The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2013, **109**, 737–746. With 2 figures

The carnivorous plant described as *Sarracenia alata* contains two cryptic species

BRYAN C. CARSTENS* and JORDAN D. SATLER

Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, OH 43210, USA

Received 17 January 2013; revised 19 February 2013; accepted for publication 19 February 2013

Modern methods for species delimitation provide biologists with the power to detect cryptic diversity in nearly any system. To illustrate the application of such methods, we collected data (21 sequence loci) from a carnivorous plant in southeastern North America and applied several recently developed methods (Gaussian clustering, Structurama, BPP, spedeSTEM). The pale pitcher plant *Sarracenia alata* inhabits the southeastern USA along the northern coast of the Gulf of Mexico. *Sarracenia alata* populations are separated by the Mississippi River and Atchafalaya Basin, a known biogeographical barrier in this region, but the cohesiveness of *S. alata* as currently classified has not been tested rigorously. Multiple analytical approaches (including allelic clustering and species trees methods) suggest that *S. alata* comprises two cryptic lineages that correspond to the eastern and western portions of the plant's distribution. That such clear genetic evidence for cryptic diversity exists within *S. alata* and is in conflict with other sources of data (e.g. morphology, environmental differentiation) illustrates a conundrum faced by those who investigate species boundaries: genetic data are often the first type of data to accumulate evidence of differentiation, but most existing taxonomic treatments are based on nongenetic data. Our results suggest that *S. alata* as currently described contains two cryptic species, and we recommend the elevation of the western populations to species status. © 2013 The Linnean Society of London, *Biological Journal of the Linnean Society*, 2013, **109**, 737–746.

ADDITIONAL KEYWORDS: carnivorous plants – phylogeography – population genetic structure – species delimitation.

INTRODUCTION

As befitting a discipline that developed at the interface between systematics and population genetics, the detection of cryptic diversity has been one of the primary aims of phylogeographical research (Avise, 2000). Although early investigations relied on qualitative examinations of phylogenies estimated from organellar genomes, several recent methodological advances have leveraged key findings from population genetics to address the problem of cryptic diversity detection. As a result, practitioners now have the ability to discover cryptic species-level diversity in a broad range of systems using a modest amount of genetic data. Given the diversity of recently developed methods, it can be difficult to intuit which to apply to a given empirical system and to infer the biological meaning in the face of incongruence across methods. In order to highlight the challenge and potential of species delimitation, we explored the species boundaries in a carnivorous plant from southeastern North America using a variety of methods. Our focal system is the pale pitcher plant *Sarracenia alata*, a long-lived perennial Angiosperm (order Ericales; family Sarraceniaceae) that occurs in pine savannahs throughout the western Gulf Coast.

Sarracenia alata has a disjunct distribution (Fig. 1), with eastern and western populations separated by nearly 200 km across either side of the Mississippi River/Atchafalaya Basin. This barrier is one of several recurring phylogeographical breaks in plants and animals in southeastern North America (e.g. Soltis *et al.*, 2006; Jackson & Austin, 2010).

*Corresponding author. E-mail: carstens.12@osu.edu

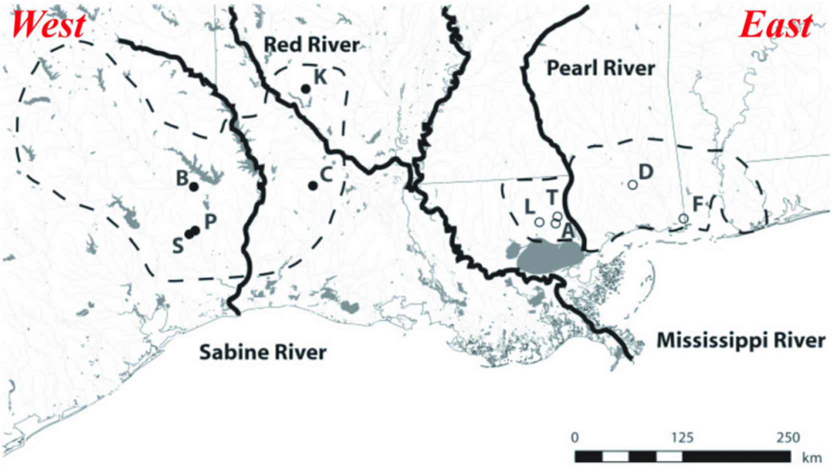

Figure 1. Distribution map of *Sarracenia alata*. Broken lines indicate the range of the species, as indicated by herbarium records. Sampling localities are abbreviated as follows: Sundew (S), Pitcher Trail (P), Bouton Lake (B), Cooter's Bog (C), Kisatchie (K), Lake Ramsay (L), Talisheek (T), Abita Springs (A), DeSoto (D) and Franklin Creek (F). Also shown are the major rivers in the region.

