MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 473–480 doi: 10.1111/j.1755-0998.2010.02947.x

TECHNICAL ADVANCES

# SpedeSTEM: a rapid and accurate method for species delimitation

DANIEL D. ENCE and BRYAN C. CARSTENS
*Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*

## Abstract

**We describe a software package (SpedeSTEM) that allows researchers to conduct a species delimitation analysis using intraspecific genetic data. Our method operates under the assumption that a priori information regarding group membership is available, for example that samples are drawn from some number of described subspecies, races or distinct morphotypes. SpedeSTEM proceeds by calculating the maximum likelihood species tree from all hierarchical arrangements of the sampled alleles and uses information theory to quantify the model probability of each permutation. SpedeSTEM is tested here against empirical and simulated data; results indicate that evolutionary lineages that diverged as few as 0.5N generations in the past can be validated as distinct using sequence data from little as five loci. This work enables speciation investigations to identify lineages that are evolutionarily distinct and thus have the potential to form new species *before* these lineages acquire secondary characteristics such as reproductive isolation or morphological differentiation that are commonly used to define species.**

*Keywords*: coalescent, STEM, species delimitation, species trees

*Received 17 June 2010; revision received 23 September 2010; 18 October 2010; accepted 22 October 2010*

## Introduction

### *Species delimitation*

Over the last decade, the widespread acquisition of genetic data at the interface between populations and species has led to the development of several methods for species delimitation (Sites & Marshall 2004; Wiens 2007). As befitting a discipline that operates at the interface between population and phylogenetics, many of these methods incorporate a coalescent model within a phylogenetic framework (Pons *et al.* 2006; Knowles & Carstens 2007; O'Meara 2010; Yang & Rannala 2010). Approaches to species delimitation can be broadly separated out into two groups (species discovery, species validation) on the basis of whether the samples are partitioned prior to the analysis. Species discovery attempts to partition the samples into species, absent any a priori information regarding species membership (e.g., O'Meara 2010). This can be valuable in the study of systems that have not been subject to prior investigation, but does not take advantage of existing data in many well-studied groups (for example, described races or subspecies). On the other hand, it is often the case that certain lines of evidence (geographical, molecular, behavioural, etc.) support the grouping of certain samples to the exclusion of others. In these instances, it is of considerable utility to quantify the evidential support for these partitions. This validation is ultimately crucial for species tree inference, as correct assignment of samples to species is a basic assumption that is made by all phylogenetic methods (Knowles & Carstens 2007; O'Meara 2010), but may be difficult in the face of cryptic speciation, ambiguous samples from a hybrid zone, or any number of other naturally occurring scenarios.

Gene trees provide evidence vital to understanding the process of speciation because they span intraspecific and interspecific evolution (Harrison 1998), a connection that is most obvious in the ancestral-descendant relationships of alleles within phylogenies and populations (Hey 1994; Templeton 1994). Consequently, gene trees from neutral loci have several advantages for species delimitation. Because the pattern of allele coalescence is stochastic and can be defined in a probabilistic manner (Tajima 1983; Takahata & Nei 1985; Hudson 1991), the rate that ancestral polymorphism is lost in a given lineage provides valuable information concerning the temporal divergence between sister lineages (Rosenberg 2002; Hudson & Turelli 2003). This divergence may be evidence that the sister lineages have not exchanged

Correspondence: Bryan C. Carstens, Fax: (1)225 578 2597;
E-mail: carstens@lsu.edu

migrants and thus can serve as the basis for lineage delimitation. Unfortunately, gene trees do not track the pattern of lineages splitting and divergence as well as one might like (Hudson & Coyne 2002), and there are certain combinations of species tree branch lengths where the most common gene tree is incongruent with the species tree (Degnan & Rosenberg 2006; Rosenberg & Tao 2008). Distressingly, these combinations are not particularly extreme and could reasonably occur in empirical systems. Because concatenation across loci is not a reliable approach to species phylogeny inference (Mossel & Vigoda 2005; Kolaczkowski & Thornton 2006), the incorporation of coalescent models into phylogenetic methods is an important recent development in systematics (Page & Sullivan 2008).

### *Problem definition*

Phylogenetic inference near the species level should incorporate aspects of population genetic theory because the genetic forces that act within a population, such as genetic drift, selection and migration, may each play an important role in speciation (Maddison & Knowles 2006). Several recent approaches to phylogeny estimation operate under the assumption that genetic drift has produced the incongruence between gene trees and species trees (Maddison & Maddison 2004; Liu & Pearl 2006; Ané *et al.* 2007; Edwards *et al.* 2007; Oliver 2008; Kubatko *et al.* 2009; Heled & Drummond 2010). Relative to species delimitation, the most important aspect of these approaches is the shift in which entities are used as the operational taxonomic units (OTUs) in the phylogenetic analysis (Carstens & Dewey 2010). Rather than using single or multiple representative samples as exemplars, methods for species tree inference overtly treat species or population lineages as the OTUs and include multiple samples within each lineage. Thus, adoption of the species tree paradigm allows the relationships among species (or population) lineages to be inferred directly while also allowing the validity of sample assignment to be explored.

