Machine learning substantiates biologically meaningful species delimitations in the phylogenetically complex North American box turtle genus *Terrapene*

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*

[2]*Department of Biology, University of Texas, Tyler, Texas, 75799, USA*

[3]*Alabama Science in Motion, Auburn University, Auburn, AL 36849, USA*

[4]*Illinois Natural History Survey, Prairie Research Institute, University of Illinois, Champaign, 61820, USA*

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*

** Corresponding Author:*

Bradley T. Martin, Voice: (479) 575-4176; E-mail: btm002@uark.edu

Abstract.—Model-based approaches to species delimitation are constrained both by computational capacities as well as by algorithmic assumptions that are frequently violated when applied to biologically complex systems. An alternate approach, demonstrated herein, employs machine learning (=ML) approaches from which species limits are derived without an explicit definition of an underlying species model. By doing so, we demonstrate the capacity of these approaches to designate phylogenomically and biologically relevant groups, using North American box turtles (*Terrapene* spp.) as an example. Several different ML-based and traditional species delimitation algorithms were invoked to parse a large SNP dataset derived from ddRAD sequencing. Our results illuminate two major findings. First, more traditional model-based approaches perform poorly, a likely reflection of systematic biases inherent in their formulation. Multispecies coalescent methods consistently over-split *Terrapene*, particularly given prior evidence and our own phylogenetic results. Second, results from ML and clustering algorithms consistently reiterated the presence of clades that were well-supported in prior species tree analyses. In summary, we highlight both the strengths and limitations of ML algorithms, and in doing so, explore appropriate approaches to data manipulation and model fit. Our study was accomplished within the context of a well-characterized empirical system that allowed a direct contrast between ML *versus* traditional approaches. It allowed the utility of ML-methods to be underscored for species delimitation and serves as a study case from which guidelines implicit to ML methods could be applied to other study systems.

Keywords: BFD, classification, multispecies coalescent, phylogenomics, species tree, ddRAD

Delineating species is undeniably crucial to our understanding of systematics, ecology, and the evolutionary process. The pivotal role of species as the currency of biodiversity is well-known, as are inconsistencies in the application of what constitutes a species (e.g. a 'multiplicity' of species definitions; Zachos 2018). This creates downstream issues for conservation and management (Mace 2004), where spurious 'splitting' or 'lumping' of taxa are impediments to equitable allocation of limited conservation resources. For example, over-splitting risks a redundant proliferation of threatened or endangered taxa (Zachos et al. 2013; Sullivan et al. 2014). It can also conflate species listings and recovery goals that are most appropriately managed at separate scales along the species-population continuum (Coates et al. 2018).

On the other hand, inappropriate lumping serves to mask potential extinctions, as well as the recognition of adaptive differentiation (Stanton et al. 2019). This can create a biased representation of '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 promote a disproportionate representation of 'species at peril' that subsequently drives an inefficient allocation of resources (Morrison et al. 2009; Garnett and Christidis 2017). This problem has received much attention, often under divisive terms such as 'taxonomic inflation' (Agapow et al. 2004; Isaac et al. 2004). Nevertheless, epistemological decisions regarding the definition and delineation of species are a critical dimension in the 'agony of choice' that pervades the allocation of resources in conservation (Vane-Wright et al. 1991; Stanton et al. 2019). Delimiting species, then, has clear implications for how we allocate finite resources across programs to delineate biodiversity, and work to recover and protect it.

Earlier work on species delimitation relied on few genes (or markers) resulting in limited scope. Although genomic approaches have shown much promise in this area (Allendorf et al. 2010), the issue of conflicting genome-wide signals that stem from incomplete lineage sorting (ILS) and gene flow (Funk and Omland 2003) must still be dealt with. Contemporary species delimitation software utilizes the multispecies coalescent (MSC) that relies upon a probabilistic approach for the modeling of gene tree conflicts (Yang and Rannala 2010). However, some models assume all such conflicts are the result of ILS, and in so doing, ignore other important sources, such as introgressive hybridization.

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

Machine learning (ML) is broadly divided into two components: supervised (SML) and unsupervised (UML). The former requires that a classification model be ‘trained’ with regard to *a priori* designations, from which a classification model is derived and optimized for the subsequent assignment of ‘unknown’ data. A popular SML approach invokes support vector machines (SVM), in an attempt to partition 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. In addition, SVM is often computationally demanding, and hence slow with respect to alternative approaches (Suryachandra and 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), t-distributed stochastic neighbor embedding (t-SNE), and variational autoencoders (VAE), 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 (i.e. over-training the model such that it does not generalize well with new data). One stipulation is that features must be of low occupancy and without undue noise (Breiman 2001; 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 and Hinton 2008)].

By contrast, VAE uses neural networks in an attempt to 'learn' or reconstruct multidimensional data patterns from a compressed, low-dimensionality (='encoded') representation. Again, the approach is similar in a conceptual way to the dimensionality-reduction employed by various ordination techniques, but without linear and orthogonal constraint being imposed upon the informative components. This approach may thus be more

appropriate for the capture of variability within highly complex data. Yet, as above, 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, in contrast to SML, do not require *a priori* designations from which to train a classification model, yet may still be sensitive to priors and parameter settings. Given this, guidelines for appropriate application must be clearly defined, particularly with regard to complex, empirical datasets. Additionally, one metric 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 noise in the data and choice of parameters. However, empirical tests that employ machine learning for species delimitation are lacking, as are guidelines for their application and interpretation. Here we demonstrate these approaches and provide necessary guidelines by evaluating a group of species historically recalcitrant to taxonomic resolution.

The Convoluted Evolutionary History of Terrapene

North American box turtles (Emydidae: *Terrapene*) are primarily terrestrial, with a common name based on a single synapomorphy, an anterior ventral hinge that allows the plastron (bottom part of shell) to be closed dorsally 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* is split into 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*) that range 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 of the U.S. and 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 and Parham 2002; Stephens and 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*. *Terrapene carolina carolina* is sister to *T. c. major*/*T. coahuila*, although potential gene flow is 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. 2019). 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 with caution 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 that occurs within a hybrid zone in the southeastern U.S., with four taxa potentially involved (Auffenberg 1958, 1959; Milstead and Tinkle 1967; Milstead 1969). Some researchers (Fritz and Havaš 2013, 2014) interpret reproductive semi-permeability as evidence for lumping the southeastern taxa as a single species. However, an adaptive capacity to reinforce species boundaries in some southeastern *Terrapene* has been suggested as a reason for re-examining their classificatory status, despite the presence of ongoing gene flow (Martin et al. 2014, 2019). Alternatively, the close phylogenetic relationship between *T. c. major* and *T. coahuila* is less well understood. This may result from the '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, yield a well-characterized phylogenetic hypothesis. Of note is the fact that our particular taxonomic situation emphasizes the potential for overfitting, as well as the necessity for defining optimality criteria across different methods. Finally, we discuss best practices for deploying a machine learning framework for species delimitation and identify available algorithms that are potentially too opaque to accomplish this task.

MATERIALS & METHODS

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, all 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) and eluted to 100uL. Extracted DNA was quantified using Qubit (Thermo Fisher Scientific) broad-range dsDNA fluorometry, and tested for high-molecular weight DNA using gel electrophoresis on 2% agarose.

DNA Library Preparation

We first estimated the expected number of loci recovered via double digest restriction-associated DNA 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 so as to retain as many loci as possible while promoting an elevated sequencing depth. We also used *in silico* digest to identify potential over-enrichment of fragments, indicative of restriction sites lying within repetitive genomic elements, with candidate size selection designed to avoid them (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 DNA of sufficient quantity (>= 50 ng/uL) were processed via ddRADseq protocol (Peterson et al. 2012). Between 500-1,000ng of genomic DNA per samples 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. 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.

Sequence Quality Control and Assembly

FASTQC v. 0.11.5 was used to assess sequence quality (Andrews 2010), with IPYRAD v. 0.7.28 employed to demultiplex the raw sequences and align reads (Eaton and 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 filtered out, with additional strict 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 so as 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 individual sample (the latter two instances indicating over-merged paralogs).

Phylogenomic Inference

To assess differences in phylogenetic inference, we generated species trees using three contemporary algorithms. Where possible, we only utilized individuals distant from contact zones (characterized in Martin et al. 2019), so as to mitigate impact of contemporary gene flow on phylogenetic inference. However, *T. c. major* occurs exclusively in the southeastern United States, and thus overlaps substantially with three co-occurring taxa. We thus partitioned *T. c. major* into subsets based on *a priori* identification of pure individuals, as modelled using NEWHYBRIDS (Anderson and Thompson 2002; detailed in Martin et al. 2019).

SVDQUARTETS.—Maximum likelihood phylogenies have been commonly produced for decades, yet the increased use of large-scale SNP datasets often inflates the bootstrap support for concatenated phylogenomic datasets (Salichos and Rokas 2013; Simmons and Goloboff 2014).

Coalescent-based approaches that account for independent gene tree histories are more applicable for SNP analysis, and given this, we employed SVDQUARTETS [(Chifman and Kubatko 2014), as 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, only one SNP from each ddRADseq locus was included in the SVDQUARTETS alignment. This was important because independent gene tree histories are assumed for each site. To assess sampling variance, we ran the analysis with 100 bootstrap replicates, and considered nodes resampled at >70% as being strongly supported. Taxon partitions were grouped at the lowest level of field identification (i.e., subspecific designations, when available), and by 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).

POMO.—We 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 replicates (Hoang et al. 2017) and a maximum virtual population size of 19. The 'safe' option was toggled to prevent numerical underflow in large datasets, and the discrete gamma rate

model was applied (N=4). Clades with bootstrap support ≥95% were considered strongly supported.

Concordance factors.—We also generated a lineage-tree supermatrix phylogeny (IQ-TREE v1.7.12; Nguyen et al. 2015) to contrast with our species-tree approaches. An edge-linked partition model with 1,000 ultrafast bootstrap 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 to include 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, in that it is 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 lineage tree, as recommended by IQ-TREE for stable sCF values. Ultrafast bootstraps ≥95% and sCF ≥50% were considered as strong support (per IQ-TREE documentation).

