5-2021

Dynamics of Hybrid Zones at a Continental Scale

Bradley T. Martin
University of Arkansas, Fayetteville

Dynamics of Hybrid Zones at a Continental Scale

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Biological Sciences

by

Bradley T. Martin
University of Texas at Tyler
Bachelor of Science in Biology, 2009
University of Texas at Tyler
Master of Science in Biology, 2012

May 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Marlis R. Douglas, Ph.D.
Dissertation Director

Michael E. Douglas, Ph.D.
Committee Member

Andrew J. Alverson, Ph.D.
Committee Member

John S. Placyk, Jr., Ph.D.
Committee Member

ABSTRACT

Hybridization has traditionally been viewed as a happenstance that negatively impacts populations, but is now recognized as an important evolutionary mechanism that can substantially impact the evolutionary trajectories of gene pools, influence adaptive capacity, and contravene or reinforce divergence. Physiographic processes are important drivers of dispersal, alternately funneling populations into isolation, promoting divergence, or facilitating secondary contact of diverged populations, increasing the potential for hybridization. In North America, glacial-interglacial cycles and geomorphological changes have provided a dynamic backdrop over the last two million years that promoted such oscillations of population contraction and expansion. These biogeographic processes have resulted in regional hybrid zones where hybridization spans generations

Herein, I explored hybrid zones in two species complexes of reptiles across Eastern, Central, and Southwestern North America. Hybrid zones can influence evolutionary trajectories, and understanding the mechanisms underlying their formation is important for defining appropriate management strategies and can help avoid actions that would inadvertently lead to new hybrid zones.

Chapter I assessed differential introgression in a complex of terrestrial turtles, the American Box Turtles (*Terrapene* spp.), from a contemporary hybrid zone in the southeastern United States. Transcriptomic loci were correlated with environmental predictors to evaluate mechanisms engendering maladapted hybrids and adaptive introgression. Selection against hybrids predominated for inter-specifics but directional introgression did so in conspecifics. Outlier loci also primarily correlated with temperature, reflecting the temperature dependency of ectotherms and underscoring their vulnerability to climate change.

Chapter II performed a robust assessment of recently developed machine learning (M-L) approaches to delimit four *Terrapene* species and evaluate the impact of data filtering and M-L parameter choices. Parameter selections were varied to determine their effects in resolving clusters. The results provide necessary recommendations on using M-L for species delimitation in species complexes defined by secondary contact. These data exemplify usage of M-L software in a phylogenetically complex group.

Chapter III describes an R package to visualize some of the analyses from Chapter I. Current software to generate genomic clines does not include functions to visualize the results. Thus, I wrote an API (application programming interface) that does so and also performs other genomic and geographic cline-related tasks.

Chapter IV examines historical and contemporary phylogeographic patterns in the Massasaugas (*Sistrurus* spp.), a type of dwarf rattlesnake found across the Southwest and Central Great Plains. In the Southwest, *S. tergeminus tergeminus* and *S. t. edwardsii* putatively diverged in the absence of strong physiographic barriers and physical glaciers, suggesting primary divergence. In contrast, a disjunct population of *S. t. tergeminus* in Missouri reflects potentially historical secondary contact with *S. catenatus*. These taxa represent contrasting examples of divergence resulting from alternative phylogeographic processes and contextualizes evolutionarily significant and management units.

Combined, the four chapters present population genomic data to elucidate impacts of phylogeographic processes on hybrid zones at a continental scale. The data will promote effective conservation management strategies, as many species in the focal regions have been affected by anthropogenic pressures. In this sense, the results can be extrapolated to co-distributed taxa with similar phylogeographic histories.

ACKNOWLEDGEMENTS

The results herein would not have been possible without the sample contributions of hundreds of volunteers, museums, agencies, and wildlife organizations. Their dedicated work enabled robust population genomic analyses that are well-represented geographically. Special thanks go to R. Birkhead, J. Greenway, S. Mackessy, C. Phillips, J. Placyk, and J. Wright, who each facilitated the collection of numerous *Terrapene* and *Sistrurus* tissues. These results also would not have been possible without funding by the University of Arkansas in the form of generous endowments. Funding sources included: The Bruker Professorship in Life Sciences (Marlis Douglas), and the 21st Century Chair in Global Change Biology (Michael Douglas). The computationally challenging analyses herein were only possible due to the exceptionally well-maintained and powerful resources from the Arkansas High Performance Computing Center (AHPCC) and the NSF-funded JetStream Cloud-Computing System (XSEDE Allocation TG-BIO160065).

Special thanks go to my Ph.D. advisors, Drs. Marlis and Michael Douglas. Their continued support has encouraged both personal and professional growth and elevated my academic skillset. I appreciate their allowing me the intellectual freedom to pursue my preferred projects and professional aspirations, despite having made plenty of mistakes. Special thanks also to current and former Douglas Lab members who have entertained countless conversations to provide advice, support, motivation, and inspiration. They include: Drs. T. Chafin, S. Mussmann, and M. Bangs as well as Z. Zbinden, K. Wangchuk, and S. Wangchuk. I also thank my dissertation committee members, Drs. A. Alverson and J. Placyk, whose expertise and commitment to the development and completion of the projects herein have been of great value.

Finally, I owe a debt of gratitude to my family, without whom I undoubtedly would not have completed this dissertation. Their sacrifices and patience while I pursued my Ph.D. degree were extraordinary and will never be forgotten.

DEDICATION

To Pamela, Aurelia, Milla, Mom, and Dad, for your unwavering love and support

TABLE OF CONTENTS

LIST OF PUBLISHED PAPERS

Chapter I:

Martin BT, Douglas MR, Chafin TK, Placyk Jr. JS., Birkhead RD, Phillips CA, Douglas ME. 2020. Contrasting signatures of introgression in North American box turtle (*Terrapene* spp.) contact zones. *Molecular Ecology*. 29(21): 4186-4202.

Chapter II:

Martin BT, Chafin TK, Douglas MR, Placyk Jr. JS, Birkhead RD, Phillips CA, Douglas ME. 2021. The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (*Terrapene* spp.). *Molecular Ecology Resources*. In press.

INTRODUCTION

Hybridization (interbreeding of distinct evolutionary lineages) is widespread, but its evolutionary impact has surprisingly long been debated. The traditional view of hybridization centered upon it as being maladaptive, and thus serving to reinforce species boundaries and the concept of the biological species. Outspoken proponents of this negative view of hybridization included Dobzhansky (1937), Muller (1942), and Mayr (1963), all of whom were involved in developing the ‘modern synthesis.’ Although this perspective was contemporary at the time, it has subsequently been argued against on numerous occasions (Hedrick 2013). In fact, the tenor of this debate has shifted markedly, such that interpretations of hybridization and its manifestations are now seen as being far more important in the evolutionary process than previously thought (e.g., Abbott *et al.* 2013).

One phenomenon of hybridization as an evolutionary process is the formation of ‘hybrid zones’, areas were interbreeding amongst distinct lineages occurs at elevated frequencies and spanning multiple generations (Anderson 1948, 1949, 1953). Such 'hybrid zones' are often defined by physiographic processes at regional scales, where climate oscillation in synergy with geomorphological shifts have modulated dispersal and funneled populations into isolation, promoting divergence. Population expansion can follow, potentially leading to secondary contact of distinct lineages, and, in turn, increasing the potential for hybridization (Hewitt 1996, 2000, 2001).

In North America, glacial-interglacial cycles and geomorphological changes over the last two million years have provided a dynamic backdrop that promoted such oscillations of population contraction and expansion (Axelrod 1979, 1985; Soltis *et al.* 2006). These biogeographic processes have resulted in regional hybrid zones in the Southeast, Central, and

Southwest (Remington 1968; Swenson & Howard 2005). Hybrid zones are fascinating natural 'experiments' of hybridization as an evolutionary process because they are characterized in both historic and contemporary contexts, are manifested across various taxonomic groups, and often involve species complexes with unresolved phylogenies. They also generally coincide with areas of high biodiversity and the juxtaposition of different habitats (Moore 1977; Arnold 1997). Thus, they offer opportunities for comparative approaches to explore evolutionary mechanisms that underly the formation and persistence of hybrid zones and provide insights into genomic consequences of hybridization (Barton & Hewitt 1985; Harrison & Larson 2016). Understanding their evolutionary underpinnings can also promote effective conservation management strategies and help avert the anthropogenically-induced formation of new hybrid zones.

For my doctoral research, I addressed several general questions involving hybrid zones. 1) How have regional biogeographic processes influenced North American hybrid zones? 2) Are the hybrid zones of interest primarily reflective of historical and/ or contemporary processes? 3) Are the underlying genomic processes occurring differentially across the genome and with respect to species boundaries? 4) Do the hybrid zones differ in a phylogeographic context? Hence, I explore hybrid zones in two species complexes of reptiles across Eastern, Central, and Southwestern North America to explore genomic signatures of historic and contemporary hybridization and their interplay with local adaptation. I also examined the correlation between taxonomic divergence, introgression, selection, and environmental factors at a continental scale.

Chapter I assessed differential introgression in a complex of terrestrial turtles, *Terrapene* spp., from a contemporary hybrid zone in the southeastern United States. Transcriptomic loci were correlated with environmental predictors to evaluate mechanisms engendering maladapted hybrids and adaptive introgression. Selection against hybrids predominated inter-specific admixture, whereas directional introgression defined intra-specific admixture. Direction introgression, as inferred from outlier loci, primarily correlated with temperature, reflecting the temperature dependency of ectotherms and underscoring their vulnerability of these species to climate change.

Chapter II demonstrated a robust assessment of recently developed species delimitation software based on machine learning (M-L) to untangle phylogenetic relationships in a species complex defined by hybridization. Data filtering and M-L parameters were varied to determine their effects in resolving clusters. Minor allele frequency filters and low-to-moderate per-individual and per-population missing data were optimal in *Terrapene*. The results provide necessary recommendations on using M-L that, in this application, is still in its infancy. Furthermore, four *Terrapene* species were delimited with secondary contact supported, likely facilitated by glaciation in the Eastern United States. These data exemplify usage of M-L software in a phylogenetically complex group.

Chapter III describes an R package to visualize some of the analyses from Chapter I. Current software to generate genomic clines does not include functions to visualize the results, for which I wrote an API (application programming interface) that also performs other genomic and geographic cline-related tasks.

Chapter IV examines historical and contemporary phylogeographic patterns in the Massasauga Rattlesnakes (*Sistrurus* spp.) across the Southwest and Central Great Plains. In the Southwest, *S. tergeminus tergeminus* and *S. t. edwardsii* putatively diverged in the absence of strong physiographic barriers and physical glaciers. The results accordingly suggest they are undergoing primary divergence. In contrast, a disjunct population of *S. t. tergeminus* in Missouri has potentially been subjected to historical secondary contact with *S. catenatus*. These taxa represent contrasting examples of divergence resulting from alternative phylogeographic processes and contextualizes evolutionarily significant and management units.

The above chapters present population genomic data to elucidate impacts of phylogeographic processes on hybrid zones at a continental scale. The data will promote effective conservation management strategies, as many species in the focal regions have been affected by anthropogenic pressures. In this sense, the results can be extrapolated to co-distributed taxa with similar phylogeographic histories. Below I provide an overview of contemporary perspectives on hybridization, as well as what is currently known – or surmised – regarding how such perspectives define various aspects of biogeography and ecology.

Traditional perspective of hybridization

Hybridization has traditionally been considered inherently maladaptive, impacting the dynamics of gene flow and reproductive output (Dowling & Secor 1997). For example, hybridizing individuals may represent negative reproductive effort if hybrids are indeed sterile or offspring inviable (Rhymer & Simberloff 1996). The latter, in particular, may be a consequence of endogenous selection such as Dobzhansky-Muller incompatibilities or of sexual selection against hybrids (Coyne & Orr 2004; Gavrilets 2004). Maladapted and/or disrupted gene complexes may also occur in fertile hybrids, particularly if some traits are intermediate whereas the environment is not (exogenous selection). In either case, reinforcement of reproductive isolation or genetic assimilation were considered the endpoints under the Biological Species Concept [(BSC) (Mayr 1942, 1963)], with selection overshadowing recombination or *vice versa* (Abbott *et al.* 2013).

Hybridization as a promoter of evolution

On the other hand, fertile hybrids can potentially backcross with one or both parental species (i.e., ‘introgressive hybridization;’ Anderson & Hubricht 1938) and introduce new alleles into lineages which, in turn, can promote adaptive potential via recombination (Lewontin & Birch 1966; Barton & Hewitt 1989). Increasingly, molecular studies have demonstrated that hybridization and introgression can facilitate adaptive evolution, thus refuting its original premise of being maladaptive (per Mayr 1942, 1963). One such case includes the introgression of darker coat color from domestic dogs into coyotes and wolves (*Canis* spp.), leading to positive selection for darker pelages in populations in forests *versus* those inhabiting more open areas (Anderson *et al.* 2009). Similarly, coyote/grey wolf introgression led to local adaptation for larger body size in coyotes of eastern North America (vonHoldt *et al.* 2016). Another case involves Darwin’s finches (*Geospiza*), where climatic change and anthropogenic disturbance have promoted the fitness of hybrids, as well as provoking reproductive isolation among hybrids and parentals (Arnold 2015). More in-depth studies into both *Geospiza* and *Canis* also revealed that some regions of their genomes were introgressed, while others exhibited reduced inter-species gene flow, demonstrating the semipermeable nature of adaptive introgression and divergence (vonHoldt *et al.* 2016; Lawson & Petren 2017). In each of these cases, introgression allowed hybrids to reach adaptive peaks of fitness in heterogeneous landscapes that otherwise may have been inaccessible to the parental forms (Barton 2001). Thus, clearly introgression can facilitate evolutionary processes such as local adaptation and speciation.

Hybridization in the context of deep history

Genetic signatures of historic introgression remain in the population and can be detected in the absence of contemporary hybridization. Identification of historic introgression has several benefits, including a perspective on past distributions, speciation events, and local adaptation. Historic hybridization can also result in introgressed alleles that are retained in a subset of populations, thus characterizing distinct evolutionarily significant units (ESUs) or management units (MUs) that may require unique conservation management strategies (e.g., Placyk *et al.* 2012; vonHoldt *et al.* 2016). Finally, a recognition of historic hybridization can be beneficial with regard to uncertain phylogenetic relationships, particularly those where gene tree discordance is apparent, but with incomplete lineage sorting (ILS) as a possible counter-argument for the observed patterns (Maddison 1997; Bangs *et al.* 2018; Chafin *et al.* 2020).

Hybridization driven by climate change

Climate oscillations are important drivers of large-scale dispersal, alternately isolating populations or promoting secondary contact among distinct lineages. In North America, Pleistocene glaciations dominated such phylogeographic processes over last two million years. Glacial expansion forced species to converge on glacial refugia, forcing the congregation of previously isolated taxa (Anderson 1949; Remington 1968). In contrast, interglacial recession opened new habitat, promoting dispersal from glacial refugia and leading to post-glacial contact with distinct lineages isolated nearby. However, if phylogeographic breaks induced multiple allopatric refugia, populations closer to the glacial margin could potentially block lineages from more distant areas from returning. A zone of secondary contact could thus emerge, with hybridization promoted (Hewitt 1996, 1999).

Hybridization and environmental complexity

In addition, a mosaic of multiple habitat types in newly available area can lead to the breakdown of previously discrete gene pools, thus instigating contact (Rhymer & Simberloff 1996). Importantly, these areas wherein hybridization occurs at elevated frequencies spanning multiple generations, or ‘hybrid zones’ (Anderson 1948, 1949, 1953), can also shift coincident with ecological, biogeographic, or climatic conditions (Barton & Hewitt 1985), facilitating local adaptation that can be heterogeneously expressed across the genome to maintain both gene flow and reproductive isolation (Harrison & Larson 2016). Hybrid zones can also be heavily influenced by anthropogenic disturbance (Anderson 1948, 1949). Thus, from a conservation and management standpoint, a better understanding of hybrid zone dynamics is certainly required.

Hybridization in a contemporary context

Recognizing how signatures of hybridization reverberate on contemporary population structure, coupled with understanding the disruption of pre-zygotic barriers in a hybrid zone, can inform management decisions (Anderson 1948; Rhymer & Simberloff 1996). Specifically, anthropogenic disturbance can facilitate contact between sympatric or parapatric taxa, thus increasing the frequency of hybridization. This, in turn, can disrupt local adaptation, or result in the formation of a hybrid swarm wherein hybrid gene pools replace those of distinct parental lineages (Allendorf *et al.* 2001). Alternatively, the reinforcement of reproductive barriers can further promote population divergence and, eventually, lead to further speciation, particularly if reduced hybrid fitness is followed by pre- or post-zygotic isolation mechanisms (Dobzhansky 1936, 1940; Orr & Turelli 2001). Finally, hybridization within a localized area spanning multiple generations can promote breakdown of ecological barriers resulting in a hybrid zone.

Contemporary hybrid zones are also expected to have either mosaic or clinal distributions. They are typically maintained by two alternative processes: 1) Selective pressures enable hybrids to better survive in intermediate habitat ('ecotonal'; Moore 1977) and 2) hybrids have lower fitness in the hybrid zone, but it persists because parental types continually disperse into it and interbreed therein (environmentally independent 'tension-zone'; Bazykin 1969; Barton & Hewitt 1985). Selection pressures, and the potential for an environmental correlation, can be assessed by inspecting geographic and genomic clines (Endler 1977; Fitzpatrick 2013). For example, whether a particular habitat promotes hybridization can be determined by gauging the movement of the hybrid zone as environmental conditions change, and how this may be driven by habitat degradation.

Hybrid zones of interest in North America

Rationale

Understanding the mechanisms underlying formation of historic and contemporary hybrid zones is important to define appropriate management strategies that prevent situations that would promote the inadvertent formation of new hybrid zones. First, the study of an historic hybrid zone such as that in midwestern North America will shed light on those processes that promoted hybrid zone formation in the past. It also will help understand mechanisms that allow locally adapted but introgressed populations to persist. Second, by clarifying how anthropogenically induced secondary contact between previously isolated taxa promote hybrid zones, management strategies can be implemented that avoid such situations from occurring in the first place. Lastly, pro-active management strategies can target anticipated effects of climate change such as altered distributions of species and/or altered environments. Climate change is predicted to enhance the frequency and extent of introgression as well as detrimental or beneficial effects on locally adapted alleles under selection (Taylor *et al.* 2014, 2015). For example, selection against hybrids can reinforce distinct gene pools (Orr & Turelli 2001), and adaptive introgression (i.e., introgression provides adaptive capacity) can allow range expansion or even facilitate evolutionary rescue from inhospitable environments (Hamilton & Miller 2016; Oziolor *et al.* 2019). These processes can be monitored by sampling transects in hybrid zones, sequencing genome-wide loci, and deriving genomic clines that regress ancestry against parental allele frequencies (Chown *et al.* 2015; Taylor *et al.* 2015; Gompert *et al.* 2017). In so doing, selection and adaptive introgression can be contrasted with environmental variables, allowing one to elucidate the manner by which individuals respond to habitat degradation, and consequently, how hybrid zones will shift accordingly.

Regional foci

Biodiversity and endemism are elevated in southeastern North America (Ricketts 1999), yet the region harbors many species of conservation concern (Lydeard & Mayden 1995). Anthropogenic disturbance, such as the clear-cutting of forests and the introduction of non-native species are also prevalent (Stapanian *et al.* 1997, 1998). In addition, multiple ecological regions are juxtaposed in this region which also is characterized by overlapping phylogeographic breaks (U.S.E.P.A. 2013). This has resulted in unique ecological, climatic, and biogeographical characteristics that facilitated formation of hybrid zones in the region (Swenson & Howard 2005), as reflected by areas that containing hybrid zones across a variety of species (Remington 1968; Swenson & Howard 2004). The region is sufficiently far south to have avoided Pleistocene glaciation, and accordingly served instead as a glacial refugia (Hewitt 2000). However, the contemporary processes that maintain the hybrid zone, and how they may change over time, certainly warrant a more thorough investigation.

Similarly, the southwestern United States also provides a worthwhile system within which to assess hybridization and introgression. It too has unique biodiversity, and much of the shortgrass prairie habitat of eastern Colorado, New Mexico, and western Texas has been lost or degraded (Samson *et al.* 2004). Furthermore, it may also represent a suture zone that formed from the convergence of post-glacial dispersal routes (Swenson & Howard 2005). Alternatively, the region has limited physiographic barriers to cause vicariant events in terrestrial species and lacked the physical presence of glaciers (Licciardi *et al.* 2004). Accordingly, the southwestern climate saw drastic shifts in temperature, precipitation, and vegetational composition during Pleistocene glacial-interglacial cycles (Axelrod 1948, 1979, 1983; Owen *et al.* 2003). These changes may have maintained mosaic hybrid zones, with gene flow persisting over time as mesic and xeric refugia expanded and contracted with the climate (Axelrod 1979; Van Devender *et al.* 1987). They may also have facilitated ecological speciation, with divergence occurring when the habitat preferences of sympatric populations became sufficiently different to restrict gene flow (Douglas *et al.* 2006).

Midwestern North America, on the other hand, represents an historic hybrid zone that stems from the contact zones developed during post-glacial recolonization. Contact between northern and southern taxa in the Midwest subsequently promoted dispersal into northeastern North America via the ‘Prairie Corridor.’ This region then evolved into a natural hybrid zone (Swenson & Howard 2005). However, some populations may have dispersed in atypical routes toward the southwest during post-glacial recession (Swenson & Howard 2005), a hypothesis that needs to be investigated further.

Overall implications

Herein, I explored hybrid zone dynamics by analyzing next-generation sequencing data across historic and contemporary perspectives and at a broad geographic scale. Data that quantify local adaptation, speciation, species distributions, selection, and hybridization/ introgression were employed. Specifically, the studies below broaden our knowledge on the evolutionary history of hybrid zones in Southeast, Central, and Southwest North America, and provide data on the impacts of climate change and anthropogenic disturbance on these processes. They also have broader implications for a wide variety of taxa because other terrestrial organisms will likely react similarly to habitat loss and anthropogenic disturbance.

REFERENCES

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, and Buggs R (2013) Hybridization and Speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Allendorf FW, Leary RF, Spruell P, and Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology and Evolution*, **16**, 613–622.

Anderson E (1948) Hybridization of the habitat. *Evolution*, **2**, 1–9.

Anderson E (1949) *Introgressive Hybridization*. John Wiley and Sons, New York City, NY, USA.

Anderson E (1953) Introgressive hybridization. *Biological Reviews*, **28**, 280–307.

Anderson TM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, and Bustamante CD (2009) Molecular and evolutionary history of melanism in North American gray wolves. *Science*, **323**, 1339–1343.

Anderson E and Hubricht L (1938) The evidence for introgressive hybridization. *American Journal of Botany*, **25**, 396–402.

Arnold ML (1997) *Natural Hybridization and Evolution*. Oxford University Press, New York, NY, USA.

Arnold ML (2015) *Divergence with Genetic Exchange*. Oxford University Press, New York City, New York, USA.

Axelrod DI (1948) Climate and evolution in western North America during middle Pliocene time. *Evolution*, **2**, 127–144.

Axelrod DI (1979) Age and origin of Sonoran Desert vegetation. *Occasional Papers of the California Academy of Sciences*, **132**, 1–74.

Axelrod DI (1983) Paleobotanical history of the western deserts. In: *Origin and Evolution of Deserts* (eds Wells SG and Haragan DR), pp. 113–129. University of New Mexico Press, Albuquerque, New Mexico, USA.

Axelrod DI (1985) Rise of the grassland biome, central North America. *Bot. Rev.*, **51**, 163–201.

Bangs MR, Douglas MR, Mussmann SM, and Douglas ME (2018) Unraveling historical introgression and resolving phylogenetic discord within *Catostomus* (Osteichthys: Catostomidae). *BMC Evolutionary Biology*, **18**, 86.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

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

Barton NH and Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.

Bazykin AD (1969) Hypothetical mechanism of speciaton. *Evolution*, **23**, 685–687.

Chafin TK, Douglas MR, Bangs MR, Martin BT, Mussmann SM, and Douglas ME (2020) Taxonomic Uncertainty and the Anomaly Zone: Phylogenomics Disentangle a Rapid Radiation to Resolve Contentious Species (*Gila robusta* complex) in the Colorado River. *bioRxiv*, 692509.

Chown SL, Hodgins KA, Griffin PC, Oakeshott JG, Byrne M, and Hoffmann AA (2015) Biological invasions, climate change and genomics. *Evolutionary Applications*, **8**, 23–46.

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

Van Devender TR, Thompson RS, and Betancourt JL (1987) Vegetation history of the deserts of southwestern North America: the nature and timing of the late Wisconsin-Holocene transition. In: *North America and Adjacent Oceans During the Last Glaciation: The Geology of North America* (eds Ruddiman WF and Wright Jr. HE), pp. 323–352. Geological Society of America, Boulder, CO, USA.

Dobzhansky TH (1936) Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics*, **21**, 113–135.

Dobzhansky T (1937) *Genetics and the Origin of Species*. Columbia University Press, New York City, NY, USA.

Dobzhansky T (1940) Speciation as a stage in evolutionary divergence. *American Naturalist*, **74**, 312–321.

Douglas ME, Douglas MR, Schuett GW, and Porras LW (2006) Evolution of rattlesnakes (Viperidae; *Crotalus*) in the warm deserts of western North America shaped by Neogene vicariance and Quaternary climate change. *Molecular Ecology*, **15**, 3353–3374.

Dowling TE and Secor CL (1997) The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics*, **28**, 593–619.

Endler JA (1977) *Geographic variation, speciation, and clines*. Princeton University Press, Princeton, NJ, USA.

Fitzpatrick BM (2013) Alternative forms for genomic clines. *Ecology and Evolution*, **3**, 1951–1966.

Gavrilets S (2004) *Fitness landscapes and the origin of species*. Princeton University Press, Princeton, NJ, USA.

Gompert Z, Mandeville EG, and Buerkle CA (2017) Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 207–229.

Hamilton JA and Miller JM (2016) Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, **30**, 33–41.

Harrison RG and Larson EL (2016) Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Molecular Ecology*, **25**, 2454–2466.

Hedrick PW (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, **22**, 4606–4618.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnaean Society*, **58**, 247–276.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnaean Society*, **68**, 87–112.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hewitt GM (2001) Speciation, hybrid zones and phylogeography—or seeing genes in space and time. *Molecular Ecology*, **10**, 537–549.

Lawson LP and Petren K (2017) The adaptive genomic landscape of beak morphology in Darwin’s finches. *Molecular Ecology*, **26**, 4978–4989.

Lewontin RC and Birch LC (1966) Hybridization as a source of variation for adaptation to new environments. *Evolution*, **20**, 315–336.

Licciardi JM, Clark PU, Brook EJ, Elmore D, and Sharma P (2004) Variable responses of western US glaciers during the last deglaciation. *Geology*, **32**, 81–84.

Lydeard C and Mayden RL (1995) A diverse and endangered aquatic ecosystem of the southeast United States. *Conservation Biology*, **9**, 800–805.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Mayr E (1942) *Systematics and the Origin of Species: From the Viewpoint of a Zoologist*. Harvard University Press, Cambridge, MA, USA.

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

Moore WS (1977) An evaluation of narrow hybrid zones in vertebrates. *Quarterly Review of Biology*, **52**, 263–277.

Muller HJ (1942) Isolating mechanisms, evolution and temperature. In: *Temperature, Evolution, Development Biological Symposia: A Series of Volumes Devoted to Current Symposia in the Field of Biology* (ed Dobzhansky T), pp. 71–125. Jaques Cattell Press, Lancaster, PA, USA.

Orr HA and Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution*, **55**, 1085–1094.

Owen LA, Finkel RC, Minnich RA, and Perez AE (2003) Extreme southwestern margin of late Quaternary glaciation in North America: timing and controls. *Geology*, **31**, 729–732.

Oziolor EM, Reid NM, Yair S, Lee KM, Guberman VerPloeg S, Bruns PC, Shaw JR, Whitehead A, and Matson CW (2019) Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science*, **364**, 455–457.

Placyk JS Jr., Casper GS, Fitzpatrick BM, Small RL, Reynolds RG, Noble DWA, Brooks RJ, Burghardt GM. 2012. Hybridization between two gartersnake species (*Thamnophis*) of conservation concern: A threat or an important natural interaction? *Conservation Genetics*, **13**, 649–663.

Remington CL (1968) Suture-zones of hybrid interaction between recently joined biotas. In: *Evolutionary Biology* (ed Dobzhansky T), pp. 321–428. Springer, New York, NY, USA.

Rhymer JM and Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Ricketts TH (1999) *Terrestrial ecoregions of North America: a conservation assessment*. Island Press, Washington, DC, USA.

Samson FB, Knopf FL, and Ostlie WR (2004) Great Plains ecosystems: past, present, and future. *Wildlife Society Bulletin*, **32**, 6–15.

Soltis DE, Morris AB, McLachlan JS, Manos PS, and Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Stapanian MA, Cassell DL, and Cline SP (1997) Regional patterns of local diversity of trees: associations with anthropogenic disturbance. *Forest Ecololgy and Management*, **93**, 33–44.

Stapanian MA, Sundberg SD, Baumgardner GA, and Liston A (1998) Alien plant species composition and associations with anthropogenic disturbance in North American forests. *Plant Ecology*, **139**, 49–62.

Swenson NG and Howard DJ (2004) Do suture zones exist? *Evolution*, **58**, 2391–2397.

Swenson NG and Howard DJ (2005) Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American Naturalist*, **166**, 581–591.

Taylor SA, Larson EL, and Harrison RG (2015) Hybrid zones: windows on climate change. *Trends in Ecology and Evolution*, **30**, 398–406.

Taylor SA, White TA, Hochachka WM, Ferretti V, Curry RL, and Lovette I (2014) Climate-mediated movement of an avian hybrid zone. *Current Biology*, **24**, 671–676.

U.S.E.P.A. (2013) *Level III ecoregions of the continental United States, map scale 1:7,500,000,*. National Health and Environmental Effects Research Laboratory, Corvallis, Oregon, USA.

vonHoldt BM, Kays R, Pollinger JP, and Wayne RK (2016) Admixture mapping identifies introgressed genomic regions in North American canids. *Molecular Ecology*, **25**, 2443–2453.

CHAPTER I

Contrasting signatures of introgression in North American box turtle (*Terrapene* spp.) contact zones

Bradley T. Martin, Marlis R. Douglas, Tyler K. Chafin, John S. Placyk, Jr., Roger D. Birkhead, Christopher A. Phillips, Michael E. Douglas

ABSTRACT

Hybridization occurs differentially across the genome in a balancing act between selection and migration. With the unprecedented resolution of contemporary sequencing technologies, selection and migration can now be effectively quantified such that researchers can identify genetic elements involved in introgression. Furthermore, genomic patterns can now be associated with ecologically relevant phenotypes, given availability of annotated reference genomes. We do so in North American box turtles (*Terrapene*) by deciphering how selection affects hybrid zones at the interface of species boundaries and identifying genetic regions potentially under selection that may relate to thermal adaptations. Such genes may impact physiological pathways involved in temperature-dependent sex determination, immune system functioning, and hypoxia tolerance. We contrasted these patterns across inter- and intra-specific hybrid zones that differ temporally and biogeographically. We demonstrate hybridization is broadly apparent in *Terrapene*, but with observed genomic cline patterns corresponding to species boundaries at loci potentially associated with thermal adaptation. These loci display signatures of directional introgression within intra-specific boundaries, despite a genome-wide selective trend against intergrades. In contrast, outlier loci for inter-specific comparisons exhibited evidence of being under selection

against hybrids. Importantly, adaptations coinciding with species boundaries in *Terrapene* overlap with climatic boundaries and highlight the vulnerability of these terrestrial ectotherms to anthropogenic pressures.

1. INTRODUCTION

Hybrid zones are natural laboratories that allow the genetic architecture of local adaptation and/or reproductive isolation to be examined. They frequently juxtapose with underlying ecological gradients, allowing researchers to quantify the how selection impacts the genome (Barton & Hewitt 1985; Payseur 2010). Here, selection might prevent introgression at loci underpinning crucial adaptations while the rest of the genome essentially homogenizes (Via 2009; Feder *et al.* 2013). Once such loci are identified, the phenotypes inferred by adaptive divergence can then be inferred via a bottom-up, "reverse-ecology" approach (Li *et al.* 2008; Tiffin & Ross-Ibarra 2014). Hybrid zones effectively become unique "windows" into the speciation process by allowing functional loci to be associated with various aspects of ecology (Taylor *et al.* 2015).

Diminishing costs associated with genomic sequencing, coupled with an upsurge in genomic annotations, has facilitated the reverse-ecology approach. Examples include adaptive divergence in seasonal growth and variability in immune responses (Rödin-Mörch *et al.* 2019), proactive responses to environmental gradients (Keller & Seehausen 2012; Guo *et al.* 2016; Waterhouse *et al.* 2018; Teske *et al.* 2019), and an upsurge in contemporary effects such as anthropogenic modulation of reproductive boundaries (Garroway *et al.* 2010; Taylor *et al.* 2014; Grabenstein & Taylor 2018). Importantly, researchers can now effectively gauge the how the genome is impacted by ecological and climatic shifts, and the how species distributions are promoted by pre-existing adaptive gradients (Rosenzweig *et al.* 2008; Taylor *et al.* 2015; Ryan *et al.* 2018). Thus, genome-scale datasets can often refine our perspectives on two major evolutionary patterns: First, reproductive boundaries among historically co-existing species can become blurred due to substantial environmental change, a situation directly analogous to the imposition of contact among otherwise allopatric taxa (Rhymer & Simberloff 1996). Second, selection and migration are balanced within hybrid zones (Key 1968; Parmesan *et al.* 1999), and this balance can shift due to rapid and/or exceptional climate change (Seehausen *et al.* 2008; Kearns *et al.* 2018). Genomic data extracted from hybrid zones may thus allow species boundaries to be defined according to their phenotypic and genetic underpinnings.

Species boundaries can either be strengthened (Ryan *et al.* 2018) or eroded (Muhlfeld *et al.* 2014) due to temperature shifts, a major component of climate change, potentially shifting species distributions and/or hybrid zone dynamics. Furthermore the effects of temperature on physiological and cellular mechanisms are well known (Kingsolver 2009), directly affecting growth, development, reproduction, locomotion, and immune response (Keller & Seehausen 2012). As a result, the manner by which thermal gradients interact with species boundaries has become a major focus (Qin *et al.* 2013). Herein, we attempt to clarify how species boundaries reflect environmental processes by quantifying the geographic and ecological foundations of two hybrid zones in the ectothermic North American box turtles (*Terrapene*).

1.1 Hybridization in North American box turtles

North American box turtles (Emydidae, *Terrapene*) are long-lived, omnivorous, and primarily terrestrial ectotherms, with a rectangular appearance defined by a dome-shaped dorsal carapace and a ventral plastron hinged to tightly close against the carapace (hence the common name) (Dodd 2001). Their North American range is characterized by two well-known zones of hybridization (Milstead 1969; Dodd 2001; Cureton *et al.* 2011) that provide excellent models from which to contrast regional patterns of hybridization and introgression. To do so, we evaluated four southeastern taxa [the Woodland (*T. carolina carolina*), Gulf Coast (*T. c. major*), Three-toed (*T. carolina triunguis*), and Florida (*T. bauri*) box turtles (Auffenberg 1958, 1959; Milstead & Tinkle 1967; Milstead 1969; Martin *et al.* 2013; Iverson *et al.* 2017)], and two midwestern (the Ornate box turtle, *T. ornata ornata* and *T. c. carolina*; Cureton *et al.* 2011). Each hybridizes regionally, and therefore we focus on inter- and intra-specific contacts within these two regions.

One focal hybrid zone is nested within southeastern North America (Ricketts 1999), where box turtles inhabit a biodiversity hotspot. Here, clear-cutting, invasive species, and altered fire regimes are widespread (Stapanian *et al.* 1997, 1998; van Lear & Harlow 2002), and impact numerous endemic species (Lydeard & Mayden 1995). The region also displays clinal intergradation (i.e., interbreeding between subspecies), as well as hybridization across a variety of taxa (Remington 1968; Swenson & Howard 2004), due largely to coincident ecological and climatic transitions (Swenson & Howard 2005).

By contrast, contact zones in midwestern North America seemingly stem from secondary contact (i.e., resumption of interbreeding following a geographic separation), as associated with postglacial recolonization/ expansion (Swenson & Howard 2005). Here, prairie-grassland habitat has been anthropogenically fragmented such that niche overlap now occurs between grassland and woodland species (Johnson 1994; Samson & Knopf 1994; Rhymer & Simberloff 1996; Samson *et al.* 2004). Furthermore, while overlapping forms in the Midwest represent distinct species, southeastern forms are taxonomically in flux. Species status has varied for *T. m. triunguis*, and *T. c. major* is now viewed as an intergrade population (Butler *et al.* 2011; Iverson *et al.* 2017), despite recent molecular work suggesting specific status for *T. m. triunguis* and phylogenetic structure in *T. c. major* (Martin *et al.* 2013, 2014, 2020). For the sake of clarity, we follow Martin *et al.*, regarding *T. m. triunguis* as a full species and *T. c. major* as a recognized subspecies.

We used ddRAD sequencing (Peterson *et al.* 2012) to contrast genome-wide patterns of clinal introgression within each hybrid zone. We then identified/ quantified loci potentially under selection by mapping them to an available genomic reference and accordingly interpreted their potential ecological associations, which provide invaluable insights into how *Terrapene* has responded to a fluctuating climate. As such, our results extend to a proactive management paradigm that underscores the conservation of co-occurring forms.

2. MATERIALS AND METHODS

2.1. Tissue and DNA collection

Tissues for *T. carolina*, *T. ornata*, and *T. mexicana triunguis* were collected by volunteers and agency collaborators (Table S1). Additional samples were provided by numerous museums and organizations. Live animals were sampled non-invasively (e.g., blood, toenails, or toe-clips), whereas road-kills were sampled indiscriminately. Isolation of genomic DNA was performed using DNeasy Blood and Tissue Kits (QIAGEN), QIAamp Fast DNA Tissue Kit (QIAGEN), and E.Z.N.A. Tissue DNA Kits (Omega Bio-tek). The presence of genomic DNA was confirmed via gel electrophoresis using a 2% agarose gel.

2.2. Library preparation

In silico digests were carried out to optimize restriction enzyme selection, using available genomic references [Painted turtle (*Chrysemys picta*), GenBank Accession #: GCA_000241765.2 (Shaffer *et al.* 2013); FRAGMATIC (Chafin *et al.* 2018); genome size=2.59 X 10^9 bp]. The distribution of fragments (from N=24 samples) were first optimized then evaluated using an Agilent 4200 TapeStation. Library preparation was conducted per standard protocol (Peterson *et al.* 2012), using *PstI* (5'-CTGCAG-3') and *MspI* (5'-CCGG-3') restriction enzymes. We digested ~500-1,000ng of DNA/ sample at 37°C, with unique DNA barcode and sequencing adapters subsequently ligated. Prior to sequencing, quality control checks were performed at the core facility, to include fragment analysis for confirmation of correct size range and quantitative real-time PCR. Individuals (N=96) were pooled per lane of single-end Illumina sequencing at the University of Oregon Genomics and Cell Characterization Core Facility (Hi-Seq 4000, 1x100bp). Populations were randomized across multiple lanes to mitigate batch effects.

2.3. Assembly and quality control

Read quality was quantified using FASTQC v. 0.11.5, then demultiplexed and aligned using IPYRAD v. 0.7.28 (Eaton & Overcast 2020), with reads mapped to the scaffold-level *T. mexicana triunguis* reference assembly (GenBank Accession #: GCA_002925995.2) at a distance threshold of 0.15. Non-mapping reads were discarded. This alignment is herein termed the "scaffold alignment" to differentiate it from a separate transcriptome-mapped alignment (see section 2.6). Barcodes and adapters were trimmed, as were the last five base pair (bp) of each read. Those exceeding five bases with low PHRED quality score (<33) were discarded, and potential paralogs were filtered by excluding loci with high heterozygosity (>75%) or >2 alleles per individual. Loci with a sequencing depth of <20X per individual or <50% presence across individuals were also discarded. Our mapping and filtering steps above were conducted in IPYRAD.

2.4. Assessing admixture and population structure

ADMIXTURE (Alexander *et al.* 2009) was used to assess contemporary hybridization. It employs a model-based ML approach that estimates the proportion of ancestry shared across the genome-wide average of each sample. K=1-13 were used for datasets containing all sequenced taxa and subsets from the Southeast and Midwest hybrid zones, with 20 independent replicates per K (ADMIXPIPE; Mussmann *et al.* 2020). Hierarchical partitioning was done because ADMIXTURE often underestimates K by detecting only the uppermost hierarchy of population structure (Evanno *et al.* 2005). SNP data were pre-filtered using VCFTOOLS (Danecek *et al.* 2011), with SNPs randomly thinned to one per locus to alleviate linkage bias and filtered by removing sites with a minor allele frequency (MAF)<1.0% to reduce bias associated with erroneous genotypes and singletons (Linck & Battey 2019). Model support across K-values was assessed using five-fold cross-validation (Alexander *et al.* 2009). ADMIXPIPE output was summarized using the CLUMPAK server (Kopelman *et al.* 2015), with each individual subsequently plotted as a stacked bar-chart (Rosenberg 2004).

We also performed Discriminate Analysis of Principal Components (DAPC) using the *adegenet* R-package (v2.0-0) with identical filtering parameters (Jombart *et al.* 2010). The *find.clusters()* function was utilized with 1,000,000 iterations to determine the optimal K with the lowest Bayesian Information Criterion (BIC). DAPC cross-validation (100 replicates, 90%

training dataset) then evaluated which principle components and discriminant functions to retain, with individuals plotted against the top three DAPC axes.

Finally, we also ran TESS3 (TESS3R R package; Caye *et al.* 2016) to estimate ancestry coefficients (as with ADMIXTURE), but also incorporate spatial proximity into ancestral genotype estimates. The TESS3 input alignments were subsets of those used in ADMIXTURE to include only individuals with GPS coordinates. Cross-validation (with 10% sites randomly masked) was performed for K=1-10 with twenty independent runs to assess optimal *K*. The output Q-matrix was interpolated using spatial kriging (Jay *et al.* 2012).