The method described here incorporates STEM (Kubatko *et al.* 2009), an analytical technique that calculates the maximum likelihood species tree from a sample of observed gene trees under the assumption that the discord between the topology of gene trees and the species phylogeny is produced by the coalescent process alone. Because the phylogeny produced by STEM is an analytical solution that maximizes the probability of the gene trees given the species tree (with branch length), the only type of phylogenetic uncertainty that can influence the estimation of the species tree is the uncertainty in the inference of the gene trees themselves; this uncertainty can be assessed using conventional approaches, such as nodal support (Felsenstein 1985) or parametric bootstrapping (Goldman 1993).

### *An information-theoretic approach to species delimitation*

Our method proceeds by recognizing that there are often several putative 'groups' within a described species, for example subspecies, distinct populations or morphotypes. It allows users to divide their data into partitions that correspond to these groups, so that the independence of the evolutionary lineages represented by these partitions can be evaluated. Partition is used here in the mathematical sense of the term as the set of nonempty, exhaustive and mutually exclusive subsets of a set of elements. The number of possible partitions in a set of N elements is equal to the Bell number for N. The sequence of Bell numbers grows at a greater than exponential rate, presenting a practical and computational limitation (O'Meara 2010). For this reason, we restrict the calculations to all hierarchical permutations, those within but not between species. For example, if a user defines three partitions within species A and 4 within species B, SpedeSTEM will calculate the likelihood of the species tree given the gene trees for the product of the Bell numbers of 3 and 4 (i.e., $5 \times 15 = 75$) rather than the Bell number of 7 (i.e., 877). Note that the hierarchical nature of the validation approach does not force the species to be monophyletic when the species tree is computed by STEM. Once the user assigns samples to some number of partitions, gene trees are estimated from each partition and stored in a format compatible with STEM. SpedeSTEM calculates the likelihood of the species tree given the set of gene trees for all hierarchical permutations of the data and then computes a series of metrics based on information theory. Essentially, each permutation is represented as a model of lineage composition. The Akaike Information Criteria (AIC) (Akaike 1973) of each arrangement are calculated, as well as the AIC differences ($\Delta_i$) and model probabilities ($w_i$), which describes the Kullback–Leibler (K-L) distance (Kullback & Liebler 1951) of model $i$ to the best model and the probability that model $i$ is the best model, respectively (Anderson 2008).

### *Subsampling*

Phylogeographic studies require large sample sizes to accurately identify population substructure (Avise 2001), and it is not uncommon for sequence data from hundreds (Wares & Cunningham 2001; Zamudio & Savage 2003) or even thousands (Bernatchez 2001) of samples to be collected. While multilocus phylogeographic investigations often sample fewer individuals to sequence more genes, the resulting data set can still be quite large. For example,

sample sizes of 6–10 genes and 50–150 samples are common (Dolman & Moritz 2006; Geraldes *et al.* 2008), but some investigations sample hundreds of individuals (Garrick *et al.* 2008; Peters *et al.* 2008) while others sample many more loci (Lee & Edwards 2008; Moeller & Tiffin 2008). The large sample sizes of phylogeographic data present two difficulties to species tree inference methods: the computational time required for gene and species tree estimation increase with sample size, and it is often the case that there are missing data from some individuals at some loci. In anticipation of these difficulties, SpedeSTEM allows users to utilize replicated subsampling to estimate the species trees (Hird *et al.* 2010). This approach draws a small number of alleles from each putative lineage, estimates gene trees from the reduced data and then uses the reduced gene trees to estimate the species tree. The process can be repeated any number of times; Hird *et al.* (2010) demonstrated that replicated subsampling produced accurate estimates of the species tree using a few as 3–5 alleles per species.

## Methods

### *Program implementation*

SpedeSTEM is a Java application that makes use the *BioJava* library (Holland *et al.* 2008) to represent nexus files as Java objects. It also utilizes an R-script that requires the *ape* library (Paradis *et al.* 2004) for R and requires UNIX or LINUX executables of STEM (Kubatko *et al.* 2009). If the user wishes to automate the estimation of the gene trees, PAUP* (Swofford 2002) is also required. This is particularly desirable when the subsampling option is invoked.

The user supplies the following set of files:

1. A set of single-locus DNA sequence alignments in nexus format (i.e., one for each locus).
2. A locus-information file containing a model of sequence evolution and a scaling factor indicating the mode of inheritance for each locus.
3. A group-information file that records the membership of alleles to groups as well as the subsampling proportion for each group.

