negate taxonomic discrepancy among analyses. We compared the results from bPTP to the recently developed mPTP model that accommodates different rates of coalescence within clades (Kapli *et al.* 2016). Discordant coalescent patterns could be due to uneven sampling intensity among species or varying degrees of genetic structure arising from differences in evolutionary processes and effective population sizes (N_e). We performed both ML and MCMC analyses on RAXML ML trees using the standalone MPTP software (v.0.1.1). MCMC analyses were run for 100 million generations, sampling every 10 000. The first 2 million generations were discarded as burn-in and analyses started from the ML species delimitation estimate (identical results were obtained when starting from both random and null delimitations). Convergence was again assessed by monitoring the plot of generation vs. log-likelihood. Both ML and MCMC analyses utilized the *–multi* option to incorporate differences in rates of coalescence among species and used a minimum branch length of 0.0001. We compared results among multiple MCMC runs (10) to assess congruence.

Because our taxonomic sampling for the four target taxa was highly uneven (i.e. *P. platyrhinos* was represented by ~3× the number of haplotypes), we reran all of the above analyses after pruning haplotypes from *P. platyrhinos* to determine the potential influence of highly heterogeneous sampling intensity on species delimitation. Haplotypes were selected (40 of 111 total) based on the previous phylogenetic analyses to maximize diversity within the species. A new model of substitution was then calculated (HKY + I + Γ), and all phylogenetic and species delimitation analyses were repeated as described above. However, only unique haplotypes (149) were included in this set of analyses.

To compare the results from the tree-based methods above, we ran Automatic Barcode Gap Discovery (ABGD; Puillandre *et al.* 2012a) on both the full and pruned data sets. ABGD is a computationally efficient distance-based method of species delimitation that has been shown to perform well when compared to tree-based coalescent methods (Puillandre *et al.* 2012b; Kekkonen & Hebert 2014; Kapli *et al.* 2016) and other threshold techniques (Ratnasingham & Hebert 2013). The method seeks to quantify the location of the barcode gap that separates intra- from interspecific distances. As the presence of singletons may bias the analysis (Puillandre *et al.* 2012a), ABGD analyses were restricted to *P. mcallii*, *P. orbiculare*, *P. platyrhinos* and the *P. douglasii* complex. Default settings were used for the prior range for maximum intraspecific divergence (0.001, 0.1). Results were compared using both JC69 and K80 corrected distances and minimum slope increase (*X*) of 1.5 (default) and 1.0.

For all analyses, we reported the number of delimited species inferred by each method along with the corresponding confidence intervals. In addition, we used current horned lizard taxonomy (Leaché & Linkem 2015; Montanucci 2015), to compare the proportion of delimited species matching taxonomic species, the proportion of taxonomic species lumped into a delimited species and the number of taxonomic species splits. We note that large values for species lumps and splits may not necessarily indicate poor performance of methods, but they do provide evidence that horned lizard taxonomy might be in need of revision. Next, we reported the match ratio (following Ahrens *et al.* 2016) using the following formula:

$$match\,ratio = 2 * \frac{N_{\text{match}}}{(N_{\text{delimited}} + N_{\text{morph}})},$$

where N_{match} is the number of delimited species exactly matching taxonomic species, $N_{\text{delimited}}$ is the total number of delimited species, and N_{morph} is the number of taxonomic, morphologically defined species. Finally, we quantified performance of methods using the recently developed Relative Taxonomic Resolving Power Index (R_{tax}) and the Taxonomic Index of Congruence (C_{tax}) following Miralles & Vences (2013). The R_{tax} index quantifies the relative power of a method to infer all estimated speciation events present in a data set (large R_{tax} = small type II error), but does not necessarily imply correct delimitations (i.e. can lead to oversplitting). R_{tax} metrics were calculated as follows:

$$R_{\text{tax}}(A) = \frac{nA}{n(A \cup B \cup C \cup D \cup E)},$$

where *A*, *B*, *C*, *D*, *E* represent the five species delimitation methods tested, the numerator (*nA*) represents the number of speciation events inferred by method *A*, and the denominator represents the cumulative number of speciation events inferred by all methods. An R_{tax} value of 1 would indicate that the method recovered all speciation events present across methods.

The C_{tax} index is a measure of congruence in species assignments among two methods, with a value of 1 indicating complete congruence. C_{tax} metrics were calculated as follows:

$$C_{\text{tax}}(AB) = \frac{n(A \cap B)}{n(A \cup B)}$$

where *A* ∩ *B* represents the number of speciation events shared by methods *A* and *B*, and *A* ∪ *B* represents the total number of speciation events inferred by method *A* and/or *B*. We refer the reader to the original publication for additional descriptions of these metrics (Miralles & Vences 2013).

Results and discussion

Uneven sampling

Inferred genealogies were congruent with previous studies based on mtDNA (Reeder & Montanucci 2001; Hodges & Zamudio 2004; Leaché & McGuire 2006) and discordant from nuclear phylogenies (Leaché & Linkem 2015), presumably due to historical introgression (Leaché & McGuire 2006; Leaché & Linkem 2015). Our inferred divergence times based on an assumed mtDNA substitution rate of 0.00805 substitutions per site per million years were congruent with times previously inferred using secondary calibration information for the crown age of phrynosomatids (Leaché & Linkem 2015). Average pairwise sequence divergence varied considerably among species, from relatively low levels in *P. mcallii* (0.94%) to high divergence within the *P. douglasii* complex (7.7%; Table 1). Similarly, effective population sizes varied by an order of magnitude.

All MCMC species delimitation analyses indicated adequate convergence based on visualizing plots of generation vs. likelihood score. Because no differences were detected when using a single run vs. 10 independent MCMC runs, we present results from single runs only. The number of horned lizard species inferred by each delimitation method was substantially different between the full (220 haplotypes) and reduced (149 haplotypes) data sets.