Constraint tree comparisons.—For statistical tests of topology, we employed IQ-TREE to generate lineage trees 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 violations, and more thoroughly explore gamma and invariant site parameters. Nodal confidence of individual trees was assessed using 1,000 ultrafast bootstraps. We then compared support for the concatenated supermatrix among the 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 and Hasegawa 1989); (d) Shimodaira-Hasegawa test (SH; Shimodaira and Hasegawa 1999); (e) Approximately Unbiased test (Shimodaira 2002); and (f) Expected Likelihood Weights (Strimmer and Rambaut 2002). Finally, 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.

Species Delimitation

BFD*.—We employed the multispecies coalescent Bayes Factor Delimitation approach [BFD*; (Leaché et al. 2014a)] as a baseline against which machine learning-based methods could be compared. 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). 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 Online 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 and Raftery 1995).

RF and t-SNE.—The RF and t-SNE algorithms (Breiman 2001; Maaten and Hinton 2008) were run on scaled principle components generated using ADEGENET v2.1.1 (Jombart and 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 output from RF was visualized using both classic and isotonic multidimensional scaling procedures (CMDS and ISOMDS; Shepard et al. 1972; Kruskal and 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.

Determining optimal K *for RF and t-SNE.*—We utilized two common clustering algorithms 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 given this, we first tested *K*=1-10 then employed both the gap statistic and mean silhouette widths [(MSW) (Rousseeuw 1987; Tibshirani et al. 2001)] to determine an 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. We implemented HAC via the MCLUST R-package (Scrucca et al. 2016).

VAE.—We also utilized a Bayesian UML approach (VAE) that employs neural networks and deep learning to infer the marginal likelihood distribution of sample means (μ) and standard deviations [(σ) (i.e. 'latent variables') (Derkarabetian et al. 2019)]. 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' (i.e. 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'). An escalating model loss in the validation dataset indicates overfitting, whereas a failure to acquire a minimum value points to underfitting (i.e. an inability to generalize across both training and unseen data).

SVM.—We derived six summary statistics for SVM using the CLADES package (Pei et al. 2018): 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 (i.e. longest string shared by two sequences). More extensive methodological descriptions of the UML and SML components of ML are found in Online Appendix 1.

RESULTS

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*. We included a minimum of 10 individuals per taxon, with occasionally fewer 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 were mapped to the *T. m. mexicana* genome, with 90,777 being parsimoniously informative. The mean per-individual coverage depth was 56.3X (Fig. S1).

Species Tree Inferences

The sCF tree contained N=214 tips (Fig. 2), whereas those from 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 subsequently 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 methodological differences 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 partitioned *T. m. triunguis*. Furthermore, SVDquartets, PoMo, and IQ-TREE each differed with respect to the placement of *T. c. bauri*, *T. coahuila*, and two previously recognized clades within *T. c. major* (Florida and Mississippi; Martin et al. 2013, 2019). 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*.

Species tree reconciliation.—Trees representing Sanger data and SVDquartets were in agreement when we contrasted our topology tests with the input supermatrix, 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 relatively small number of loci (Fig. S2), whereas the remaining majority was relatively uninformative.

Species Delimitation Methods Compared

Multi-species coalescent.—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 effective samples sizes (ESS) were >300 (Table S2).

Unsupervised machine learning.—Similarly, the majority of the RF and t-SNE runs (Fig. 4) 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 those using RF proximity scores.

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. Additionally, 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, 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 total 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.

Supervised machine learning.—CLADES yielded optimal *K*=2 (P=1.44e^{-4}; Fig. 4; Table S3), but with highly discordant clusters when 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.

Relative performance among approaches.—All UML species delimitation methods converged on *K*=3 when RF and t-SNE groups were parsed via intra-group variation. Three *Terrapene* species (plus *T. nelsoni*) were corroborated (per Martin et al. 2013, 2019), whereas the grouping of *T. mexicana* and *T. carolina* (Minx 1996) was rejected. ML 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.

DISCUSSION

We observed substantial heterogeneity among traditional and ML approaches in resolving the southeastern *Terrapene* taxa, echoing previous evaluations on the basis of morphological and single-gene data (Milstead 1967, 1969; Milstead and Tinkle 1967; Butler et al. 2011; Martin et al. 2013). However, a collective consensus was provided by the ML approaches with respect to model-based methods employing the multispecies coalescent. And—as argued below—these were interpreted as a more appropriate reflection of taxon-specific biological patterns. The results of our analyses represent for these methods a compendium of best practices that allow their *de novo* application to other taxonomically-complex systems.

Species Delimitation Approaches Reconciled in Terrapene

Species trees are a necessary requisite for species delimitation, in that they provide a phylogenetic context within which to build hypothetical species compositions and identities. In our case, species trees (Figs. 2-3) underscored classic discordance, previously hypothesized via single-gene sequencing (Martin et al. 2013). Differences were apparent among approaches 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*. Additionally, two individuals were placed in unexpected clades, though this 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. 2019). In turn, we suspect these 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 and Oaks 2017). PoMo, for example, constrains all nodes to the same N_e, a relatively poor assumption given the demographic histories of our study species (Martin et al. 2013, 2019). An examination of the species trees alone essentially reiterates previous taxonomic assessments provided by single-gene methods. However, powerful species delimitation assessments were derived from Next Generation Sequencing (NGS) methods that, taken together, provide a far more robust phylogenetic classification of *Terrapene*.

Traditional MSC species delimitation approaches seemingly promoted over-splitting, particularly when contrasted with prior results (Martin et al. 2013, 2019), and other approaches outlined herein. This opinion is well-characterized by results derived from empirically-structured and *in silico* systems (Sukumaran and Knowles 2017; Chambers and Hillis 2019). Likewise, several assumptions inherent to MSC-based delimitation are not congruent with the contemporarily occurring introgression seen in *Terrapene* (Martin et al. 2019). Thus, we posit the best-fitting BFD* model (i.e. virtually all taxa, save *T. o. ornata/luteola* as separate species) are systematic artefacts that reflect violations of the model rather than biological reality. Our interpretation is seemingly supported as well by the morphological overlap found among subspecific groups (as previously diagnosed; Milstead and Tinkle 1967; Milstead 1969; Minx 1992, 1996; Dodd 2001).

UML-based inference, on the other hand, collectively corroborated 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 describe as species-level variants, thereby encompassing hierarchical diversity. *Terrapene m. yucatana* is placed within *T. mexicana*, and *T. nelsoni* as sister to *T. ornata*, although extreme rarity and concomitantly limited sampling (N=1) excluded them from species delimitation. The variability in RF and t-SNE results primarily echoed uncertainty found in the species tree analyses, including the distinctiveness of *T. o. luteola* and subspecific relationships within *T. carolina*. This variation also corresponded with the characteristic properties of each clustering algorithm. For example, RF invokes a randomized process, perhaps exacerbated by the phylogenetic uncertainty 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 genetic similarities among Mississippi and Florida *T. c. major*. Finally, delimitations for both were strongly impacted by clustering algorithm, which closely paralleled their algorithmic tendencies. For example, the gap statistic often over-estimates *K* (Dudoit and Fridlyand 2002; Yan and Ye 2007), MSW under-splits (Şenbabaoğlu et al. 2014), and outliers and noise particularly impact HAC (Kim et al. 2009; Şenbabaoğlu et al. 2014). This, in turn, may explain the varying extent of agreement between HAC and either MSW or the gap statistic.

Much like RF and t-SNE, VAE also delimited *K*=3 (i.e. *Terrapene mexicana*, *T. carolina*, and *T. ornata* ssp.). It was much less amenable to replicating numerous runs, so we were unable to robustly test how prone it was to stochasticity. However, we ran VAE three times with the same delimitations resulting, a situation that did not occur for RF and t-SNE.

We thus suggest the variability we observed among RF and t-SNE runs was due to a lack of divergence within the more variable groups. In this sense, mixed classification was not observed among the *T. mexicana*, *T. ornata*, and *T. carolina* groups, but was instead restricted to their respective subspecific units. Thus, the lack of cohesive clusters across UML and clustering algorithms does not indicate separate species-level designations. We specifically refer to the *T. carolina*/*bauri*/*major*/*coahuila* cluster that was 'unidentifiable' (i.e. unassignable), at least with respect to the classification algorithms herein.

Machine Learning and the Limits of Model-based Inference

A seemingly recurring theme is that the complexities of the evolutionary processes are often too byzantine for our current mathematical models to parse. This was seemingly the case with *Terrapene*, where ML methods delimited more phylogenetically relevant groups than did their model-based counterparts. Given computational constraints, the performance discrepancy between the two is at least partly defined by the complexity of the data. In this sense, it is difficult for a model-based statistical process to exhaustively search a complex parameter space. Instead, a heuristic approach must be employed. Yet, as model complexity increases, so also does the demand for computational resources, especially given the step-wise or linear manner by which common algorithms (e.g. MCMC) explore this space (Adams et al. 2019). Consequently, multi-dimensional data with many local optima are problematic, especially when many components of the data matrix have little information content (Money and Whelan 2012).

On the other hand, the ML approach allows predictions to be extracted without substantial information loss when these data do not conform to the assumptions of an explicit

model (Bzdok et al. 2018). Predictive accuracy is optimized via relationships derived from training data (Adams et al. 2019), rather than from a parameter-rich model with assumptions inadequate for the parameters involved. ML might also be more prone to efficiently explore parameter space (e.g. of hyperparameters) due to the reduced computational burden they impose. Consequently, high-dimensional data can be processed without the need for explicit assumptions regarding the underlying and often convoluted intricacies of model-based inference.

Statistical models, in contrast, often become increasingly intractable as the number of variables increase, due to the elevated probability of sampling unmodelled processes. Given this, NGS studies create a scenario in which modelled relationships display, on one hand, little stochastic variation, yet on the other, widespread systematic bias (largely due to violations of the model). As a result, incorrect interpretations ultimately emerge, despite the presence of elevated statistical support (Philippe et al. 2011; Kumar et al. 2012). Although ML does have an interpretive cost, it also reflects numerous advantages as well, with a capacity to extend the field of evolutionary biology beyond more traditional perspectives as engendered by model-based approaches (Schrider and Kern 2018; Flagel et al. 2019).

