Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The effects of sampling on delimiting species from multi-locus sequence data

Eric N. Rittmeyer *, Christopher C. Austin

Department of Biological Sciences, Museum of Natural Science, 119 Foster Hall, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 6 January 2012
Revised 4 June 2012
Accepted 30 June 2012
Available online 20 July 2012

Keywords:
Species delimitation
Sampling strategy
Structurama
Nonparametric delimitation
Gaussian clustering

ABSTRACT

As a fundamental unit in biology, species are used in a wide variety of studies, and their delimitation impacts every subfield of the life sciences. Thus, it is of utmost importance that species are delimited in an accurate and biologically meaningful way. However, due to morphologically similar, cryptic species, and processes such as incomplete lineage sorting, this is far from a trivial task. Here, we examine the accuracy and sensitivity to sampling strategy of three recently developed methods that aim to delimit species from multi-locus DNA sequence data without *a priori* assignments of samples to putative species. Specifically, we simulate data at two species tree depths and a variety of sampling strategies ranging from five alleles per species and five loci to 20 alleles per species and 100 loci to test (1) Structurama, (2) Gaussian clustering, and (3) nonparametric delimitation. We find that Structurama accurately delimits even relatively recently diverged (greater than 1.5 N generations) species when sampling 10 or more loci. We also find that Gaussian clustering delimits more deeply divergent species (greater than 2.5 N generations) relatively well, but is not sufficiently sensitive to delimit more recently diverged species. Finally, we find that nonparametric delimitation performs well with 25 or more loci if gene trees are known without error, but performs poorly with estimated gene genealogies, frequently over-splitting species and mis-assigning samples. We thus suggest that Structurama represents a powerful tool for use in species delimitation. It should be noted, however, that intraspecific population structure may be delimited using this or any of the methods tested herein. We argue that other methods, such as other species delimitation methods requiring *a priori* putative species assignments (e.g. SpeDeSTEM, Bayesian species delimitation), and other types of data (e.g. morphological, ecological, behavioral) be incorporated in conjunction with these methods in studies attempting to delimit species.

1. Introduction

Species are a fundamental unit in biology important to every subfield in biology, and the inaccurate delimitation of species can compromise the integrity, relevance, and conclusions of research (Bickford et al., 2007; Bortolus, 2008; Coyne and Orr, 2004; de Queiroz, 2007). Despite the importance of species and of delimiting species in a biologically meaningful manner, species concepts remain a controversial topic subject to extensive debate (Coyne and Orr, 2004; de Queiroz, 2007). A variety of criteria, including reproductive isolation (biological species concept; Mayr, 1942, 1995), reciprocal monophyly (genealogical species concept; Baum and Donoghue, 1995; Baum and Shaw, 1995), and diagnostic characters (phylogenetic species concept; Cracraft, 1989), among others, have been proposed for delimiting species; however, it is unlikely for many of these criteria to evolve instantaneously with speciation and the order in which they evolve is likely to vary among systems (de Queiroz, 2007). Arguably the most inclusive species concept is the unified species concept, which defines species as independently evolving metapopulation lineages, and argues that many species concepts represent criteria that evolve as lineages diverge and that may be used to help delimit species, rather than definitions of species (de Queiroz, 2007). Regardless of the specific species concept used, errors in species delimitation may come in three forms: over-splitting (i.e. a single species is treated as multiple species), over-lumping (i.e. multiple species are treated as a single species), or incorrect assignment of individuals or populations (i.e. samples of one species is treated as a member of a different, though valid, species). Over-splitting of species inflates measures of biodiversity, potentially biasing harvest or conservation strategies (Bickford et al., 2007). Over-splitting can also result in underestimates of intraspecific variation and viability, and overestimates of interspecific gene flow (Funk and Omland, 2003). Over-lumping of species can cause the opposite problems, and potentially the failure to recognize and protect species of conservation concern (Bickford et al., 2007). Depending on the sampling strategy and questions investigated, all of these problems may also arise as a result of the incorrect assignment of populations to species. With increased use of molecular markers and

* Corresponding author.
E-mail address: erittm1@lsu.edu (E.N. Rittmeyer).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.06.031

increased sophistication of analyses for molecular data, it is becoming apparent that many traditionally recognized species, particularly those with broad geographic distributions, are actually complexes of multiple species with little or no gene flow among them, often recently diverged and morphologically conservative (Bickford et al., 2007). Thus, of the three delimitation errors, recent research suggests that over-lumping is of major concern to biodiversity research (Bickford et al., 2007).

Further, recently developed species tree inference methods (e.g. *BEAST, Heled and Drummond, 2010; BEST, Liu, 2008; minimize deep coalescences, Maddison, 1997; STEM Kubatko et al., 2009) attempt to identify the underlying species-level phylogeny while accounting for heterogeneity among gene genealogies due to incomplete lineage sorting (Edwards et al., 2007). However, these methods assume accurate delimitations of species *a priori*, and errors in these assignments are likely to result in unreliable species tree estimates, particularly if mis-assigned or over-lumped samples involve non-sister species.

Several methods that attempt to delimit species from molecular data rely on fixed divergence thresholds (Hebert et al., 2004; Lefébure et al., 2006) or reciprocal monophyly (Sites and Marshall, 2004). Many such methods, such as generalized mixed Yule coalescent model (Monaghan et al., 2009; Pons et al., 2006) and statistical parsimony networks (Clement et al., 2000; Templeton et al., 1992), also use only single markers. Selection of a threshold of divergence for species delimitation is highly subjective, and a single threshold is unlikely to be appropriate for all systems (Knowles and Carstens, 2007; Moritz and Cicero, 2004). Additionally, while reciprocal monophyly may be useful for identifying species with older divergences, it may take a substantial amount of time for lineages to sort to reciprocal monophyly, particularly at multiple loci and in species with large effective population sizes (Degnan and Rosenberg, 2006, 2009). Thus, reciprocal monophyly is highly conservative and likely to fail to identify recently diverged species (Hudson and Coyne, 2002; Hudson and Turelli, 2003; Knowles and Carstens, 2007). Further, processes such as incomplete lineage sorting can result in a single marker not accurately representing the species phylogeny and species boundaries, particularly for recently diverged species, and in rapid radiations where the interval between speciation events is short (Degnan and Rosenberg, 2006, 2009). Therefore, a more accurate method of delimiting species from molecular data would incorporate multiple loci and account for the stochasticity of the coalescent process.

While some powerful methods are available that delimit species from multi-locus data under a coalescent framework (e.g. SpedeSTEM, Ence and Carstens, 2011; Bayesian species delimitation, Yang and Rannala, 2010), these and other methods require *a priori* assignment of samples to putative species. These species validation methods are not be appropriate in all situations; even in well-studied systems, processes such as convergent evolution or morphological conservatism may make it impossible to accurately and objectively assign all populations to putative species (e.g. *Pantherophis obsoletus* complex, Burbrink et al., 2000; *Sceloporus undulatus* complex, Leaché, 2009; Leaché and Reeder, 2002; *Carlia fusca* group, Austin et al., 2011). In such situations, errors in assignment would likely result in errors in species delimitation, potentially resulting in misleading inferences.

Several methods are available for delimiting species from multilocus molecular data without *a priori* putative species assignments. Of particular promise are Structurama (Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011), Gaussian clustering (Hausdorf and Hennig, 2010), and nonparametric delimitation (O'Meara, 2010). One additional method for delimiting species from multi-locus molecular data without *a priori* assignments is fields for recombination (FFR), which attempts to delimit species from non-overlapping sets of heterozygous individuals (Doyle, 1995; Sites and Marshall, 2003). However, this method performed extremely poorly in a previous test, correctly assigning less than 27% of individuals (Hausdorf and Hennig, 2010). We thus exclude FFR from our analyses.

Structurama was developed for the purpose of detecting intraspecific population structure from genetic data (Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011) by combining the Bayesian clustering algorithm implemented in Structure (Falush et al., 2003; Pritchard et al., 2000) with a Dirichlet process prior that allows the number of populations to be treated as a random variable (Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011). The algorithm thus aims to estimate both the number of populations and the composition of these populations by minimizing linkage disequilibrium and maximizing Hardy–Weinberg equilibrium (Falush et al., 2003; Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011; Pritchard et al., 2000). The processes driving neutral genetic differentiation among species are similar to those driving neutral differentiation among intraspecific populations (i.e. genetic drift coupled with restricted gene flow); thus Structurama may also be useful for species delimitation. Indeed, the method has previously been shown to be informative for this application (Hausdorf and Hennig, 2010; Leaché and Fujita, 2010; Pinzón and LaJeunesse, 2011; Salicini et al., 2011). While both Structure and Structurama utilize the same algorithm, Structure assumes a fixed number of populations (K), whereas Structurama treats the number of populations as a random variable estimated via a Dirichlet process prior. Although metrics have been proposed to estimate the most appropriate K using Structure (Evanno et al., 2005), estimating K remains difficult and may be somewhat ambiguous (Hausdorf and Hennig, 2010), and confidence in the values of K generated by these methods cannot be statistically assessed (Evanno et al., 2005; Huelsenbeck et al., 2011; Pritchard et al., 2000). Thus, it seems unlikely that Structure would significantly outperform Structurama, and we here focus on testing Structurama and do not include Structure in this study.

Gaussian clustering was first applied to the problem of species delimitation by Hausdorf and Hennig (2010) for use with dominant and co-dominant allelic data (e.g., AFLPs, microsatellites) by using multidimensional scaling to convert a genetic distance matrix to a series of similarity vectors, from which clusters (i.e., species) are estimated. In the previous implementation of this method for species delimitation, it performed relatively well (Hausdorf and Hennig, 2010), correctly assigning 73–93% of individuals. We include this method in this study to more thoroughly test its accuracy using multilocus DNA sequence data.