Although earlier investigations into *S. alata* did not identify fixed morphological (Sheridan, 1991) or genetic (Neyland, 2008) variation across this barrier, next-generation sequencing conducted by Zellmer *et al*. (2012) provided evidence that populations on either side of the river are genetically differentiated. Their genetic data were clustered into three partitions: one composed of all eastern samples, a second composed of all samples west of the Red River and a third composed of samples from the Kisatchie site between the Red and Mississippi Rivers. In addition, phylogenetic analysis indicated that the deepest split in the population tree occurs between clades composed of eastern and western populations, albeit with low posterior probability ($P = 0.46$). Although these results were suggestive of cryptic diversity between the eastern and western populations, Zellmer *et al*. (2012) did not offer a comprehensive investigation into this question. Rather, they demonstrated that the main axis of environmental variation was inland to coastal, rather than east to west, and proposed that the genetic structure evident in *S. alata* was promoted by the major rivers in the region, which may serve as barriers to gene flow. Here, we expand on the work of Zellmer *et al*. (2012) to explicitly test for diversity in cryptic lineages in *S. alata*.

The diversity of methodological approaches to species delimitation has increased dramatically in the last decade. In general, these approaches can be separated into two classes – discovery approaches that estimate partitions from the data and thus divide the samples into groups as a part of the analysis, and validation approaches that test partitions in the samples derived from other sources of data (Ence & Carstens, 2011). In general, the discovery approaches are nonphylogenetic and based on clustering algorithms (but see Pons *et al*., 2006; O'Meara, 2010). In contrast, most validation approaches adopt the species tree framework and model lineage composition probabilistically. We operate here under the assumption that each of these approaches has merit, that multiple discovery and validation approaches should be used in any investigation, and that, if a strong signal is present in the data, we should observe congruent results across methods.

MATERIAL AND METHODS

Data acquisition and preliminary analysis

Zellmer *et al*. (2012) collected 82 *S. alata* samples from 10 populations, sequenced a reduced representation library on a Roche 454 and identified 76 variable loci. However, their data matrices contained many gaps, such that the majority of these loci were not sequenced in all individuals. In order to address this shortcoming, we designed polymerase chain reaction (PCR) primers from 20 loci and sequenced individuals with data missing from the matrices using Sanger sequencing. This number was chosen because simulations generally indicate that the accuracy of species delimitation methods does not increase past this number of loci (Yang & Rannala, 2010; Ence & Carstens, 2011; Camargo *et al*., 2012; Rittmeyer & Austin, 2012). In addition, we collected data from the chloroplast gene *rps*16-*trn*k following the protocol

Table 1. Information about the 21 loci used for species delimitation. The name of each locus, the number of individuals sequenced, the model of sequence evolution, the number of variable sites (vs), the number of informative sites (is) and the length of each locus (bp) are shown

Locus	# sequenced	Model	vs	is	bp
cpDNA	50	F81	2	2	354
Sa135	64	JC	10	10	298
Sa14	47	JC	13	10	399
Sa144	51	K80+I	19	15	170
Sa163	28	K80+G	12	10	354
Sa181	31	JC	2	1	187
Sa220	41	K80+I	22	20	441
Sa230	36	HKY	14	8	386
Sa242	50	HKY	7	4	325
Sa297	57	F81+I	6	6	279
Sa302	66	JC	7	7	262
Sa323	55	JC	4	3	401
Sa334	63	JC	5	4	249
Sa340	12	HKY	20	14	439
Sa36	34	JC	14	10	430
Sa39	33	HKY	15	11	431
Sa4	65	F81	3	0	164
Sa40	42	JC	8	7	413
Sa405	42	JC+G	13	10	403
Sa548	60	JC+I	3	3	250
Sa80	30	JC	10	5	412
Total	**957**		**209**	**160**	**7047**

described by Koopman & Carstens (2010). In total, data were collected from 20 autosomal loci and one chloroplast locus (Table 1). Heterozygous sites were phased using PHASE v2.1 (Stephens, Smith & Donnelly, 2001; Stephens & Donnelly, 2003) with alleles called in the earlier analysis included. Alleles phased at $P = 0.90$ or higher were retained; standard ambiguity codes were used for alleles below this threshold.

Species delimitation using discovery approaches

Two discovery methods were used for species delimitation. We used Gaussian clustering (Hausdorf & Hennig, 2010), a phenetic approach that clusters individuals on the basis of genetic distance. Distance matrices for each locus were estimated in PAUP* v4.0b10 (Swofford, 2002) and corrected using maximum likelihood (ML) and an appropriate model of sequence evolution (see Table 1). These matrices were then combined to construct a multilocus distance matrix using POFAD v1.03 (Joly & Bruneau, 2006), which was converted into similarity vectors using nonmetric multidimensional scaling (Kruskal, 1964). Gaussian clustering was conducted in *R* using two packages, *prabclus* (Hausdorf & Hennig, 2010) and *mclust* (Fraley & Raftery, 2006). We also used Structurama v2.0 (Huelsenbeck & Andolfatto, 2007; Huelsenbeck, Andolfatto & Huelsenbeck, 2011) to discover species limits. This approach utilizes the Bayesian clustering algorithm introduced in the widely used Structure (Pritchard, Stephens & Donnelly, 2000) package, but treats the number of clusters (K) as a random variable. A Dirichlet process prior is used to propose different clustering levels, and thus the probability of the data given the model (i.e. a particular cluster composition) and K is estimated. In order to ensure that the proposal distributions were not influencing the posterior, we conducted multiple analyses varying the probability distribution for the alpha parameter. All analyses were run for 1×10^7 generations and sampled every 1×10^3 generations (with a burn-in of 1000 generations).