The Application of Machine Learning to Species Delimitation: Best Practices

ML has proven useful in other evolutionary arenas, namely: 1) Delineating loci under selection (Schrider and Kern 2016; Kern and Schrider 2018; Koropoulis et al. 2020); 2) Identifying/ interpreting occurrence of selection within population demography (Sheehan and Song 2016; Torada et al. 2019); and 3) Documenting differential introgression across a genome (Schrider et al. 2018). Clearly, ML can elevate the resolution by which biological problems are

defined, a capacity that will become more extensive moving forward. ML is also computationally-benign when compared with MSC implementations, an increasingly attractive option that will resonate as whole genome sequencing supplants SNP discovery methods. Thus, given these caveats, UML approaches will be useful to a broad range of researchers. However, their accessibility is currently limited due to an opacity with regard to parameter choice, as well as the steep 'learning curve' regarding analytical implementation and statistical interpretation.

We found multiple instances in our compendium of analyses where hyperparameter settings were substantially impacted by sensitivity settings. Thus, we emphasize that users employ the same level of caution as with a more typical model-based statistical analysis. As shown herein, altering the perplexity settings in t-SNE strongly influenced the dispersion of clusters, which resulted in the most closely related taxa being variably differentiated. The number of trees or features sampled at each 'split' in an RF decision tree can also have large consequences on the resulting downstream classification (Breiman 2001; Probst and Boulesteix 2017), particularly if too few trees are used. Likewise, we encountered a variety of sensitive elements in executing VAE: 1) The number of neural network dimensions and number of encoder/decoder 'neurons' employed; 2) The neural activation function; 3) The learning rate; as well as 4) Batch size (Hu and Greene 2019). For *Terrapene*, an increase or reduction in neural network dimensions and neurons did not noticeably alter the resulting delimitations, but the impact of these settings likely depends on data dimensionality.

Finally, supervised SVM algorithms are sensitive to model-training parameters, such as kernel type, cost, and gamma, each of which affects the mathematical functions used to draw the vector (Eitrich and Lang 2006). In our case, the CLADES SVM model may have benefitted from a grid search to optimize these parameters, though we suspect the fit would be improved by a model

that also incorporated migration, particularly given the demographic history of *Terrapene* and its proclivity for admixture in contact zones.

Some considerations overlap with model-based approaches, such as the number of epochs (i.e. analogous to 'proposals' in a likelihood search), and the concepts of model fit, algorithmic limitations, and data quality/ biases. Overly large numbers of 'iterations' (epochs) can also cause overfitting, in a manner similar to how a prior distribution may overly constrain MCMC searches (Lambert et al. 2005). This may result in the model being strongly influenced by the training data, a situation that does not juxtapose well with data as yet unseen, and a predictive error with high variance (Cawley and Talbot 2010), as seen in our study (Fig. 5a,b; Fig. S3). Such an interpretation also explains our over-splitting in *Terrapene*, the result of a large maximum epoch size with VAE. Interestingly, this was not a built-in software function, but became apparent upon our examination of training *versus* test model loss. The results would have been easily interpreted as an erroneous K=8, stemming from the VAE overfit. However, excess iterations waste only computation time when using model-based approaches, with elevated ESS that characterize the posterior often being increased so as to obtain sampling lengths. In comparison, underfitting in ML models often occurs when features are uninformative, or too few features are actually employed, thereby biasing predictions.

CONCLUSIONS

UML approaches attempt to identify groups based on inherent structure in the data, and given this, are a natural extension to the species delimitation problem. However, they do come with interpretability as a cost, insofar as their internal mechanics may operate unseen, and their

sensitivity in the context of species delimitation has not been extensively explored. Given this, side-by-side tests against traditional methods are necessary to understand their limitations and considerations (as herein). In our case, a consensus among UML approaches corroborated other axes of differentiation (i.e. phylogenetic, morphological, and reproductive), whereas MSC-based delimitations over-partitioned the data. Thus, we posit that UML-based algorithms are a more computationally tractable alternative to model-based statistical approaches. While ML has only recently been adopted by evolutionary biologists, it provides a promising alternative from which to resolve long-standing biological problems for which traditional methods have seemingly plateaued.

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 Box Turtle Conservation Committee (BTCC), 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.

LITERATURE CITED

Adams R.H., Schield D.R., Castoe T.A. 2019. Recent Advances in the Inference of Gene Flow from Population Genomic Data. Curr. Mol. Biol. Reports. 5:107–115.

Agapow P.M., Bininda-Emonds O.R.P., Crandall K.A., Gittleman J.L., Mace G.M., Marshall J.C., Purvis A. 2004. The impact of species concept on biodiversity studies. Q. Rev. Biol. 79:161–179.

Allendorf F.W., Hohenlohe P.A., Luikart G. 2010. Genomics and the future of conservation genetics. Nat. Rev. Genet. 11:697–709.

Anderson E.C., Thompson E.A. 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. .

Auffenberg W. 1958. Fossil turtles of the genus *Terrapene* in Florida. Bull. Florida State Museum. 3:53–92.

Auffenberg W. 1959. A Pleistocene *Terrapene* hibernaculum, with remarks on a second complete box turtle skull from Florida. Q. J. Florida Acad. Sci. 22:49–53.

Breiman L. 2001. Random Forests. Mach. Learn. 45:5–32.

Butler J.M., Dodd Jr. C.K., Aresco M., Austin J.D. 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. Biol. J. Linn. Soc. 102:889–901.

Bzdok D., Altman N., Krzywinski M. 2018. Points of Significance: Statistics versus machine learning. Nat. Methods. 15:233–234.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Cawley G.C., Talbot N.L.C. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11:2079–2107.

Chafin T.K., Douglas M.R., Martin B.T., Douglas M.E. 2019. Hybridization drives genetic erosion in sympatric desert fishes of western North America. Heredity (Edinb). 123:759–773.

Chafin T.K., Martin B.T., Mussmann S.M., Douglas M.R., Douglas M.E. 2018. FRAGMATIC: in silico locus prediction and its utility in optimizing ddRADseq projects. Conserv. Genet. Resour. 10:325–328.

Chambers E.A., Hillis D.M. 2019. The multispecies coalescent over-splits species in the case of geographically widespread taxa. Syst. Biol. 69:184–193.

Chifman J., 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., Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics. 16:S2.

Coates D.J., Byrne M., Moritz C. 2018. Genetic Diversity and Conservation Units: Dealing With the Species-Population Continuum in the Age of Genomics. Front. Ecol. Evol. 6:165.

Derkarabetian S., Castillo S., Koo P.K., Ovchinnikov S., Hedin M. 2019. A demonstration of unsupervised machine learning in species delimitation. Mol. Phylogenet. Evol. 139:106562.

Dodd K.C. 2001. North American Box Turtles, A Natural History. Norman, OK, USA: University of Oklahoma Press.

Dudoit S., Fridlyand J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. 3:research0036.1.

Eaton D.A.R., Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics. 36:2592–2594.

Eaton D.A.R., Spriggs E.L., Park B., Donoghue M.J. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66:399–412.

Eitrich T., Lang B. 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. J. Comput. Appl. Math. 196:425–436.

Feldman C.R., Parham J.F. 2002. Molecular phylogenetics of emydine turtles: Taxonomic revision and the evolution of shell kinesis. Mol. Phylogenet. Evol. 22:388–398.

Flagel L., Brandvain Y., Schrider D.R. 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol. Biol. Evol. 36:220–238.

Fraley C., Raftery A.E. 1998. MCLUST: Software for model-based cluster and discriminant analysis. Dep. Stat. Univ. Washingt. Tech. Rep.

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

Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.

Garnett S.T., Christidis L. 2017. Taxonomy anarchy hampers conservation. Nature. 546:25–27.

Gippoliti S., Cotterill F.P.D., Zinner D., Groves C.P. 2018. Impacts of taxonomic inertia for the conservation of African ungulate diversity: an overview. Biol. Rev. 93:115–130.

Herrmann H., Rosen P.C. 2009. Conservation of aridlands turtles III: preliminary genetic studies of the desert box turtle and yaqui slider. Son. Herpetol. 22:38–43.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2017. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Hu Q., Greene C.S. 2019. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. Pacific Symp. Biocomput. 2019:362–373.

Isaac N.J.B., Mallet J., Mace G.M. 2004. Taxonomic inflation: Its influence on macroecology and conservation. Trends Ecol. Evol. 19:464–469.

Iverson J.B., Meylan P.A., Seidel M.E. 2017. Scientific and Standard English Names of Amphibians and Reptiles of North America North of Mexico, with Comments Regarding Confidence in Our Understanding. SSAR Herpetological Circular 43. In: Crother B.I.,

editor. p. 82–91.

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

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587.

Kass R.E., Raftery A.E. 1995. Bayes Factors. J. Am. Stat. Assoc. 90:773–795.

Kaufman L., Rousseeuw P. 1987. Clustering by means of medoids. Stat. Data Anal. Based L1-Norm Relat. Methods.:405–416.

Kern A.D., Schrider D.R. 2018. DiploS/HIC: An updated approach to classifying selective sweeps. G3 Genes, Genomes, Genet. 8:1959–1970.

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

Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29:170–179.

Kishino H., Miyata T., Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

Koropoulis A., Alachiotis N., Pavlidis P. 2020. Detecting positive selection in populations using genetic data. In: Julien Y. Dutheil, editor. Statistical Population Genomics. Humana, New York, NY. p. 87–123.

Kruskal J.B., Wish M. 1978. Multidimensional scaling. Sage.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Lambert P.C., Sutton A.J., Burton P.R., Abrams K.R., Jones D.R. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat. Med. 24:2401–2428.

Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014a. Species delimitation using genome-wide SNP data. Syst. Biol. 63:534–542.

Leaché A.D., Harris R.B., Rannala B., Yang Z. 2014b. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.

Leaché A.D., Oaks J.R. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 48:69–84.

Livingstone D.J., Manallack D.T., Tetko I. V. 1997. Data modelling with neural networks: Advantages and limitations. J. Comput. Aided. Mol. Des. 11:135–142.

Maaten L. van der, Hinton G. 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9:2579–2605.

Mace G.M. 2004. The role of taxonomy in species conservation. Philos. Trans. R. Soc. B Biol. Sci. 359:711–719.

