The effect of missing data on coalescent species delimitation and a taxonomic revision of whipsnakes (Colubridae: *Masticophis*)

Kyle A. O'Connell[a,b,*], Eric N. Smith[a]

[a] *Department of Biology, The University of Texas at Arlington, Arlington, TX 76019, USA*
[b] *Department of Vertebrate Zoology; National Museum of Natural History, Washington, DC 20007, USA*

A B S T R A C T

A stable alpha taxonomy is essential to understanding evolutionary processes and achieving effective conservation aims. Taxonomy depends on the identification of independently evolving lineages, and the delimitation of these lineages based on multiple lines of evidence. Coalescent species delimitation within an integrative framework has increased the rigor of the delimitation process. Here we use genome-wide SNP data and coalescent species delimitation to explore lineage relationships within several North American whipsnake species, test the species status of several lineages, and test the effect of missing data on species delimitation. We find support for the elevation of several previously recognized subspecies to full species status, and formally elevate two species. This study demonstrates the power of molecular data and model-based delimitation methods to identify evolutionary relationships, and finds that missing data have little impact on the outcome of delimitation analyses.

1. Introduction

The field of species delimitation has received increased attention in recent years (Sites and Marshall, 2003). Since the foundational work of de Queiroz (2007), the definition of the general lineage species concept has decoupled species conceptualization from species delimitation. As such, various lines of evidence can be used to assess lineage independence, but the status of the species is not dependent on any one type of evidence (de Queiroz, 2007). In the pre-molecular era, species delimitation primarily depended on morphological data, although ecological, distributional, or other types of data were used support a species' status (Padial et al., 2010, and Sites and Marshal, 2004). In the case of allopatrically distributed species, reproductive isolation is demonstrable, but in species with overlapping ranges, researchers traditionally relied on morphological differences as a proxy for reproductive isolation (Fujita et al., 2012). However, morphological or ecological variation may not accurately represent the evolutionary history of a species (Ruane et al., 2014). The advent of molecular data revolutionized taxonomy and species delimitation, but a dependence on a small number of loci often misled inferences of phylogeny due to incomplete lineage sorting and hybridization (Knowles and Carstens, 2007, and Streicher et al., 2016). Fortunately, genomic data, and the subsequent increase in available loci, have helped to mitigate many of these shortcomings by estimating species trees more accurately, and by

allowing for robust testing of species hypotheses (Liu et al., 2015, Leaché et al., 2014, and Faircloth et al., 2012).