For the full data set, sGMYC was the most conservative method, inferring a total of 10 species with wide confidence intervals (Table 1; Fig. 1). The comparatively small number of species inferred with sGMYC was likely a result of the method having difficulty in locating the threshold point in the data (Fig. S1A, Supporting information). sGMYC is expected to work well when there is

Table 1 Number of horned lizard species (*Phrynosoma*) inferred by each single-locus species delimitation method tested on the full data set of unique haplotypes

Taxon	*n*	Mean Tamura–Nei distance	Watterson's theta	GMYC single	GMYC multiple	bPTP	mPTP	ABGD
P. douglasii complex	41	0.0774	115.6936	2	29	26	2	4
P. orbiculare	34	0.0633	56.985	1	14	11	6	5
P. mcallii	29	0.0094	12.4772	1	2	1	1	1
*P. platyrhinos**	111	0.0211	35.2124	1	31	4	4	2
P. asio	1	NA	NA	1	1	1	1	—
P. cornutum	1	NA	NA	1	1	1	1	—
P. coronatum	1	NA	NA	1	1	1	1	—
P. solare	1	NA	NA	1	1	1	1	—
P. Taurus	1	NA	NA	1	1	1	1	—
Total	220	0.1258††	121.1346†	10 (7–98)‡	81 (69–103)§	52.40 (37–84)¶	18 (16–21)**	12
Species matched (percentage of total)	—	—	—	0.80	0.07	0.15	0.33	0.17
Species lumped (percentage of total)	—	—	—	0.50	0.02	0.04	0.28	0.42
Species splits (percentage of total)	—	—	—	0.20	0.95	0.74	0.67	0.83
Match ratio	—	—	—	0.70	0.13	0.24	0.39	0.20

Results from GMYC and PTP are from genealogies containing 220 ND4 haplotypes (*n*) with highly heterogeneous sampling intensity of target taxa. All bPTP and mPTP results are from Bayesian MCMC analyses. Confidence intervals for totals are in parentheses. ABGD results are based on the initial partitioning scheme with a maximum intraspecific diversity value of 0.012915 (K80 distances). Singletons were pruned prior to ABGD analysis. Taxonomy for *P. douglasii* complex follows Montanucci (2015) and includes *P. douglasii, P. hernandesi*, *P. bauri*, *P. brevirostris* and *P. ornatissimum.* Also shown are average corrected pairwise distances (substitutions/site) for each species containing multiple haplotypes and Watterson's theta ($\theta = 4N_e\mu$). NA, not applicable. 'Species matched' refers to the proportion of delimited species matching defined taxonomic species; 'Species lumped' indicates the proportion of taxonomic species classified within a delimited species; 'Species splits' represents the proportion of taxonomic species splits by each delimitation method.

*~3× number of haplotypes vs. *P. douglasii* complex, *P. orbiculare*, *P. mcallii*

†Theta for all 220 sequences

‡Likelihood ratio vs. null model 10.90506, $P = 0.0043$.

§Likelihood ratio vs. null model 23.85364, $P < 0.001$.

¶Mean number of species.

**Central Credible Interval.

††Mean Tamura–Nei distance for all 220 sequences.

– : Not Applicable.

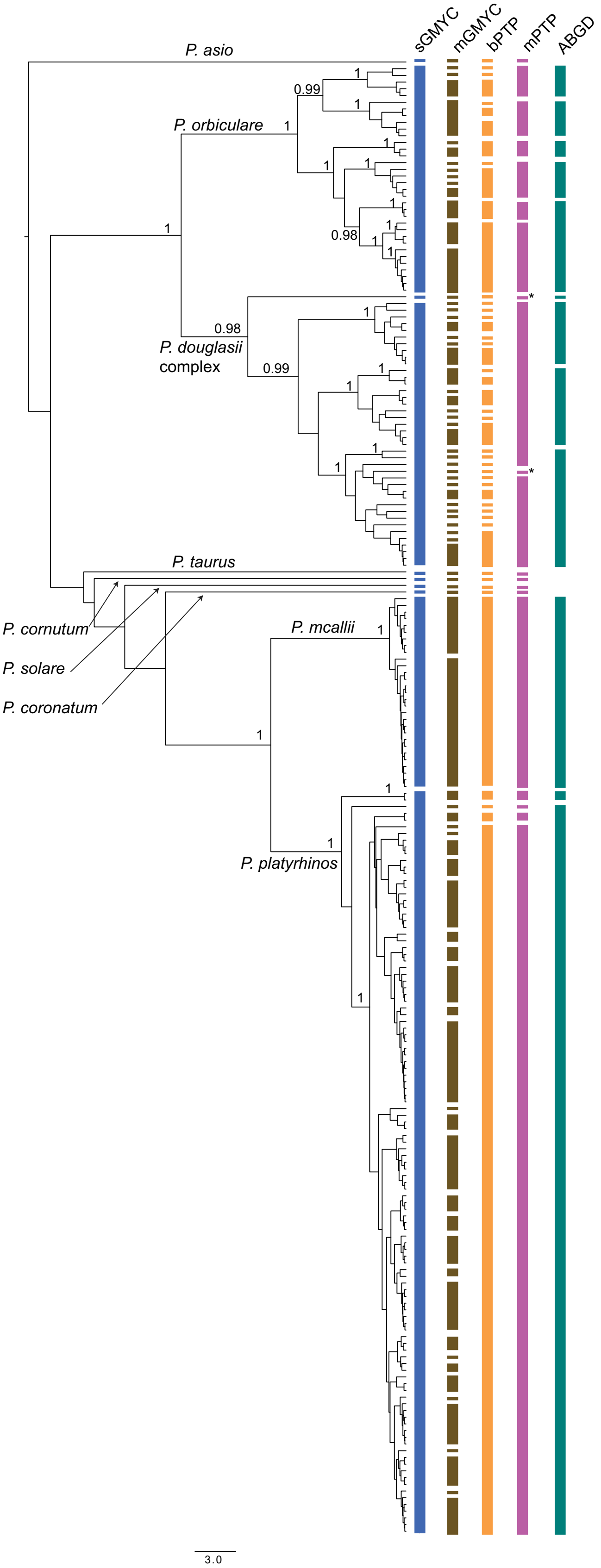

Fig. 1 Comparison of species delimitation results of horned lizards (*Phrynosoma*) based on analysis of 220 unique ND4 haplotypes and highly heterogeneous taxonomic sampling. Each coloured bar represents a species delimited by each method tested. Gene tree is from a BEAST analysis under a strict clock and constant-size coalescent tree prior. Node height was determined using mean heights across the posterior distribution. Node values represent Bayesian posterior probabilities (>0.95) for major clades. sGMYC, single-threshold GMYC; mGMYC, multiple-threshold GMYC; bPTP, single-rate Poisson tree processes model (MCMC analysis of a RAXML gene tree) fitting a single branch length distribution to coalescent events; mPTP, multirate Poisson tree processes model (MCMC analysis of a RAXML gene tree) fitting multiple branch length distributions to coalescent events across distinct species. ABGD, Automatic Barcode Gap Discovery. Note that singletons (i.e. *P. asio*, *P. taurus*, *P. cornutum*, *P. solare*, *P. coronatum*) were removed prior to ABGD analysis. *OTUs (PDU71587, PDU71589) clustered together as a single species based on mPTP analysis of a RAXML gene tree, which yielded a slightly different topology than the BEAST genealogy shown. See Fig. 3 for additional details. Refer to online version for a full colour representation of figure. [Colour figure can be viewed at wileyonlinelibrary.com]