Species delimitation using validation approaches

Validation approaches adopt a phylogenetic framework for species delimitation by modelling the data as a phylogenetic tree and evaluating the probability of models that vary the number of lineages (Knowles & Carstens, 2007). Validation approaches require users to partition the samples prior to analysis, and these choices can have important ramifications. For our data, we explored several partitions based on the results of the above discovery analyses and previous research. We tested three levels of partitioning: $K = 2$, as identified using Structurama (below); $K = 3$ (with eastern and western groups and the Kisatchie population separate), as identified by Zellmer *et al*. (2012) using Structure and all 76 of their loci; and $K = 6$, as predicted by the riverine barrier hypothesis proposed by Zellmer *et al*. (2012). In addition, we also considered $K = 10$, which treats all sampling localities as putative lineages to be validated. Although this last partitioning level is probably not biologically reasonable because some sampled populations are separated by less than 20 km, our use here follows Leaché & Fujita (2010), who also considered all sampling localities, and is motivated by a desire to explore the effect of over-splitting the data using validation approaches.

Phylogenetic approaches to species delimitation could potentially include a very large parameter space that could contain all the parameters associated with the estimation of gene trees, the species tree and the species delimitations. As a result of the size of parameter space, the methods used here, spedeSTEM v1.9 beta (Ence & Carstens, 2011) and BPP v2.1 (Yang & Rannala, 2010), adopt different strategies for the simplification of the parameter space of species delimitation. Rather than estimating gene trees, spedeSTEM takes previously estimated gene trees as input, calculates the ML species tree for a given partition of samples using STEM v2.0 (Kubatko, Carstens & Knowles, 2009) and identifies the best of these partitions using information theory (Burnham & Anderson, 2002). Because it computes the likelihood of species trees representing the ML estimate of all possible permutations of lineages, spedeSTEM is robust to phylogenetic error. However, the results are dependent on the quality of the gene tree estimates, and the accuracy of spedeSTEM is expected to decrease as the quality of the gene tree estimates decreases. However, it is both easy to assess the quality of gene tree estimates and clear from simulations that, when spedeSTEM is inaccurate, it fails to delimit what are, in reality, separate lineages (Ence & Carstens, 2011); therefore, we believe that we are unlikely to falsely delimit as independent entities that are actually part of the same evolutionary lineage.

Gene trees were estimated using PAUP* with models of DNA sequence substitution selected using DTmodsel (Minin *et al*., 2003). Our data average 9.95 variable and 7.62 informative sites per locus, which suggests that poor gene tree estimates are likely to be problematic with our data. The nodal support is generally poor in the estimated gene trees, with bootstrap support values of less than 80% in the majority of nodes across gene trees estimated from different loci (not shown). Accordingly, we utilized replicated subsampling in an attempt to circumvent this difficulty, because such subsampling will increases the ratio of variable sites per tip in the gene tree (i.e. by reducing the number of tips), and thus improve gene tree estimates. Previous work has demonstrated that estimates of the species tree are accurate when subsampling is used with STEM – for example, replicated sampling of five alleles per species produces an accurate species tree (Hird, Kubatko & Carstens, 2010). As the number of species is in question for a delimitation analysis, we subsampled 20 alleles from *S. alata* (i.e. two per sampled population) and repeated this analysis 100 times. Log-likelihood scores were averaged across replicates prior to calculation of information theoretic metrics. Subsampling was conducted using python scripts available from: http://carstenslab.org.ohio-state.edu/software.html. As a phylogenetic approach, spedeSTEM requires samples to be divided into a minimum of two lineages to calculate the probability of the species tree | gene tree. We sampled 10 populations, but, in order to evaluate the model in which all *S. alata* are part of the same lineage, an outgroup is required, and we were only able to amplify eight of the loci described above in *Sarracenia rubra*. We therefore conducted spedeSTEM analyses at all levels using this reduced dataset, and at the $K = 3$, $K = 6$ and $K = 10$ levels using all data. For both approaches, STEM was used to calculate the probability of the species tree | gene tree using the gene trees estimated above and assuming $\theta = 0.126$ (Zellmer *et al*., 2012).

In contrast with spedeSTEM, which does not consider error in the gene tree estimates, the Bayesian approach BPP utilizes Markov chain Monte Carlo (MCMC) methods to integrate over the uncertainty in the gene tree parameter space. However, BPP does this at the expense of considering uncertainty in the species tree (i.e. it does not integrate over species tree parameter space). BPP relies on the user to supply a topology to guide the analysis. Reversible jump MCMC is used to evaluate whether a given node in the guide tree should be collapsed or retained. Like most users of BPP, we followed Leaché & Fujita (2010) in using the Bayesian program *BEAST v1.7.4 (Heled & Drummond, 2010) to estimate a species tree for use as a guide tree. All loci were set to a strict clock model using models of DNA sequence substitution described above. Analyses were run for 5×10^8 generations and sampled every 5×10^4 generations (with a burn-in of 1000 trees). Convergence was assessed using Tracer v1.5 (Rambaut & Drummond, 2007), and a maximum clade credibility (MCC) tree was assembled from the posterior distribution and used to guide the analysis. For BPP, priors were informed by results in Zellmer *et al*. (2012), with the gamma prior of θ (population size) set to 2497 and τ_0 (species tree root) set to 2511. Each analysis was replicated to confirm consistent results.