Nonparametric delimitation and KC delimitation are two additional approaches that attempt to jointly estimate species assignments and species trees without *a priori* data on putative species (O'Meara, 2010). Unlike the methods described above, which use either a distance matrix or genetic data directly in a non-genealogical context, both of these methods are topology-based; that is, these methods attempt to delimit species from a set of gene genealogies. Nonparametric delimitation attempts to identify the species tree and the species delimitations that minimize both excess structure within species and the number of deep coalescent events among species. As this method uses gene genealogies as input, errors in species delimitation may result from one of two sources: errors due to gene tree uncertainty, and errors due to the algorithm itself. Therefore, to both test the empirical utility of the method and to tease apart the sources of error, we apply nonparametric delimitation both on estimated gene trees and on simulated coalescent gene trees.

While KC delimitation is a theoretically intriguing method that attempts to identify the species delimitations and species tree that maximize the probability of a set of gene genealogies, the method is extremely computationally intensive, and is thus unfeasible for use with datasets larger than a few samples or loci (O'Meara, 2010). Further, in previous tests of the method, KC delimitation performed poorly (O'Meara, 2010), possibly due to an inability to efficiently search the parameter space. Thus, we focus instead on nonparametric delimitation and do not further test KC delimitation in this study.

While all the methods discussed above have been applied to the problem of species delimitation, their accuracy and sensitivity to sampling intensity has not been thoroughly examined. We use simulated datasets at a variety of sampling intensities to assess the performance of a variety of species delimitation methods and to investigate their robustness to a range of sampling strategies. Specifically, we focus on testing Structurama (Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011), Gaussian clustering (Hausdorf and Hennig, 2010), and nonparametric delimitation (O'Meara, 2010) due to their fulfillment of two primary criteria: (1) they can be applied to multi-locus data, and (2) they do not require *a priori* assignments to putative species.

2. Methods

2.1. Data simulations

To represent groups in which lineage sorting is expected to be complete for many loci between the most divergent species, as well as groups in which lineage sorting is expected to be incomplete at many loci among all species, we simulated data for two different levels of divergence or tree depths: 6 N and 12 N generations, where N is the effective population size. The mean time for lineage sorting to complete for a given locus is 4 N ± 2 N generations (Degnan and Rosenberg, 2006, 2009); thus, the shallower trees (6 N) represent the mean time to lineage sorting + one standard deviation, and some lineage sorting would be expected among all species, whereas for the deeper trees, lineage sorting should be complete for the majority of loci across the deeper divergences. For each of the two total tree depths, 100 species trees, each with five species, were simulated under a uniform Yule speciation model in Mesquite v.2.73 (Maddison and Maddison., 2010). Simulated species trees are provided as Supplementary material. For each species tree, 100 gene genealogies were simulated with 20 tips per species (i.e. 100 operational taxonomic units total) and θ equal to 0.01 in ms (Hudson, 2002). In our simulations, we assume no migration following speciation, thus the species simulated represent reproductively isolated species (biological species concept, (Mayr, 1942, 1995)). DNA sequence data, 500 bp in length for each gene, were then simulated on each gene genealogy in Seq-Gen (Rambaut and Grassly, 1997) under an HKY + G model with a transition–transversion ratio of 3.0, base frequencies as 0.3 A, 0.2 C, 0.3 T, 0.2 G, and a discrete gamma distribution with a shape parameter α of 0.8, as in (McCormack et al., 2009). We refer to these simulated sequences as alleles, regardless of whether each is unique, such that alleles refers to the number of sampled sequences, and datasets with 20 alleles sample per species may (and typically do) include less than 20 unique sequences. To test the sensitivity of each method of species delimitation to sampling effort, we randomly reduced the number of alleles and loci sampled to obtain 18 total datasets per species tree: 5, 10 or 20 alleles per species sampled at 5, 10, 25, 50, 75 or 100 loci. While we do not explicitly assign alleles to diploid individuals in this study, the methods tested herein do not take intra-individual variation (i.e. heterozygosity) into account in delimiting species, thus our results are still applicable to diploid or higher ploidy level organisms. To prevent biases due to particularly informative loci or alleles, random reductions were performed such that all samples included in the smaller datasets were included in all larger datasets (i.e. datasets were nested).

To examine the extent of incomplete lineage sorting in the simulated data, genealogical sorting indices (gsi, Cummings et al., 2008) were calculated for each species based on the true coalescent gene genealogies using the genealogicalSorting package (Cummings et al., 2008; Bazinet et al., unpublished) in R v. 2.14.1. The gsi quantifies the amount of common ancestry of a group of operational taxonomic units (OTUs) on a phylogenetic tree, and varies from 0 to 1, where larger values represent more complete lineage sorting, up to a maximum value of 1.0 for a monophyletic group (i.e. complete lineage sorting). To examine the extent of variation within the simulated sequence data, we calculated the number of unique alleles for each locus both for each species and for each species tree (i.e. combining the five simulated species). We similarly calculated the number of segregating sites for each locus both for each species and for each species tree. The numbers of unique alleles were calculated using the pegas package (Paradis, 2010) in R; the numbers of segregating sites were calculated using the ape package (Paradis et al., 2004) in R.

2.2. Species delimitation using structurama

Structurama (Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011) assumes loci are unlinked allelic markers and thus requires multi-locus sequence data to be converted to alleles (coded by integers). We used SNAP Map (Aylor et al., 2006; Price and Carbone, 2005) to convert each locus to numbered alleles. The number of populations (*K*) was set as a random variable to implement the Dirichlet process prior; the prior distribution on the number of populations was set as a gamma distribution with a shape of 1.0 and a scale of 1.0. Markov chains were each run for one million iterations, sampling every 100 iterations; the first 1000 samples (10%) were discarded as burn-in. To ensure consistency of the results, a subset of 360 analyses (10 at each sampling intensity and tree depth) were repeated.

2.3. Species delimitation using Gaussian clustering

Genetic distance matrices for each locus were calculated using maximum likelihood as implemented in PAUP* ver. 4.0b10 (Swofford, 2003) and the model of sequence evolution under which the data was simulated (HKY + G). Single locus distance matrices were then combined using standardized distances to create a multi-locus distance matrix in pofad ver. 1.03 (Joly and Bruneau, 2006). This method scales distance matrices for each locus by the largest distance at that locus to prevent highly variable loci from having an excessive impact on the combined distance matrix. Gaussian clustering was then implemented in R v. 2.12.0 using the prabclus (Hausdorf and Hennig, 2010) and mclust (Fraley and Raftery, 2006) packages. Kruskal's non-metric multidimensional scaling (Kruskal, 1964) was used to convert the multi-locus distance matrix into similarity vectors, with a tuning constant of 4 (as suggested by Hausdorf and Hennig (2010) for identifying clusters containing a minimum of five individuals). Nearest neighbor-based noise detection was used with a tuning constant equal to the smallest integer greater than or equal to the number of samples divided by 40, as suggested by Hausdorf and Hennig (2010). Gaussian clustering was then implemented for all clustering models implemented in mclust; the best-fit model was selected using the Bayesian information criterion. To ensure consistency of the results, a subset of 360 analyses (10 at each sampling intensity and tree depth) were repeated.

2.4. Species delimitation using nonparametric delimitation

Because nonparametric delimitation (NP) is a topology-based species delimitation method, we first estimated gene genealogies for each locus using maximum likelihood in RAxML ver. 7.0.3 (Stamatakis, 2007). The model of sequence evolution was set to GTR + G, as the simpler, HKY + G model under which the data were simulated cannot be implemented in RAxML. Three search replicates were conducted for each locus, and the tree with the highest log likelihood was retained for subsequent analyses.

While nonparametric delimitation based on these estimated gene genealogies would be more comparable to empirical implementations of the method, it would remain unclear if errors in species delimitation based on these estimated genealogies were due to errors in gene genealogy estimation or to poor performance of the nonparametric delimitation method. Thus, to control for errors in species delimitation due to gene tree uncertainty, we also implemented nonparametric delimitation using the true coalescent gene genealogies from which the sequence data were simulated.

Nonparametric delimitation was implemented with both estimated (NP.E) and coalescent (NP.C) gene genealogies in Brownie ver. 2.1.3 (O'Meara, 2010) under default parameters of a structure weight of 0.5, and a *P* threshold of 1.0. All nonparametrc delimitation analyses consisted of five search replicates to ensure the best solution had been found. Although nonparametric delimitation jointly estimates the species tree and species delimitations, we here focus only on the accuracy of the species assignments, as the accuracy of a number of species tree estimation methods have previously been examined elsewhere (Heled and Drummond, 2010; Leaché and Rannala, 2011; Linnen and Farrell, 2008; McCormack et al., 2009).

2.5. Statistical tests

We calculated the accuracy of each delimitation method for every sampling strategy by calculating the percent of the samples correctly assigned to species, and averaging these values across all species trees at each sampling intensity for each of the two total species tree depths. Thus in a case with 100 alleles, if two species are lumped into a single species, and no samples are mis-assigned to a different species, the accuracy would be 0.8 (80/100), but the proportion of incorrectly assigned samples would be 0 (0/100). Similarly, in a case with 50 alleles, if two species are lumped into a single species, and two samples from a third species are lumped within this single lumped species, the accuracy would be 0.76 (38/50), but the proportion inaccurate would be 0.04 (2/50). To further examine the specific sources of error in species delimitations, we also calculated the number of over-split species, the number of over-lumped species, and the proportion of incorrectly assigned samples. We considered species as over-split if greater 20% of the alleles (i.e. at least two for tests with five alleles sampled, at least three for tests with 10 alleles sampled, or at least five for tests with 20 alleles sampled) were assigned to each of two distinct species. Similarly, we considered species as over-lumped if greater than 20% of the alleles from two different species were assigned to the same species. We calculated the proportion of incorrectly assigned alleles as the proportion of alleles assigned to a cluster along with only 20% or less of the conspecific alleles,

To test for significant differences among methods on the accuracy of species delimitations, as well as to determine specific impact of sampling intensity of the accuracy of species delimitation, we conducted pairwise *t*-tests in R ver. 2.14.1. *P*-values were adjusted for multiple comparisons via Bonferroni correction, that is, by multiplying the *p*-values by the number of comparisons: two for examining the impact of number of alleles sampled (i.e. increases from 5 to 10 alleles and from 10 to 20 alleles), five for examining the impact of the number of loci sampled (i.e. increases from 5 to 10 loci, 10 to 25 loci, 25 to 50 loci, 50 to 75 loci, and 75 to 100 loci).