Additional parameters supplied by the user include the following:

- The number of subsampling replicates to be run (optional).
- A value for $\theta = 4N_e\mu$ to be used by STEM for all loci.
- A flag to indicate the search strategy for gene-tree estimation in PAUP* (optional; subtree pruning-regrafting, nearest-neighbour interchange, tree bisection-reconnection, or Branch and Bound can be used)
- A path to a folder that will hold all the intermediate and result files created by SpedeSTEM.

If subsampling is specified, SpedeSTEM begins by creating a series of nexus files from the subsampled alleles. The algorithm then conducts gene tree estimation using PAUP* (we used Portable version 4.0b10 for Unix). Note that any gene tree estimation program that takes nexus files as input can be used (e.g., GARLI; Zwickl 2006), provided that the estimated gene trees meet the requirements of STEM (ultrametric, consistent with the molecular clock, midpoint-rooted). As part of the process of generating the gene-tree file, any polytomous nodes in the gene trees are resolved using the *ape* library for R. Next, the algorithm generates gene tree and settings files for STEM, estimates the $-ln\text{L}$ (model | data) for each permutation using STEM and collects the output in a 'STEM output' folder. If indicated in the command line, SpedeSTEM will also generate all hierarchical permutations of lineage grouping using the information provided by the user in the group-information file. The results across permutations are collected in a file called 'AICMatrix', also written to the 'STEM output' folder. This file allows the user to examine the relative evidential support for competing hypotheses using information theory (Anderson 2008).

### *Program testing*

*Empirical data.* To evaluate the performance of SpedeSTEM across a range of scenarios, including the difficulties that routinely arise in empirical data sets, we evaluated SpedeSTEM using an empirical data sets from *Myotis* bats (Carstens & Dewey 2010) as well as data simulated under a variety of demographical scenarios. The empirical *Myotis* data consist of six loci (phased) collected from 37 individuals; these data are sampled from eleven described subspecies belonging to four species, and an outgroup (*Myotis volans*). We conducted the species delimitation analysis by subsampling three alleles from each of four described subspecies within *Myotis lucifugus* and two within *Myotis evotis*. Because of a limited number of individuals from *Myotis thysanodes*, we did not partition this species into its three described subspecies.

*Simulation studies—data simulation.* MS (Hudson 2002) and SEQ-GEN (Rambaut & Grassly 1997) were used to generate data under four broad scenarios (Fig. 1), hereafter referred to as 'treatments'. For Treatment I, the depth of node 1 ranged from 0.25N to 6N. For Treatment II, the depths of nodes 1 and node 2 ranged from 0.25N to 4N. Treatment III and treatment IV correspond to Treatments I and II, respectively, with the addition of reciprocal migration between taxa A and B at relatively low ($m = 0.01$) and moderate ($m = 0.1$) levels. One hundred simulated data sets for each combination of parameter settings were generated. Twenty loci were generated for

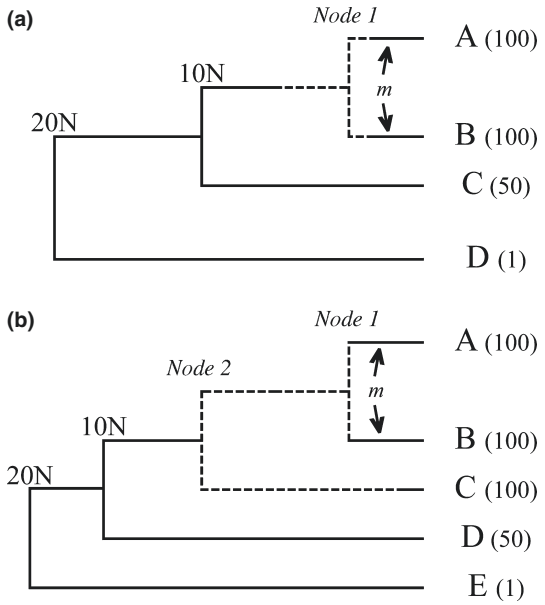


**Fig. 1** Models used in data simulation. Data were simulated under the model of lineage divergence shown in (a). for Treatments I and III. The depth of node 1 was varied between 0.25N and 6N, with the depths of the other nodes held constant. The numbers in parentheses represent the number of simulated alleles for each lineage, and the m with arrows mark the lineages that exchange alleles via gene flow in Treatment III. The model shown in (b). was used for Treatments II and IV. Node 1 was varied between 0.25N and 4N, and node 2 ranged from 0.5N to 8N.

each replicate. Sequences were generated with 50 variable sites under the HKY model with base pair frequencies of 0.3, 0.2, 0.2, 0.3, a Ti/Tv ratio of 3.0, and $\theta = 4N_e\mu = 10$.