Martin B.T., Bernstein N.P., Birkhead R.D., Koukl J.F., Mussmann S.M., Placyk J.S. 2013. Sequence-based molecular phylogenetics and phylogeography of the American box turtles (*Terrapene* spp.) with support from DNA barcoding. Mol. Phylogenet. Evol. 68:119–134.

Martin B.T., Bernstein N.P., Birkhead R.D., Koukl J.F., Mussmann S.M., Placyk Jr J.S. 2014. On the reclassification of the *Terrapene* (Testudines: Emydidae): a response to Fritz & Havaš. Zootaxa. 3835:292–294.

Martin B.T., Douglas M.R., Chafin T.K., Placyk J.S., Birkhead R.D., Phillips C.A., Douglas M.E. 2019. Differential introgression reveals thermal adaptation and candidate genes shaping species boundaries in North American box turtles (*Terrapene* spp.). bioRxiv.:752196.

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

Milstead W.W. 1969. Studies on the evolution of the box turtles (genus *Terrapene*). Bull. Florida State Museum, Biol. Sci. Ser. 14:1–113.

Milstead W.W., Tinkle D.W. 1967. *Terrapene* of Western Mexico, with comments on species groups in the genus. Copeia. 1967:180–187.

Minh B.Q., Hahn M.W., 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*. J. Herpetol. 26:234–238.

Minx P. 1996. Phylogenetic relationships among the box turtles, Genus *Terrapene*.

Herpetologica. 52:584–597.

Money D., Whelan S. 2012. Characterizing the phylogenetic tree-search problem. Syst. Biol. 61:228–239.

Morrison W.R., Lohr J.L., Duchen P., Wilches R., Trujillo D., Mair M., Renner S.S. 2009. The impact of taxonomic change on conservation: Does it kill, can it save, or is it just irrelevant? Biol. Conserv. 142:3201–3206.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nielsen R., Paul J.S., Albrechtsen A., Song Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443.

Pei J., Chu C., Li X., Lu B., Wu Y. 2018. CLADES: A classification-based machine learning method for species delimitation from population genetic data. Mol. Ecol. Resour. 18:1144–1156.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 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 D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Probst P., Boulesteix A.-L. 2017. To tune or not to tune the number of trees in random forest. J.

Mach. Learn. Res. 18:6673–6690.

R Development Core Team. 2018. R: A language and environment for statistical computing. Available at: https://www.R-project.org/.

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

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.

Rodriguez-Galiano V.F., Ghimire B., Rogan J., Chica-Olmo M., Rigol-Sanchez J.P. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67:93–104.

Rousseeuw P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20:53–65.

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

Schrempf D., Minh B.Q., De Maio N., von Haeseler A., Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. J. Theor. Biol. 407:362–370.

Schrider D.R., Ayroles J., Matute D.R., Kern A.D. 2018. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. PLoS Genet. 14:e1007341.

Schrider D.R., Kern A.D. 2016. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. 12.

Schrider D.R., Kern A.D. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. 34:301–312.

Scrucca L., Fop M., Murphy T.B., Raftery A.E. 2016. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. 8:289–317.

Şenbabaoğlu Y., Michailidis G., Li J.Z. 2014. Critical limitations of consensus clustering in class discovery. Sci. Rep. 4:1–13.

Shaffer H.B., Minx P., Warren D.E., Shedlock A.M., Thomson R.C., Valenzuela N., Abramyan J., Amemiya C.T., Badenhorst D., Biggar K.K. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14:R28.

Sheehan S., Song Y.S. 2016. Deep Learning for Population Genetic Inference. PLoS Comput. Biol. 12:e1004845.

Shepard R.N., Romney A.K., Nerlove S.B. 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. Syst. Biol. 51:492–508.

Shimodaira H., Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Simmons M.P., Goloboff P.A. 2014. Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. Mol. Phylogenet. Evol. 78:334–348.

Stanton D.W.G., Frandsen P., Waples R.K., Heller R., Russo I.R.M., Orozco-terWengel P.A., Pedersen C.E.T., Siegismund H.R., Bruford M.W. 2019. More grist for the mill? Species delimitation in the genomic era and its implications for conservation. Conserv. Genet. 20:101–113.

Stephens P.R., Wiens J.J. 2003. Ecological diversification and phylogeny of emydid turtles. Biol. J. Linnaean Soc. 79:577–610.

Strimmer K., Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. R. Soc. London. Ser. B Biol. Sci. 269:137–142.

Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. U. S. A. 114:1607–1611.

Sullivan B.K., Douglas M.R., Walker J.M., Cordes J.E., Davis M.A., Anthonysamy W.J.B., Sullivan K.O., Douglas M.E. 2014. Conservation and management of polytypic species: The Little Striped Whiptail Complex (*Aspidoscelis inornata*) as a case study. Copeia. 2014:519–529.

Suryachandra P., Reddy P.V.S. 2016. Comparison of machine learning algorithms for breast cancer. Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016.

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

Tibshirani R., Walther G., Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. 63:411–423.

Torada L., Lorenzon L., Beddis A., Isildak U., Pattini L., Mathieson S., Fumagalli M. 2019.

ImaGene: A convolutional neural network to quantify natural selection from genomic data. BMC Bioinformatics. 20:1–12.

Vane-Wright R.I., Humphries C.J., Williams P.H. 1991. What to protect? Systemations and the agony of choice. Biol. Conserv. 55:235–254.

Yan M., Ye K. 2007. Determining the Number of Clusters Using the Weighted Gap Statistic. Biometrics. 63:1031–1037.

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. 107:9264–9269.

Zachos F.E. 2018. Species concepts, species delimitation and the inherent limitations of taxonomy. J. Genet. 97:811–815.

Zachos F.E., Apollonio M., Bärmann E. V., Festa-Bianchet M., Göhlich U., Habel J.C., Haring E., Kruckenhauser L., Lovari S., McDevitt A.D., Pertoldi C., Rössner G.E., Sánchez-Villagra M.R., Scandura M., Suchentrunk F. 2013. Species inflation and taxonomic artefacts-A critical comment on recent trends in mammalian classification. Mamm. Biol. 78:1–6.

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 genome-wide ddRADseq data (Fig. 3, this study). Bolded *P*-values with an asterisk (*) indicate supported trees (P>0.05 or highly weighted).

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; "+" indicates taxa that were grouped together; "/" delineates multiple groupings. DS=Desert (*T. o. luteola*), ON=Ornate (*T. o. ornata*), EA=Eastern (*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*).

Figure 1: Range map for North American box turtles, *Terrapene,* depicting sample localities. Cross-hatched areas indicate areas of sympatry. Circles represent samples used only for the species trees, whereas squares were also used in species delimitation analyses. Headings and subheadings correspond to species and subspecies, respectively, following (Martin et al. 2013),

and include Ornate (*T. ornata ornata*), Desert (*T. o. luteola*), Spotted (*T. nelsoni*), Three-toed *(T. mexicana triunguis*), Mexican (*T. m. mexicana*), Yucatan (*T. m. yucatana*), Eastern (*T. carolina carolina*), Florida (*T. c. bauri*), Gulf Coast from distinct Florida and Mississippi populations (*T. c. major*), and Coahuilan (*T. coahuila*) box turtles. The black localities indicate species-level field identification of *T. carolina*.

Figure 2: Maximum likelihood lineage tree reflecting relationships among 214 North American Box Turtle (*Terrapene* spp.) samples. The tree was generated from 11,962 unlinked ddRADseq loci (N_{sites} = 1,043,153) using IQ-TREE v1.7.12 with 1,000 ultrafast bootstrap (UFB) replicates. The substitution model was set to general time reversible (GTR), with per-locus parameters estimated via MODELFINDER (IQ-TREE). Site concordance-factors (sCF) were calculated from 100 quartets randomly sampled from internal branches. Branch support values represent UFB replicates, with sCF to the left and right of each vertical line, respectively. Bootstraps ≥95% and sCF ≥50% were considered strong support. UFBs <90 and sCF <40 were omitted for visual clarity, and the latter was rounded to the nearest integer. Asterisks (*) indicate 100% support. The 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), Eastern (*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. The * and + characters above branches represent nodes with 100% and ≥95% bootstrap support, respectively

Figure 4: Species delimitations among 37 North American box turtle (*Terrapene* spp.) 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 traditional multispecies coalescent (MSC) approach. UML algorithms included 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 the colors indicating percent group assignments per method. Mixed colors indicate clustering variation among runs. t-SNE was performed 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 (PAM) using mean silhouette widths (MSW) and the gap statistic (GS), and PAM using RF proximity scores (PAMp), whereas standard deviation (σ) overlap was used for VAE. CLA=CLADES (SML) and BFD=Bayes Factor Delimitation (MSC) optimal *K*'s were determined using probabilities (*P*) and Bayes Factors (BF).

Figure 5: Results from variational autoencoders (VAE) machine learning species delimitation among 37 North American box turtle (*Terrapene* spp.) 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 deviation (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 *Terrapene* subspecies across *T. mexicana*, *T. carolina*, and *T. ornata*, plus monotypic *T. coahuila*.

Table S1: Metadata for 214 North American box turtle (*Terrapene* spp.) samples. Fields with a "-" indicate metadata that is unknown or was not provided by the collector(s). Taxonomic IDs are as designated in the field. Geographic coordinates are in decimal degrees. Collection dates are generally in the format mm/year.

Table S2: Summary statistics from twenty BFD* models (Bayes Factor Delimitation, *with genomic data) among 37 North American box turtle (*Terrapene* spp.) samples and 179 unlinked ddRADseq single nucleotide polymorphism (SNP) variants. The standard deviations reflect error in the calculation of the marginal likelihood estimate (MLE) from ten path sampling cross-validation runs.

Table S3: Pairwise probabilities for CLADES supervised machine learning species delimitation among 37 North American box turtle (*Terrapene* spp.) individuals. Six machine learning features

were derived from 7,395 unlinked ddRADseq loci. The classification labels indicate probabilities for splitting (+1) or clumping (-1) the taxa in each pairwise comparison. Taxon codes include CH=Coahuilan (*Terrapene coahuila*), DS=Desert (*T. ornata luteola*), EA=Eastern (*T. carolina carolina*), FL=Florida (*T. c. bauri*), GUFL=Gulf Coast (from Florida; *T. c. major*), GUMS=Gulf Coast (from Mississippi; *T. c. major*), MX=Mexican (*T. mexicana mexicana*), ON=Ornate (*T. o. ornata*), TT=Three-toed (*T. m. triunguis*) box turtles.