3. Results

Incomplete lineage sorting was extensive in the simulated datasets, as expected given the depth of the simulated species trees. The shallower, 6 N species trees had an average gsi of 0.808 (±0.192 standard deviation, SD). Further, an average 16.4% of simulated loci had gsi less than 1.0 for all five species (as expected given that most simulated divergences were more recent than 6 N generations, and many were much more recent), and only 1.4% of the simulated loci were monophyletic (i.e. gsi = 1.0) for all five simulated species. Each locus showed complete lineage sorting for an average of 1.6 (±1.1 SD) species for the shallower species trees. Lineage sorting was much more complete for the deeper, 12 N species trees, though still prevalent. The average gsi for these trees was 0.877 (±0.162 SD). Only an average of 4.1% of simulated loci had gsi of less than 1.0 for all five species, and lineage sorting was complete (i.e. gsi = 1.0) for all five species for 5.4% of the simulated loci in the deeper species trees. Lineage sorting had completed for an average of 2.4 (±1.2 SD) species on the deeper species trees.

The simulated sequence data included an average of 39.9 (±7.8 SD) segregating sites per locus for the complete datasets (i.e. 100 alleles) for the shallower species trees, but an average of only 8.7 (±4.2 SD) segregating sites within each simulated species. The total number of unique alleles for the complete datasets averaged 26.7 (±4.5 SD), whereas within each species, the average number of unique alleles was 6.0 (±1.8 SD). The simulated sequence data for the deeper species trees averaged 50.6 (±8.6 SD) segregating sites per locus for the complete datasets, but only 8.7 (±4.2 SD) average segregating sites per locus within each simulated species. The complete datasets included an average of 27.9 (±4.3 SD) unique alleles for the deeper species trees, and each species contained, on average, 6.0 (±1.8 SD) unique alleles.

Structurama and NP.C performed significantly better ($p < 0.001$) than other tested methods under nearly all sampling strategies, and under both tree depths (Fig. 1), though Structurama only moderately outperformed NP.E under the lowest sampling intensity (5 alleles, 5 loci) for the shallower trees ($p = 0.047$). Three exceptions to this are the largest datasets (20 alleles, 100 loci) for the deeper tree (in which NP.C was not significantly better than Gaussian clustering, $p = 0.463$), the smallest datasets for the deeper tree (in which Structurama was not significantly better than NP.E, $p = 0.126$), and datasets including 10 alleles and 100 loci for the deeper tree (in which NP.C was not significantly better than Gaussian clustering, $p = 0.105$). However, in all these cases, these equivalent methods were significantly outperformed by another method (i.e. Structurama or NP.C, $p < 0.001$). With smaller numbers of alleles sampled per locus, NP.C typically outperformed Structurama, whereas with larger numbers of alleles, Structurama typically outperformed NP.C. Similarly, with smaller numbers of alleles, NP.E typically outperformed Gaussian clustering, but with larger numbers of alleles Gaussian clustering outperformed NP.E.

Despite the higher accuracy of NP.C at lower numbers of sampled alleles, Structurama had the lowest percent of incorrectly assigned samples, regardless of sampling strategy (Fig. 2). Further, all methods except Structurama failed in some cases to delimit even the most deeply divergent species (i.e. those that diverged from all other species 6 N or 12 N generations ago), lumping them with other species at the exclusion of lineages more closely related to the latter species. However, Structurama only failed to delimit these deeply divergent species under the least intense sampling strategies (i.e. five loci for any number of alleles or 10 loci and five alleles). In all sampling strategies with 10 or more loci sequenced, except when five alleles were sequenced for 10 loci, Structurama successfully detected all divergences greater than approximately 2 N generations. With at least 25 loci se-

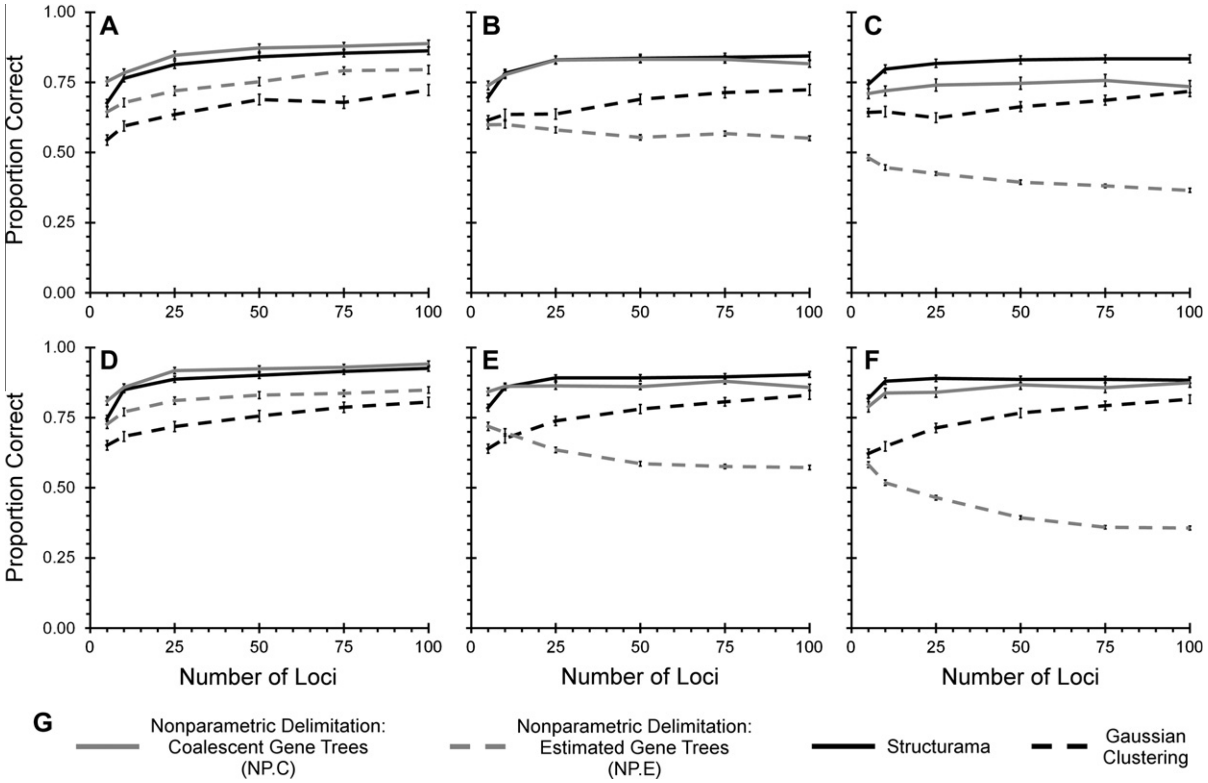

Fig. 1. Proportion of samples correctly assigned to species by each of the tested methods for the various tested sampling strategies. For all panes, line colors correspond to species delimitation method: solid grey, NP.C; dashed grey, NP.E; solid black, Structurama; dashed black, Gaussian clustering. (A) 6 N total tree depth, 5 alleles per species. (B) 6 N total tree depth, 10 alleles per species. (C) 6 N total tree depth, 20 alleles per species. (D) 12 N total tree depth, 5 alleles per species. (E) 12 N total tree depth, 10 alleles per species. (F) 12 N total tree depth, 20 alleles per species. (G) Legend indicating the method indicated by each line style and color.

