ELSEVIER

Convergence of multiple markers and analysis methods define the genetic distinctiveness of cryptic species ☆

Mrinalini [a,*], Roger S. Thorpe [a], Simon Creer [a], Delphine Lallias [a], Louise Dawnay [a], Bryan L. Stuart [b], Anita Malhotra [a,*]

[a] School of Biological Sciences, College of Natural Sciences, Bangor University, Gwynedd LL57 2UW, UK
[b] North Carolina Museum of Natural Sciences, Raleigh, NC, USA

ABSTRACT

Using multiple markers and multiple analytical approaches is critical for establishing species boundaries reliably, especially so in the case of cryptic species. Despite development of new and powerful analytical methods, most studies continue to adopt a few, with the choice often being subjective. One such example is routine analysis of Amplified Fragment Length Polymorphism (AFLP) data using population genetic models despite disparity between method assumptions and data properties. The application of newly developed methods for analyzing this dominant marker may not be entirely clear in the context of species delimitation. In this study, we use AFLPs and mtDNA to investigate cryptic speciation in the Trimeresurus macrops complex that belongs to a taxonomically difficult lineage of Asian pitvipers. We analyze AFLPs using population genetic, phylogenetic, multivariate statistical, and Bayes Factor Delimitation methods. A gene tree from three mtDNA markers provided additional evidence. Our results show that the inferences about species boundaries that can be derived from population genetic analysis of AFLPs have certain limitations. In contrast, four multivariate statistical analyses produced clear clusters that are consistent with each other, as well as with Bayes Factor Delimitation results, and with mtDNA and total evidence phylogenies. Furthermore, our results concur with allopatric distributions and patterns of variation in individual morphological characters previously identified in the three proposed species: T. macrops sensu stricto, T. cardamomensis, and T. rubeus. Our study provides evidence for reproductive isolation and genetic distinctiveness that define these taxa as full species. In addition, we re-emphasize the importance of examining congruence of results from multiple methods of AFLP analysis for inferring species diversity.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Current efforts to discover and delimit species are usually facilitated by DNA sequence-based evolutionary reconstructions. However, this can be biased by locus-specific evolutionary constraints and genomic non-representation. In the case of recently diverged species, short nuclear sequences are often phylogenetically uninformative (e.g. Bardeleben et al., 2005; Weisrock et al., 2010). Multilocus markers are therefore highly recommended and offer quantitative advantages and genome-wide coverage (Zhang and Hewitt, 2003; Meyer and Paulay, 2005; Brito and Edwards, 2009; Dupuis et al., 2012; Leaché et al., 2014). Routine

phylogenomic analysis is still constrained by issues such as differences in results across methods, the need to integrate evolutionary histories of multiple loci, the lack of guidelines for best practices, and extensive computational requirements (Song et al., 2012; Gatesy and Springer, 2013; Dell'Ampio et al., 2014; Faria et al., 2014; Leaché et al., 2014). Nonetheless, new methods have been developed and high-throughput sequence analysis is gaining popularity in evolutionary and speciation research (Morin et al., 2010; Springer et al., 2012; McCormack et al., 2013; Misof et al., 2014).

Alternatively, the use of non-sequence-based multilocus markers has also increased, often revealing surprisingly clear, fine-scale genetic structure undetected by morphology and sometimes even by mtDNA markers (Brown et al., 2007; Egger et al., 2007; Kingston et al., 2009; Meudt et al., 2009; Milá et al., 2010). Among these, Amplified Fragment Length Polymorphism (AFLP) (Vos et al., 1995) is a time-tested, cost-effective, and powerful technique requiring no sequence knowledge. AFLPs have continued to prove useful for

resolving species-level taxonomy, recovering patterns of speciation, evolutionary histories and inter-relationships, inferring population structure and genetic diversity analyses in a wide-range of animal species, such as butterflies (Kronforst and Gilbert, 2008; Quek et al., 2010), cichlids (Albertson et al., 1999), salamanders (Wooten et al., 2010), lizards (Ogden and Thorpe, 2002), dolphins (Kingston et al., 2009), and pinnipeds (Dasmahapatra et al., 2009).

1.1. Dominant marker analysis

Usually, multilocus markers are analyzed using genetic clustering and diversity analysis methods. These are implemented in population genetics models using F-statistics calculations based on allele-frequencies. The uses of these methods with respect to co-dominant datasets have been evaluated in both spatial and non-spatial models (Latch et al., 2006; Chen et al., 2007; Frantz et al., 2009). AFLPs, however, are dominant markers and do not allow distinction between homozygous and heterozygous states of an allele. Therefore, AFLP analysis using allele frequency-based population genetics models requires several assumptions to be made. As this results in analytical limitations, it is highly recommended that multiple analysis methods are applied, and inferences are made with high confidence only when results show congruence across methods (Carstens et al., 2013). However, the majority of AFLP studies continue to apply population genetic methods and derive biological inferences with rare discussion of possible analytical biases (Hollingsworth and Ennos, 2004; Bonin et al., 2007). One such example is the routine use of the popular, non-spatial clustering program, STRUCTURE that uses a Bayesian MCMC algorithm to infer K – the number of populations. Other such programs include TESS, GENECLUST, and GENELAND that perform Bayesian cluster analysis under spatial models (Guillot et al., 2005a, 2005b; François et al., 2006; Chen et al., 2007; Guillot, 2008; Guillot et al., 2008).