*Program execution.* The requisite loci-information and group-information files were generated using Perl scripts, with all loci treated as autosomal (i.e., STEM scaling factor = 1.0). Lineages A and B were grouped together for treatments I and III, and lineages A, B and C were treated as three groups within one species for treatments II and IV. Replicated subsampling (100 replicates; 5 alleles for all lineages except the outgroup) was used. The gene trees were estimated using PAUP* (Swofford 2002) with the TBR search strategy. To examine the effect of sampling more loci on the results, SpedeSTEM was executed using a randomly selected five, ten, fifteen and twenty of the simulated loci.

## Results and discussion

### *Empirical data*

Unlike traditional approaches to statistics, which aim to test hypotheses (i.e., to reject or fail to reject the null), the information-theoretic approach utilized by SpedeSTEM aims to quantify the evidential support for a set of models given the data. For analyses that do not include subsampling, the calculations of AIC, Akaike differences ($\Delta_i$) and model probabilities ($w_i$) follow Anderson (2008). These metrics can be used to rank the hypotheses and to quantify the evidential support for each; the $w_i$ can be interpreted as the probability that a given model is the shortest K-L distance of those contained within the set of models to the 'true' model. For analyses that implement subsampling, we advocate averaging the $-lnL$ (model | data) across replicates with a subsequent calculation of the information-theoretic metrics from these averaged values.

An example can be seen in the *Myotis* data (Table 1). The AIC score of the model that treats each of the subspecies within *M. lucifugus* and *M. evotis* as an independent lineage encompasses over 97% of the total model probability, indicating that this model is closer to the true model than the other models in the comparison (Anderson 2008). These results are consistent with those of Carstens & Dewey (2010) and easily interpreted. Notably, the replicated subsampling decreases the computational effort required to conduct the delimitation analysis; 100 replicates drawing three alleles from each of eight *Myotis* lineages required ~120 min on a laptop with an older MacBook Pro (2.4 GHz Intel Core 2 Duo processor).

### *Simulation study*

The implementation of the information-theoretic metrics prevents us from reducing the results from the simulation study to a table of *P*-values. To simplify discussion, we measure accuracy using the median value (across replicates) of the model probabilities of the correct model of lineage composition. This value approaches 1.0 as $-lnL$ ($model_{true}$ | data) improves.

Treatment I represents the simplest possible scenario, where the validity of a single pair of putative lineages (A & B) is evaluated. In this case, the accuracy of SpedeSTEM is initially high, even at shallow (i.e., 0.25N) levels of divergence, and improves as the depth of the node increases and as data (in the form of loci) are added. Across all treatments, the correct model is always supported by the data, in most of these cases the support is overwhelming (Fig. 2; Supplemental Table S1). When complexity is added and three lineages are included (Treatment II), the accuracy of SpedeSTEM decreases slightly (Supplemental Table S2). For example, species are easily delimited given a node depth of 2N generations for two diverging lineages, but when a third lineage is added at a slightly older node depth (i.e., 2N for node I, 0.5N for node 2), the accuracy decreases from a median model weight of 0.997 (five loci) to a median weight of

**Table 1** Information-theoretic metrics from reanalysis of the *Myotis* empirical data. Results from 100 replicates (2–97 omitted for sake of brevity) based on a partitioning of the described species *Myotis lucifugus* into four subspecies (*alascensis* = ala; *carissima* = car, *relictus* = rel; *lucifugus* = luc) and the described species *Myotis evotis* into two subspecies (*chrysonotus* = chr; *jonesorum* = jon)

