Species delimitation using machine learning recovers a phylogenomically consistent classification for North American box turtles (*Terrapene* spp.)

Running title: Machine learning species delimitation

Authors:

Bradley T. Martin[1,5,*], Tyler K. Chafin[1], Marlis R. Douglas[1], John S. Placyk Jr.[2,6], Roger D. Birkhead[3], Chris A. Phillips[4], Michael E. Douglas[1]

[1]Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas 72701, USA; Email: (BTM) btm002@uark.edu (send reprint requests to this address); (TKC): tkchafin@uark.edu; (MRD): mrd1@uark.edu; (MED): med1@uark.edu.

[2]Department of Biology, University of Texas, Tyler, Texas, 75799, USA; Email: japlacyk@gmail.com

[3]Alabama Science in Motion, Auburn University, Auburn, AL 36849, USA; Email: birkhrd@auburn.edu

[4]Illinois Natural History Survey, Prairie Research Institute, University of Illinois, Champaign, IL 61820; Email: caphilli@illinois.edu

Present Address:

[5]Global Campus, University of Arkansas, 2 E. Center St., Fayetteville, Arkansas 72701, USA

[6]Science Division, Trinity Valley Community College, Athens, Texas 75751, USA

Disclosure statement: Authors have nothing to disclose

ABSTRACT

Model-based approaches to species delimitation are constrained by computational capacities as well as frequently violated algorithmic assumptions applied to biologically complex systems. An alternate approach employs machine learning to derive species limits without explicitly defining an underlying species model. Herein, we demonstrate the capacity of these approaches to identify phylogenomically relevant groups in North American box turtles (*Terrapene* spp.). We invoked several machine learning-based species delimitation algorithms and a multispecies coalescent approach to parse a large ddRAD sequencing SNP dataset. We highlight two major findings: 1) Machine learning delimitations were variable among replicates, but heterogeneity only occurred within major species tree clades; 2) in this sense unsupported splits echoed patterns of phylogenetic discordance among several species-tree methods. Discordance, as corroborated by previously observed patterns of differential introgression, may reflect biogeographic history, gene flow, incomplete lineage sorting, or their combinations. Our study underscores machine learning as a species delimitation method, and provides insight into how commonly observed patterns of phylogenetic discordance may similarly affect machine learning classification.

Keywords: discordance, species tree, VAE, t-SNE, phylogenomics, ddRAD

1. INTRODUCTION

Delineating species is undeniably crucial for systematics, ecology, and the evolutionary process. Species are the currency of biodiversity, as are inconsistencies in the application of what constitutes a species (‘multiplicity’ of species definitions; Zachos 2018). This creates downstream issues for conservation (Mace 2004), where spurious ‘splitting’ or ‘lumping’ of taxa are impediments to equitable allocation of limited resources. For example, over-splitting may redundantly allow threatened/endangered taxa to proliferate (Zachos *et al.* 2013; Sullivan *et al.* 2014), or conflate recovery goals more appropriately managed at separate scales along the species-population continuum (Coates *et al.* 2018).

On the other hand, inappropriate lumping can mask potential extinctions and the recognition of adaptive differentiation (Stanton *et al.* 2019). This can bias ‘true’ diversity, as reflected by regional or clade-specific differences in taxonomic ‘culture’ (e.g. biases in trait-delimitation or species-concepts), or ‘inertia’ (i.e. persistent knowledge gaps; Gippoliti *et al.* 2018). Both disproportionately promote ‘species at peril’ and subsequently drive inefficient resource allocation (Morrison *et al.* 2009; Garnett & Christidis 2017), viewed divisively as ‘taxonomic inflation’ (Agapow *et al.* 2004; Isaac *et al.* 2004). Nevertheless, species definitions/delineations are a critical dimension in conservation's ‘agony of choice’ regarding resource allocation (Vane-Wright *et al.* 1991; Stanton *et al.* 2019). Delimiting species impacts not only finite resource allocation across programs but also efforts to recover and protect biodiversity.

Earlier work on species delimitation relied on few genes (or markers), resulting in limited scope. Although genomic approaches have shown promise (Allendorf *et al.* 2010), conflicting

genome-wide signals from incomplete lineage sorting (ILS) and gene flow (Funk & Omland 2003) are still apparent. Contemporary species delimitation relies upon a probabilistic approach to model gene tree conflicts (i.e. multispecies coalescent; MSC) (Yang & Rannala 2010). However, some models assume all such conflicts stem from ILS, and thus ignore other sources such as introgressive hybridization.

Two popular packages, BPP and BFD*/SNAPP (Yang & Rannala 2010; Leaché *et al.* 2014a), are not only intractable with large datasets, but also seemingly over-split in the presence of high population structure (Sukumaran & Knowles 2017) or when broad, continuous geographic distributions are involved (Chambers & Hillis 2019). Therein lies the difficulty when species delimitation explicitly assumes an underlying process of speciation (i.e. not effectively modeled as an aspect of high-dimensionality data; Chafin *et al.* 2019). Here, we advocate recently developed machine learning algorithms as an alternative that does not rely upon *a priori* assumptions regarding the speciation process, but instead evaluates the process in a relatively unrestricted manner.

Machine learning is broadly divided into two components: supervised (SML) and unsupervised (UML). The former requires a classification model be 'trained' with *a priori* designations, from which a classification model is derived and optimized for assignment of 'unknown' data. A popular SML approach invokes support vector machines (SVM) that partitions groups using linear or non-linear vectors in multi-dimensional space. However, the requirement of an *a priori* classification scheme from which to train the model limits its applicability, particularly when the purpose is to define groups, as in species delimitation. Additionally, SVM is often computationally demanding, and hence slow with respect to

alternatives (Suryachandra & Reddy 2016). UML, on the other hand, requires no *a priori* classification, and relies instead upon inherent patterns in the data.

Several popular UML classifiers lend themselves to the species delimitation problem, including: Random Forest (RF; Breiman 2001), t-distributed stochastic neighbor embedding (t-SNE; Maaten & Hinton 2008), and variational autoencoders (VAE; Derkarabetian *et al.* 2019), each with inherent strengths and weaknesses. For example, RF uses randomly replicated data subsets (in the form of pairwise distances) as a mechanism to develop binary 'decision trees' for a classification model. All randomly seeded decision trees are aggregated (='forest'), with classification decisions parsed as a majority vote amongst all trees. The random sub-setting approach is relatively robust to correlations among features (=summary statistics or principle components used for prediction) and model overfitting (=over-training the model where it does not generalize well with new data). One stipulation is that features must be of low occupancy and without undue noise (Rodriguez-Galiano *et al.* 2012). By contrast, the goal of t-SNE is to create diagnosable clusters in reduced-dimension space, typically a 2D plane extracted from a distillation of multi-dimensional data. Thus, it conceptually resembles methods such as principle components analysis [(PCA) (Maaten & Hinton 2008)].

Alternatively, VAE uses neural networks in an attempt to 'learn' or reconstruct multidimensional data patterns from a compressed, low-dimensionality ('encoded') representation. Again, the approach conceptually resembles the dimensionality-reduction employed by various ordination techniques, but without linear and orthogonal constraint being imposed upon the informative components. This approach may also be more statistically interpretable (Derkarabetian *et al.* 2019), and thus more appropriate for the capture of variability

within highly complex data. Yet, careful consideration must be paid to the derivation of parameters (e.g. neural network 'depth') that controls the encoding process (Livingstone *et al.* 1997).

UML methods do not require *a priori* designations from which to train a classification model yet may still be sensitive to priors and parameter settings. Thus, guidelines for appropriate application must be clearly defined, particularly regarding complex, empirical datasets. Two metrics that can influence the support of a given species delimitation hypothesis is concordance among algorithms (Carstens *et al.* 2013), and the susceptibility of the underlying algorithms to common sources of phylogenetic discordance. Some machine learning algorithms are robust to processes such as gene flow (Derkarabetian *et al.* 2019; Newton *et al.* 2020; Smith & Carstens 2020), but more empirical tests in complex systems are warranted. For example, performance can vary among datasets, with potential influences including data quality (e.g. missing data proportions) and size (Newton *et al.* 2020), historical demography, evolutionary history, and coalescent processes such as incomplete lineage sorting (Austerlitz *et al.* 2009). Thus, we empirically apply some recently developed software packages (CLADES: Pei *et al.* 2018; RF, t-SNE, VAE: Derkarabetian *et al.* 2019) and discuss their capacity for evaluating a group of species historically recalcitrant to taxonomic resolution.

1.1. *The convoluted evolutionary history of* Terrapene

North American box turtles (Emydidae: *Terrapene*) are primarily terrestrial, with a common name based on an anterior ventral hinge that allows the plastron (bottom part of shell) to

dorsally close against the carapace (Dodd 2001). There are five currently recognized species (Minx 1996; Iverson *et al*. 2017): Eastern (*Terrapene carolina*), Ornate (*T. ornata*), Florida (*T. bauri*), Coahuilan (*T. coahuila*), and Spotted (*T. nelsoni*), with a sixth (*T. mexicana*) proposed (Martin *et al*. 2013, 2014). *Terrapene carolina* includes two subspecies (Woodland: *T. c. carolina*; Gulf Coast: *T. c. major*) that inhabit the eastern U.S. from the Mississippi River to the Atlantic Ocean, and south through the Gulf Coastal Plain (Fig. 1). The putative *T. mexicana* contains three subspecies (Three-toed: *T. m. triunguis*; Mexican: *T. m. mexicana*; Yucatan: *T. m. yucatana*) ranging across the southeastern and midwestern United States, the Mexican state of Tamaulipas, and the Yucatan Peninsula. The Ornate (*T. ornata ornata*) and Desert (*T. o. luteola*) box turtles inhabit the Midwest and Southwest U.S. plus the Northwest corner of México, while the Southern and Northern Spotted box turtles (*T. nelsoni nelsoni* and *T. n. klauberi*) occupy the Sonoran Desert in western México. *Terrapene coahuila* is semi-aquatic and restricted to Cuatro Ciénegas (Coahuila, México), and the Florida box turtles occur in Peninsular Florida.