Finally, because the question of the root placement of the phylogeny of populations in *S. alata* has important ramifications for species delimitation, we re-estimated species trees with STEM and *BEAST using the loci for which we had sequences from the related species *S. rubra* (eight loci).

RESULTS

Data acquisition and preliminary analysis

In total, 796 Sanger sequences were added to the existing data, with an additional 50 alleles sequenced from cpDNA (Table S1). Novel data were edited using Sequencher 4.8 (GeneCodes, Ann Arbor, MI, USA), manually aligned in MacClade v4.08 (Maddison & Maddison, 2005) and subsequently phased. The data contain a total of 209 single nucleotide polymorphisms (SNPs) in 7047 bp, or an average of 9.95 variable sites per locus with a length of 335 bp (Table 1).

Species delimitation using discovery approaches

The Gaussian clustering indicates that most samples are members of one of two clusters. Although these clusters are largely associated with the eastern and western populations, there are two samples from the eastern Abita Springs locality that join a cluster with the majority of western samples, and several samples from the western Cooter's Bog and Kisatchie localities that cluster with the majority of eastern samples. In addition, a single sample from Abita Springs does not fit into either of the two large clusters (Table S2). We do not consider these results to be biologically realistic, because it seems unlikely that a single population could harbour representatives of each genetic cluster, and we do not use it as the basis for further analysis. However, the results from the Structurama analysis appear to have clear biological meaning. They support a clustering level of $K = 2$ with PPs > 0.95 across a wide range of proposal distributions (Table 2), and with one cluster composed of all eastern samples and the other of all western samples. We thus base the $K = 2$ validation analyses (below) on the results of Structurama.

Species delimitation using validation approaches

The two validation approaches produced contradictory results. In the spedeSTEM analysis, results consistently found no evidence of cryptic species within *S. alata* (Table 3; Table S3), regardless of whether we used all 21 loci or the eight loci with outgroup sequence. Clearly, these results indicate that *S. alata*, as described, is a single species, and are consistent with the lack of well-supported nodes in gene trees from across these loci. In contrast with the spedeSTEM results, results from the BPP analyses suggest that there are multiple independent lineages within both the eastern and western populations, regardless of the assumed clustering level (Table 3; Fig. 2). However, we are skeptical of these results at the higher clustering levels for two reasons. First, we have considerable uncertainty in the species tree space of our *BEAST analysis, which we discuss below. A related concern at the higher clustering levels is the number of tips in our species tree, which (particularly for the $K = 10$ level) is higher than BPP was intended to analyse (B. Rannala, UC-Davis, pers. comm.), and its accuracy with larger numbers of tips (as in the *S. alata* data) has not been thoroughly explored.

DISCUSSION

As currently described, *S. alata* is probably composed of two cryptic species – one each in the eastern and western portions of its range. The clearest evidence in support of this assertion is the results of the Structurama and BPP analysis, each of which supports these partitions with high PP (Tables 2 and 3). Although we find similarly strong support for additional lineages using BPP, we suspect that these are false delimitations (see below). We do not find evidence for cryptic diversity using spedeSTEM, a result probably caused by the relative paucity of SNPs in our data, which do not allow gene or species trees to be estimated accurately (Fig. S1). The results from the Gaussian clustering are slightly incongruent with the other methods, as several eastern individuals were assigned to the western cluster and vice versa. This finding is consistent with simulation results (Rittmeyer & Austin, 2012), which suggest that the proportion of misassigned individual samples is higher with Gaussian clustering than with Structurama. However, the most curious result from the Gaussian analysis was the division of samples from the Abita Springs population; it seems unlikely that a site of < 305 hectares in size harbours three cryptic species.

In empirical systems, it is possible to be misled by false delimitations as well as by failures to detect cryptic lineages, and congruence across multiple analyses can help to prevent this error (Camargo *et al.*, 2012). Although some analyses delimit more than two lineages, our confidence in a particular

Table 2. Results from Structurama analysis. The marginal likelihoods and model probabilities are shown for several proposal distributions. In the $K = 2$ level, one cluster is composed of eastern samples and the other of western samples

	K	Alpha (1,1)	Alpha (1,5)	Alpha (1,10)	Alpha (0.1,1)	Alpha (0.1,5)	Alpha (0.1,10)
Marginal likelihood		–1139.54	–1147.88	–1139.75	–1142.06	–1141.82	–1143.92
Model probability (# pops)	1	0	0	0	0	0	0
East–west	2	0.95	0.97	0.98	0.98	0.98	0.99
	3	0.05	0.03	0.02	0.02	0.01	0.01
	4	0	0	–	0	0	–