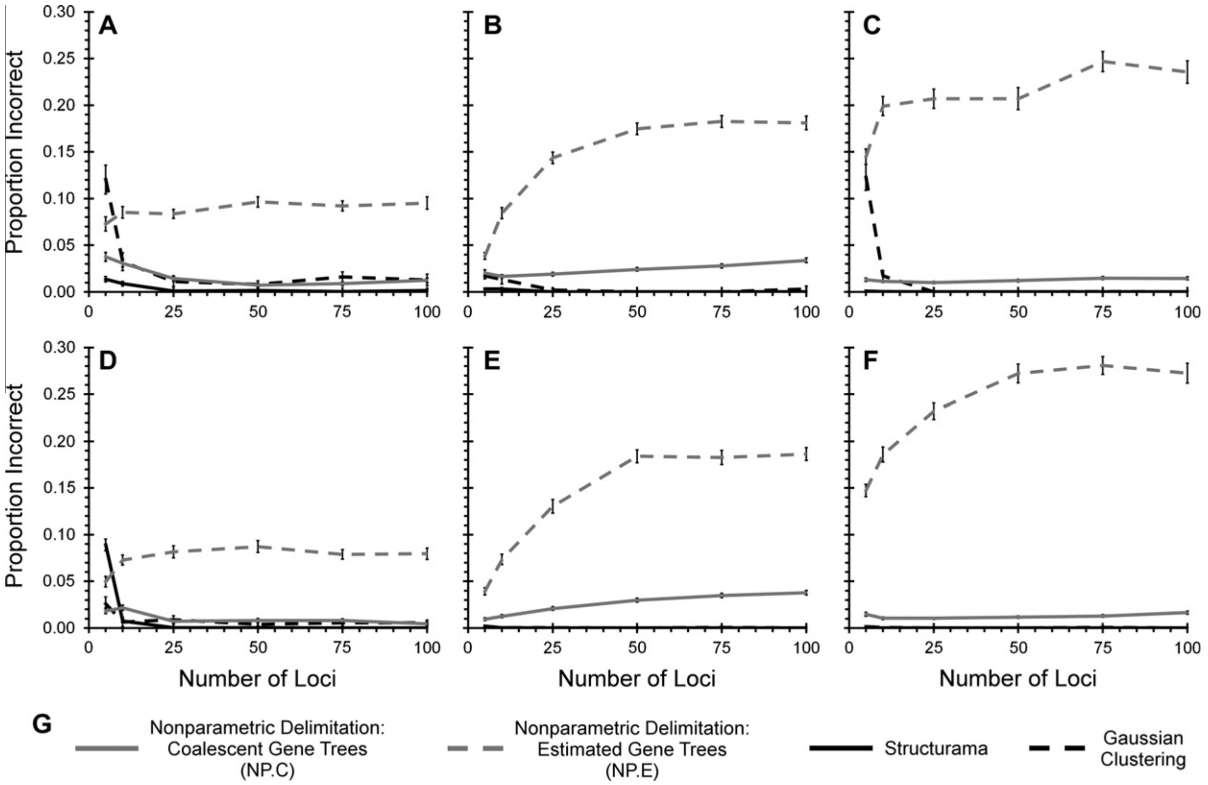

Fig. 2. Proportion of samples incorrectly assigned to species by each of the tested methods for the various tested sampling strategies. For all panes, line colors correspond to species delimitation method: solid grey, NP.C; dashed grey, NP.E; solid black, Structurama; dashed black, Gaussian clustering. (A) 6 N total tree depth, 5 alleles per species. (B) 6 N total tree depth, 10 alleles per species. (C) 6 N total tree depth, 20 alleles per species. (D) 12 N total tree depth, 5 alleles per species. (E) 12 N total tree depth, 10 alleles per species. (F) 12 N total tree depth, 20 alleles per species. (G) Legend indicating the method indicated by each line style and color.

quenced, Structurama successfully detected all divergences greater than approximately 1 N.

3.1. Species delimitation using structurama

The effect of number of alleles on the accuracy of species delimitation by Structurama varied depending on the number of loci sampled (Fig. 3, Tables 1 and 2). With small numbers of loci (<25), accuracy generally increases with increasing numbers of alleles. However, when sampling a large number of loci (>50), the accuracy generally decreases with increasing numbers of alleles. When sampling only five alleles, there was generally a greater sensitivity to the number of loci sampled, and accuracy increased significantly for all increases in numbers of loci except from 75 to 100 loci ($p < 0.024$; Fig. 3, Tables 3 and 4). With greater than five alleles sampled, a significant increase in accuracy was detected for the increases from five to 10 loci ($p < 0.001$), and no significant increase in accuracy was detected by increasing the number of loci beyond 25 ($p > 0.514$).

The majority of the errors in species delimitations with Structurama were a result of over-lumping of species (Figs. 2, 4, 5 and 6), typically involving recently diverged (<1.5 N generations) sister species. Over 90% of the species over-lumped by Structurama were sister taxa, and most other cases of over-lumped species involved lumping of closely related, three-species clades. We detected some instances of over-splitting of species and of incorrectly assigned samples with small datasets (five alleles and <25 loci, or 10 alleles and <10 loci); however, when sampling larger datasets, we found almost no instances of incorrectly assigned samples or of over-split species (Figs. 2, 5 and 6).

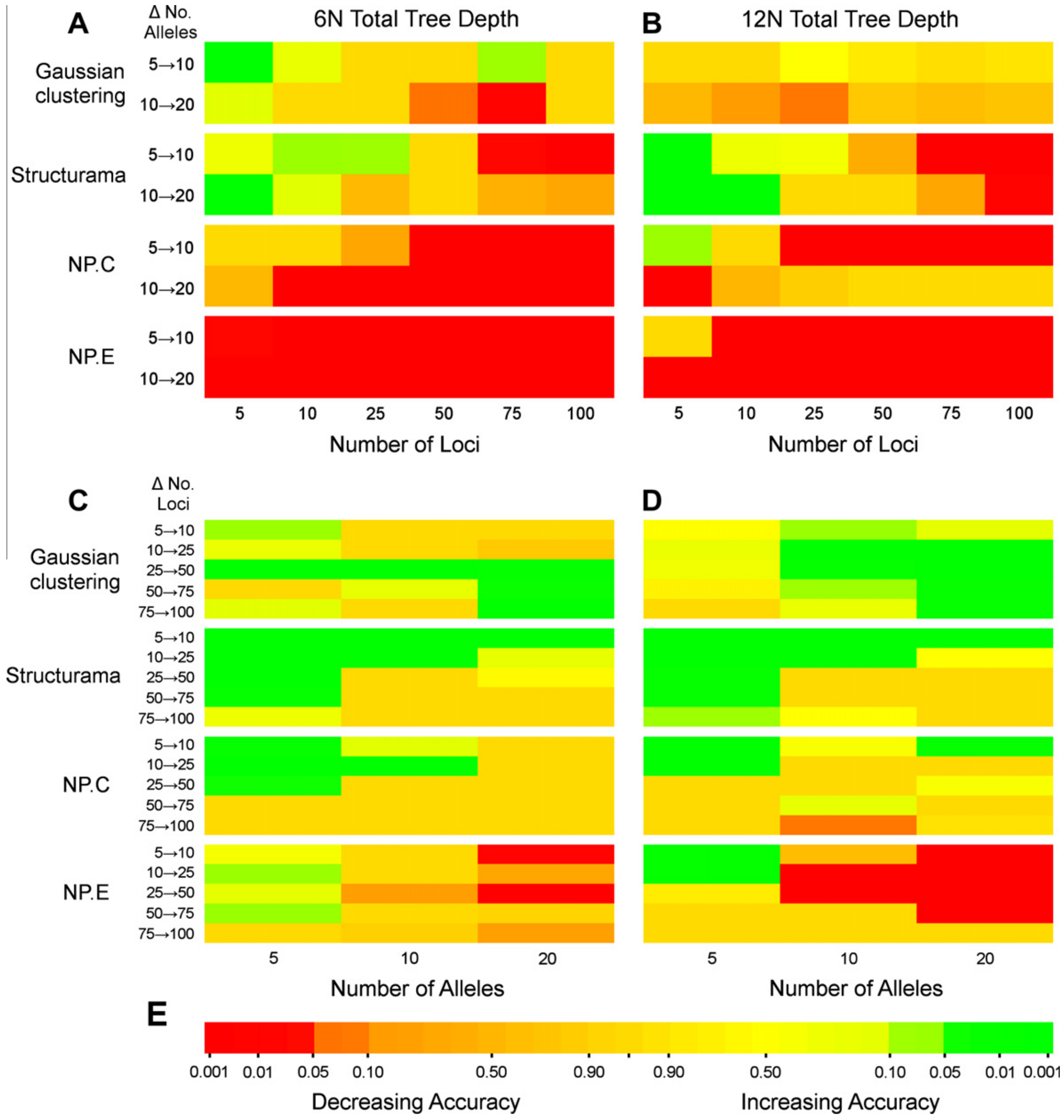

Fig. 3. Significance of change in accuracy of species delimitation with changes in sampling intensity. Red indicates significant decrease in accuracy, green indicates significant decrease in accuracy, yellow indicates no significant change. (A) Impact of increasing number of alleles sampled on 6 N total tree depth. (B) Impact of increasing number of alleles sampled on 12 N total tree depth. (C) Impact of increasing number of loci sampled on 6 N total tree depth. (D) Impact of increasing number of loci on 12 N total tree depth. (E) Legend indicating the Bonferroni corrected p^{*}-value (p^{*}-value = p-value × number of comparisons) indicated by each color.

Table 1
Significance of change in accuracy of species delimitation with increasing numbers of alleles per species for shallower species trees, 6 N total tree depth. *T*-scores greater than zero indicate an increase in accuracy with increased sampling intensity, *t*-scores less than zero indicate decreased sampling intensity. *P*-values are corrected for multiple comparisons via Bonferroni correction (p^*-value = *p*-value × number of comparisons (2). Values significant at the $\alpha = 0.05$ level after Bonferroni correction are in bold.

Δ No. alleles	5 Loci		10 Loci		25 Loci		50 Loci		75 Loci		100 Loci		All #s Loci	
	t	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value
Gaussian clustering														
5 → 10	**4.013**	**0.001**	2.041	0.219	0.100	1.000	0.089	1.000	2.422	0.086	0.072	1.000	3.705	0.001
10 → 20	2.314	0.114	0.776	1.000	−1.258	1.000	−**2.625**	**0.050**	−**3.002**	**0.017**	−0.620	1.000	−1.349	0.889
Structurama														
5 → 10	1.920	0.289	2.468	0.077	2.411	0.089	−0.907	1.000	−**2.805**	**0.030**	−**3.262**	**0.008**	0.922	1.000
10 → 20	**4.813**	**<0.001**	2.352	0.103	−1.616	0.546	−1.136	1.000	−1.750	0.416	−1.915	0.292	1.626	0.523
NP.C														
5 → 10	−0.895	1.000	−0.379	1.000	−1.946	0.273	−**3.398**	**0.005**	−**3.481**	**0.004**	−**5.163**	**<0.001**	−**5.944**	**<0.001**
10 → 20	−1.631	0.531	−**3.553**	**0.003**	−**3.971**	**0.001**	−**4.004**	**0.001**	−**3.334**	**0.006**	−**3.549**	**0.003**	−**8.234**	**<0.001**
NP.E														
5 → 10	−**2.818**	**0.029**	−**4.930**	**<0.001**	−**8.370**	**<0.001**	−**13.450**	**<0.001**	−**13.604**	**<0.001**	−**14.095**	**<0.001**	−**21.370**	**<0.001**
10 → 20	−**7.593**	**<0.001**	−**11.115**	**<0.001**	−**13.826**	**<0.001**	−**13.681**	**<0.001**	−**17.020**	**<0.001**	−**16.884**	**<0.001**	−**30.959**	**<0.001**

Table 2
Significance of change in accuracy of species delimitation with increasing numbers of alleles per species for deeper species trees, 12 N total tree depth. *T*-scores greater than zero indicate an increase in accuracy with increased sampling intensity, *t*-scores less than zero indicate decreased sampling intensity. *P*-values are corrected for multiple comparisons via Bonferroni correction (p^*-value = *p*-value × number of comparisons (2). Values significant at the $\alpha = 0.05$ level after Bonferroni correction are in bold.