a clear demarcation in branching rates between vs. within species (Pons *et al.* 2006; Esselstyn *et al.* 2012; Reid & Carstens 2012; Fujisawa & Barraclough 2013), which was not the case in the horned lizard data as the threshold was placed relatively deep in the genealogy. Conversely, mGMYC suggested an unrealistically large number of species (81), a high proportion of splits (0.95) and low match ratio (0.13) and appeared to have an equally difficult time placing transition points between inter- and intraspecific branching processes (Fig. S1B, Supporting information). mGMYC suggested up to 31 species within *P. platyrhinos* alone (Fig. 1), which is unlikely given the low levels of divergence within the species (Table 1). bPTP also suggested an unrealistically high number of horned lizard species (52) with wide confidence intervals from MCMC analyses. However, unlike mGMYC, bPTP inferred only four species within *P. platyrhinos* (Fig. 1). Both mGMYC and bPTP inferred a large number of species within the *P. douglasii* complex (29 and 26 species, respectively). mPTP analyses suggested an intermediate number of total species (18) that was the most congruent with the relative levels of structure in the inferred genealogies (Table 1; Fig. 1). These results were consistent with previous findings that the original PTP model tends to oversplit, whereas mPTP is more conservative and likely to represent true species clusters (Kapli *et al.* 2016).

ABGD analysis suggested a total of 12 species based on initial partitioning over a range of prior values for maximum intraspecific divergence (Table 1; Fig. S2, Supporting information). Results based on JC69 and K80 corrected distances were identical. The number of species decreased to five with a maximum intraspecific divergence prior value (P) of 0.021554 and to one species with a value of 0.035938. Although there is still a lack of consensus of how to interpret discordant ABGD results (Kekkonen & Hebert 2014), previous studies advocate using a P value of ~0.01 (Puillandre *et al.* 2012a), which in our data would result in the recognition of 12 species of horned lizards (excluding the singletons). The relatively low value for the proportion of species matches (0.17) and match ratio (0.20) was likely a result of both excluding singletons from the analysis and the likely presence of multiple undescribed species in the data. Applying the recursive algorithm resulted in a maximum of 20 species when X = 1.5 (Fig. S2A, Supporting information). This value increased to a maximum of 36 species when X = 1.0 (Fig. S2B, Supporting information). To be conservative, we focus on the results from the initial partitioning as this scheme has been shown to be more stable across parameter settings and congruent with other species delimitation methods (Puillandre *et al.* 2012a,b; Kekkonen & Hebert 2014), including those examined in this study.

R_{tax} values for the full 220 unique haplotype data set ranged from 0.05 for sGMYC to 0.96 for mGMYC (Table 2), further illustrating the tendency of mGMYC to delimit a large number of (likely erroneous) species. Congruence among methods, based on C_{tax} values, was highest between mPTP and ABGD (0.64) and lowest between sGMYC and mGMYC (0.05). mPTP and ABGD also showed the largest mean C_{tax} among methods (Table 2).

Results of the full 368 sequence data sets (including identical sequences) revealed varying degrees of sensitivity of methods to the presence of duplicates, with most algorithms suggesting additional species (Table S2, Supporting information). mGMYC was by far the most sensitive of the methods compared, inferring a total of 164 species vs. 81 species in the unique 220 haplotype data set. In general, the performance of all methods (except bPTP) was reduced based on match ratios. Interestingly, mPTP failed to distinguish between *P. cornutum* and *P. solare* even though these species are distantly related (Leaché & Linkem 2015). The performance of ABGD also seemed to be impacted by the inclusion of identical sequences, concordant with other recent findings (Ahrens *et al.* 2016). Thus, although previous studies

Table 2 Calculation of the taxonomic index of congruence (C_{tax}) and relative taxonomic resolving power index (R_{tax}) for all species delimitation methods compared based on the full 220 unique haplotype data set

	C_{tax}							
	sGMYC	mGMYC	bPTP	mPTP	ABGD	Mean C_{tax}	R_{tax}	# species
sGMYC	—	—	—	—	—	0.22	0.05	5
mGMYC	0.05	—	—	—	—	0.22	0.96	76
bPTP	0.12	0.50	—	—	—	0.3	0.53	42
mPTP	0.33	0.16	0.29	—	—	0.36	0.15	13
ABGD	0.36	0.15	0.27	0.64	—	0.36	0.14	12

Total number of speciation events across all methods (excluding singletons) = 78; Mean C_{tax} = average value for method across all pairwise comparisons. Singletons were excluded from calculations to allow fair comparisons (all singletons were pruned prior to ABGD analyses). See text for additional details.

suggest that the GMYC model may be robust to the inclusion of identical sequences (Talavera *et al.* 2013), our results suggest that additional bias may be introduced when duplicate haplotypes are not removed.

Even sampling

Results from our analyses using a relatively even sampling scheme (after pruning the number of *P. platyrhinos* haplotypes to 40) revealed varying levels of sensitivity among the species delimitation methods to uneven sampling among species. sGMYC results were virtually identical to the full analysis due to the same position of the inflection point (Fig. S3A, Supporting information), but the confidence interval was substantially reduced. In contrast, mGMYC was extremely sensitive to sampling regime (Figs 1,2; Table 3) and placed the four thresholds at different times in the evenly sampled genealogy (Fig. S3B, Supporting information). The proportion of species splits inferred by mGMYC was substantially reduced in the pruned data set (0.71 vs. 0.95), and the match ratio was increased from 0.13 to 0.39. For example, only a single species within *P. platyrhinos* was inferred by mGMYC for the reduced data set (Fig. 2), compared to 31 species for the full data (Fig. 1). These results were surprising since we pruned *P. platyrhinos* haplotypes evenly throughout the original genealogy. mGMYC also inferred different numbers of species within *P. orbiculare* and the *P. douglasii* complex. Thus, in contrast to sGMYC, mGMYC appears quite sensitive to sampling conditions, which may further limit the utility of the method (Esselstyn *et al.* 2012; Fujisawa & Barraclough 2013; Talavera *et al.* 2013). Conversely, bPTP and mPTP appear less sensitive to sampling issues as the number of inferred species, proportion of matches, lumps and splits and match ratios were similar between both sets of analyses (Tables 1,3). mPTP was the most consistent, with 18 species inferred from the full data set and 17 in the reduced data set (Figs 1,2). The discrepancy in the single species arose from a slightly different gene tree for the reduced data set within the *P. platyrhinos* clade. Zhang *et al.* (2013) also tested for the influence of sampling evenness on species delimitation results and found that PTP outperformed GMYC with even sampling, whereas GMYC was slightly more accurate with uneven sampling.

Due to the extreme sensitivity of mGMYC to sampling conditions, the largest R_{tax} value was obtained from bPTP (1.00 vs. 0.59 for mGMYC) in the pruned data set (Table 4). Similar to the full 220 haplotype analysis, pairwise C_{tax} was highest between mPTP and ABGD (0.69) and in this case lowest between sGMYC and bPTP (0.11; Table 4).