| Replicate | -*ln*L | -*ln*L | -*ln*L | -*ln*L | -*ln*L | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Permutation | 0 | 1 | 98 | 99 | (ave) | k | AIC | $\Delta_i$ | Model likelihood | $w_i$ |
| ala, car, rel, luc, chr, jon, thy, ke, vo | 138.60 | 144.65 | 142.87 | 145.41 | 143.22 | 8 | 302.435 | 0.000 | 1.000000 | 0.970 |
| ala, car_luc, rel, chr, jon, thy, ke, vo | 145.49 | 151.48 | 149.91 | 150.85 | 149.33 | 7 | 312.670 | 10.235 | 0.005992 | 0.006 |
| ala, car_rel, luc, chr, jon, thy, ke, vo | 145.48 | 151.54 | 149.76 | 149.47 | 149.52 | 7 | 313.043 | 10.608 | 0.004971 | 0.005 |
| ala_luc, car, rel, chr, jon, thy, ke, vo | 145.49 | 151.63 | 149.97 | 150.82 | 149.70 | 7 | 313.395 | 10.960 | 0.004169 | 0.004 |
| ala_rel, car, luc, chr, jon, thy, ke, vo | 145.48 | 151.42 | 149.76 | 150.83 | 149.76 | 7 | 313.510 | 11.075 | 0.003936 | 0.004 |
| ala_car, rel, luc, chr, jon, thy, ke, vo | 145.60 | 150.09 | 149.85 | 148.20 | 149.83 | 7 | 313.651 | 11.216 | 0.003669 | 0.004 |
| ala, car, rel_luc, chr, jon, thy, ke, vo | 144.15 | 151.42 | 149.78 | 152.37 | 149.87 | 7 | 313.737 | 11.302 | 0.003514 | 0.003 |
| ala, car, rel, luc, chr_jon, thy, ke, vo | 145.59 | 151.65 | 149.86 | 152.40 | 149.93 | 7 | 313.855 | 11.420 | 0.003313 | 0.003 |
| ala_rel, car_luc, chr, jon, thy, ke, vo | 148.94 | 156.21 | 154.66 | 154.36 | 153.81 | 6 | 319.614 | 17.179 | 0.000186 | 0.000 |
| ala_luc, car_rel, chr, jon, thy, ke, vo | 148.94 | 156.53 | 154.84 | 152.97 | 153.97 | 6 | 319.934 | 17.499 | 0.000159 | 0.000 |
| ala_rel_luc, car, chr, jon, thy, ke, vo | 148.94 | 154.96 | 154.79 | 154.30 | 153.99 | 6 | 319.989 | 17.554 | 0.000154 | 0.000 |
| ala, car_luc, rel, chr_jon, thy, ke, vo | 150.47 | 156.48 | 154.90 | 155.84 | 154.04 | 6 | 320.084 | 17.649 | 0.000147 | 0.000 |
| ala, car_rel, luc, chr_jon, thy, ke, vo | 150.47 | 156.54 | 154.75 | 154.46 | 154.23 | 6 | 320.463 | 18.028 | 0.000122 | 0.000 |
| ala_luc, car, rel, chr_jon, thy, ke, vo | 150.47 | 156.63 | 154.96 | 155.81 | 154.41 | 6 | 320.810 | 18.375 | 0.000102 | 0.000 |
| ala_rel, car, luc, chr_jon, thy, ke, vo | 150.47 | 156.42 | 154.75 | 155.82 | 154.46 | 6 | 320.926 | 18.491 | 0.000097 | 0.000 |
| ala_car, rel_luc, chr, jon, thy, ke, vo | 149.15 | 154.82 | 154.76 | 153.16 | 154.47 | 6 | 320.945 | 18.510 | 0.000096 | 0.000 |
| ala_car, rel, luc, chr_jon, thy, ke, vo | 150.59 | 155.09 | 154.84 | 153.19 | 154.53 | 6 | 321.068 | 18.633 | 0.000090 | 0.000 |
| ala, car_rel_luc, chr, jon, thy, ke, vo | 148.94 | 156.20 | 154.64 | 152.96 | 153.55 | 7 | 321.098 | 18.663 | 0.000089 | 0.000 |
| ala, car, rel_luc, chr_jon, thy, ke, vo | 149.14 | 156.42 | 154.77 | 157.36 | 154.58 | 6 | 321.153 | 18.718 | 0.000086 | 0.000 |
| ala_car_luc, rel, chr, jon, thy, ke, vo | 150.49 | 155.07 | 154.87 | 151.59 | 153.93 | 7 | 321.853 | 19.418 | 0.000061 | 0.000 |
| ala_car_rel, luc, chr, jon, thy, ke, vo | 150.48 | 154.81 | 154.73 | 150.21 | 154.13 | 7 | 322.265 | 19.830 | 0.000049 | 0.000 |
| ala, car_rel_luc, chr_jon, thy, ke, vo | 153.92 | 161.20 | 159.63 | 157.95 | 158.26 | 5 | 326.512 | 24.077 | 0.000006 | 0.000 |
| ala_car_luc, rel, chr_jon, thy, ke, vo | 155.47 | 160.07 | 159.86 | 156.58 | 158.63 | 5 | 327.264 | 24.829 | 0.000004 | 0.000 |
| ala_luc, car_rel, chr_jon, thy, ke, vo | 153.92 | 161.53 | 159.83 | 157.96 | 158.67 | 5 | 327.347 | 24.912 | 0.000004 | 0.000 |
| ala_rel_luc, car, chr_jon, thy, ke, vo | 153.92 | 159.96 | 159.78 | 159.29 | 158.70 | 5 | 327.397 | 24.962 | 0.000004 | 0.000 |
| ala_car_rel_luc, chr, jon, thy, ke, vo | 153.94 | 159.77 | 159.59 | 153.65 | 158.12 | 6 | 328.248 | 25.813 | 0.000002 | 0.000 |
| ala_car, rel_luc, chr_jon, thy, ke, vo | 154.14 | 159.82 | 159.75 | 158.15 | 159.18 | 5 | 328.361 | 25.926 | 0.000002 | 0.000 |
| ala_car_rel, luc, chr_jon, thy, ke, vo | 155.47 | 159.81 | 159.72 | 155.20 | 158.84 | 6 | 329.684 | 27.249 | 0.000001 | 0.000 |
| ala_rel, car_luc, chr_jon, thy, ke, vo | 153.92 | 161.21 | 159.65 | 159.35 | 158.51 | 7 | 331.026 | 28.591 | 0.000001 | 0.000 |
| ala_car_rel_luc, chr_jon, thy, ke, vo | 158.92 | 164.77 | 164.58 | 158.64 | 162.83 | 5 | 335.657 | 33.222 | 0.000000 | 0.000 |
| | | | | | | | | Σ | 1.031024326 | |

Results indicate strong support (e.g., $w_i$ values) for the model that treats each of these subspecies as a distinct evolutionary lineage. The other models collapse two or more subspecies into a single lineage (indicated by the '_' between abbreviations), have far lower probabilities and thus lower Akaike Information Criteria (AIC) scores indicating that they exist at a greater distance to the true model.