Δ No. alleles	5 Loci		10 Loci		25 Loci		50 Loci		75 Loci		100 Loci		All #s Loci	
	t	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value
Gaussian clustering														
5 → 10	−0.817	1.000	−0.471	1.000	1.656	0.505	1.423	0.790	1.322	0.946	1.369	0.870	1.834	0.335
10 → 20	−1.641	0.519	−2.247	0.134	−2.442	0.082	−1.460	0.737	−1.564	0.605	−1.492	0.694	−**4.474**	**<0.001**
Structurama														
5 → 10	**3.606**	**0.002**	1.926	0.285	1.830	0.351	−1.817	0.361	−**3.288**	**0.007**	−**3.498**	**0.004**	0.160	1.000
10 → 20	**3.543**	**0.003**	**3.282**	**0.007**	−0.445	1.000	−1.136	1.000	−1.915	0.292	−**3.000**	**0.017**	0.991	1.000
NP.C														
5 → 10	2.481	0.074	0.391	1.000	−**4.927**	**<0.001**	−**6.271**	**<0.001**	−**4.465**	**<0.001**	−**7.288**	**<0.001**	−**7.359**	**<0.001**
10 → 20	−**3.418**	**0.005**	−1.662	0.498	−1.397	0.828	0.402	1.000	−1.295	0.992	1.025	1.000	−2.517	0.060
NP.E														
5 → 10	−0.528	1.000	−**4.987**	**<0.001**	−**12.395**	**<0.001**	−**16.801**	**<0.001**	−**18.165**	**<0.001**	−**20.181**	**<0.001**	−**24.346**	**<0.001**
10 → 20	−**11.807**	**<0.001**	−**14.434**	**<0.001**	−**13.808**	**<0.001**	−**18.857**	**<0.001**	−**23.304**	**<0.001**	−**20.781**	**<0.001**	−**39.820**	**<0.001**

Table 3
Significance of change in accuracy of species delimitation with increasing numbers of loci for shallower species trees, 6 N total tree depth. *T*-scores greater than zero indicate an increase in accuracy with increased sampling intensity, *t*-scores less than zero indicate decreased sampling intensity. *P*-values are corrected for multiple comparisons via Bonferroni correction ($p*$-value = p-value × number of comparisons (5). Values significant at the $\alpha = 0.05$ level after Bonferroni correction are in bold.

Δ No. loci	5 Alleles		10 Alleles		20 Alleles		All #s Alleles	
	t	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value
Gaussian clustering								
5 → 10	2.490	0.072	1.211	1.000	0.193	1.000	2.421	0.080
10 → 25	2.019	0.231	0.103	1.000	−1.418	0.797	0.721	1.000
25 → 50	3.317	0.006	3.978	0.001	3.210	0.009	6.030	<0.001
50 → 75	−0.621	1.000	2.140	0.174	2.820	0.029	1.729	0.424
75 → 100	2.348	0.104	1.068	1.000	3.329	0.006	3.713	0.001
Structurama								
5 → 10	9.451	<0.001	8.133	<0.001	4.942	<0.001	12.653	<0.001
10 → 25	5.599	<0.001	5.181	<0.001	2.116	0.184	7.299	<0.001
25 → 50	4.214	<0.001	1.136	1.000	1.616	0.546	4.037	<0.001
50 → 75	2.879	0.024	0.705	1.000	1.000	1.000	2.547	0.057
75 → 100	1.990	0.247	0.815	1.000	0.000	1.000	1.496	0.679
NP.C								
5 → 10	2.869	0.025	2.352	0.103	0.529	1.000	2.939	0.018
10 → 25	6.114	<0.001	4.255	<0.001	0.991	1.000	5.190	<0.001
25 → 50	2.627	0.050	0.219	1.000	0.295	1.000	1.271	1.000
50 → 75	0.623	1.000	−0.018	1.000	0.478	1.000	0.625	1.000
75 → 100	0.887	1.000	−1.256	1.000	−0.897	1.000	−0.976	1.000
NP.E								
5 → 10	1.819	0.359	0.024	1.000	−2.779	0.033	−0.060	1.000
10 → 25	2.505	0.069	−1.284	1.000	−1.911	0.295	0.043	1.000
25 → 50	2.305	0.116	−2.094	0.194	−3.220	0.009	−1.108	1.000
50 → 75	2.624	0.050	1.065	1.000	−1.339	0.917	1.765	0.393
75 → 100	0.262	1.000	−1.362	0.881	−2.033	0.224	−1.356	0.880

Table 4
Significance of change in accuracy of species delimitation with increasing numbers of loci for deeper species trees, 12 N total tree depth. *T*-scores greater than zero indicate an increase in accuracy with increased sampling intensity, *t*-scores less than zero indicate decreased sampling intensity. *P*-values are corrected for multiple comparisons via Bonferroni correction ($p*$-value = p-value × number of comparisons (5). Values significant at the $\alpha = 0.05$ level after Bonferroni correction are in bold.

Δ No. loci	5 Alleles		10 Alleles		20 Alleles		All #s Alleles	
	t	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value	*t*	*p*∗-Value
Gaussian clustering								
5 → 10	1.642	0.519	2.592	0.055	2.227	0.141	3.549	0.002
10 → 25	2.024	0.228	4.775	<0.001	4.439	<0.001	6.198	<0.001
25 → 50	1.877	0.317	3.091	0.013	4.395	<0.001	4.925	<0.001
50 → 75	1.498	0.687	2.391	0.093	2.875	0.025	3.328	0.005
75 → 100	0.819	1.000	2.073	0.204	2.932	0.021	2.478	0.069
Structurama								
5 → 10	10.325	<0.001	7.212	<0.001	5.898	<0.001	13.321	<0.001
10 → 25	5.105	<0.001	4.573	<0.001	1.682	0.479	6.689	<0.001
25 → 50	2.855	0.026	0.000	1.000	−0.815	1.000	1.129	1.000
50 → 75	2.915	0.022	0.705	1.000	0.000	1.000	2.148	0.162
75 → 100	2.435	0.083	1.647	0.514	−0.575	1.000	2.247	0.127
NP.C								
5 → 10	3.692	0.002	1.722	0.441	2.903	0.023	4.821	<0.001
10 → 25	5.267	<0.001	0.212	1.000	0.155	1.000	2.898	0.020
25 → 50	0.937	1.000	−0.284	1.000	1.741	0.424	1.454	0.735
50 → 75	0.571	1.000	2.094	0.194	−0.680	1.000	0.680	1.000
75 → 100	1.276	1.000	−2.458	0.079	1.348	0.904	0.447	1.000
NP.E								
5 → 10	3.185	0.010	−1.579	0.588	−6.430	<0.001	−1.798	0.366
10 → 25	3.103	0.012	−4.849	<0.001	−4.673	<0.001	−3.306	0.005
25 → 50	1.440	0.766	−3.928	0.001	−7.148	<0.001	−4.570	<0.001
50 → 75	0.493	1.000	−0.934	1.000	−4.316	<0.001	−2.097	0.184
75 → 100	1.136	1.000	−0.365	1.000	−0.474	1.000	0.346	1.000

3.2. Species delimitation using nonparametric delimitation

In the case of nonparametric delimitation using the true coalescent gene genealogies, the effect of increasing the number of alleles varied depending on the scale of the increase, and, to a lesser extent, the tree depth (Fig. 3, Tables 1 and 2). In the case of the shallower trees, accuracy generally decreased with increasing numbers of alleles, though this effect was not significant for the increase from 5 to 10 alleles when sampling less than fifty loci ($p > 0.077$), or when sampling only five loci ($p > 0.531$). For the deeper tree depths, the increase from 5 to 10 alleles increased accuracy when sampling 5 or 10 loci (though non-significantly, $p > 0.074$), but decreased accuracy when sampling large numbers of loci (>25). The increase from 10 to 20 alleles significantly decreased accuracy when sampling five loci ($p = 0.005$), but had no significant impact on accuracy when sampling greater numbers of loci ($p > 0.498$).

The effect of number of loci sampled on delimitation via NP.C also varied dependent on the tree depth and the number of alleles sampled (Fig. 3, Tables 3 and 4). For the shallower trees, the number of loci generally had a stronger effect when sampling a small number of alleles: with five alleles sampled, accuracy increased up to fifty loci, whereas with 20 alleles sampled, the number of loci had no significant effect. Results were similar for the deeper trees, though with five alleles sampled, accuracy improved up to 25 loci.

The most frequent source of error in species delimitations with NP.C was over-lumped species (Figs. 2, 5 and 6). As with other delimitation methods, many of these over-lumped species involved recently diverged sister species. However, many cases of species over-lumping with NP.C also involved more deeply divergent, non-sister species (occasionally involving even the deepest divergences in the simulated species trees of 6 N or 12 N generations), often at the exclusion of other, more closely related species. While less prevalent than over-lumping species, over-splitting species and incorrectly assigning species were also common sources of error in species delimitations based on NP.C, regardless of the sampling strategy (Figs. 2 and 6).