Morphological analyses delineate *T. carolina*/*mexicana* as a single species, sister to *T. coahuila* (Minx 1992, 1996), with anecdotal support from a subset of genetic studies (Feldman & Parham 2002; Stephens & Wiens 2003). Alternatively, Martin et al. (2013) proposed the elevation of *T. mexicana* as a separate species, with *T. coahuila* as a subgroup within *T. carolina*. In this latter study, *T. c. carolina* was sister to *T. c. major*/*T. coahuila*, although potential gene flow was suspected between *T. c. carolina* and *T. c. major* due to mito-nuclear discordance. Accordingly, *T. c. major* was recently demoted to an intergrade population and its subspecific status removed (Butler *et al.* 2011; Iverson *et al.* 2017), but Martin et al. (2013) disagreed and a more recent study identified two potentially pure *T. c. major* populations in the Florida and

Mississippi panhandles (Martin *et al.* 2020). Likewise, *T. bauri* (formerly *T. carolina bauri*) was recently elevated to a distinct species (Butler *et al.* 2011; Iverson *et al.* 2017). a possibility that Martin *et al.* (2013) acknowledged, albeit cautiously as weak statistical support and inconsistent phylogenetic placement were evident. For the sake of clarity, we herein follow the recommendations of Martin et al. (2013, 2014), considering *T. c. major* a distinct entity and *bauri* as a subspecies within *T. carolina*. The monophyly of *T. o. ornata*/*luteola* has also been questioned; Herrmann and Rosen (2009) suggested distinct lineages using microsatellite analyses, whereas Martin et al. (2013) suggested polyphyly and a lack of phylogenetic structure using mitochondrial (mt)DNA and nuclear (n)DNA sequences.

One likely reason for the historically enigmatic classification of *T. carolina* and *T. mexicana* includes contemporary hybridization and introgression occurring within a hybrid zone in the southeastern U.S., with four taxa potentially involved (Auffenberg 1958, 1959; Milstead & Tinkle 1967; Milstead 1969). Some researchers (Fritz & Havaš 2013, 2014) interpret reproductive semi-permeability as evidence for lumping the southeastern taxa as a single species. However, divergent selection reinforcing species boundaries in some southeastern *Terrapene* has been suggested as a reason for re-examining their classificatory status, despite ongoing gene flow (Martin *et al.* 2014, 2020). Alternatively, the close phylogenetic relationship between *T. c. major* and *T. coahuila* is less well understood. This may result from ‘ghost’ admixture of *T. coahuila* and/or *T. c. major* with the extinct *T. c. putnami* (Martin *et al.* 2013).

Herein, we evaluate the classification of *Terrapene* within the context of both UML and coalescent model-based species delimitation approaches. In doing so, we empirically validate the use of machine learning approaches with complex genetic datasets that, upon analysis, support a

well-characterized phylogenetic hypothesis. Of note, observed species delimitation classifications are consistent with patterns of phylogenetic discordance, demonstrating an empirical application where the sources for such discordance may similarly affect machine learning.

2. MATERIALS AND METHODS

2.1. *Sample collection, storage, and DNA extraction*

Tissue samples were obtained from various museums, organizations, agencies, and volunteers (Table S1), then stored in 70%-95% ethanol or DMSO (di-methyl sulfoxide) buffer. Non-invasive samples were also acquired from live specimens, with those more invasive (e.g. toes, muscle) taken from road-kills. Upon receipt, samples were stored at -20°C. Genomic DNA was extracted via the following spin-column kits: DNeasy Blood and Tissue Kits (QIAGEN), QIAamp Fast DNA Tissue Kit (QIAGEN), and E.Z.N.A. Tissue DNA Kits (Omega Bio-tek). Extracted DNA was quantified using Qubit (Thermo Fisher Scientific) broad-range dsDNA fluorometry and tested for high-molecular weight DNA using gel electrophoresis.

2.2. *DNA library preparation*

We first estimated the expected number of loci recovered via ddRAD sequencing (ddRADseq) through *in silico* digestion (Chafin *et al.* 2018) of the painted turtle (*Chrysemys picta*) genome (Shaffer *et al.* 2013). This was done to optimize choice of base-cutters, size-selection bounds, and multiplex-size, thus maximizing loci coverage while promoting high sequencing depth. We also used the *in silico* digest to identify a candidate size-selection that

avoids restriction sites lying within repetitive genomic elements (Chafin *et al*. 2018). The expected number of ddRADseq loci and depth of coverage were empirically verified by performing a restriction enzyme digest on 1,000ng of DNA for a representative panel of 24 samples, followed by fragment analysis (Agilent 4200 TapeStation).

Samples with sufficient DNA quantity (≥50 ng/uL) were processed via ddRADseq protocol (Peterson *et al.* 2012). Between 500-1,000ng of genomic DNA per sample was digested using two restriction enzymes, *PstI* (5'-CTGCA|G-3') and *MspI* (5'-C|CGG-3'). Following a digestion at 37°C for 24 hours, 5ul of each sample was visualized on a 2% agarose gel via electrophoresis to verify DNA fragmentation. Samples were purified using an AMPure XP (Beckman Coulter) solution at a concentration of 1.5X (relative to DNA volume), then standardized at 100ng of DNA per sample. Unique barcoded adapters were ligated to each individual before pooling 48 samples into a library. Taxa were spread across multiple libraries to mitigate potential batch effects, and libraries were size-selected on a Pippin Prep (Sage Science) using the *in silico* optimized range [378-433 base pairs (bp), excluding adapters]. Lastly, a twelve-cycle polymerase chain reaction (PCR) was run with Phusion DNA Polymerase (New England BioLabs), followed by 1x100 single-end sequencing on the Illumina Hi-Seq 4000, pooling two indexed libraries (=96 individuals) per lane. Sequencing and additional quality control (fragment visualization and qPCR) were performed at the Genomics and Cell Characterization Core Facility, University of Oregon/Eugene.

2.3. *Sequence quality control and assembly*

FastQC v. 0.11.5 was used to assess sequence quality (Andrews 2010), with ipyrad v0.7.28 employed to demultiplex the raw sequences and align reads (Eaton & Overcast 2020). Demultiplexed reads were allowed a strict maximum of one barcode mismatch, given that barcodes were designed with a minimum two-base distance. Reads with low PHRED quality scores (<33) were excluded, with additional filtering to remove adapter sequences. We then performed reference-guided assembly using the *Terrapene m. mexicana* reference genome (GenBank Accession #: GCA_002925995.2) with a minimum identity threshold of 0.85. Unmapped reads were removed, and retained loci exhibited ≥20X coverage depth to reduce sequencing error bias (Nielsen *et al.* 2011) and maximize phylogenetically informative sites in the alignment (Eaton *et al.* 2017). Loci were further excluded if they displayed <50% individual occupancy, excessive heterozygosity (≥75% of individual SNPs), or more than two alleles per sample (the latter two instances indicating over-merged paralogs).

2.4. *Phylogenomic inference*

To assess differences in phylogenetic inference, we generated species trees using three contemporary algorithms. Admixture across *Terrapene* hybrid zones has been well-characterized (Butler *et al.* 2011; Martin *et al.* 2013, 2020). Thus, to mitigate the impact of contemporary gene flow on phylogenetic inference, we only utilized individuals confirmed to be parental types (characterized in Martin et al. 2019), as modelled using NewHybrids (Anderson & Thompson

2002). In so doing, we partitioned *T. c. major* into two subsets comprising two putative parental populations.

Maximum likelihood phylogenies have been commonly produced for decades, yet the increased use of large-scale SNP datasets often inflates bootstrap support for concatenated phylogenomic datasets (Salichos & Rokas 2013; Simmons & Goloboff 2014). Coalescent-based approaches that account for independent gene tree histories are more applicable for SNP analysis, and thus we employed SVDQUARTETS [(Chifman & Kubatko 2014), implemented in PAUP* v4.0a164 (Swofford 2003)] to produce a species tree with individuals grouped into populations. Unrooted four-taxon gene trees were generated to assess legitimate splits, then assembled to form the full species tree. SVDQUARTETS performs better for concatenated SNP datasets than do species tree methods utilizing summary statistics (Chou *et al.* 2015), and importantly works well with the large amount of missing data typically produced by ddRADseq (Leaché *et al.* 2015).

To reduce linkage bias and because independent gene tree histories are assumed for each site, only one SNP from each ddRADseq locus was included in the SVDQUARTETS alignment. To assess sampling variance, we ran 100 bootstrap replicates and considered nodes resampled at >70% as strongly supported. Taxon partitions were grouped at the lowest level of field identification (i.e. subspecific designations, when available), and by U.S. and Mexican state locality. Blanding's (*Emydoidea blandingii*) and spotted (*Clemmys guttata*) turtles were included as outgroups. An exhaustive search of all possible quartets was performed, with the consensus tree visualized in FIGTREE v1.4.2 (Rambaut 2014).

We also employed a polymorphism aware model (POMO: Schrempf *et al.* 2016), as implemented in IQ-TREE v1.6.9 (Nguyen *et al.* 2015), to generate a second species tree. We did

so because PoMo allows within-population polymorphism to account for ILS. The full IPYRAD alignment, including invariant sites, was input into PoMo and executed with 1,000 ultrafast bootstrap (UFB) replicates (Hoang *et al.* 2017) and a maximum virtual population size of 19. The discrete gamma rate model was applied (N=4), and clades with bootstrap support ≥95% were considered strongly supported.

Finally, we generated a lineage-tree phylogeny (IQ-TREE v1.7.12; Nguyen et al. 2015) to contrast with our species-trees. An edge-linked partition model with 1,000 UFB replicates was run using ModelFinder (Kalyaanamoorthy *et al.* 2017) to determine the optimal substitution model for each separate ddRADseq locus. Given computational constraints, model selection was restricted only the general time reversible (GTR) model. Following tree reconstruction, IQ-TREE was used to calculate site-wise concordance factors (sCF; Minh *et al.* 2018) for each branch because they are less susceptible than traditional bootstrapping to over-inflation (Philippe *et al.* 2011). The sCF were calculated from 100 quartets randomly sampled from internal branches of the tree, as recommended by IQ-TREE for stable sCF values. UFB≥95% and sCF≥50% were considered as strong support (per IQ-TREE documentation).

For statistical topology tests, we generated lineage trees with IQ-TREE under the topological constraints supported by four species-tree hypotheses derived from: (a) SVDquartets and (b) PoMo topologies, as generated herein; (c) Sanger sequencing with mtDNA and nuclear introns (Martin *et al.* 2013); and (d) Morphological data (Minx 1996). ModelFinder was again employed to optimize substitution models for each locus, as partitioned in a concatenated supermatrix, using a hierarchical clustering algorithm to minimize computational burden in IQ-TREE (*-rcluster*). We also toggled the *-bnni* and *-opt-gamma-inv*

options to reduce the impact of severe model violation and more thoroughly explore gamma and invariant site parameters. Nodal confidence of individual trees was assessed using 1,000 UFB. We then compared support for the concatenated supermatrix among constraint trees using seven topological tests and 10,000 re-samplings: (a) Raw log-likelihoods; (b) bootstrap proportion test using the RELL approximation (Kishino *et al.* 1990); (c) Kishino-Hasegawa test (Kishino & Hasegawa 1989); (d) Shimodaira-Hasegawa test (SH; Shimodaira & Hasegawa 1999); (e) Approximately Unbiased test (Shimodaira 2002); and (f) Expected Likelihood Weights (Strimmer & Rambaut 2002). To visualize support for each topology across the genome, site-likelihood probabilities and pairwise site-likelihood score differences (ΔSLS) were calculated between the best-supported *versus* remaining trees.