2.5. Identifying hybrids

NEWHYBRIDS (Anderson & Thompson 2002) was used to assign statistically-supported hybrids to genotype frequency classes (i.e., Pure, F_1, F_2, and backcrosses between F_1 and parental types). The *getTopLoc()* function in HYBRIDDETECTIVE (Wringe *et al.* 2017a) reduced the data to 300 loci containing the highest among-population differentiation (F_{ST}) and lowest linkage disequilibrium correlation ($r^2<0.2$). Burn-in was 500,000 MCMC generations followed by 2,000,000 post burn-in sweeps. Seeds were randomized and the analysis employed the Jeffrey's prior for θ and π. To train the data, individuals sampled outside the focal hybrid zones with ADMIXTURE proportions=100% were pre-assigned as parentals. The following combinations of taxa were employed: *Terrapene carolina carolina* X *T. c. major*, *T. c. carolina* X *T. m. triunguis, T. c. major* X *T. m. triunguis*, and *T. c. carolina* X *T. o. ornata*. A posterior probability threshold >0.8 was required for assignment into the genotype frequency classes, as determined using a power analysis conducted with HYBRIDDETECTIVE and PARALLELNEWHYBRID pipelines (Wringe *et al.* 2017b; a).

2.6. Genomic clines among scaffolds and mRNA mapping

In addition to the "scaffold alignment," IPYRAD was rerun with reads mapped to the *T. m. triunguis* reference transcriptome (GenBank Accession: GCA_002925995.2), with identical filtering parameters. Three subsets of the resulting "transcriptomic alignment" were generated to retain only individuals per each pairwise combination of southeastern taxa (*T. c. carolina*, *T. c. major*, *T. m. triunguis*). The "scaffold" and "transcriptome" alignments were then independently examined for patterns of differential introgression using INTROGRESS (Gompert & Buerkle 2010) and Bayesian Genomic Clines (Gompert & Buerkle 2012). Both generate genomic clines, which assess locus-specific ancestry to identify outliers *versus* the genome-wide average and can identify outlier SNPs having cline shapes divergent from neutral expectations. Parental reference populations were determined *a priori* via population structure and NEWHYBRIDS results. Parental status was considered only for samples with ADMIXTURE ancestry coefficients=100%.

INTROGRESS derived neutral expectations from 1,000 parametric simulations (Gompert & Buerkle 2010). Genomic clines were only generated for SNPs with a high allele frequency differential (δ) between parental types (Andrés *et al.* 2013). Outlier SNPs were defined using a Bonferroni-corrected α-significance threshold.

Prior to running BGC, sites with a MAF<5% were removed because an over-abundance of uninformative loci inhibited parameter convergence. Five independent BGC runs were conducted for both scaffold and transcriptomic alignments, which included 1,000,000 and

1,800,000 burn-in, respectively, each with 200,000 post-burn-in generations. Samples were thinned with every 50 iterations retained to mitigate auto-correlation. Genotype uncertainty corrections were applied to each locus, with the sequencing error rate prior computed in IPYRAD (ranging from 0.1-0.2%). The BGC linkage model was tested but found to be computationally intractable (i.e., >1Tb memory and unreasonable running times). Upon run completion, parameter traces were visually inspected for convergence. Replicate runs were subsequently combined.

Two parameters [i.e., genomic cline center (α) and rate (ß)] represented BGC output (Gompert & Buerkle 2011). The α-parameter indicates the direction of introgression, with negative and positive outliers depicting excess P_1 and P_2 ancestry, respectively. ß characterizes the rate of change, with negative values reflecting a wider genomic cline where loci more freely introgress, and a steeper cline indicated by positive values with relatively sharp transitions from P_1 to P_2 ancestry. BGC outliers were considered significant if they met either of two criteria: 1) the 95% credible intervals for α or ß did not overlap zero, or 2) the median of the posterior distribution exceeded the probability distribution's quantile interval [$(\frac{1-0.975}{2}, \frac{0.975}{2})$; (Gompert & Buerkle 2011)].

2.7. Mapping BGC outliers to chromosomes

BGC parameters α and ß were plotted onto assembled chromosomes to visualize the distribution of outliers across the genome. The *Terrapene* reference is a scaffold-level assembly, so MINIMAP2 (Li 2018) and PAFSCAFF (https://github.com/slimsuite/pafscaff) were used to map the *T. m. triunguis* reference to a closely related chromosome-level assembly (*Trachemys* *scripta*; Simison *et al.* 2020; GenBank accession: GCA_013100865.1). The ASM20 MINIMAP2 preset was chosen to accommodate the ~5-10% sequence divergence expected between *Trachemys* and *Terrapene* (Feldman & Parham 2002). The mapping connected *Terrapene* BGC outliers with *Trachemys* chromosomal positions, allowing the putative BGC outlier locations to be visualized at the chromosome-level, with plots subsequently generated (RIDEOGRAM R package; Hao *et al.* 2020).

2.8. Correlating outliers with environmental variables

We also independently identified SNP outliers by environmental association using redundancy analysis (RDA). Outlier SNPs were correlated with the standard WorldClim v2 Bioclimatic variables that included 19 raster layers of temperature and precipitation, plus mean annual solar radiation, mean annual wind speed, and elevation (Fick & Hijmans 2017). The finest available scale (30 arc-seconds) was chosen for each raster. Layers were projected to WGS84 and cropped to the sampling extent (RASTER R package; Hijmans & Van Etten 2016). Raster values at each sampling location were then extracted.

Each predictor variable was scaled, centered, and tested for normality with a Shapiro-Wilks test. Non-normal distributions were transformed using the BESTNORMALIZE R package (Peterson & Cavanaugh 2019) per RDA's assumptions. The environmental layers were assayed for predictive capabilities with uncorrelated variables retained (ADESPATIAL R package, forward selection with 10,000 permutations). Predictors that failed forward selection were removed. To account for underlying spatial influence, distance-based Moran's eigenvector maps (dbMEM) were generated using sample coordinates (*quickMEM()* R function; Borcard *et al.* 2018). The dbMEMs are a matrix of axes that capture spatial patterns from multiple angles rather than just a latitudinal or longitudinal vector. To reduce overfitting, informative, non-redundant dbMEM axes were subset, using forward selection with 1,000 permutations. Finally, the SNP matrix was imported into R using ADEGENET, and missing data were imputed as the most frequent allele per population, following RDA assumptions.

A partial RDA (pRDA) conditioned on the dbMEM spatial matrix was then conducted using genotypes as the response variables, with 1,000 permutations (VEGAN R package; Oksanen *et al.* 2019). This approach "partialed out" spatial autocorrelative effects that could yield false negative SNP-environment associations. Significant RDA axes were determined using VEGAN's *anova.cca*() function, and SNPs with loadings +/- 3 standard deviations from the mean on a significant axis were considered outliers. A full RDA and a pRDA conditioned on environment were also conducted to estimate the contributions of spatial *versus* environmental predictors. Each SNP was then correlated pairwise with all environmental variables using Pearson's correlation coefficient (*r*), and those with the strongest correlations represented the best supported SNP-environment association.

3. RESULTS

A total of 368 individuals (Tables S1, S2) were retained across 12,052 (combined alignment), 10,338 (Midwest-only), and 11,308 (Southeast-only) unlinked reference-mapped loci. This, a result of quality control steps and post-alignment filters that eliminated individuals with >90% missing data, and sites with MAF<1.0%. The scaffold alignment included 134,607 variable and 90,777 parsimoniously informative sites. The transcriptome-guided alignment contained 2,741 bi-allelic SNPs across 247 individuals, with subsets generated for *T. c. carolina*

X *T. c. major* (EAxGU), *T. c. carolina* X *T. m. triunguis* (EAxTT), and *T. c. major* X *T. m. triunguis* (GUxTT).

3.1. ADMIXTURE across the hybrid zones

The combined east/ west ADMIXTURE CV (Fig. S1) supported *K*=6 (x̄= 0.19279, SD = 0.00017), followed by *K*=4 (x̄= 0.19490, SD = 0.0016) and *K*=5 (x̄= 0.19738, SD = 0.00013). The analysis indicated population structure for *T. c. carolina*, *T. m. triunguis*, two distinct *T. c. major* subpopulations (Alabama/ Mississippi and Florida panhandles), and northern and southern *T. o. ornata* subpopulations from Illinois+Wisconsin+Iowa and Kansas+Texas+Colorado+Nebraska (Fig. S2). *Terrapene bauri* was excluded due to limited sampling (N=4). Admixture occurred between *T. c. carolina* X *T. c. major* (EAxGU), *T. c. carolina* X *T. m. triunguis* (EAxTT), *T. c. major* X *T. m. triunguis* (GUxTT), *T. o. ornata* X *T. c. carolina* (EAxON), and the two *T. o. ornata* subpopulations.

The lowest CV score for the southeastern ADMIXTURE (N=259 individuals) was at *K*=4 (x̄= 0.21851, SD=0.00016), with *K*=3 (x̄=0.22134, SD=0.00015) and *K*=5 (x̄= 0.22519, SD=0.00082) trailing (Fig. S3). Southeastern taxa included *T. c. carolina*, *T. c. major*, and *T. m. triunguis,* and their analysis concurred with the all-taxa dataset in terms of both population structure and admixed taxa (Figs. 1, S4). Admixture primarily occurred throughout Alabama and the Florida panhandle (EAxGU), Georgia and South Carolina (EAxTT), and Mississippi/ southern Alabama (GUxTT). The two *T. c. major* subpopulations in the all-taxa analysis were corroborated. The same four southeastern groups plus *T. bauri* were also produced by DAPC

(*K*=5; Fig. S5). We found *T. bauri* highly differentiated along axis 1 (71.9% variance explained), whereas axes 2-3 delineated the remaining southeastern taxa (17.5% and 5.73%, respectively).

The midwestern analysis (Figs. 1, S6) included an optimal *K*=2 (x̄= 0.24069, SD= 0.00023), followed by *K*=3 (x̄=0.25703, SD=0.00454) and *K*=4 (x̄=0.25861, SD=0.00374) (Fig. S7). The *K*=2 groups consisted of *T. c. carolina* and *T. o. ornata*, with only a few individuals indicating admixture. At *K*=3, *T. c. carolina* from Illinois split as a distinct group, although only a few of the admixture proportions approached 100%. At *K*=4, the northern and southern *T. o. ornata* subpopulations produced by the all-taxa analysis also materialized.

TESS3 corroborated both midwestern and southeastern ADMIXTURE analyses, with *T. o. ornata*, *T. c. carolina T. m. triunguis*, and the two *T. c. major* subpopulations in Alabama/ Mississippi and Florida being delineated (Fig. 1). The Kriging interpolation also spatially highlighted ancestry gradients consistent with ADMIXTURE (FIG. S8), with lower surface prediction scores concordant with areas that contained frequently mixed ancestry.

3.2. Genealogical hybrid classification

HYBRIDDETECTIVE confirmed convergence for inter- and intra-simulation replicates (Fig. S9) with 500,000 burn-in and 2,000,000 post-burn-in sweeps (the EAxON analysis required 4,000,000 sweeps with 1,000,000 burn-in). Our power analyses suggested 90% assignment accuracy (+/- SD) for all genotype classes at a critical threshold of 0.8 (Figs. S10, S12, S14, S16). Statistical power was also elevated (≥0.8), although some genotype classes for EAxGU displayed relatively lower power (<0.8) (Figs. S11, S13, S15, S17).

The southeastern hybrid zone consisted entirely of backcrosses (F_1 hybrids X parental types), F_2-generation, and unassigned (>F_2) hybrids (Fig. 2, Table S3). Specifically, the EAxGU analysis identified backcrosses with parental *T. c. major* and F_2 hybrids in the Florida panhandle and southern Alabama (Fig. 2A). Similarly, all hybrid-generation EAxTT individuals from Georgia were identified as backcrosses with both parental types, whereas South Carolina hybrids were backcrosses with *T. c. carolina* (Fig. 2B).

Second-generation hybrids and backcrosses with both parental types were evident among GUxTT (Fig. 2C). Mississippi contained individuals with all three hybrid genotype classes (F_2, B_1, and B_2), but with backcrosses to parental *T. m. triunguis* at the greatest frequency. Alabama and Florida GUxTT were only represented by *T. m. triunguis* backcrosses. Finally, *T. o. ornata* and *T. c. carolina* in Illinois displayed relatively few hybrid genotypes (5%) but all were F_1, in contrast to the southeastern analyses (Fig. 2D).

3.3. Selective signatures at transcriptomic loci

Using INTROGRESS, only SNPs with a high allelic frequency differential (δ>0.8) were retained (Andrés *et al.* 2013). One exception was the EAxGU analysis where δ>0.7 was applied because no loci at the higher threshold were recovered.

The INTROGRESS genomic cline analysis recovered three outlier mRNA loci for EAxGU, and five each for EAxTT and GUxTT, with thirteen total and nine unique outliers among the three pairwise comparisons (Table 1; Fig. 3). Clines were inconsistent in that some pairwise taxon comparisons displayed rapid transitions from P_1 to P_2, whereas others demonstrated patterns indicative of directional introgression.

For example, all three EAxGU outlier loci displayed an over-representation of EA alleles in the hybrid zone, concomitant with an under-representation of heterozygotes and GU alleles (Fig. 3A). The *SULT* locus was an extreme example, with excess EA alleles below a hybrid index of ~0.8 (=80% assignment to GU at diagnostic loci). This pattern was replicated to a lesser degree in *ZNF236*, whereas *TLR9* was more sigmoidal. These genotypic proportions, coupled with the non-sigmoidal cline shape in *SULT* and *ZNF236*, suggest that introgression may be driven by a directional shift towards homozygous P_1 genotypes. In contrast, a steep, sigmoidal cline represented the genomic trend among scaffold assembly (and putatively non-functional) loci (Fig. 18). Taken together, these results suggest underlying directional introgression facilitating exchange of EA alleles despite divergence being maintained at most loci.

Cline shape was inconsistent within the EAxTT hybrid zone (Fig. 3B). Three (of five) outlier loci (*SASH3*, *SYPL2*, and *TLR9*) were significantly under-represented with regards to heterozygotes. Their clines displayed steep slopes, suggesting rapid transition among parental genotypes. An additional locus (*CITED4*) reflected an overrepresentation of P_2 (TT), and a fifth (*FAM89B*) displayed three equally-represented genotypes. Of note, *FAM89B* did not differ significantly from neutral expectations following Bonferroni correction (*P*=0.036).

By contrast, neutral expectations were rejected in all five GUxTT clines (*P*=0; α=0.007), with four (i.e., *SASH3*, *SYPL2*, *ACAD11*, *FAM89B*) suggesting a pattern of restricted introgression (Fig. 3C). A fifth (i.e., *TMEM214*) displayed directional introgression, with the homozygous P_2 (TT) genotype being overrepresented. Both the GUxTT and EAxTT analyses showed a ubiquitous signal of steep clines in non-transcriptomic loci (Fig. S18).

A greater number of SNP outliers (N=81) were identified with BGC rather than INTROGRESS, likely reflecting the larger number of loci included in the former. All INTROGRESS outliers were also corroborated by BGC (Fig. 4; Table 1). For EAxGU (Fig. 4A), all three exhibited excess EA ancestry (negative α=excess P_1; positive α=excess P_2), but there were no cline rate (ß) outliers in either direction. In contrast, four out of five EAxTT loci (*SYPL2*, *SASH3*, *TLR9*, and *CITED4*) were positive ß outliers, indicating steep clines and thus restricted introgression, with *SYPL2* also being an α outlier with excess EA (P_1) ancestry (Fig. 4B). The fifth locus (*FAM89B*) was an α (not ß) outlier that favored EA alleles. Finally, GUxTT (Fig. 4C) included two loci that were both α and ß outliers with steep clines and excess GU ancestry (*SYPL2* and *FAM89B*). Two others were ß-only outliers with steep clines (*SASH3* and *ACAD11*), and one (*TMEM214*) an α-only outlier with excess TT ancestry.

3.4. Environmental correlations with outliers

Shapiro-Wilks tests confirmed normality for all layers, following *OrderNorm* transformation. Forward selection retained ten uncorrelated and predictive layers (Table 2), and pRDAs revealed 49.2% environmental and 40.9% spatial contributions to model inertia. After controlling for spatial autocorrelation, the pRDA ANOVA identified four significant axes (*P*<0.05) explaining 21.5%, 10.6%, 10.3%, and 9.6% of the variance (cumulative 52.1%). The pRDA also identified 56 annotated outlier SNPs correlated with environmental predictors (Fig. 5). Twenty-eight pRDA outliers overlapped with INTROGRESS and/or BGC analyses (Fig. 6). Many of the overlapping SNPs, including all nine INTROGRESS loci that remained as outliers across all three analyses, were most strongly correlated with temperature variables rather than precipitation or wind speed (Table 3).

4. DISCUSSION

Our analyses characterized introgression in two North American box turtle hybrid zones (i.e., midwestern and southeastern North America). The midwestern hybrid zone showed no evidence of introgression, with hybrids restricted at low frequency to F_1, whereas southeastern hybridization was introgressive in nature, as evidenced by numerous backcrosses and F_2 individuals and a conspicuous lack of F_1 hybrids. Furthermore, contrasting intra- and inter-specific southeastern hybrid zones revealed they not only varied in the genealogical composition of their hybrids but also in the shapes and widths of locus-specific clines. We propose such a contrast provides insight into the evolutionary histories of the taxa involved and serves to delineate their appropriate taxonomic designations. Specifically, recent phylogenetic research indicates that *T. m. triunguis* is a separate species from *T. c. carolina* and *T. c. major* (Martin *et al.* 2013, 2014, 2020). The genomic clines herein are consistent with those phylogenetic results in that transcriptomic loci with steep clines are only found in inter-specific comparisons (i.e., *T. mexicana versus T. carolina*). The candidate genes at these transcriptomic loci may be targets of selection and directional introgression, as they deviated from neutral expectations not only with respect to genome-wide ancestry (i.e., genomic clines) but also multivariate environmental associations. Below we consider the impact of these results on the evolutionary history, genomic architecture, and species boundaries of *Terrapene*.

4.1. Regional and taxon-specific perspectives

Several of our conclusions have implications for *Terrapene* systematics. First, our ADMIXTURE analyses substantiate the presence of discrete *T. c. major* populations in Florida and Mississippi, with the Alabama and Apalachicola river drainages potentially serving as biogeographic barriers (Fig. S8). This consideration was markedly absent in previous morphological analyses (Butler *et al.* 2011) that concluded *T. c. major* merely represented an area of admixture between other *Terrapene* in the region. We did indeed detect considerable admixture between *T. c. major*, *T. c. carolina*, and *T. m. triunguis*, but the presence of two apparently non-admixed *T. c. major* populations demonstrated the existence of cryptic genetic variation (Douglas *et al.* 2009). We interpret these populations as representing distinct evolutionary significant units (ESUs) or (at worst) management units (MUs).

Second, admixture is apparent among all southeastern *Terrapene* except *T. bauri*, which absorbed the greatest amount of DAPC variation. This is likely attributed to Pliocene vicariance in the Florida Peninsula, where the Okefenokee Trough divided northern from southern Florida (Bert 1986; Douglas *et al.* 2009). Each of these aspects will require careful consideration when conservation efforts are planned or implemented, particularly given that *Terrapene* are in decline throughout their range (Dodd 2001).

On the other hand, *T. ornata* and *T. carolina* are separated by greater genetic distances than are the southeastern taxa (Martin *et al.* 2013), which may suggest the presence of intrinsic genetic incompatibilities (Barton 2001; Abbott *et al.* 2013) and is consistent with the lack of hybrids beyond the F_1 generation. Furthermore, the low frequency of F_1 hybrids observed in the Illinois ONxEA population may have resulted from recent degradation of the prairie grassland habitat (Manning 2001; Mussmann *et al.* 2017), which subsequently initiated increased heterospecific contact or otherwise disturbed reproductive boundaries by altering the fitness consequences of hybridization (Chafin et al. 2019; Grabenstein and Taylor 2018). However, this hypothesis cannot be explicitly tested herein. Although we did not identify contemporary back-crossed individuals, Cureton *et al.* (2011) did depict potential introgressive hybridization between ONxEA based on mitochondrial DNA and multiple microsatellite markers. These discrepancies may reflect either historical admixture as a source of introgression, which we did not explicitly test for, or back-crossed hybrids as a rarity not encountered in our sampling.

4.2. Biogeography and hybrid zone formation in *Terrapene*

The disparity in early and late-generation hybrids between the midwestern (ONxEA) and southeastern (EAxGU, GUxTT, and EAxTT) hybrid zones suggests differences in the underlying evolutionary processes. Such differences could involve regional variability in the extent or nature of reproductive isolation, or simply their respective biogeographic histories. Pleistocene glaciation precipitated numerous widespread distributional shifts across many taxa in the Midwest (King 1981; Webb 1981). Subsequent postglacial "shuffling" has also been implicated in as an historical driver of introgressive hybridization in populations of *Sistrurus* rattlesnakes now allopatric in the same region (Sovic *et al.* 2016). The same process could explain evidence for past introgression in the ONxEA hybrid zone (Cureton *et al* 2011b), despite a lack of contemporary hybrids. Here, a rapid re-colonization [especially from northern refugia such as the "Driftless Area" of Wisconsin, Iowa, and Illinois during the last glacial maximum (Holliday *et al.* 2002)] resulted in north-south contact zones near the glacial maximum (i.e., "leading-edge" hypothesis; Hewitt 1996, 2000). However, our purported range overlap also represents a broad interdigitation of "prairie" and "interior highland" habitats (Ennen *et al.* 2017), indicating that later-generation hybrids simply requires a finer-scale sampling than ours to be detected.

The southeastern United States has long been known for the of co-occurrence of contact zones (Remington 1968; Avise 2000), with migration from refugia in southern Florida and east

Texas/ west Louisiana as an hypothesized mechanism (Swenson & Howard 2005). The divergence of these lineages prior to Pleistocene glaciation (Martin *et al.* 2013) seemingly indicates postglacial expansion as a potential mechanism underlying southeastern hybrid zone formation. Likewise, finer-scale phylogeographic breaks that corroborate those in *Terrapene* have also been detected in numerous other turtles (Walker & Avise 1998). For example, *Sternotherus minor* and *S. odoratus* show deep east-west phylogeographic breaks approximately centered on Alabama, with unique lineages in peninsular Florida (Iverson 1977; Walker & Avise 1998). *Kinosternon subrubrum* mirrors this pattern, but with an additional unique lineage in the panhandle region (Walker *et al.* 1998). Similar breaks again appear in *Trachemys scripta*, *Macroclemys temminckii* (Walker & Avise 1998; Roman *et al.* 1999), and *Gopherus polyphemus* (Lamb *et al.* 1989). It is thus no surprise for the region to be identified as a hotspot for inter-specific contact and phylogeographic concordance (Soltis *et al.* 2006; Rissler & Smith 2010), also reflected in *Terrapene*. As always, it becomes inherently difficult to separate the relative importance of historical processes from contemporary physiographic or ecological features.

4.3. Functional genomic architecture in the southeastern hybrid zone

Regardless of the biogeographic scenarios invoked, a clear hotspot for hybridization exists in the southeastern United States. The variance among locus-specific patterns of differentiation or exchange allows for the potential interpretation of adaptive processes, at least in the context of neutral expectations. SNPs located in several mRNA loci are implicated as potentially contributing to selection in *Terrapene* from three southeastern hybrid zones. Among inter-specific comparisons (Figs. 3B, 3C, 4B, 4C), the dominant pattern was a steep, sigmoidal cline (GUxTT and EAxTT, but most clearly apparent in the latter). This accordingly points to selection against interspecific heterozygotes (Fitzpatrick 2013). In contrast, the selective advantage of EA alleles in the EAxGU hybrid zone (Figs. 3A, 4A) fails to agree with the general genome-wide pattern of underdominance (Fig. S18), which suggests directional introgression within which EA alleles are favored in hybrids under contemporary conditions. Mapping BGC outliers to *Trachemys* chromosomes (Fig. 4) also revealed their ubiquitous rather than finely concentrated distribution across the genome, and this in turn highlighted the differential nature of the observed interspecific introgression. Inaccuracies regarding mapping a *Terrapene* assembly against a *Trachemys* reference genome is also a possibility, although we attempted to minimize this by resolving conflicts via PAFSCAFF.

Loci with steep genomic clines were most strongly correlated with temperature predictors, suggesting the importance of thermal adaptations in maintaining species boundaries between southeastern *Terrapene*. In contrast, neither precipitation nor wind-associated outliers followed this pattern. Given the positive relationship between outlier genes and thermal predictors, a natural extrapolation would be that a thermal gradient drives differential introgression. This has multiple implications regarding the integrity of species boundaries if *Terrapene* are subjected to future environmental changes. In one scenario, a shifting adaptive landscape may promote hybridization by contravening long-term reproductive isolation (EAxGU), with subsequent introgression at specific loci (as herein). Alternatively, rapid environmental change could simply outpace the selective filtering of maladaptive variants, with a subsequent decrease in fitness (Kokko *et al.* 2017). This would be particularly evident when effective population sizes are already depressed following a population bottleneck (Chafin *et al.* 2019). Here, extreme rates of change may also link with a genetic swamping effect (i.e., replacement of local genotypes by hybrids; Todesco *et al.* 2016). Both scenarios implicate anthropogenic pressures as governing the fates of diverse taxa across hybrid zones (Taylor *et al.* 2015).

The putative functions of the nine outlier loci supported by genomic clines and RDA provide additional support for the strong impact of thermal selection. Two loci potentially relate to TSD during embryonic development, while others seemingly associate with molecular pathways in skeletal muscle and nervous tissues that involve tolerance to anoxia and hypoxia (N=6), and immune response to pathogens [(N=2); see Table 1 for sources]. Anoxia/ hypoxia-related genes have been associated with freeze tolerance in hibernating turtles (Storey 2006), thus supporting an obvious association with thermal gradients. Here, three loci (*SYPL2, ACAD11*, and *TMEM214*) may regulate brain function and metabolism by up-regulating Ca^{2+} concentrations (Takeshima *et al.* 1998; Pamenter *et al.* 2016), inducing lipid metabolism (He *et al.* 2011; Gomez & Richards 2018), and initiating stress-induced apoptosis (Kesaraju *et al.* 2009; Li *et al.* 2013). Similarly, the *CITED4* gene (EAxTT) potentially inhibits hypoxia-related transcription factors (Fox *et al.* 2004), whereas *FAM89B* (GUxTT) may become up-regulated when physiological conditions turn hypoxic (Goyal & Longo 2014). The regulation of immune function is less clearly associated with underlying thermal gradients, but may associate instead with behavioral thermoregulation during infection, given that infection resistance increases at warmer temperatures (Dodd 2001; Agha *et al.* 2017). However, we remain cautious in that these genes have not been associated with specific functions in *Terrapene,* and to do so remains speculative. Nevertheless, their potential connection with thermal adaptations is consistent with the RDA.

We emphasize that ectothermic vertebrates are exceptionally vulnerable to contemporary pressures, and reflect an elevated extinction risk due to a strong reliance on environmental thermoregulation and a dependence on suitable habitat (Gibbons *et al.* 2000; Sinervo *et al.* 2010; Winter *et al.* 2016). Indeed, ectotherms can exhibit reduced fitness and growth-rates when environmental conditions exceed their thermal optima (Deutsch *et al.* 2008; Martin & Huey 2008; McCallum *et al.* 2009; Huey *et al.* 2012; Huey & Kingsolver 2019). Increased temperatures also impact physiological pathways, and these are putatively comparable with the genes described herein, such as increasing metabolic rates (Dillon *et al.* 2010), intensifying hypoxic stress despite higher temperature-driven O_2 demands (Huey & Ward 2005), heightening disease transmission (Pounds *et al.* 2006), or even the over-extension of thermal tolerances (Sinervo *et al.* 2010; Ceia-Hasse *et al.* 2014). Going forward, climate change may facilitate evolutionary responses to changing thermal conditions, potentially including local adaptation (Holt 1990; Norberg *et al.* 2012; Bush *et al.* 2016), physiological and behavioral mechanisms (e.g., thermoregulation, phenology), plasticity (Urban *et al.* 2014; Sgrò *et al.* 2016), of shifts in species distributions (Parmesan *et al.* 1999; Parmesan & Yohe 2003; Moreno-Rueda *et al.* 2012). Accordingly, ectothermic species boundaries may be particularly susceptible if indeed governed by thermal conditions, as seemingly exemplified herein with *Terrapene*.

4.4. Conclusions

Our study suggests that reproductive isolation in turtles involves numerous mechanisms regulated by thermally induced selective pressures. Clearly, such selective pressures play a prominent role in chelonian ecology. Many turtle species are at elevated risks from climate change due to the imposition of TSD during embryonic development, as well as prolonged generation times. Similarly, long generation times can also restrict the adaptive capacity of turtles in a rapidly changing climate (Hoffmann *et al.* 2017). Here, climate change can shift sexratios and promote demographic collapse, with warmer temperatures initiating male bias in biological sex, and vice versa (Janzen 1994).

Two important evolutionary implications are evident in our data. First, we demonstrated differential introgression along an ecological gradient in three taxa inhabiting a North American hybrid zone. We then assessed scaffold-aligned and transcriptomic SNPs to identify several genes whose functions are consistent with physiological processes related to thermal ecology, and as such are capable of promoting adaptive divergence. In this sense, they potentially describe ecological gradients related to TSD, anoxia/ hypoxia tolerance, and immune response in a hybrid zone encompassing three taxa. While we acknowledge that the observed clinal patterns could represent "molecular spandrels" reflecting an underlying neutral process, such as isolation-by-distance, we sought environmental associations by actively controlling for spatial autocorrelation as a corroboration of clinal loci with a putative adaptive role (Vasemägi 2006; Barrett & Hoekstra 2011). We also underscore specific loci displaying genomic cline patterns consistent with directional introgression and selection, and these loci are potentially sustaining divergence across species.

Second, we characterized a southeastern North American hybrid zone representing a variety of biodiversity elements as being susceptible to anthropogenic and environmental changes (Remington 1968; Swenson & Howard 2005; Rissler & Smith 2010). Our results demonstrate that NGS population genomic methods can clearly identify population structure and detect introgression, whereas traditional Sanger sequencing methods are inadequate to do so (Butler *et al.* 2011; Martin *et al.* 2013). We also underscore specific loci prone to directional introgression and selection that may potentially sustain divergence across species.

ACKNOWLEDGEMENTS

The research herein was conducted in partial fulfillment of the Ph.D. degree in Biological Sciences at the University of Arkansas (BTM). We would like to extend our gratitude to the countless volunteers and organizations who collected and/or provided tissue samples (Table S1). We also thank the current and former members of the Douglas Lab and University of Arkansas faculty who provided advice and support, especially A. Alverson, W. Anthonysamy, M. Bangs, J. Koukl, S. Mussmann, J. Pummill, and Z. Zbinden. Finally, many thanks to the reviewers, whose comments led to a greatly improved manuscript. Sample collections were approved under the University of Texas-Tyler Animal Care and Use Committee (IACUC) permit #113 and University of Illinois IACUC protocols 16160 and 18000. Funding was provided by the Lucille F. Stickle Fund of the North American Box Turtle Committee, the American Turtle Observatory (ATO), and the following endowments: The Bruker Professorship in Life Sciences (MRD) and the Twenty-First Century Chair in Global Change Biology (MED). Analytical resources were provided by the Arkansas High Performance Computing Cluster (AHPCC) and an NSF-XSEDE Research Allocation (TG-BIO160065) that allowed access the Jetstream cloud service.

DATA ACCESSIBILITY

The demultiplexed ddRADseq reads are deposited as FASTQ files to NCBI's sequence read archive (https://www.ncbi.nlm.nih.gov/sra); Accessions: SAMN12668545-SAMN12668981 (BioProject ID: PRJNA563121). The R scripts, metadata, and input files for each analysis are available from a Dryad Digital Repository (https://doi.org/10.5061/dryad.brv15dv7k).

5. REFERENCES

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, and Buggs R (2013) Hybridization and Speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Agha M, Price S, Nowakowski A, Augustine B, and Todd B (2017) Mass mortality of eastern box turtles with upper respiratory disease following atypical cold weather. *Diseases of Aquatic Organisms*, **124**, 91–100.

Alexander DH, Novembre J, and Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

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

Andrés JA, Larson EL, Bogdanowicz SM, and Harrison RG (2013) Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. *Genetics*, **193**, 501–13.

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.

Avise JC (2000) *Phylogeography*. Harvard University Press, Cambridge, MA.

Babik W, Dudek K, Fijarczyk A, Pabijan M, Stuglik M, Szkotak R, and Zieliński P (2015) Constraint and Adaptation in newt Toll-Like Receptor Genes. *Genome Biology and Evolution*, **7**, 81–95.

Barrett RDH and Hoekstra HE (2011) Molecular spandrels: Tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

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

Bert T (1986) Speciation in western Atlantic stone crabs (genus *Menippe*): the role of geological processes and climatic events in the formation and distribution of species. *Marine Biology*, **93**, 157–170.

Borcard D, Gillet F, and Legendre P (2018) Spatial analysis of ecological data. In: *Numerical ecology with R*, pp. 299–367. Springer.

Bush A, Mokany K, Catullo R, Hoffmann A, Kellermann V, Sgrò C, McEvey S, and Ferrier S (2016) Incorporating evolutionary adaptation in species distribution modelling reduces projected vulnerability to climate change. *Ecology Letters*, **19**, 1468–1478.

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.

Caye K, Deist TM, Martins H, Michel O, and François O (2016) TESS3: Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, **16**, 540–548.

Ceia-Hasse A, Sinervo B, Vicente L, and Pereira HM (2014) Integrating ecophysiological models into species distribution projections of European reptile range shifts in response to climate change. *Ecography*, **37**, 679–688.

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.

Cureton JC, Buchman AB, Deaton R, and Lutterschmidt WI (2011) Molecular analysis of hybridization between the box turtles *Terrapene carolina* and *T. ornata*. *Copeia*, **2011**, 270–277.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, and Group 1000 Genomes Project Analysis (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Deutsch CA, Tewksbury JJ, Huey RB, Sheldon KS, Ghalambor CK, Haak DC, and Martin PR (2008) Impacts of climate warming on terrestrial ectotherms across latitude. *Proceedings of the National Academy of Sciences*, **105**, 6668–6672.

Dillon ME, Wang G, and Huey RB (2010) Global metabolic impacts of recent climate warming. *Nature*, **467**, 704–706.

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

Douglas ME, Douglas MR, Schuett GW, and Porras LW (2009) Climate change and evolution of the New World pitviper genus *Agkistrodon* (Viperidae). *Journal of Biogeography*, **36**, 1164–1180.

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

Ennen JR, Matamoros WA, Agha M, Lovich JE, Sweat SC, and Hoagstrom CW (2017) Hierarchical, quantitative biogeographic provinces for all North American turtles and their contribution to the biogeography of turtles and the continent. *Herpetological Monographs*, **31**, 114–140.

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

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.

Fick SE and Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, **37**, 4302–4315.

Fitzpatrick BM (2013) Alternative forms for genomic clines. *Ecology and Evolution*, **3**, 1951–1966.

Fox SB, Braganca J, Turley H, Campo L, Han C, Gatter KC, Bhattacharya S, and Harris AL (2004) CITED4 inhibits hypoxia-activated transcription in cancer cells, and its cytoplasmic location in breast cancer is associated with elevated expression of tumor cell hypoxia-inducible factor 1α. *Cancer Research*, **64**, 6075–6081.

Garroway CJ, Bowman J, Cascaden TJ, Holloway GL, Mahan CG, Malcolm JR, Steele MA, Turner G, and Wilson PJ (2010) Climate change induced hybridization in flying squirrels. *Global Change Biology*, **16**, 113–121.

Gibbons JW, Scott DE, Ryan TJ, Buhlmann KA, Tuberville TD, Metts BS, Greene JL, Mills T, Leiden Y, Poppy S, and Winne CT (2000) The global decline of reptiles, déjà vu amphibians: reptile species are declining on a global scale. Six significant threats to reptile populations are habitat loss and degradation, introduced invasive species, environmental pollution, disease, unsustaina. *Bioscience*, **50**, 653–666.

Gomez CR and Richards JG (2018) Mitochondrial responses to anoxia exposure in red eared sliders (*Trachemys scripta*). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, **224**, 71–78.

Gompert Z and Buerkle CA (2010) INTROGRESS: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, **10**, 378–384.

Gompert Z and Buerkle CA (2011) Bayesian estimation of genomic clines. *Molecular Ecology*, **20**, 2111–2127.

Gompert Z and Buerkle CA (2012) BGC: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, **12**, 1168–1176.

Goyal R and Longo LD (2014) Acclimatization to long-term hypoxia: gene expression in ovine carotid arteries. *Physiological Genomics*, **46**, 725–734.

Grabenstein KC and Taylor SA (2018) Breaking barriers: causes, consequences, and experimental utility of human-mediated hybridization. *Trends in Ecology and Evolution*, **33**, 198–212.

Guo B, Lu D, Liao WB, and Merilä J (2016) Genomewide scan for adaptive differentiation along altitudinal gradient in the Andrew's toad *Bufo andrewsi*. *Molecular Ecology*, **25**, 3884–3900.

Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, and Chen J (2020) RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the ideograms. *PeerJ Computer Science*, **6**, e251.

He M, Pei Z, Mohsen A-W, Watkins P, Murdoch G, Van Veldhoven PP, Ensenauer R, and Vockley J (2011) Identification and characterization of new long chain acyl-CoA dehydrogenases. *Molecular Genetics and Metabolism*, **102**, 418–429.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnaean Society*, **58**, 247–276.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hijmans RJ and Van Etten J (2016) raster: Geographic Data Analysis and Modeling. R package version 2.5-8.

Hoffmann AA, Sgrò CM, and Kristensen TN (2017) Revisiting adaptive potential, population size, and conservation. *Trends in Ecology and Evolution*, **32**, 506–517.

Holliday VT, Knox JC, Running I V., Mandel RD, and Ferring CR (2002) The central lowlands. In: *The physical geography of North America* (ed Orme AR), pp. 335–362. Oxford University Press, New York.

Holt RD (1990) The microevolutionary consequences of climate change. *Trends in Ecology and Evolution*, **5**, 311–315.

Huey RB, Kearney MR, Krockenberger A, Holtum JAM, Jess M, and Williams SE (2012) Predicting organismal vulnerability to climate warming: roles of behaviour, physiology and adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 1665–1679.

Huey RB and Kingsolver JG (2019) Climate warming, resource availability, and the metabolic meltdown of ectotherms. *The American Naturalist*, **194**, E140–E150.

Huey RB and Ward PD (2005) Hypoxia, global warming, and terrestrial late Permian extinctions. *Science*, **308**, 398–401.

Iverson JB (1977) Geographic variation in the musk turtle, *Sternotherus minor*. *Copeia*, **1977**, 502.

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.

Janzen FJ (1994) Climate change and temperature-dependent sex determination in reptiles. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 7487–90.

Jay F, Manel S, Alvarez N, Durand EY, Thuiller W, Holderegger R, Taberlet P, and François O (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, **21**, 2354–2368.

Johnson W (1994) Woodland expansion in the Platte River, Nebraska: patterns and causes. *Ecological Monographs*, **64**, 45–84.

Jombart T, Devillard S, and Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Kearns AM, Restani M, Szabo I, Schrøder-Nielsen A, Kim JA, Richardson HM, Marzluff JM, Fleischer RC, Johnsen A, and Omland KE (2018) Genomic evidence of speciation reversal in ravens. *Nature Communications*, **9**, 906.

Keller I and Seehausen O (2012) Thermal adaptation and ecological speciation. *Molecular Ecology*, **21**, 782–799.

Kesaraju S, Schmidt-Kastner R, Prentice HM, and Milton SL (2009) Modulation of stress proteins and apoptotic regulators in the anoxia tolerant turtle brain. *Journal of Neurochemistry*, **109**, 1413–1426.

Key KHL (1968) The concept of stasipatric speciation. *Systematic Biology*, **17**, 14–22.

King JE (1981) Late Quaternary vegetational history of Illinois. *Ecological Monographs*, **51**, 43–62.

Kingsolver JG (2009) The well-temperatured biologist. (American Society of Naturalists Presidential Address). *The American Naturalist*, **174**, 755–68.

Kokko H, Chaturvedi A, Croll D, Fischer MC, Guillaume F, Karrenberg S, Kerr B, Rolshausen G, and Stapley J (2017) Can evolution supply what ecology demands? *Trends in Ecology and Evolution*, **32**, 187–197.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, and Mayrose I (2015) CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across *K*. *Molecular Ecology Resources*, **15**, 1179–1191.

Kumaresan V, Pasupuleti M, Arasu MV, Al-Dhabi NA, Arshad A, Amin SMN, Yusoff FM, and Arockiaraj J (2018) A comparative transcriptome approach for identification of molecular changes in Aphanomyces invadans infected Channa striatus. *Molecular Biology Reports*, **45**, 2511–2523.

Lamb T, Avise JC, and Gibbons JW (1989) Phylogeographic patterns in mitochondrial-dna of the desert tortoise (Xerobates-agassizi), and evolutionary relationships among the north-american gopher tortoises. *Evolution*, **43**, 76–87.

van Lear DH and Harlow RF (2002) Fire in the eastern United States: influence on wildlife habitat. In: *Proceedings: the role of fire for nongame wildlife management and community restoration: traditional uses and new directions. General Technical Report 288* (eds Ford W., Russell KR and Moorman CE), pp. 2–10. US Dept. of Agriculture, Forest Service, Northeastern Research Station.

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li YF, Costello JC, Holloway AK, and Hahn MW (2008) “Reverse ecology” and the power of population genomics. *Evolution*, **62**, 2984–2994.

Li C, Wei J, Li Y, He X, Zhou Q, Yan J, Zhang J, Liu Y, Liu Y, and Shu H-B (2013) Transmembrane Protein 214 (TMEM214) mediates endoplasmic reticulum stress-induced caspase 4 enzyme activation and apoptosis. *The Journal of Biological Chemistry*, **288**, 17908–17.

Linck EB and Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. *Molecular Ecology Resources*, **19**, 639–647.

Lydeard C and Mayden RL (1995) A diverse and endangered aquatic ecosystem of the southeast United States. *Conservation Biology,* **9**, 800–805.

Manning B (2001) *Critical Trends in Illinois Ecosystems*. Illinois Department of Natural Resources. Springfield, IL.

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, Chafin TK, Douglas MR, Placyk JS, Birkhead RD, Phillips CA, and Douglas ME (2020) Machine learning substantiates biologically meaningful species delimitations in the phylogenetically complex North American box turtle genus *Terrapene*. *bioRxiv, doi: https://doi.org/10.1101/2020.05.19.103598*.

Martin TL and Huey RB (2008) Why suboptimal is optimal: Jensen’s inequality and ectotherm thermal preferences. *The American Naturalist*, **171**, E102-18.

McCallum M, McCallum JL, and Trauth SE (2009) Predicted climate change may spark box turtle declines. *Amphibia-Replilia*, **30**, 259–264.

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.

Moreno-Rueda G, Pleguezuelos JM, Pizarro M, and Montori A (2012) Northward shifts of the distributions of Spanish reptiles in association with climate change. *Conservation Biology*, **26**, 278–283.