Table 3. Species validation results using two, three and six operational taxonomic units (OTUs). Populations are abbreviated as in the legend of Figure 1. In the rows below, all eastern populations are abbreviated east (east = FDALT), all western populations are abbreviated west (west = CKBSP), eastern Louisiana populations are abbreviated eLA (eLA = ALT) and eastern Texas populations are abbreviated eTX (eTX = BSP). For the spedeSTEM results, the number of nodes, the average Akaike information criterion (avAIC) across 100 replicates, the AIC differences (Δ_i), model likelihoods (w_i) and model likelihoods of the models except the $K = 1$ model (w_i*) are shown. For the BPP results at the same clustering levels, the guide topology and posterior probabilities of each node ('*P*') are shown

spedeSTEM results	*K*	avAIC	Δ_i	w_i	w_i*
***K* = 2**					
(rubra, east+west)	1	4283.27	0	1	
(rubr,(west,east))	2	10384.64	6101.37	0	
***K* = 3**					
(rubr, east+west)	1	4283.27	0	1	n/a
(rubr,(westK,east))	2	10384.64	6101.37	0	1
(rubr,(eastwest,K))	2	11799.06	7515.79	0	0
(rubr,(eastK,west))	2	11890.35	7607.08	0	0
(rubr,(K,(west,east)))	3	11900.71	7617.44	0	0
***K* = 6**					
(rubra, east+west)	1	4283.27	0	1	n/a
(rubr,(CK eTX, eLA FD))	2	10384.64	6101.37	0	1
(rubr,(eLA F,(CK eTX,D)))	3	11889.44	7606.17	0	0
(rubr,(C eTX,(eLA FD,K)))	3	11900.69	7617.42	0	0
(rubr,(F,(CK eTX, eLA D)))	3	11901.11	7617.84	0	0
(rubr,(K eTX,(eLA FD,C)))	3	11901.19	7617.92	0	0
(rubr,(eLA,(CK eTX, FD)))	3	11902.71	7619.44	0	0
(rubr,(eTX,(CK, eLA FD)))	3	11907.97	7624.7	0	0
(rubr,(F,(eLA,(CK eTX,D))))	4	11910.04	7626.77	0	0
(rubr,(C,(K,(eTX, eLA FD))))	4	11915.08	7631.81	0	0
(rubr,(C eTX,(eLA F,(K,D))))	4	11917.26	7633.99	0	0
(rubr,(K eTX,(eLA F,(C,D))))	4	11917.83	7634.56	0	0
(rubr,(eTX,(eLA F,(CK,D))))	4	11924.78	7641.51	0	0
(rubr,(C eTX,(F,(K, eLA D))))	4	11929.62	7646.36	0	0
(rubr,(F,(K eTX,(eLA D,C))))	4	11930.08	7646.81	0	0
(rubr,(C eTX,(eLA,(K,FD))))	4	11931.48	7648.21	0	0
(rubr,(C,(K,(eLA F,(eTX,D)))))	5	11932.03	7648.76	0	0
(rubr,(K eTX,(eLA,(C,FD))))	4	11932.03	7648.76	0	0
(rubr,(eTX,(F,(CK, eLA D))))	4	11937.32	7654.05	0	0
(rubr,(C eTX,(F,(eLA,(K,D)))))	5	11938.94	7655.67	0	0
(rubr,(eTX,(eLA,(CK,FD))))	4	11939.34	7656.07	0	0
(rubr,(K eTX,(F,(eLA,(C,D)))))	5	11939.48	7656.21	0	0
(rubr,(C,(K,(F,(eTX, eLA D)))))	5	11944.67	7661.4	0	0
(rubr,(C,(K,(eLA,(eTX,FD)))))	5	11946.74	7663.47	0	0
(rubr,(eTX,(F,(eLA,(CK,D)))))	5	11946.86	7663.59	0	0
(rubr,(C,(K,(F,(eLA,(eTX,D))))))	6	11954.3	7671.03	0	0

BPP results:

***K* = 2** – (east, west) '*P* = 1.0'.

***K* = 3** – (K, west) '*P* = 0.992',east) '*P* = 1.0'.

***K* = 6** – ((eTX, C) '*P* = 0.999', K) '*P* = 1.0', ((eLA, D) '*P* = 1.0', F) '*P* = 1.0') '*P* = 1.0'.

delimitation is proportional to the congruence in the results across methods. Based on the findings of the four methods utilized here, we favour the delimitation of the eastern and western populations, because three of the four generally support this interpretation, and because the lack of resolution in the gene trees offers a good explanation for the conflicting results in spedeSTEM. However, on the basis of such findings,