2.5. *Species delimitation*

We employed the multispecies coalescent Bayes Factor Delimitation approach [BFD*; (Leaché *et al.* 2014a)] as a baseline to compare the machine learning-based methods. Because BFD* is computationally intensive, taxa were subset to a maximum of five individuals that contained the least amount of missing data (N=37, plus outgroups), with sampling locations varied (excepting *T. c. bauri* and the extremely rare *T. m. mexicana* and *T. coahuila*, which occur exclusively in Peninsular Florida and the Mexican states of Tamaulipas and Cuatro Ciénegas). For consistency, the same subset of individuals was used across all approaches. Details for BFD* prior selection and additional data filtering steps can be found in Supplemental Appendix 1.

For each BFD* model, SNAPP employed 48 path-sampling steps, 200,000 burn-in, plus 400,000 MCMC iterations, with sampling every 1,000 generations. The path-sampling steps were conducted with 200,000 burn-in, 300,000 MCMC generations, α=0.3, 10 cross-validation replicates, and 100 repeats. Trace plots were visualized (TRACER v1.7.1) to confirm parameter convergence and compute effective sample sizes (ESS; Rambaut *et al.* 2018). Bayes factors (BF) were calculated as [2 X (MLE_1 - MLE_2)] from the normalized marginal likelihood estimates (MLE). We considered the following scheme for BF model support: 0<BF<2=no model differentiation; 2<BF<6=positive; 6<BF<10=strong; and BF>10=decisive support (Kass & Raftery 1995).

The RF and t-SNE algorithms (Breiman 2001; Maaten & Hinton 2008) were run and visualized using an R script developed by Derkarabetian *et al.* (2019). The data were represented as scaled principle components (N=37 axes) generated in ADEGENET v2.1.1 (Jombart & Ahmed 2011) in R v3.5.1 (R Development Core Team 2018). We averaged 100,000 majority-vote decision trees over 10,000 bootstrap replicates to generate RF predictions. Clustered RF output was visualized using both classic and isotonic multidimensional scaling procedures (CMDS and ISOMDS; Shepard *et al.* 1972; Kruskal & Wish 1978). We ran t-SNE for 10,000 iterations within which equilibria of the clusters was visually confirmed. Perplexity, which limits the effective number of t-SNE neighbors, was tested at values of five and ten.

2.6. *Determining optimal K for random forests and t-SNE*

Two common clustering algorithms, as implemented in the aforementioned R scripts (Derkarabetian *et al*. 2019), were used to derive optimal *K* for both the RF and t-SNE analyses. The first [Partitioning Around Medoids (PAM); Kaufman and Rousseeuw 1987] attempts to minimize the distance between the center point *versus* all other points of *K* clusters. The program requires *K* to be defined *a priori*, and thus *K*=1-10 were tested, with the gap statistic and highest mean silhouette widths [(MSW) (Rousseeuw 1987; Tibshirani *et al*. 2001)] determining optimal *K*. The second [Hierarchical Agglomerative Clustering (HAC); Fraley and Raftery 1998] merges points with minimal dissimilarity metrics (based on pairwise distances) until all are clustered.

2.7. *Variational autoencoders*

The VAE UML approach (Derkarabetian *et al.* 2019) employs neural networks and deep learning to infer the marginal likelihood distribution of sample means (μ) and standard deviations [(σ) (i.e. 'latent variables')]. Clusters with non-overlapping σ are interpreted as distinct clusters, or 'species.' Data were input as 80% training/20% validation, with model loss (~error) visualized to determine the optimal number of 'epochs' (=cycles through the training dataset). VAE should ideally be terminated when model loss converges on a minimum value between training and validation datasets [(i.e. the 'Goldilocks zone'; Fig. S1) (Al'Aref *et al.* 2019)]. An escalating model loss in the validation dataset indicates overfitting, whereas a failure to acquire a minimum value points to underfitting (i.e. inability to generalize across both training and unseen data).

2.8. *Support vector machines*

The CLADES software (Pei *et al*. 2018) derives six summary statistics for SVM: 1) Proportion of private alleles; 2) a folded site-frequency spectrum (SFS); 3) pairwise F_{ST} values within populations; 4) pairwise F_{ST} values among populations; 5) the pairwise difference ratio ($d_{between}/d_{within}$); and 6) the longest shared tract (longest string shared by two sequences). More extensive methodological descriptions of the UML and SML components of machine learning are found in Supplemental Appendix 1.

3. RESULTS

3.1. *Sampling and data processing*

We sequenced 214 geographically widespread *Terrapene* (Fig. 1; Table S1) including all recognized species and subspecies, save the exceptionally rare *T. nelsoni klauberi*. When possible, we included a minimum of 10 individuals per taxon, though fewer were used per rare clade (*T. m. yucatana*, *T. m. mexicana*, *T. coahuila*, *T. n. nelsoni*, *T. o. luteola*, and *T. c. bauri*). The IPYRAD pipeline recovered 134,607 variable sites across 13,353 loci that mapped to the *T. m. mexicana* genome, with 90,777 being parsimoniously informative. The mean per-individual coverage depth was 56.3X (Fig. S2).

3.2. *Species tree inferences*

The sCF tree contained N=214 tips (Fig. 2), whereas SVDQUARTETS and POMO (Fig. 3) grouped individuals into N=26 populations, again based on locality and subspecies (when provided). The SVDQUARTETS alignment contained 10,299 unlinked SNPs, with 87,395,061 quartets employed to assemble the species tree (Fig. 3a). Concatenated ddRADseq loci were included in the POMO tree (Fig. 3b), to include both invariable and variable sites (N_{sites}=1,163,463). All trees clearly delineated eastern *versus* western clades, with *T. mexicana*, *T. carolina*, and *T. coahuila* composing the eastern clade and the west represented by the monophyletic *T. ornata* and *T. nelsoni*. However, some differences among methodologies were apparent within these clades.

All phylogenies clearly delineated the western *T. ornata* and *T. nelsoni*. However, SVDQUARTETS paraphyletically nested *T. o. luteola* within *T. o. ornata*, whereas IQ-TREE and POMO represented them as distinct monophyletic clades. In the eastern clade, SVDQUARTETS displayed two subdivisions: *Terrapene mexicana* (all subspecies) and *T. carolina* (all subspecies) + *T. coahuila*. POMO did likewise, but also placed *T. m. triunguis* as paraphyletic in *T. mexicana*. Furthermore, SVDQUARTETS, POMO, and IQ-TREE each differed regarding the placement of *T. c. bauri*, *T. coahuila*, and two previously recognized clades within *T. c. major* (Martin *et al.* 2013, 2020). Specifically, SVDQUARTETS depicted *T. c. bauri* as ancestral in the *bauri*/*major*/*coahuila*/*carolina* clade, whereas POMO placed *T. c. major* from MS/*coahuila* as ancestor to *T. c. major* (FL)/*bauri/carolina*. However, IQ-TREE placed 1) *T. c. bauri* sister to all of *T. carolina*/*T. mexicana*, and 2) *T. coahuila*/*T. c. major* (MS) sister to *T. c. carolina*/*T. c.*

major (FL). IQ-TREE also placed one *T. c. major* individual within the *T. m. triunguis* clade, and one *T. c. carolina* as ancestral to the Floridian *T. c. major* and remaining *T. c. carolina*.

3.3. *Species tree reconciliation*

Trees representing Sanger data and SVDQUARTETS were in agreement when we contrasted our topology tests, whereas morphology-based and POMO trees were both significantly rejected (Table 1). Although the SVDQUARTETS tree was ranked the highest, site-likelihood scores indicated that each topology was determined by a small number of loci (Fig. S3), whereas the remaining majority was relatively uninformative.

3.4. *Species delimitation methods compared*

BFD* supported two top models (Table 2): All taxa delimited (*K*=9), and all as distinct save *T. o. ornata*/*T. o. luteola* (*K*=8; Fig. 4). BF did not distinguish between the top models (<2), although both were decisively better than all others (BF>10). Convergence was confirmed for the likelihood traces, and the mean per-model ESS were >300 (Table S2).

The majority of the RF and t-SNE runs (Fig. 4) also grouped *T. o. ornata* and *T. o. luteola*. However, the remaining clusters were split conservatively relative to BFD*. All runs clearly delineated *T. ornata*, *T. carolina* and *T. mexicana* ssp., with some also delimiting as distinct entities *T. c. carolina*, two *T. c. bauri* clusters, and *T. m. mexicana*. Of note, the runs and clustering algorithms exhibited high within- but not among-clade variability for *T. carolina* and *T. mexicana*, excepting MSW using ISOMDS.

Each clustering algorithm and ordination technique displayed its own inherent characteristics. Essentially, CMDS and the gap statistic were inclined to split subclades of *T. carolina* and *T. mexicana*, ISOMDS and MSW were the most conservative, and t-SNE and HAC were intermediate, though HAC oscillated in agreement with MSW and the gap statistic (Fig. 4). RF, but not t-SNE, varied among the 100 replicates, which was most pronounced for CMDS. Heightened CMDS run variation highlights its inherent sensitivity to low among-group variability (Olteanu *et al.* 2013). Finally, t-SNE optimal *K* increased with perplexity.

VAE initially agreed with BFD* in recognizing *K*=8, clumping *T. o. ornata*/*T. o. luteola* and splitting all other taxa (Fig. 5a). However, assessments of model-loss indicated overfitting in the sense that given enough epochs, the predictive model can perfectly 'learn' the training dataset, with predictive capacity rapidly decreasing for unseen test data. To mitigate, we identified in the model loss plot the transition point, or 'elbow' (Fig. 5b), where predictive accuracy falls off for the test data, such that test *versus* training sets diverge in accuracy. This occurred at a much lower number of sampled epochs (N=2,000) and was subsequently re-initiated at a new termination point. Once overfitting was eliminated, an optimal *K*=3 was derived (Figs. 4, 5c, 5d), in agreement with other UML methods. The model was also tested with N=1,000 epochs (not shown), for which *K*=3 clusters again persisted.

3.5. *Supervised machine learning*

CLADES yielded optimal *K*=2 (P=1.44e^{-4}; Fig. 4; Table S3), but with highly discordant clusters compared with prior results and phylogenomic findings: *Terrapene c. carolina*/*T. c.*

bauri emerged as one species, and the remaining seven taxa (*T. ornata*, *T. mexicana*, and the remaining *T. carolina*) as a consistently paraphyletic second species (Figs. 2-3). The possibility of outliers misleading the delimitations was also explored by removing two *T. c. bauri* and North Carolina *T. c. carolina* that, in a subset of UML runs either formed a potential second cluster or clustered instead with *T. c. bauri*. However, CLADES provided similar output without phylogenetic cohesiveness (K=2; P=6.88e^{-6}) with *T. c. bauri*/*T. c. major* (MS population) as one species, and the remainder forming the second. In both cases, the estimated probability for optimal K was quite low.