Collectively, these results suggest mGMYC and bPTP were more sensitive to sampling regime and that the large difference in the inferred number of species between bPTP and mPTP is likely due to the latter fitting multiple exponential branch length distributions to species to account for different rates of coalescence in heterogeneous data sets containing species with different N_e and demographic histories. Thus, our results provide further empirical evidence that mPTP may be a good choice for single-locus species delimitation based on accuracy, consistency and speed (Kapli *et al.* 2016).

ABGD analyses on the evenly sampled (pruned) data set also indicated 12 species for most values of P using the initial partition (Table 3; Fig. S4A,B, Supporting information). Once again, the recursive partition suggested a substantially higher number of species, particularly when $P < 0.0129$. There were also slight differences in the recursive partition between K80 and JC69 corrected distances (Fig. S4A,B, Supporting information). However, K80 distances indicated the same number of groups (12) among the initial and recursive partition when $P = 0.0129$.

Performance of methods

Many studies have examined the GMYC model in detail using both simulated and empirical data (Esselstyn *et al.* 2012; Reid & Carstens 2012; Fujisawa & Barraclough 2013; Talavera *et al.* 2013; Tang *et al.* 2014; Ahrens *et al.* 2016). Tang *et al.* (2014) quantified the influence of gene tree reconstruction method and rate smoothing technique on the performance of both GMYC and PTP and found that the former was generally more sensitive to the selected model. They found that most of the sensitivity was likely due to errors during the smoothing step and subsequently advocated the use of BEAST to generate ultrametric gene trees. Talavera *et al.* (2013) used a large butterfly data set to test the performance of GMYC and suggested that the model is highly stable under a variety of conditions including tree reconstruction method, the number of singletons included, the number of species included in the gene tree and sampling coverage. They provided a summary table and chart with recommendations for running GMYC on empirical data sets. Interestingly, their analysis suggested that sGMYC often overestimates the number of species, in contrast to our analysis where the model may be underestimating true diversity. However, our results were concordant in the sense that species delimitations did not change significantly with different sample coverage, although it was slightly impacted by the inclusion of identical sequences. Studies have also indicated that GMYC may be sensitive to general phylogenetic history, sampling intensity, DNA sequence length, speciation rate, N_e and

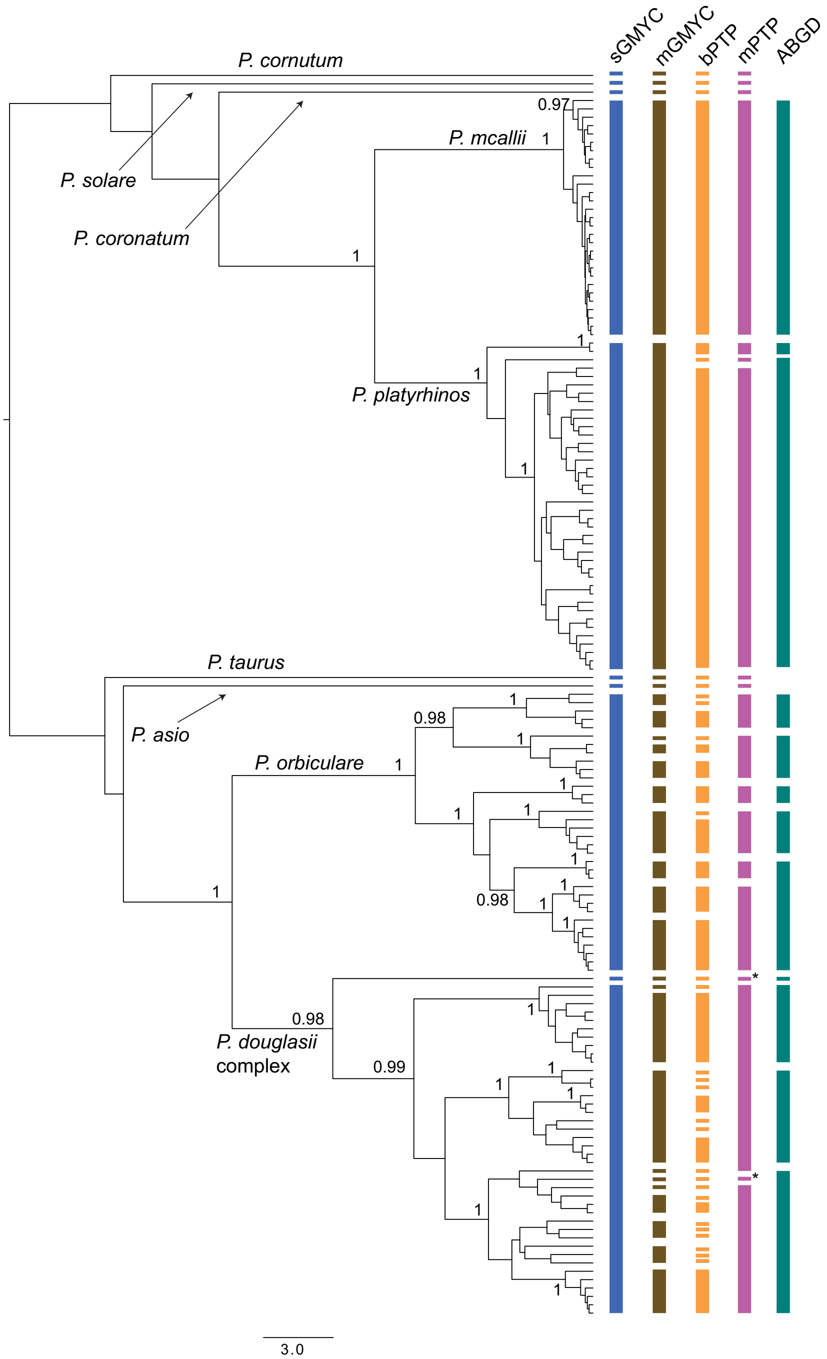

Fig. 2 Comparison of species delimitation results of horned lizards (*Phrynosoma*) based on analysis of 149 unique ND4 haplotypes with relatively even intraspecific sampling. Each coloured bar represents a species delimited by each method tested. Gene tree is from a BEAST analysis under a strict clock and constant-size coalescent tree prior. Node height was determined using mean heights across the posterior distribution. Node values represent Bayesian posterior probabilities (>0.95) for major clades. sGMYC, single-threshold GMYC; mGMYC, multiple-threshold GMYC; bPTP, single-rate Poisson tree processes model (MCMC analysis of a RAXML gene tree) fitting a single branch length distribution to coalescent events; mPTP, multirate Poisson tree processes model (MCMC analysis of a RAXML gene tree) fitting multiple branch length distributions to coalescent events across distinct species. ABGD, Automatic Barcode Gap Discovery. Note that singletons (i.e. *P. asio*, *P. taurus*, *P. cornutum*, *P. solare*, *P. coronatum*) were removed prior to ABGD analysis. *OTUs (PDU71587, PDU71589) clustered together as a single species based on mPTP analysis of a RAXML gene tree, which yielded a slightly different topology than the BEAST genealogy shown. See Fig. 3 for additional details. Refer to online version for a full colour representation of figure. [Colour figure can be viewed at wileyonlinelibrary.com]