Muhlfeld CC, Kovach RP, Jones LA, Al-Chokhachy R, Boyer MC, Leary RF, Lowe WH, Luikart G, and Allendorf FW (2014) Invasive hybridization in a threatened species is accelerated by climate change. *Nature Climate Change*, **4**, 620–624.

Mussmann SM, Douglas MR, Anthonysamy WJB, Davis MA, Simpson SA, Louis W, and Douglas ME (2017) Genetic rescue, the greater prairie chicken and the problem of conservation reliance in the Anthropocene. *Royal Society Open Science*, **4**, 160736.

Mussmann SM, Douglas MR, Chafin TK, and Douglas ME (2020) AdmixPipe: population analyses in Admixture for non-model organisms. *BMC Bioinformatics*, **21**, 1–9.

Norberg J, Urban MC, Vellend M, Klausmeier CA, and Loeuille N (2012) Eco-evolutionary responses of biodiversity to climate change. *Nature Climate Change*, **2**, 747–751.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, and Wagner H (2019) vegan: Community Ecology Package.

Paitz RT and Bowden RM (2008) A proposed role of the sulfotransferase/sulfatase pathway in modulating yolk steroid effects. *Integrative and Comparative Biology*, **48**, 419–427.

Pamenter ME, Gomez CR, Richards JG, and Milsom WK (2016) Mitochondrial responses to prolonged anoxia in brain of red-eared slider turtles. *Biology Letters*, **12**, 20150797.

Parmesan C, Ryrholm N, Stefanescu C, Hill JK, Thomas CD, Descimon H, Huntley B, Kaila L, Kullberg J, Tammaru T, Tennent WJ, Thomas JA, and Warren M (1999) Poleward shifts in geographical ranges of butterfly species associated with regional warming. *Nature*, **399**, 579–583.

Parmesan C and Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37–42.

Payseur BA (2010) Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources*, **10**, 806–820.

Peterson RA and Cavanaugh JE (2019) Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, **2019**, 1–16.

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.

Pounds AJ, Bustamante MR, Coloma LA, Consuegra JA, Fogden MPL, Foster PN, La Marca E, Masters KL, Merino-Viteri A, Puschendorf R, Ron SR, Sánchez-Azofeifa GA, Still CJ, and Young BE (2006) Widespread amphibian extinctions from epidemic disease driven by global warming. *Nature*, **439**, 161–167.

Qin D, Plattner G, Tignor M, Allen S, Boschung J, Nauels A, Xia Y, Bex V, and Midgley P (2013) *Summary for policymakers. Climate change 2013: the physical science basis. Contribution of Working Group I to the fifth assessment report of the Intergovernmental Panel on Climate Change, eds Stocker, TF et al.* Cambridge University Press, Cambridge, UK.

Remington CL (1968) Suture-zones of hybrid interaction between recently joined biotas. In: *Evolutionary Biology* (ed Dobzhansky T), pp. 321–428. Springer, New York, NY, USA.

Rhymer JM and Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Ricketts TH (1999) *Terrestrial ecoregions of North America: a conservation assessment*. Island Press, Washington, DC, USA.

Rissler LJ and Smith WH (2010) Mapping amphibian contact zones and phylogeographical break hotspots across the United States. *Molecular Ecology*, **19**, 5404–5416.

Rödin-Mörch P, Luquet E, Meyer-Lucht Y, Richter-Boix A, Höglund J, and Laurila A (2019) Latitudinal divergence in a wide-spread amphibian: contrasting patterns of neutral and adaptive genomic variation. *Molecular Ecology*, **28**, 2996–3011.

Roman J, Santhuff SD, Moler PE, and Bowen BW (1999) Population structure and cryptic evolutionary units in the alligator snapping turtle. *Conservation Biology*, **13**, 135–142.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Rosenzweig C, Karoly D, Vicarelli M, Neofotis P, Wu Q, Casassa G, Menzel A, Root TL, Estrella N, Seguin B, Tryjanowski P, Liu C, Rawlins S, and Imeson A (2008) Attributing physical and biological impacts to anthropogenic climate change. *Nature*, **453**, 353–357.

Ryan SF, Deines JM, Scriber JM, Pfrender ME, Jones SE, Emrich SJ, and Hellmann JJ (2018) Climate-mediated hybrid zone movement revealed with genomics, museum collection, and simulation modeling. *Proceedings of the National Academy of Sciences*, 2017–14950.

Samson F and Knopf F (1994) Prairie conservation in North America. *Bioscience*, **44**, 418–421.

Samson FB, Knopf FL, and Ostlie WR (2004) Great Plains ecosystems: past, present, and future. *Wildlife Society Bulletin*, **32**, 6–15.

Seehausen O, Takimoto G, Roy D, and Jokela J (2008) Speciation reversal and biodiversity dynamics with hybridization in changing environments. *Molecular Ecology*, **17**, 30–44.

Sgrò CM, Terblanche JS, and Hoffmann AA (2016) What Can Plasticity Contribute to Insect Responses to Climate Change? *Annual Review of Entomology*, **61**, 433–451.

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.

Simison W, Parham J, Papenfuss T, Lam A, and Henderson J (2020) Annotated chromosome-level reference genome of the red-eared slider turtle (*Trachemys scripta elegans*). *Genome Biology and Evolution*, **12**, 456–462.

Sinervo B, Méndez-de-la-Cruz F, Miles DB, Heulin B, Bastiaans E, Villagrán-Santa Cruz M, Lara-Resendiz R, Martínez-Méndez N, Calderón-Espinosa ML, Meza-Lázaro RN, Gadsden H, Avila LJ, Morando M, De la Riva IJ, Victoriano Sepulveda P, Rocha CFD, Ibargüengoytía N, Aguilar Puntriano C, Massot M *et al.* (2010) Erosion of lizard diversity by climate change and altered thermal niches. *Science*, **328**, 894–9.

Soltis DE, Morris AB, McLachlan JS, Manos PS, and Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Sovic MG, Fries AC, and Gibbs HL (2016) Origin of a cryptic lineage in a threatened reptile through isolation and historical hybridization. *Heredity*, **117**, 358–366.

Stapanian MA, Cassell DL, and Cline SP (1997) Regional patterns of local diversity of trees: associations with anthropogenic disturbance. *Forest Ecololgy and Management*, **93**, 33–44.

Stapanian MA, Sundberg SD, Baumgardner GA, and Liston A (1998) Alien plant species composition and associations with anthropogenic disturbance in North American forests. *Plant Ecology*, **139**, 49–62.

Storey KB (2006) Reptile freeze tolerance: metabolism and gene expression. *Cryobiology*, **52**, 1–16.

Swenson NG and Howard DJ (2004) Do suture zones exist? *Evolution*, **58**, 2391–2397.

Swenson NG and Howard DJ (2005) Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American Naturalist*, **166**, 581–591.

Takeshima H, Shimuta M, Komazaki S, Ohmi K, Nishi M, Iino M, Miyata A, and Kangawa K (1998) Mitsugumin29, a novel synaptophysin family member from the triad junction in skeletal muscle. *The Biochemical Journal*, **331**, 317–22.

Taylor SA, Larson EL, and Harrison RG (2015) Hybrid zones: windows on climate change. *Trends in Ecology and Evolution*, **30**, 398–406.

Taylor SA, White TA, Hochachka WM, Ferretti V, Curry RL, and Lovette I (2014) Climate-mediated movement of an avian hybrid zone. *Current Biology*, **24**, 671–676.

Teske PR, Sandoval-Castillo J, Golla TR, Emami-Khoyi A, Tine M, von der Heyden S, and Beheregaray LB (2019) Thermal selection as a driver of marine ecological speciation. *Proceedings of the Royal Society B: Biological Sciences*, **286**, 20182023.

Tiffin P and Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology and Evolution*, **29**, 673–680.

Todesco M, Pascual MA, Owens GL, Ostevik KL, Moyers BT, Hübner S, Heredia SM, Hahn MA, Caseys C, Bock DG, and Rieseberg LH (2016) Hybridization and extinction. *Evolutionary Applications*, **9**, 892–908.

Urban MC, Richardson JL, and Freidenfelds NA (2014) Plasticity and genetic adaptation mediate amphibian and reptile responses to climate change. *Evolutionary Applications*, **7**, 88–103.

Vasemägi A (2006) The adaptive hypothesis of clinal variation revisited: Single-locus clines as a result of spatially restricted gene flow. *Genetics*, **173**, 2411–2414.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, **106**, 9939–9946.

Walker DE and Avise JC (1998) Principles of phylogeography as illustrated by freshwater and terrestrial turtles in the southeastern United States. *Annual Review of Ecology and Systematics*, **29**, 23–58.

Walker DE, Moler PE, Buhlmann KA, and Avise JC (1998) Phylogeographic patterns in *Kinosternon subrubrum* and *K. baurii* based on mitochondrial DNA restriction analyses. *Herpetologica*, **54**, 174–184.

Waterhouse MD, Erb LP, Beever EA, and Russello MA (2018) Adaptive population divergence and directional gene flow across steep elevational gradients in a climate-sensitive mammal. *Molecular Ecology*, **27**, 2512–2528.

Webb T (1981) The past 11,000 years of vegetational change in eastern North America. *BioScience,* **31**, 501–506.

Winter M, Fiedler W, Hochachka WM, Koehncke A, Meiri S, and De la Riva I (2016) Patterns and biases in climate change research on amphibians and reptiles: a systematic review. *Royal Society Open Science*, **3**, 160158.

Wringe BF, Stanley RRE, Jeffery NW, Anderson EC, and Bradbury IR (2017a) HYBRIDDETECTIVE: a workflow and package to facilitate the detection of hybridization using genomic data in R. *Molecular Ecology Resources*, **17**, e275–e284.

Wringe BF, Stanley RRE, Jeffery NW, Anderson EC, and Bradbury IR (2017b) *parallelnewhybrid*: an R package for the parallelization of hybrid detection using NEWHYBRIDS. *Molecular Ecology Resources*, **17**, 91–95.

TABLES AND FIGURES

Table 1: Annotation information for overlapping Introgress and BGC (Bayesian Genomic Cline) outliers (Figs. 3, 4). Each outlier was derived from an alignment mapped to the *Terrapene* transcriptome, with three pairwise taxa combinations performed. The significance threshold (α_B) was determined using a Bonferroni correction for multiple tests. Bold Gene abbreviations with an asterisk (*) differ significantly from neutral expectations. EA=Woodland (*T. c. carolina*), GU=Gulf Coast (*T. c. major*), TT=Three-toed (*T. m. triunguis*).

Gene Abbr. (P-value)	**BGC (α/β)**	**Outlier (α/β)‡**	**Full Gene Name**	**Possible Function(s)**	**Source(s)**
EAxGU (α_B=0.01)					
SULT (P=0)*	-1.59/-0.22	-/N	Amine Sulfotransferase-like	Regulates yolk steroids during TSD†	(Paitz & Bowden 2008)
TLR9 (P=0)*	-0.74/1.14	-/N	Toll-like Receptor 9	Immune Response to Pathogens	(Babik *et al.* 2015)
ZNF236 (P=0)*	-0.94/0.14	-/N	Zinc Finger Protein 236	Unknown	
EAxTT (α_B=0.007)					
SASH3 (P=0)*	-0.49/1.35	N/+	SAM and SH3 Domain Containing 3	Immune response	(Kumaresan *et al.* 2018)
SYPL2 (P=0)*	-0.68/1.65	-/+	Synaptophysin-like Protein 2 (AKA Mitsugumin 29)	Maintenance of $[Ca^{2+}]$ during anoxia	(Takeshima *et al.* 1998; Pamenter *et al.* 2016)
FAM89B (P=0.036)	-0.59/0.23	-/N	Family with Sequence Similarity 89, member B	Upregulated in hypoxic conditions	(Goyal & Longo 2014)
CITED4 (P=0)*	0.26/0.97	N/+	Cbp/p300 Interacting Transactivator, Domain 4	Inhibits hypoxia-activated transcription	(Fox *et al.* 2004)
TLR9 (P=0)*	-0.38/1.63	N/+	Toll-like Receptor 9	Immune Response to Pathogens	(Babik *et al.* 2015)
GUxTT (α_B=0.007)					
SASH3 (P=0)*	-0.22/1.86	N/+	SAM and SH3 Domain Containing 3	Immune response	(Kumaresan *et al.* 2018)
SYPL2 (P=0)*	-0.48/2.23	-/+	Synaptophysin-like Protein 2	Maintenance of $[Ca^{2+}]$ during anoxia	(Takeshima *et al.* 1998; Pamenter *et al.* 2016)
FAM89B (P=0)*	-0.46/0.66	-/+	Family with Sequence Similarity 89, member B	Upregulated in hypoxic conditions	(Goyal & Longo 2014)
ACAD11 (P=0)*	-0.24/1.16	N/+	Acyl-CoA Dehydrogenase, family member 11-like	Lipid metabolism	(He *et al.* 2011; Gomez & Richards 2018)
TMEM214 (P=0)*	0.77/0.11	+/N	Transmembrane Protein 214	Stress-induced apoptosis during anoxia	(Kesaraju *et al.* 2009; Li *et al.* 2013)

†TSD=Temperature-dependent sex determination

‡α outliers=excess P_1 (-) or P_2 (+) ancestry; ß outliers=slow (-) or rapid (+) genomic cline rates; N=neutral

Table 2: WorldClim environmental predictor abbreviations for the redundancy analysis (RDA). The raster layers were obtained from https://worldclim.org. Additional variables were excluded via forward selection due to low predictive capacity or correlation with the remaining layers.

Abbr.	Full Name	BioClim No.
tmDR2	Mean diurnal range [Monthly (Max temp - Min temp)]	2
tS4	Temperature seasonality (standard deviation X 100)	4
mxtWM5	Max temperature of warmest month	5
mintCM6	Min temperature coldest month	6
mtWQ8	Mean temperature of wettest quarter	8
pAM12	Annual precipitation	12
pDM14	Precipitation of driest month	14
pWQ16	Precipitation of wettest quarter	16
pCQ19	Precipitation of coldest quarter	19
Wind	Mean annual wind speed	N/A

Table 3: *Terrapene* ddRAD SNP outliers (N=118) most strongly associated with ten predictive and uncorrelated WorldClim variables (Table 2), collapsed into temperature, precipitation, or wind speed categories. Percentages include the total for only redundancy analysis (% RDA) and for all three outlier methods (% All): INTROGRESS, Bayesian genomic cline (BGC), and RDA.

Environment Type	**No. Outliers**	**% (RDA)**	**% (All)**
Temperature	44	78.6	34.7
Precipitation	10	17.9	8.5
Wind	2	3.6	1.7
Total	56	100.0	44.9

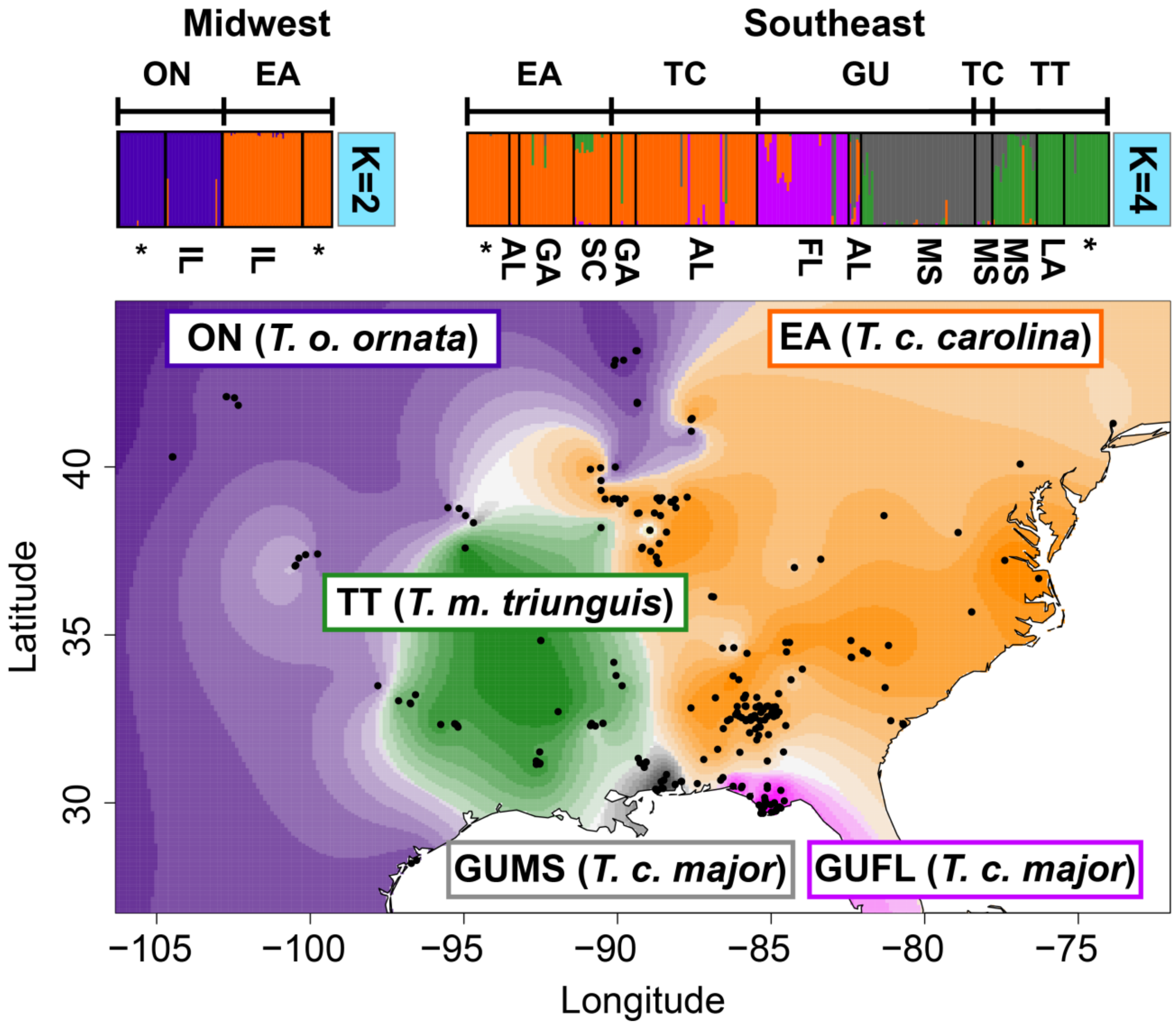


Figure 1: ADMIXTURE (barplots) and TESS3 (map) analyses for 11,308 *Terrapene* ddRADseq loci. All plots include individuals (N=320; black circles on map) from the Midwest and Southeast hybrid zones, with the optimal number of clusters (*K*) determined via cross-validation across 20 independent runs. Labels above ADMIXTURE plots indicate subspecies (when available) as identified in the field: EA=Woodland (*T. carolina carolina*), ON=Ornate (*T. o. ornata*), TC=*T. carolina* (subspecies identification unavailable), GU=Gulf Coast (*T. c. major*), TT=Three-toed (*T. mexicana triunguis*). Bottom labels show hybrid zone localities by U.S. state: IL=Illinois, AL=Alabama, GA=Georgia, FL=Florida, MS=Mississippi, LA=Louisiana. A * represents a group of "pure" individuals from multiple localities outside the hybrid zones. TESS3 ancestry coefficients (Q) are predicted across the spatial surface via Kriging interpolation (θ=10) and are color-coded with the ADMIXTURE plots. Lighter/ darker gradient shades depict lower/ higher Q.

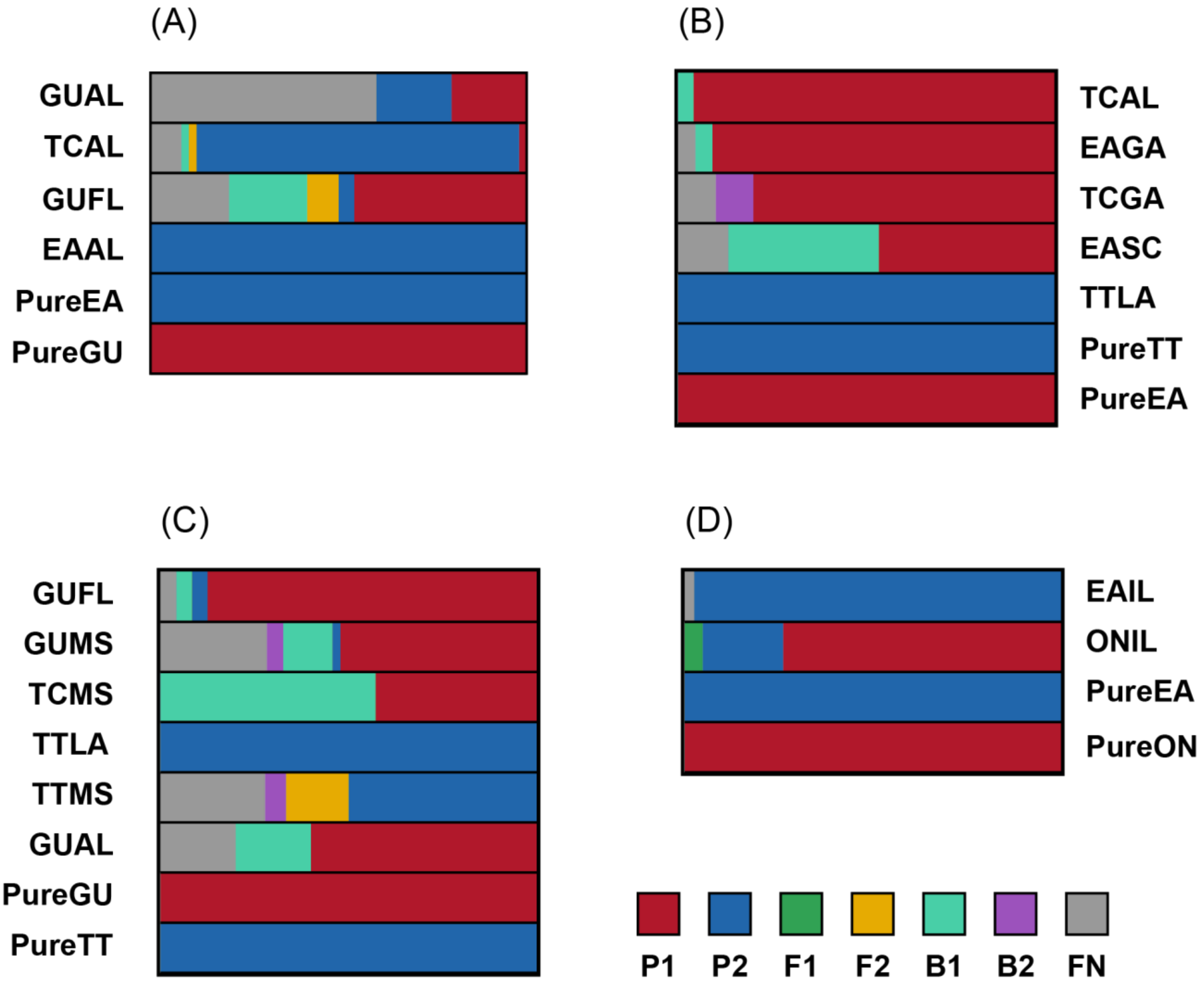


Figure 2: Population-level NEWHYBRIDS plots for four pairs of southeastern and midwestern *Terrapene* taxa. Individuals were collapsed into populations based on field identification at the subspecific level. The first two characters represent: GU=Gulf Coast, *T. c. major*; EA=Woodland, *T. c. carolina*; TT=Three-toed, *T. m. triunguis*; ON=Ornate, *T. o. ornata*; TC=*T. carolina* (subspecies-level field identification unavailable). The last two characters represent U.S. state: AL=Alabama, FL=Florida, MS=Mississippi, SC=South Carolina, GA=Georgia, LA=Louisiana, IL=Illinois). Each plot corresponds to tests between parental groups (A) EAxGU (N=109), (B) EAxTT (N=135), (C) GUxTT (N=139), and (D) EAxON (N=112). A posterior probability threshold >0.8 was required for genotype frequency class assignments, which included P_1 and P_2 (parental types), F_1 and F_2 (first and second-generation hybrids), backcrosses (B_1 and B_2), and F_N (unclassified).

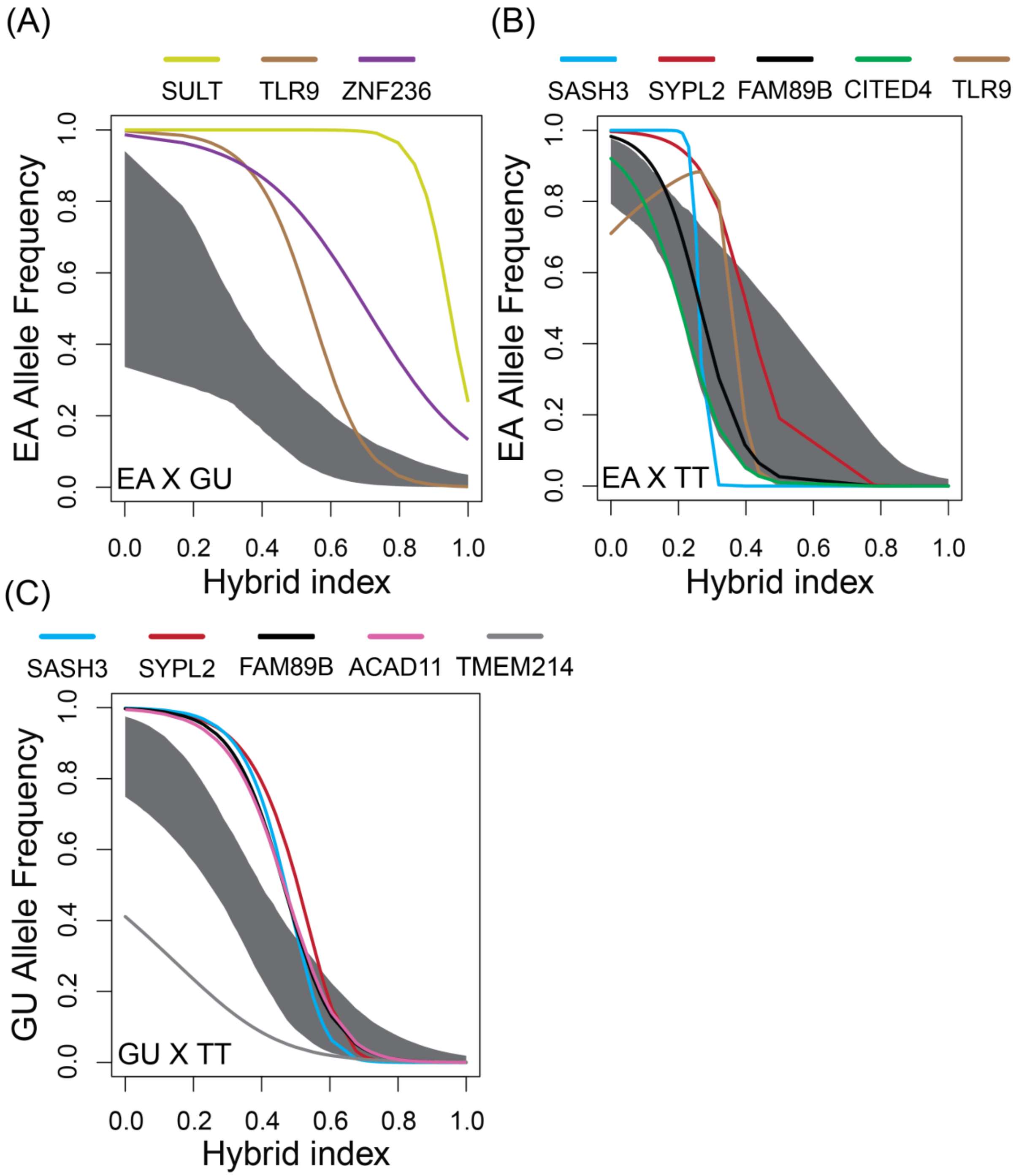


Figure 3: *Terrapene* Genomic clines depicting outlier SNPs found in transcriptome-aligned ddRAD loci. Pairwise comparisons are between EA=Woodland (*T. c. carolina*), GU=Gulf Coast (*T. c. major*), and TT=Three-toed (*T. mexicana triunguis*) box turtles. The gray area represents neutral expectations based on 2,660 (EAxGU), 2,623 (EAxTT), and 2,622 (GUxTT) transcriptome-aligned SNPs, and each line is a genomic cline for one outlier locus (abbreviations defined in Table 1).

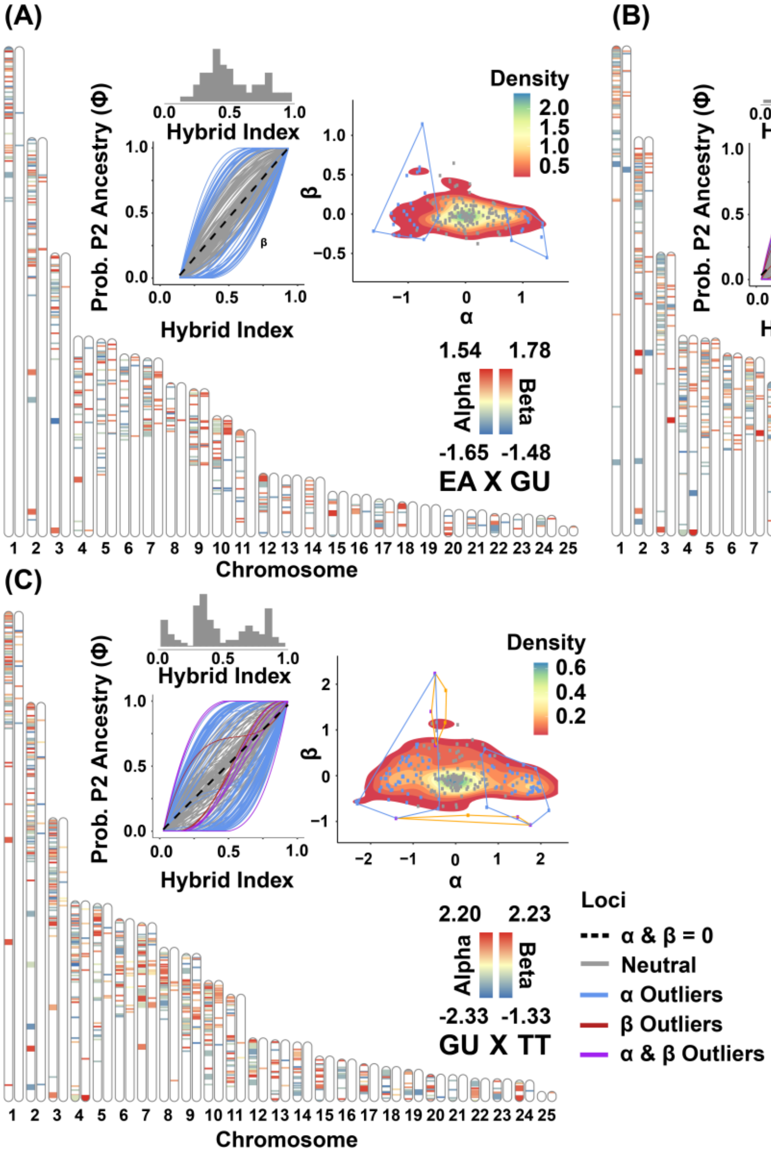


Figure 4: Bayesian genomic cline (BGC) outliers for *Terrapene* ddRADseq SNPs, plotted as heatmaps mapped to *Trachemys scripta* chromosomes. Each chromosome repeats to display significant outliers from the genomic cline center (α; left) and rate (ß; right) BGC parameters. Thinner and thicker bands represent SNPs from unknown scaffolds and annotated genes. BGC was run pairwise for three taxa: EA=*T. carolina carolina*, GU=*T. c. major,* and TT=*T. mexicana triunguis.* Outliers were significant if they had a 95% CI excluding zero or exceeding the probability distribution's quantile interval $(\frac{1-0.975}{2}, \frac{0.975}{2})$. The ϕ plots depict transcriptome-aligned BGC genomic clines with hybrid index histograms above, and each line represents a genomic cline for one locus. The αXß plots illustrate the BGC parameters as a function of density. Polygons define density space for significant α (blue), ß (orange), and both (purple) outliers.

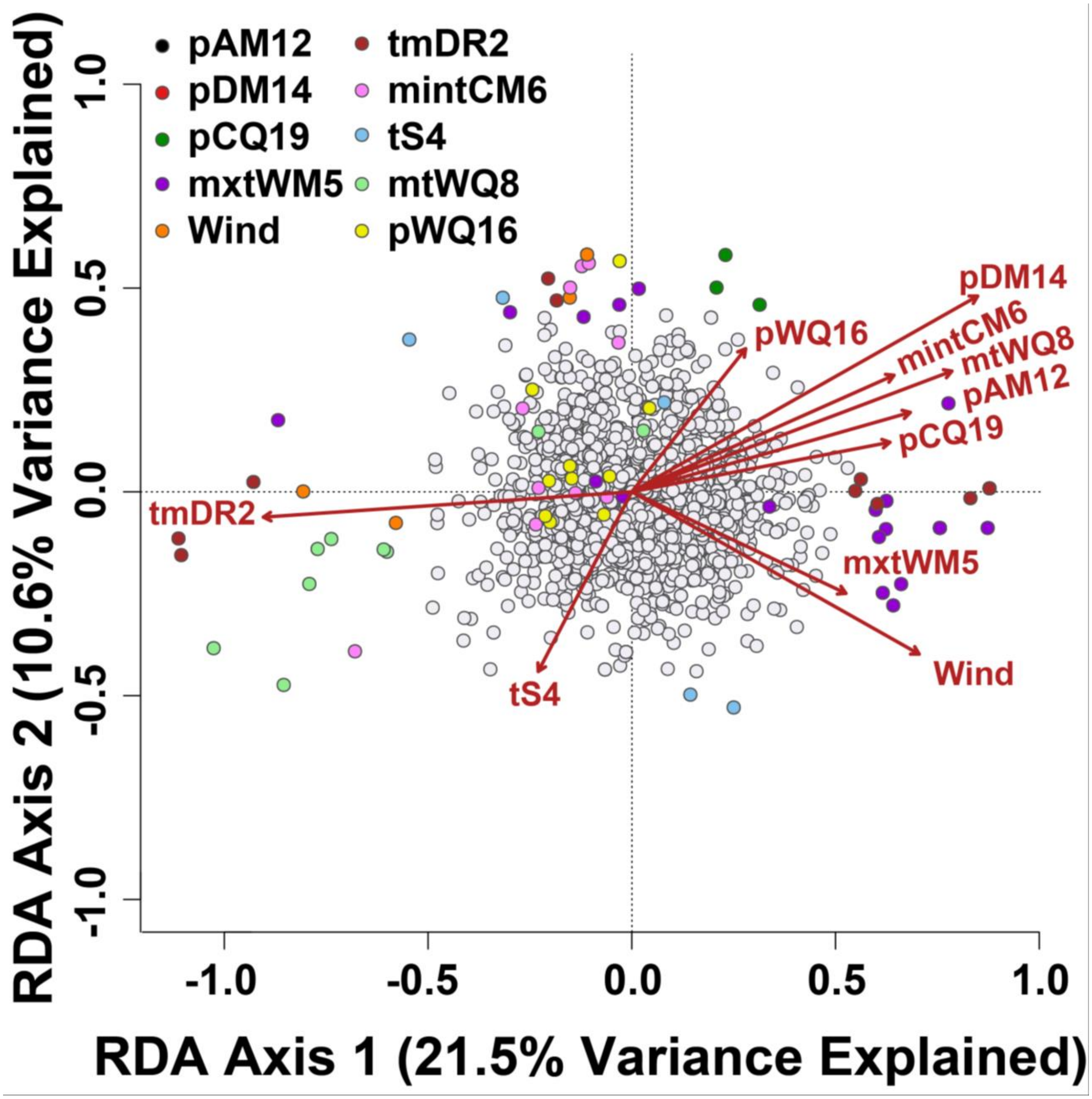


Figure 5: Redundancy analysis (RDA) representing outlier *Terrapene* SNPs correlated with ten predictive, non-redundant BioClim environmental variables (see Table 2 for predictor abbreviations). Significant outliers were designated as being +/- 3 standard deviations from the RDA axis loading means. Pairwise Pearson's correlations between each outlier and environmental variable were performed, and the strongest correlation coefficient (r) determined the best-supported predictor.

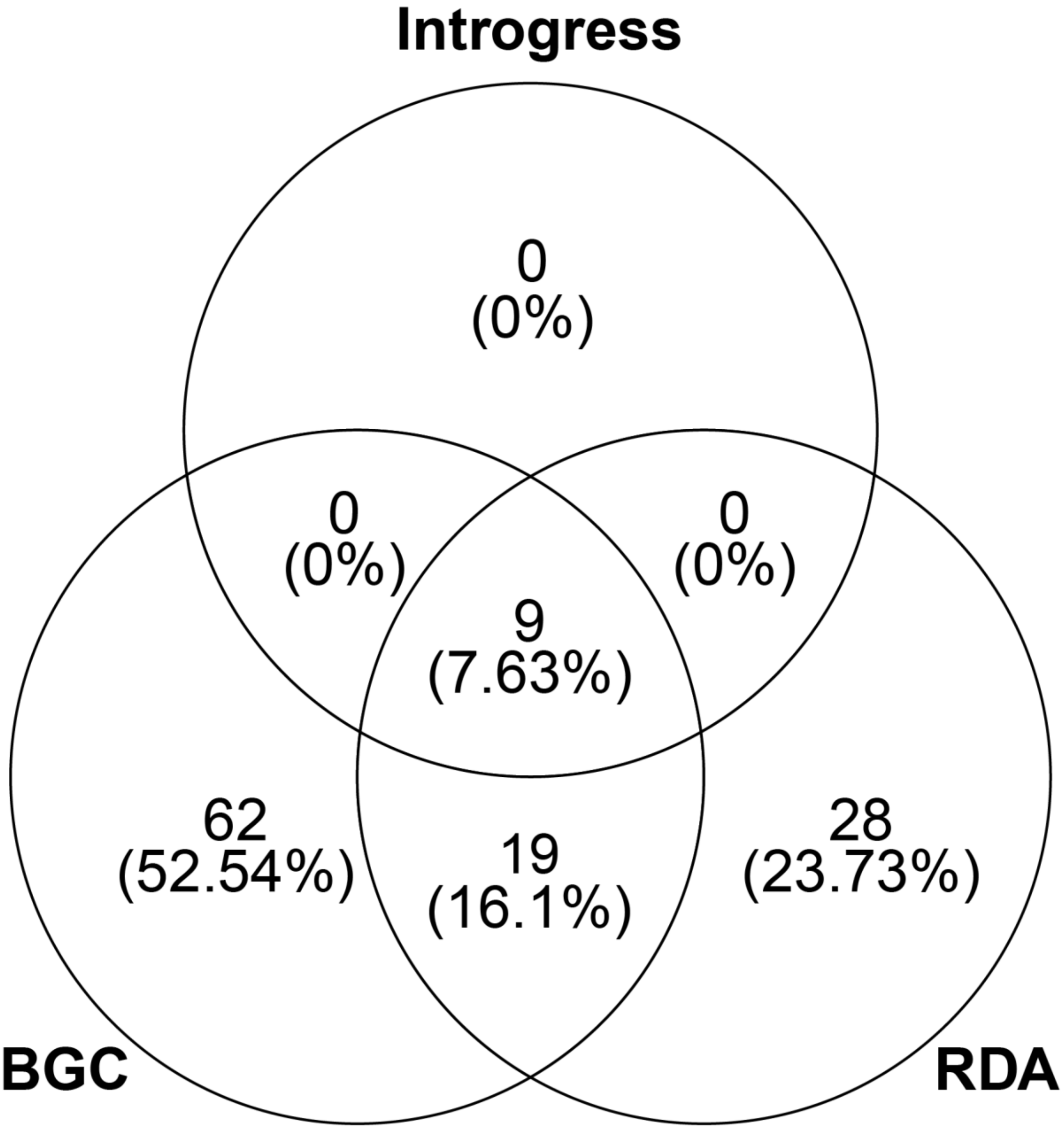


Figure 6: Venn Diagram depicting overlap between *Terrapene* transcriptomic outliers identified in INTROGRESS, Bayesian genomic cline (BGC), and redundancy analysis (RDA). Each value includes raw counts (top) and percentages (bottom).

SUPPLEMENTARY TABLES AND FIGURES

Table S1: Terrapene sample metadata. 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 generally follow the format "mm/year", unless only the year was known. Population codes precede the sample IDs with underscores as delimiters, with the first two characters representing subspecies (when available), and the second two U.S. state locality.