With one exception (the smallest datasets for the deeper trees), the accuracy of nonparametric delimitation based on estimated gene genealogies decreased significantly with increasing number of alleles sampled, regardless of the number of loci sampled or the total tree depths ($p < 0.029$; Fig. 3, Tables 3 and 4). In general, when sampling a small number of alleles, increasing the number of loci increased the accuracy of NP.E, whereas when sampling a large number of alleles, the accuracy generally decreased with increasing numbers of loci (Fig. 3, Tables 3 and 4).

Errors in species delimitations from NP.E varied dependent upon the sampling strategy, but frequently involved over-lumped species, over-split species, and incorrectly assigned samples (Figs. 2, 5 and 6). When sampling only five alleles per species, most of the errors in species delimitations from NP.E resulted from over-lumping of species or incorrectly assigning samples, though over-split species were also frequently detected. With larger datasets (10 or 20 alleles per species), over-lumping of species was still a common source of error; however, over-splitting species and incorrectly assigning samples were increasingly common. The prevalence of these errors of over-splitting species and incorrectly assigning samples increases with larger numbers of loci, to the point that over 20% of the samples were incorrectly assigned with the larger datasets (Fig. 2). Errors in species delimitations from NP.E were also frequently combined, with several samples from each of multiple species lumped into a single species. Further, unlike other methods of species delimitation, over-lumping of species in NP.E analyses frequently involved non-sister species, and often lumped species across the deepest divergences (6 N or 12 N generations) simulated in the species trees, regardless of the sampling strategy.

3.3. Species delimitation using gaussian clustering

The effect of sampling strategy on species delimitation by Gaussian clustering is somewhat more complicated than other examined methods. For the deeper total tree depths, there was generally no significant effect of increasing the number of alleles, though the increase from 10 to 20 alleles tended to decrease accuracy (Fig. 3, Table 2). This decrease was significant when sampling a moderate number of loci (10–25). For the shallower trees, the increase from 5 to 10 alleles tended to increase accuracy (though not significant for all numbers of alleles examined; Fig. 3, Table 1). However, the increase from 10 to 20 alleles tended to increase accuracy of species delimitation when sampling a small number of loci (<25; Fig. 3, Table 1), but decrease accuracy when sampling a larger number of alleles (>25; Fig. 3, Table 1). Increasing the number of loci generally resulted in an increase in the accuracy of delimitations, regardless of the total tree depth or the number of alleles sampled (Fig. 3, Tables 3 and 4). However, these increases were not significant in a number of cases.

As with species delimitation analyses using Structurama, the most prevalent source of error for species delimitation with Gaussian clustering was over-lumping of relatively recently diverged (<2.5 N generations) species. Over 80% of the species over-lumped by Gaussian clustering were sister species, and many other over-lumped species were grouped with one or both other members of a relatively recently diverged, three species clade, though more deeply divergent species, even those over the deepest divergences in the species tree (i.e. 6 N or 12 N generations) were lumped in some, albeit rare, instances. With smaller datasets (<25 loci), over-splitting of species and incorrectly assigning samples was also an important source of error in species delimitations via Gaussian clustering (Figs. 2, 5 and 6). While both over-splitting and incorrect assignments were both detected at all sampling intensities, both these sources of error were rare when sampling larger numbers (25 or more) of loci.

4. Discussion

When sampling 25 or more loci, Structurama always successfully delimited species greater than approximately 1 N generations divergent and typically delimited species greater than 1.5 N generations divergent, regardless of the sampling strategy. When sampling 25 or more loci, at least 90% of the species not properly delimited by Structurama were sister taxa, typically with shallow divergences. Similarly, while Gaussian clustering occasionally failed to delimit even the most divergent species, species greater than 2.5 N generations divergent were typically delimited successfully when sampling 25 or more loci, and at least 80% of those species not detected were sister species with relatively shallow divergences. Thus, the imperfect performance of these methods is largely due to over-lumping of extremely shallow (less than 2 N) divergences between sister species. Lineage sorting is expected to take an average of 4 N ± 2 N generations per locus (Degnan and Rosenberg, 2006, 2009), thus the shallow divergences examined here would be expected to exhibit extensive incomplete lineage sorting for nearly all loci. Indeed, incomplete lineage sorting was abundant in the simulated data: the average gsi for the shallower species trees was 0.808, and only an average of 1.6 species were monophyletic per locus, while the average gsi was only 0.877 for the deeper species trees, and each locus had, on average, 2.4 monophyletic species. Further, with the exception of NP.E, the most frequent source of error in species delimitations was over-lumping of closely related species. The failure of these methods to delimit species with shallow divergences is likely the result of insufficient time for lineage sorting to occur and therefore a lack of detectable differences between species. Thus, delimiting species with extremely shallow divergences should rely on other types of data, such as morphology, ecology, and reproductive isolation, or on identifying specific diagnostic loci responsible for maintaining and driving lineage divergence.

Nonparametric delimitation performs relatively well when the true, coalescent gene genealogies are known. Indeed, when sampling only five alleles, NP.C generally outperforms all other tested methods. However, when using estimated gene genealogies, nonparametric delimitation performs rather poorly, and, when sampling 10 or more alleles, performs significantly worse than any other examined method of species delimitation. The poor

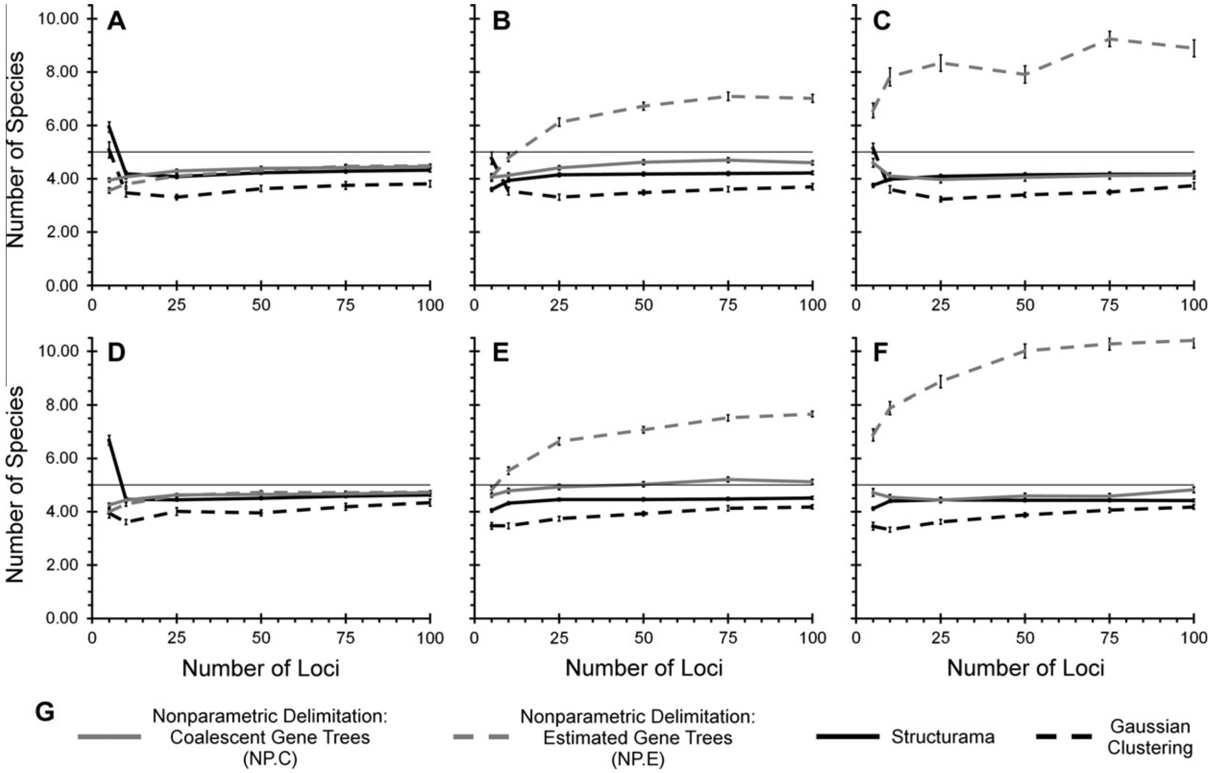

Fig. 4. Number of species identified by each of the tested methods for the various tested sampling strategies. For all panes, line colors correspond to species delimitation method: solid grey, NP.C; dashed grey, NP.E; solid black, Structurama; dashed black, Gaussian clustering. Fine black line represents true number of species (5). (A) 6 N total tree depth, 5 alleles per species. (B) 6 N total tree depth, 10 alleles per species. (C) 6 N total tree depth, 20 alleles per species. (D) 12 N total tree depth, 5 alleles per species. (E) 12 N total tree depth, 10 alleles per species. (F) 12 N total tree depth, 20 alleles per species. (G) Legend indicating the method indicated by each line style and color.

performance of NP.E therefore appears to be a result of errors in gene tree estimation and gene tree uncertainty, rather than poor performance of the nonparametric delimitation method itself. Regardless, nonparametric delimitation's use in empirical study systems is limited, since all researchers will only have estimated gene trees. True coalescent gene trees can never be known with certainty and are particularly difficult to accurately estimate in recently diverged species, where species delimitation is likely to be most problematic. As such, NP.C is empirically impossible, and the problems caused by uncertainty or errors in gene tree estimation suggest that nonparametric delimitation is an ineffective method for species delimitation. As nonparametric delimitation assumes accurate point estimates of gene trees, relaxing this assumption to accommodate gene tree uncertainty, such as through repeated sampling from a distribution of gene trees rather than using a single fixed topology per locus, may improve the utility of nonparametric delimitation and improve its accuracy when using estimated gene genealogies.