Figure S1: Mean per-individual read depth (points) ± standard deviation (gray bars) for 214 North American box turtle (*Terrapene* spp.) individuals using ddRAD sequencing.

Figure S2: Constraint trees and respective changes in site-likelihood scores (**Δ**SLS) representing genome-wide support for the a) SVDQUARTETS, b) POMO, C) Sanger, and d) Morphological phylogenetic hypotheses. The SVDQUARTETS and POMO trees are derived in this study, whereas the Sanger and morphological hypotheses are results previously published (Minx 1996; Martin et al. 2013).

Figure S3: Prediction error *versus* model complexity for machine learning. Ideally, training should stop when the validation and training prediction error are at their lowest point (‘Goldilocks zone’, indicated by gray dashed lines). High bias and underfitting occur when training ends prior to reaching the Goldilocks zone, and high variance and overfitting occur when validation error begins to climb.

Tree	logL	deltaL	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***

bp-RELL = Bootstrap proportions using the RELL method (weights, sum to 1)

p-KH = Kishino-Hasegawa test (

p-SH = Shimodaira-Hasegawa test

c-ELW = Expected likelihood weight (weights, sum to 1)

p-AU = Approximately unbiased test

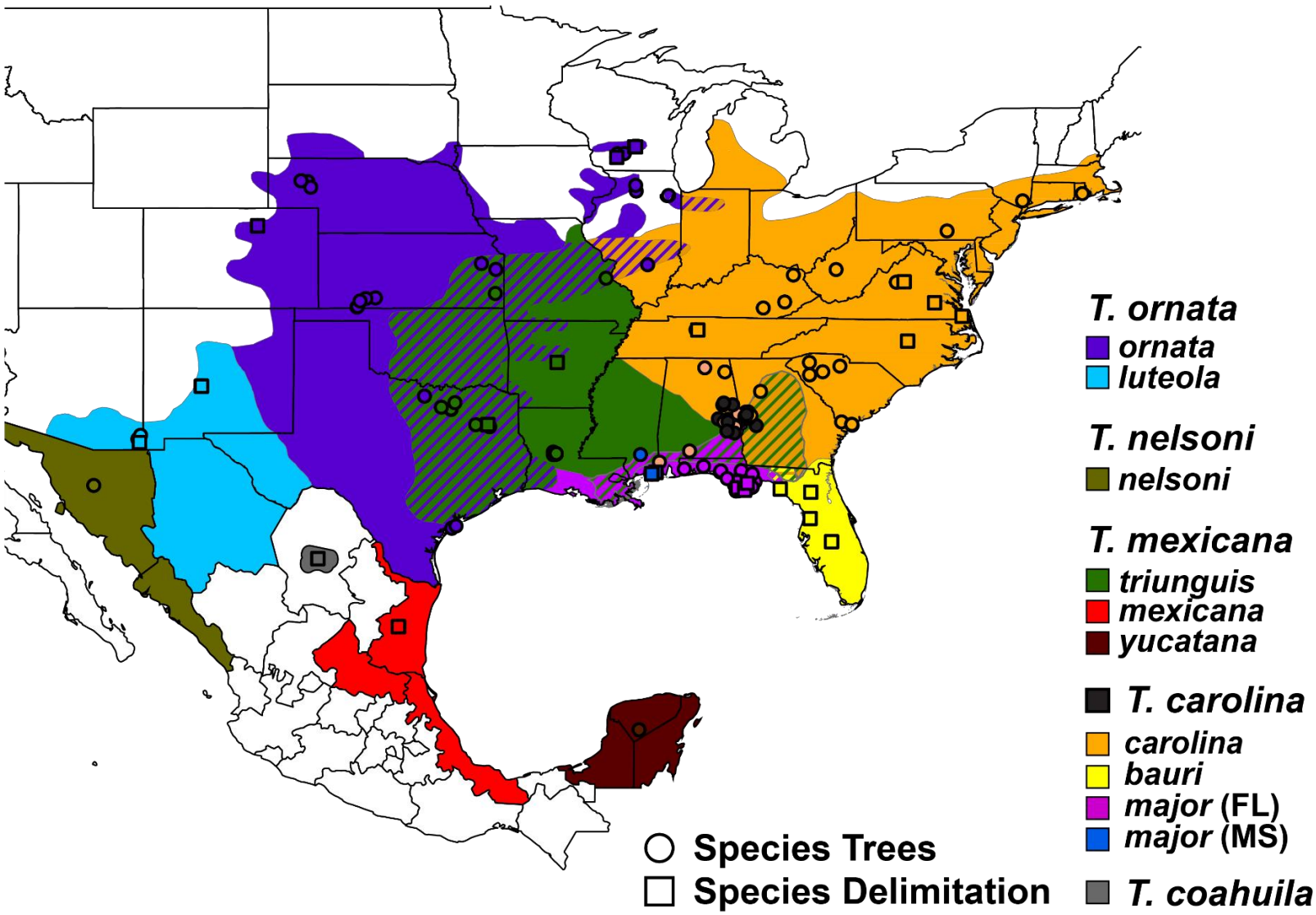
BFD* Model	**MLE**	***K***	**Rank**	**BF**
All Separate*	-2403.39	10	1	-
DS+ON*	-2404.34	9	2	1.90
EA+GUFL	-2417.84	9	3	28.91
GUMS+GUFL	-2427.58	9	4	48.39
GUMS+CH	-2448.61	9	5	90.44
GUMS+CH/GUFL+EA	-2461.28	8	6	115.79
GUMS+GUFL+CH	-2489.62	8	7	172.45
EA+FL	-2511.83	9	8	216.89
GUMS+GUFL+CH+EA	-2514.86	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