Sample ID	**Collector(s) (Affiliation)**	**Taxonomic ID**	**Sex**	**Tissue**	**State**	**County**	**Date**	**Lat**	**Long**
EAAL_BX1380	R Birkhead	*T. carolina carolina*	M	muscle	AL	Russell	06/2010	32.47	-85.20
EAAL_BX1387	R Birkhead	*T. carolina carolina*	M	toe tips	AL	Bullock	06/2010	32.08	-85.69
EAAL_BX211	R Birkhead	*T. carolina carolina*	-	toenails	AL	Lee	05/2009	32.44	-85.35
EAAL_BXEA27	D O'Halloran, R Birkhead	*T. carolina carolina*	F	tail tip	AL	Dekalb	05/2010	34.45	-85.78
EAGA_BX217	R Batts	*T. carolina carolina*	M	toenails	GA	Harris	05/2009	32.61	-84.82
EAGA_BX219	W Birkhead	*T. carolina carolina*	M	toenails	GA	Harris	05/2009	32.63	-85.00
EAGA_BX220	W Birkhead	*T. carolina carolina*	-	toenails	GA	Muscogee	05/2009	32.56	-84.91
EAGA_BX301	W Birkhead	*T. carolina carolina*	M	toenails	GA	Harris	07/2009	32.74	-84.92
EAGA_BX346	W Birkhead	*T. carolina carolina*	M	toenails	GA	Harris	07/2009	32.71	-84.96
EAGA_BX472	W Birkhead	*T. carolina carolina*	F	toenails	GA	Marion	10/2009	32.30	-84.52
EAGA_BX473	W Birkhead	*T. carolina carolina*	-	toenails	GA	Troup	10/2009	32.87	-85.14
EAGA_BX660	W Birkhead	*T. carolina carolina*	F	toenails	GA	Harris	06/2009	32.62	-84.82
EAGA_BXEA14	R Birkhead	*T. carolina carolina*	F	toe tips	GA	Dekalb	05/2010	33.67	-84.35
EAGA_BXEA15_654	W Birkhead	*T. carolina carolina*	-	toenails	GA	Harris	05/2009	32.85	-84.85
EAGA_BXEA17	W Birkhead	*T. carolina carolina*	-	toenails	GA	Harris	05/2009	32.78	-84.87
EAGA_BXEA19	W Birkhead	*T. carolina carolina*	M	toenails	GA	Harris	06/2009	32.76	-84.90
EAGA_BXEA21	W Birkhead	*T. carolina carolina*	F	toenails	GA	Harris	05/2009	32.79	-84.96
EAGA_BXEA25	W Birkhead	*T. carolina carolina*	F	toenails	GA	Troup	06/2014	32.75	-84.90
EAGA_BXEA26	R Birkhead	*T. carolina carolina*	M	toe tip	GA	Harris	05/2009	32.88	-85.09
EAGA_BXEA29_655	W Birkhead	*T. carolina carolina*	-	toenails	GA	Harris	05/2009	32.69	-84.96
EAGA_BXEA31_659	W Birkhead	*T. carolina carolina*	F	toenails	GA	Harris	05/2009	32.77	-84.91
EAGA_BXEA32_662	W Birkhead	*T. carolina carolina*	-	toenails	GA	Harris	06/2009	32.67	-84.96
EAGA_BXEA33_663	W Birkhead	*T. carolina carolina*	F	toenails	GA	Harris	06/2009	32.60	-84.83

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
EAGA_BXEA34_665	W Birkhead	*T. carolina carolina*	-	toenails	GA	Harris	06/2009	32.80	-84.92
EAGA_BXEA35_666	R Birkhead and W Birkhead	*T. carolina carolina*	M	toenails	GA	Harris	06/2009	32.84	-84.94
EAGA_BXEA49_564	R Birkhead	*T. carolina carolina*	-	toenails	GA	Harris	10/2009	32.70	-84.74
EAIL_BX24	C Phillips (INHS)	*T. carolina carolina*	-	blood	IL	Will	2004	41.44	-87.57
EAIL_BX25	C Phillips (INHS)	*T. carolina carolina*	-	blood	IL	Will	2004	41.40	-87.60
EAIL_BX28	C Phillips (INHS)	*T. carolina carolina*	-	blood	IL	Clinton	-	-	-
EAIL_BX33	C Phillips (INHS)	*T. carolina carolina*	-	blood	IL	Clinton	-	-	-
EAIL_BX34	C Phillips (INHS)	*T. carolina carolina*	-	blood	IL	Pope	2006	37.32	-88.73
EAIL_BXIL02	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Vermillion	-	-	-
EAIL_BXIL03	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Macoupin	2009	39.06	-89.75
EAIL_BXIL04	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Marion	2009	38.63	-88.79
EAIL_BXIL05	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Green	2009	39.29	-90.53
EAIL_BXIL06	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Clinton	2008	38.61	-89.35
EAIL_BXIL07	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Macoupin	2009	39.04	-89.98
EAIL_BXIL08	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Effingham	2009	39.08	-88.67
EAIL_BXIL09	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jersey	2009	39.97	-90.54
EAIL_BXIL12	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Richland	2012	38.78	-88.09
EAIL_BXIL13	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Wayne	2012	38.54	-88.59
EAIL_BXIL14	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Hamilton	2013	38.05	-88.40
EAIL_BXIL15	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Crawford	2013	39.10	-87.73
EAIL_BXIL16	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Massac	2013	37.16	-88.70
EAIL_BXIL18	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jackson	2013	37.56	-89.21
EAIL_BXIL19	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jasper	2012	39.03	-88.11
EAIL_BXIL20	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Johnson	2013	37.49	-88.91
EAIL_BXIL21	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Massac	2013	37.13	-88.65
EAIL_BXIL22	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jackson	2013	37.60	-89.18
EAIL_BXIL23	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jersey	2009	39.04	-90.14
EAIL_BXIL24	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Marion	-	-	-

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
EAIL_BXIL33	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Vermillion	-	-	-
EAIL_BXIL36	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Vermillion	-	-	-
EAIL_BXIL39	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Vermillion	-	-	-
EAIL_BXIL42	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Vermillion	-	-	-
EAIL_BXIL45	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Marion	-	-	-
EAIL_BXIL46	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Effingham	2012	39.06	-88.70
EAIL_BXIL47	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Clinton	2012	38.62	-89.29
EAIL_BXIL51	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jasper	2012	38.95	-88.25
EAIL_BXIL53	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Brown	2014	39.92	-90.88
EAIL_BXIL54	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jersey	-	39.05	-90.10
EAIL_BXIL56	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Effingham	2012	38.99	-88.62
EAIL_BXIL57	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Jersey	-	39.04	-90.40
EAIL_BXIL61	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Madison	2009	38.90	-89.93
EAIL_BXIL62	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Will	2004	41.44	-87.57
EAIL_BXIL70	C Phillips, M Allender (INHS)	*T. carolina carolina*	-	blood	IL	Saline	2014	37.72	-88.63
EAKY_BX1027	unknown	*T. carolina carolina*	M	toenails	KY	Carter	-	-	-
EAKY_BX26	GJ Watkins-Colwell, A.A. Leenders, B.T. Roach, L. Colwell (PMNH, Yale)	*T. carolina carolina*	-	-	KY	Laurel	08/2004	37.00	-84.24
EAKY_BX27	GJ Watkins-Colwell, A.A. Leenders, B.T. Roach, L. Colwell (PMNH, Yale)	*T. carolina carolina*	-	-	KY	Leslie	08/2004	37.25	-83.38
EAMS_BXEA28	J Lee (Nature Conservancy, Camp Shelby)	*T. carolina carolina*	M	toenails	MS	Tishomingo	07/2008	-	-
EANC_BX316	J Reynolds (NC State Museum of Natural Sciences)	*T. carolina carolina*	-	-	NC	Johnston	07/2009	35.68	-78.47
EANC_BX318	J Reynolds (NC State Museum of Natural Sciences)	*T. carolina carolina*	-	-	NC	Johnston	08/2009	35.68	-78.46
EANY_BXEA11	E Smithes-Baker	*T. carolina carolina*	M	tail tip	NY	Westchester	-	41.29	-73.87
EAPA_BXEA13	S Ray	*T. carolina carolina*	F	tail tip	PA	-	-	40.09	-76.89
EARI_BX1608	Rhode Island Dept. of Environmental Management	*T. carolina carolina*	M	toenails	RI	Washington	06/2011	-	-
EASC_BX1108	M Martin	*T. carolina carolina*	M	shell shavings	SC	Beaufort	06/2010	32.33	80.70

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
EASC_BX1109	M Martin	*T. carolina carolina*	M	toenails	SC	Beaufort	06/2010	32.34	80.70
EASC_BX1110	M Martin	*T. carolina carolina*	M	muscle	SC	Jasper	06/2010	32.45	-81.11
EASC_BX1111	M Martin	*T. carolina carolina*	F	toenails	SC	Barnwell	06/2010	33.43	-81.29
EASC_BX1112	M Martin	*T. carolina carolina*	F	toenails	SC	Beaufort	06/2010	32.34	-80.70
EASC_BX1114	M Martin	*T. carolina carolina*	F	toenails	SC	Beaufort	06/2010	32.36	-80.71
EASC_BX1115	M Martin	*T. carolina carolina*	M	toenails	SC	Beaufort	06/2010	32.32	-80.69
EASC_BX1116	M Martin	*T. carolina carolina*	F	toenails	SC	Beaufort	06/2010	32.35	-80.70
EASC_BX231	J Smith	*T. carolina carolina*	-	muscle	SC	Laurens	06/2009	34.45	-81.86
EASC_BX232	J Smith	*T. carolina carolina*	-	toenails	SC	Laurens	06/2009	34.52	-81.99
EASC_BX235	J Smith	*T. carolina carolina*	-	foot tissue	SC	Laurens	07/2009	-	-
EASC_BXEA40_1304	ML Edwards (Erskine College)	*T. carolina carolina*	M	toenails	SC	Abbeville	10/2010	34.33	-82.38
EASC_BXEA41_1305	ML Edwards (Erskine College)	*T. carolina carolina*	-	toenails	SC	Chester	10/2010	34.68	-81.18
EASC_BXEA42_1306	ML Edwards (Erskine College)	*T. carolina carolina*	F	toenails	SC	Greenville	10/2010	34.83	-82.39
EASC_BXEA43_1307	ML Edwards (Erskine College)	*T. carolina carolina*	M	toenails	SC	Abbeville	10/2010	34.33	-82.39
EATN_BX35	W Duzak	*T. carolina carolina*	-	tail tip	TN	Davidson	-	36.13	-86.93
EATN_BXEA02_36x2	W Duzak	*T. carolina carolina*	-	tail tip	TN	Davidson	-	36.13	-86.87
EAVA_BX101	Wildlife Center of Virginia	*T. carolina carolina*	M	-	VA	Albemarle	-	-	-
EAVA_BX103	Wildlife Center of Virginia	*T. carolina carolina*	M	-	VA	Fluvanna	-	38.04	-78.91
EAVA_BX104	Wildlife Center of Virginia	*T. carolina carolina*	M	-	VA	Fluvanna	-	38.04	-78.91
EAVA_BX320	E Winther	*T. carolina carolina*	M	toenails	VA	Norfolk	-	36.68	-76.29
EAVA_BX321	E Winther	*T. carolina carolina*	M	toenails	VA	Dinwiddie	-	37.22	-77.39
EAWV_BX449	A Gooley	*T. carolina carolina*	-	tail tip	WV	Roane	07/2009	38.54	-81.33
GUAL_BX275	R Birkhead and V Jo	*T. carolina major*	F	tail tip	AL	Houston	08/2009	31.24	-85.12
GUAL_BXGU02	R Birkhead	*T. carolina major*	M	toe tips	AL	Mobile	06/2010	30.55	-88.12
GUAL_BXGU03	R Birkhead	*T. carolina major*	F	tail tip	AL	Mobile	06/2010	30.54	-88.12
GUAL_BXGU04	R Birkhead	*T. carolina major*	F	toe tips	AL	Baldwin	06/2010	30.63	-87.91
GUAL_BXGU05	R Birkhead	*T. carolina major*	M	tail tip	AL	Baldwin	06/2010	30.64	-87.91

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
GUFL_BX503	M Michelsohn	*T. carolina major*	-	scutes	FL	Franklin	05/2009	29.94	-85.01
GUFL_BX504	M Michelsohn	*T. carolina major*	-	toenails	FL	Franklin	05/2009	29.80	-84.83
GUFL_BX509	Museum of Vertebrate Zoology	*T. carolina major*	-	muscle	FL	Wakulla	-	30.06	-84.57
GUFL_BX626	M Greene	*T. carolina major*	M	muscle	FL	Walton	04/2010	30.44	-85.96
GUFL_BX627	M Aresco	*T. carolina major*	F	muscle	FL	Walton	06/2009	30.49	-85.94
GUFL_BX628	B Walker	*T. carolina major*	M	muscle	FL	Calhoun	04/2009	30.43	-85.12
GUFL_BX684	W Birkhead	*T. carolina major*	M	toenails	FL	Gulf	06/2009	30.06	-85.19
GUFL_BX685	W Birkhead	*T. carolina major*	F	toenails	FL	Gulf	06/2009	29.82	-85.28
GUFL_BX908	D Steen	*T. carolina major*	F	toenails	FL	Okaloosa	07/2009	30.75	-86.56
GUFL_BX909	D Steen	*T. carolina major*	-	toe tips	FL	Okaloosa	07/2009	30.67	-86.63
GUFL_BX910	D Steen	*T. carolina major*	M	toe tips	FL	Okaloosa	08/2009	30.75	-86.56
GUFL_BX911	D Steen	*T. carolina major*	-	toe tips	FL	Okaloosa	05/2009	-	-
GUFL_BX929	K Krysko, PE Moler	*T. carolina major*	-	muscle	FL	Holmes	05/2007	-	-
GUFL_BX931	K Krysko	*T. carolina major*	-	muscle	FL	Escambia	-	30.57	-87.40
GUFL_BX933	K Krysko	*T. carolina major*	-	muscle	FL	Gulf	03/2007	29.85	-85.26
GUFL_BXGU07	R Birkhead	*T. carolina major*	M	toenails	FL	Gulf	09/2013	30.08	-85.19
GUFL_BXGU08	R Birkhead	*T. carolina major*	F	tail tip	FL	Gulf	07/2013	29.99	-85.17
GUFL_BXGU10	R Birkhead	*T. carolina major*	M	toenails	FL	Gulf	07/2013	29.68	-85.33
GUFL_BXGU11	R Birkhead	*T. carolina major*	-	tail tip	FL	Gulf	07/2012	30.16	-85.21
GUFL_BXGU13	R Birkhead, J McGuire	*T. carolina major*	F	tail tip	FL	Gulf	07/2015	29.69	-85.25
GUFL_BXGU25	R Birkhead, J McGuire	*T. carolina major*	M	tail tip	FL	Gulf	07/2015	29.89	-85.22
GUFL_BXGU26	R Birkhead, C Ward	*T. carolina major*	F	foot tissue	FL	Gulf	05/2015	29.87	-85.23
GUFL_BXGU27	R Birkhead, C Ward	*T. carolina major*	F	toenails	FL	Calhoun	05/2015	30.50	-85.12
GUFL_BXGU28	R Birkhead, C Ward	*T. carolina major*	F	toenails, muscle	FL	Gulf	05/2015	29.84	-85.27
GUFL_BXGU29	R Birkhead, C Ward	*T. carolina major*	F	toenail	FL	Gulf	05/2015	30.00	-85.17
GUFL_BXGU31	R Birkhead	*T. carolina major*	F	toenails, toe	FL	Walton	07/2016	30.49	-86.23

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
GUFL_BXGU32	R Birkhead	*T. carolina major*	F	tail tip	FL	Franklin	07/2016	29.72	-84.99
GUFL_BXGU33	R Birkhead	*T. carolina major*	F	toenails	FL	Gulf	07/2016	29.93	-85.39
GUFL_BXGU35_AA13	D Alix	*T. carolina major*	M	toenails	FL	Liberty	03/2018	30.37	-84.68
GUFL_BXGU36_AA14	D Alix	*T. carolina major*	F	toenails	FL	Franklin	04/2018	30.00	-84.89
GUFL_BXGU37_1391	R Birkhead, J Westmoreland	*T. carolina major*	F	toenails	FL	Bay	05/2010	30.19	-85.68
GUFL_BXGU38_502	M Michelsohn	*T. carolina major*	-	scutes	FL	Franklin	05/2009	29.86	-84.75
GUFL_BXGU61_U57	R Birkhead	*T. carolina major*	M	toenails, skin	FL	Gulf	09/2013	30.08	-85.19
GUFL_BXGU62_AA36	C Matechik	*T. carolina major*	F	toenails	FL	Franklin	02/2018	29.88	-84.73
GUFL_BXGU63_AA37	C Matechik	*T. carolina major*	M	toenails	FL	Franklin	03/2018	29.84	-84.68
GUFL_BXGU65_AA39	C Matechik	*T. carolina major*	-	toenails	FL	Franklin	03/2018	29.75	-84.84
GUFL_BXGU66_AA40	C Matechik	*T. carolina major*	-	toenails	FL	Franklin	04/2018	29.85	-84.69
GULA_BX762	A Bass	*T. carolina major*	-	toenails	LA	Bastrob	-	32.71	-91.93
GUMS_BX188	J Lee (Nature Conservancy, Camp Shelby)	*T. carolina major*	F	tail tip	MS	Forrest	07/2008	31.15	-89.18
GUMS_BX190	J Lee (Nature Conservancy, Camp Shelby)	*T. carolina major*	F	tail tip	MS	Perry	07/2008	31.06	-89.12
GUMS_BX193	J Lee (Nature Conservancy, Camp Shelby)	*T. carolina major*	M	tail tip	MS	Perry	07/2008	31.14	-89.15
GUMS_BX200	J Lee (Nature Conservancy, Camp Shelby)	*T. carolina major*	M	tail tip	MS	Perry	04/2009	31.14	-89.15
GUMS_BX201	J Lee (Nature Conservancy, Camp Shelby)	*T. carolina major*	M	tail tip	MS	Perry	05/2009	31.14	-89.15
GUMS_BXGU15	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU17	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU18	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU20	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU21	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	07/2015	30.44	-88.55
GUMS_BXGU22	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	06/2015	30.44	-88.55
GUMS_BXGU23	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU24	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
GUMS_BXGU30	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU39_T34	A Lynn McCoy	*T. carolina major*	M	toenails	MS	Jackson	04/2015	30.42	-88.52
GUMS_BXGU40_T46	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU43_T55	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	07/2015	30.44	-88.55
GUMS_BXGU44_T56	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	32.55	-85.64
GUMS_BXGU45_T59	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU46_T60	A Lynn McCoy	*T. carolina major*	M	toenails	MS	Jackson	04/2015	30.42	-88.52
GUMS_BXGU47_T61	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU48_T62	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU49_T66	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	07/2015	30.44	-88.55
GUMS_BXGU50_T69	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU51_T70	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU52_T71	A Lynn McCoy	*T. carolina major*	M	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU53_T72	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU54_T73	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU55_T78	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.44	-88.55
GUMS_BXGU56_T83	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU58_T92	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.44	-88.55
GUMS_BXGU59_T93	A Lynn McCoy	*T. carolina major*	F	tail tip	MS	Jackson	05/2015	30.43	-88.54
GUMS_BXGU67_AA41	J Wright	*T. carolina major*	-	tail tip	MS	Jackson	04/2018	30.38	-88.71
GUMS_BXGU68_AA42	J Wright	*T. carolina major*	F	tail tip	MS	Jackson	05/2018	30.38	-88.71
GUMS_BXGU69_AA43	J Wright	*T. carolina major*	M	foot tissue	MS	Jackson	05/2018	30.38	-88.71
GUMS_BXGU70_AA44	J Wright	*T. carolina major*	F	tail tip	MS	Jackson	05/2018	30.38	-88.70
GUMS_BXGU71_AA45	J Wright	*T. carolina major*	M	tail tip	MS	Jackson	05/2018	30.38	-88.72
GUMS_BXGU72_AA46	J Wright	*T. carolina major*	-	tail tip	MS	Jackson	05/2018	30.40	-88.73
GUMS_BXGU73_AA47	J Wright	*T. carolina major*	F	tail tip	MS	Jackson	05/2018	30.36	-88.71
GUMS_BXGU74_AA48	J Wright	*T. carolina major*	M	skin	MS	Jackson	05/2018	30.37	-88.69

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
GUMS_BXGU75_AA49	J Wright	*T. carolina major*	M	tail tip	MS	Jackson	05/2018	30.36	-88.69
GUMS_BXGU76_AA50	J Wright	*T. carolina major*	F	tail tip	MS	Jackson	05/2018	30.38	-88.72
GUMS_BXGU77_AA51	J Wright	*T. carolina major*	F	toe	MS	Jackson	06/2018	30.38	-88.71
GUMS_BXGU78_AA52	J Wright	*T. carolina major*	M	tail tip	MS	Jackson	06/2018	30.38	-88.71
GUMS_BXGU79_AA73	J Wright	*T. carolina major*	M	tail tip	MS	Jackson	06/2018	30.40	-88.75
GUMS_BXGU80_AA54	J Wright	*T. carolina major*	M	foot tissue	MS	Jackson	06/2018	30.37	-88.71
ONCO_BX580	AE Nash (Colorado Reptile Human Society)	*T. ornata ornata*	F	toenails	CO	Weld	06/2009	40.30	-104.47
ONCO_BX588	AE Nash (Colorado Reptile Human Society)	*T. ornata ornata*	M	toenails	CO	Weld	06/2009	40.29	-104.48
ONCO_BX601	AE Nash (Colorado Reptile Human Society)	*T. ornata ornata*	F	toenails	CO	Weld	05/2009	40.30	-104.48
ONCO_BX602	AE Nash (Colorado Reptile Human Society)	*T. ornata ornata*	M	toenails	CO	Weld	05/2009	40.29	-104.48
ONIA_BX1435	F Janzen	*T. ornata ornata*	-	blood	IA	-	04/2011	-	-
ONIL_BX32	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Franklin	2005	38.11	-88.94
ONIL_BXON01	C Phillips, L Adamovicz (INHS)	*T. ornata ornata*	-	blood	IL	Marion	2017	-	-
ONIL_BXON02	C Phillips, L Adamovicz (INHS)	*T. ornata ornata*	-	blood	IL	Marion	2017	-	-
ONIL_BXON03	C Phillips, L Adamovicz (INHS)	*T. ornata ornata*	-	blood	IL	Marion	2016	-	-
ONIL_BXON04	C Phillips, L Adamovicz (INHS)	*T. ornata ornata*	-	blood	IL	Marion	2016	-	-
ONIL_BXON05	C Phillips, L Adamovicz (INHS)	*T. ornata ornata*	-	blood	IL	Marion	2016	-	-
ONIL_BXON06	C Phillips, L Adamovicz (INHS)	*T. ornata ornata*	-	blood	IL	Marion	2016	-	-
ONIL_BXON07	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON08	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON09	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON10	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON11	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON12	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON13	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON14	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
ONIL_BXON15	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON16	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON17	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON18	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON20	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON21	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON22	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2016	-	-
ONIL_BXON23	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON25	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON26	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON27	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON28	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON29	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON30	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON31	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2013	41.91	-89.35
ONIL_BXON32	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON33	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	-	-	-	-
ONIL_BXON34	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2012	41.89	-89.35
ONIL_BXON35	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Effingham	2008	39.07	-88.54
ONIL_BXON36	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2013	41.92	-89.36
ONIL_BXON37	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Scott	2008	39.59	-90.53
ONIL_BXON38	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Lee	2012	41.89	-89.35
ONIL_BXON40	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Jasper	-	-	-
ONIL_BXON41	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Franklin	2005	38.11	-88.94
ONIL_BXON42	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Jasper	2012	38.93	-88.19
ONIL_BXON43	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Cass	2013	40.00	-90.07
ONIL_BXON44	C Phillips (INHS)	*T. ornata ornata*	-	blood	IL	Will	2013	41.06	-87.59

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
ONKS_BXON47_160	J Jagels (Kansas Dept. of Wildlife and Parks)	*T. ornata ornata*	-	tail tip	KS	Clark	06/2009	37.41	-99.76
ONKS_BXON50_168	J Jagels (Kansas Dept. of Wildlife and Parks)	*T. ornata ornata*	F	tail tip	KS	Meade	06/2009	37.38	-100.14
ONKS_BXON52_171	J Jagels (Kansas Dept. of Wildlife and Parks)	*T. ornata ornata*	-	tail tip	KS	Meade	06/2009	37.04	-100.49
ONKS_BXON53_172	J Jagels (Kansas Dept. of Wildlife and Parks)	*T. ornata ornata*	-	tail tip	KS	Meade	06/2009	37.07	-100.47
ONKS_BXON54_173	J Jagels (Kansas Dept. of Wildlife and Parks)	*T. ornata ornata*	-	tail tip	KS	Meade	06/2009	37.29	-100.37
ONKS_BXON61_133	L Schmidt	*T. ornata ornata*	F	tail tip	KS	Miami	-	38.55	-94.94
ONKS_BXON62_134	L Schmidt	*T. ornata ornata*	F	tail tip	KS	Douglas	-	38.77	-95.15
ONKS_BXON64_136	L Schmidt	*T. ornata ornata*	F	tail tip	KS	Osage	-	38.78	-95.51
ONNE_BXON56_431	J Iverson	*T. ornata ornata*	-	tail tip	NE	Box Butte	06/2009	42.09	-102.72
ONNE_BXON57_432	J Iverson	*T. ornata ornata*	-	tail tip	NE	Sheridan	06/2009	42.06	-102.46
ONNE_BXON58_433x2	J Iverson	*T. ornata ornata*	-	blood	NE	Garden	06/2009	41.83	-102.34
ONNE_BXON59_439	J Iverson	*T. ornata ornata*	-	tail tip	NE	Box Butte	06/2009	42.09	-102.74
ONTX_BX765	C Franklin	*T. ornata ornata*	-	muscle	TX	Montague	10/2009	33.48	-97.79
ONTX_BXON45_150	A Inslee (Aransas/Matagorda Island National Wildlife Refuge Complex)	*T. ornata ornata*	M	-	TX	Calhoun	05/2009	28.20	-96.70
ONTX_BXON46_153	A Inslee (Aransas/Matagorda Island National Wildlife Refuge Complex)	*T. ornata ornata*	M	-	TX	Calhoun	06/2009	28.29	-96.53
ONWI_BX486	B Hay	*T. ornata ornata*	-	toenails	WI	Sauk	-	43.18	-90.07
ONWI_BX489	B Hay	*T. ornata ornata*	F	toenails	WI	Iowa	-	43.03	-90.11
ONWI_BX490	B Hay	*T. ornata ornata*	F	toenails	WI	Iowa	-	43.03	-90.11
ONWI_BX491	B Hay	*T. ornata ornata*	F	toenails	WI	Iowa	-	43.03	-90.11
ONWI_BX492	B Hay	*T. ornata ornata*	F	toenails	WI	Iowa	-	43.03	-90.11
ONWI_BX493	B Hay	*T. ornata ornata*	F	toenails	WI	Dane	-	43.18	-89.80
ONWI_BX495	B Hay	*T. ornata ornata*	F	toenails	WI	Columbia	-	43.46	-89.39

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
ONWI_BX497	B Hay	*T. ornata ornata*	F	toenails	WI	Columbia	-	43.45	-89.35
TCAL_BX1614	R Birkhead	*T. carolina*	M	tail tip	AL	Barbour	04/2011	31.88	-85.46
TCAL_BX1616	R Birkhead	*T. carolina*	F	toenails	AL	Barbour	04/2011	32.03	-85.09
TCAL_BX1618	R Birkhead	*T. carolina*	M	tail tip	AL	Russell	09/2011	32.26	-85.42
TCAL_BX271	R Birkhead	*T. carolina*	M	toenails	AL	Lee/Macon	07/2009	32.54	-85.59
TCAL_BX272	R Birkhead	*T. carolina*	M	tail tip	AL	Lowndes	07/2009	32.21	-86.55
TCAL_BX273	R Birkhead	*T. carolina*	-	toenails	AL	Macon	07/2009	32.48	-85.80
TCAL_BX279	R Birkhead	*T. carolina*	-	tail tip	AL	Macon	08/2009	32.51	-85.61
TCAL_BX280	R Birkhead	*T. carolina*	M	muscle	AL	Tallapoosa	08/2009	32.87	-85.81
TCAL_BX281	R Birkhead	*T. carolina*	F	toenails, muscle	AL	Elmore	08/2009	32.57	-86.03
TCAL_BX282	R Birkhead	*T. carolina*	M	tail tip	AL	Macon	08/2009	32.45	-85.81
TCAL_BX283	R Birkhead	*T. carolina*	M	tail tip	AL	Tallapoosa	08/2009	32.88	-85.82
TCAL_BX289	F Scott, R Birkhead	*T. carolina*	-	scutes	AL	Tallapoosa	2009	32.88	-85.84
TCAL_BX302	R Birkhead	*T. carolina*	F	tail tip	AL	Chambers	08/2009	32.89	-85.38
TCAL_BX304	R Birkhead	*T. carolina*	M	toenails	AL	Lee	09/2009	32.60	-85.53
TCAL_BX305	R Birkhead	*T. carolina*	M	toenails	AL	Elmore	09/2009	32.49	-86.33
TCAL_BX310	R Birkhead	*T. carolina*	-	tail tip	AL	Chambers	09/2009	32.84	-85.48
TCAL_BX326	R Birkhead	*T. carolina*	F	tail tip	AL	Lee	06/2009	32.54	-85.50
TCAL_BX327	R Birkhead	*T. carolina*	F	tail tip	AL	Lee	06/2009	32.69	-85.32
TCAL_BX329	R Birkhead	*T. carolina*	F	tail tip	AL	Chambers	06/2009	32.77	-85.26
TCAL_BX592	R Birkhead	*T. carolina*	M	toenails	AL	Russell	08/2009	32.26	-85.35
TCAL_BX612	R Birkhead, S Graham	*T. carolina*	F	toenails	AL	Barbour	03/2010	32.01	-85.40
TCAL_BXTC01	R Birkhead	*T. carolina*	F	tail tip	AL	Macon	07/2009	32.50	-85.61
TCAL_BXTC09	R Birkhead	*T. carolina*	F	toes	AL	Autauga	05/2010	32.44	-86.41
TCAL_BXTC103	R Birkhead	*T. carolina*	-	toenails, muscle	AL	Hale	08/2013	32.83	-87.60
TCAL_BXTC104	F Scott, R Birkhead	*T. carolina*	-	toenails	AL	Coosa	05/2011	32.87	-86.10
TCAL_BXTC109	R Birkhead	*T. carolina*	-	tail tip	AL	Butler	07/2016	31.59	-86.73

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
TCAL_BXTC11	R Birkhead	*T. carolina*	F	muscle	AL	Macon	07/2010	32.46	-85.66
TCAL_BXTC110	R Birkhead	*T. carolina*	-	tail tip	AL	Conecuh	08/2016	31.29	-87.19
TCAL_BXTC111_U52	R Birkhead	*T. carolina*	F	toenails, muscle	AL	Bullock	07/2013	32.08	-85.69
TCAL_BXTC12	R Birkhead	*T. carolina*	F	liver tissue	AL	Lee	04/2015	32.57	-85.12
TCAL_BXTC21	R Birkhead	*T. carolina*	F	tail tip	AL	Lee	08/2009	32.55	-85.55
TCAL_BXTC24	R Birkhead	*T. carolina*	F	toes	AL	Elmore	05/2010	32.54	-85.95
TCAL_BXTC29	R Birkhead	*T. carolina*	-	toenails, muscle	AL	Lee	03/2012	32.57	-85.53
TCAL_BXTC33	R Birkhead	*T. carolina*	-	tail tip	AL	Macon	04/2015	32.55	-85.64
TCAL_BXTC39	D O'Halloran, R Birkhead	*T. carolina*	F	tail tip	AL	Jackson	08/2011	34.62	-86.20
TCAL_BXTC45	R Birkhead	*T. carolina*	-	toenails, skin	AL	Coffee	08/2014	31.50	-86.01
TCAL_BXTC63	R Birkhead, J McGuire	*T. carolina*	F	toenails	AL	Mobile	04/2014	30.84	-88.40
TCAL_BXTC65	R Birkhead	*T. carolina*	-	toenails, skin	AL	Randolph	07/2013	33.14	-85.46
TCAL_BXTC79	R Birkhead	*T. carolina*	M	muscle	AL	Bullock	04/2011	32.20	-85.50
TCAL_BXTC80	R Birkhead	*T. carolina*	F	toenails, skin	AL	Bullock	07/2013	32.08	-85.69
TCAL_BXTC86	D O'Halloran, S Dery, R Birkhead	*T. carolina*	M	toenails	AL	Madison	07/2010	34.61	-86.58
TCAL_BXTC90	R Birkhead	*T. carolina*	M	tail tip	AL	Coosa	05/2010	33.67	-86.05
TCAL_BXTC91	R Birkhead	*T. carolina*	-	muscle	AL	Clay	05/2010	33.11	-85.89
TCAL_BXTC92	R Birkhead	*T. carolina*	M	muscle	AL	Clay	05/2010	33.13	-85.86
TCAL_BXTC93	R Birkhead	*T. carolina*	F	muscle	AL	Clay	05/2010	33.21	-85.82
TCAL_BXTC94	S Graham	*T. carolina*	-	scutes	AL	St. Claire	04/2010	33.78	-86.24
TCAL_BXTC97	R Birkhead	*T. carolina*	F	muscle	AL	Elmore	10/2009	32.69	-86.10
TCAL_BXTC98	R Birkhead	*T. carolina*	F	tail tip	AL	Elmore	10/2009	32.62	-86.15
TCAL_BXTC99	R Birkhead	*T. carolina*	M	muscle	AL	Shelby	05/2010	33.12	-86.81
TCGA_BX300	W Birkhead	*T. carolina*	F	toenails	GA	Harris	07/2009	32.63	-85.00
TCGA_BX343	W Birkhead	*T. carolina*	F	toenails	GA	Harris	06/2009	32.66	-84.98

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
TCGA_BX344	W Birkhead	*T. carolina*	M	toenails	GA	Harris	06/2009	32.74	-84.90
TCGA_BX345	W Birkhead	*T. carolina*	M	toenails	GA	Calhoun	07/2009	31.51	-84.59
TCGA_BX589	R Birkhead	*T. carolina*	-	tail tip	GA	Gwinett	-	33.98	-83.97
TCGA_BXTC34	J Greenway	*T. carolina*	F	whole foot	GA	Gilmer	06/2015	34.78	-84.39
TCGA_BXTC35	J Greenway	*T. carolina*	-	whole foot	GA	Gilmer	02/2015	34.78	-84.52
TCGA_BXTC36	J Greenway	*T. carolina*	-	whole foot	GA	Pickens	08/2015	34.50	-84.50
TCGA_BXTC46	R Birkhead	*T. carolina*	F	toenails, skin	GA	Coweta	10/2012	33.25	-84.76
TCMS_BXTC13	A Lynn McCoy	*T. carolina*	F	tail tip	MS	Jackson	07/2015	30.63	-88.57
TCMS_BXTC14	A Lynn McCoy	*T. carolina*	F	tail tip	MS	Jackson	07/2015	30.67	-88.49
TCMS_BXTC15	A Lynn McCoy	*T. carolina*	M	tail tip	MS	Jackson	08/2015	30.63	-88.57
TCMS_BXTC16	A Lynn McCoy	*T. carolina*	M	tail tip	MS	Jackson	08/2015	30.67	-88.49
TCMS_BXTC17	A Lynn McCoy	*T. carolina*	M	toe tip	MS	Jackson	05/2015	30.44	-88.55
TCMS_BXTC18	A Lynn McCoy	*T. carolina*	F	tail tip	MS	Jackson	07/2015	30.63	-88.57
TCMS_BXTC84	A Lynn McCoy	*T. carolina*	M	toenails	MS	Jackson	04/2015	30.44	-88.55
TTAR_BX507	B Millig	*T. mexicana triunguis*	M	toenails	AR	Pulaski	08/2009	34.83	-92.49
TTAR_BX984	B Millig	*T. mexicana triunguis*	F	toenails	AR	Pulaski	07/2009	34.83	-92.49
TTAR_BX987	B Millig	*T. mexicana triunguis*	M	toenails	AR	Pulaski	07/2009	34.83	-92.49
TTKS_BXTT20_78	J Jagels (Kansas Dept. of Wildlife and Parks)	*T. mexicana triunguis*	F	tail tip	KS	Crawford	05/2009	37.59	-94.96
TTKS_BXTT23_132	L Schmidt	*T. mexicana triunguis*	F	tail tip	KS	Linn	-	38.34	-94.68
TTLA_BX421	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	M	toenails	LA	Rapides	06/2009	31.20	-92.58
TTLA_BX422	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	F	shell shavings	LA	Rapides	06/2009	31.18	-92.56
TTLA_BX775	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	F	toenails	LA	Grant	08/2009	31.52	-92.53

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
TTLA_BXTT13	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	F	toenails	LA	Rapides	06/2010	31.16	-92.52
TTLA_BXTT34_1482	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	F	toenails	LA	Rapides	06/2010	31.25	-92.64
TTLA_BXTT35_1486	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	M	toenails	LA	Rapides	05/2010	31.14	-92.64
TTLA_BXTT36_1491	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	M	toenails	LA	Rapides	05/2010	31.21	-92.58
TTLA_BXTT37_1492	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	F	toenails	LA	Rapides	05/2010	31.21	-92.58
TTLA_BXTT38_1493	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	M	toenails	LA	Rapides	04/2010	31.16	-92.52
TTLA_BXTT39_1498	S Shively (Calcasieu Ranger District)	*T. mexicana triunguis*	F	toenails	LA	Rapides	04/2010	31.18	-92.52
TTMO_BX109	M Brodt	*T. mexicana triunguis*	M	toenails	MO	Jefferson	-	38.20	-90.53
TTMS_BX191	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	M	tail tip	MS	Perry	07/2008	31.21	-89.07
TTMS_BX192	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	F	tail tip	MS	Perry	08/2008	31.21	-89.07
TTMS_BX195	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	F	tail tip	MS	Perry	08/2008	31.21	-89.07
TTMS_BX196	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	M	tail tip	MS	Perry	08/2008	31.21	-89.07
TTMS_BX198	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	F	tail tip	MS	Forrest	09/2008	31.32	-89.31
TTMS_BX199	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	F	tail tip	MS	Perry	09/2008	31.21	-89.07
TTMS_BX236	J Hoover	*T. mexicana triunguis*	M	toenails	MS	Warren	-	32.36	-90.84
TTMS_BX238	J Hoover	*T. mexicana triunguis*	M	toenails	MS	Hinds	-	32.36	-90.47
TTMS_BX239	J Hoover	*T. mexicana triunguis*	M	toenails	MS	Hinds	-	32.36	-90.47
TTMS_BX240	J Hoover	*T. mexicana triunguis*	F	toenails	MS	Hinds	-	32.36	-90.47
TTMS_BX241	J Hoover	*T. mexicana triunguis*	M	toenails	MS	Hinds	-	32.36	-90.47
TTMS_BX242	J Hoover	*T. mexicana triunguis*	F	toenails	MS	Hinds	-	32.30	-90.87

Table S1 (Cont.)

Sample ID	Collector(s) (Affiliation)	Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
TTMS_BX243	J Hoover	*T. carolina carolina*	F	toenails	MS	Warren	-	32.29	-90.71
TTMS_BX465	B Rosamond (US Fish and Wildlife Service)	*T. mexicana triunguis*	-	scutes	MS	Carroll	-	33.49	-89.85
TTMS_BX467	B Rosamond (US Fish and Wildlife Service)	*T. mexicana triunguis*	M	toenails	MS	Grenada	05/2009	33.78	-90.04
TTMS_BX471	B Rosamond (US Fish and Wildlife Service)	*T. mexicana triunguis*	F	toenails	MS	Panola	09/2009	34.18	-90.11
TTMS_BXTT07	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	F	toe tip	MS	Forrest	07/2008	31.19	-89.25
TTMS_BXTT08	J Lee (Nature Conservancy, Camp Shelby)	*T. mexicana triunguis*	M	tail tip	MS	Forrest	09/2008	31.19	-89.25
TTTX_BX11	J Koukl	*T. mexicana triunguis*	F	toenails	TX	Smith	05/2008	32.30	-95.21
TTTX_BX15	J Koukl	*T. mexicana triunguis*	M	toenails	TX	Dallas	05/2008	32.95	-96.73
TTTX_BX16	J Koukl	*T. mexicana triunguis*	M	toenails	TX	Smith	05/2008	32.35	-95.30
TTTX_BX19	J Koukl	*T. mexicana triunguis*	F	toenails	TX	Dallas	05/2008	32.97	-96.74
TTTX_BX22	J Koukl	*T. mexicana triunguis*	M	toenails	TX	Tarrant	05/2008	33.04	-97.12
TTTX_BX222	J Placyk	*T. mexicana triunguis*	-	tail tip	TX	Henderson	08/2009	32.34	-95.75
TTTX_BX223	J Koukl	*T. mexicana triunguis*	-	toenails	TX	Smith	-	32.34	-95.27
TTTX_BX224	C Samuelson	*T. mexicana triunguis*	-	toenails	TX	Smith	08/2009	32.26	-95.19
TTTX_BX225	C Samuelson	*T. mexicana triunguis*	-	toenails	TX	Smith	08/2009	32.34	-95.27
TTTX_BX227	J Placyk	*T. mexicana triunguis*	-	toenails	TX	Smith	08/2009	32.34	-95.27
TTTX_BX228	J Placyk	*T. mexicana triunguis*	-	toenails	TX	Smith	08/2009	32.34	-95.27
TTTX_BX23	J Koukl	*T. mexicana triunguis*	M	toenails	TX	Collin	05/2006	33.22	-96.57

Table S2: Number of sequenced individuals (N) per *Terrapene* taxon, as identified in the field. *T. carolina carolina*=Woodland, *T. c. major*=Gulf Coast, *T. c. bauri*=Florida, *T. carolina*=field identification limited to species-level, *T. m. triunguis*=Three-toed, and *T. o. ornata*=Ornate box turtles.

Taxonomic ID	**N**
T. carolina carolina	106
T. carolina major	88
T. carolina bauri	4
T. carolina	65
T. mexicana triunguis	47
T. ornata ornata	62
Total	368

Table S3: Genotype frequency proportions from four NewHybrids analyses involving the GU=Gulf Coast (*T. c. major*), EA=Woodland (*T. c. carolina*), TT=Three-toed (*T. m. triunguis*), and ON=Ornate (*T. o. ornata*) box turtles; TC=*T. carolina* (subspecies unidentified). The second two letters in the population ID correspond to U.S. state locality (AL=Alabama, FL=Florida, LA=Louisiana, SC=South Carolina, GA=Georgia, MS=Mississippi, IL=Illinois). Columns depict the proportion of assignment to parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), backcross (B_1 and B_2), and unassigned (FN) genotype frequency classes.