0.888 (five loci). Unsurprisingly, the validity of shallower nodes (i.e., more recent cladogenesis) is the most difficult to establish. These results illustrate the complexity involved with delimiting species that are members of multispecies radiations; we attribute much of this difficulty to the presence of unsorted ancestral polymorphism in each of the three descendent lineages (Fig. 1). It should be noted that it is possible to validate more than two species even among lineages that have rapidly radiated, although more data are required. For example, three lineages can be validated as distinct even if they formed as recently as 0.75N generations in the past, so long as 20 loci are used. As in Treatment 1, the performance of SpedeSTEM improves as data are added and the node depth increases.

Because gene flow can lead to a decrease in accuracy for species tree estimation (Eckert & Carstens 2008), we also simulated data under the topologies used in the first two treatments with low ($m = 0.01$) and moderate ($m = 0.1$) amounts of gene flow. When gene flow occurs between two diverging sister lineages (i.e., Treatment III), the accuracy of SpedeSTEM is decreased (Supplemental Table S3), but the method is still quite capable of validating lineages that are independent across all depths of divergence. At shallow nodes, however, as the rate of gene flow increases, more data are required to successfully validate two lineages as distinct. For example, when lineages separated by 0.25N divergence exchange and $m = 0.1$, 10 loci are required before the median $w_i$ value surpasses 0.95. The decreased accuracy is most easily

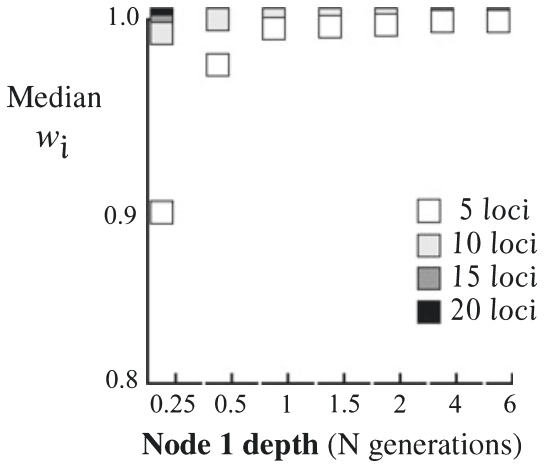


**Fig. 2** Results from Treatment I. The median model probability ($w_i$) across 100 replicates is reported at seven levels of divergence for 5, 10, 15 and 20 loci.

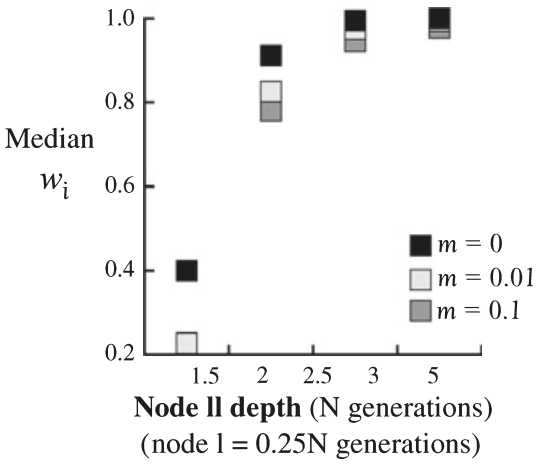


**Fig. 3** Results from Treatment IV. The median model probability ($w_i$) across 100 replicates is reported at five levels of divergence for differing levels of gene flow. In all cases, data from five loci were used.

noticed at deeper levels of divergence, consistent with the general finding that ongoing gene flow can inhibit divergence (Wright 1931). Gene flow has the largest impact on the accuracy of SpedeSTEM when there are three lineages to delimit (i.e., Treatment IV), particularly at low levels of divergence and moderate levels of gene flow. (Fig. 3). Gene flow inhibits the ability of SpedeSTEM to validate independent evolutionary lineages in a manner that is proportional level of gene flow, as well as the depth of divergence (Supplemental Table S4). For both low and moderate amounts of gene flow, validation becomes more accurate as data in the form of additional loci are added (Fig. 4). However, accuracy values that do not consistently surpass 0.9 until 15 loci are analysed (Supplemental Table S4).