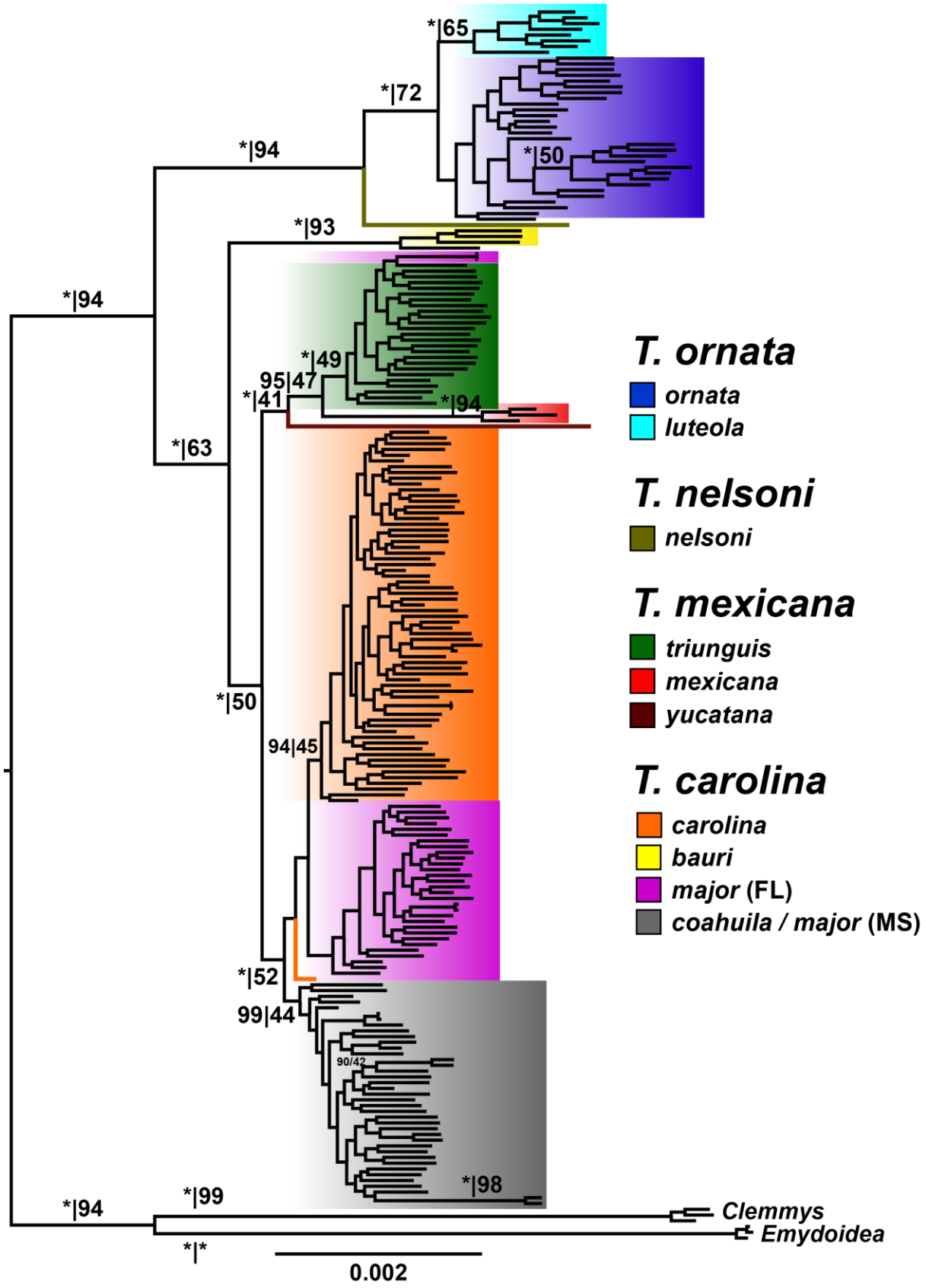
MLE = Marginal likelihood estimates

BF = Bayes factors

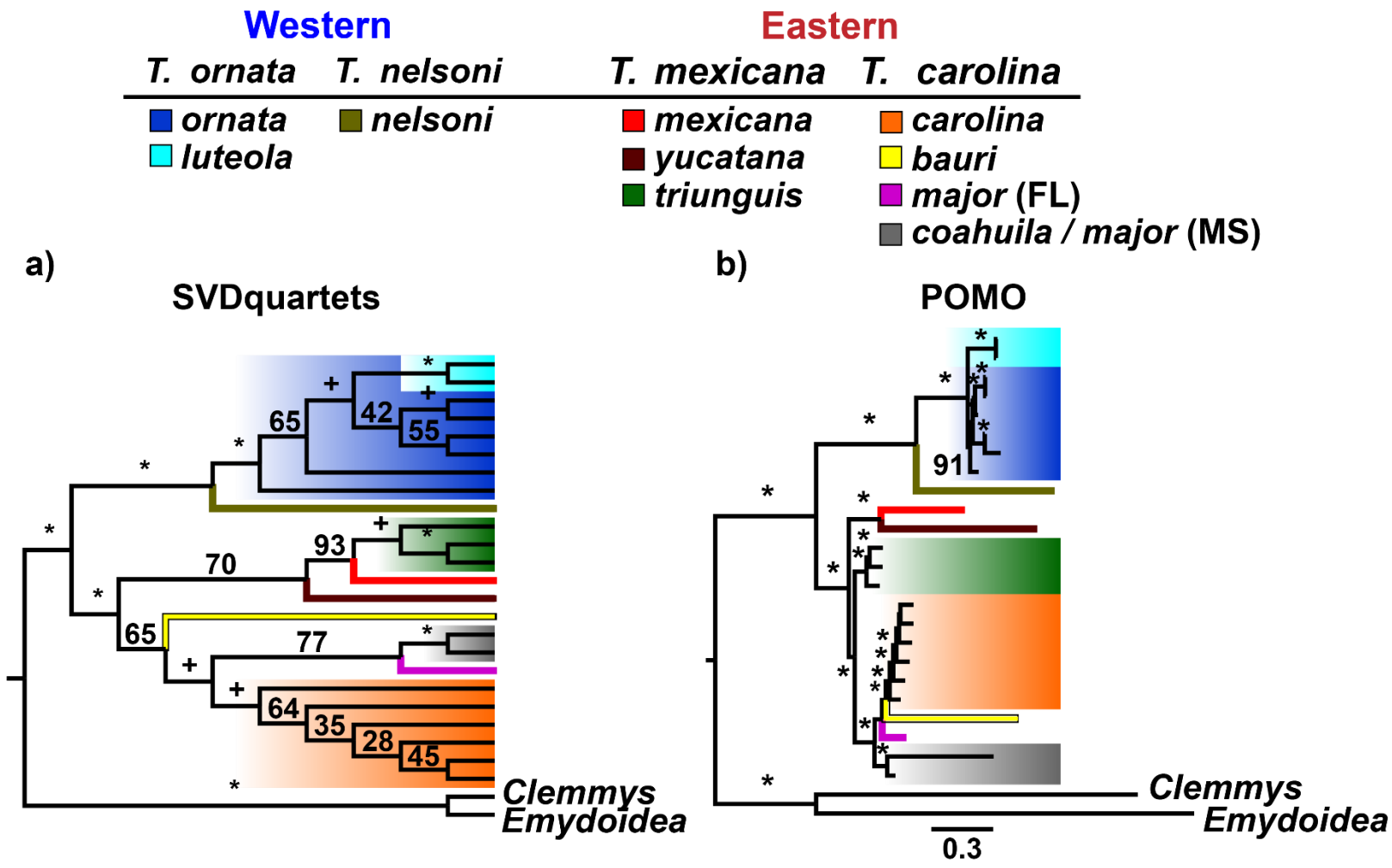
K = # tips

Rank = model ranking based on MLE (lower=better)

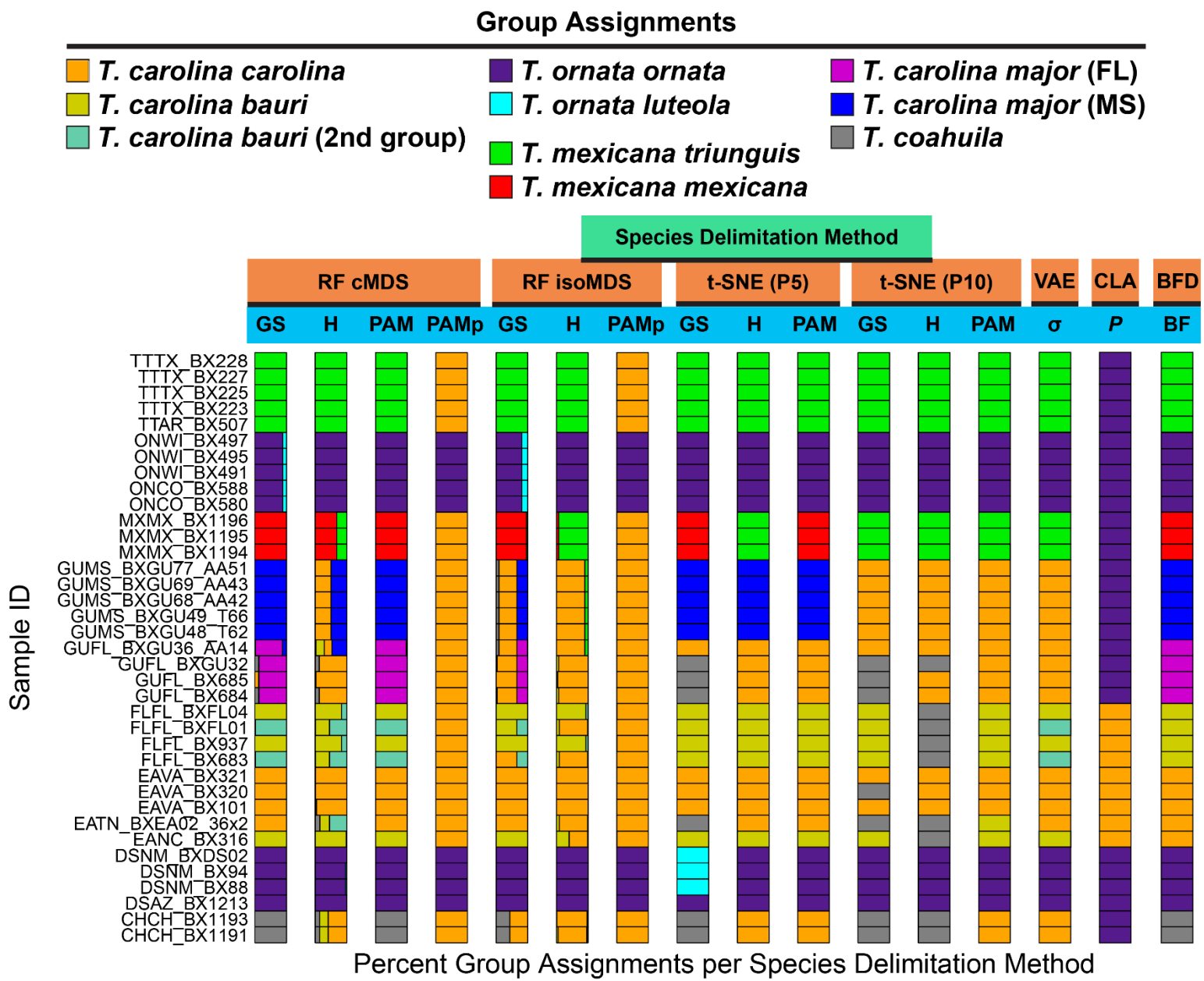

T. ornata
ornata
luteola
T. nelsoni
nelsoni
T. mexicana
triunguis
mexicana
yucatana
T. carolina
carolina
bauri
major (FL)
major (MS)
T. coahuila
Species Trees
Species Delimitation