Population	**P1**	**P2**	**F1**	**F2**	**B1**	**B2**	**FN**
GUxEA							
PureGU	1.00	0.00	0.00	0.00	0.00	0.00	0.00
PureEA	0.00	1.00	0.00	0.00	0.00	0.00	0.00
EAAL	0.00	1.00	0.00	0.00	0.00	0.00	0.00
GUFL	0.46	0.04	0.00	0.08	0.21	0.00	0.21
TCAL	0.02	0.86	0.00	0.02	0.02	0.00	0.08
GUAL	0.20	0.20	0.00	0.00	0.00	0.00	0.60
EAxTT							
PureEA	1.00	0.00	0.00	0.00	0.00	0.00	0.00
PureTT	0.00	1.00	0.00	0.00	0.00	0.00	0.00
TTLA	0.00	1.00	0.00	0.00	0.00	0.00	0.00
EASC	0.47	0.00	0.00	0.00	0.40	0.00	0.13
TCGA	0.80	0.00	0.00	0.00	0.00	0.10	0.10
EAGA	0.91	0.00	0.00	0.00	0.05	0.00	0.05
TCAL	0.96	0.00	0.00	0.00	0.04	0.00	0.00
TTxGU							
PureTT	0.00	1.00	0.00	0.00	0.00	0.00	0.00
PureGU	1.00	0.00	0.00	0.00	0.00	0.00	0.00
GUAL	0.60	0.00	0.00	0.00	0.20	0.00	0.20
TTMS	0.00	0.50	0.00	0.17	0.00	0.06	0.28
TTLA	0.00	1.00	0.00	0.00	0.00	0.00	0.00
TCMS	0.43	0.00	0.00	0.00	0.57	0.00	0.00
GUMS	0.52	0.02	0.00	0.00	0.13	0.04	0.28
GUFL	0.88	0.04	0.00	0.00	0.04	0.00	0.04
ONxEA							
PureON	1.00	0.00	0.00	0.00	0.00	0.00	0.00
PureEA	0.00	1.00	0.00	0.00	0.00	0.00	0.00
ONIL	0.74	0.21	0.05	0.00	0.00	0.00	0.00
EAIL	0.00	0.98	0.00	0.00	0.00	0.00	0.03

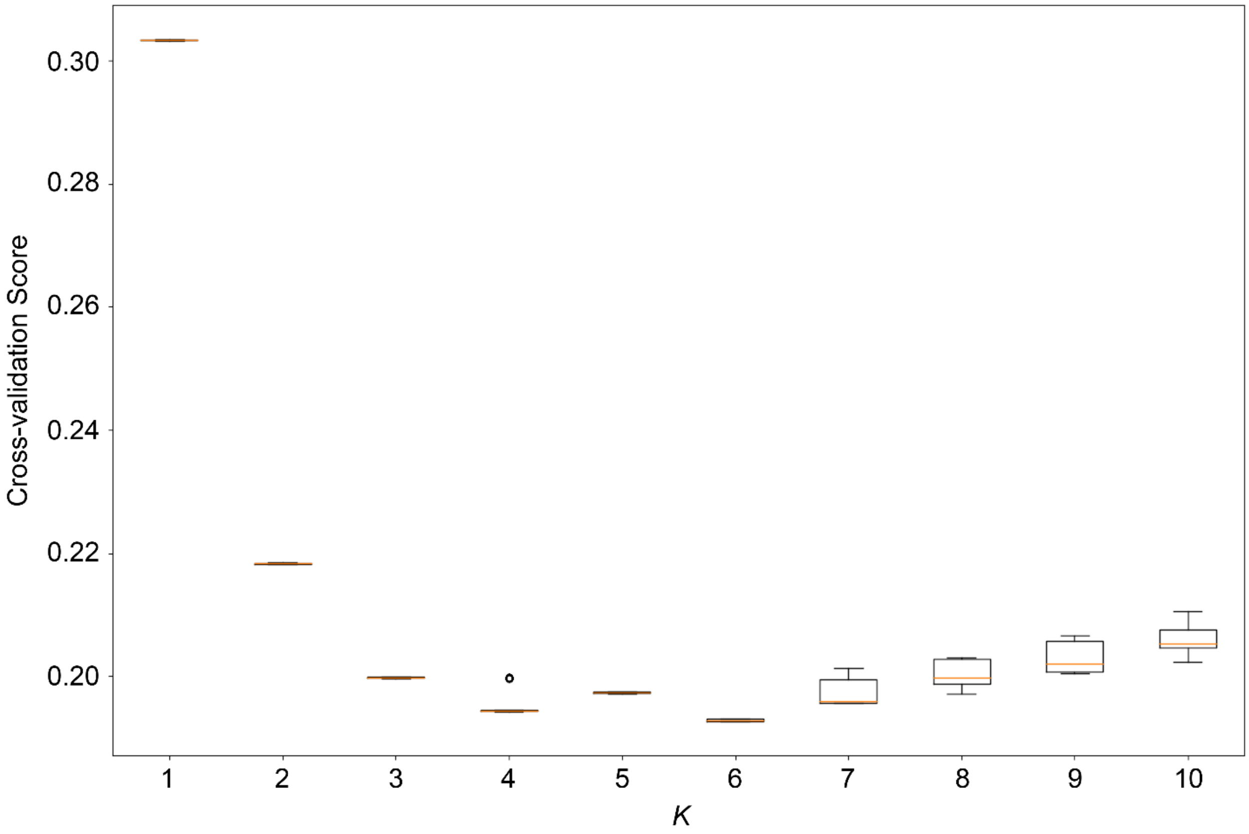


Figure S1: Cross-validation (CV) scores across *K-values* (*K*=1-10) for ADMIXTURE runs containing all sequenced samples (N=368). Lower CV values indicate less error and stronger support for the corresponding *K*.

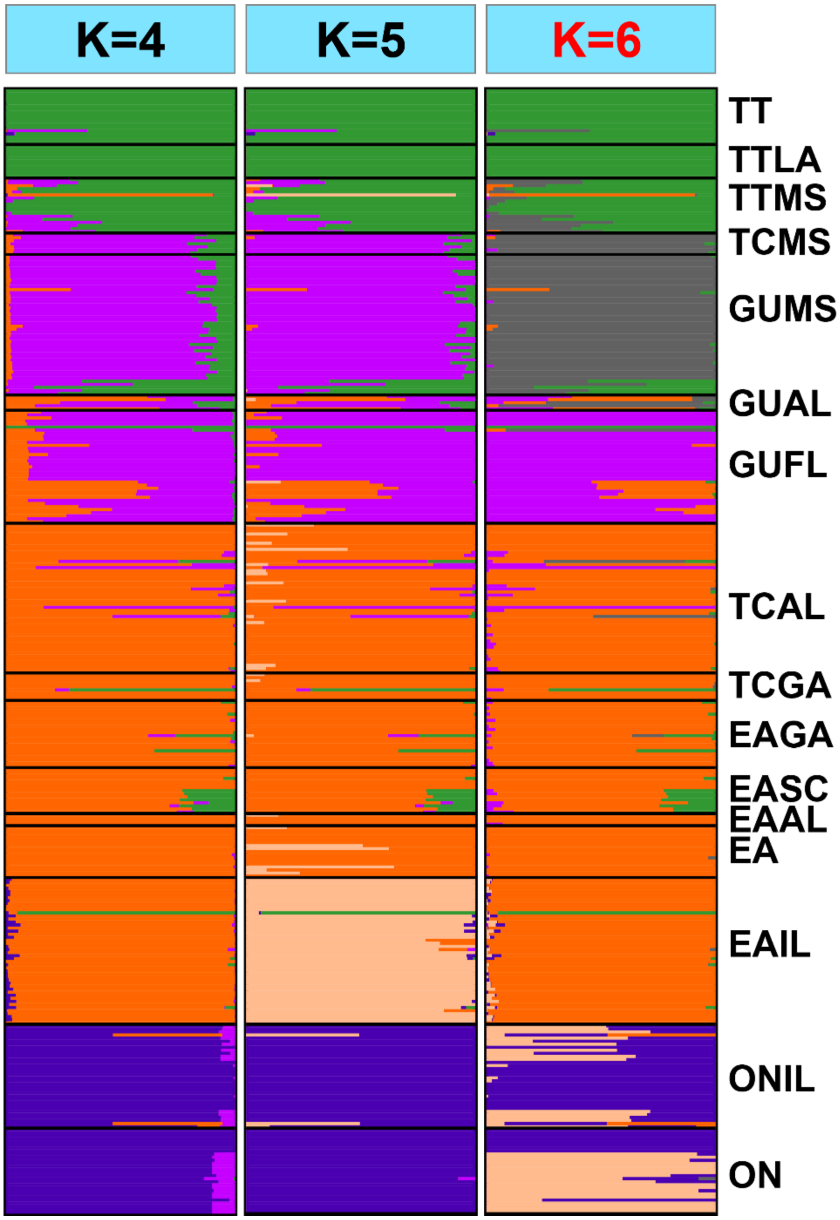


Figure S2: Top three *Terrapene* ADMIXTURE plots representing 12,052 unlinked ddRAD SNPs across all sampled populations. The lowest cross-validation score was for *K*=6 (depicted at right in red), followed by *K*=4 and then *K*=5. Each bar represents a unique individual, and bars with mixed colors represent admixed ancestry. The first two letters of the populations correspond to subspecific field identification (ON=Ornate, *T. ornata ornata*; EA=Woodland, *T. carolina carolina*; GU=Gulf Coast, *T. c. major*; TT=Three-toed, *T. mexicana triunguis*; TC=*Terrapene carolina*, with subspecies unidentified in the field). The second two letters (if present) represent locality codes for U.S. or Mexican state (IL=Illinois; AL=Alabama; GA=Georgia; SC=South Carolina; FL=Florida; MS=Mississippi; LA=Louisiana). Populations lacking a state locality code consisted of multiple localities sampled outside hybrid zones.

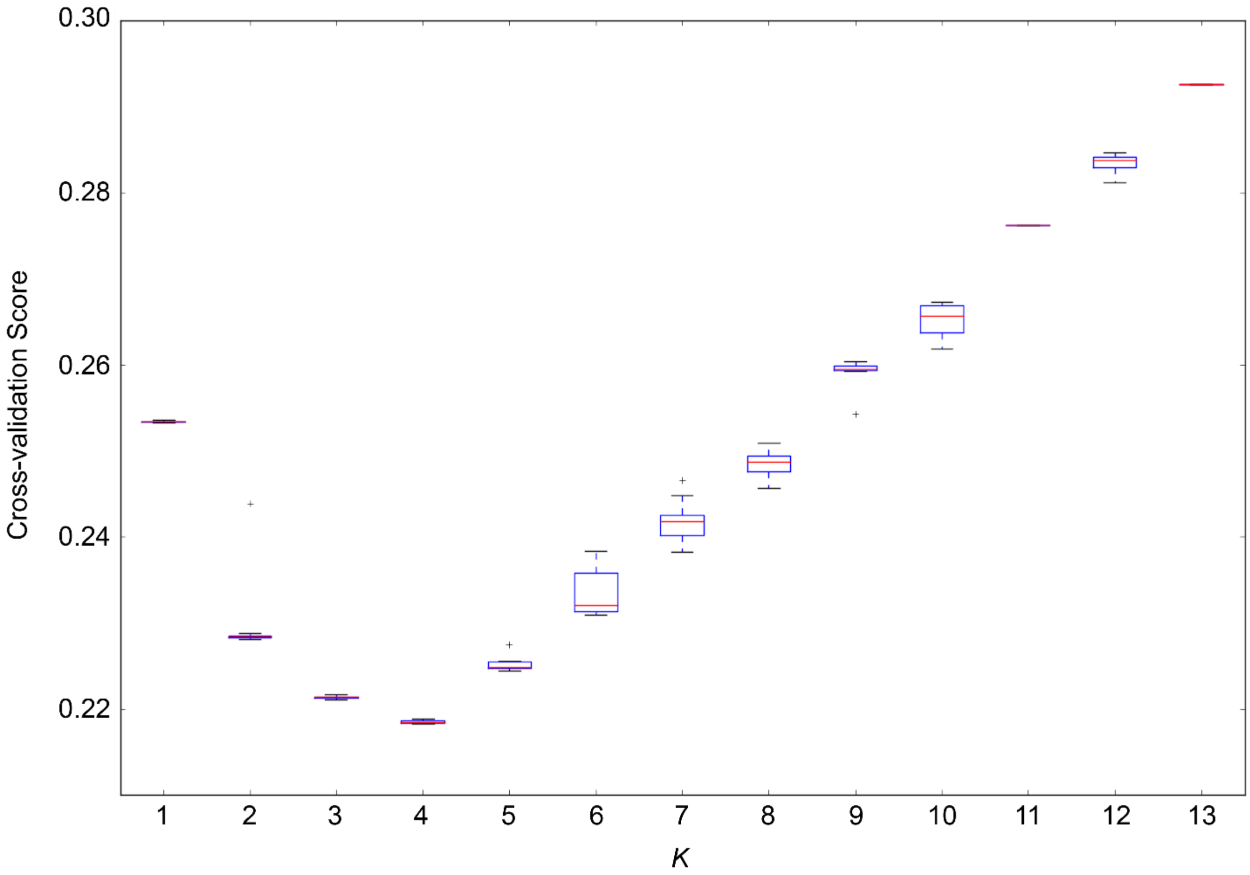


Figure S3: Cross-validation (CV) scores across *K-values* (*K*=1-13) for ADMIXTURE runs containing samples from southeastern North America (N=259). Lower CV values indicate less error and stronger support for the corresponding *K*.

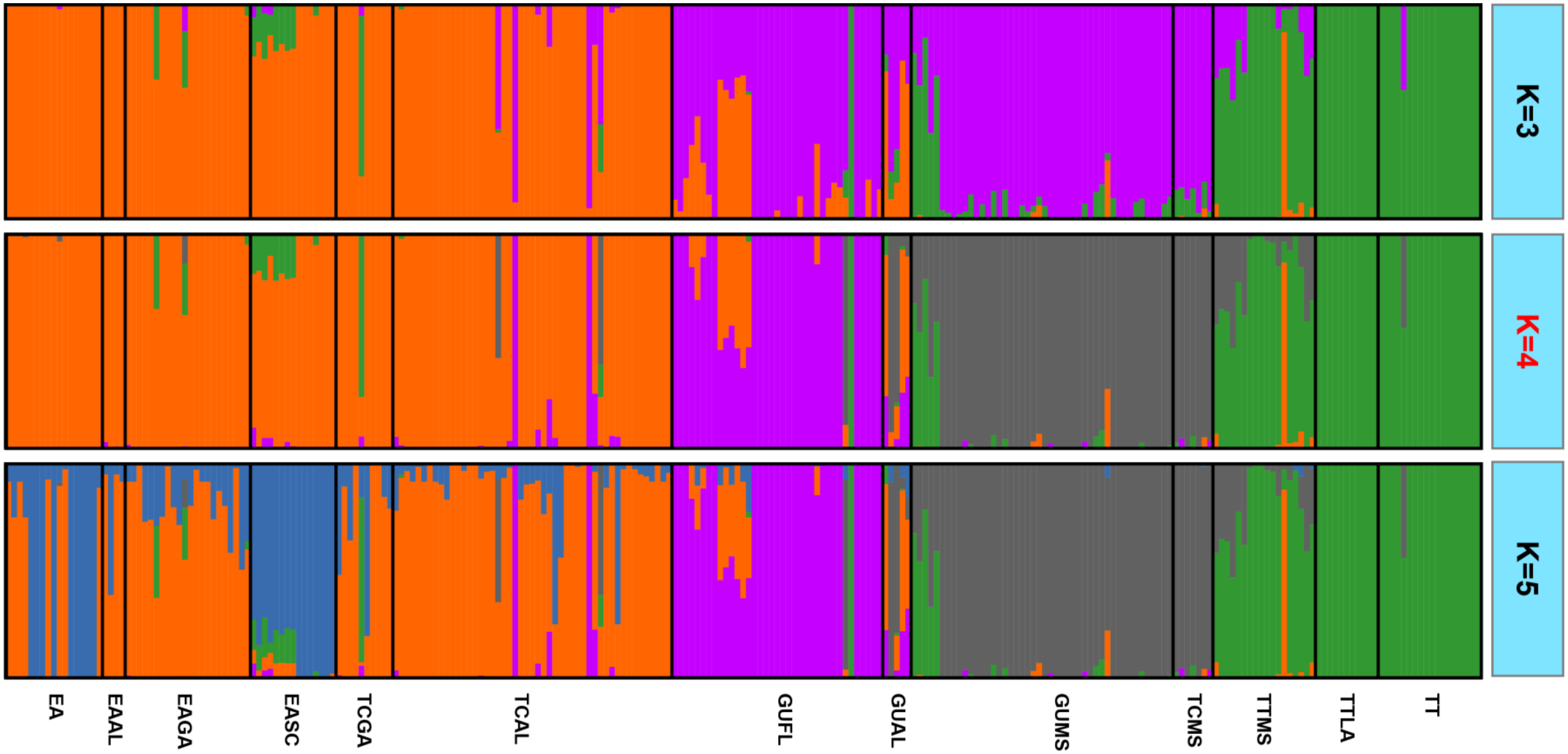


Figure S4: Top three southeastern *Terrapene* ADMIXTURE plots representing 11,308 unlinked ddRAD SNPs. The lowest cross-validation score was for *K*=4 (depicted in red), followed by *K*=3 then *K*=5. Each bar represents a unique individual, and bars with mixed colors depict admixed ancestry. The first two population code letters correspond to subspecific field identification (EA=Woodland, *T. c. carolina*; GU=Gulf Coast, *T. c. major*; TT=Three-toed, *T. m. triunguis*; TC=*Terrapene carolina*, with subspecies unidentified). The second two letters represent locality codes for U.S. states (AL=Alabama; GA=Georgia; SC=South Carolina;FL=Florida; MS=Mississippi; LA=Louisiana). Populations lacking a state code consisted of multiple localities sampled outside the hybrid zone.

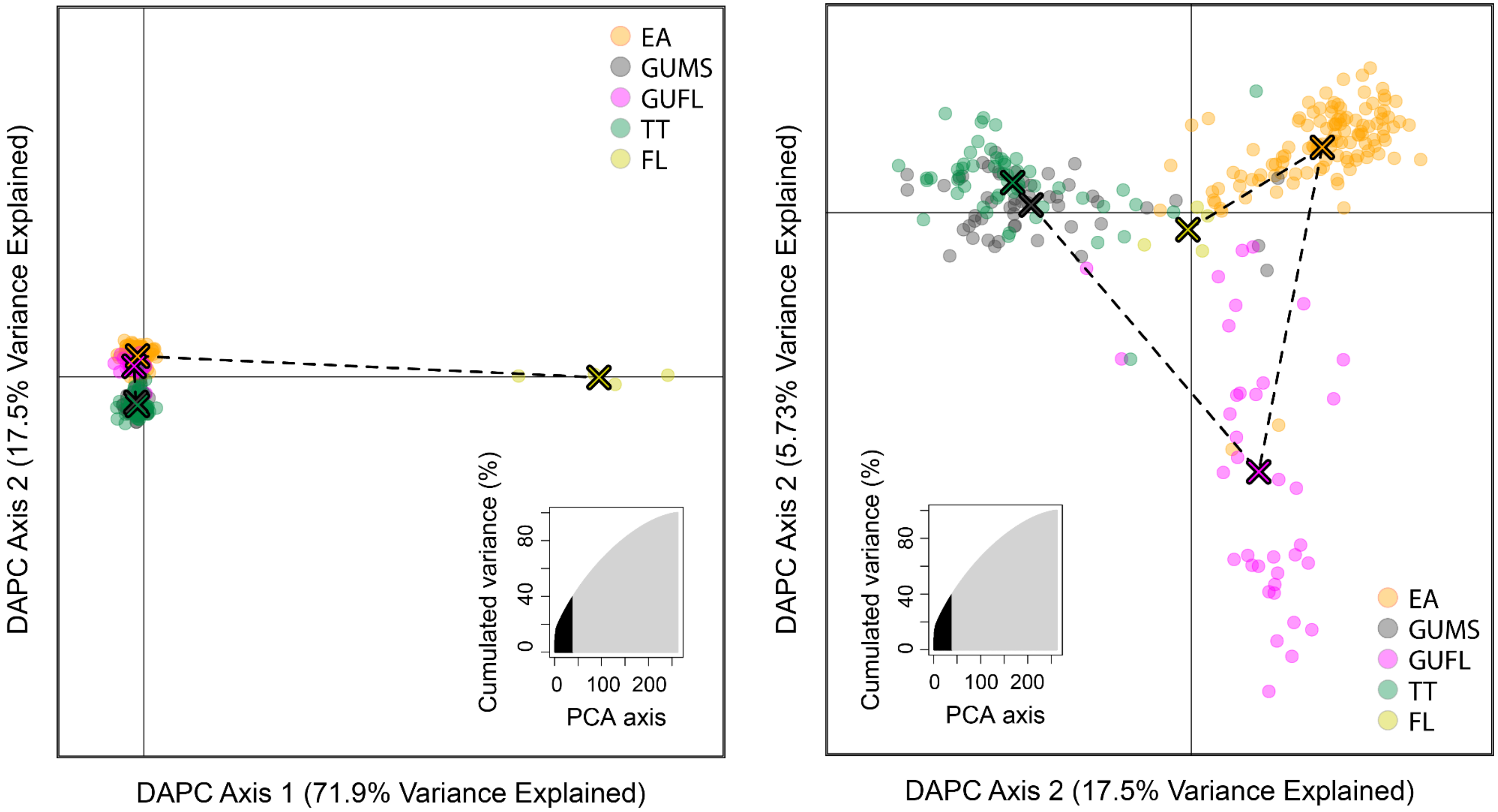


Figure S5: Discriminant Analysis of Principle Components (DAPC) for southeastern *Terrapene*. Each circle represents one individual, and each "X" delineates cluster centroids. The clusters (*K*=5, determined via Bayesian Information Criterion) represent: *T. c. carolina* (EA=Eastern), *T. c. major* (GU=Gulf Coast) from the Mississippi (GUMS) and Florida (GUFL) Panhandles, *T. m. triunguis* (TT=Three-toed), and *T. c. bauri* (FL=Florida). Inset plots demonstrate the number of retained principle components (PCs; N=40; shaded area), as determined using cross-validation (100 replicates with 90% of the dataset partitioned for training), versus the non-retained PCs (light gray area).

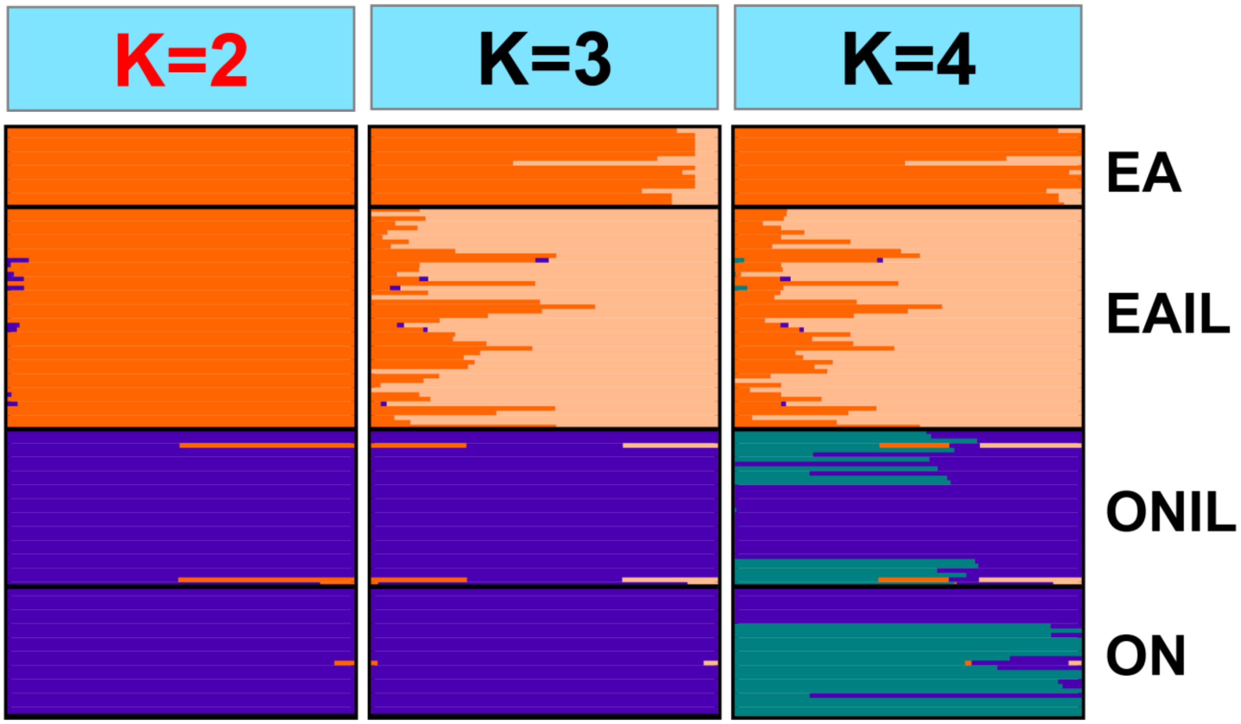


Figure S6: Top three midwestern *Terrapene* ADMIXTURE plots representing 10,338 unlinked ddRAD SNPs. The lowest cross-validation score was for *K*=2 (depicted in red at right), followed by *K*=4 then *K*=3. The first two letters of the population codes correspond to subspecific field identification (EA=Woodland, *T. c. carolina*; ON=Ornate, *T. o. ornata*). The second two letters (if present) represent locality codes for U.S. state (IL=Illinois). Populations lacking state locality code consist of multiple localities sampled outside the hybrid zone.

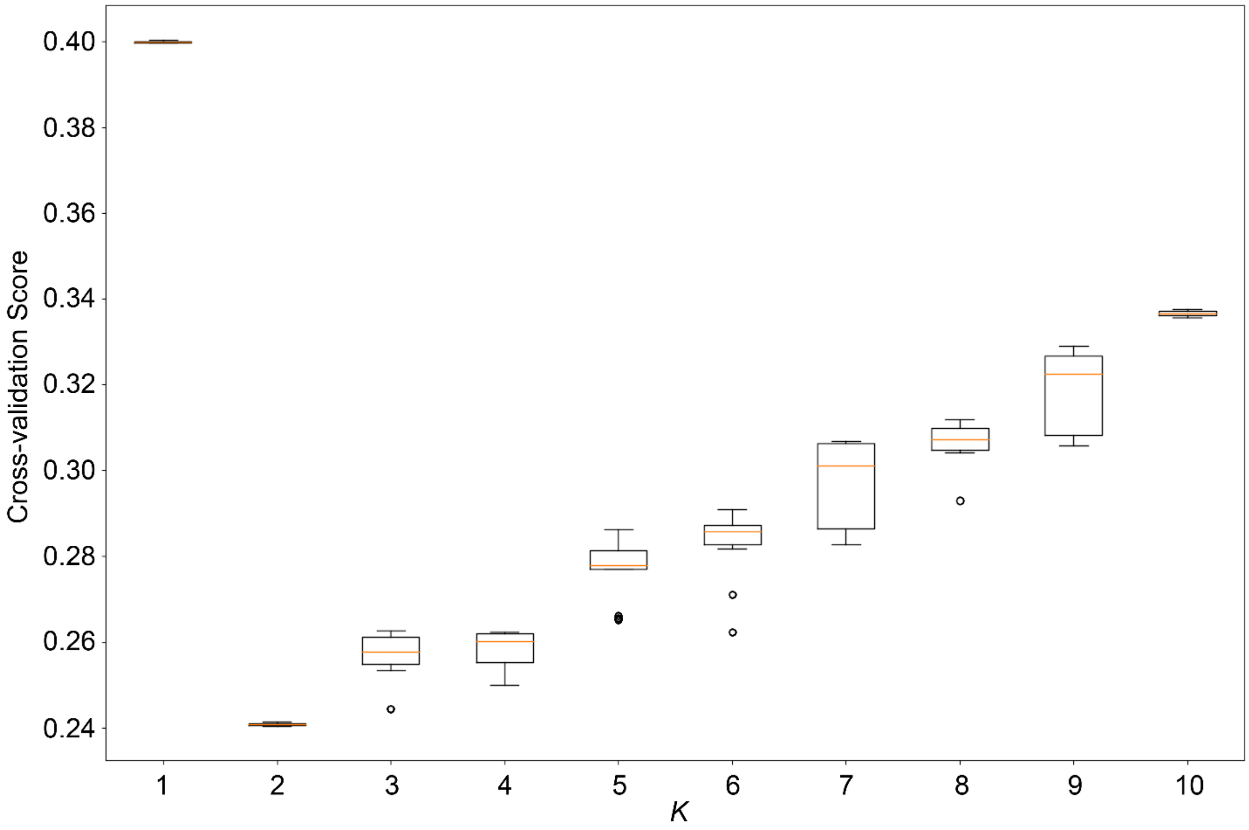


Figure S7: Cross-validation (CV) scores across all *K-values* (*K*=1-10) for ADMIXTURE runs containing samples from midwestern North America (N=135). Lower CV values indicate less error and stronger support for the corresponding *K*.

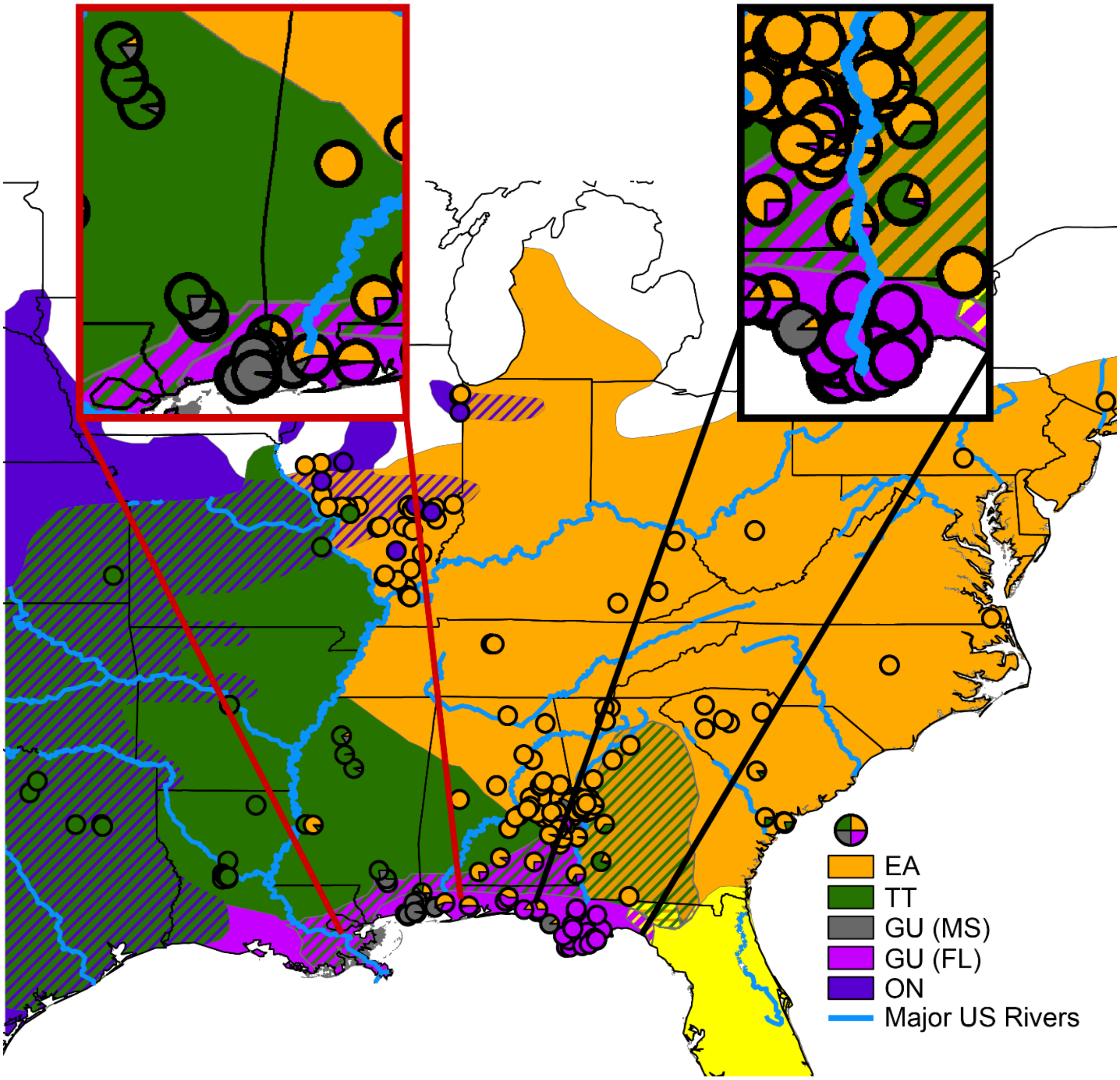


Figure S8: *Terrapene* distribution map. Cross-hatched areas represent contact zones. Circles indicate individual sampling localities, and the accompanying pie charts depict admixture proportions from the all-taxon *K*=5 (for midwestern individuals) and southeastern *K*=4 ADMIXTURE analyses (Fig. 1, 2). The expanded regions highlight two distinct *T. c. major* populations in the panhandles of Mississippi (red box) and Florida (black box), located in the Alabama and Apalachicola river basins, respectively. EA=Woodland (*T. carolina carolina*), GU=Gulf Coast (*T. c. major*), TT=Three-toed (*T. mexicana triunguis*), , ON=Ornate (*T. ornata ornata*).

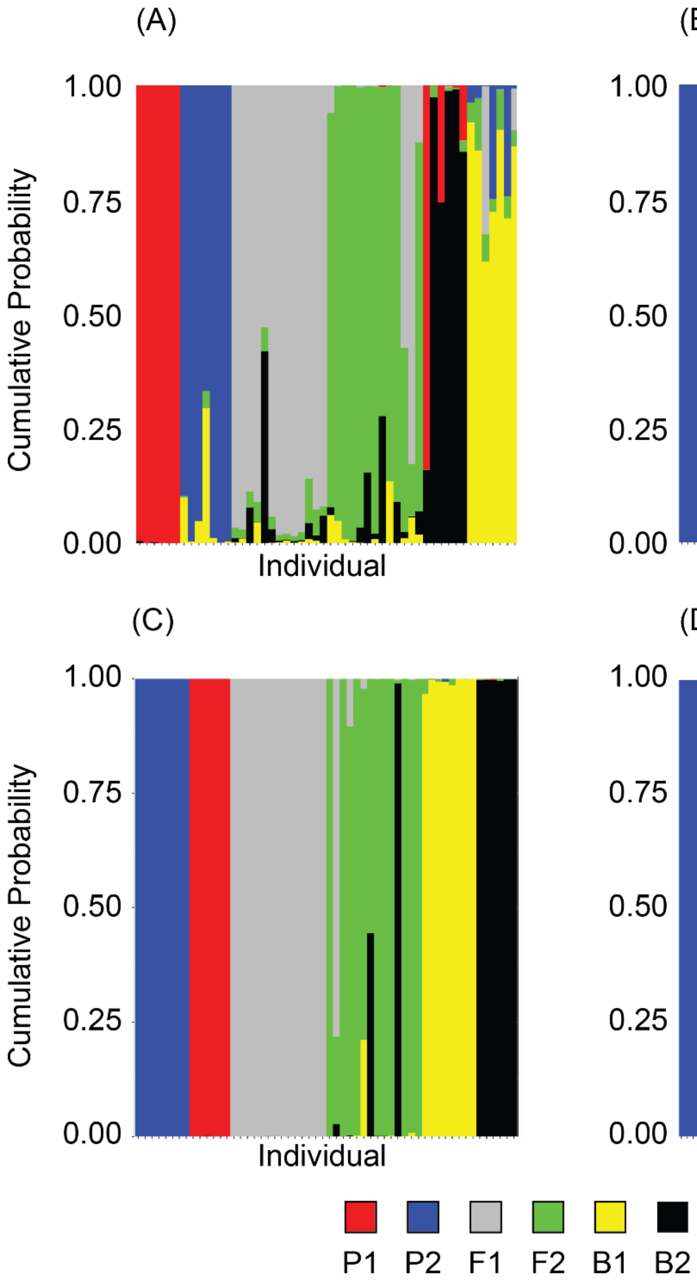


Figure S9: Results for *Terrapene* NEWHYBRIDS simulations that tested for convergence between inter- and intra-simulation replicates. Convergence was confirmed by the HYBRIDDETECTIVE pipeline, thus only one of the virtually identical simulation replicates is presented for each of (A) *T. carolina carolina* (Woodland) X *T. c. major* (Gulf Coast), (B) *T. c. carolina* X *T. mexicana triunguis* (Three-toed), (C) *T. c. major* X *T. m. triunguis*, and (D) *T. c. carolina* X *T. o. ornata* (Ornate) is shown here. The genotype frequencies included parental groups (P_1 and P_2), first and second-generation hybrids (F_1 and F_2), and backcross (B_1 and B_2) generations.

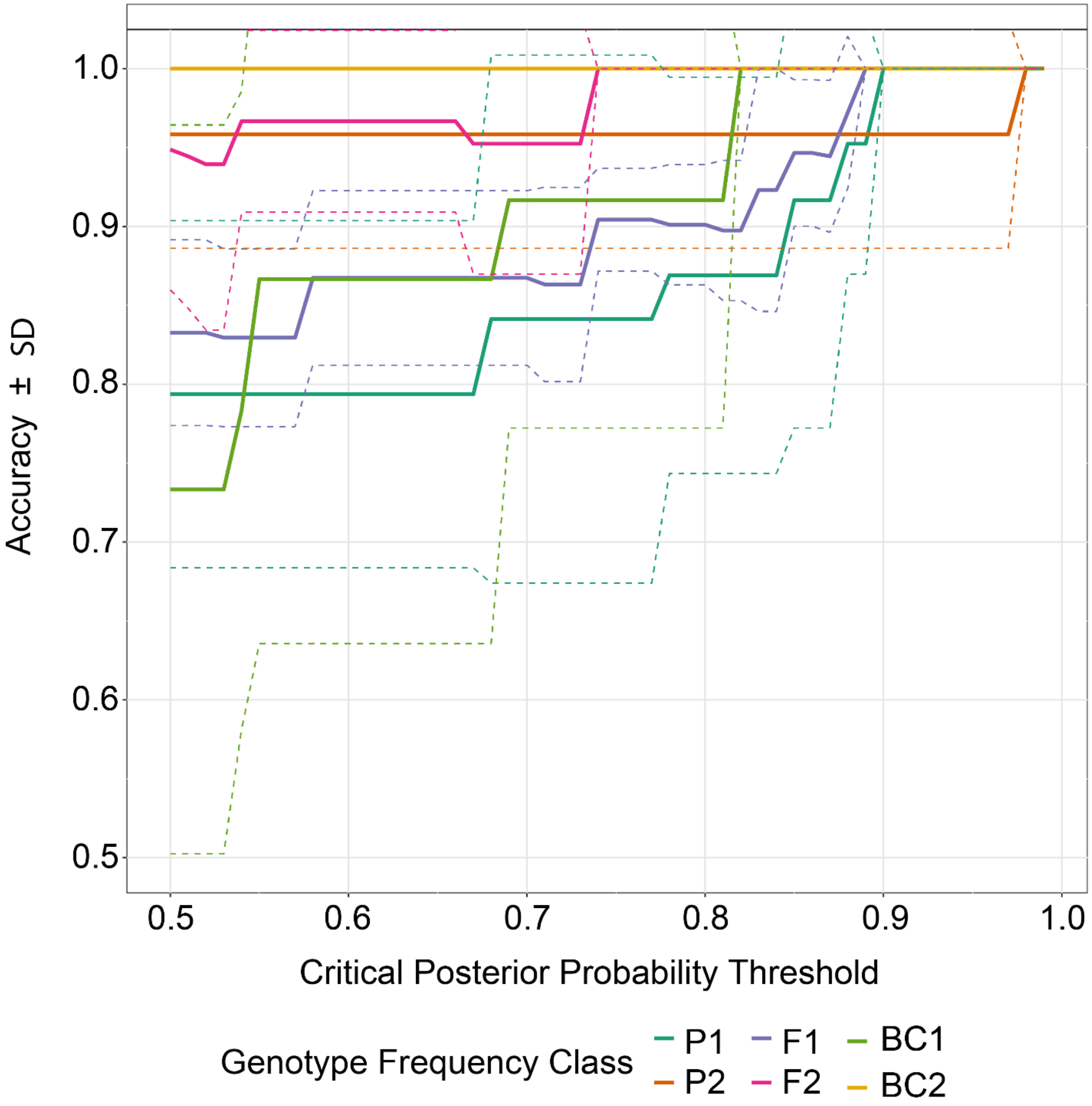


Figure S10: NEWHYBRIDS power analysis for *Terrapene carolina carolina* (Woodland) X *T. c. major* (Gulf Coast; EAxGU) showing predicted accuracy plotted against posterior probability thresholds. Accuracy was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

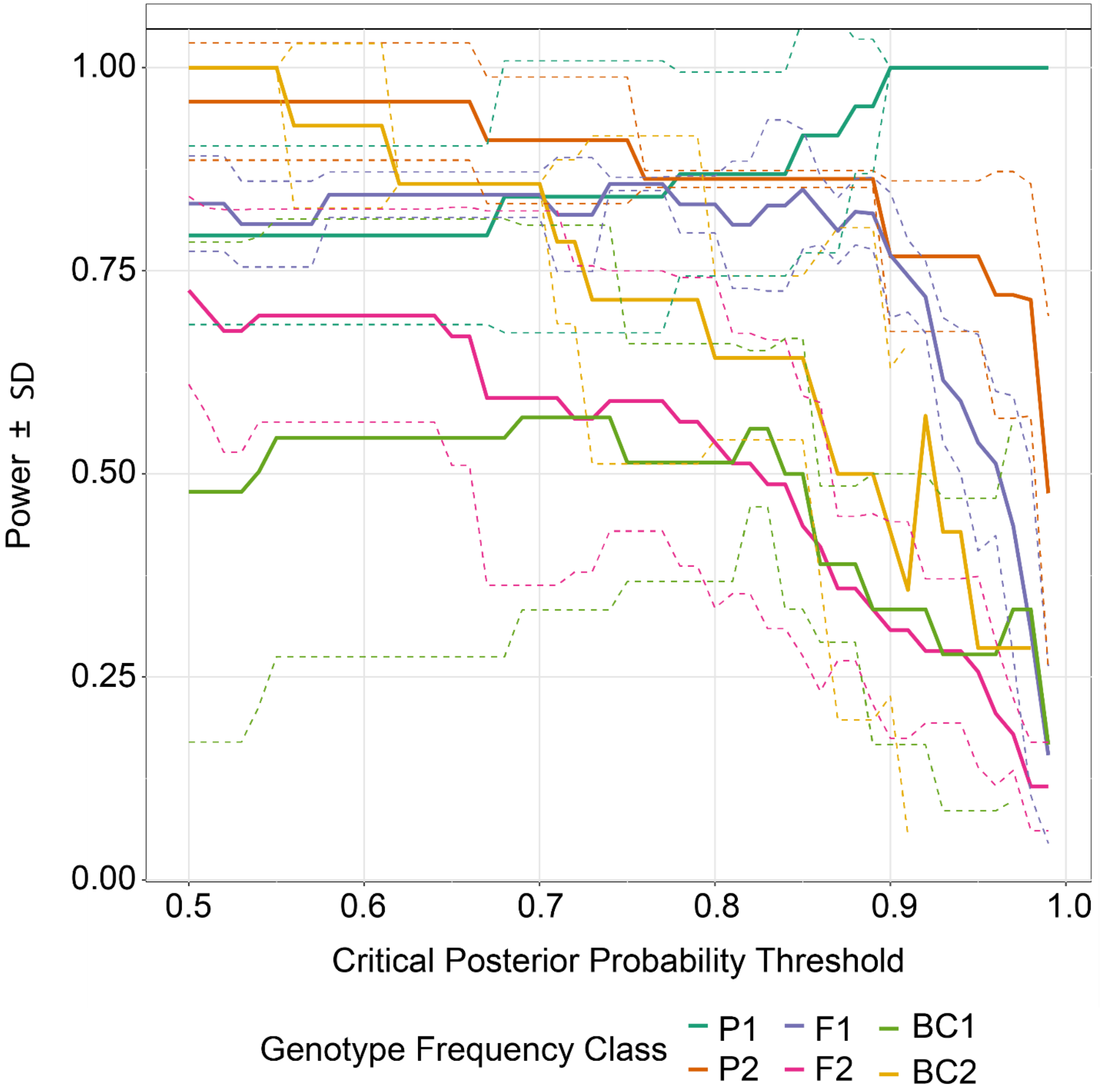


Figure S11: NEWHYBRIDS power analysis for *Terrapene carolina carolina* (Woodland) X *T. c. major* (Gulf Coast; EAxGU) showing predicted power plotted against posterior probability thresholds. Power was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