differences in N_e among species (Esselstyn *et al.* 2012; Reid & Carstens 2012) and may sometimes underestimate (sGMYC) or overestimate (mGMYC) the true number of species (Esselstyn *et al.* 2012). This is a likely explanation for our sGMYC results that estimated only 10 species of horned lizards, as there was no abrupt change in branching rates between vs. within species. More recently, Ahrens *et al.* (2016) used both simulated and empirical data to better understand potential biases in GMYC due to sampling and population genetic artefacts. Their results suggested that the majority of bias is introduced by variation in N_e among species, which can be exacerbated by uneven species abundance/sampling. In these cases, sGMYC tends to lump species and return wide confidence intervals, which is consistent with our results for horned lizards. To help overcome these issues, they suggest increasing the number of clades examined to balance out the large skew in N_e among species.

Table 3 Number of horned lizard species (*Phrynosoma*) inferred by each single-locus species delimitation method tested on the pruned data set

Taxon	*n*	Mean Tamura–Nei distance	Watterson's theta	GMYC single	GMYC multiple	bPTP	mPTP	ABGD
P. douglasii complex	41	0.0774	115.6936	2	11	22	2	4
P. orbiculare	34	0.0633	56.985	1	10	12	6	5
P. mcallii	29	0.0094	12.4772	1	1	1	1	1
P. platyrhinos	40	0.0256	34.7945	1	1	3	3	2
P. asio	1	NA	NA	1	1	1	1	—
P. cornutum	1	NA	NA	1	1	1	1	—
P. coronatum	1	NA	NA	1	1	1	1	—
P. solare	1	NA	NA	1	1	1	1	—
P. taurus	1	NA	NA	1	1	1	1	—
Total	149	0.1468*	129.0831†	10 (3–19)‡	28 (9–66)§	46.74 (34–68)¶	17 (11–19)**	12
Species matched (percentage of total)	—	—	—	0.80	0.29	0.19	0.35	0.17
Species lumped (percentage of total)	—	—	—	0.50	0.14	0.00	0.29	0.42
Species splits (percentage of total)	—	—	—	0.20	0.71	0.73	0.65	0.83
Match ratio	—	—	—	0.70	0.39	0.30	0.40	0.20

Results from GMYC and PTP are from genealogies containing 149 ND4 haplotypes (*n*) with approximately even sampling intensity of target taxa. All bPTP and mPTP results are from Bayesian MCMC analyses. Confidence intervals for totals are in parentheses. ABGD results are based on the initial partitioning scheme with a maximum intraspecific diversity value of 0.012915 (K80 distances). Singletons were pruned prior to ABGD analysis. Taxonomy for *P. douglasii* complex follows Montanucci (2015) and includes *P. douglasii*, *P. hernandesi*, *P. bauri*, *P. brevirostris* and *P. ornatissimum*. Also shown are average corrected pairwise distances (substitutions/site) for each species containing multiple haplotypes and Watterson's theta ($\theta = 4N_e\mu$). NA, Not applicable. 'Species matched' refers to the proportion of delimited species matching defined taxonomic species; 'Species lumped' indicates the proportion of taxonomic species classified within a delimited species; 'Species splits' represents the proportion of taxonomic species splits by each delimitation method.

*Mean Tamura–Nei distance for all 149 sequences.

†Theta for all 149 sequences.

‡Likelihood ratio vs. null model 10.02172, $P = 0.0067$.

§Likelihood ratio vs. null model 12.4399, $P = 0.002$.

¶Mean number of species.

**Central credible interval.

–: Not Applicable.

Table 4 Calculation of the taxonomic index of congruence (C_{tax}) and relative taxonomic resolving power index (R_{tax}) for all species delimitation methods compared based on the pruned 149 unique haplotype data set

	C_{tax}							
	sGMYC	mGMYC	bPTP	mPTP	ABGD	Mean C_{tax}	R_{tax}	#Species
sGMYC	—	—	—	—	—	0.25	0.11	5
mGMYC	0.18	—	—	—	—	0.40	0.59	23
bPTP	0.11	0.59	—	—	—	0.33	1.00	38
mPTP	0.36	0.38	0.30	—	—	0.43	0.30	12
ABGD	0.36	0.43	0.30	0.69	—	0.45	0.30	12

Total number of speciation events across all methods (excluding singletons) = 37; Mean C_{tax} = average value for method across all pairwise comparisons. Singletons were excluded from calculations to allow fair comparisons (all singletons were pruned prior to ABGD analyses). See text for additional details.

Although this solution may alleviate some of the issues with sGMYC, mPTP may be more reliable in such cases as the method can explicitly account for differences in N_e and rates of coalescence among species.

To our knowledge, few studies have examined the potential influence of methods for summarizing node height information in BEAST analyses for subsequent species delimitation using GMYC. To provide preliminary data on this issue, we performed a suite of additional sGMYC and mGMYC analyses on both the full (368 sequences) and unique haplotype (220 sequences) data sets. Three methods of summarizing node heights were compared: mean heights, median heights, common ancestor heights. Different results were detected depending on whether the single- or multiple-threshold model was used (Table 5). Using common ancestor node heights with sGMYC dramatically increased the number of delimited species, due to the threshold point be pushed closer to the present (Fig. S5, Supporting information). This effect was negligible with mGMYC, which appeared to be more affected by the inclusion of identical sequences (Fig. S6, Supporting information) resulting in a doubling of the number of inferred species (Table 5). Thus, there appears to be additional nuances of GMYC that should be considered when utilizing these methods on empirical data. Based on relative concordance with our mPTP and ABGD analyses, using sGMYC with mean or median node heights may be a good approach. Additional simulation studies will be needed to test this prediction further.