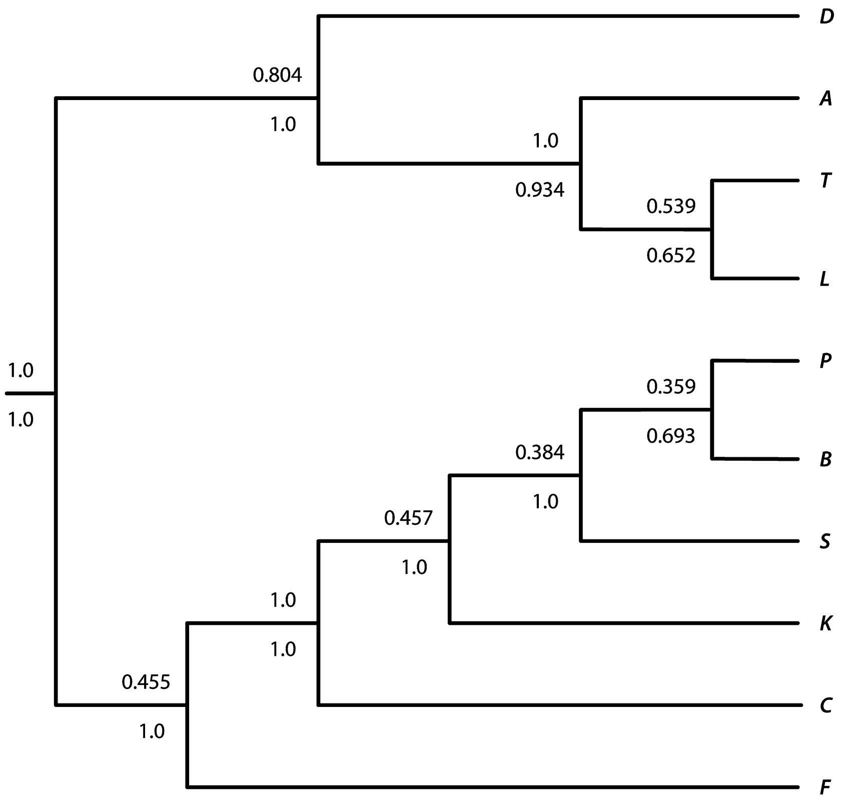

Figure 2. Results from *BEAST analysis are presented as a maximum clade credibility tree. Numbers above the nodes represent the posterior probability of that node, and numbers below the nodes represent the posterior probability of a BPP run using this tree as a guide tree. Midpoint rooting is used. Please see the legend of Figure 1 for a key to locality abbreviations. Note that the placement of the Franklin population (F) as sister to the clade composed of BSPCK is probably an artefact of the midpoint rooting, as this population is included in the DALT clade when an eight-locus dataset is analysed with an outgroup sequence (see Fig. S2).

the best course of action is less than clear. Other recent investigations have had a variety of responses to the detection of cryptic diversity using methods for species delimitation. Although some have made clear taxonomic recommendations (e.g. Burbrink *et al.*, 2011; Niemiller, Near & Fitzpatrick, 2012) or have described new species (Leaché & Fujita, 2010), others have treated the delimited lineages as evolutionary significant units (ESUs) (Barrett & Freudenstein, 2011) or have not made a formal recommendation (Setiadi *et al.*, 2011; Camargo *et al.*, 2012; Zhou *et al.*, 2012). In the case of *S. alata*, the taxonomic implications are clear: as the type locality of *S. alata* is in Mississippi, we recommend the elevation of the western populations to species status, and plan to formally describe this species in a forthcoming article. We operate here under the assumption of the general lineage concept of species (de Queiroz, 2005), and interpret the results from the genetic analysis as evidence of evolutionary independence between the eastern and western populations.

Phylogenetic uncertainty in the species tree and BPP

Although coalescent stochasticity in the gene trees is accounted for by BPP, we are uncomfortable with the use of a single guide tree for the delimitation analysis, given the phylogenetic uncertainty in the topology of the species tree. Failing to consider this uncertainty could lead to overly liberal delimitation. Consider our analysis at the $K = 10$ level, which treats each sampled population as a putative lineage. When *BEAST is used to estimate a guide tree, a naïve interpretation of the BPP results would indicate at least five independent lineages (Fig. 2). However, it is clear that the guide tree has low nodal support and, when we selected 20 trees at random from the posterior distribution of the *BEAST analysis (post burn-in), and used each as a guide tree in a subsequent BPP analysis, support (as summarized by the computation of the posterior probability of nodes represented across these draws from the region of highly probable species tree space) was generally low for the delimitations shown in Figure 2 (see Table S4). Our point is not to argue that BPP is inaccurate because we proposed a greater than reasonable number of putative lineages. However, the choice of which partitions to validate with BPP (or any validation method) is critical, precisely because previous work has demonstrated that BPP is prone to over-splitting if the guide tree is misspecified (Leaché & Fujita, 2010). In this way, spedeSTEM is a vital component of our study, because it does not require a guide tree, but only validates the partitions of the data. spedeSTEM finds that there is much higher support given the data for the level at which all *S. alata* populations are treated as the same lineage, regardless of the number of a priori lineages, because there are few fixed differences between any populations, and this pattern in the data allows for a single modelled divergence time to fit the data well (i.e. $-\ln L$ of the two-species model is much greater than that of any model with more than two species). However, if we focus on the information theoretic rankings of the remaining models, we find clear support for the east–west delimitation regardless of the a priori partitioning level (Table 3). To reiterate, despite the limitations of our data (and the generally poor estimates of the gene trees), the second-best model in all analyses (and regardless of an a priori assumption of two, three, six or ten putative lineages) is the model that is consistent with the east–west disjunction and the Structurama results. spedeSTEM is an important tool for species validation, and should be paired with BPP, because these methods have complementary strengths and weaknesses.