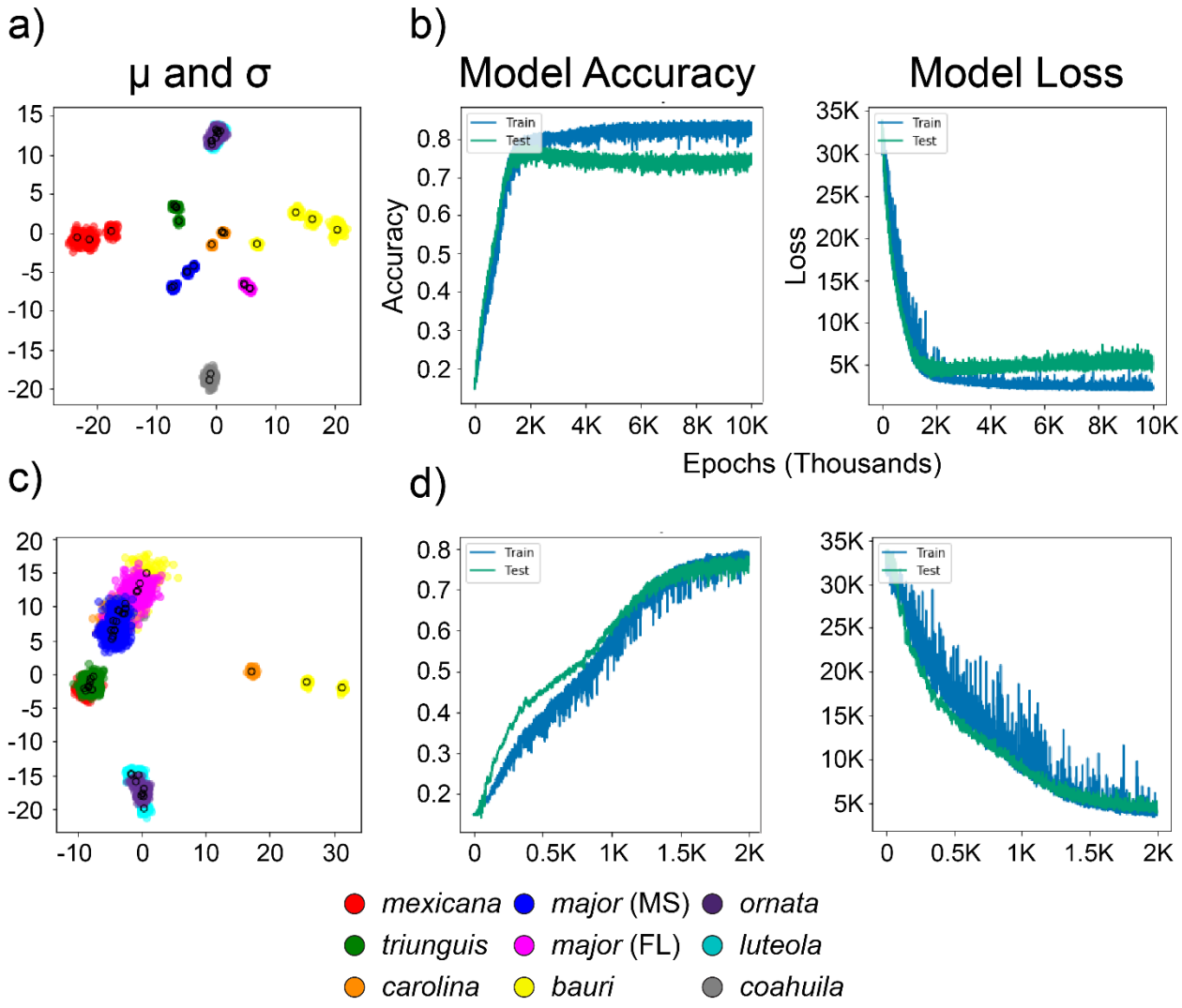

*|65
*|72
*|94
*|50
*|93
*|94
*|49
95|47
*|41
*|94
*|63
*|50
94|45
*|52
99|44
90/42
*|98
*|99
*|94
|
Clemmys
Emydoidea
0.002
T. ornata
ornata
luteola
T. nelsoni
nelsoni
T. mexicana
triunguis
mexicana
yucatana
T. carolina
carolina
bauri
major (FL)
coahuila / major (MS)

998


Western
Eastern
T. ornata
T. nelsoni
T. mexicana
T. carolina
ornata
luteola
nelsoni
mexicana
yucatana
triunguis
carolina
bauri
major (FL)
coahuila / major (MS)
a)
SVDquartets
b)
POMO
65
42
55
93
70
65
77
64
35
28
45
91
Clemmys
Emydoidea
0.3

.


Group Assignments
T. carolina carolina
T. carolina bauri
T. carolina bauri (2nd group)
T. ornata ornata
T. ornata luteola
T. mexicana triunguis
T. mexicana mexicana
T. carolina major (FL)
T. carolina major (MS)
T. coahuila
Species Delimitation Method
RF cMDS
RF isoMDS
t-SNE (P5)
t-SNE (P10)
VAE
CLA
BFD
GS H PAM PAMp GS H PAMp GS H PAM GS H PAM σ P BF
Sample ID
TTTX_BX228
TTTX_BX227
TTTX_BX225
TTTX_BX223
TTAR_BX507
ONWI_BX497
ONWI_BX495
ONWI_BX491
ONCO_BX588
ONCO_BX580
MXMX_BX1196
MXMX_BX1195
MXMX_BX1194
GUMS_BXGU77_AA51
GUMS_BXGU69_AA43
GUMS_BXGU68_AA42
GUMS_BXGU49_T66
GUMS_BXGU48_T62
GUFL_BXGU36_AA14
GUFL_BXGU32
GUFL_BX685
GUFL_BX684
FLFL_BXFL04
FLFL_BXFL01
FLFL_BX937
FLFL_BX683
EAVA_BX321
EAVA_BX320
EAVA_BX101
EATN_BXEA02_36x2
EANC_BX316
DSNM_BXDS02
DSNM_BX94
DSNM_BX88
DSAZ_BX1213
CHCH_BX1193
CHCH_BX1191
Percent Group Assignments per Species Delimitation Method


a)
μ and σ
b)
Model Accuracy
Model Loss
Train
Test
Accuracy
Loss
Epochs (Thousands)
c)
d)
mexicana
triunguis
carolina
major (MS)
major (FL)
bauri
ornata
luteola
coahuila

.

SUPPLEMENTARY MATERIALS

Table S2

BFD Model	ESS Mean	ESS Median (Min, Max)	MLE	Std. Dev.
run1, East/West	649.18	506.72755 (122.6519, 1334)	-2926.56	0.06
run2, EA+FL+CH+GUMS+GUFL+TT+MX	515.20	422.0864 (105.0803, 1289.1458)	-2926.20	0.08
run3, ON+DS/EA+TT+MX+CH+GUMS+GUFL/FL	593.63	459.24485 (71.6897, 1334)	-2720.23	0.07
run4, EA+FL+CH+GUMS+GUFL+TT	487.48	372.5475 (21.7602, 1264.0155)	-2800.56	0.09
run5, EA+CH+GUMS+GUFL+TT+MX	493.91	492.88135 (56.1857, 1115.7808)	-2719.02	0.08
run6, EA+FL+CH+GUMS+GUFL	484.56	433.14245 (74.3832, 1236.2148)	-2693.37	0.10
run7, EA+CH+GUMS+GUFL+MX	478.60	380.1236 (68.1691, 1153.8154)	-2657.48	0.09
run8, EA+CH+GUMS+GUFL+TT	455.44	379.51645 (28.2256, 1217.8904)	-2607.72	0.10
run9, GUMS+GUFL+CH+EA	420.68	369.88705 (28.1414, 1295.5895)	-2514.86	0.11
run10, EA+FL+GUFL/CH+GUMS	407.49	322.56155 (38.1882, 1181.6271)	-2594.91	0.11
run11, GUMS+GUFL+CH	367.29	305.75535 (62.5265, 1097.3091)	-2489.62	0.12
run12, GUMS+CH/GUFL+EA	340.61	285.44545 (42.0809, 968.308)	-2461.28	0.13
run13, EA+FL/CH+GUMS	365.06	307.52505 (73.946, 913.1602)	-2555.16	0.13
run14, EA+FL+GUFL	321.74	251.91375 (31.6129, 1127.1272)	-2552.22	0.11
run15, GUMS+GUFL	288.79	274.3423 (59.7872, 815.3268)	-2427.58	0.12
run16, GUMS+CH	330.86	261.6958 (68.1546, 880.6262)	-2448.61	0.13
run17, EA+GUFL	314.11	212.32105 (49.5533, 1158.6918)	-2417.84	0.12
run18, EA+FL	343.54	299.22845 (28.5148, 1291.1445)	-2511.83	0.14
run19, DS+ON	386.57	265.64725 (54.6711, 1181.5824)	-2404.34	0.12
run20, All Separate	305.72	223.09945 (81.8985, 1066.5475)	-2403.39	0.13

ESS = Effective sample size
MLE = Marginal likelihood estimate

Table S3

Taxa	+1	-1
CH-DS	0.3622	0.6378
CH-EA	0.4819	0.5181
CH-FL	0.5999	0.4001
CH-GUFL	0.3894	0.6106
CH-GUMS	0.3681	0.6319
CH-MX	0.2513	0.7487
CH-ON	0.3262	0.6738
CH-TT	0.2956	0.7044
DS-EA	0.4917	0.5083
DS-FL	0.5863	0.4137
DS-GUFL	0.4102	0.5898
DS-GUMS	0.4342	0.5658
DS-MX	0.3792	0.6208
DS-ON	0.3087	0.6913
DS-TT	0.3917	0.6083
EA-FL	0.3747	0.6253
EA-GUFL	0.4299	0.5701
EA-GUMS	0.5229	0.4771
EA-MX	0.4988	0.5012
EA-ON	0.5334	0.4666
EA-TT	0.5076	0.4924
FL-GUFL	0.5654	0.4346
FL-GUMS	0.6748	0.3252
FL-MX	0.6042	0.3958
FL-ON	0.6401	0.3599
FL-TT	0.6167	0.3833
GUFL-GUMS	0.3668	0.6332
GUFL-MX	0.4042	0.5958
GUFL-ON	0.4413	0.5587
GUFL-TT	0.3859	0.6141
GUMS-MX	0.3796	0.6204
GUMS-ON	0.3978	0.6022
GUMS-TT	0.3465	0.6535
MX-ON	0.3335	0.6665
MX-TT	0.2952	0.7048
ON-TT	0.3229	0.6771