Given the high levels of discordance observed among the methods tested, how should researchers use these algorithms to discover and delimit diversity? As detailed above, mPTP has numerous advantages over other methods. The consistency of mPTP to delimit putative species in our study despite varied sampling depths and effective population sizes provides additional evidence suggesting that the model may be appropriate for a wide variety of empirical data sets (Kapli *et al.* 2016). Further, mPTP may be a good choice in data sets such as ours where sGMYC has difficulty in placing the threshold point due to a more gradual slope in branching times, possibly as a result of sampling a low species-to-individual ratio or due to large differences in N_e among species (Ahrens *et al.* 2016). However, we agree with previous authors in that taxonomic changes should not be made solely on the results of these methods (Lohse 2009; Esselstyn *et al.* 2012; Puillandre *et al.* 2012a; Talavera *et al.* 2013; Zhang *et al.* 2013), although concordance using multiple analyses does tend to increase reliability (Puillandre *et al.* 2012b; Carstens *et al.* 2013; Satler *et al.* 2013). Rather, robust single-locus approaches should be used to form primary taxonomic hypotheses that are subsequently tested with other types of data as part of an integrative taxonomic framework (Fujita *et al.* 2012). Although ABGD has the potential to offer a rapid and robust framework for assessing concordance (Puillandre *et al.* 2012a,b; Ahrens *et al.* 2016), additional work is needed to determine optimal parameter settings and whether the recursive partition tends to oversplit. Moreover, additional studies are needed to compare the performance of RESL through BOLD against some of the newer tree-based methods (e.g. bPTP and mPTP).

Table 5 Comparison of the number of delimited horned lizard (*Phrynosoma*) species by the single- (sGMYC) and multiple-threshold (mGMYC) GMYC models based on different methods of annotating node height and sampling regimes

Data set	Node heights	sGMYC	mGMYC
Unique	Mean heights	10 (7–98)**	81 (69–103)***
Unique	Common ancestor heights	64 (58–72)***	66 (61–66)***
Unique	Median heights	9 (7–69)**	76 (68–103)***
All	Mean heights	13 (8–138)*	164 (157–164)***
All	Common ancestor heights	79 (76–87)***	100 (83–100)***
All	Median heights	9 (7–16)*	191 (13–299)***

'Unique' = only unique haplotypes (220); 'All' = all sequences (368). Three ways to summarize node heights on BEAST maximum clade credibility trees were evaluated ('Mean heights', 'Common ancestor heights', 'Median heights'). Values represent the number of ML entities with confidence intervals.

*$P < 0.05$.

**$P < 0.005$.

***$P < 0.001$.

The adoption of single-locus species delimitation methods to biodiversity research seems particularly relevant to large metabarcoding studies (including microbial 16S rRNA sequencing) as a rapid and cost-effective means to target groups for additional investigation. High-throughput, multiplex amplicon sequencing using next-generation sequencing platforms allows for the rapid generation of single-locus data from a large number of samples for primary species delineation. Due to its speed and accuracy, mPTP seems to be an ideal analytical tool for these large heterogeneous data sets consisting of species with different coalescent histories. We anticipate that empiricists will continue to explore the utility of single-locus methods as sequencing technologies improve and new analytical tools are developed.

Taxonomy of horned lizards

Based on the performance and consistency of single-locus species delimitation methods tested in this study, species-level diversity within *Phrynosoma* might be underestimated. Analyses of the mitochondrial gene ND4 provide evidence to suggest at least 11 species could be present *P. mcallii* (1 spp.), *P. platyrhinos* (2 spp.), *P. orbiculare* (5 spp.) and *P. douglasii* complex (3 spp.). We note that this interpretation is conservative and based on congruence between mPTP and ABGD analyses. Within *P. orbiculare*, both methods suggest the presence of one species in the northern Sierra Madre Occidental of Chihuahua, one species in the southern Sierra Madre

Occidental, one in the northern Sierra Madre Oriental and Central Mexican Plateau, one in the southern Sierra Madre Oriental and adjacent Trans-Mexican Volcanic Belt (Veracruz and Puebla) and one in the central Trans-Mexican Volcanic Belt (Estado de México and Distrito Federal). mPTP also suggested an additional species in the southern Sierra Madre Oriental in Hidalgo (SMOr-H; Fig. 3).

Species delimitation scenarios for the *P. douglasii* complex varied among the methods. In our RAXML gene trees, a strongly supported *P. douglasii* clade (bootstrap = 93%) was nested deep within *P. hernandesi* sensu lato (s.l.), albeit with weak support (<50% bootstrap; Fig. 3). However, this relationship resulted in nonmonophyly of *P. hernandesi* in RAXML gene trees. In contrast, BEAST analyses reconstructed two major clades, one consisting of haplotypes from *P. hernandesi* s.l. and one containing haplotypes from *P. douglasii*. This is one reason why the tree-based species delimitation methods failed to differentiate the two species. In all phylogenetic analyses, however, a single divergent haplotype was found that may represent a third species within this clade (PDU71587 from Arizona), a finding that warrants further evaluation in future studies. ABGD further split *P. hernandesi* into two putative species that correspond to the 'SER' and 'GB/CP' lineages of Zamudio *et al.* (1997). There was weak evidence to support the recently proposed taxonomy of Montanucci (2015), who resurrected the names *P. brevirostris* and *P. ornatissimum* and described two new taxa (*P. bauri* and *P. diminutum*) based on morphological comparisons. *Phrynosoma bauri* and *P. ornatissimum* were represented in our data by one haplotype each, which were both strongly nested within *P. hernandesi* sensu stricto (s.s.) in all phylogenetic analyses. In addition, nearly all species delimitation analyses (including tree-based and distance-based) suggested that these taxa are not distinct from *P. hernandesi*. Only bPTP and mGMYC analyses suggested that these species may be valid, but these analyses generally tended to split *P. hernandesi* s.l. into a large number of singletons that