3.6. *Relative performance among approaches*

All UML species delimitation methods converged on K=3 if considering RF and t-SNE classifications that did not inter-mix. Three *Terrapene* species (plus *T. nelsoni*) were corroborated (Martin *et al*. 2013, 2020), whereas the clumping of *T. mexicana* and *T. carolina* (Minx 1996) was rejected. Machine learning approaches were also markedly faster than BFD*. For example, RF, t-SNE, and VAE required ~10-30 min run time on a Desktop PC utilizing one Intel i5-3570 CPU core and 16 GB RAM. Comparatively, the twenty BFD* runs required ~4,000 total wall-time hours (~200 hours/model), parallelized across 24-48 threads and utilizing 200 GB RAM/model.

4. DISCUSSION

We observed substantial heterogeneity among machine learning species delimitation approaches in resolving the southeastern *Terrapene* taxa, echoing previous morphological and single-gene results (Milstead 1967, 1969; Milstead & Tinkle 1967; Butler *et al.* 2011; Martin *et al.* 2013). However, groups exhibiting such heterogeneity may indicate the involved taxa are one species, whereas deficit groups may support distinctiveness. Additionally—as argued below—these were interpreted as a more appropriate reflection of taxon-specific biological patterns. Our results represent an empirical test for the *de novo* application of these software packages to other taxonomically-complex systems.

4.1. *Species Delimitation Approaches Reconciled in Terrapene*

Species trees provide a necessary phylogenetic context for species delimitation by outlining hypothetical species compositions and identities. In our case, they underscored classic discordance (Figs. 2-3), previously hypothesized via single-gene sequencing (Martin *et al.* 2013). Differences were apparent in the ancestral progression of taxa, and in transitions between monophyly *versus* paraphyly. Persistent uncertainties include: 1) Placement of *T. c. bauri*; 2) monophyly of *T. mexicana* and 3) *T. o. luteola* subspecies status. Additionally, two individuals were placed in unexpected clades, which was a far smaller proportion than previously seen in single-gene datasets. We suggest the latter are examples of admixture, as both were collected near a southeastern US hybrid zone (Martin *et al.* 2020), and suspect the other idiosyncrasies represent either violations of the model or methodological artifacts.

Impacts of interspecific gene flow on species tree inference are well-characterized, yet surprisingly, seldom modeled explicitly (Leaché *et al.* 2014b; Leaché & Oaks 2017). PoMo, for example, constrains all nodes to the same N_e, a potentially poor assumption given contemporary and possibly historical admixture (Martin *et al.* 2013, 2020). An examination of the species trees alone reiterates previous single-gene taxonomic assessments. However, powerful species delimitation assessments were utilized that provide a far more robust phylogenetic classification.

UML species delimitation inferences were consistent with the most recent phylogenetic hypotheses (Martin *et al.* 2013; this study). *Terrapene o. ornata*/*luteola*, *T. c. carolina*/*bauri*/*major*/*coahuila*, and *T. m. mexicana*/*triunguis* represent what we would consider as species-level variants, within which each encompasses group assignment heterogeneity. *Terrapene m. yucatana* falls within *T. mexicana*, and *T. nelsoni* as sister to *T. ornata*, although their extreme rarity and concomitantly limited sampling (N=1) precluded them from species delimitation analysis. Importantly, the variability in RF and t-SNE results primarily echoed uncertainty found in the species tree analyses, including the distinctiveness of *T. o. luteola* subspecific relationships within *T. carolina*. This variation also corresponded with the proclivities of each algorithm. RF, for example, invokes a randomized process, with stochasticity perhaps exacerbated by the phylogenetic discordance within *T. carolina*. t-SNE was influenced by its perplexity parameter, with a second *T. c. major* group from Mississippi being a minor addition for perplexity=5. This could underscore population structure among Mississippi and Florida *T. c. major*. Finally, delimitations for both were strongly impacted by clustering algorithm, which closely paralleled their own algorithmic tendencies. For example, the gap statistic often over-estimates *K* (Dudoit & Fridlyand 2002; Yan & Ye 2007), MSW under-splits (Şenbabao□lu *et al.*

2014), and outliers and noise particularly impact HAC (Kim *et al*. 2009; Şenbabao□lu *et al.* 2014). This, in turn, may explain the varying extent of agreement between HAC and either MSW or the gap statistic.

We suggest the variability observed among RF and t-SNE runs was due to a lack of divergence within the more variable groups. Mixed classification was not observed among the *T. mexicana*, *T. ornata*, and *T. carolina* groups, excepting RF ISOMDS based on MSW that only differentiated *T. ornata* versus *T. carolina* (Fig. 4). The more conservative nature of ISOMDS was reflected in several original empirical tests, which suggested a restriction to two clustering dimensions may be more sensitive to higher genetic divergences (Derkarabetian *et al.* 2019), as seems to be the case here. Otherwise, variability among taxa was constrained within respective subspecific units.

VAE initially recovered results identical to BFD* (*K*=8), delimiting all taxa except *T. o. ornata*/*T. o. luteola*. To ensure model training occurred appropriately, we more closely inspected model loss and observed overfitting (Fig. 5b). The VAE script includes dropout regularization methods, which randomly thin neural network nodes during model training to reduce overfitting (Srivastava *et al*. 2014). However, regularization parameters can be sensitive to dataset properties (e.g. large *versus* small/noisy *versus* tidy), and may not perform well for every dataset (Gal & Ghahramani 2016; Derkarabetian *et al.* 2019). In model loss exploration, overfitting was mitigated by early termination of model training when loss was at its minimum, though this could also be accomplished by tuning dropout parameters. After correcting for overfitting, VAE also delimited *K*=3 (i.e. *Terrapene mexicana*, *T. carolina*, and *T. ornata* ssp.), much like RF and t-SNE if considering classification heterogeneity to indicate intra-specific relationships.

4.2. *Phylogenetic and biological support of species delimitations*

We suggest that identifying machine learning groups that consistently lack classification overlap is one criterion to delimit species. In our case, these were corroborated as major species tree clades, highlighting their complementary nature. In contrast, inconsistent species delimitation assignments reflect many of the phylogenetic discordances observed in this and previous studies (Butler *et al.* 2011; Martin *et al.* 2013). Potential underlying biological processes include incomplete lineage sorting, ongoing primary divergence, hybridization, and/or complex phylogeographic history [(e.g. isolation followed by secondary contact) (Mayr 1963; Barton & Hewitt 1985; Rieseberg *et al.* 1999, 2007; Coyne & Orr 2004; Sousa & Hey 2013)]. Divergent selection can counteract such processes and reinforce species boundaries (Feder *et al.* 2013). Our species delimitation results are consistent with previously observed divergent selection at candidate loci across *T. carolina* and *T. mexicana*, whereas it was absent for *T. c. carolina* and *T. c. major* (Martin *et al.* 2020). Thus, *T. carolina* and *T. mexicana* may exhibit signatures of secondary contact, whereas *T. c. major* and *T. c. carolina* may be earlier in the divergence process. Alternatively, *T. c. major* could be an intergrade population between *T. c. carolina* and *T. m. triunguis* (Butler *et al.* 2011), though the species trees disagree and two putative parental populations persist (Martin *et al.* 2020). In this sense, *T. c. major* displays fairly disparate habitat preferences, favoring salt marshes on the Gulf Coastal Plain, whereas *T. c. carolina* and *T. m. triunguis* occupy mesic woodlands. The low differentiation between *T. c. major* and *T. c. carolina* may result from *T. c. major* being restricted to the southeast. Here, *T. c. carolina* possibly blocked northward expansion of *T. c. major*, with gene flow persisting across much of *T.*

c. major's smaller range. Alternatively, it may have diverged more recently and now reflects ongoing primary divergence.

4.3. *Comparisons to other empirical studies*

The capability of machine learning species delimitation algorithms to discount population structure while isolating higher-level differentiation is corroborated by other recent studies (Derkarabetian *et al*. 2019; Hedin *et al.* 2020). However, and Derkarabetian *et al.* (2019) and Newton *et al*. (2020) emphasized the importance of integrative approaches, as they were able to identify cryptic species by considering both VAE species delimitation and ecological niche modeling. Given the increasing availability of geological resources, such integrative taxonomic considerations may prove to be invaluable.

Excepting CLADES, the machine learning software used herein also seem robust to hierarchical levels of genetic variation, having differentiated *T. carolina versus T. ornata* and the less divergent *T. m. triunguis*. However, this hierarchical robustness may have limits, as one recent geometric morphometric image-based deep learning study favored inter-generic over inter-specific delimitations (Boer & Vos 2018). On the contrary, another recent study was more accurate in recovering species-level delimitations rather than across genera, which they suggested stemmed from less informed model training in low-diversity families with many unique species. Recent and future work may also illuminate the impact of gene flow and population demography on observed delimitations, processes that MSC approaches do not consider. For example, DELIMITR incorporates models of secondary contact and divergence with gene flow into RF

classifiers for species delimitation (Smith & Carstens 2020). However, empirical tests of DELIMITR tended to agree with the species delimited by BFD* and BP&P, whereas for *Terrapene* RF, t-SNE, and VAE were more conservative than the MSC approaches. It may be that *Terrapene* exhibits stronger population structure than the species included in the DELIMITR applications, which can influence BFD* (Sukumaran & Knowles 2017). Finally, machine learning frameworks may illuminate other potential sources of species tree discordance, with recent applications predicting discordant species trees (Roettger *et al.* 2009), assessing historical introgression despite ongoing gene flow (Burbrink & Gehara 2018), and identifying ILS (Burbrink *et al*. 2020). Nevertheless, RF, t-SNE, and VAE are reported to at least be robust to gene flow, with recent applications showing that they place admixed individuals between parental clusters (Derkarabetian *et al*. 2019; Hedin *et al*. 2020; Newton *et al.* 2020). However, in our case we avoided hybrids (as characterized by NewHybrids; Martin *et al.* 2020) due to frequent introgression in the southeastern *Terrapene* hybrid zone.

4.4. *Conclusions*

UML approaches attempt to identify groups based on inherent structure in the data, and accordingly are a natural extension to the species delimitation problem. In our case, a consensus among UML approaches corroborated other axes of differentiation, whereas MSC-based delimitations over-partitioned the data. Specifically, groups that were not supported by RF, t-SNE, and VAE echoed classic patterns of phylogenetic uncertainty seen among our species trees, which may be affected by previously observed genome-wide differential introgression. Furthermore, in our case it seems likely that the phylogenetic signals affecting discordance are

similarly affecting the machine learning algorithms, which may include a combination of historical biogeographic processes, gene flow, and incomplete lineage sorting. What is clear is that delimiting almost every *Terrapene* taxon, as supported by BFD*, is probably not biologically appropriate. Though MSC methods are undoubtedly still extremely useful, machine learning provides a promising alternative for resolving long-standing biological problems. This may particularly be the case for species that violate MSC model assumptions, as demonstrated by our study system.