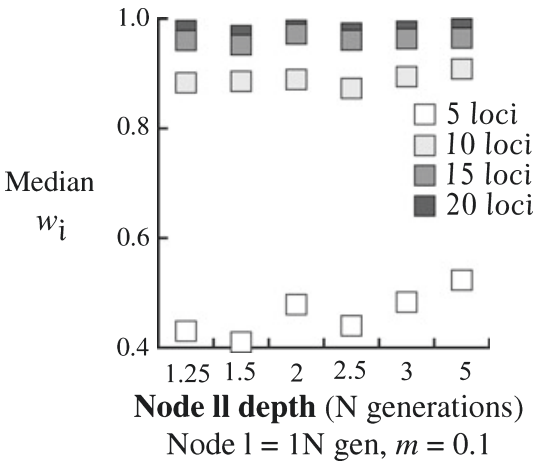


**Fig. 4** Results from Treatment IV with moderate levels of gene flow (m = 0.1). The median model probability ($w_i$) across 100 replicates is reported at five levels of divergence for differing amounts of data.

## Conclusion

This study introduces a novel approach to lineage delimitation (SpedeSTEM) that enables biologists to identify distinct evolutionary lineages shortly after they form. Our motivation for this software is illustrated by the example from the *Myotis* example; in this case as in numerous others there are described subspecies that are essentially hypothesized evolutionary lineages. SpedeSTEM allows biologists to evaluate whether the pattern of genetic variation exhibited by the lineages that compose such systems are consistent with a model of a single or of multiple species. In this manner, we envision SpedeSTEM as a tool that will aid taxonomists in their endeavour to identify and describe species.

SpedeSTEM employs replicated subsampling (Hird *et al.* 2010), an approach that allows large phylogeographic data sets to be evaluated. By increasing the rigor with which phylogenetic reasoning can be applied to the analysis of population genetic variation, and by facilitating the incorporation of population level processes into phylogenetic analysis, SpedeSTEM offers dramatic practical benefits to biologists investigating empirical systems at the interface between population genetics and systematics, the very region where speciation occurs.

Our results demonstrate that the number of lineages and the amount of migration are both factors that influence the accuracy of SpedeSTEM. For example, when there are only two putative lineages, SpedeSTEM is able to validate lineages as independent using ≥10 or more loci, even when the divergence between these lineages is

0.25N generations (Fig. 2). At this level, ancestral polymorphism will be broadly distributed across sister lineages ((Hudson 2002), and thus, a visual inspection of the gene trees would not indicate that cryptic lineages are present. While these results are promising, factors such as gene flow and multiple lineages will negatively influence the accuracy of SpedeSTEM. For example, in the case where the depth of Node I is 1N generations and gene flow is moderate ($m$ = 0.1), SpedeSTEM only consistently surpasses the 0.9 threshold for accuracy when 15 or more loci are used (Fig. 3). The important factor in the accuracy of SpedeSTEM is the depth of Node II; when this node is short, the ancestral polymorphism is more or less evenly distributed among the three lineages, making it difficult to extract signal, but as this depth increases the accuracy improves. These findings are consistent regardless of the level of migration (Fig. 4).

The above results demonstrate that empiricists will be able to identify cryptic lineages in their focal taxa with a modest amount of data, an ability that will allow them to more easily identify the environmental and landscape forces that produce lineage diverge (and that may lead to speciation). Because lineage divergence can be detected at an earlier stage, these events will have occurred more recently in the past and thus will be more generally detectable.

## Acknowledgements

Portions of this research were conducted with high performance computational resources provided by the Louisiana Optical Network Initiative (http://www.loni.org). Funding was provided by a grant from the National Science Foundation to B. Carstens (DEB: 0918212). We thank AE Gaggiotti, three anonymous reviewers for comments that vastly improved this manuscript. We thank S. Hird, M. Koopman, J. McVay, T. Pelletier, N. Reid, Y. E. Tsai, and A. Zellmer for discussion and ideas related to this work.

## References

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory*. (eds. Petrov BN & Csaki F), pp. 267–281. Akademiai Kiado, Budapest.

Anderson DR (2008) *Model Based Inference in the Life Sciences*. Springer, New York.

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, **24**, 412–426.

Avise JC (2001) *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, MA.

Bernatchez L (2001) The evolutionary history of brown trout (*Salmo trutta* L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation. *Evolution*, **55**, 351–379.

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.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, 0762.

Dolman G, Moritz C (2006) A multilocus perspective on refugial isolation and divergence in rainforest skinks (*Carlia*). *Evolution*, **60**, 573–582.

Eckert AJ, Carstens BC (2008) Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics and Evolution*, **49**, 832–842.

Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the USA*, **104**, 5936–5941.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Garrick RG, Rowell DM, Simmons CS, Hillis DM, Sunnucks P (2008) Fine-scale phylogeographic congruence despite demographic incongruence in two low-mobility saproxylic springtails. *Evolution*, **62**, 1103–1118.