The topological uncertainty seen in the STEM phylogeny estimates (Fig. S1) and the posterior distribution of the *BEAST runs begs the question of whether the species tree model is appropriate for our data. Systems to which species delimitation methods are applied are inherently emergent phylogenetic systems; as such, the empirical data from such systems approach the lower limits of appropriateness for a phylogenetic approach. In this sense, the ambiguous results seen in the BPP analyses are not surprising. This method has been interpreted as being more powerful at detecting recently diverged lineages than other methods (e.g. Camargo *et al*., 2012), and performs well in simulation studies (Yang & Rannala, 2010; Camargo *et al*., 2012). However, it relies on an accurate guide tree, an accuracy that is itself dependent on the correct identification of the species boundaries (Carstens & Dewey, 2010; O'Meara, 2010). Given the uncertainty in species tree estimation, and the fact that BPP does not integrate over this tree space, we suggest that BPP can potentially be misleading and could reasonably lead us to over-estimate species diversity if it is used to validate data that are overly divided. A conservative approach should be adopted for the description of cryptic lineage diversity. We are cautious when interpreting our BPP results, especially when analyses are conducted using a single guide tree where incorrect phylogenetic relationships can drive false positives in species delimitation.

CONCLUSIONS

Next-generation sequencing technology has led to the detection of cryptic lineage diversity within *S. alata*, and the populations west of the Mississippi River should be elevated to species status. These results have been tested and confirmed through multiple approaches to species delimitation using multilocus genetic data, including the exploration of both discovery and validation approaches. Although we generally agree with Fujita *et al*. (2012), who argue that coalescent-based species delimitation should be central to integrative taxonomy, species delimitation is best conducted using *both* discovery and validation approaches, with congruence across multiple methods necessary for recognizing the presence of cryptic lineage diversity. However, the challenge occurs in systems such as *S. alata*, in which genetic evidence supports cryptic diversity, but other sources of data do not. Given the influx of genomic data available for systematists, the number of cryptic species discovered will probably continue to increase.

DATA ACCESSIBILITY

All sequence alignments have been deposited in GenBank under accession numbers KC835522–KC835594. Data deposited in the Dryad repository: doi: 10.5061/dryad.f51bb (Carstens & Satler, 2013).

ACKNOWLEDGEMENTS

We thank Danielle Fusilier, Margaret Hanes, Sarah Hird and Amanda Zellmer for their collection of samples and contributions to the *S. alata* system. Funding was provided by the National Science Foundation (DEB-0918212; DEB-0956069) and the Louisiana Board of Regents. We thank Tara Pelletier for reading several drafts of the manuscript, and members of the Weisrock laboratory at the University of Kentucky and participants in the NimBios Species Delimitation Working Group for discussions related to this research. We thank John Allen, Adam Leaché and two anonymous reviewers for comments that improved the manuscript.

REFERENCES

Avise JC. 2000. *Phylogeography: the history and formation of species*. Cambridge, MA: Harvard University Press.

Barrett CF, Freudenstein JV. 2011. An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. *Molecular Ecology* **20:** 2771–2786.

Burbrink FT, Yao H, Ingrasci M, Bryson JRW, Guiher TJ, Ruane S. 2011. Speciation at the mogollon rim in the Arizona mountain kingsnake (*Lampropeltis pyromelana*). *Molecular Phylogenetics & Evolution* **60:** 445–454.

Burnham KP, Anderson DA. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edn. New York: Springer-Verlag.

Camargo A, Morando M, Avila LJ, Sites JW. 2012. Species delimitation with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the *Liolaemus darwinii* complex (Squamata: Liolaemidae). *Evolution* **66:** 2834–2849.

Carstens BC, Dewey TA. 2010. Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. *Systematic Biology* **59:** 400–414.

Carstens BC, Satler JD. 2013. Data from: The carnivorous plant described as *Sarracenia alata* contains two cryptic species. *Dryad Digital Repository* doi:10.5061/dryad.f51bb

Ence DD, Carstens BC. 2011. SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources* **11:** 473–480.

Fraley C, Raftery AE. 2006. MCLUST Version 3 for R: Normal mixture modeling and model-based clustering. Technical Report no. 504, Department of Statistics, University of Washington.

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.

Hausdorf B, Hennig C. 2010. Species delimitation using dominant and codominant multilocus markers. *Systematic Biology* **59:** 491–503.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology & Evolution* **27:** 570–580.

Hird S, Kubatko L, Carstens BC. 2010. Replicated subsampling enables accurate species tree estimation in empirical systems. *Molecular Phylogenetics & Evolution* **57:** 888–898.

Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. *Genetics* **175:** 1787–1802.