ACKNOWLEDGEMENTS

We would like to thank the numerous volunteers, organizations, and agencies that contributed tissue samples (Table S1). We also thank both current students and alumni of the Douglas Lab, as well as University of Arkansas faculty for support, advice, and guidance, to include: A. Alverson, W. Anthonysamy, M. Bangs, J. Beaulieu, J. Koukl, S. Mussmann, J. Pummill, and Z. Zbinden. Sample collections were approved under three Animal Care and Use Committee (IACUC) protocols: #113 (University of Texas at Tyler), and #16160 and #18000 (University of Illinois/Champaign-Urbana). Funding sources included the Lucille F. Stickle Fund of the North American Box Turtle Committee, the American Turtle Observatory (ATO), and two University of Arkansas endowments [the Bruker Professorship in Life Sciences (MRD), and the 21st Century Chair in Global Change Biology (MED)]. The Arkansas High Performance Computing Cluster (AHPCC) and the Jetstream Cloud Service (NSF-XSEDE Research Allocation TG-BIO160065) graciously supplied analytical resources.

5. REFERENCES

Agapow PM, Bininda-Emonds ORP, Crandall KA, Gittleman JL, Mace GM, Marshall JC, and Purvis A (2004) The impact of species concept on biodiversity studies. *Quarterly Review of Biology*, **79**, 161–179.

Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, Van Rosendael AR, and Beecy AN (2019) Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European Heart Journal*, **40**, 1975–1986.

Allendorf FW, Hohenlohe PA, and Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Anderson EC and Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. https://www.bibsonomy.org/bibtex/2b6052877491828ab53d3449be9b293b3/ozborn.

Auffenberg W (1958) Fossil turtles of the genus *Terrapene* in Florida. *Bulletin of the Florida State Museum*, **3**, 53–92.

Auffenberg W (1959) A Pleistocene *Terrapene* hibernaculum, with remarks on a second complete box turtle skull from Florida. *Quarterly Journal of the Florida Academy of Science*, **22**, 49–53.

Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, Veuille M, and Laredo C (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, **10**, S10.

Barton NH and Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Boer MJA and Vos RA (2018) Taxonomic classification of ants (Formicidae) from images using deep learning. *BioRxiv*, 407452.

Breiman L (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Burbrink FT and Gehara M (2018) The biogeography of deep time phylogenetic reticulation. *Systematic Biology*, **67**, 743–755.

Burbrink FT, Grazziotin FG, Pyron RA, Cundall D, Donnellan S, Irish F, Keogh JS, Kraus F, Murphy RW, and Noonan B (2020) Interrogating genomic-scale data for Squamata (lizards, snakes, and amphisbaenians) shows no support for key traditional morphological relationships. *Systematic Biology*, **69**, 502–520.

Butler JM, Dodd Jr. CK, Aresco M, and Austin JD (2011) Morphological and molecular evidence indicates that the Gulf Coast box turtle (*Terrapene carolina major*) is not a distinct evolutionary lineage in the Florida Panhandle. *Biological Journal of the Linnean Society*, **102**, 889–901.

Carstens BC, Pelletier TA, Reid NM, and Satler JD (2013) How to fail at species delimitation. *Molecular Ecology*, **22**, 4369–4383.

Chafin TK, Douglas MR, Martin BT, and Douglas ME (2019) Hybridization drives genetic erosion in sympatric desert fishes of western North America. *Heredity*, **123**, 759–773.

Chafin TK, Martin BT, Mussmann SM, Douglas MR, and Douglas ME (2018) FRAGMATIC: in silico locus prediction and its utility in optimizing ddRADseq projects. *Conservation Genetics Resources*, **10**, 325–328.

Chambers EA and Hillis DM (2019) The multispecies coalescent over-splits species in the case of geographically widespread taxa. *Systematic Biology*, **69**, 184–193.

Chifman J and Kubatko L (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics*, **30**, 3317–3324.

Chou J, Gupta A, Yaduvanshi S, Davidson R, Nute M, Mirarab S, and Warnow T (2015) A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, **16**, S2.

Coates DJ, Byrne M, and Moritz C (2018) Genetic Diversity and Conservation Units: Dealing With the Species-Population Continuum in the Age of Genomics. *Frontiers in Ecology and Evolution*, **6**, 165.

Coyne JA and Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, MA, USA.

Derkarabetian S, Castillo S, Koo PK, Ovchinnikov S, and Hedin M (2019) A demonstration of unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution*, **139**, 106562.

Dodd KC (2001) *North American Box Turtles, A Natural History*. University of Oklahoma Press, Norman, OK, USA.

Dudoit S and Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, **3**, research0036.1.

Eaton DAR and Overcast I (2020) ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, **36**, 2592–2594.

Eaton DAR, Spriggs EL, Park B, and Donoghue MJ (2017) Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology*, **66**, 399–412.

Feder JL, Flaxman SM, Egan SP, and Nosil P (2013) Hybridization and the build□up of genomic divergence during speciation. *Journal of Evolutionary Biology*, **26**, 261–266.

Feldman CR and Parham JF (2002) Molecular phylogenetics of emydine turtles: Taxonomic revision and the evolution of shell kinesis. *Molecular Phylogenetics and Evolution*, **22**, 388–398.

Fraley C and Raftery AE (2002) *MCLUST: Software for model-based cluster and discriminant analysis. Department of Statistics, University of Washington: Technical Report 415, Department of Statistics, University of Washington, Seattle, WA, USA. http://www.stat.washington.edu/raftery*.

Fritz U and Havaš P (2013) Order Testudines: 2013 update. In: Zhang, Z.-Q. (Ed.) Animal Biodiversity: An Outline of Higher-level Classification and Survey of Taxonomic Richness (Addenda 2013). *Zootaxa*, **3703**, 12–14.

Fritz U and Havaš P (2014) On the reclassification of Box Turtles (*Terrapene*): A response to Martin et al. (2014). *Zootaxa*, **3835**, 295–298.

Funk DJ and Omland KE (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.

Gal Y and Ghahramani Z (2016) A theoretically grounded application of dropout in recurrent neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1019–1027.

Garnett ST and Christidis L (2017) Taxonomy anarchy hampers conservation. *Nature*, **546**, 25–27.

Gippoliti S, Cotterill FPD, Zinner D, and Groves CP (2018) Impacts of taxonomic inertia for the conservation of African ungulate diversity: an overview. *Biological Reviews*, **93**, 115–130.

Hedin M, Foldi S, and Rajah-Boyer B (2020) Evolutionary divergences mirror Pleistocene paleodrainages in a rapidly-evolving complex of oasis-dwelling jumping spiders (Salticidae, *Habronattus tarsalis*). *Molecular Phylogenetics and Evolution*, **144**, 106696.

Herrmann H and Rosen PC (2009) Conservation of aridlands turtles III: preliminary genetic studies of the desert box turtle and yaqui slider. *Sonoran Herpetologist*, **22**, 38–43.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, and Vinh LS (2017) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, **35**, 518–522.

Isaac NJB, Mallet J, and Mace GM (2004) Taxonomic inflation: Its influence on macroecology and conservation. *Trends in Ecology and Evolution*, **19**, 464–469.

Iverson JB, Meylan PA, and Seidel ME (2017) Testudines—Turtles. In: *Scientific and Standard English Names of Amphibians and Reptiles of North America North of Mexico, with Comments Regarding Confidence in Our Understanding* (ed Crother BI), pp. 82-91. SSAR Herpetological Circular 43.

Jombart T and Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070–3071.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, and Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.

Kass RE and Raftery AE (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kaufman L and Rousseeuw P (1987) Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, 405–416.

Kim E-Y, Kim S-Y, Ashlock D, and Nam D (2009) MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, **10**, 260.

Kishino H and Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, **29**, 170–179.

Kishino H, Miyata T, and Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, **31**, 151–160.

Kruskal JB and Wish M (1978) *Multidimensional Scaling*. Sage Publisiinig, Thousand Oaks, CA, USA.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, and Linkem CW (2015) Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution*, **7**, 706–719.

Leaché AD, Fujita MK, Minin VN, and Bouckaert RR (2014a) Species delimitation using genome-wide SNP data. *Systematic Biology*, **63**, 534–542.

Leaché AD, Harris RB, Rannala B, and Yang Z (2014b) The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, **63**, 17–30.

Leaché AD and Oaks JR (2017) The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 69–84.

Livingstone DJ, Manallack DT, and Tetko I V. (1997) Data modelling with neural networks: Advantages and limitations. *Journal of Computer-Aided Molecular Design*, **11**, 135–142.

Maaten L van der and Hinton G (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Mace GM (2004) The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359**, 711–719.

Martin BT, Bernstein NP, Birkhead RD, Koukl JF, Mussmann SM, and Placyk JS (2013) Sequence-based molecular phylogenetics and phylogeography of the American box turtles (*Terrapene* spp.) with support from DNA barcoding. *Molecular Phylogenetics and Evolution*, **68**, 119–134.

Martin BT, Bernstein NP, Birkhead RD, Koukl JF, Mussmann SM, and Placyk Jr JS (2014) On

the reclassification of the *Terrapene* (Testudines: Emydidae): a response to Fritz & Havaš. *Zootaxa*, **3835**, 292–294.

Martin BT, Douglas MR, Chafin TK, Placyk JS, Birkhead RD, Phillips CA, and Douglas ME (2020) Differential introgression supports thermal adaptation and candidate genes shaping species boundaries in North American box turtles (*Terrapene* spp.). *bioRxiv*, 752196.

Mayr E (1963) *Animal Species and Evolution*. Belknap Press at Harvard University Press, Cambridge, MA.

Milstead WW (1967) Fossil box turtles (*Terrapene*) from central North America, and box turtles of eastern Mexico. *Copeia*, **1967**, 168–179.

Milstead WW (1969) Studies on the evolution of the box turtles (genus *Terrapene*). *Bulletin of the Florida State Museum, Biological Science Series*, **14**, 1–113.

Milstead WW and Tinkle DW (1967) *Terrapene* of Western Mexico, with comments on species groups in the genus. *Copeia*, **1967**, 180–187.

Minh BQ, Hahn MW, and Lanfear R (2018) New methods to calculate concordance factors for phylogenomic datasets. *bioRxiv*, 487801.

Minx P (1992) Variation in phalangeal formulas in the turtle genus *Terrapene*. *Journal of Herpetology*, **26**, 234–238.

Minx P (1996) Phylogenetic relationships among the box turtles, Genus *Terrapene*. *Herpetologica*, **52**, 584–597.

Morrison WR, Lohr JL, Duchen P, Wilches R, Trujillo D, Mair M, and Renner SS (2009) The impact of taxonomic change on conservation: Does it kill, can it save, or is it just irrelevant? *Biological Conservation*, **142**, 3201–3206.