In general, the accuracy of NP.E decreases with increased sampling (Fig. 1), particularly when sampling a large number of alleles, a somewhat unexpected and troubling observation, as with an accurate and powerful method, accuracy should increasing with increasing amounts of data. Further, these decreases in accuracy occur in a complex, non-linear pattern. For example, for the shallower species trees, when sampling 20 alleles per species, the increases from 5 to 10 and from 25 to 50 loci result in significant decreases in accuracy, whereas the increase from 10 to 25 loci, while still resulting in decreased accuracy, is not significant. The cause of this complex pattern is not entirely clear, and may be a result of particularly misleading loci resulting in substantial decreases in accuracy, or, perhaps more likely, it may be the result of stochasticity and noise in the dataset overpowering any signal of species identity. Regardless, the decreasing accuracy of species delimitations from NP.E with increased sampling intensity is apparently due to the accumulation of errors in estimated gene genealogies, resulting in a combination of increased over-splitting of species (Fig. 6), increased over-lumping of species (Fig. 5), and increased numbers of incorrectly assigned samples (Fig. 2). The mean number of species identified by NP.E when sampling 100 loci and 20 alleles is 8.890 ± 0.32 for the shallower trees or 10.410 ± 0.287 for the deeper trees (Fig. 4), far higher than the true number of five species, or the number of species identified by any other method: the largest number of species identified by any other method is 5.940 ± 0.194 for the shallower trees or 6.680 ± 0.183 for the deeper trees (Fig. 4; both from Structurama with very small datasets of five loci and five alleles).

Similarly, when sampling 100 loci and 20 alleles, the average proportion of samples incorrectly assigned by NP.E is 27.3% for the deeper trees or 23.6% with the shallower trees. The only other methods with an average proportion of incorrectly assigned samples higher than 5% error rates are Structurama when sampling five alleles and five loci for the deeper trees (8.9%) or Gaussian clustering when sampling five loci at 5 or 20 alleles for the shallower trees (12.0% and 12.3%, respectively). The only method to, on average, incorrectly assign more than 3.8% of samples when sampling greater than five loci was NP.E, and, on average, Structurama incorrectly assigns less than 1% of samples when sampling 10 or more loci.

Despite performing significantly worse than Structurama and NP.C under most sampling strategies, Gaussian clustering performs moderately well, as most species not delimited properly are sister taxa with relatively shallow divergences. Further, while the proportions of incorrectly assigned samples are generally somewhat higher than Structurama, they are still relatively low, particularly when sampling greater than five loci. Proportions of samples

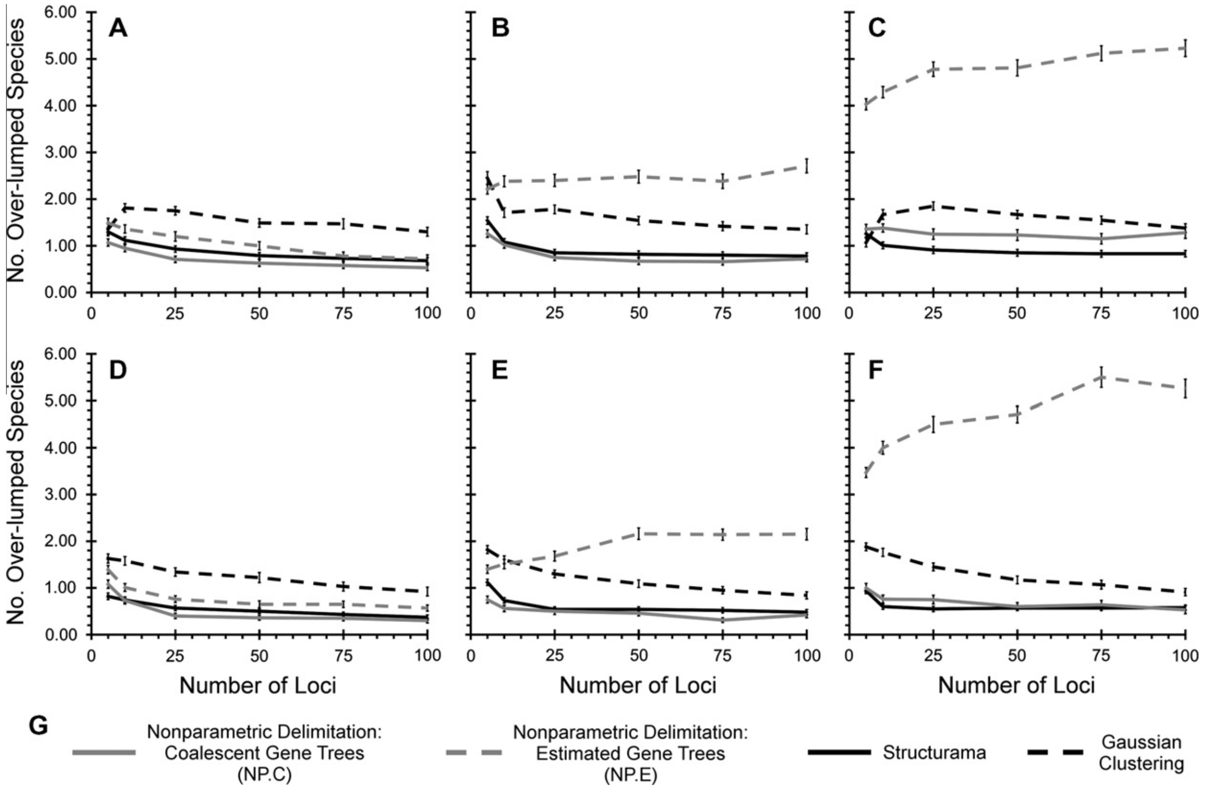

Fig. 5. Number of species over-lumped by each method for the various tested sampling strategies. For all panes, line colors correspond to species delimitation method: solid grey, NP.C; dashed grey, NP.E; solid black, Structurama; dashed black, Gaussian clustering. (A) 6 N total tree depth, 5 alleles per species. (B) 6 N total tree depth, 10 alleles per species. (C) 6 N total tree depth, 20 alleles per species. (D) 12 N total tree depth, 5 alleles per species. (E) 12 N total tree depth, 10 alleles per species. (F) 12 N total tree depth, 20 alleles per species. (G) Legend indicating the method indicated by each line style and color.

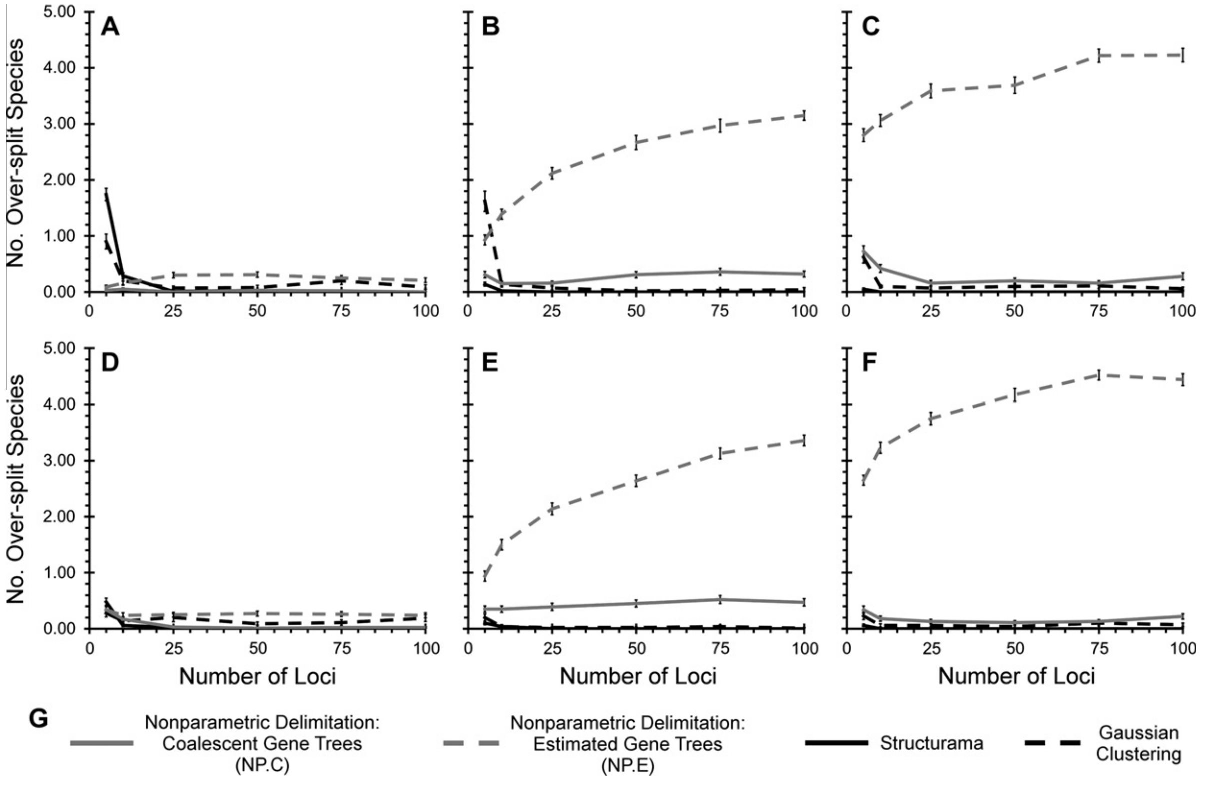

Fig. 6. Number of species over-split by each method for the various tested sampling strategies. For all panes, line colors correspond to species delimitation method: solid grey, NP.C; dashed grey, NP.E; solid black, Structurama; dashed black, Gaussian clustering. (A) 6 N total tree depth, 5 alleles per species. (B) 6 N total tree depth, 10 alleles per species. (C) 6 N total tree depth, 20 alleles per species. (D) 12 N total tree depth, 5 alleles per species. (E) 12 N total tree depth, 10 alleles per species. (F) 12 N total tree depth, 20 alleles per species. (G) Legend indicating the method indicated by each line style and color.

incorrectly assigned by Gaussian clustering are also generally lower than the proportions with NP.C when sampling more than five alleles, or comparable to those with NP.C when sampling only five alleles and 10 or more loci (incorrect assignments are, however, generally rather high with Gaussian clustering when sampling only five loci). Similarly, while the number of over-split species were higher for Gaussian clustering than for Structurama, this number was still low under most sampling strategies, and was far lower than for NP.E. The lower accuracy of Gaussian clustering is thus apparently largely a result of lower sensitivity of the method, as evidenced by the general failure to detect divergences between 1 N and 2.5 N generations divergent, that are generally detected by Structurama, as well as the occasional failure to delimit deeply divergent species at the exclusion of other, more closely related species. However, the relatively complex response of the method to sampling strategy suggests it may be highly sensitive to the amount of information in the loci included. Loci with higher levels of incomplete lineage sorting may cause a strong response in terms of decreased accuracy, whereas those with low levels of incomplete lineage sorting may cause a similarly strong response in increased accuracy.