Species delimitation methods attempt to accurately quantify independently evolving lineages (Knowles and Carstens, 2007, Petit and Excoffier, 2009, and Sites and Marshall, 2003). The species delimitation process is comprised of two steps: lineage identification and hypothesis testing (Carstens et al., 2013). Lineage identification relies on a variety of methods, including morphological or ecological variation, disjunct geographic distributions, or molecular phylogenies (Wiens, 2007). However, lineages identified by one or more of these methods may not reflect the accurate evolutionary history of lineages, creating the need to test hypotheses regarding species composition and relationships (Fontaneto et al., 2015). Several recent techniques leverage coalescent theory to test species delimitation hypotheses (Fujita et al., 2012, and Pante et al., 2015). Bayes Factor Delimitation (with genomic data; BFD*) is one method for testing hypotheses of species relationships that utilizes genome-wide SNP data (Leaché et al., 2014). This method is advantageous to other coalescent species delimitation methods because it does not require a guide tree, but rather directly estimates the species tree from biallelic markers, and can calculate a marginal likelihood estimate (MLE) for each species model (Leaché et al., 2014). Nonetheless, recent criticism of this method suggests that it may 'over-split' populations instead of species, and that integrative taxonomic approaches should be used to balance these shortcomings (Sukumaran
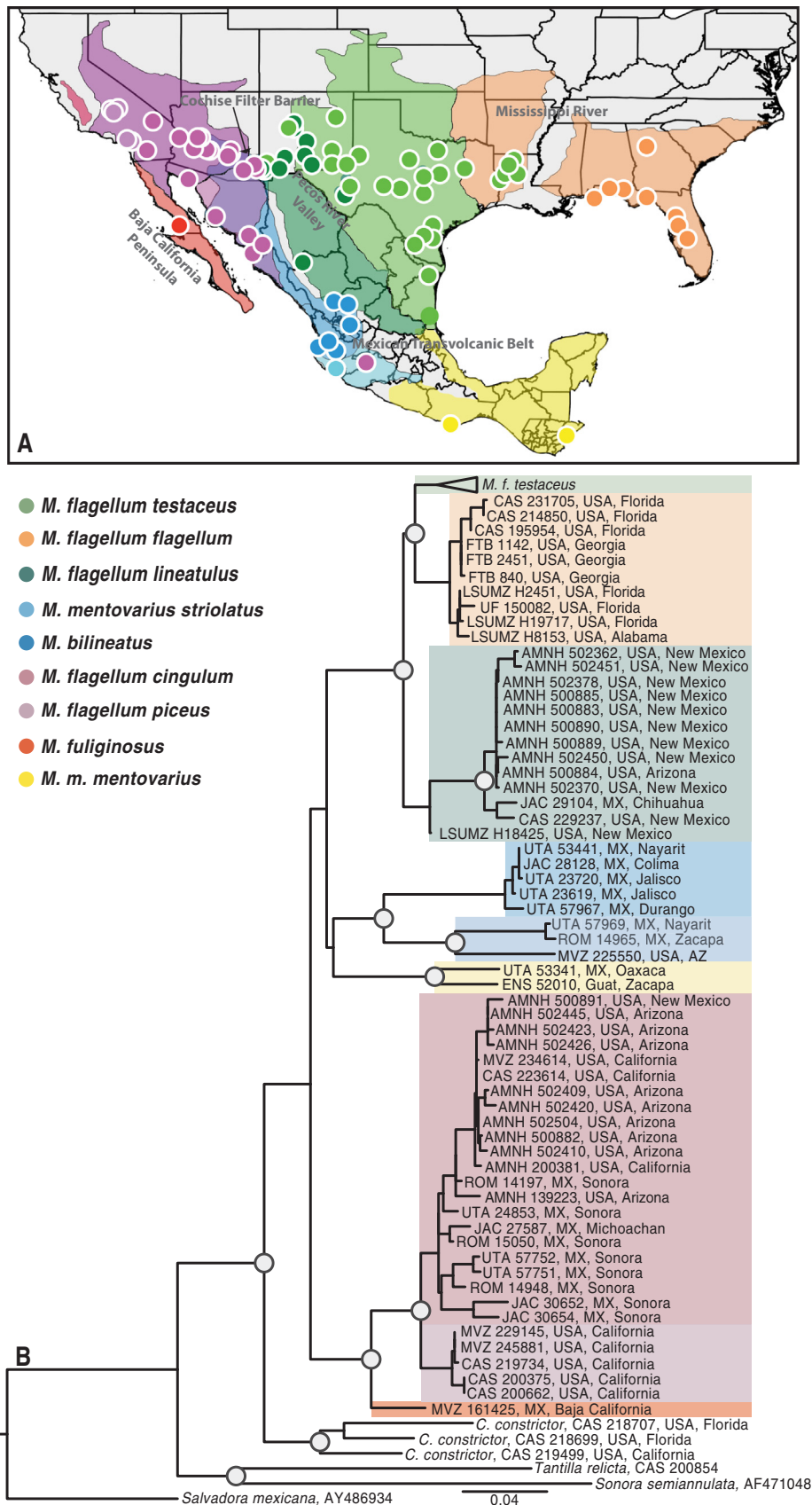
Fig. 1. (A) Map showing the approximate distributional ranges of each subspecies investigated in this study as described by Wilson (1970) and Stebbins (2003). Circles represent sampling localities for mitochondrial data. (B) Maximum likelihood phylogeny including several species of whipsnakes. Grey circles show nodes with at least 70% bootstrap support. Colors on each clade correspond to the colors used in the range map. We collapsed the clade pertaining to *M. flagellum testaceus* to save space.