Huelsenbeck JP, Andolfatto P, Huelsenbeck ET. 2011. Structurama: Bayesian inference of population structure. *Evolutionary Bioinformatics* **7:** 55–59.

Jackson ND, Austin CC. 2010. The combined effects of rivers and refugia generate extreme cryptic fragmentation within the common ground skink (*Scincella lateralis*). *Evolution* **64:** 409–428.

Joly S, Bruneau A. 2006. Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from Rosa in North America. *Systematic Biology* **55:** 623–636.

Knowles LL, Carstens BC. 2007. Delimiting species without monophyletic gene trees. *Systematic Biology* **56:** 887–895.

Koopman MM, Carstens BC. 2010. Inferring population structure and demographic parameters across a riverine barrier in the carnivorous plant *Sarracenia alata* (Sarraceniaceae). *Conservation Genetics* **11:** 2027–2038.

Kruskal JB. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29:** 1–27.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* **25:** 971–973.

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

Maddison DR, Maddison WP. 2005. *MacClade 4: analysis of phylogeny and character evolution Version 408a*. Available at: http://macclade.org (accessed November 2012).

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology* **52:** 674–680.

Neyland R. 2008. Intraspecific systematic relationships of *Sarracenia alata* Wood (Sarraceniaceae) inferred from nuclear ribosomal DNA sequences. *Journal of the Mississippi Academy of Sciences* **53:** 238–245.

Niemiller ML, Near TJ, Fitzpatrick BM. 2012. Delimiting species using multilocus data: diagnosing cryptic diversity in the southern cavefish *Typhlichthys subterraneus* (Teleostei: Amblyopsidae). *Evolution* **66:** 846–866.

O'Meara BC. 2010. New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology* **59:** 59–73.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* **55:** 595–609.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

de Queiroz K. 2005. Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 6600–6607.

Rambaut A, Drummond AJ. 2007. *Tracer v1.5*. Available at: http://beast.bio.ed.ac.uk/Tracer (accessed November 2012).

Rittmeyer EN, Austin CC. 2012. The effects of sampling on delimiting species from multi-locus sequence data. *Molecular Phylogenetics & Evolution* **65:** 451–463.

Setiadi MI, McGuire JA, Brown RM, Zubairi M, Iskandar DT, Andayani N, Supriatna J, Evans B. 2011. Adaptive radiation and ecological opportunity in Sulawesi and Philippine fanged frog (Limnonectes) communities. *American Naturalist* **178:** 221–240.

Sheridan PM. 1991. What is the identity of the West Gulf Coastal pitcher plant *Sarracenia alata*? *Carnivorous Plant News* **20:** 102–110.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS. 2006. Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15:** 4261–4293.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73:** 1162–1169.

Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68:** 978–989.

Swofford DL. 2002. *PAUP**: phylogenetic analysis using parsimony (*and other methods)*. Version 4. Sunderland, MA: Sinauer Associates.

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

Zellmer AJ, Hanes MM, Hird S, Carstens BC. 2012. Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. *Systematic Biology* **61:** 763–777.

Zhou W-W, Wen Y, Fu J, Xu Y-B JJ-Q, Ding L, Min M-S, Che J, Zhang Y. 2012. Speciation in the *Rana chensinensis* species complex and its relationship to the uplift of the Qinghai–Tibetan Plateau. *Molecular Ecology* **21:** 960–973.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Maximum likelihood estimate of the phylogeny from the STEM analysis. A consensus tree from 100 replicates shows one clade consisting of all eastern populations present in 84% of the replicates, and no other clades are present in more than 50% of the replicates. Please see the legend of Table S1 for population abbreviations.

Figure S2. *BEAST Markov chain Monte Carlo (MCMC) phylogeny for *Sarracenia alata* using *Sarracenia rubra* as an outgroup. Eight loci were used to estimate this tree. Please see the legend of Table S1 for population abbreviations.

Table S1. For each of 21 loci, the number of samples sequenced is shown. Populations are denoted with abbreviations as follows: Abita Springs (A), Buton Lake (B), Cooter's Bog (C), DeSoto (D), Franklin (F), Lake Ramsay (L), Pitcher Trail (P), Sundew Trail (S), Talisheek (T).

Table S2. Results from Gaussian clustering. The assignments of individuals to clusters are shown. Samples are grouped into eastern and western populations, and clusters are shaded for ease of viewing. Please see the legend of Table S1 for a key to locality abbreviations.

Table S3. spedeSTEM results from *Sarracenia alata* with the analysis of all 21 loci. The results across 100 replicates for the $K = 10$ level are shown, with eastern and western populations divided. From left to right, columns show the model, the average $-\ln L$ across replicates, the average Akaike information criterion (AIC), AIC differences and model probabilities. Also shown are the results for the $K = 6$ level. Please see the legend of Table S1 for a key to locality abbreviations.

Table S4. Summary of results from BPP analysis which considered the uncertainty in species tree space. The posterior probabilities of delimitation are shown averaged over 20 trees selected at random from the posterior distribution. Please see the legend of Table S1 for population abbreviations.