In conclusion, our study suggests that Structurama may be the most promising method among those tested herein for species delimitation. While NP.C significantly outperforms Structurama when the number of alleles sampled is small, the true coalescent gene trees are never known in empirical studies, thus NP.C is not empirically applicable. Further, Structurama has the lowest rates of incorrectly assigned samples and of over-split species among tested methods, and deeply divergent species were always detected when sampling at least 10 loci, unlike any other method examined. We acknowledge an important caveat, however. The algorithm implemented in Structurama was designed to detect intraspecific population structure by defining clusters in a way that minimizes linkage disequilibrium and maximizes Hardy–Weinberg equilibrium (Falush et al., 2003; Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2011; Pritchard et al., 2000). The simulation strategy implemented herein did not include intraspecific phylogeographic structure, yet it is probable that in some empirical applications, divergent but conspecific populations may be identified as distinct clusters. Thus, clusters defined by Structurama (and other methods tested in this study) are perhaps most appropriately treated as putative genetic lineages that should be further tested, such as using methods of species verification (e.g. Bayesian species delimitation, SpeDeSTEM). Additionally, Structurama and the other methods tested herein provide a means to identify distinct species – i.e. independently evolving lineages – based on available genetic data. However, genetic data alone should not be used for the identification and description of cryptic species. Wherever possible and informative, we recommend combining the genetic species delimitation methods examined herein with other types of data, such as morphology, ecology, sonograms, behavior, and reproductive data, as perhaps the most promising approach to species delimitation in taxonomically difficult complexes.

Acknowledgements

We thank Bryan C. Carstens, Robb T. Brumfield, and Michael E. Hellberg for discussions related to this project. We also thank the Austin lab group, Noah M. Reid, Sarah M. Hird, Tara A. Pelletier, Jordan D. Satler and an anonymous reviewer for comments on previous versions of this manuscript. Structurama analyses were carried out using the resources of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation. Gene genealogy estimation and nonparametric delimitation analyses were conducted using the High Performance Computing cluster at Louisiana State University. This research was funded by National Science Foundation grants DEB 0445213 and 1146033 to CCA.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2012.06.031.

References

Austin, C.C., Rittmeyer, E.N., Oliver, L.A., Anderman, J.O., Zug, G.R., Rodda, G.H., Jackson, N.D., 2011. The bioinvasion of Guam: inferring geographic origin, pace, pattern and process of an invasive lizard in the Pacific using multi-locus genomic data. Biological Invasions 13, 1951–1967.

Aylor, D.L., Price, E.W., Carbone, I., 2006. SNAP: combine and Map modules for multilocus population genetic analysis. Bioinformatics 22, 1399–1401.

Baum, D.A., Donoghue, M.J., 1995. Choosing among alternative species concepts. Systematic Botany 20, 560–573.

Baum, D.A., Shaw, K.L., 1995. Genealogical perspectives on the species problem. In: Hoch, P.C., Stephenson, A.G. (Eds.), Experimental and Molecular Approaches to Plant Biosystematics. Missouri Botanical Garden, St. Louis, MO, pp. 289–303.

Bazinet, A.L., Neel, M.C., Shaw, K.L., Cummings, M.P., unpublished. The genealogical sorting index: software and web site for quantifying the exclusivity of lineages.

Bickford, D., Lohman, D.J., Sodhi, N.S., Ng, P.K.L., Meier, R., Winker, K., Ingram, K.K., Das, I., 2007. Cryptic species as a window on diversity and conservation. Trends in Ecology and Evolution 22, 148–155.

Bortolus, A., 2008. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. Ambio 37, 114–118.

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

Clement, M., Posada, D., Crandall, K., 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology 9, 1657–1660.

Coyne, J.A., Orr, H.A., 2004. Speciation. Sinauer Associates, Inc., Sunderland, MA, USA.

Cracraft, J., 1989. Speciation and its ontogeny: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In: Otte, D., Endler, J.A. (Eds.), Speciation and Its Consequences. Sinauer Associates, Sunderland, MA, pp. 28–59.

Cummings, M.P., Neel, M.C., Shaw, K.L., 2008. A genealogical approach to quantifying lineage divergence. Evolution 62, 2411–2422.

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

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics 2, 762–768.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordace, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution 24, 332–340.

Doyle, J.J., 1995. The irrelevance of allele tree topologies for species delimitation, and a non-topological alternative. Systematic Botany 20, 574–588.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences of the United States of America 104, 5936–5941.

Ence, D.D., Carstens, B.C., 2011. SpedeSTEM: a rapid and accurate method for species delimitation. Molecular Ecology Resources 11, 473–480.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14, 2611–2620.

Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587.

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

Funk, D.J., Omland, K.E., 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution, and Systematics 34, 397–423.

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

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., Francis, C.M., 2004. Identification of birds through DNA Barcodes. PLoS Biology 2, 1657–1663.

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

Hudson, R.R., 2002. Generating samples under a Wright-Fisher neutral model. Bioinformatics 18, 337–338.

Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. Evolution 56, 1557–1565.

Hudson, R.R., Turelli, M., 2003. Stochasticity overrules the “three-times rule”: genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57, 182–190.

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

Huelsenbeck, J.P., Andolfatto, P., Huelsenbeck, E.T., 2011. Structurama: bayesian inference of population structure. Evolutionary Bioinformatics 7, 55–59.

Joly, S., Bruneau, 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, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. Systematic Biology 56, 887–895.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.

Leaché, A.D., 2009. Species tree discordance traces to phylogeographic clade boundaries in North American Fence Lizards (*Sceloporus*). Systematic Biology 58, 547–559.

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

Leaché, A.D., Rannala, B., 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology 60, 126–137.

Leaché, A.D., Reeder, T.W., 2002. Molecular systematics of the Easter fence lizard (*Sceloporus undulatus*): a comparison of parsimony, likelihood, and bayesian approaches. Systematic Biology 51, 44–68.

Lefébure, T., Douady, C.J., Gouy, M., Gibert, J., 2006. Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. Molecular Phylogenetics and Evolution 40, 435–447.

Linnen, C.R., Farrell, B.D., 2008. Comparison of methods for species-tree inference in the Sawfly Genus *Neodiprion* (Hymenoptera: Diprionidae). Systematic Biology 57, 876–890.

Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.

Maddison, W., 1997. Gene trees in species trees. Systematic Biology 46, 523–536.

Maddison, W.P., Maddison., D.R., 2010. Mesquite: a modular system for evolutionary analysis. Version 2.73 <http://mesquiteproject.org>.

Mayr, E., 1942. Systematics and the Origin of Species. Columbia University Press, New York.

Mayr, E., 1995. Species, classification, and evolution. In: Arai, R., Kato, M., Doi, Y. (Eds.), Biodiversity and Evolution. National Science Museum Foundation, Tokyo, pp. 3–12.

McCormack, J.E., Huang, H., Knowles, L.L., 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Systematic Biology 58, 501–508.

Monaghan, M.T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D.J.G., Lees, D.C., Ranaivosolo, R., Eggleton, P., Barraclough, T.G., Vogler, A.P., 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delimitation. Systematic Biology 58, 298–311.

Moritz, C., Cicero, C., 2004. DNA barcoding: promise and pitfalls. PLoS Biology 2, 1529–1531.

O'Meara, B.C., 2010. New heuristic methods for joint species delimitation and species tree inference. Systematic Biology 59, 59–73.

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.

Pinzón, J.H., LaJeunesse, T.C., 2011. Species delimitation of common reef corals in the genus *Pocillopora* using nucleotide sequence phylogenies, population genetics and symbiosis ecology. Molecular Ecology 20, 311–325.

Pons, J., Barraclough, T.G., Gomex-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D., Vogler, A.P., 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic Biology 55, 595–609.

Price, E.W., Carbone, I., 2005. SNAP: workbench management tool for evolutionary population genetic analysis. Bioinformatics 21, 402–404.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of Population structure using multilocus genotype data. Genetics 155, 945–959.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications In The Biosciences 13, 235–238.

Salicini, I., Ibáñez, C., Juste, J., 2011. Multilocus phylogeny and species delimitation within the Natterer's bat species complex in the Western Palearctic. Molecular Phylogenetics and Evolution 61, 888–898.

Sites Jr., J.W., Marshall, J.C., 2003. Delimiting species: a Renaissance issue in systematic biology. Trends in Ecology and Evolution 18, 462–470.

Sites Jr., J.W., Marshall, J.C., 2004. Operational criteria for delimiting species. Annual Review in Ecology, Evolution and Systematics 35, 199–227.

Stamatakis, A., 2007. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of Taxa and mixed models. Bioinformatics 22, 2688–2690.

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Templeton, A.R., Crandall, K.A., Sing, C.F., 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease and sequencing data. III. Cladogram estimation. Genetics 132, 619–633.

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.