Newton LG, Starrett J, Hendrixson BE, Derkarabetian S, and Bond JE (2020) Integrative species delimitation reveals cryptic diversity in the southern Appalachian Antrodiaetus unicolor (Araneae: Antrodiaetidae) species complex. *Molecular Ecology*, **29**, 2269–2287.

Nguyen L-T, Schmidt HA, von Haeseler A, and Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.

Nielsen R, Paul JS, Albrechtsen A, and Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443.

Olteanu M, Nicolas V, Schaeffer B, Denys C, Missoup A, Kennis J, and Laredo C (2013) Nonlinear projection methods for visualizing barcode data and application on two data sets. *Molecular Ecology Resources*, **13**, 976–990.

Pei J, Chu C, Li X, Lu B, and Wu Y (2018) CLADES: A classification-based machine learning method for species delimitation from population genetic data. *Molecular Ecology Resources*, **18**, 1144–1156.

Peterson BK, Weber JN, Kay EH, Fisher HS, and Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLOS ONE*, **7**, e37135.

Philippe H, Brinkmann H, Lavrov D V., Littlewood DTJ, Manuel M, Wörheide G, and Baurain D (2011) Resolving difficult phylogenetic questions: why more sequences are not enough (D Penny, Ed,). *PLOS Biology*, **9**, e1000602.

R Development Core Team (2018) R: A language and environment for statistical computing. https://cran.r-project.org/.

Rambaut A (2014) FigTree v1.4.2. http://tree.bio.ed.ac.uk/software/figtree/.

Rambaut A, Drummond AJ, Xie D, Baele G, and Suchard MA (2018) Posterior summarization in bayesian phylogenetics using Tracer 1.7 (E Susko, Ed,). *Systematic Biology*, **67**, 901–904.

Rieseberg LH, Kim S-C, Randell RA, Whitney KD, Gross BL, Lexer C, and Clay K (2007) Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, **129**, 149–165.

Rieseberg LH, Whitton J, and Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, and Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, **67**, 93–104.

Roettger M, Martin W, and Dagan T (2009) A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Molecular Biology and Evolution*, **26**, 1931–1939.

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Salichos L and Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, **497**, 327–333.

Schrempf D, Minh BQ, De Maio N, von Haeseler A, and Kosiol C (2016) Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, **407**, 362–370.

Şenbabao□lu Y, Michailidis G, and Li JZ (2014) Critical limitations of consensus clustering in class discovery. *Scientific Reports*, **4**, 1–13.

Shaffer HB, Minx P, Warren DE, Shedlock AM, Thomson RC, Valenzuela N, Abramyan J, Amemiya CT, Badenhorst D, and Biggar KK (2013) The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology*, **14**, R28.

Shepard RN, Romney AK, and Nerlove SB (1972) *Multidimensional Scaling: Theory and*

Applications in the Behavioral Sciences: I. Theory. Seminar Press.

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, **51**, 492–508.

Shimodaira H and Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, **16**, 1114–1116.

Simmons MP and Goloboff PA (2014) Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. *Molecular Phylogenetics and Evolution*, **78**, 334–348.

Smith ML and Carstens BC (2020) Process based species delimitation leads to identification of more biologically relevant species. *Evolution*, **74**, 216–229.

Sousa V and Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, **15**, 1929–1958.

Stanton DWG, Frandsen P, Waples RK, Heller R, Russo IRM, Orozco-terWengel PA, Pedersen CET, Siegismund HR, and Bruford MW (2019) More grist for the mill? Species delimitation in the genomic era and its implications for conservation. *Conservation Genetics*, **20**, 101–113.

Stephens PR and Wiens JJ (2003) Ecological diversification and phylogeny of emydid turtles. *Biological Journal of the Linnaean Society*, **79**, 577–610.

Strimmer K and Rambaut A (2002) Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **269**, 137–142.

Sukumaran J and Knowles LL (2017) Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, 1607–1611.

Sullivan BK, Douglas MR, Walker JM, Cordes JE, Davis MA, Anthonysamy WJB, Sullivan KO, and Douglas ME (2014) Conservation and management of polytypic species: The Little Striped Whiptail Complex (*Aspidoscelis inornata*) as a case study. *Copeia*, **2014**, 519–529.

Suryachandra P and Reddy PVS (2016) Comparison of machine learning algorithms for breast cancer. In: *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*, pp. 1–6. Institute of Electrical and Electronics Engineers Inc.

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

Tibshirani R, Walther G, and Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),*

63, 411–423.

Vane-Wright RI, Humphries CJ, and Williams PH (1991) What to protect? Systemations and the agony of choice. *Biological Conservation*, **55**, 235–254.

Yan M and Ye K (2007) Determining the number of clusters using the weighted gap statistic. *Biometrics*, **63**, 1031–1037.

Yang Z and Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, **107**, 9264–9269.

Zachos FE (2018) Species concepts, species delimitation and the inherent limitations of taxonomy. *Journal of Genetics*, **97**, 811–815.

Zachos FE, Apollonio M, Bärmann E V., Festa-Bianchet M, Göhlich U, Habel JC, Haring E, Kruckenhauser L, Lovari S, McDevitt AD, Pertoldi C, Rössner GE, Sánchez-Villagra MR, Scandura M, and Suchentrunk F (2013) Species inflation and taxonomic artefacts-A critical comment on recent trends in mammalian classification. *Mammalian Biology*, **78**, 1–6.

DATA ACCESSIBILITY

The raw ddRADseq data is available on the GenBank Nucleotide Database at https://www.ncbi.nlm.nih.gov/bioproject/563121 (BioProject ID: 563121) [to be made public upon publication]. Additional sequence alignments, Supplemental Appendix 1, and supplementary materials will be available from a Dryad Digital Repository.

AUTHOR CONTRIBUTIONS

BTM and TKC designed the research, laboratory protocols, and scripts. BTM conducted the lab work and bioinformatic analyses, analyzed the data, and wrote the manuscript. MRD and MED were the study supervisors, guided the study design, and provided funding. JSP facilitated the collection of thousands of *Terrapene* tissues and provided methodological expertise. RDB collected hundreds of *Terrapene* tissues from southeastern North America and facilitated the collection of many additional individuals. CAP provided many of the *T. ornata* tissues and provided sampling expertise. All authors edited and revised the manuscript

Table 1: Topology tests for four hypothesized North American box turtle (*Terrapene*) phylogenies. The morphology and Sanger sequencing trees are based on previously published data (Minx 1996; Martin *et al*. 2013), whereas trees representing SVDQUARTETS and POMO (Polymorphism-Aware Model) were generated from ddRADseq data (Fig. 3, this study). Bolded *P*-values with an asterisk (*) indicate supported trees (P>0.05 or highly weighted).

Guide Tree	Log-likelihood	ΔLL	bp-RELL	p-KH	p-SH	c-ELW	p-AU
Morphology	-2639307.9	601.5	0.00	0.01	0.02	0.00	0.01
PoMo	-2639200.2	493.8	0.01	0.03	**0.06***	0.01	0.03
Sanger	-2638898.4	192.0	**0.23***	**0.24***	**0.41***	**0.23***	**0.26***
SVDquartets	-2638706.4	0.0	**0.75***	**0.76***	**1.00***	**0.75***	**0.81***

ΔLL=change in log-likelihood
bp-RELL=Bootstrap proportions using RELL method (weights sum to 1)
p-KH=Kishino-Hasegawa test
p-SH=Shimodaira-Hasegawa test
c-ELW=Expected likelihood weight (sum to 1)
p-AU=Approximately unbiased test

Table 2: Species delimitation results from BFD* (Bayes Factor Delimitation, *with genomic data) in 37 North American box turtle (*Terrapene* spp.) individuals. BFD* was run with 179 unlinked, bi-allelic single nucleotide polymorphisms (SNPs) generated using ddRADseq. Bayes factors (BF) were used to identify support among models and were calculated as 2 X (MLE_1 - MLE_2). An * indicates the best supported models; "+" shows taxa that were grouped together; "/" delineates multiple groupings. DS=Desert (*T. o. luteola*), ON=Ornate (*T. o. ornata*), EA=Woodland (*T. c. carolina*), GUFL=Gulf Coast (*T. c. major*) from Florida, GUMS=Gulf Coast (*T. c. major*) from Mississippi, CH=Coahuilan (*T. coahuila*), FL=Florida (*T. c. bauri*), TT=Three-toed (*T. m. triunguis*), and MX=Mexican (*T. m. mexicana*) box turtles. East=all *T. carolina* and *T. mexicana*, West=all *T. ornata*. The outgroup (not shown) included the spotted turtles (*Clemmys guttata*).

BFD* Model	**MLE†**	***K*‡**	**Rank§**	**BF¶**
All Separate*	-2403.39	10	1	-
DS+ON*	-2404.34	9	2	1.90
EA+GUFL	-2417.84	9	3	28.91
GUMS+GUFL	-2427.58	9	4	48.39
GUMS+CH	-2448.61	9	5	90.44
GUMS+CH/GUFL+EA	-2461.28	8	6	115.79
GUMS+GUFL+CH	-2489.62	8	7	172.45
EA+FL	-2511.83	9	8	216.89
GUMS+GUFL+CH+EA	-2514.86	7	9	222.94
EA+FL+GUFL	-2552.22	8	10	297.66
EA+FL/CH+GUMS	-2555.16	8	11	303.53
EA+FL+GUFL/CH+GUMS	-2594.91	7	12	383.04
EA+CH+GUMS+GUFL+TT	-2607.72	6	13	408.66
EA+CH+GUMS+GUFL+MX	-2657.48	6	14	508.19
EA+FL+CH+GUMS+GUFL	-2693.37	6	15	579.96
EA+CH+GUMS+GUFL+TT+MX	-2719.02	5	16	631.27
ON+DS/EA+TT+MX+CH+GUMS+GUFL/FL	-2720.23	4	17	633.69
EA+FL+CH+GUMS+GUFL+TT	-2800.56	5	18	794.35
EA+FL+CH+GUMS+GUFL+TT+MX	-2926.20	4	19	1045.62
East/West	-2926.56	3	20	1046.35

†MLE=Marginal likelihood estimates
‡*K*=# tips
§Rank=model ranking based on MLE (lower=better)
¶BF=Bayes factors