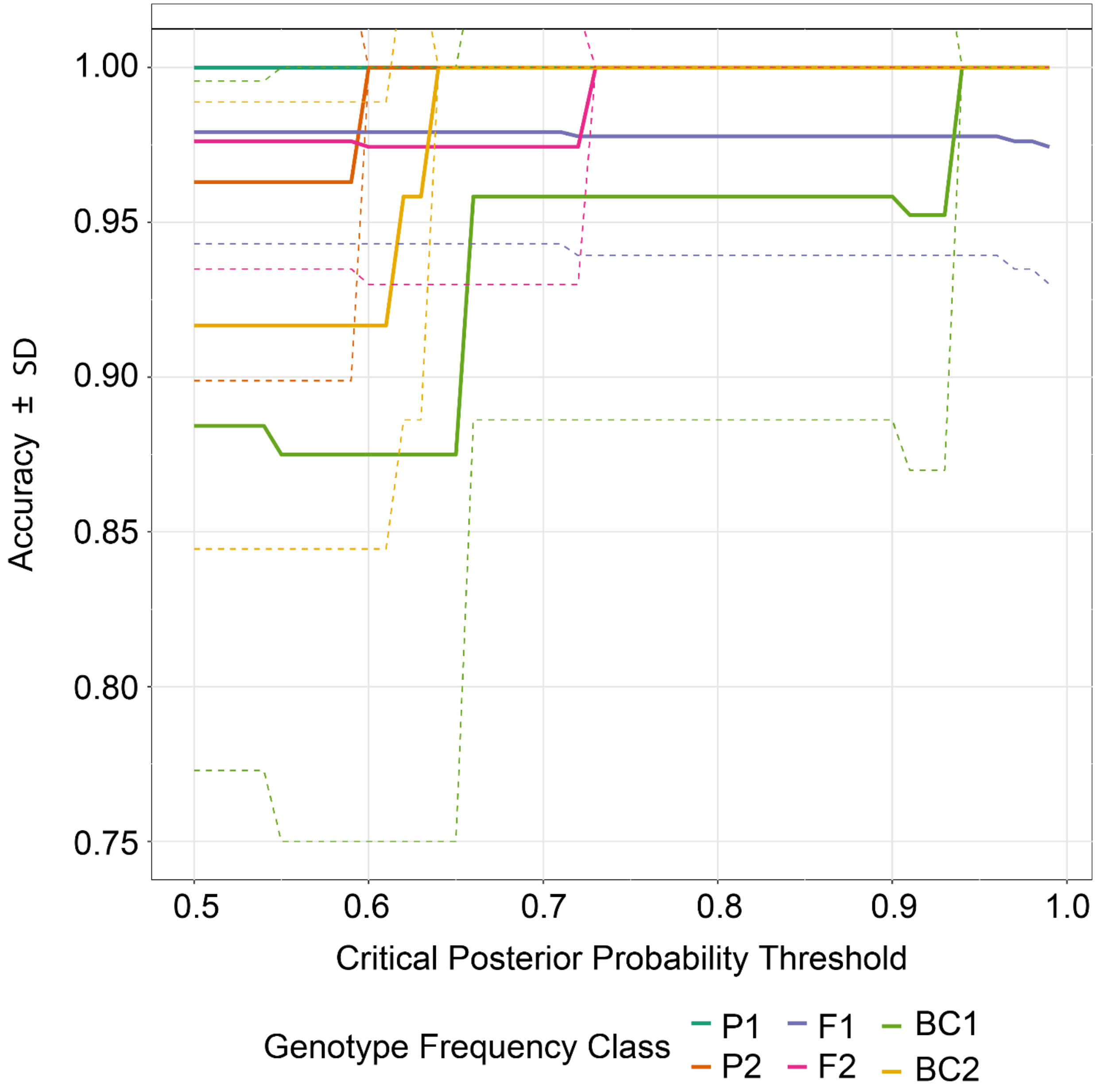


Figure S12: NewHybrids power analysis for *Terrapene carolina carolina* (Woodland) X *T. mexicana triunguis* (Three-toed; EAxTT) showing predicted accuracy plotted against posterior probability thresholds. Accuracy was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

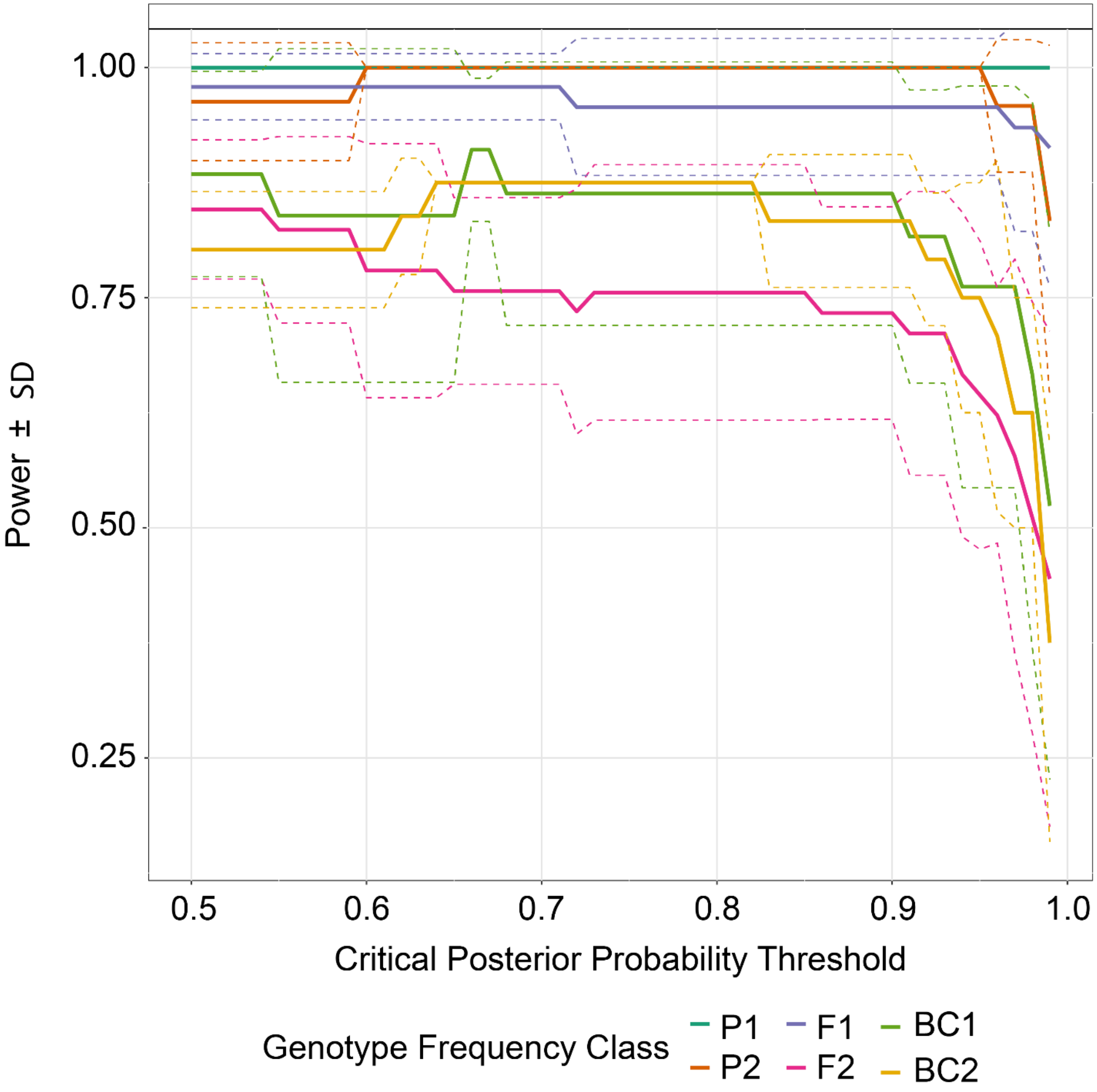


Figure S13: NewHybrids power analysis for *Terrapene carolina carolina* (Woodland) X *T. mexciana triunguis* (Three-toed; EAxTT) showing predicted power plotted against posterior probability thresholds. Power was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

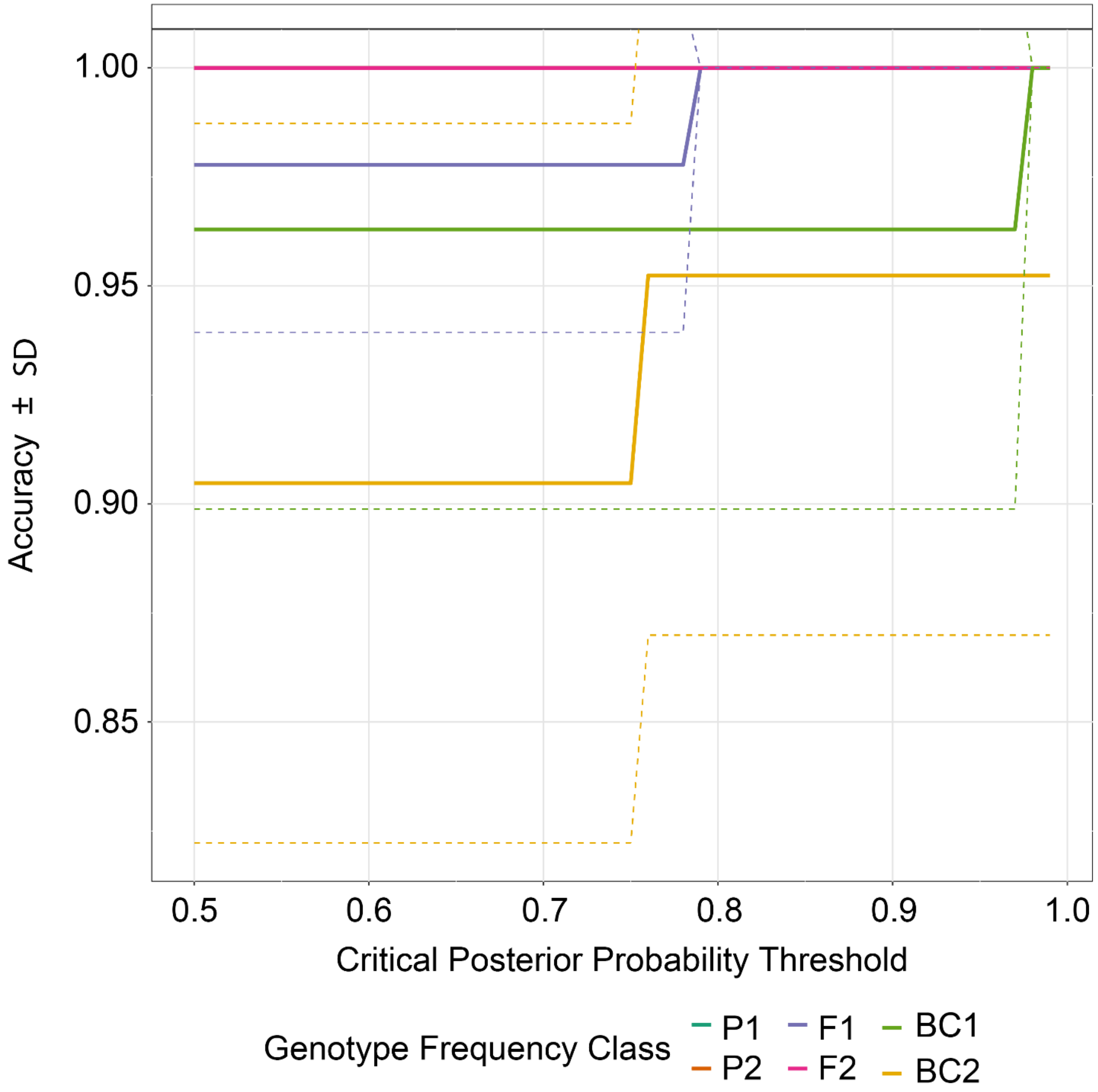


Figure S14: NEWHYBRIDS power analysis for *Terrapene carolina major* (Gulf Coast) X *T. mexicana triunguis* (Three-toed; GUxTT) showing predicted accuracy plotted against posterior probability thresholds. Accuracy was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

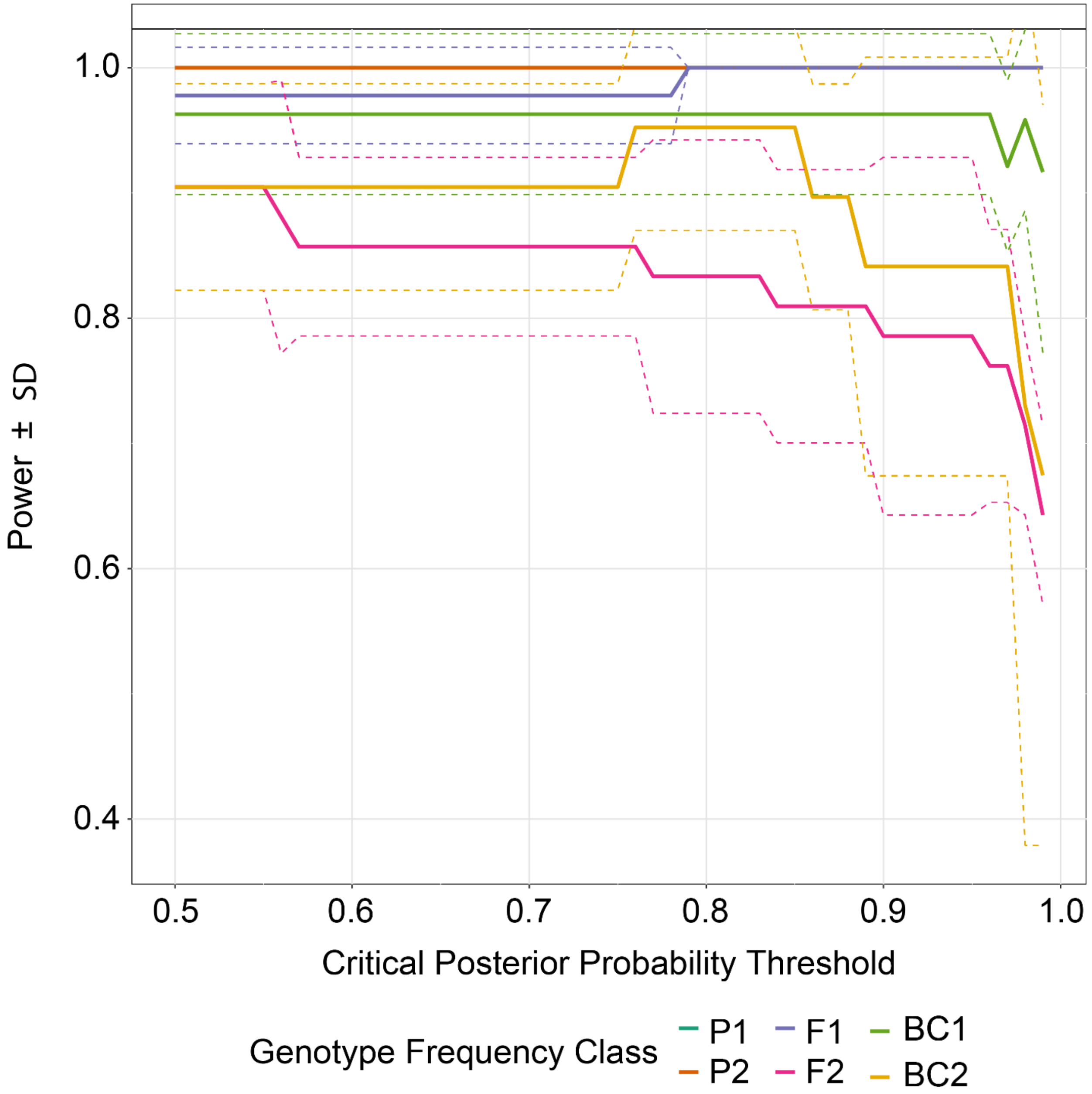


Figure S15: NEWHYBRIDS power analysis for *Terrapene carolina major* (Gulf Coast) X *T. mexicana triunguis* (Three-toed; GUxTT) showing predicted power plotted against posterior probability thresholds. Power was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

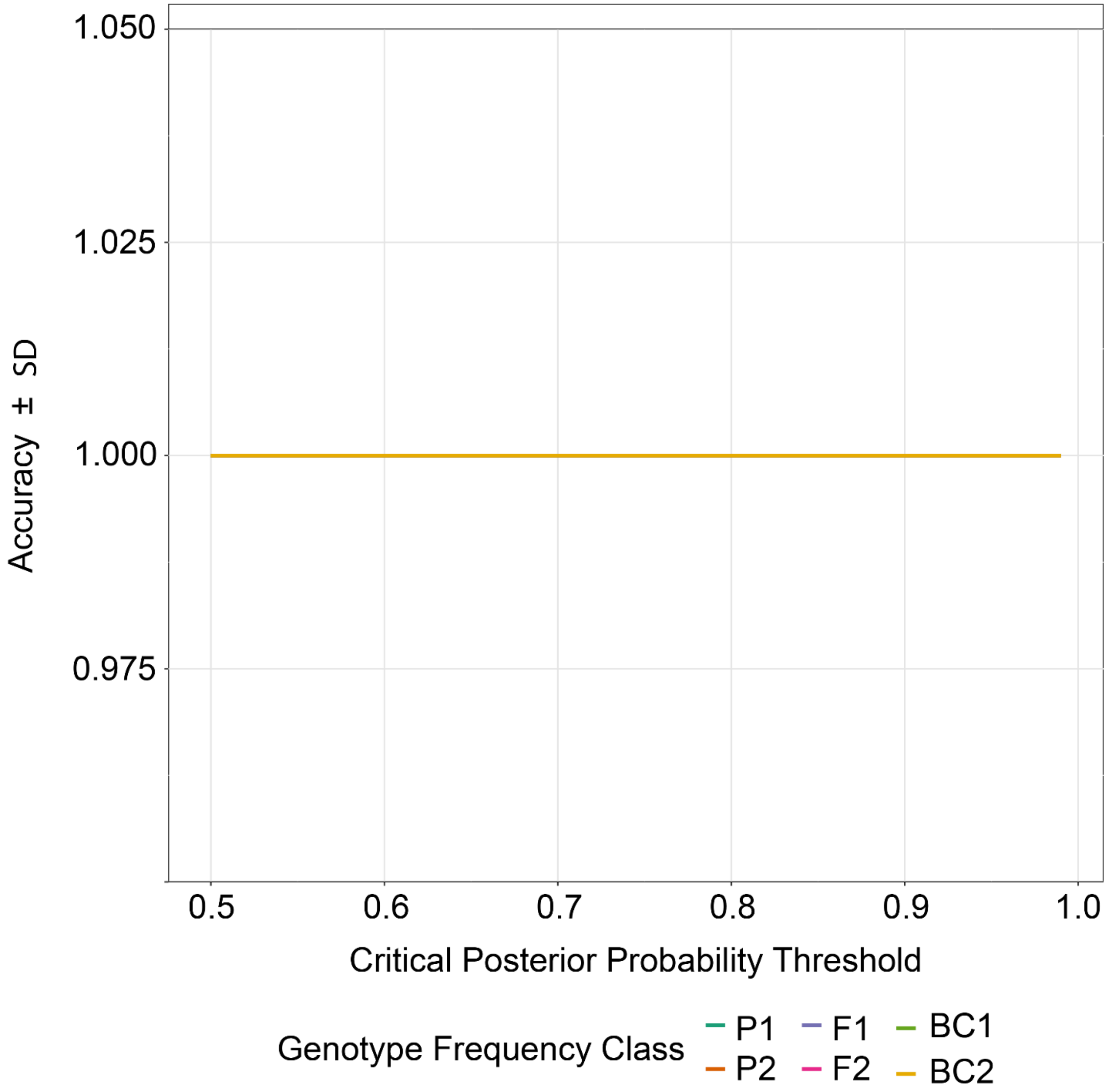


Figure S16: NEWHYBRIDS power analysis for *Terrapene carolina carolina* (Woodland) X *T. ornata ornata* (Ornate; EAxON) showing predicted accuracy plotted against posterior probability thresholds. Accuracy was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

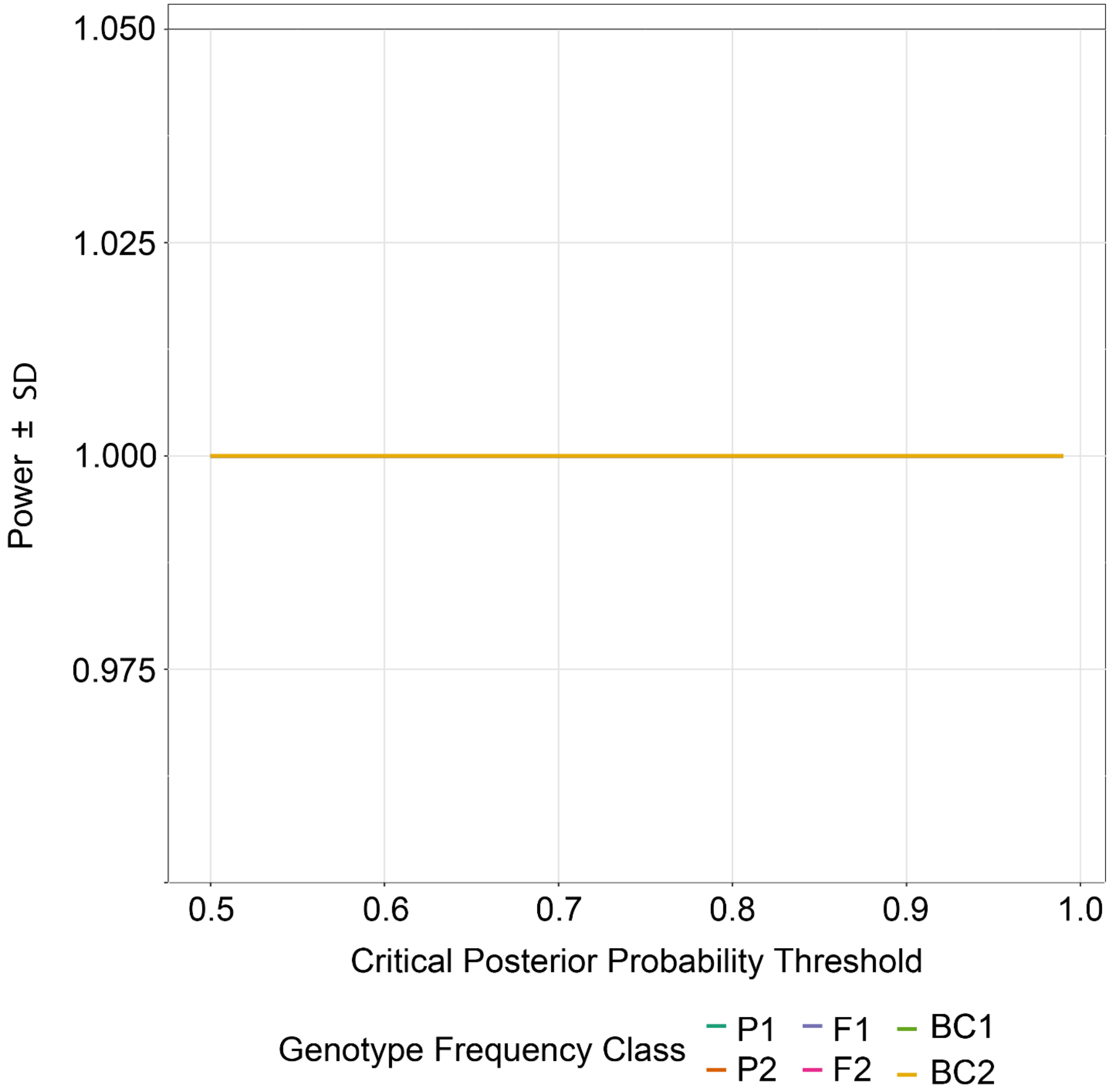


Figure S17: NEWHYBRIDS power analysis for *Terrapene carolina carolina* (Woodland) X *T. ornata ornata* (Ornate; EAxON) showing predicted power plotted against posterior probability thresholds. Power was calcuated using simulated datasets for parental (P_1 and P_2), first and second-generation hybrid (F_1 and F_2), and backcross (B_1 and B_2) generations.

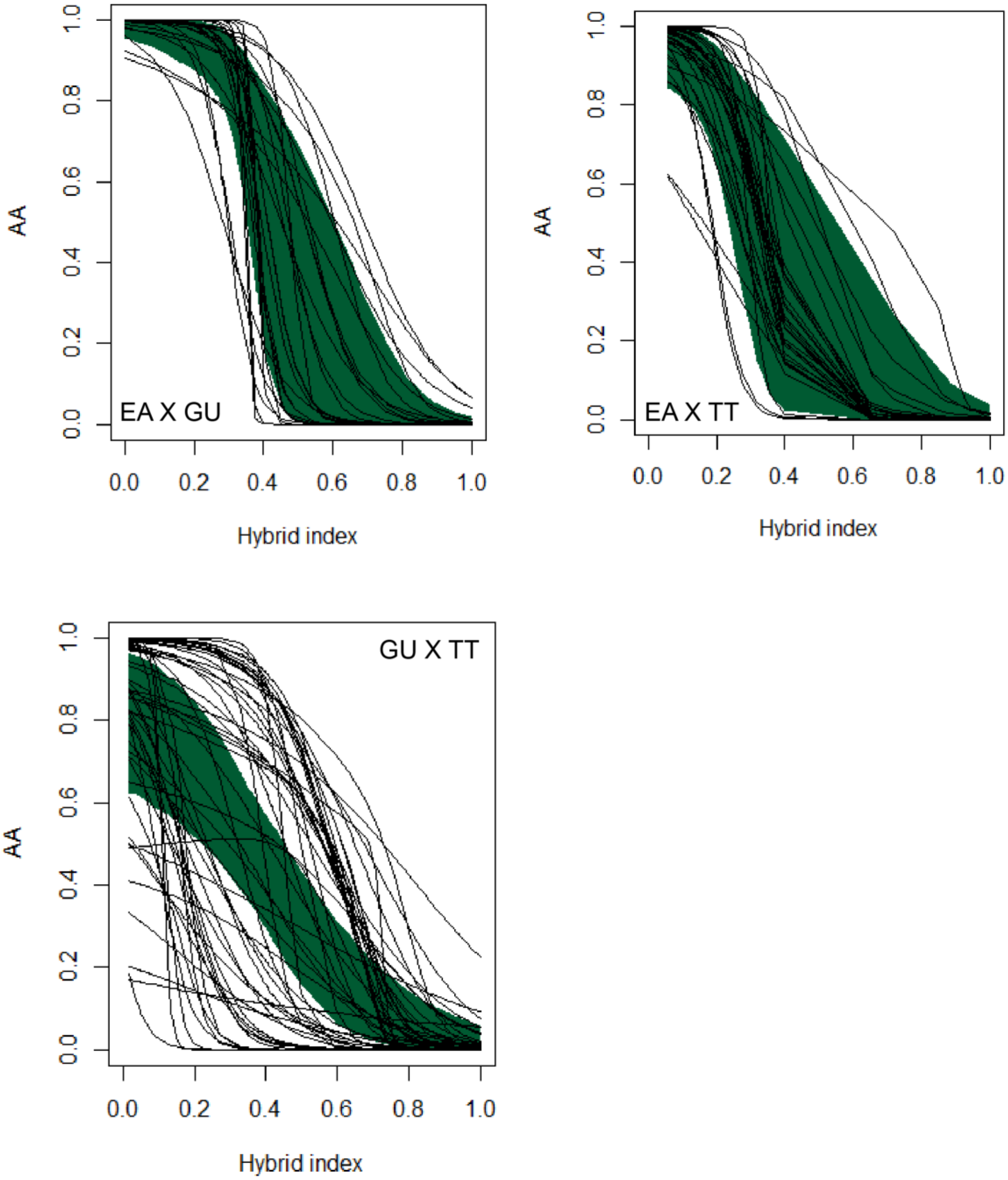


Figure S18: Genomic clines depicting outlier SNPs for all *Terrapene* ddRAD loci. Pairwise comparisons are between *T. carolina carolina* (EA=Woodland), *T. c. major* (GU=Gulf Coast), and *T. mexicana triunguis* (TT=Three-toed), with the number of loci per comparison being: N=10,106 (EAxGU); N=11,390 (EAxTT); and N=10,786 (GUxTT). The dark green area represents null expectations and each line is a genomic cline at one outlier locus. In each analysis, the P_1 genotype represents EA, EA, and GU, respectively.

CHAPTER II

The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (*Terrapene* spp.)

Bradley T. Martin, Tyler K. Chafin, Marlis R. Douglas, John S. Placyk Jr., Roger D. Birkhead, Christopher A. Phillips, Michael E. Douglas

ABSTRACT

Model-based approaches that attempt to delimit species are hampered by computational limitations as well as the unfortunate tendency by users to disregard algorithmic assumptions. Alternatives are clearly needed, and machine-learning (M-L) is attractive in this regard as it functions without the need to explicitly define a species concept. Unfortunately, its performance will vary according to which (of several) bioinformatic parameters are invoked. Herein, we gauge the effectiveness of M-L-based species-delimitation algorithms by parsing 64 variably-filtered versions of a ddRAD-derived SNP dataset collected from North American box turtles (*Terrapene* spp.). Our filtering strategies included: (A) minor allele frequencies (MAF) of 5%, 3%, 1%, and 0% (=none), and (B) maximum missing data per-individual/per-population at 25%, 50%, 75%, and 100% (=no filtering). We found that species-delimitation via unsupervised M-L impacted the signal-to-noise ratio in our data, as well as the discordance among resolved clades. The latter may also reflect biogeographic history, gene flow, incomplete lineage sorting, or combinations thereof (as corroborated from previously observed patterns of differential introgression). Our results substantiate M-L as a viable species-delimitation method, but also

demonstrate how commonly observed patterns of phylogenetic discordance can seriously impact M-L-classification.

1. INTRODUCTION

Species are recognized as the currency of biodiversity, yet defining what constitutes a species has been hampered by subjective interpretations. This in turn creates downstream issues for conservation (Mace 2004), where spurious ‘splitting’ or ‘lumping’ impede an equitable allocation of limited resources. Although genomic approaches based on the multispecies coalescent (MSC) are promising and have been commonly applied to the species problem (Allendorf *et al.* 2010), conflicting genome-wide signals are widely apparent due to incomplete lineage sorting (ILS) and gene flow (Funk & Omland 2003). Two MSC methods, BPP and BFD* (Yang & Rannala 2010; Leaché *et al.* 2014), seemingly over-split in the presence of strong population structure (Sukumaran & Knowles 2017) or with continuous geographic distributions (Chambers & Hillis 2019). Both are also computationally limited when applied to large datasets. As model complexity and data expand concomitantly, so also do: 1) efforts required to computationally explore appropriate parameter space; and 2) the probabilities that models fail to accommodate process. Herein, we explore alternative approaches for the parsing of high-dimensionality data by evaluating the performance of recently developed machine-learning (M-L) algorithms and classificatory approaches in successfully adjudicating variably-filtered versions of a ddRAD-derived SNP dataset.

‘Unsupervised’ machine learning methods (UML) are of particular interest for group delimitation, in that they do not require *a priori* designations to train the classification model. Several UML classifiers lend themselves to species delimitation, including: Random Forest (RF; Breiman 2001), t-distributed stochastic neighbor embedding (T-SNE; Maaten & Hinton 2008), and variational autoencoders (VAE; Kingma & Welling 2013). Each has distinct advantages: RF uses randomly replicated data subsets to develop ‘decision trees’ that are subsequently aggregated (='forest'), with classificatory decisions parsed as a majority vote. The random subsetting approach is robust to correlations among features (=summary statistics or principal components used for prediction) as well as model overfitting (i.e., over-training the model such that it does not generalize to new data). One stipulation is that features must lack undue noise (Rodriguez-Galiano *et al.* 2012). By contrast, T-SNE creates clusters in reduced-dimension space, typically a 2D plane distilled from multi-dimensional data, and as such conceptually resembles principal components analysis (Maaten & Hinton 2008). On the other hand, VAE employs neural networks to 'learn' patterns within multidimensional data extracted from a compressed, low-dimensionality (='encoded') representation. Again, an ordination technique is simulated but without imposing linear/orthogonal constraints, such that a statistically interpretable result emerges that is appropriate for highly-complex data (Derkarabetian *et al.* 2019).

Some algorithms are robust to gene flow (Derkarabetian *et al.* 2019; Newton *et al.* 2020; Smith & Carstens 2020), yet a greater number of tests must be performed across diverse systems so as to understand which parameters impinge upon performance. Potentials include: Data quantity (Newton *et al.* 2020), the proportion of missing data (Mussmann *et al.* 2020), and evolutionary complexity (Austerlitz *et al.* 2009). Here, we employ M-L algorithms alongside coalescent methods such as BFD* (Leaché *et al.* 2014) as vehicles to parse a taxonomically recalcitrant clade. Included algorithms are: Process-based RF (DELIMITR; Smith *et al.* 2017; Smith & Carstens 2020) and unsupervised RF, T-SNE, and VAE, as implemented in Derkarabetian *et al.* (2019).

1.1. Species concepts and their evolution in *Terrapene*

North American box turtles (Emydidae: *Terrapene*) are a primarily terrestrial group that includes 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). *Terrapene carolina* is split into two subspecies east of the Mississippi River and south through the Gulf Coast [Woodland (*T. c. carolina*) and Gulf Coast (*T. c. major*); Figure 1]. *Terrapene mexicana* contains three subspecies: Three-toed (*T. m. triunguis*); Mexican (*T. m. mexicana*); and Yucatan (*T. m. yucatana*) that range across southeastern and midwestern United States, the Mexican state of Tamaulipas, and the Yucatan Peninsula. Ornate (*T. ornata ornata*) and Desert (*T. o. luteola*) inhabit the Midwest and Southwest U.S. and Northwest México, while 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), while Florida box turtle occurs in Peninsular Florida.

Morphological analyses delineate *T. carolina*/*mexicana* as a single species, sister to *T. coahuila* (Minx 1992, 1996), as supported by genetic studies (Feldman & Parham 2002; Stephens & Wiens 2003). Martin et al. (2013) elevated *T. mexicana*, and nested *T. coahuila* within *T. carolina*. *Terrapene carolina carolina* is sister to *T. c. major*/*T. coahuila*, although gene flow was suspected with *T. c. major*. *Terrapene carolina major* was recently demoted to an intergrade with subsequent loss of subspecific status (Butler *et al.* 2011; Iverson *et al.* 2017). However a recent genomic study supported pure *T. c. major* populations in Florida and Mississippi (Martin *et al*. 2020). Similarly, *T. bauri* (formerly *T. carolina bauri*) was recently elevated (Butler *et al.* 2011; Iverson *et al.* 2017), but more substantial evidence is needed (Martin *et al.* 2013). For clarity, we retain the nomenclature of Martin *et al.* (2013, 2014), with *T. c. major* and *bauri* representing *T. carolina* subspecies.

One explanation for the enigmatic classification of *T. carolina* and *T. mexicana* involves hybridization (Auffenberg 1958, 1959; Milstead & Tinkle 1967; Milstead 1969). Some researchers (Fritz & Havaš 2013, 2014) interpreted reproductive semi-permeability as justification sufficient to collapse the southeastern taxa. However, their classificatory status must be re-examined, as indicated by results modulating the species boundaries of southeastern *Terrapene* (Martin *et al.* 2020).

Taxonomic disputes in *Terrapene* highlight the philosophical disparity among species definitions [e.g., biological (Mayr 1963) versus phylogenetic (Eldredge & Cracraft 1980)]. The approach advocated herein acknowledges that operational criteria among concepts are intimately related. Specifically, reproductive barriers (through time) beget genealogical concordance, while contemporary evaluations of gene flow are contextualized via phylogenetic/phylogeographic perspectives (Avise 2000a; b). We thus subscribe to a 'unified species concept' (De Queiroz 2007) wherein the primary criterion for formal taxonomic rank is the existence of evolutionary lineages (e.g., as distinct metapopulations), with evidence via reproductive isolation, phylogenetic-phylogeographic resolution, and phenotypic adaptation, with all acknowledged as being inherently linked. Here, our clustering and classificatory approaches define molecular diagnosability, and as such variably place *Terrapene* lineages along a speciation continuum (Via 2009; Nosil & Feder 2012; Edwards *et al.* 2016; Martin *et al.* 2020).

2. MATERIALS AND METHODS

2.1. DNA extraction and library preparation

Tissue samples were obtained from museums, agencies, and volunteers (Supplementary Information Table S1) and stored at -20°C. Genomic DNA was extracted via spin-column kits: DNeasy Blood and Tissue (QIAGEN), QIAamp Fast DNA (QIAGEN), and E.Z.N.A. Tissue DNA Kits (Omega Bio-tek). Extracted DNA was quantified using Qubit fluorometry (Thermo Fisher Scientific), and characterized using gel electrophoresis on 2% agarose.

Samples were processed via ddRADseq (Peterson *et al.* 2012), with ~500-1,000ng of genomic DNA/sample digested with PstI and MspI at 37°C for 24 hours. Samples were bead-purified (Beckman-Coulter) at 1.5X concentration then standardized at 100ng. Barcoded adapters were ligated before pooling 48 samples per library. Taxa were spread across libraries to mitigate batch effects then size-selected (454-509 bp, including ligated adapters) on a Pippin Prep (Sage Science). Adapter-extension was performed via twelve-cycle PCR, followed by 1×100 sequencing on the Illumina Hi-Seq 4000 (University of Oregon/Eugene), with two indexed libraries pooled/lane.

2.2. Quality control and assembly

FASTQCv.0.11.5 was used to assess sequence quality (Andrews 2010), with raw reads demultiplexed via IPYRAD v.0.7.28 (Eaton & Overcast 2020), allowing for one barcode mismatch as a maximum. Low quality sequences (>5 bases with Q<33) and adapters were removed. Assembly was reference-guided using *Terrapene mexicana* (GCA_002925995.2), with unmapped reads discarded. To reduce error, only loci exhibiting ≥20X coverage were retained

(Nielsen *et al.* 2011). We also excluded loci with excessive heterozygosity (≥75% of individual SNPs), <50% global occupancy, or >two alleles/sample.

2.3. Phylogenomic inferences

F_1 and F_2-generation hybrids previously identified in a population-level analysis (Martin *et al.* 2020) were excluded as a means of mitigating impacts of contemporary gene flow on species tree inference (Long & Kubatko 2018). We then employed SVDQUARTETS (Chifman & Kubatko 2014) filtered to one SNP per locus to reduce linkage bias, with exhaustive quartet sampling and 100 bootstrap pseudo-replicates. Taxon partitions were grouped by subspecies and U.S./Mexican state locality, with *Emydoidea blandingii* and *Clemmys guttata* as outgroups.

We also employed a polymorphism-aware model (POMO: Schrempf *et al.* 2016), as implemented in IQ-TREE v1.6.9 (Nguyen *et al.* 2015), with full-locus alignments and 1,000 ultrafast bootstrap (UFBOOT) replicates (Hoang *et al.* 2017). The maximum virtual population size was 19, with discrete gamma-distributed rates=4.

Using ten-thousand re-samplings, we performed topology tests (IQ-TREE) with seven statistical criteria on the SVDQUARTETS and POMO trees, as well as a previously published morphological (Minx 1996) and a molecular hypothesis (Martin *et al.* 2013). Additional details are in Supplementary Information Appendix A.1.1.

A lineage tree was generated (IQ-TREE v2.0.6; Minh *et al.* 2020) and full-locus partitions merged (Chernomor *et al.* 2016), with the top 10% of combinations employed and a per-partition model search (MODELFINDER: Kalyaanamoorthy *et al.* 2017). Node support was assessed using 1,000 UFBOOT replicates and site-wise concordance factors (SCF; Minh *et al.* 2018). The SCF values were calculated from 10,000 randomly sampled quartets.

2.4. Divergence dating

A full concatenation tree was time-calibrated via least square dating (LSD2), as implemented in IQ-TREE (To *et al.* 2016). Four fossil calibration points were used (Holman & Fritz 2005; Spinks & Shaffer 2009), including the following most recent common ancestors (MRCAs): (1) *T. ornata* and *T. carolina/T. mexicana*, minimally constrained to 13 million years ago (Mya); (2) *T. o. ornata* and *T. o. luteola* (9.0-13.0 Mya); (3) *T. carolina* and *T. mexicana* (9.0-11.0 Mya); and (4) *Terrapene* and *Clemmys*/*Emydoidea* [(maximally constrained to 29.4 Mya) (per Martin *et al.* 2013)]. Branch lengths were simulated from a Poisson distribution with 1,000 replicates to assess 95% confidence intervals.

2.5. Species delimitation using BFD*

We employed Bayes Factor Delimitation (BFD*; Leaché *et al.* 2014) as a comparative baseline. Given its computationally-intense process, each taxon was subset to a maximum of five individuals containing the least missing data (N=37+outgroups). Sites with >50% missing data in any population were removed (see Supplementary Information Appendix A.2.1 for prior selection and data formatting steps for BFD*).

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

2.6. Preparing and executing UML datasets

To assess the influence of bioinformatic choices on M-L species delimitation, we performed missing data filtering sweeps to produce 64 datasets across three filtering options. Missing data was filtered per-individual and per-population, with the maximum permitted occupancy set to 25%, 50%, 75%, and no filtering (=100%). Datasets were also filtered by minor allele frequency (MAF) at values of 5%, 3%, 1%, and 0% (=no MAF filter). Custom scripts were employed for all filtering steps (https://github.com/tkchafin/scripts).

RF and T-SNE (Breiman 2001; Maaten & Hinton 2008) were executed and visualized using an R script [Derkarabetian *et al.* (2019); https://github.com/shahanderkarabetian/uml_species_delim]. We ran 100 replicates for each of the 64 datasets, with data subsequently represented as scaled principal components (ADEGENETv2.1.1; Jombart & Ahmed 2011) in Rv3.5.1 (R Development Core Team 2018). To generate RF predictions, we averaged 10,000 majority-vote decision trees. Clustered RF output was visualized using both classic and isotonic multidimensional scaling (CMDS and ISOMDS; Shepard *et al.* 1972; Kruskal & Wish 1978). We ran T-SNE for 20,000 iterations, with equilibria of the clusters visually observed. Perplexity, which limits the effective number of T-SNE neighbors, was subjected to a grid search with values from 5-50, incremented by five.