The algorithm used in STRUCTURE accommodates dominant data by assuming the presence of recessive alleles at a subset of loci that provide partial information about diploid genotypes for the entire dataset (Falush et al., 2007). GENELAND uses geographic coordinates and identifies groups of individuals in Hardy–Weinberg Equilibrium (HWE) (Guillot et al., 2005a, 2005b). GENELAND was recently upgraded to correct allele frequency estimates from dominant data by taking into account observed genotypes and estimating unknown genotypes using model-based MCMC simulations (Guillot and Santos, 2010). However, both STRUCTURE and GENELAND still assume that AFLP null-alleles (i.e., band absences) are recessive alleles for allele frequency calculations and subsequent K estimation. Common assumptions that drive these analyses (such as HWE in a population and linkage disequilibrium between populations but not within populations) are conceptually not applicable to dominant data. The models and assumptions used for K estimation in STRUCTURE are less than straightforward and need to be used with caution as they could yield inaccurate results (Pritchard et al., 2000). Furthermore, the accuracy of K estimation using dominant datasets in GENELAND is lower than in co-dominant datasets (Guillot and Santos, 2010). In some cases, tree-building analysis of AFLPs has performed better at cluster identification due to the absence of population genetics model assumptions (Meudt et al., 2009). A certain degree of uncertainty is therefore unavoidable when using population genetic models to estimate the number of K, to assign individuals to each K, and to assess the genetic structure of each K. Hence, the need for new and more appropriate methods for dominant marker analysis has been identified (Hollingsworth and Ennos, 2004; Excoffier and Heckel, 2006; Bonin et al., 2007; Meudt et al., 2009).

Multivariate methods (such as factor and cluster analysis, principal component analysis, Multi-Dimensional Scaling, Molecular Analysis of Variance) implemented outside the confines of population genetics models have been extensively used to analyze AFLPs. More recently, two tools that use a combination of multivariate procedures to analyze multilocus genetic data were developed. Discriminant Analysis of Principal Components (DAPC) was developed in *adegenet* (R package) as a method for inferring genetic clusters and genetic diversity using dominant data (Jombart et al., 2010). Hausdorf and Hennig (2010) developed *prabclus*, also an R package, for species delimitation and ordination-cluster analysis using both dominant and co-dominant datasets. Both *adegenet* and *prabclus*, have performed better than STRUCTURE in initial studies (Hausdorf and Hennig, 2010; Jombart et al., 2010). These methods could be useful for cluster and population structure analysis and speciation research employing dominant markers.

Finally, Leaché et al. (2014) developed a new approach for AFLP and SNP based species delimitation by adapting a method called Single Nucleotide Polymorphism and AFLP Phylogenies (SNAPP) (Bryant et al., 2012). SNAPP produces posterior probability distributions of allele frequency changes and allows species tree estimation without the need for gene tree reconstruction and integration (Bryant et al., 2012). Grummer et al. (2013) first developed sequence-based Bayes Factor Delimitation (BFD) to perform marginal likelihood estimations (MLE) and test multiple species delimitation hypotheses. Species delimitation models are tested at the same time as species tree estimation, forgoing the need to specify a guide species tree (Grummer et al., 2013). Leaché et al. (2014) modified this and developed SNAPP BFD for species delimitation using SNPs and AFLPs. SNAPP BFD is implemented using MLE path sampling analysis in version 2 of Bayesian Evolutionary Analysis Sampling Trees (BEAST) software (Drummond et al., 2012; Bouckaert et al., 2014).

1.2. The study group – Trimeresurus (Trimeresurus) macrops

Asian green pitvipers from the genus *Trimeresurus* (Serpentes: Crotalidae: Crotalinae) (Lacépède, 1804) are well known for cryptic speciation (e.g. Malhotra and Thorpe, 2000, 2004a; Vogel et al., 2004). The genus was divided into several genera in 2004 (Malhotra and Thorpe, 2004b), among which was *Cryptelytrops* (Cope, 1860). Recently, as a result of new information on the type species of *Trimeresurus*, *Cryptelytrops* was shown to be a junior synonym of *Trimeresurus* (David et al., 2011). Therefore, the species placed in *Cryptelytrops* by Malhotra and Thorpe (2004b) are now correctly placed within *Trimeresurus*, whether defined in a broader (by subsuming Malhotra and Thorpe's proposed genera as subgenera) or narrower sense (continuing to accept the existence of several well-defined, ecologically, genetically and morphologically diagnosable generic-level units within the former larger genus).

Trimeresurus macrops sensu lato (*s.l.*) is distributed across Thailand, Laos, Cambodia, and Viet Nam, and was shown to consist of three cryptic species with disjunct geographic ranges in the highlands of Cambodia (Fig. 1), distinguished by variations in several individual morphological characters corresponding to their allopatric distributions (Malhotra et al., 2011a). The populations have been proposed as three distinct species: (i) *T. macrops sensu stricto* (*s.s.*) found in Thailand, south & central Laos, and northeast Cambodia, (ii) *T. cardamomensis* (Cardamom Mountains green pitviper), from southeast Thailand and the Cardamom mountains of southwest Cambodia, and (iii) *T. rubeus* (Ruby-eyed green pitviper), found in southern Viet Nam and eastern Cambodia (Malhotra et al., 2011a). The morphology and species ranges for each of these putative species have been fully described (Malhotra et al., 2011a). A multivariate morphometric analysis, however, was not completely successful in separating the three species (Fig. A.1), possibly due to geographic variation within each of the species. Thus,