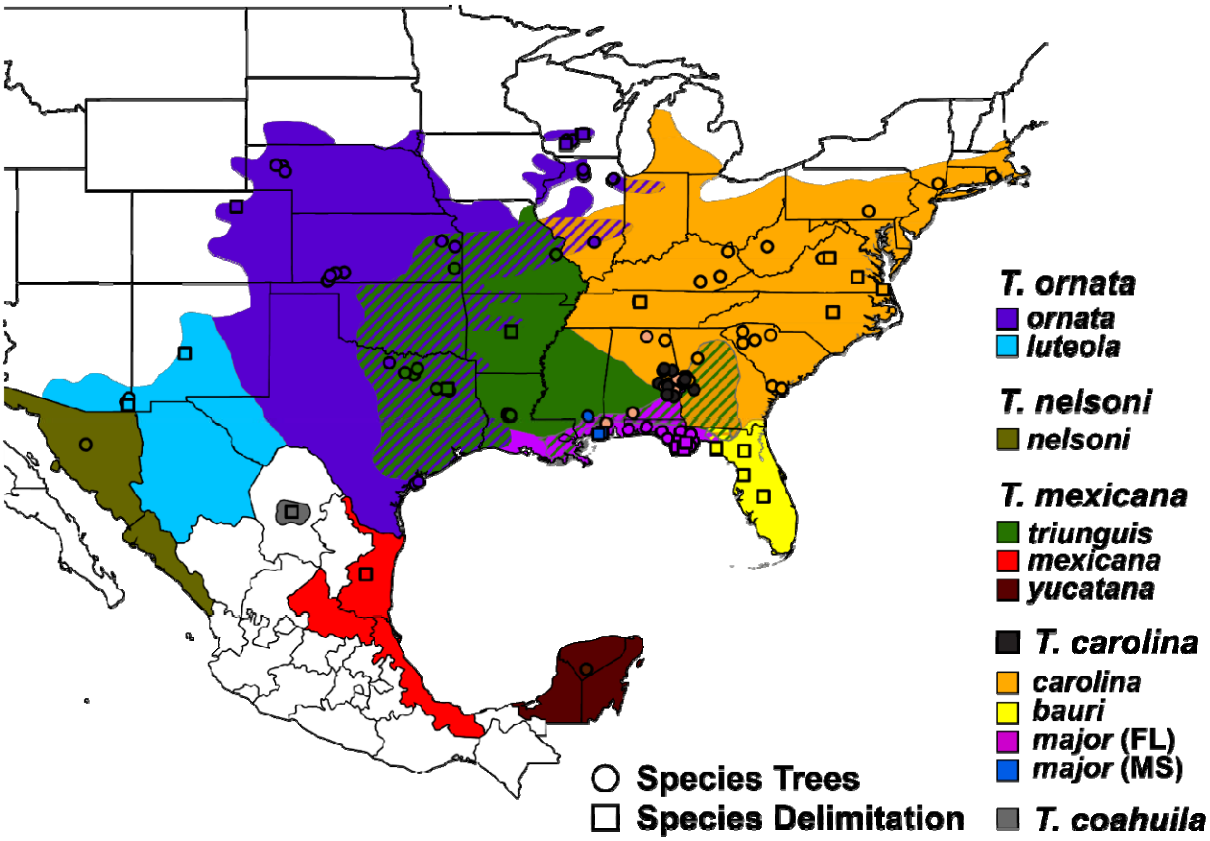

Figure 1: Range map for North American box turtles, *Terrapene,* depicting sample localities. Cross-hatched areas indicate known hybrid zones. Circles represent samples used for the species trees, whereas squares were also used only in species delimitation analyses. Headings and subheadings correspond to species and subspecies, respectively, following (Martin *et al.* 2013), and include the Ornate (*T. ornata ornata*), Desert (*T. o. luteola*), Spotted (*T. nelsoni*), Three-toed (*T. mexicana triunguis*), Mexican (*T. m. mexicana*), Yucatan (*T. m. yucatana*), Woodland (*T. carolina carolina*), Florida (*T. c. bauri*), Gulf Coast from distinct Florida and Mississippi populations (*T. c. major*), and Coahuilan (*T. coahuila*) box turtles. Localities with black circles indicate *T. carolina* samples lacking subspecific field identifications.

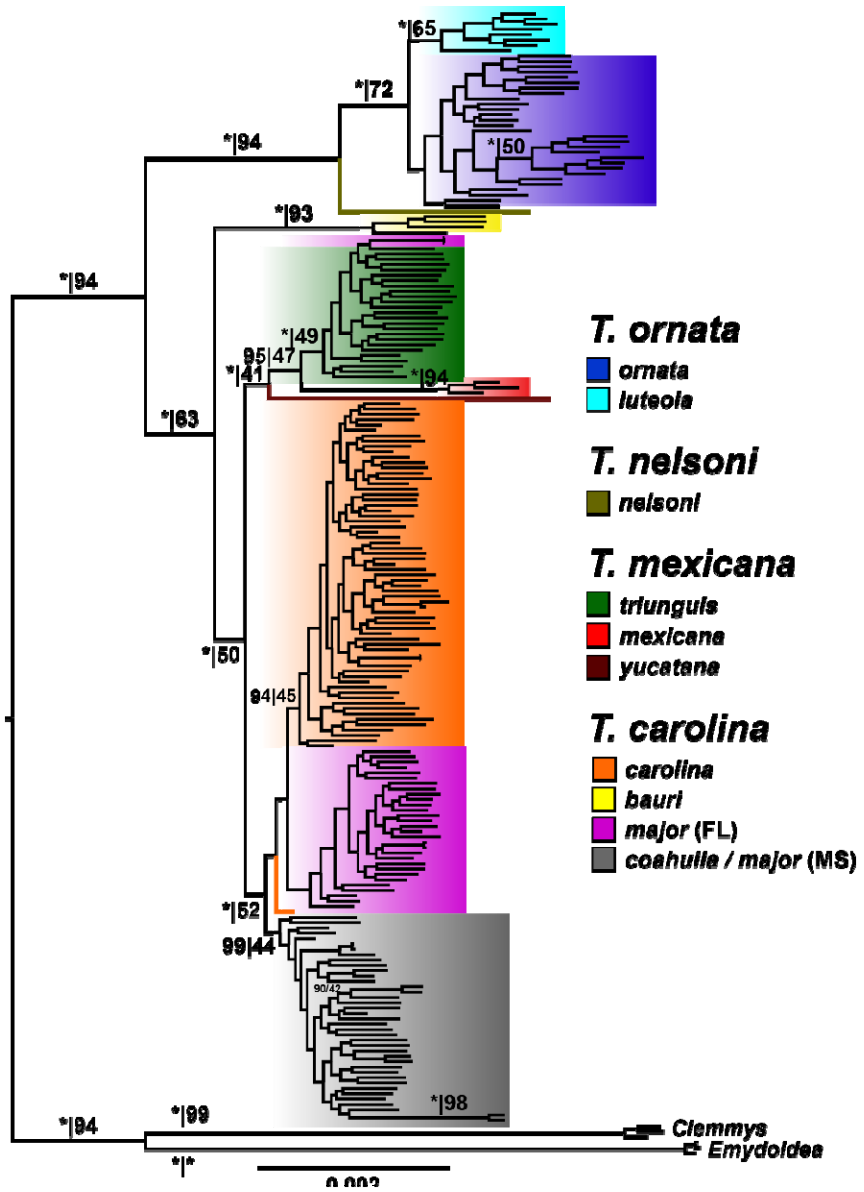

Figure 2: Maximum likelihood phylogeny (IQ-TREE) reflecting relationships among 214 *Terrapene* samples. The tree was generated from 11,962 unlinked ddRADseq loci with 1,000 ultrafast bootstrap (UFB) replicates. Site concordance-factors (sCF) were calculated from 100 quartets randomly sampled from internal branches. Branch support values represent UFB replicates and sCF on the left and right of each vertical line, respectively. UFBs≥95% and sCF≥50% were considered strong support. UFBs<90 and sCF<40 were omitted for visual clarity, with the latter rounded to the nearest integer. Asterisks (*) indicate 100% support. Legend headers and subheadings depict species and subspecies from Martin et al. (2013): Ornate (*Terrapene ornata ornata*), Desert (*T. o. luteola*), Spotted *(T. nelsoni*), Three-toed (*T. mexicana triunguis*), Mexican (*T. m. mexicana*), Yucatan (*T. m. yucatana*), Florida *(T. carolina bauri*), Gulf Coast (*T. c. major*; two populations from Florida and Mississippi), Woodland (*T. c. carolina*), and Coahuilan (*T. coahuila*) box turtles. The Spotted (*Clemmys guttata*) and Blanding's (*Emydoidea blandingii*) turtles were used as outgroups.

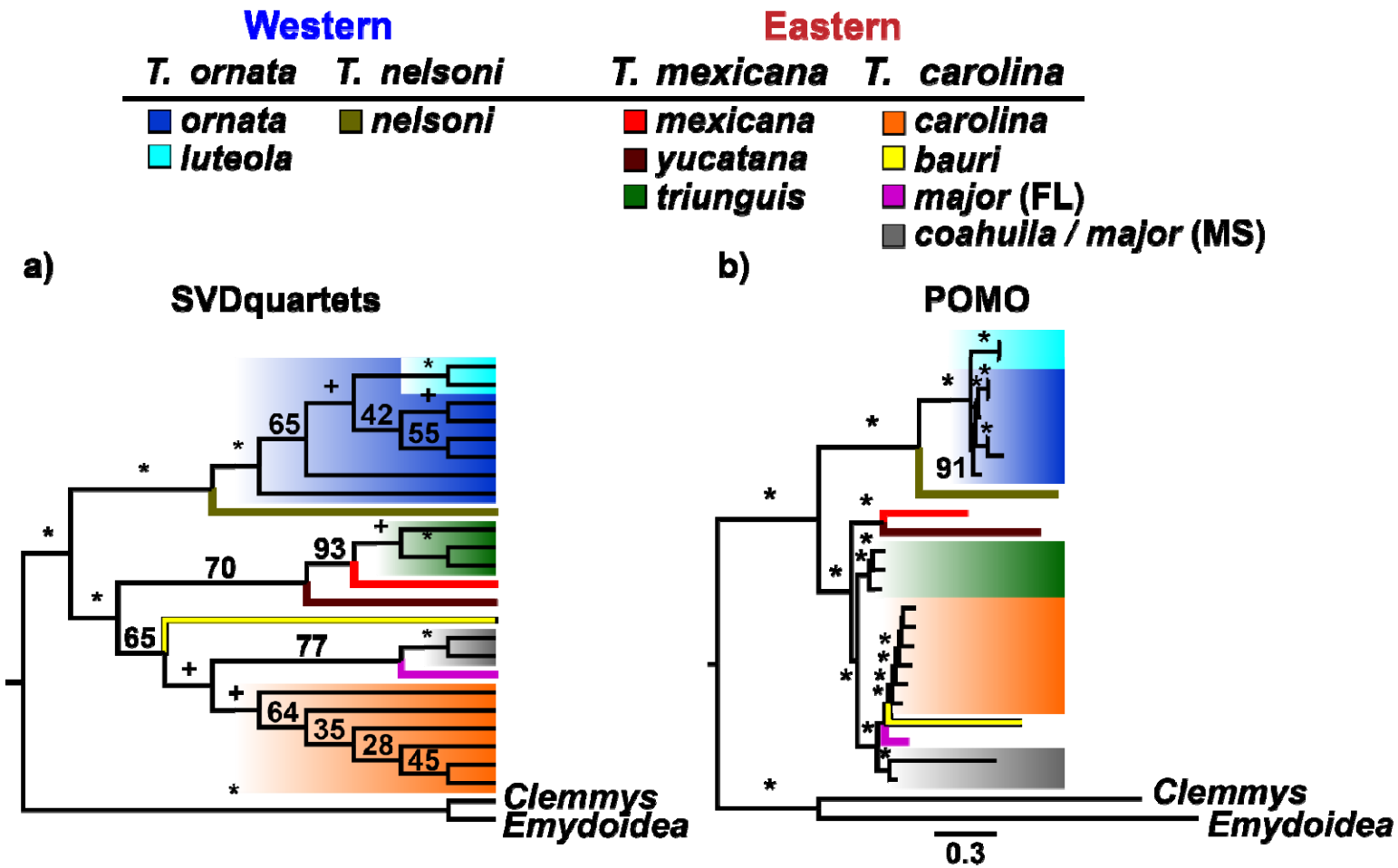

Figure 3: Species trees based on 10,299 unlinked ddRADseq loci depicting phylogenetic relationships of 214 North American box turtle (*Terrapene* spp.) samples. Species trees were generated using a) SVDQUARTETS and b) POMO (Polymorphism Aware Model). Each tree contained 26 populations grouped by specific or subspecific designations (if available), and U.S. or Mexican State locality. Spotted (*Clemmys guttata*) and Blanding's (*Emydoidea blandingii*) turtles were used as outgroups to root the trees. * and + above branches represent nodes with 100% and ≥95% bootstrap support.