VAE (Derkarabetian *et al.* 2019) employs neural networks to infer the marginal likelihood distribution of sample means (μ) and standard deviations [(σ) (i.e. 'latent variables')]. As with RF and T-SNE analyses, VAE was also run with 100 replicates to assess cluster stochasticity. Each of the 64 datasets were split into 80% training/20% validation datasets using the *train_test_split* module (*scikit-learn*: Pedregosa *et al.* 2011), with model loss (~error) visualized to determine the optimal number of 'epochs' (=cycles through the training dataset). VAE should ideally be terminated when loss converges on a minimal difference between training and validation datasets [the 'Goldilocks zone'; Supplementary Information Figure S1 (Al'Aref *et al.* 2019)].

Overfitting is indicated when model loss in the validation dataset escalates, whereas underfitting is a failure to reach minimum points (=inability to generalize to unseen data). Thus, we added minor modifications to the original Python script (Derkarabetian *et al.* 2019) by implementing an early stopping callback (*keras.callbacks* Python module; Chollet 2015), which terminates training when model loss fails to improve for 50 epochs, then restores the best model prior to the tolerance period (see Supplementary Information Appendix A.2).

2.7. K-selection for RF, t-SNE, and VAE

Two clustering algorithms (R-scripts: Derkarabetian *et al.* 2019), were used to identify clusters and derive optimal *K* for RF and T-SNE analyses. The first [Partitioning Around Medoids (PAM); Kaufman and Rousseeuw 1987] minimizes the distance of intra-cluster points to a centroid. The program requires *K* to be defined *a priori*, and thus *K*=1-10 were tested. The second (hierarchical clustering, HC; Fraley & Raftery 1998) iteratively merges points with minimal dissimilarity. After clustering, optimal *K* was chosen using the gap statistic (GS) and highest mean silhouette width [HMSW; Rousseeuw (1987), Tibshirani *et al.* (2001)].

VAE used DBSCAN (Ester *et al.* 1996), as implemented in a custom Python script (*vae_dbscan.py*), to derive clusters using a distance threshold (ε) rather than *a priori* setting of *K*.

Here we used 2 × the standard deviation, but averaged globally across all samples (following Derkarabetian *et al.* 2019).

For plotting, we implemented a permutation-based heuristic search to align *K* across all replicates and the 64 datasets [‘Cluster Markov Packager Across K;’ Kopelman *et al.* (2015) implemented in POPHELPER (Francis 2017)]. Assignment probabilities were then visualized as stacked bar plots for each method (via a custom script: *plotUML_missData_maf.R*). For each dataset, we plotted as heatmaps the optimal *K* and standard deviation (SD) among replicates [(*plot_missData_comparison_maf.R*) (Scripts deposited at: https://github.com/btmartin721/mecr_boxturtle)].

2.8. Demography, migration history, and species-delimitation

We tested for reticulation in our phylogenomic dataset, as complementary to a range-wide evaluation of introgression in *Terrapene* (Martin *et al.* 2020). We first explored reticulation by identifying candidate edges (TREEMIX; Pickrell & Pritchard 2012), with populations having but one sample (*T. nelsoni* and *T. m. yucatana*) being excluded from input, which was then thinned to bi-allelic SNPs. TREEMIX was run 10X with subsets of SNPs randomly sampled per locus at 1,000 bootstrap replicates using the ‘global search’ option. The optimal number of admixture edges (*m*) was determined by running for *m*=1-10 and choosing the inflection point of log-likelihood scores.

TREEMIX results and introgression (Martin *et al.* 2020) were used to generate gene flow hypotheses in a species-delimitation framework (DELIMITR: Smith *et al.* 2017; Smith & Carstens 2020). DELIMITR uses the joint site-frequency spectrum (JSFS) and FASTSIMCOALV2.6 (Excoffier *et al.* 2013) to simulate demographic models, including possible variations of lumping/splitting taxa and primary divergence, secondary contact, or no gene flow. The program then builds an RF-classifier trained with the simulated models (i.e., 'supervised' M-L) to predict the best model. Input was generated using EASYSFS (https://github.com/isaacovercast/easySFS), with taxa reduced to N=6 given computational resources required by larger datasets. Those excluded (*T. m. mexicana*, *T. m. yucatana*, *T. o. luteola*, *T. coahuila*, *T. nelsoni*) were either limited in sample size or had clear taxonomic identities in the other analyses.

To improve efficiency, we also used EASYSFS to down-project the JSFS to six alleles for *T. c. bauri*, and ten each for the remaining taxa. Samples were selected to maximize per-individual occupancy, followed by a maximum 50% per-population missing data filter. The SVDQUARTETS result served as our topological prior for DELIMITR. Models considered were: No gene flow, primary divergence, secondary contact, and up to four migration edges. Migration was permitted between: *T. c. carolina* x *T. c. major*, *T. c. carolina* x *T. c. bauri*, *T. c. major* x *T. m. triunguis*, and *T. m. triunguis* x *T. o. ornata*. Population size priors were set broadly (1,000-100,000) and divergence times were obtained from LSD2 results. We defined a rule set that ranked overlapping coalescence times for *T. c. bauri*/*T. m. triunguis* and *T. c. major* from Mississippi/Florida. The migration rate prior range (1.96×10^{-6}–9.78×10^{-5}) was estimated from the number of migrants (GENEPOP v4.7.5; Rousset 2008). We applied three JSFS binning classes and 5,000 RF trees to build the classifier and predict the models.

3. RESULTS

3.1. Sampling and data processing

We sequenced 214 geographically-widespread *Terrapene* (Figure 1; Supplementary Information Table S1) including all recognized species and subspecies save the rare *T. nelsoni klauberi*.

IPYRAD recovered 134,607 variable sites (of 1,163,463 total) across 14,760 retained loci, with 90,777 as parsimony informative. The mean per-individual depth was 56.3X (Supplementary Information Figure S2).

3.2. Species tree inference

The lineage tree contained N=214 tips (Figure 2), whereas those from SVDQUARTETS (Figure 3a) and POMO (Figure 3b) grouped individuals into N=26 populations, again per locality and subspecies. SVDQUARTETS examined 10,299 unlinked SNPs and the species tree was assembled from 87,395,061 quartets. Full loci were used for POMO. All trees clearly delineated eastern *versus* western clades, with *T. mexicana*, *T. carolina*, and *T. coahuila* composing the eastern clade, with western represented by *T. ornata* and *T. nelsoni*.

All phylogenies delineated *T. ornata* and *T. nelsoni*. However, SVDQUARTETS nested *T. o. luteola* within a paraphyletic *T. o. ornata*, whereas IQ-TREE and POMO represented them as reciprocally monophyletic. In the eastern clade, SVDQUARTETS displayed two subdivisions: *Terrapene mexicana* (all subspecies) and *T. carolina*+*T. coahuila*. POMO included *T. m. triunguis* as sister to *T. c. carolina*+*T. c. major* but paraphyletic with respect to *T. m. mexicana*+*T. m. yucatana*. Furthermore, SVDQUARTETS, POMO, and IQ-TREE each differed with respect to the placement of *T. c. bauri*, *T. coahuila*, and two previously recognized populations within *T. c. major* (Martin *et al.* 2013, 2020). SVDQUARTETS depicted *T. c. bauri* as sister to the *major*/*coahuila*/*carolina* clade, whereas POMO placed *T. c. major* from Mississippi/*coahuila* as sister to *T. c. major* (FL)/*bauri/carolina*. IQ-TREE placed *T. c. bauri* sister to *T. carolina*/*T. mexicana*, and *T. coahuila*/*T. c. major* (MS) sister to *T. c. carolina*/*T. c. major* (FL).

The topology tests failed to reject either Martin *et al*. (2013) or the SVDQUARTETS trees, whereas morphology-based and POMO trees were significantly rejected (Table 1). Although the SVDQUARTETS tree was ranked highest, site-likelihood scores indicated a minority of sites drove those topologies (Supplementary Information Figure S3).

3.3. Species delimitation via BFD* and DELIMITR

TREEMIX converged upon four migration edges (Figure 3c; Supplementary Information Figure S4), with gene flow identified between: *Terrapene m. mexicana* × *T. o. ornata+T. o. luteola*; T. *c. carolina* × *T. c. bauri*; *T. m. triunguis* × *T. c. major* (MS); and *T. coahuila* × *T. c. major* (FL). To target specific reticulation hypotheses, DELIMITR was run with a reduced set of sub-species, in compliance with computational constraints. The best-fitting DELIMITR model within selected taxa (*T. m. triunguis*, *T. o. ornata*, *T. c. major*, *T. c. bauri*, and *T. c. carolina*) was *K*=4 (posterior probability=0.98; Table 3; Figure 3d). Also, *T. c. major* and *T. c. carolina* were collapsed, and three secondary contact migration edges were apparent: *T. o. ornata* × *T. c. carolina+T. c. major*; *T. c. bauri* × *T. c. carolina+T. c. major*; and *T. o. ornata* × *T. m. triunguis*. The second-best model was identical save for excluding the latter migration, although it also had the highest error (Table 3).

BFD* supported two top models (Table 2), each delimited (*K*=9), and all distinct except *T. o. ornata*/*T. o. luteola* (*K*=8; Figure 3d). Although not statistically distinguishable (BF<2), both were decisively better than others (BF>10). Convergence was confirmed for the likelihood traces, with mean per-model ESS>300 (Supplementary Information Table S2).

3.4. UML species delimitation

UML results varied considerably (Figures 4, 5; Supplementary Information Figures S5-S10), with mean optimal *K* greatest for t-SNE, followed by cMDS, VAE, and isoMDS (Figures 4a, 5a). Across datasets, PAM clustering with the gap statistic (PAM+GS) exhibited the largest *K*, whereas PAM with the highest mean silhouette width (PAM+HMSW) was lowest (Figure 5b). Hierarchical clustering (HC)+HMSW and VAE were intermediate (Figures 4a, 5a; Supplementary Information Figure S5). Each algorithm delimited *T. ornata* from *T. carolina*+*T. mexicana* in most datasets, save PAM+HMSW in some of the larger datasets, and among some t-SNE replicates (e.g., Supplementary Information Appendix B, B1). In all cases, cMDS with PAM+GS and HC+HMSW further delimited *T. m. triunguis*+*T. m. mexicana* from *T. carolina*, whereas cMDS with PAM+HMSW did not. Whether the remaining algorithms did so depended upon filtering parameters. Finally, cMDS with PAM+GS and HC+HMSW further partitioned subgroups within *T. carolina* in most datasets, whereas isoMDS did so in a limited fashion, and t-SNE split *T. carolina* into multiple clusters without a phylogenetic pattern. Bar plots for 64 filtered datasets are in Supplementary Information Appendix B1-B60.

We present representative results (Figure 3d) that displayed minimal inconsistencies among replicates and with respect to the phylogeny, with parameter choice also reflecting how each algorithm interacted with filtering values (below). This included 25% per-individual and per-population filters for all algorithms, a 5% MAF filter for cMDS, t-SNE, and VAE, and a 1% MAF filter for isoMDS. Five groups were delineated by cMDS with PAM+GS: *T. o. ornata* (ON)+*T. o. luteola* (DS), *T. c. major* from Mississippi (GUMS), *T. c. major* from Florida (GUFL), *T. c. carolina* (EA), and *T. m. mexicana* (MX)+*T. m. triunguis* (TT). However, *T. c. bauri* displayed mixed assignment between *T. c. carolina* and GUMS. cMDS with HC+HMSW also delimited *K*=5 but lumped the two populations of *T. c. major*, splitting *T. c. bauri*, and grouped some *T. c. carolina* individuals with *T. c. bauri*. It also split *T. ornata* and *T. carolina*+*T. mexicana*. While ISOMDS with PAM+GS resembled CMDS with HC+HMSW, it clustered *T. c. bauri* with *T. c. carolina*. Similarly, ISOMDS with HC+HMSW showed *T. o. ornata*+*T. o. luteola*, *T. c. carolina*+GUMS+GUFL, and *T. m. mexicana*+*T. m. triunguis*. However, ISOMDS with PAM+HMSW only delimited *T. ornata* from *T. carolina*+*T. mexicana*. The model T-SNE (at perplexity=15) clearly partitioned *T. ornata*, *T. carolina*, and *T. mexicana*, though the PAM+GS algorithm exhibited spurious groupings within *T. carolina*. However, T-SNE with HC+HMSW clustered many *T. c. carolina* with GUFL and the remaining with GUMS. We found VAE and T-SNE with PAM+HMSW only delimited *T. ornata*, *T. carolina*, and *T. mexicana*.

3.5. Effects of data filtering

Among all dimensionality reduction and clustering algorithms, greater per-individual and per-population missing data generally increased mean optimal *K* and SD (Figures 4a-b and 5a-b; Supplementary Information Figure S5). PAM+HMSW deviated due to low *K*, regardless of filtering. This was manifested as two types of noise in the bar plots (Supplementary Information Appendix B1-B60): ‘vertical striping’ (inconsistency of assignment among replicates) and ‘horizontal striping’ (groupings inconsistent with phylogeny). We found the former largely driven by increased missing data per-locus, whereas the latter by increased missing data per-individual. However, performance varied among algorithms in how they interacted with both missing data parameters.

We found that T-SNE consistently resolved *T. ornata* and *T. carolina*+*T. mexicana*, but *T. mexicana* was only partitioned when per-population filtering was 25%. However, T-SNE did not further partition *T. carolina* in any dataset and displayed a tendency to form phylogenetically spurious groupings (=horizontal striping). The perplexity grid search (Figures 4c-d and 5b; Supplementary Information Figures S6-S10) suggested that the highest *K* and SD among replicates was at perplexity=5-10, with a plateau at higher perplexities.

We also found CMDS with PAM+GS and HC+HMSW delineated most clades, save for inconsistency amongst the *T. c. major* populations and *T. coahuila*. In contrast, CMDS and ISOMDS with PAM+HMSW typically displayed *K*=2 or 3 and contained no phylogenetically meaningful clusters with ≥75% missing data per-individual (e.g., Supplementary Information Appendix B58). Finally, VAE partitioned *T. ornata* from *T. carolina*+*T. mexicana* in all datasets, but *T. mexicana* was only delineated from *T. carolina* when per-individual missing data was ≤50% and with MAF filter.

Filtering by MAF ubiquitously reduced noise, although results varied by algorithm (Supplementary Information Appendix B1-B60). For T-SNE, optimal *K* and SD were reduced. In contrast, the clusters yielded by CMDS with PAM+GS and HC+HMSW were only marginally affected. We found CMDS and ISOMDS with PAM+HMSW and MAF filters ≥3% were less noisy, but for ISOMDS with PAM+GS and HC+HMSW the MAF filter effect was dependent on the number of individuals present in the dataset. With a maximum of 25% per-individual missing data (N=117), a 1% MAF filter shows minimal striping and higher *K* than did a >1% MAF filter. However, larger MAF filters have a greater effect above 25% per-individual filtering. Lastly, optimal *K*, SD, and striping in VAE were strongly influenced by MAF filters (Figures 4e-f, 5a, Supplementary Information Figure S5). With lower per-individual filters (≤50%) and a 5% MAF filter, VAE consistently delineated *T. mexicana* from *T. carolina*, even with high per-population filters. However, lower MAF and higher per-individual (>50%) filters introduced progressively more noise and grouped *T. carolina* and *T. mexicana*.

3.6. Relative performance among approaches

The cMDS model with PAM+GS and HC+HMSW consistently displayed the highest *K* and was less susceptible to data filtering. However, isoMDS with PAM+GS and HC+HMSW were more influenced by filtering parameters, but still consistently resolved the highest level of hierarchical structure (*T. ornata/T. carolina+T. mexicana*). Both cMDS and isoMDS with PAM+HMSW consistently displayed the lowest *K* at the top hierarchy and were usually in complete agreement. We note that t-SNE was highly susceptible to horizontal and vertical striping, and only partitioned *T. mexicana* from *T. carolina* ssp. at 25% per-individual filtering. Similarly, VAE performed far more consistently with a 5% MAF filter and ≤50% per-individual filtering. VAE also consistently hovered between K=2 and K=3, making it the second most conservative algorithm next to PAM+HMSW. In contrast, BFD* delimited the most taxa among all the approaches, splitting all save *T. o. luteola* and *T. o. ornata*, and delimitR partitioned *T. ornata*, *T. carolina*, *T. mexicana*, and *T. c. bauri*.

In terms of computational resources, the UML algorithms were far less intensive than BFD* and delimitR, enabling stochasticity to be assessed in many replicates. Each UML algorithm needed ~1-3GB RAM per replicate and ~2-3 days runtime for 100 replicates. Comparatively, BFD* required the greatest memory and time, often using >200GB RAM (with 16 CPU threads) and a ~10-day runtime per model. We note delimitR used much less memory and was faster than BFD*, but output ~3.2 TB with six tips and 51 models.

4. DISCUSSION

We observed substantial heterogeneity in resolving *Terrapene* via M-L approaches, which echoed previous morphological and single-gene results (Milstead 1967, 1969; Milstead & Tinkle 1967; Butler *et al.* 2011; Martin *et al.* 2013). We interpret this variability as reflecting inherent differences in dimensionality-reduction, clustering, and *K*-selection, as well how methodologies interact with biological aspects of the data and user-defined filtering.

4.1. Delimitation hypotheses and biological interpretations reconciled

Two factors likely contribute to the observed heterogeneity: 1) An hierarchical arrangement of phylogenetic signal (Martin *et al.* 2013); and 2) Phylogenetic discord (Martin *et al.* 2020). Both reverberate noticeably within prior literature and phylogenetic evaluations.

The most consistent grouping was eastern (*T. carolina+T. mexicana*) versus western (*T. ornata*) clades, representing the deepest *Terrapene* divergence (Figures 3a-b). This is unsurprising given it is the most prominent axis of molecular variation (morphologically corroborated; Milstead & Tinkle 1967; Dodd 2001) Nominal species have been identifiable since late Miocene (Holman & Fritz 2005), as corroborated by molecular dating (Figure 2).

4.1.1. Terrapene ornata

Although introgression between *T. o. ornata* and *T. m. triunguis* occurred during secondary contact (Table 3; Figure 3d), no contemporary evidence for introgression among these clades emerged from previous evaluations, except rare F_1 hybrids between *T. o. ornata* and *T. carolina* (Martin *et al.* 2020). TREEMIX also suggested introgression between *T. ornata* and *T. m. mexicana* (Figure 3c). Although contact with *T. mexicana* was certainty possible during glacial expansion-contraction (Martin *et al.* 2020), we echo earlier conclusions that hybridization lacks justifiable taxonomic implications, per hybridization between *T. ornata* and *T. carolina* (Martin *et al.* 2020).

Regarding *T. ornata*, algorithms failed to further partition *T. o. ornata*/*T. o. luteola*, suggesting a lack of diagnosability at our most recent scale. Notably, both also lack reciprocal monophyly in some phylogenomic (Figure 3a) and single-gene analyses (Martin *et al.* 2013). They also lack clear morphological synapomorphies (Minx 1996). Although *T. o. luteola* exhibits habitat and movement patterns markedly different from mesic conspecifics (Nieuwolt 1996), few investigations have similarly compared *T. ornata* subspecies, such that inferences regarding reproductive isolation (or potential thereof) are difficult. Populations of *T. o. luteola* also do not exhibit thermal adaptations that are mutually exclusive from *T. o. ornata*, as might be surmised given other desert-dwelling tortoises (Plummer 2003).

Previous authors hypothesized *T. o. luteola* as a relict population (Milstead & Tinkle 1967). Weak differentiation [molecular: Martin *et al.* (2013); morphological: Dodd (2001)], as well as possible paraphyly of *T. o. ornata* (Figure 3a) suggest isolation was recent. Although phylogenetic structuring was present in some analyses (e.g., Figure 2), it is insufficient to mandate recognition beyond the subspecific level. However, special guidelines that delineate relictual lineages may be warranted (Mussmann *et al.* 2020), particularly given the isolation and reduced N_e in *T. o. luteola* (Nieuwolt 1996).

4.1.2. **Terrapene mexicana**

The second most frequent split (Figures 2, 3a) divided *T. mexicana* and *T. carolina*, corresponding to the second-deepest phylogenetic node (Figures 2, 3a). This lends further support to a prior elevation of *T. mexicana* (Martin *et al.* 2013). Conspecifics of *T. mexicana* also share multiple morphological characteristics, such as carapace coloration and a degree of concavity to the posterior plastron, that separate the group from *T. carolina* (Minx 1996). *Terrapene mexicana mexicana* (as well as *T. m. yucatana*, excluded due to sample size) have isolated, allopatric ranges (Smith & Smith 1980; Ernst & Lovich 2009), with reproductive isolation difficult to assume.

Evidence for interbreeding of *T. m. triunguis* with *T. carolina* subspecies in the southeastern United States (Butler *et al.* 2011) has led some to conclude that species-level recognition of *T. mexicana sensu lato* is unwarranted (Fritz & Havaš 2014). Indeed, our own results suggest introgression between *T. m. triunguis* and *T. carolina* in secondary contact (Figure 3d). Martin *et al.* (2020) confirmed hybridization of *T. m. triunguis* with both *T. c. major* and *T. c. carolina* in the southeast, yet found genetic exchange was restricted, given that: (1) Genetically 'pure' individuals are predominant throughout the contact zone; and (2) patterns of gene-level exchange exhibit strong sigmoidal patterns, suggesting selection against interspecific heterozygotes. Additionally, the sigmoidal pattern was strongest within a subset of genes involved in thermal adaptation (Martin *et al.* 2020), suggesting species boundaries are modulated by an adaptive barrier between co-occurring *T. mexicana* and *T. carolina* sub-species. This functional perspective corroborates the proposed taxonomy herein, and by Martin et al. (2013).

4.1.3. **Terrapene carolina**

Partitioning within *T. carolina* echoed inconsistencies in our phylogenies (Figures 2, 3a-b), and seemingly depended upon algorithm and filtering regime (Figure 3d; Supplementary Information B). *Terrapene carolina major*, for example, occasionally split from the remaining *T. carolina* (usually including *T. coahuila*; cMDS+HC, Figure 3d), whereas in other cases, *T. c. major* (FL and MS) were separated (with the former grouped into *T. c. carolina*) (t-SNE+HC, Fig. 3d) .

In contrast to steep clines in interspecific comparisons (Martin *et al.* 2020; see above), a transect of the *T. c. carolina* and *T. c. major* contact zone revealed shallow genetic transition, with multiple loci showing potential signatures of selection-driven introgression. Previous authors have hypothesized either direct ancestry (Bentley & Knight 1998) or historic admixture with a now extinct taxon, [*T. c. putnami*; Butler *et al.* (2011)]. While such 'ghost' admixture can mislead population structure (Lawson *et al.* 2018), such a signal is unlikely manufactured in entirety. In contrast to Butler *et al*. (2011), Martin *et al.* (2020) found a pervasive signal of population structure and strong molecular diagnosability in *T. c. major*, with a cryptic east-west division roughly defined by the Apalachicola River [a recurring phylogeographic discontinuity reflecting recolonization from disparate Gulf Coast refugia; Soltis *et al.* (2006)]. Our interpretations refuted the 'genetic melting pot' assertion (Fritz & Havaš 2014) and favored instead recognition of the two as distinct evolutionarily significant units (ESUs). Additionally, differences in habitat use and movement patterns distinguish *T. c. major* (Meck *et al.* 2020), which spends greater time in mesic habitats (e.g., floodplain swamps). In support, early studies observed a distinct webbing of the hind foot in *T. c. major* (Taylor 1895). Given the genetic data herein, we reject the taxonomic coalescence of *T. c. major*.

Terrapene carolina bauri was similarly resistant to straightforward classification, although generally grouping with *T. c. major* (when the latter was separated from *T. c. carolina*; Figure 3b). We found *T. c. bauri* as sister to either the remaining *T. carolina* group, *T. c. carolina*+*T. c. major*, or only *T. c. carolina* (Figures 2-3; Martin *et al*. 2013). This argues against it being sister to *T. m. triunguis* (per Spinks *et al.* 2009). Osteologically, it alone shares a complete zygomatic arch with *T. c. major* (Taylor 1895; Ditmars 1934), although other morphological investigations have allied it more closely with *T. c. carolina* (Minx 1996). Thus, phylogenetic inconsistency for *T. c. bauri* clearly extends beyond our results.

Although hybridization likely contributes to this issue (as with *T. c. major*), the biogeography of the region may provide insight, with peninsular Florida recognized as a distinct biogeographic province (Ennen *et al.* 2017). Intraspecific division are recognized in multiple species [e.g., *Chelydra serpentina*, *Deirochelys reticularia* (Walker & Avise 1998)], a phylogenetic legacy likely reflecting periodic isolation from the mainland that may have inflated genetic divergences (Douglas *et al.* 2006), and facilitated secondary contact. This scenario is supported by DELIMITR and TREEMIX (Figures 3c-d). Here, we again stress that evidence is sufficient to support continued recognition, yet not for taxonomic elevation.

4.1.4. Terrapene coahuila

Terrapene coahuila represents a persistent phylogenetic uncertainty (Spinks *et al.* 2009; Wiens *et al.* 2010; Martin *et al.* 2013). It is unique in that it occupies streams, ponds, and marshes, with terrestrial movements restricted to the rainy seasons (Webb *et al.* 1963). Milstead (1967) postulated that *T. coahuila* evolved as a relictual population of a *Terrapene* ancestor (potentially the extinct *T. c. putnami*) during pluvial periods associated with Pleistocene glacial-interglacial cycles across the broad eastern coastal plain of Mexico. In this scenario, relictual populations are what remains from those north-south migrations, as hypothesized for *T. m. mexicana* and *T. m. yucatana*. The scenario is plausible, given semi-aquatic adaptations in the presumed ancestor (*T. c. putnami*) and closely related *T. c. major*, as well as shared morphologies between extinct *T. c. putnami* and modern *T. coahuila* (Milstead 1967). The phylogenetic placement of *T. coahuila*, as nested within *T. c. major*, offers further evidence (Figure 2-3), as does the almost unanimous UML grouping in our results (Figure 3d; Supplementary Information Appendix B1-B60). As with *T. o. luteola*, small, isolated populations that differ in evolutionary rates could contribute to a lack of molecular similarity with extant *T. c. major*, despite a unique functional morphology (Brown 1971).

4.2. Relative performance of species-delimitation methods

As with prior studies (Derkarabetian *et al.* 2019; Mussmann *et al.* 2020), we also found considerable variation among methods, some of which can be attributed either to idiosyncrasies in the data or to algorithms and their implementation. First, among RF methods cMDS with PAM+GS and HC+HMSW displayed higher *K* and isoMDS generally yielded smaller *K* (Figure 3d), with the latter being attributed by Derkarabetian *et al.* (2019) to the retention of only two dimensions. PAM+HMSW (Figure 3d) also trended towards a small *K*=2, corresponding to the deepest *Terrapene* bifurcation, and suggesting a potential failure in identifying hierarchical clusters. Here, a solution might include partitioning divergent subtrees for separate analyses.

In contrast to Derkarabetian *et al.* (2019), we found t-SNE the most inclined to produce inconsistent groupings, a pattern most prevalent with the gap statistic (Supplementary Information Appendix B1-B60). Mussmann *et al.* (2020) concurred, although in their case it was PAM+HMSW. We see this as an inherent problem relating to data structure. Previous comparisons of t-SNE found low fidelity with global data patterns, and latent space distances were poor proxies for ‘true’ among-group distances, particularly when compared to VAE (Becht *et al.* 2019; Battey *et al.* 2020). This potentially explains our observed ‘plateau’ of mean optimal *K* and SD in the t-SNE perplexity grid-search, in that perplexity defines relative weighting of local versus global components (Wattenberg *et al.* 2016). It may also explain the formation of spurious clusters even at higher perplexities, in that clusters are formed *post hoc* (PAM or HC). Thus, T-SNE may perform poorly when inter-cluster distances/dispersion in global data structure are skewed, although it is not clear to what degree hyperparameter choice and initializations contribute (Belkina *et al.* 2019; Kobak & Berens 2019).

In our case, VAE with DBSCAN yielded higher fidelity to the underlying phylogeny (Figure 3a) and was also more robust to missing data (Figures 4e-f). A particular benefit of the VAE approach is the output of a standard deviation around samples in latent space (Derkarabetian *et al.* 2019). Our DBSCAN hyperparameters were informed directly from latent variable uncertainties, and in so doing, we circumvented the issue of *K*-selection that drove heterogeneity in the RF and T-SNE methods [also recognized with other clustering approaches (Janes *et al.* 2017)].

By comparison, BFD* partitioned all groups, which may reflect a vulnerability to local structure at the population level, as reported by others for MSC methods (Sukumaran & Knowles 2017). BFD* and VAE partitioned equally in Mussmann *et al*. (2020), although their populations were relictual and without contemporary connectivity, whereas *Terrapene* reflects both historical (Figure 3d) and contemporary gene flow (Martin *et al.* 2020). In corroboration, other studies have also demonstrated reticulation to condense VAE clusters (Derkarabetian *et al.* 2019; Newton *et al.* 2020). Although not run on a full dataset, DELIMITR formed clusters consistent with (or similar to) several of the UML methods (e.g., ISOMDS+GS; Figure 3d, Table 3). The latter displayed a particular utility regarding testing targeted hypotheses relating to demographic processes such as migration, whereas these must be applied to UML results *post hoc*.

4.3. Data treatment and assignment consistency

We generally found a tendency for UML methods to 'over-split' given large amounts of missing data, and phylogenetically inconsistent groupings ('horizontal striping') were most pronounced when missing data was elevated per-individual (Supplementary Information Appendix B1-B60). However, low-level, undetected introgression could also drive such a pattern. Mussmann *et al*. (2020) noted a similar pattern with the RF methods, possibly reflecting an artificial similarity among samples generated by a non-random distribution of missing data. A similar 'vertical striping' effect was seen when missing data was elevated per-locus (e.g., Supplementary Information Appendix B13), often manifested as inconsistency among replicates. However, effects varied across methods, as per previous analyses [phylogeographic: Graham *et al.* (2020); phylogenetic: Molloy & Warnow (2018)].

Missing-data bias is a particular concern when patterns are non-random (i.e., presence or absence of observations are data-dependent; Rubin 1976). Here, the temptation is to filter stringently, yet we found highly filtered datasets were biased towards smaller *K*, generally retaining only nodes deepest within the phylogeny. The same pattern was identified using the VAE method (Newton *et al.* 2020), and is intuitive given expectations that a major subset of missing ddRAD data are systematically distributed [defined by mutation-disruption of restriction sites: Gautier *et al.* (2013); Eaton *et al.* (2017)]. Thus, indiscriminate exclusion may unintendedly bias information content leading to the underestimation of diversity (Arnold *et al.* 2013; Leaché *et al.* 2015; Huang & Knowles 2016). Again, care must be taken to filter the data such that sufficient discriminatory signal remains, while also being mindful of the signal-to-noise ratio, and the underlying biases driving interactions of sparse data versus information content (Nakagawa & Freckleton 2008).

A potential solution involves the input of genotypes to fill in missing values (per Howie *et al.* 2009; Durbin 2014; Das *et al.* 2016). However, a cautious *a priori* designation of population references is needed, particularly when group-delimitation is the goal. It may be appropriate to employ phylogenetically-informed methods previously applied in comparative studies (e.g., Goolsby *et al.* 2017).

We found MAF filters dampened the effect of missing data, likely by removing sequencing errors and uninformative variants at low-frequency (Mathieson & McVean 2012; Jakobsson *et al.* 2013). In a similar context, Linck & Battey (2019) found MAF filters to significantly increase in the discriminatory capacity of assignment-test methods (STRUCTURE; Pritchard *et al.* 2000). In our case, MAF filtering reduced noise and improved group differentiation (e.g., resulting in lower variability among replicates; Figures 4-5, Supplementary Information Figures S5-S6), although this might prompt the M-L algorithms to miss low levels of introgression. Thus, we view it as a parameter in need of further empirical exploration.

4.4. Conclusions

UML approaches identify groups based on the structure of the data, and as such, represent a natural extension to species-delimitation approaches. However, we found idiosyncrasies regarding: Phylogenetic context of the study system (e.g., hierarchical structure, reticulation); the manner by which clustering and *K*-selection approaches were applied *post hoc*; and the bioinformatic treatment of the data. We particularly note that lax filtering, performed to maximize size and information content, actually promote spurious groupings and inflate variability among replicates. An alternate method, i.e., filtering via MAF to promote informative characters, favorably altered the signal-to-noise ratio and increased the consistency of our delimitations. Thus, we recommend that UML practitioners test multiple algorithms, veer away from high levels of missing data, and utilize MAF filters. We conclude that UML approaches, when applied to formulate taxonomic hypotheses and reduce dimensionality of complex data, are valuable and computationally efficient tools for integrative species-delimitation, as demonstrated within our study system.

ACKNOWLEDGEMENTS

Many thanks to those volunteers, organizations, and agencies that contributed tissue samples (Supplementary Information Table S1). We also thank colleagues for guidance at various stages of this project: A. Alverson, W. Anthonysamy, M. Bangs, J. Beaulieu, R. Birkhead, T. Chafin, J. Koukl, P. Martin, S. Mussmann, C. Phillips, J. Placyk, J. Pummill, and Z. Zbinden. Sample collections were approved under Animal Care and Use Committee (IACUC) protocols: #113 (University of Texas/Tyler), #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, and University of Arkansas endowments to MRD and MED. The Arkansas High Performance Computing Cluster (AHPCC) and Jetstream cloud (XSEDE #TG-BIO160065) provided computational resources.

DATA AVAILABILITY STATEMENT

Raw ddRADseq data are available on the GenBank Nucleotide Database at https://www.ncbi.nlm.nih.gov/bioproject/563121 (BioProject ID: 563121). Scripts for parsing and plotting UML output are available on GitHub at https://github.com/btmartin721/mecr_boxturtle. Input and output files for all analyses can be found in a Dryad Digital Repository (DOI: https://doi.org/10.5061/dryad.xgxd254fc).

5. REFERENCES

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.

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

Arnold B, Corbett-Detig RB, Hartl D, and Bomblies K (2013) RAD seq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

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.

Avise JC (2000a) Cladists in Wonderland. *Evolution*, **54**, 1828–1832.

Avise JC (2000b) *Phylogeography: the history and formation of species*. Harvard University Press, Cambridge, MA, USA.

Battey CJ, Coffing GC, and Kern AD (2020) Visualizing population structure with variational autoencoders. *bioRxiv*, 248278.

Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, and Newell EW (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, **37**, 38–44.

Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, and Snyder-Cappione JE (2019) Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications*, **10**, 1–12.

Bentley CC and Knight JL (1998) Turtles (Reptilia: Testudines) of the Ardis local fauna late Pleistocene (Rancholabrean) of South Carolina. *Brimleyana*, **25**, 1–33.

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

Brown WS (1971) Morphometrics of *Terrapene coahuila* (Chelonia, Emydidae), with comments on its evolutionary status. *The Southwestern Naturalist*, **16**, 171–184.

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.

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

Chernomor O, Von Haeseler A, and Minh BQ (2016) Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, **65**, 997–1008.

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

Chollet F (2015) Keras. https://keras.io.

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, and McGue M (2016) Next-generation genotype imputation service and methods. *Nature Genetics*, **48**, 1284–1287.

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.

Ditmars RL (1934) A review of the box turtles. *Zoologica*, **17**, 1–44.

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

Douglas ME, Douglas MR, Schuett GW, and Porras LW (2006) Evolution of rattlesnakes (Viperidae; *Crotalus*) in the warm deserts of western North America shaped by Neogene vicariance and Quaternary climate change. *Molecular Ecology*, **15**, 3353–3374.

Durbin R (2014) Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, **30**, 1266–1272.

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.

Edwards S V, Potter S, Schmitt CJ, Bragg JG, and Moritz C (2016) Reticulation, divergence, and the phylogeography–phylogenetics continuum. *Proceedings of the National Academy of Sciences*, **113**, 8025–8032.

Eldredge N and Cracraft J (1980) *Phytigenetic Patterns and the Evolutinary Process: Methods and Theory in Comparative Biology*. Columbia University Press, New York, NY, USA.

Ennen JR, Matamoros WA, Agha M, Lovich JE, Sweat SC, and Hoagstrom CW (2017) Hierarchical, quantitative biogeographic provinces for all North American turtles and their contribution to the biogeography of turtles and the continent. *Herpetological Monographs*, **31**, 114–140.

Ernst CH and Lovich JE (2009) *Turtles of the united states and Canada, 2nd Edition*. The John Hopkins University Press, Baltimore, MD, USA.

Ester M, Kriegel H-P, Sander J, and Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, and Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.

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 (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578–588.

Francis RM (2017) pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, **17**, 27–32.

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.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, and Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Goolsby EW, Bruggeman J, and Ané C (2017) Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, **8**, 22–27.

Graham MR, Santibáñez-López CE, Derkarabetian S, and Hendrixson BE (2020) Pleistocene persistence and expansion in tarantulas on the Colorado Plateau and the effects of missing data on phylogeographical inferences from RADseq. *Molecular Ecology*, **29**, 3684–3701.

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.

Holman JA and Fritz U (2005) The box turtle genus Terrapene (Testudines : Emydidae) in the Miocene of the USA. *Journal of Herpetology*, **15**, 81–90.

Howie BN, Donnelly P, and Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**, e1000529.

Huang H and Knowles LL (2016) Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic Biology*, **65**, 357–365.

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.

Jakobsson M, Edge MD, and Rosenberg NA (2013) The relationship between FST and the frequency of the most frequent allele. *Genetics*, **193**, 515–528.

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, and Andrew RL (2017) The K = 2 conundrum. *Molecular Ecology*, **26**, 3594–3602.

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.

Kingma DP and Welling M (2013) Auto-encoding variational bayes. *In: Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:1312.6114 [stat.ML].

Kobak D and Berens P (2019) The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, **10**, 1–14.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, and Mayrose I (2015) CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across *K*. *Molecular Ecology Resources*, **15**, 1179–1191.

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

Lawson DJ, van Dorp L, and Falush D (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, **9**, 3258.

Leaché AD, Banbury BL, Felsenstein J, De Oca AN-M, and Stamatakis A (2015) Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, **64**, 1032–1047.

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

Linck EB and Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. *Molecular Ecology Resources*, **19**, 639–647.

Long C and Kubatko L (2018) The effect of gene flow on coalescent-based species-tree inference. *Systematic Biology*, **67**, 770–785.

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) Contrasting signatures of introgression in North American box turtle (*Terrapene* spp.) contact zones. *Molecular Ecology*, **29**, 4186–4202.

Mathieson I and McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, **44**, 243–246.

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

Meck JR, Jones MT, Willey LL, and Mays JD (2020) Autecological study of Gulf Coast box turtles (*Terrapene carolina major*) in the Florida Panhandle, USA, reveals unique spatial and behavioral characteristics. *Herpetological Conservation and Biology*, **15**, 293–305.

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.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, and Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, **37**, 1530–1534.

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.

Molloy EK and Warnow T (2018) To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, **67**, 285–303.

Mussmann SM, Douglas MR, Oakey DD, and Douglas ME (2020) Defining relictual biodiversity: Conservation units in speckled dace (Leuciscidae: *Rhinichthys osculus*) of the Greater Death Valley ecosystem. *Ecology and Evolution*, **10**, 10798–10817.

Nakagawa S and Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, **23**, 592–596.

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.

Nieuwolt PM (1996) Movement, activity, and microhabitat selection in the western box turtle, *Terrapene ornata luteola*, in New Mexico. *Herpetologica*, 487–495.

Nosil P and Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 332–342.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, and Dubourg V (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

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.

Pickrell JK and Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Plummer M V (2003) Activity and thermal ecology of the box turtle, *Terrapene ornata*, at its southwestern range limit in Arizona. *Chelonian Conservation and Biology*, **4**, 569–577.

Pritchard JK, Stephens M, and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–886.

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

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.

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.

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

Rousset F (2008) genepop '007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

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.

Shepard RN, Romney AK, and Nerlove SB (1972) *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences: I. Theory.* Seminar Press, New York City, NY, USA.

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

Smith ML, Ruffley M, Espíndola A, Tank DC, Sullivan J, and Carstens BC (2017) Demographic model selection using random forests and the site frequency spectrum. *Molecular Ecology*, **26**, 4562–4573.

Smith HM and Smith RB (1980) Synopsis of the herpetofauna of Mexico: Volume VI, guide to Mexican turtles, bibliographic addendum III. John Johnson, North Bennington, Vermont ("1979"), xviii + 1044 pp.

Soltis DE, Morris AB, McLachlan JS, Manos PS, and Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Spinks PQ and Shaffer HB (2009) Conflicting mitochondrial and nuclear phylogenies for the widely disjunct Emys (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization. *Systematic Biology*, **58**, 1–20.

Spinks PQ, Thomson RC, Lovely GA, and Shaffer HB (2009) Assessing what is needed to resolve a molecular phylogeny: Simulations and empirical data from emydid turtles. *BMC Evolutionary Biology*, **9**, 56.

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

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.

Taylor WE (1895) The box tortoises of North America. *Proceedings of the United States National Museum*, **17**, 573–588.

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.

To T-H, Jung M, Lycett S, and Gascuel O (2016) Fast dating using least-squares criteria and algorithms. *Systematic Biology*, **65**, 82–97.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, **106**, 9939–9946.

Walker DE and Avise JC (1998) Principles of phylogeography as illustrated by freshwater and terrestrial turtles in the southeastern United States. *Annual Review of Ecology and Systematics*, **29**, 23–58.

Wattenberg M, Viégas F, and Johnson I (2016) How to use t-SNE effectively. *Distill*, **1**, e2.

Webb RG, Minckley WL, and Craddock JE (1963) Remarks on the Coahuilan box turtle, Terrapene coahuila (Testudines, Emydidae). *The Southwestern Naturalist*, **8**, 89–99.

Wiens JJ, Kuczynski CA, and Stephens PR (2010) Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation. *Biological Journal of the Linnaean Society*, **99**, 445–461.