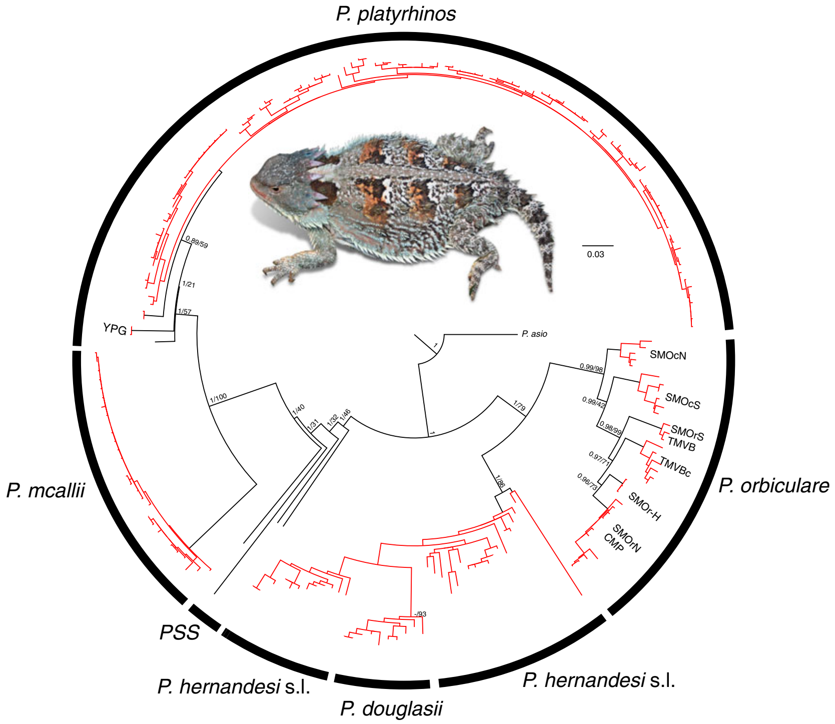

Fig. 3 Species delimitation of horned lizards (*Phrynosoma*) based on MCMC mPTP analysis of a RAXML gene tree constructed with 220 ND4 sequences under a GTRGAMMA model. Branches are colour-coded to represent speciation (black) or coalescence (red) events. Values at nodes indicate probability of a speciation event based on mPTP MCMC analysis (first value) and maximum-likelihood bootstrap proportions using the autoMRE bootstopping criterion in RAXML (second value). YPG, samples LVT9951 and LVT9952 from Yuma Proving Ground, Arizona. SMOcN, northern Sierra Madre Occidental; SMOcS, southern Sierra Madre Occidental; SMOrS, southern Sierra Madre Oriental; TMVB, Trans-Mexican Volcanic Belt; TMVBc, central Trans-Mexican Volcanic Belt; SMOr-H, Sierra Madre Oriental-Hidalgo; SMOrN, northern Sierra Madre Oriental; CMP, Central Mexican Plateau. *PSS*, *Phrynosoma* singleton species (*P. cornutum*, *P. coronatum*, *P. solare*, *P. taurus*). *P. hernandesi* s.l., *P. hernandesi*, *P. bauri*, *P. brevirostris*, *P. ornatissimum*. Horned lizard shown is a *P. orbiculare* from Coahuila, Mexico. Refer to online version for a full colour representation of figure. [Colour figure can be viewed at wileyonlinelibrary.com]

are not consistent with the recovered genealogies, and thus, we view these results as unlikely. *Phrynosoma brevirostris* was recovered as a clade deeply nested within *P. hernandesi* s.s. in all the genealogies. Once again, only bPTP and mGMYC suggested that *P. brevirostris* is a distinct species. Interestingly, mGMYC on the full 220 haplotype data set split *P. brevirostris* into multiple species, whereas mGMYC on the evenly sampled data set lumped all haplotypes into a single entity. Much more taxonomic work remains for the *P. douglasii* complex, and it is likely that nuclear data will be required to resolve the discrepancies between morphology and mtDNA data.

Within *P. platyrhinos*, our results suggested between two (ABGD) and four (mPTP) species, with strong support from mPTP. Concordance among results suggests two samples from Yuma Proving Ground in La Paz County, Arizona (LVT9951 and LVT9952), may represent a new species, consistent with previous suggestions (Mulcahy *et al.* 2006; Jezkova *et al.* 2016). mPTP results also suggested that sample LVT818 was a separate taxon as was a clade containing samples DGM478 and DGM481, although the latter clade was weakly placed in alternative positions in all of our ML and Bayesian analyses and unlikely to represent a distinct species. Finally, the majority of analyses suggested that *P. mcallii* consists of a single species, concordant with the low level of genetic diversity in the species (Mulcahy *et al.* 2006). However, mGMYC on the full 368 sequence data set suggested a highly unrealistic estimate of 18 species within *P. mcallii* (Table S2, Supporting information), further illustrating the propensity of this method to oversplit.

We note that although many of the tree-based methods we tested are robust under a variety of scenarios, poorly supported nodes and/or nonmonophyly in gene trees may render results unreliable until further data are collected – a broad limitation of the single-locus approach (Esselstyn *et al.* 2012; Fujisawa & Barraclough 2013; Kapli *et al.* 2016). Thus, when delimiting putative species based on single-locus data, researchers should consider using both tree- and threshold-based methods like ABGD to account for the shortcomings of each type of method (Hamilton *et al.* 2011; Puillandre *et al.* 2012b). Prior to formal taxonomic changes, results should be subsequently tested with additional types of data and analyses, such as morphological and/or ecological data and multilocus coalescent-based methods (Puillandre *et al.* 2012b; Satler *et al.* 2013). This multiple-lines-of-evidence approach can therefore account for gene tree/species tree discordance before formal taxonomic changes are implemented (Fujita *et al.* 2012; Carstens *et al.* 2013). Nevertheless, single-locus methods such as mPTP have the potential to quickly and accurately target lineages that warrant additional investigation.

Acknowledgements

We would like to thank A. Stamatakis, P. Kapli, J. Zhang and members of the PTP Discussion Board for their assistance running all bPTP and mPTP analyses; N. Puillandre and S. Brouillet for suggestions on ABGD analyses; and J. Jones for help with creating the *P. orbiculare* image. We would also like to thank four reviewers who provided valuable comments to improve the manuscript.

References

Ahrens D, Fujisawa T, Krammer HJ, Eberle J, Fabrizi S, Vogler AP (2016) Rarity and incomplete sampling in DNA-based species delimitation. *Systematic Biology*, **65**, 478–494.

Barraclough TG, Hughes M, Ashford-Hodges N, Fujisawa T (2009) Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data. *Biology Letters*, **5**, 425–428.

Blair C, Méndez de la Cruz FR, Law C, Murphy RW (2015) Molecular phylogenetics and species delimitation of leaf-toed geckos (Phyllodactylidae: *Phyllodactylus*) throughout the Mexican tropical dry forest. *Molecular Phylogenetics and Evolution*, **84**, 254–265.

Blaxter ML (2004) The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **359**, 669–679.

Bouckaert R, Heled J, Kühnert D *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, **10**, e1003537.

Bryson RW, García-Vázquez UO, Riddle BR (2012) Diversification in the Mexican horned lizard *Phrynosoma orbiculare* across a dynamic landscape. *Molecular Phylogenetics and Evolution*, **62**, 87–96.

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. *Molecular Ecology*, **22**, 4369–4383.

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

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

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Esselstyn JA, Evans BJ, Sedlock JL, Anwarali Khan FA, Heaney LR (2012) Single-locus species delimitation: a test of the mixed Yule-coalescent model, with an empirical application to Philippine round-leaf bats. *Proceedings of the Royal Society of London B: Biological Sciences*, **279**, 3678–3686.

Ezard T, Fujisawa T, Barraclough TG (2009) *SPLITS: Species' Limits by Threshold Statistics. R package version*, **1.**

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.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C (2012) Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*, **27**, 480–488.

Gebiola M, Gómez-Zurita J, Monti MM, Navone P, Bernardo U (2012) Integration of molecular, ecological, morphological and endosymbiont data for species delimitation within the *Pnigalio soemius* complex (Hymenoptera: Eulophidae). *Molecular Ecology*, **21**, 1190–1208.