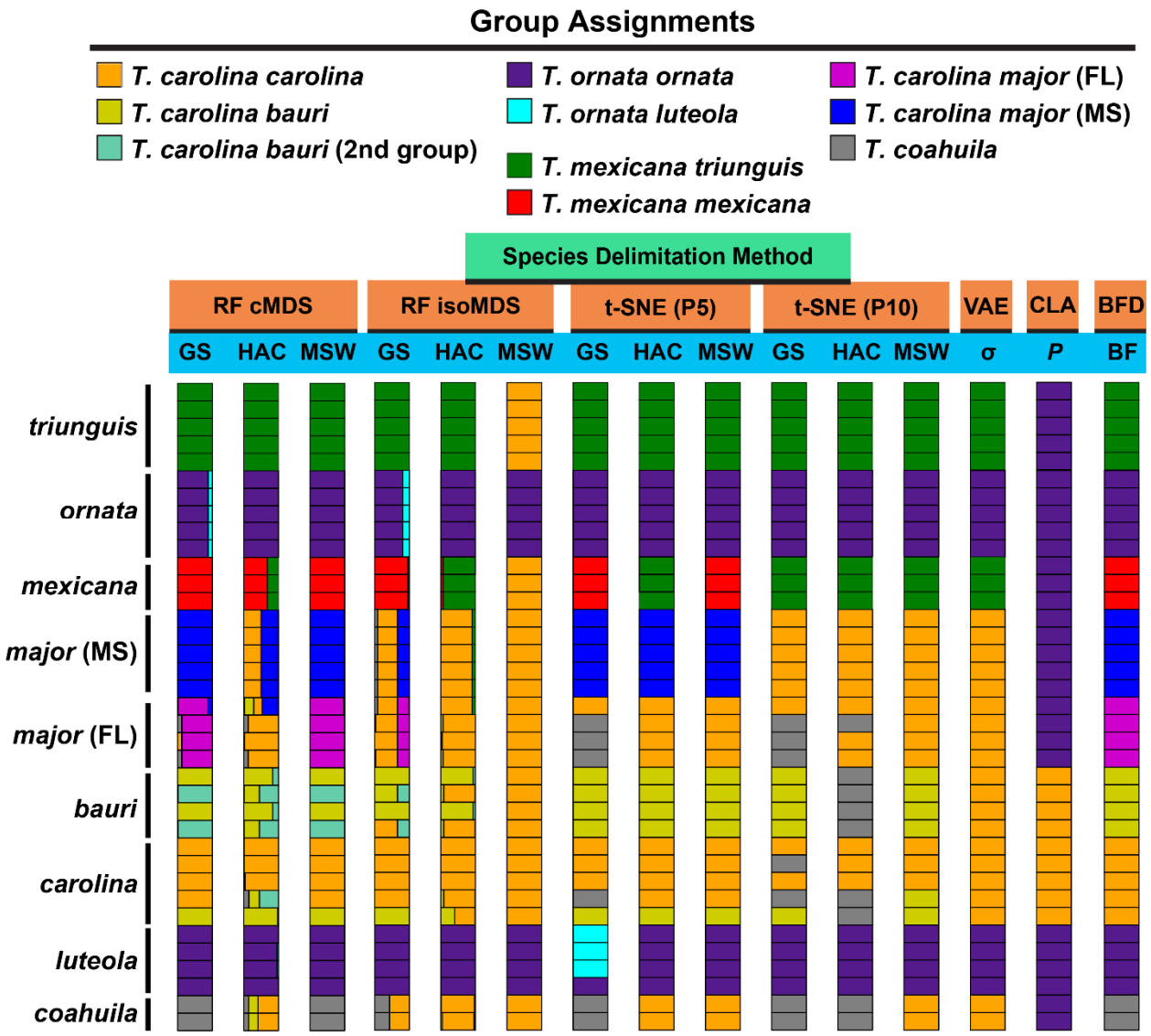

Figure 4: Species delimitations among 37 *Terrapene* samples using 7,395 unlinked ddRADseq single nucleotide polymorphisms (SNPs). Species delimitation groups were derived using unsupervised (UML) and supervised (SML) machine learning algorithms, plus a multispecies coalescent (MSC) approach. UML algorithms include RF=random forest, visualized using CMDS and ISOMDS (classic and isotonic multidimensional scaling) ordination, t-SNE=t-distributed stochastic neighbor embedding, and VAE=variational autoencoders. Each CMDS, ISOMDS, and t-SNE column represents a summarization of 100 independent runs, with colors indicating percent group assignments per method. Mixed colors show clustering variation among runs. t-SNE was run with perplexity settings of five and ten (P5 and P10). RF and t-SNE optimal *K*'s were assessed using hierarchical agglomerative clustering (HAC), partition around medoids using mean silhouette widths (MSW) and the gap statistic (GS), whereas standard deviation (σ) overlap was used for VAE. Optimal *K*'s for CLA=CLADES (SML) and BFD=Bayes Factor Delimitation were determined using probabilities (*P*) and Bayes Factors (BF).

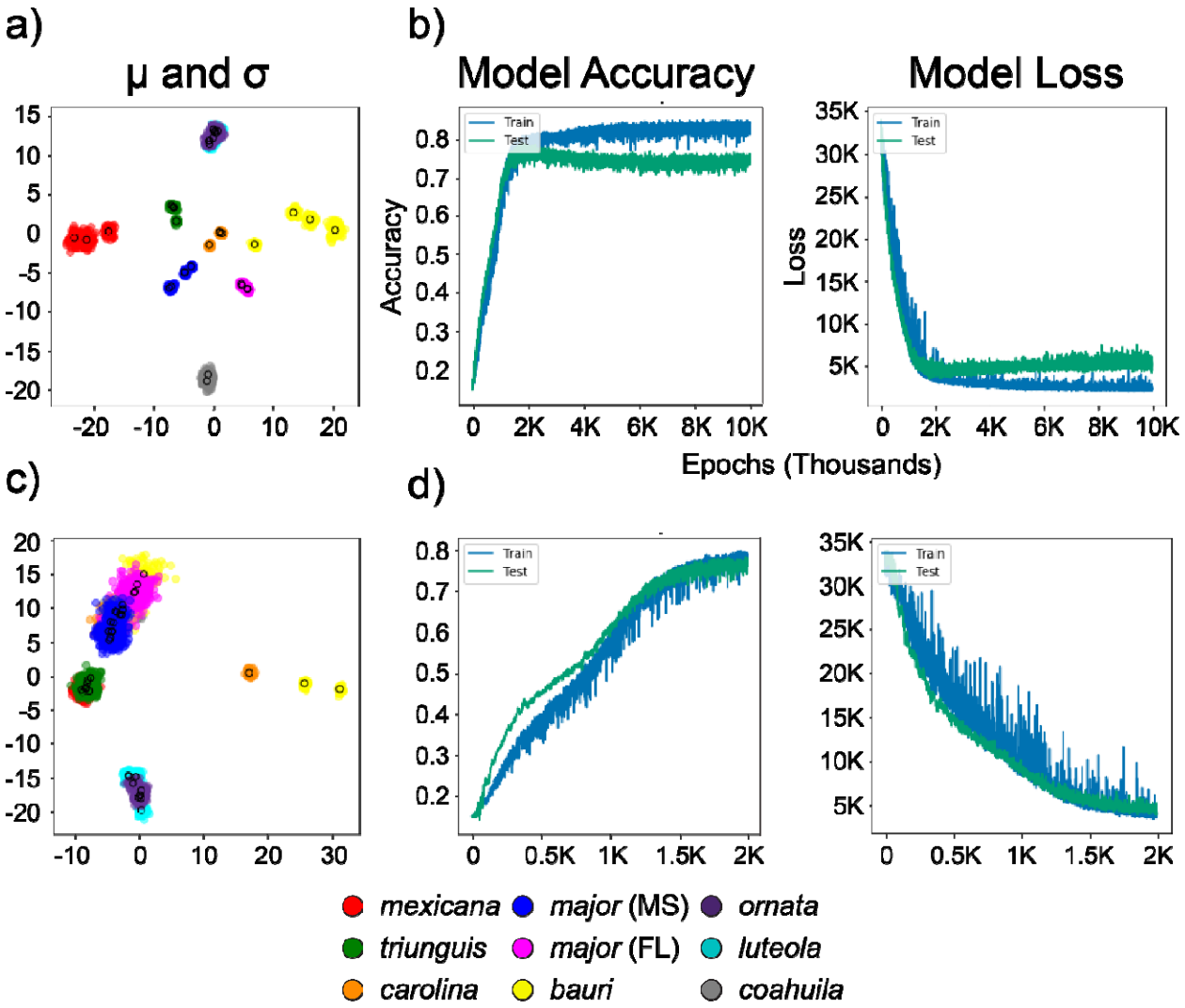

Figure 5: Results from variational autoencoder (VAE) machine learning species delimitation among 37 *Terrapene* samples and 7,395 unlinked ddRADseq single nucleotide polymorphisms (SNPs). Each circle represents the mean (μ) of one individual in the reconstructed parameter space, and the surrounding amorphous area are the standard deviations (σ) across a) 10,000 and c) 2,000 epochs. The model accuracy and loss traces depict the fit of the model to test (green) and training (blue) data across b) 10,000 and d) 2,000 epochs. The colors depict eight subspecies across *T. mexicana*, *T. carolina*, and *T. ornata*, plus monotypic *T. coahuila*.