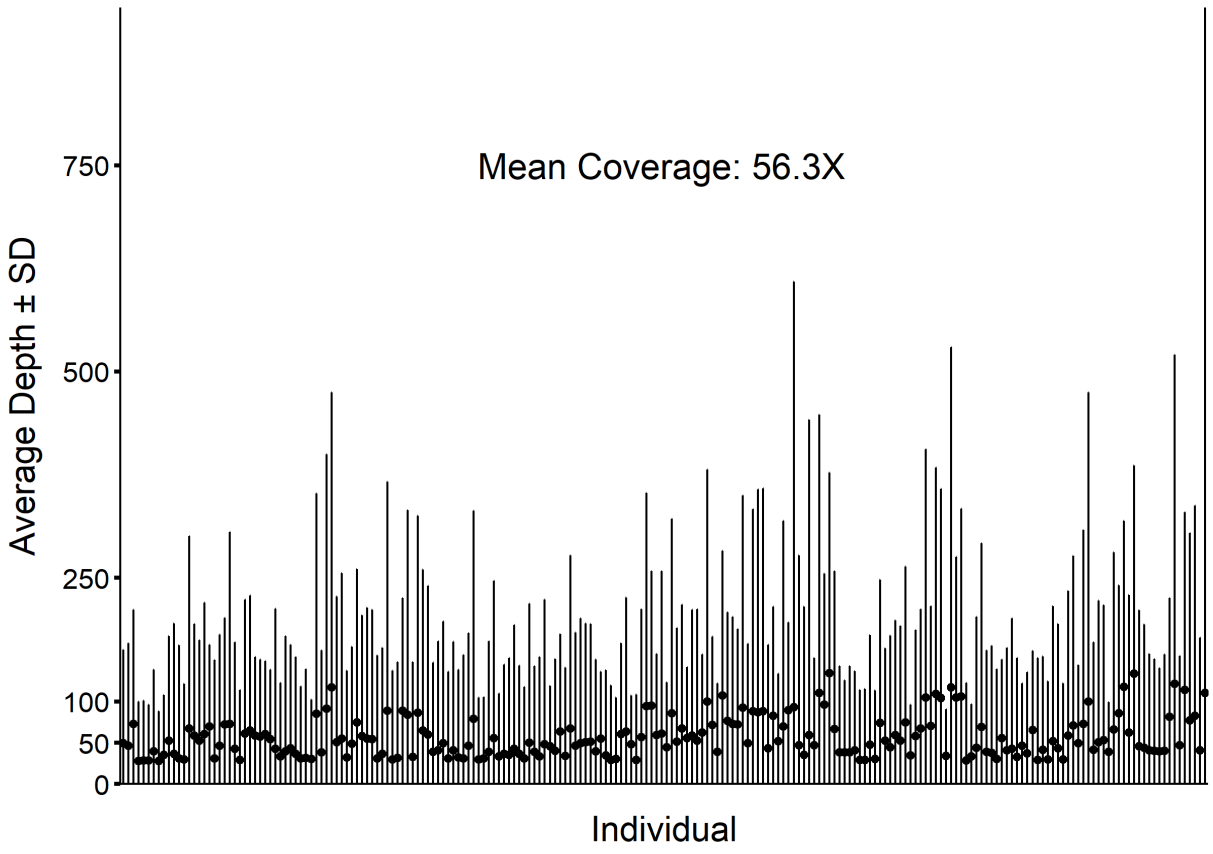

Mean Coverage: 56.3X
Average Depth ± SD
750
500
250
100
50
0
Individual

Figure S1

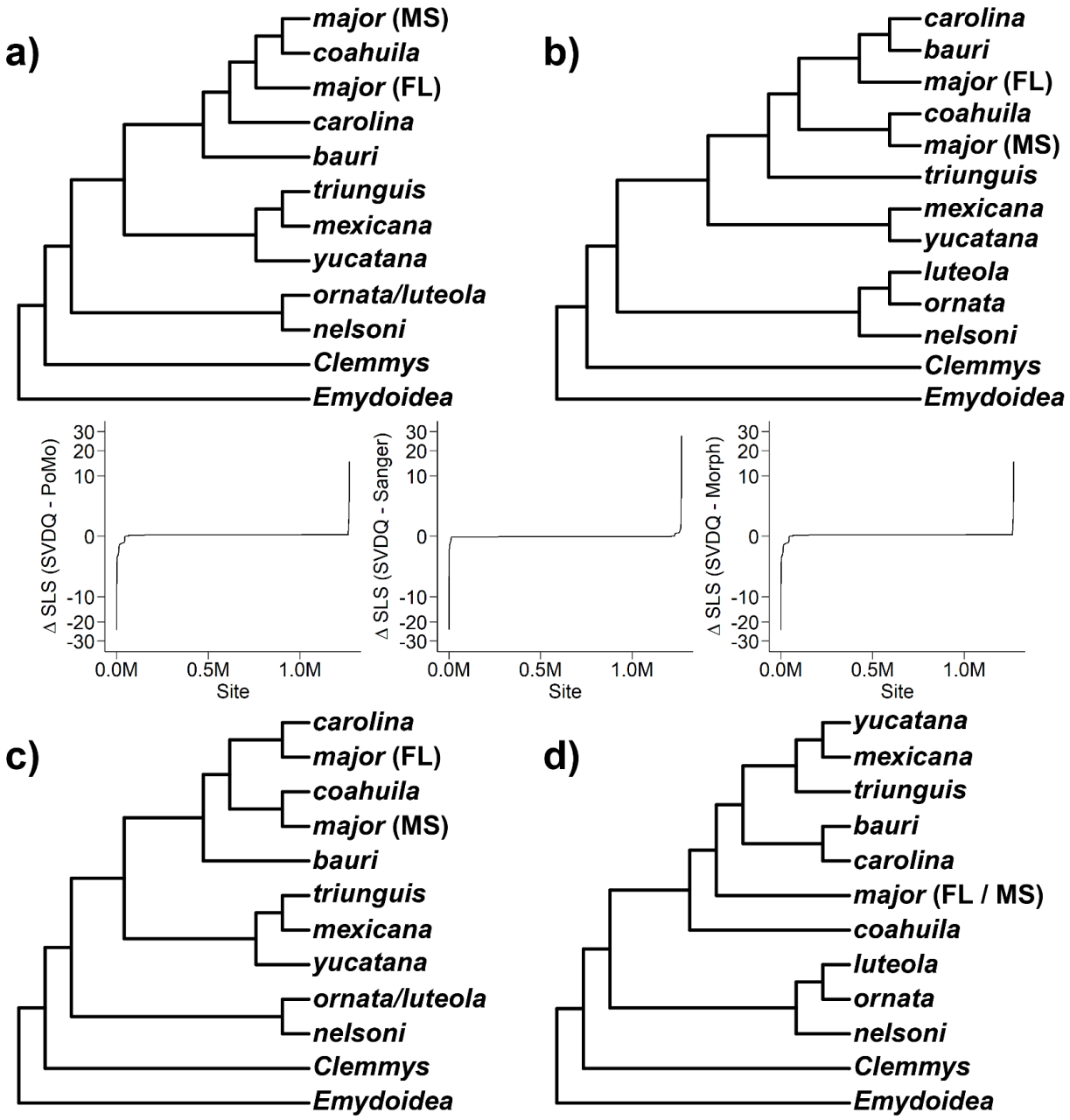

a)
major (MS)
coahuila
major (FL)
carolina
bauri
triunguis
mexicana
yucatana
ornata/luteola
nelsoni
Clemmys
Emydoidea
b)
carolina
bauri
major (FL)
coahuila
major (MS)
triunguis
mexicana
yucatana
luteola
ornata
nelsoni
Clemmys
Emydoidea
Δ SLS (SVDQ - PoMo)
Δ SLS (SVDQ - Sanger)
Δ SLS (SVDQ - Morph)
30
20
10
0
-10
-20
-30
0.0M
0.5M
1.0M
Site
c)
carolina
major (FL)
coahuila
major (MS)
bauri
triunguis
mexicana
yucatana
ornata/luteola
nelsoni
Clemmys
Emydoidea
d)
yucatana
mexicana
triunguis
bauri
carolina
major (FL / MS)
coahuila
luteola
ornata
nelsoni
Clemmys
Emydoidea

Figure S2

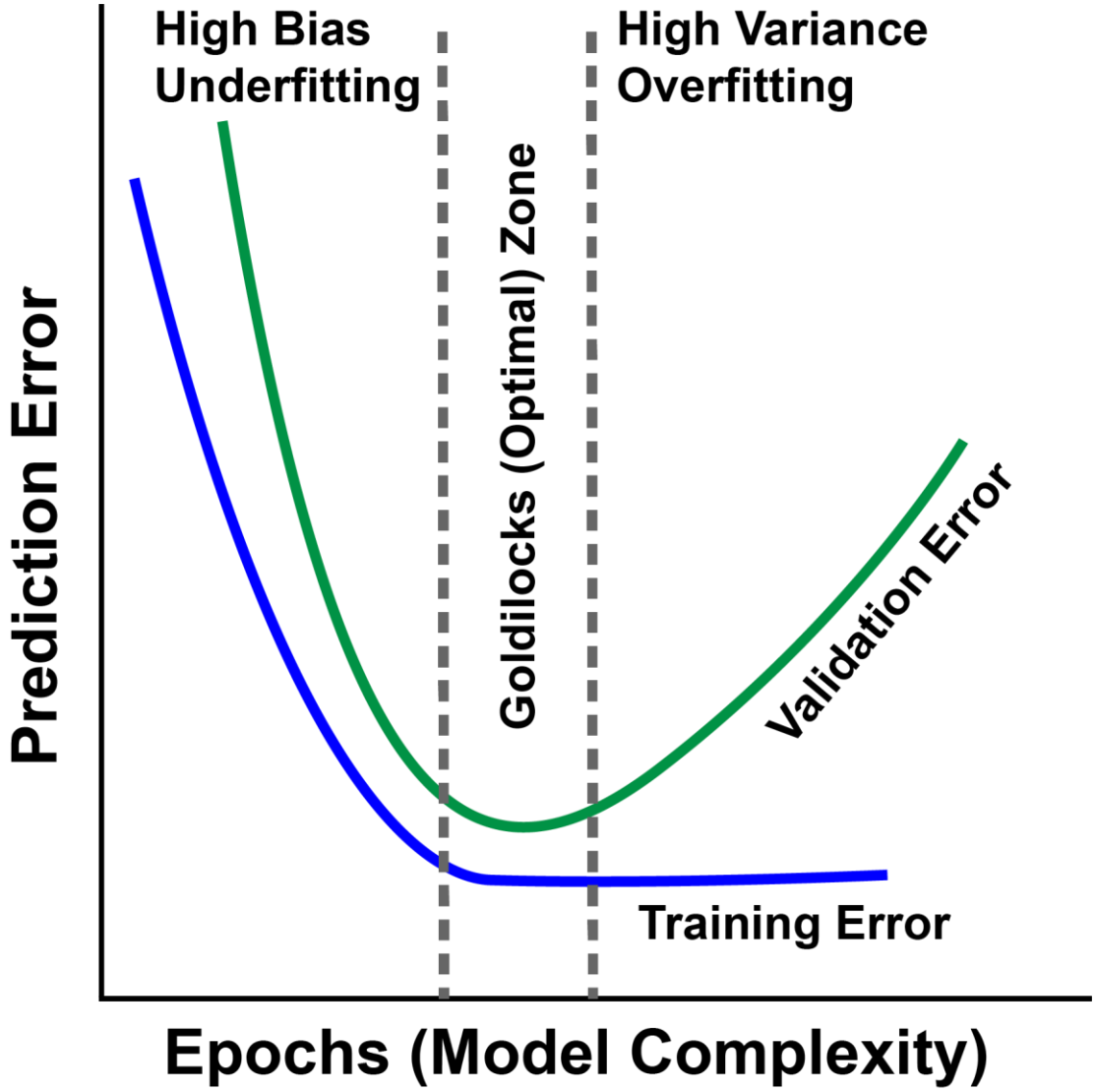

High Bias
Underfitting
High Variance
Overfitting
Goldilocks (Optimal) Zone
Validation Error
Training Error
Prediction Error
Epochs (Model Complexity)

Figure S3