Geraldes A, Basset P, Gibson B *et al.* (2008) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology*, **17**, 5349–5363.

Goldman NJ (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, **36**, 182–198.

Harrison RG (1998) *Linking Evolutionary Pattern and Process: The Relevance of Species Concepts for the Study of Speciation*. Oxford University Press, New York, NY.

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

Hey J (1994) Bridging phylogenetics and population genetics with gene tree models. In: *Molecular Ecology and Evolution: Approaches and Applications*. (eds. Schierwater B, Streit B, Wagner GP & DeSalle R), pp. 435–449. Birdhauser Verlag, Basel.

Hird S, Kubatko LS, Carstens BC (2010) Rapid and accurate species tree estimation for phylogeographic investigation using replicated subsampling. *Molecular Phylogenetics & Evolution*, **57**, 888–898.

Holland RCG, Down TA, Pocock M *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

Hudson RR (1991) Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (eds Futuyma D & Antonovics J), pp. 1–44. Oxford University Press, New York.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution*, **56**, 1557–1565.

Hudson RR, Turelli M (2003) Stochasticity overrules the ''three-times rule'': genetic drift, genetic draft, and coalescent times for nuclear loci versus mitochondrial DNA. *Evolution*, **57**, 182–190.

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

Kolaczkowski B, Thornton JW (2006) Is there a star tree paradox? *Molecular Biology and Evolution*, **23**, 1819–1823.

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

Kullback S, Liebler RA (1951) On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Lee JY, Edwards SV (2008) Divergence across Australia's Carpentarian Barrier: statistical phylogeography of the Red-backed Fairy Wren (*Malurus melanocephalus*). *Evolution*, **62**, 3117–3134.

Liu L, Pearl DK (2006) Species trees from gene trees: reconstructing Bayesian posterior distributions of species phylogeny using estimated gene tree distributions. In: *Mathematical Biosciences Institue Technical Report*, pp. 1–24. Ohio State University, Columbus, OH.

Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55**, 21–30.

Maddison WP, Maddison DR (2004) Mesquite: A modular system for evolutionary analysis. Version 1.01. Available at http://mesquiteproject.org.

Moeller DA, Tiffin P (2008) Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. *Evolution*, **62**, 3069–3081.

Mossel E, Vigoda E (2005) Phylogenetic MCMC algorithmns are misleading on mixture of trees. *Science*, **309**, 2207–2209.

Oliver JC (2008) AUGIST: inferring species trees while accounting for gene tree uncertainty. *Bioinformatics*, **24**, 2932.

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

Page RDM, Sullivan J (2008) The expanding contributions of systematic biology. *Systematic Biology*, **57**, 1–3.

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

Peters JL, Zhuravlev YN, Fefelov I, Humphries EM, Omland KE (2008) Multilocus phylogeography of a holarctic duck: colonization of North America from Eurasia by Gadwall (*Anas strepera*). *Evolution*, **62**, 1469–1483.

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.

Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computational Applications to Biosciences*, **13**, 235–238.

Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, **61**, 225–247.

Rosenberg NA, Tao R (2008) Discordance of species trees with their most likely gene trees: the case of five taxa. *Systematic Biology*, **57**, 131–140.

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

Swofford DL (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (and other methods). Version 4*. Sinnauer Associates, Sunderland, MA.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, **110**, 325–344.

Templeton AR (1994) The role of molecular genetics in speciation studies. In: *Molecular Ecology and Evolution: Approaches and Applications* (eds Schierwater B, Streit B, Wagner GP & DeSalle R), pp. 455–477, Birkhauser-Verlag, Basel.

Wares JP, Cunningham CW (2001) Phylogeography and historical ecology of the North Atlantic intertidal. *Evolution*, **55**, 2455–2469.

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

Wright SG (1931) Evolution in Mendelian populations. *Genetics*, **16**, 114–138.

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

Zamudio KR, Savage WK (2003) Historical isolation, range expansion, and secondary contact of two highly divergent mitochondrial lineages in spotted salamanders (*Ambystoma maculatum*). *Evolution*, **57**, 1631–1652.

Zwickl DJ (2006) *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion*. University of Texas, Austin.

## Supporting Information

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

**Table S1** Treatment I: For each node depth (in N generations), results show the median model weight ($w_i$) across 100 replicates, as well as the wi at the lower and upper quartiles.

**Table S2** Treatment II: For each node depth (in N generations), results show the median model weight ($w_i$) across 100 replicates, as well as the wi at the lower and upper quartiles.

**Table S3** Treatment III: For each node depth (shown in N generations) and migration rate, results show the median model weight ($w_i$) across 100 replicates, as well as the wi at the lower and upper quartiles.

**Table S4** Treatment IV: For each node depth (shown in N generations) and migration rate, results show the median model weight ($w_i$) across 100 replicates, as well as the wi at the lower and upper quartiles.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.