Hamilton CA, Formanowicz DR, Bond JE (2011) Species delimitation and phylogeography of *Aphonopelma hentzi* (Araneae, Mygalomorphae, Theraphosidae): cryptic diversity in North American tarantulas. *PLoS ONE*, **6**, e26207.

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology*, **54**, 852–859.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, **270**, 313–321.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences*, **101**, 14812–14817.

Hodges WL, Zamudio KR (2004) Horned lizard (*Phrynosoma*) phylogeny inferred from mitochondrial genes and morphological characters: understanding conflicts using multiple approaches. *Molecular Phylogenetics and Evolution*, **31**, 961–971.

Jezkova T, Jaeger JR, Oláh-Hemmings V *et al.* (2016) Range and niche shifts in response to past climate change in the desert horned lizard *Phrynosoma platyrhinos*. *Ecography*, **39**, 437–448.

Jones GR (2014) STACEY: species delimitation and phylogeny estimation under the multispecies coalescent. bioRxiv, 010199.

Kapli P, Lutteropp S, Zhang J *et al.* (2016) Multi-rate Poisson Tree Processes for single-locus species delimitation under Maximum Likelihood and Markov Chain Monte Carlo. bioRxiv, 063875.

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

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**, 1870–1874.

Larsson A (2014) ALIVIEW: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**, 3276–3278, btu531.

Leaché AD, Fujita MK (2010) Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 3071–3077, rspb20100662.

Leaché AD, Linkem CW (2015) Phylogenomics of horned lizards (Genus: *Phrynosoma*) using targeted sequence capture data. *Copeia*, **103**, 586–594.

Leaché AD, McGuire JA (2006) Phylogenetic relationships of horned lizards (*Phrynosoma*) based on nuclear and mitochondrial data: evidence for a misleading mitochondrial gene tree. *Molecular Phylogenetics and Evolution*, **39**, 628–644.

Leaché AD, Koo MS, Spencer CL *et al.* (2009) Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proceedings of the National Academy of Sciences*, **106**, 12418–12423.

Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using genome-wide SNP data. *Systematic Biology*, **63**, 534–542.

Lohse K. 2009) Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). *Systematic Biology*, **58**, 439–442.

Macey JR, Wang Y, Ananjeva NB, Larson A, Papenfuss TJ (1999) Vicariant patterns of fragmentation among gekkonid lizards of the genus Teratoscincus produced by the Indian Collision: a molecular phylogenetic perspective and an area cladogram for Central Asia. *Molecular Phylogenetics and Evolution*, **12**, 320–332.

Miralles A, Vences M (2013) New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. *PLoS ONE*, **8**, e68242.

Monaghan MT, Wild R, Elliot M *et al.* (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology*, **58**, 298–311.

Montanucci RR (2015) A taxonomic revision of the *Phrynsoma douglasii* species complex (Squamata: Phrynosomatidae). *Zootaxa*, **4015**, 001–177.

Mulcahy DG, Spaulding AW, Mendelson JR, Brodie ED (2006) Phylogeography of the flat-tailed horned lizard (*Phrynosoma mcallii*) and systematics of the *P. mcallii-platyrhinos* mtDNA complex. *Molecular Ecology*, **15**, 1807–1826.

Nieto-Montes de Oca A, Arenas-Moreno D, Beltrán-Sánchez E, Leaché AD (2014) A new species of horned lizard (Genus *Phrynosoma*) from Guerrero, México, with an updated multilocus phylogeny. *Herpetologica*, **70**, 241–257.

Paradis E (2010) PEGAS: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.

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

Pons J, Barraclough TG, Gomez-Zurita J *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, **55**, 595–609.

Puillandre N, Lambert A, Brouillet S, Achaz G (2012a) ABGD, automatic barcode gap discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.

Puillandre N, Modica MV, Zhang Y *et al.* (2012b) Large-scale species delimitation method for hyperdiverse groups. *Molecular Ecology*, **21**, 2671–2691.

de Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–886.

Rambaut A, Suchard MA, Xie D, Drummond AJ (2014) TRACER v1.6.

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

Reeder TW, Montanucci RR (2001) Phylogenetic analysis of the horned lizards (Phrynosomatidae: *Phrynosoma*): evidence from mitochondrial DNA and morphology. *Copeia*, **2001**, 309–323.

Reid NM, Carstens BC (2012) Phylogenetic estimation error can decrease the accuracy of species delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. *BMC Evolutionary Biology*, **12**, 196.

Satler JD, Carstens BC, Hedin M (2013) Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, *Aliatypus*). *Systematic Biology*, **62**, 805–823.

Sherbrooke WC (2003) *Introduction to Horned Lizards of North America*. University of California Press.

Stamatakis A (2014) RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAXML Web servers. *Systematic Biology*, **57**, 758–771.

Talavera G, Dincă V, Vila R (2013) Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. *Methods in Ecology and Evolution*, **4**, 1101–1110.

Tang CQ, Humphreys AM, Fontaneto D, Barraclough TG, Paradis E (2014) Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods in Ecology and Evolution*, **5**, 1086–1094.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, **107**, 9264–9269.

Yang Z, Rannala B (2014) Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution*, **31**, 3125–3135, msu279.

Zamudio KR, Jones KB, Ward RH (1997) Molecular systematics of short-horned lizards: biogeography and taxonomy of a widespread species complex. *Systematic Biology*, **46**, 284–305.

Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, **29**, 2869–2876.

C.B. designed study and performed statistical analyses; C.B. and R.W.B. contributed to manuscript writing and figure preparation.

Data accessibility

All multiple sequence alignments and gene trees can be found on the Dryad Digital Repository (doi:10.5061/dryad.r7989).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Voucher information and GenBank Accession Numbers for all horned lizard (Phrynosoma) samples included in this study. ND = no data.

Table S2 Number of horned lizard species (*Phrynosoma*) inferred by each single-locus species delimitation method tested using all sequences.

Fig. S1 Lineage through time (LTT) plots based on the single-threshold GMYC model (A) and multiple threshold GMYC model (B) using the full data set consisting of 220 haplotypes with uneven sampling (~3× as many haplotypes for *P. platyrhinos*).

Fig. S2 Results from ABGD analysis on the full data set (220 haplotypes) with singletons pruned.

Fig. S3 Lineage through time (LTT) plots based on the single-threshold GMYC model (A) and multiple threshold GMYC model (B) using the reduced, evenly sampled data set consisting of 149 haplotypes (40/111 haplotypes from *P. platyrhinos*).

Fig. S4 Results from ABGD analysis on the evenly sampled data set (149 haplotypes) with singletons pruned.

Fig. S5 Lineage through time (LTT) plots based on the single-threshold (sGMYC) model.

Fig. S6 Lineage through time (LTT) plots based on the multiple-threshold (mGMYC) model.