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

TABLES AND FIGURES

Table 1: Topology tests for hypothesized *Terrapene* phylogenies. Sanger sequencing and morphology trees are based on previously published data whereas those representing SVDQUARTETS and POMO (Polymorphism-Aware Model) were generated in this study from ddRADseq data. *P*-values in bold with '*' indicate significance (P>0.05/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 Bayes Factor Delimitation (BFD) in *Terrapene*. Bayes factors (BF) depict support among models and were calculated as 2 × (MLE_1-MLE_2). '*'=best supported models; '+'=taxa grouped together; '/'=multiple groupings. DS=*T. o. luteola*, ON=*T. o. ornata*, EA=*T. c. carolina*, GUFL=*T. c. major* from Florida, GUMS=Mississippi *T. c. major*, CH=*T. coahuila*, FL=*T. c. bauri*, TT=*T. m. triunguis*, and MX=*T. m. mexicana*. East=all *T. carolina* and *T. mexicana*, West=all *T. ornata*. Outgroup (not shown) included *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

Table 3: The top five (of 51) DELIMITR models describing six *Terrapene* taxa. Model=rank determined by random forest (RF) vote counts (=# Votes). '*'=best supported model. Lumped taxa are grouped by '+', whereas '/' delimits taxa. '×' indicates migration events promoting secondary contact, with multiple migrations per model separated by commas. ON=*T. o. ornata*, TT=*T. m. triunguis*, FL=*T. c. bauri*, GUMS=*T. c. major* from Mississippi, GUFL=Florida *T. c. major*, EA=*T. c. carolina*. Error=proportion of incorrect model choices.

Model	**# Votes**	**Species (# delimited)**	**Secondary Contact**	**Error**
17*****	464	ON/TT/FL/GUMS+GUFL+EA (4)	ON × TT, TT × GU+EA, FL × GU+EA	0.017
14	445	ON/TT/FL/GUMS+GUFL+EA (4)	TT × GU+EA, FL × GU+EA	0.036
3	441	ON/TT+FL+GUMS+GUFL+EA (2)	ON × TT+FL+GU+EA	0.009
8	359	ON/TT/FL+GUMS+GUFL+EA (3)	ON × TT, TT × FL+GU+EA	0.009
30	218	ON/TT/FL/GUMS+GUFL/EA (5)	TT × GU, FL × EA, GU × EA	0.007

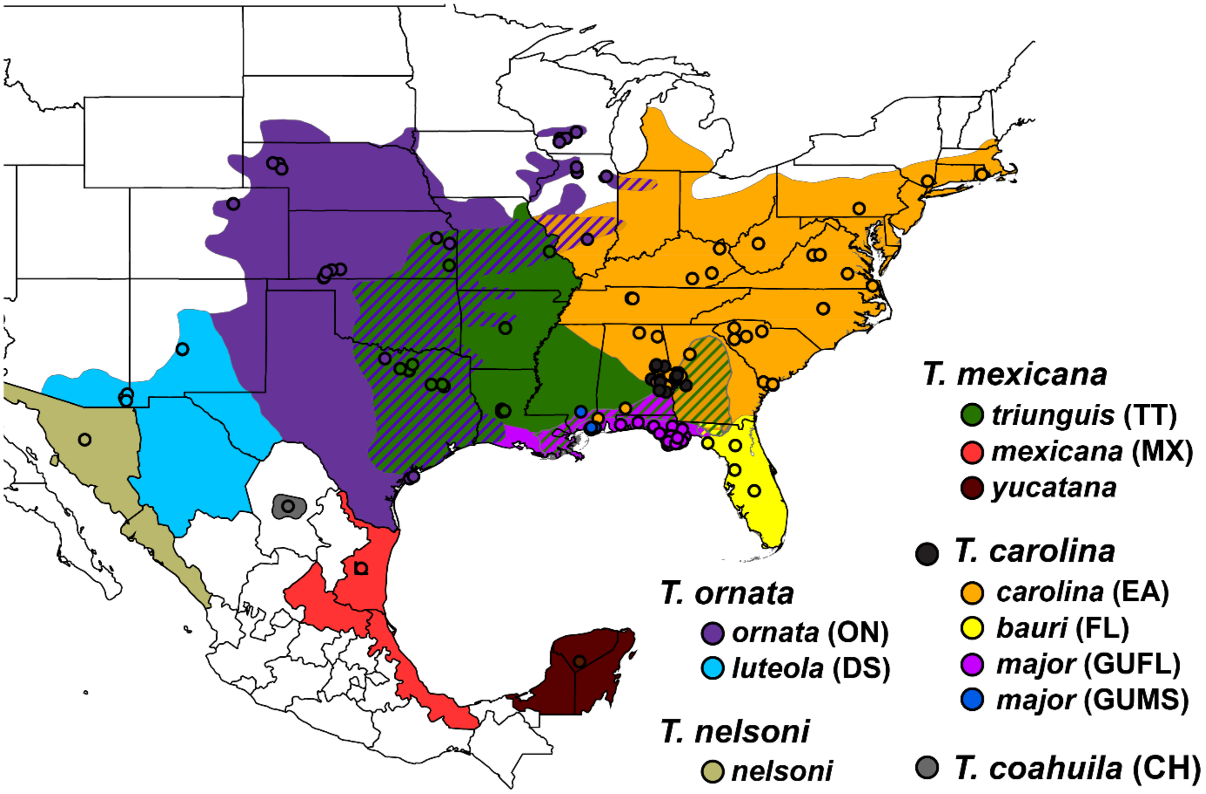


Figure 1: Range map and sample localities (=circles) for N=214 *Terrapene*. Closed circles=*T. carolina* samples without subspecific identification in the field. Cross-hatched areas=known hybrid zones. Headings and subheadings represent species and subspecies. *Terrapene carolina major*=*T. carolina major* and includes distinct subpopulations from Mississippi (GUMS) and Florida panhandle (GUFL). Parenthetical legend abbreviations correspond to Tables 2 and 3.

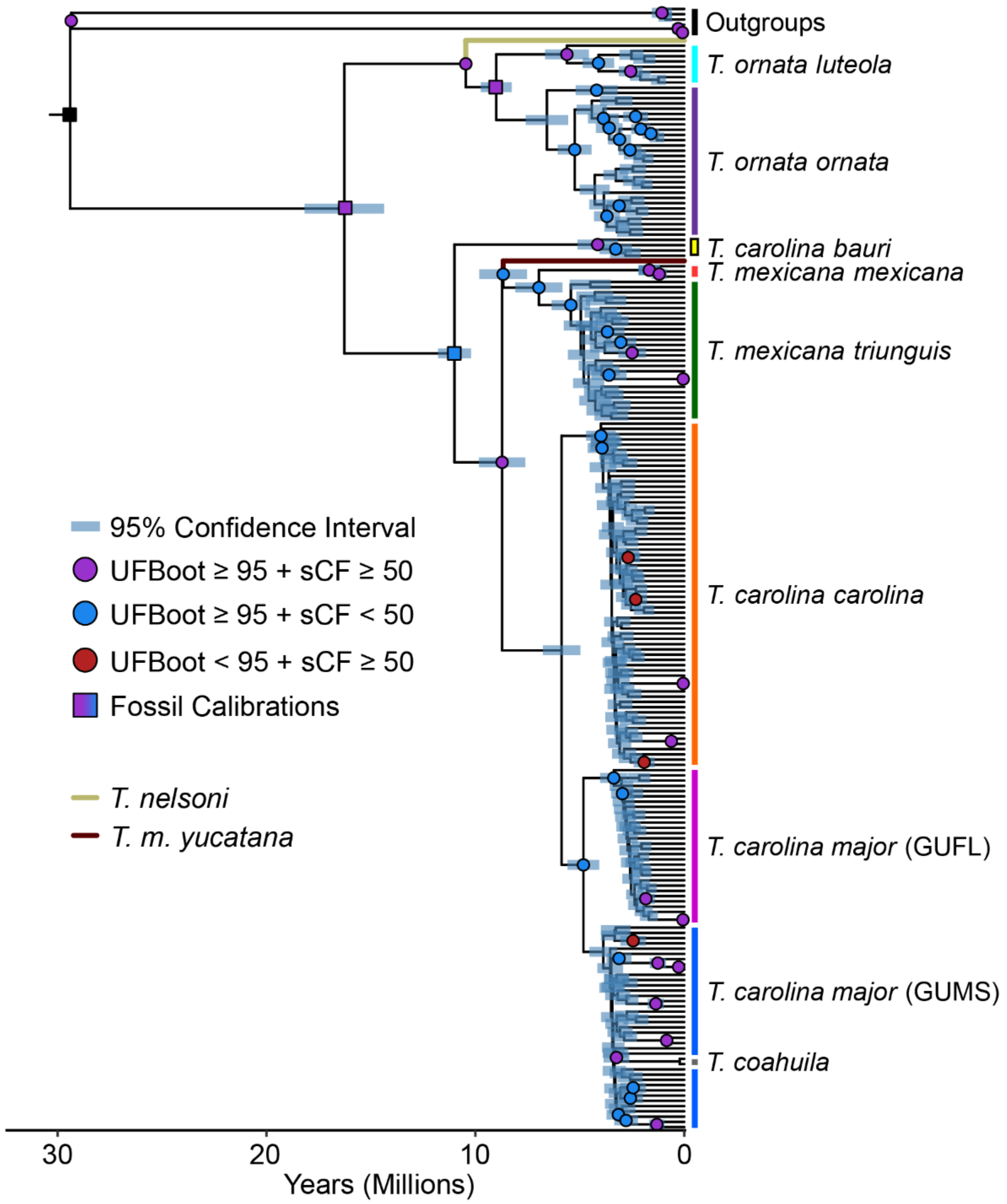


Figure 2: Chronogram reflecting relationships among 214 *Terrapene* ddRADseq samples as generated in IQ-TREE v2.1.2 and time-calibrated using LSD2. Node support was assessed with 1,000 ultrafast bootstrap (UFBOOT) replicates, and site concordance-factors (sCF) calculated from 10,000 randomly-sampled quartets. Well-supported nodes (UFBOOT≥95%, sCF≥50%) are represented by color-coded circles or squares, with squares showing fossil calibration points. Node bars reflect 95% confidence intervals based on 1,000 simulated trees. *Clemmys guttata* and *Emydoidea blandingii* represent outgroups.

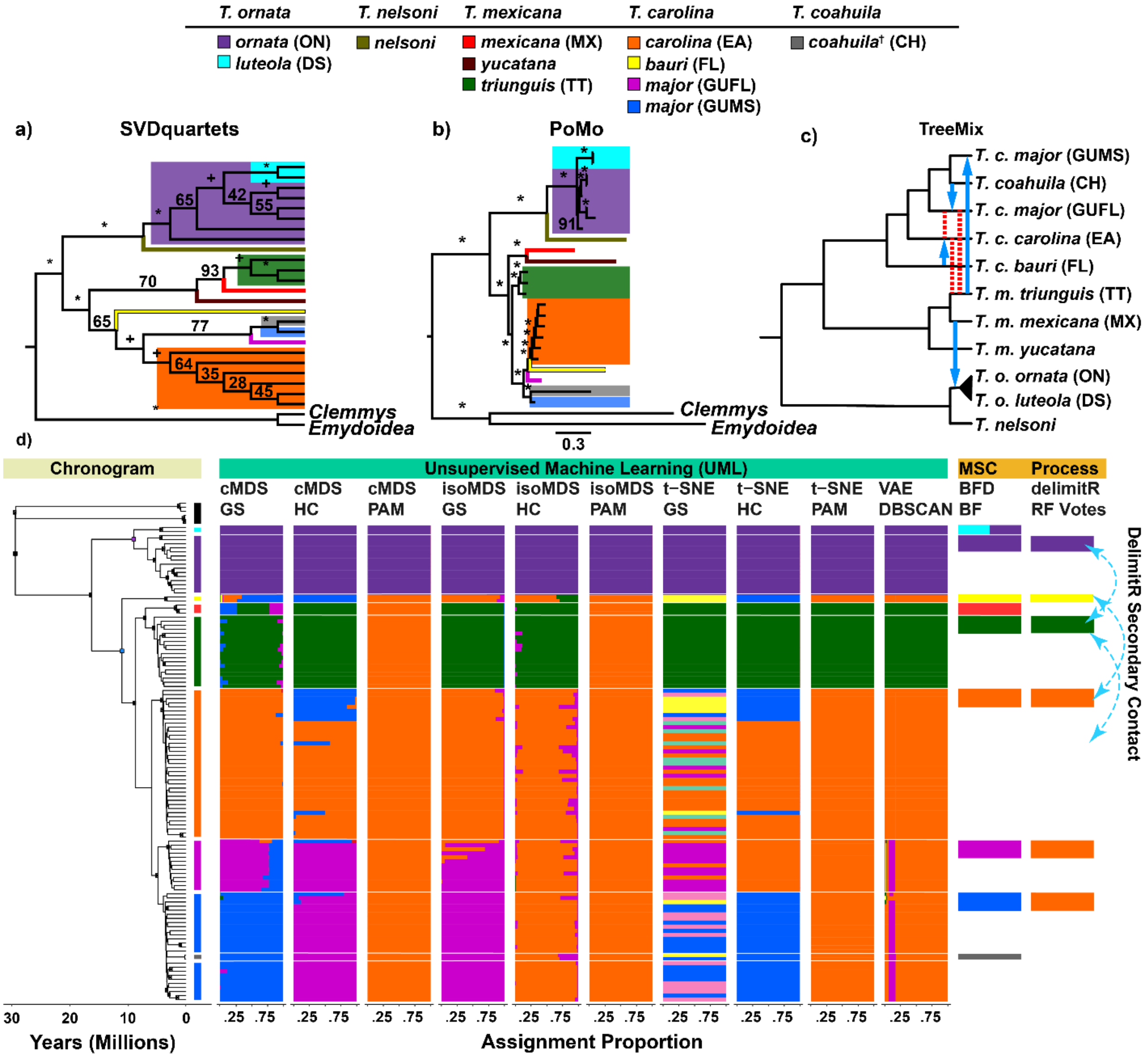


Figure 3: Species trees, TreeMix, and species delimitation results among *Terrapene* ddRADseq samples. Parenthetical legend abbreviations correspond to Tables 2 and 3. Phylogenies (N=214) were generated by (a) SVDquartets and (b) PoMo with 26 populations grouped by subspecies and state locality. '*' and '+' indicate 100% and ≥95% bootstrap support. (c) Migration supported by TreeMix (blue arrows) and previously published results (red/dashed lines; Martin *et al.* 2020). Outgroups were omitted for clarity. (d) Species delimitations for UML (N=117), multispecies coalescent (MSC; BFD=Bayes Factor Delimitation; N=37), and process-based (delimitR; N=28) methods. UML data filtering allowed ≤25% missing data per-individual and per-population, with minor allele frequency filters=5% (cMDS/t-SNE/VAE) and 1% (isoMDS), and t-SNE perplexity=15. UML includes RF=random forest, visualized with cMDS and isoMDS ordination, t-SNE, and VAE, with bar plots depicting assignment proportions among 100 replicates and aligning with chronogram tips. RF and t-SNE optimal *K* were assessed using partition around medoids (PAM)+gap statistic (GS), PAM+highest mean silhouette width (HMSW), and hierarchical clustering (HC)+HMSW, whereas VAE, BFD, and delimitR used DBSCAN, Bayes Factors (BF) and RF votes. Blue/dashed arrows show gene flow supported by delimitR. '†' indicates a monotypic *T. coahuila*.

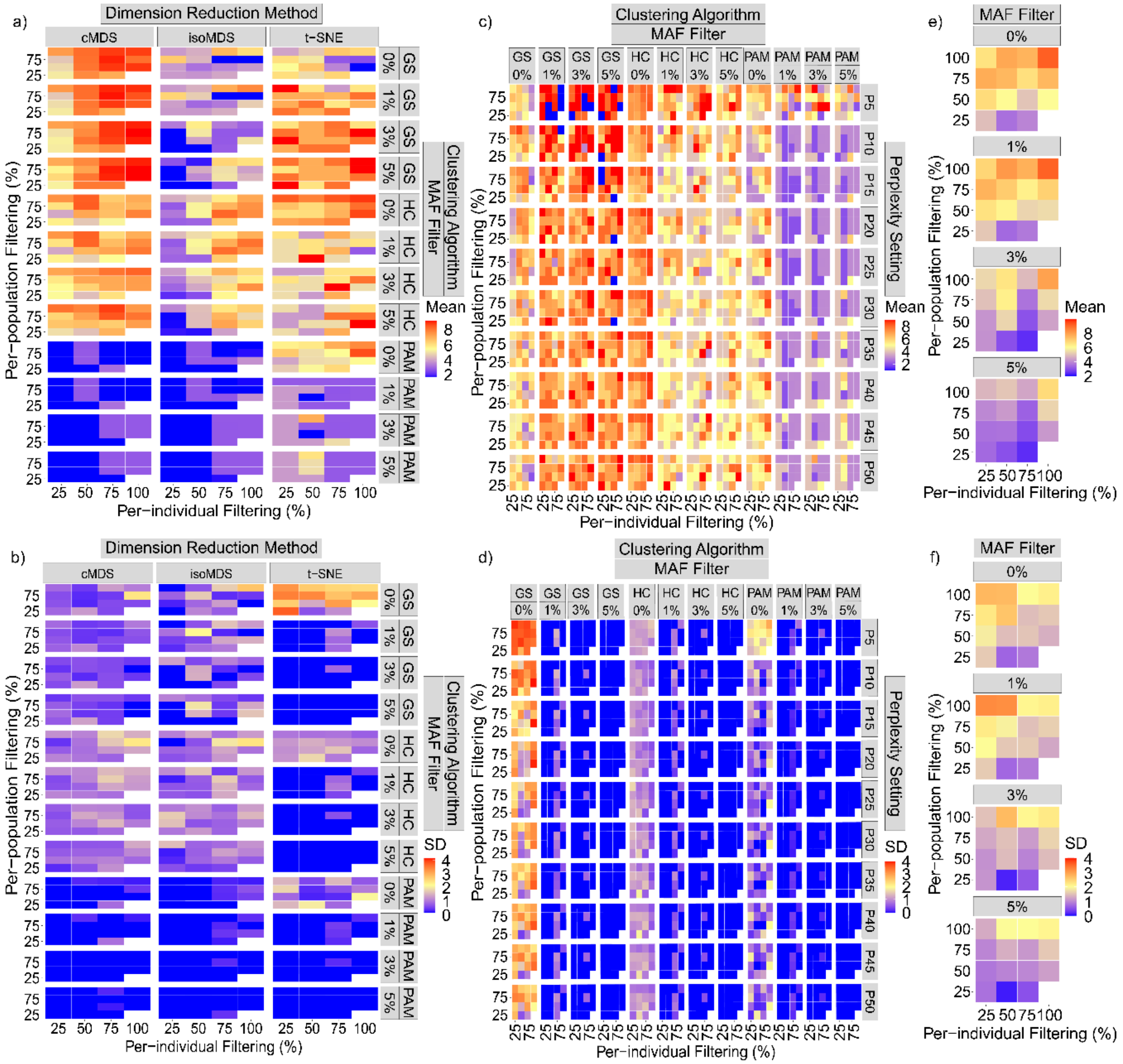


Figure 4: Heatmaps depicting mean and standard deviation (SD) of optimal *K* among 100 unsupervised machine learning species-delimitation replicates. Input ddRADseq alignments were filtered with a maximum of 25%, 50%, 75%, and 100% (=no filter) missing data allowed per-individual and per-population, and with minor allele frequency (MAF) filters as 5%, 3%, 1%, and 0% (=no filter). (a) and (b)=Pairwise missing data heatmaps for three dimensionality-reduction methods (CMDS and ISOMDS=classical and isotonic multidimensional scaling), T-SNE=t-distributed stochastic neighbor embedding *versus* three clustering algorithms [(partition around medoids+gap statistic (GS)]; HC=hierarchical clustering+highest mean silhouette width (HMSW); PAM=partition around medoids+HMSW. (c) and (d)=T-SNE heatmap panels comparing clustering algorithms with ten perplexity (P) settings. (e) and (f)=VAE (variational autoencoder) heatmaps with optimal *K* chosen via DBSCAN.

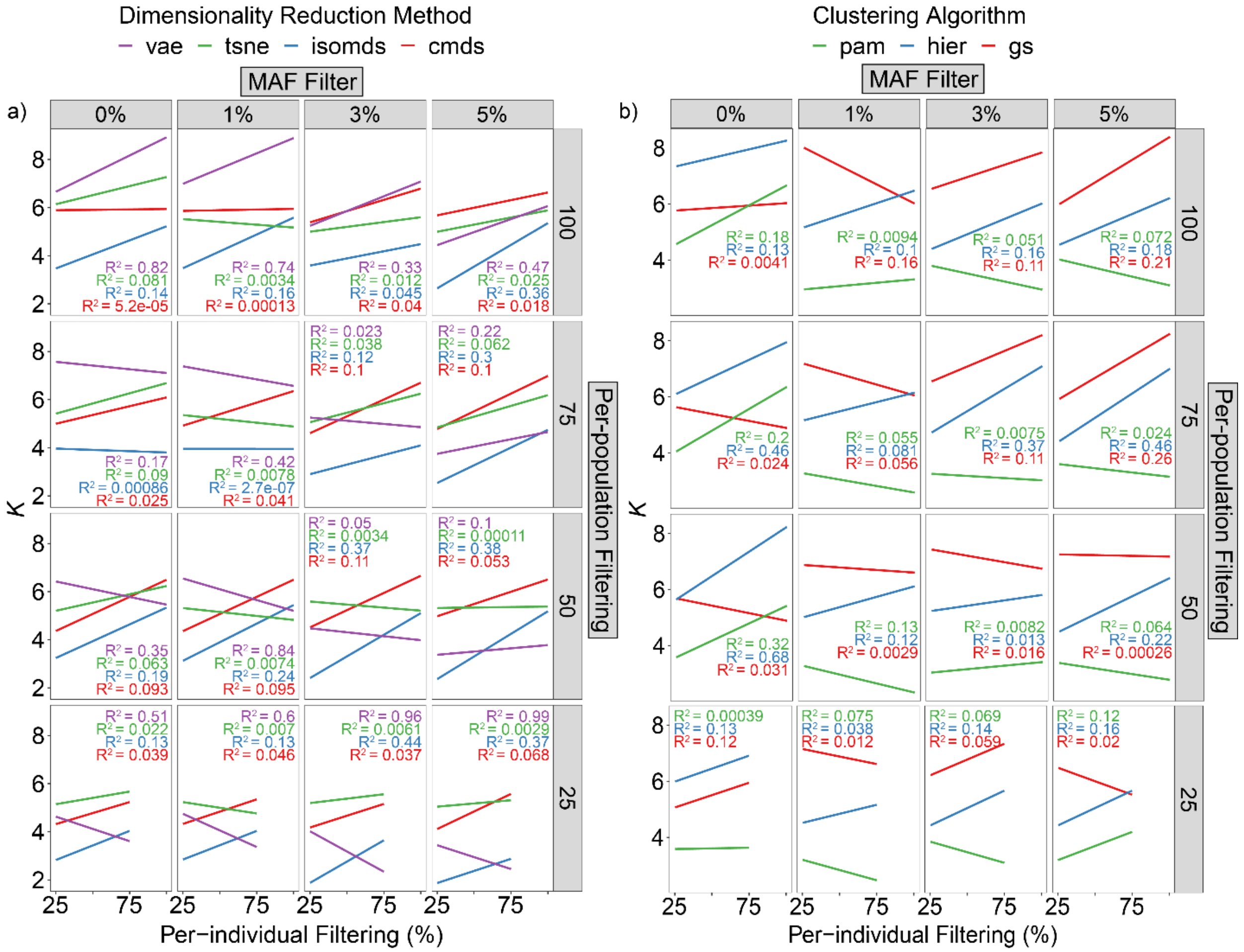


Figure 5: Regressions showing relationship between mean optimal *K* (y-axes), missing data, and minor allele frequency (MAF) filtering parameters. Missing data was filtered both per-individual (x-axes) and per-population (panel rows), with a maximum allowed of 25%, 50%, 75%, and 100% (=no filtering). Minor allele frequency (MAF) filters of 5%, 3%, 1%, and 0% (=no filtering) were also applied (panel columns). (a) Colors correspond to the dimensionality-reduction methods: cMDS and isoMDS=classical and isotonic multidimensional scaling, t-SNE=t-distributed stochastic neighbor embedding, VAE=variational autoencoder. (b) Colors indicate three clustering algorithms: GS=partition around medoids+gap statistic, HC=hierarchical clustering+highest mean silhouette width (HMSW), PAM=partition around medoids+HMSW.

SUPPLEMENTARY TABLES AND FIGURES

Table S1: *Terrapene* sample metadata. 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 generally follow the format "mm/year", unless only the year was known. Population codes precede the sample IDs with underscores as delimiters, with the first two characters representing subspecies (when available), and the second two U.S. state locality. "y" and "n" in the BFD column indicate individuals that were or were not used in the BFD analyses, respectively.

Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
T. coahuila	-	toenails	Coahuila, MX	-	01/2011	-	-
T. coahuila	-	toenails	Coahuila, MX	-	01/2011	-	-
T. ornata luteola	-	muscle	AZ	Cochise	-	31.6307	-109.198
T. ornata luteola	-	muscle	AZ	Cochise	-	31.6307	-109.198
T. ornata luteola	-	shell shavings	AZ	Cochise	-	31.7614	-109.26
T. ornata luteola	-	shell shavings	AZ	Cochise	-	31.9126	-109.151
T. ornata luteola	F	tail tip	NM	Socorro	08/2007	-	-
T. ornata luteola	-	tail tip	NM	Socorro	05/2007	-	-
T. ornata luteola	-	tail tip	NM	Socorro	-	-	-
T. ornata luteola	-	dried muscle	NM	Socorro	06/2008	-	-
T. carolina carolina	M	muscle	AL	Russell	06/2010	32.4654	-85.1998
T. carolina carolina	M	toe tips	AL	Bullock	06/2010	32.0841	-85.6902
T. carolina carolina	-	toenails	AL	Lee	05/2009	32.4399	-85.352
T. carolina carolina	F	tail tip	AL	Dekalb	05/2010	34.4455	-85.7772
T. carolina carolina	M	toenails	GA	Harris	07/2009	32.7381	-84.915
T. carolina carolina	M	toenails	GA	Harris	07/2009	32.7074	-84.9566
T. carolina carolina	F	toenails	GA	Marion	10/2009	32.3004	-84.5171
T. carolina carolina	F	toenails	GA	Harris	06/2009	32.6151	-84.8216
T. carolina carolina	F	toe tips	GA	Dekalb	05/2010	33.6654	-84.3467
T. carolina carolina	-	toenails	GA	Harris	05/2009	32.8514	-84.8459
T. carolina carolina	-	toenails	GA	Harris	05/2009	32.7795	-84.8739
T. carolina carolina	F	toenails	GA	Harris	05/2009	32.7859	-84.9566
T. carolina carolina	F	toenails	GA	Troup	06/2014	32.753	-84.9003
T. carolina carolina	-	toenails	GA	Harris	05/2009	32.695	-84.959
T. carolina carolina	F	toenails	GA	Harris	05/2009	32.766	-84.9082
T. carolina carolina	-	toenails	GA	Harris	06/2009	32.6683	-84.956
T. carolina carolina	F	toenails	GA	Harris	06/2009	32.597	-84.8266
T. carolina carolina	-	toenails	GA	Harris	06/2009	32.8013	-84.9217
T. carolina carolina	M	toenails	GA	Harris	06/2009	32.8436	-84.9379
T. carolina carolina	-	toenails	GA	Harris	10/2009	32.7038	-84.7357
T. carolina carolina	M	toenails	KY	Carter	-	-	-

Table S1 (Cont.)

Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
T. carolina carolina	-	-	KY	Laurel	08/2004	37.0029	-84.2375
T. carolina carolina	-	-	KY	Leslie	08/2004	37.2459	-83.3816
T. carolina carolina	-	-	NC	Johnston	07/2009	35.6838	-78.4682
T. carolina carolina	-	-	NC	Johnston	08/2009	35.6832	-78.4621
T. carolina carolina	M	tail tip	NY	Westchester	-	41.2884	-73.8658
T. carolina carolina	F	tail tip	PA	-	-	40.0851	-76.8906
T. carolina carolina	M	toenails	RI	Washington	06/2011	-	-
T. carolina carolina	M	shell shavings	SC	Beaufort	06/2010	32.3261	80.69663
T. carolina carolina	M	toenails	SC	Beaufort	06/2010	32.3402	80.69961
T. carolina carolina	M	muscle	SC	Jasper	06/2010	32.447	-81.1053
T. carolina carolina	F	toenails	SC	Beaufort	06/2010	32.3393	-80.7006
T. carolina carolina	F	toenails	SC	Beaufort	06/2010	32.3575	-80.7093
T. carolina carolina	M	toenails	SC	Beaufort	06/2010	32.318	-80.6866
T. carolina carolina	F	toenails	SC	Beaufort	06/2010	32.3509	-80.696
T. carolina carolina	-	muscle	SC	Laurens	06/2009	34.4524	-81.8601
T. carolina carolina	M	toenails	SC	Abbeville	10/2010	34.3276	-82.3797
T. carolina carolina	-	toenails	SC	Chester	10/2010	34.6807	-81.1752
T. carolina carolina	F	toenails	SC	Greenville	10/2010	34.8298	-82.394
T. carolina carolina	M	toenails	SC	Abbeville	10/2010	34.3319	-82.3865
T. carolina carolina	-	tail tip	TN	Davidson	-	36.1346	-86.9305
T. carolina carolina	-	tail tip	TN	Davidson	-	36.1302	-86.8651
T. carolina carolina	M	-	VA	Albemarle	-	-	-
T. carolina carolina	M	-	VA	Fluvanna	-	38.0382	-78.9138
T. carolina carolina	M	-	VA	Fluvanna	-	38.0382	-78.9138
T. carolina carolina	M	toenails	VA	Norfolk	-	36.6787	-76.2937
T. carolina carolina	M	toenails	VA	Dinwiddie	-	37.2176	-77.3915
T. carolina carolina	-	tail tip	WV	Roane	07/2009	38.542	-81.3251
T. carolina bauri	F	toenails	FL	Alachua	06/2009	29.6436	-82.3457
T. carolina bauri	-	muscle	FL	Taylor	04/2008	29.7745	-83.5711
T. carolina bauri	F	toenails	FL	-	06/2009	-	-
T. carolina bauri	-	muscle	FL	Hernando	06/2007	28.593	-82.3708
T. carolina major	-	scutes	FL	Franklin	05/2009	29.9422	-85.0068
T. carolina major	-	toenails	FL	Franklin	05/2009	29.7981	-84.8308
T. carolina major	-	muscle	FL	Wakulla	-	30.0588	-84.5688
T. carolina major	M	muscle	FL	Walton	04/2010	30.4428	-85.9583
T. carolina major	F	muscle	FL	Walton	06/2009	30.4932	-85.9365
T. carolina major	M	muscle	FL	Calhoun	04/2009	30.4261	-85.1194
T. carolina major	M	toenails	FL	Gulf	06/2009	30.0581	-85.1898
T. carolina major	F	toenails	FL	Gulf	06/2009	29.8199	-85.2836
T. carolina major	-	toe tips	FL	Okaloosa	07/2009	30.6725	-86.6317

Table S1 (Cont.)

Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
T. carolina major	-	muscle	FL	Escambia	-	30.5694	-87.3991
T. carolina major	-	muscle	FL	Gulf	03/2007	29.8473	-85.2616
T. carolina major	M	toenails	FL	Gulf	09/2013	30.0785	-85.1943
T. carolina major	F	tail tip	FL	Gulf	07/2013	29.9909	-85.1696
T. carolina major	M	toenails	FL	Gulf	07/2013	29.685	-85.3277
T. carolina major	-	tail tip	FL	Gulf	07/2012	30.1579	-85.2093
T. carolina major	F	tail tip	FL	Gulf	07/2015	29.6914	-85.2511
T. carolina major	M	tail tip	FL	Gulf	07/2015	29.8875	-85.2184
T. carolina major	F	foot tissue	FL	Gulf	05/2015	29.8731	-85.2297
T. carolina major	F	toenails	FL	Calhoun	05/2015	30.5039	-85.1173
T. carolina major	F	toenails, muscle	FL	Gulf	05/2015	29.8375	-85.273
T. carolina major	F	toenail	FL	Gulf	05/2015	30.0036	-85.1735
T. carolina major	F	tail tip	FL	Franklin	07/2016	29.7224	-84.9897
T. carolina major	F	toenails	FL	Gulf	07/2016	29.9291	-85.3929
T. carolina major	M	toenails	FL	Liberty	03/2018	30.3715	-84.6813
T. carolina major	F	toenails	FL	Franklin	04/2018	30.0049	-84.886
T. carolina major	F	toenails	FL	Bay	05/2010	30.1851	-85.6819
T. carolina major	-	scutes	FL	Franklin	05/2009	29.8647	-84.7533
T. carolina major	M	toenails with skin	FL	Gulf	09/2013	30.0785	-85.1943
T. carolina major	F	toenails	FL	Franklin	02/2018	29.8814	-84.7286
T. carolina major	M	toenails	FL	Franklin	03/2018	29.8398	-84.6809
T. carolina major	-	toenails	FL	Franklin	03/2018	29.7537	-84.8423
T. carolina major	-	toenails	FL	Franklin	04/2018	29.85	-84.6881
T. carolina major	M	tail tip	MS	Perry	04/2009	31.1418	-89.1517
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	toenails	MS	Jackson	04/2015	30.4241	-88.5167
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	tail tip	MS	Jackson	07/2015	30.4401	-88.5494
T. carolina major	F	tail tip	MS	Jackson	05/2015	32.5546	-85.6424
T. carolina major	M	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	toenails	MS	Jackson	04/2015	30.4241	-88.5167
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	tail tip	MS	Jackson	07/2015	30.4401	-88.5494
T. carolina major	M	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	M	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4408	-88.5512
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381

Table S1 (Cont.)

Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4381	-88.5475
T. carolina major	F	tail tip	MS	Jackson	05/2015	30.4317	-88.5381
T. carolina major	-	tail tip	MS	Jackson	04/2018	30.376	-88.7107
T. carolina major	F	tail tip	MS	Jackson	05/2018	30.3756	-88.7095
T. carolina major	M	foot tissue	MS	Jackson	05/2018	30.3756	-88.7124
T. carolina major	F	tail tip	MS	Jackson	05/2018	30.3757	-88.7049
T. carolina major	M	tail tip	MS	Jackson	05/2018	30.3758	-88.7179
T. carolina major	-	tail tip	MS	Jackson	05/2018	30.3955	-88.7309
T. carolina major	F	tail tip	MS	Jackson	05/2018	30.3558	-88.7095
T. carolina major	M	skin	MS	Jackson	05/2018	30.3662	-88.6905
T. carolina major	M	tail tip	MS	Jackson	05/2018	30.3557	-88.6893
T. carolina major	F	tail tip	MS	Jackson	05/2018	30.3757	-88.7187
T. carolina major	F	toe	MS	Jackson	06/2018	30.3757	-88.7056
T. carolina major	M	tail tip	MS	Jackson	06/2018	30.3754	-88.7077
T. carolina major	M	tail tip	MS	Jackson	06/2018	30.4005	-88.7543
T. carolina major	M	foot tissue	MS	Jackson	06/2018	30.375	-88.7076
T. mexicana mexicana	-	toenails	Tamaulipas, MX	-	01/2011	-	-
T. mexicana mexicana	-	toenails	Tamaulipas, MX	-	01/2011	-	-
T. mexicana mexicana	-	toenails	Tamaulipas, MX	-	01/2011	-	-
Clemmys guttata	M	blood	IL	Will	-	-	-
Clemmys guttata	M	blood	IL	Will	-	-	-
Clemmys guttata	M	blood	IL	Will	-	-	-
Emydoidea blandingii	-	blood	IL	Will	-	-	-
Emydoidea blandingii	-	blood	IL	Will	-	-	-
Emydoidea blandingii	-	blood	IL	Will	-	-	-
T. ornata ornata	F	toenails	CO	Weld	06/2009	40.2992	-104.475
T. ornata ornata	M	toenails	CO	Weld	06/2009	40.2947	-104.476
T. ornata ornata	F	toenails	CO	Weld	05/2009	40.2989	-104.479
T. ornata ornata	M	toenails	CO	Weld	05/2009	40.2935	-104.481
T. ornata ornata	-	blood	IL	Marion	2016	-	-
T. ornata ornata	-	blood	IL	Lee	2016	-	-
T. ornata ornata	-	blood	IL	Lee	2013	41.9087	-89.3451
T. ornata ornata	-	tail tip	KS	Clark	06/2009	37.4071	-99.7555
T. ornata ornata	F	tail tip	KS	Meade	06/2009	37.3809	-100.141
T. ornata ornata	-	tail tip	KS	Meade	06/2009	37.0441	-100.494
T. ornata ornata	-	tail tip	KS	Meade	06/2009	37.0656	-100.471
T. ornata ornata	-	tail tip	KS	Meade	06/2009	37.2855	-100.369
T. ornata ornata	F	tail tip	KS	Miami	-	38.5476	-94.9369
T. ornata ornata	F	tail tip	KS	Osage	-	38.7823	-95.5135
T. ornata ornata	-	tail tip	NE	Box Butte	06/2009	42.0886	-102.723
T. ornata ornata	-	tail tip	NE	Sheridan	06/2009	42.059	-102.461

Table S1 (Cont.)

Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
T. ornata ornata	-	blood	NE	Garden	06/2009	41.8305	-102.342
T. ornata ornata	-	tail tip	NE	Box Butte	06/2009	42.0898	-102.736
T. ornata ornata	-	muscle	TX	Montague	10/2009	33.4824	-97.7948
T. ornata ornata	M	-	TX	Calhoun	05/2009	28.1958	-96.7004
T. ornata ornata	M	-	TX	Calhoun	06/2009	28.2896	-96.5278
T. ornata ornata	-	toenails	WI	Sauk	-	43.1753	-90.0711
T. ornata ornata	F	toenails	WI	Iowa	-	43.0305	-90.1094
T. ornata ornata	F	toenails	WI	Iowa	-	43.0305	-90.1094
T. ornata ornata	F	toenails	WI	Iowa	-	43.0305	-90.1094
T. ornata ornata	F	toenails	WI	Iowa	-	43.0305	-90.1094
T. ornata ornata	F	toenails	WI	Dane	-	43.1767	-89.7994
T. ornata ornata	F	toenails	WI	Columbia	-	43.458	-89.3883
T. ornata ornata	F	toenails	WI	Columbia	-	43.4514	-89.3532
T. nelsoni	-	toenails	Sonora, MX	-	10/2011	29.9113	-111.027
T. carolina	M	tail tip	AL	Russell	09/2011	32.255	-85.4165
T. carolina	-	toenails	AL	Macon	07/2009	32.4777	-85.7977
T. carolina	-	tail tip	AL	Macon	08/2009	32.5139	-85.6096
T. carolina	F	toenails with muscle	AL	Elmore	08/2009	32.5733	-86.0344
T. carolina	M	tail tip	AL	Macon	08/2009	32.445	-85.8103
T. carolina	F	tail tip	AL	Lee	06/2009	32.5377	-85.5042
T. carolina	F	tail tip	AL	Chambers	06/2009	32.7701	-85.259
T. carolina	M	toenails	AL	Russell	08/2009	32.2568	-85.354
T. carolina	F	toenails	AL	Barbour	03/2010	32.0095	-85.404
T. carolina	F	muscle	AL	Macon	07/2010	32.4553	-85.6562
T. carolina	-	tail tip	AL	Conecuh	08/2016	31.2888	-87.1884
T. carolina	F	toenails, muscle	AL	Bullock	07/2013	32.0826	-85.6897
T. carolina	F	toenails	AL	Mobile	04/2014	30.8447	-88.3953
T. carolina	-	toenails with skin	AL	Randolph	07/2013	33.1354	-85.465
T. carolina	F	toenails with skin	AL	Bullock	07/2013	32.0826	-85.6897
T. carolina	M	toenails	AL	Madison	07/2010	34.6109	-86.5778
T. carolina	M	muscle	AL	Clay	05/2010	33.1264	-85.8595
T. carolina	F	muscle	AL	Clay	05/2010	33.2063	-85.8222
T. carolina	F	toenails	GA	Harris	06/2009	32.6649	-84.9763
T. carolina	M	toenails	GA	Harris	06/2009	32.7432	-84.9036
T. mexicana triunguis	M	toenails	AR	Pulaski	08/2009	34.8347	-92.4916
T. mexicana triunguis	F	toenails	AR	Pulaski	07/2009	34.8347	-92.4915
T. mexicana triunguis	M	toenails	AR	Pulaski	07/2009	34.8347	-92.4915
T. mexicana triunguis	F	tail tip	KS	Crawford	05/2009	37.5874	-94.9584
T. mexicana triunguis	M	toenails	LA	Rapides	06/2009	31.2036	-92.5784
T. mexicana triunguis	F	shell shavings	LA	Rapides	06/2009	31.1811	-92.5562
T. mexicana triunguis	F	toenails	LA	Rapides	06/2010	31.1552	-92.5231

Table S1 (Cont.)

Taxonomic ID	Sex	Tissue	State	County	Date	Lat	Long
T. mexicana triunguis	F	toenails	LA	Rapides	06/2010	31.247	-92.6409
T. mexicana triunguis	M	toenails	LA	Rapides	05/2010	31.1387	-92.6377
T. mexicana triunguis	M	toenails	LA	Rapides	05/2010	31.2058	-92.5836
T. mexicana triunguis	F	toenails	LA	Rapides	05/2010	31.2058	-92.5836
T. mexicana triunguis	M	toenails	LA	Rapides	04/2010	31.156	-92.5226
T. mexicana triunguis	F	toenails	LA	Rapides	04/2010	31.1827	-92.5198
T. mexicana triunguis	M	toenails	MO	Jefferson	-	38.1959	-90.5324
T. mexicana triunguis	F	toenails	TX	Smith	05/2008	32.2963	-95.2086
T. mexicana triunguis	M	toenails	TX	Dallas	05/2008	32.9483	-96.7299
T. mexicana triunguis	M	toenails	TX	Smith	05/2008	32.3513	-95.3011
T. mexicana triunguis	F	toenails	TX	Dallas	05/2008	32.9702	-96.7364
T. mexicana triunguis	M	toenails	TX	Tarrant	05/2008	33.0352	-97.1173
T. mexicana triunguis	-	tail tip	TX	Henderson	08/2009	32.3373	-95.7455
T. mexicana triunguis	-	toenails	TX	Smith	-	32.345	-95.2668
T. mexicana triunguis	-	toenails	TX	Smith	08/2009	32.2564	-95.1869
T. mexicana triunguis	-	toenails	TX	Smith	08/2009	32.345	-95.2668
T. mexicana triunguis	-	toenails	TX	Smith	08/2009	32.345	-95.2668
T. mexicana triunguis	-	toenails	TX	Smith	08/2009	32.345	-95.2668
T. mexicana triunguis	M	toenails	TX	Collin	05/2006	33.2162	-96.5723
T. mexicana yucatana	F	toenails	Yucatan, MX	-	05/2010	20.1417	-89.2092

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.

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

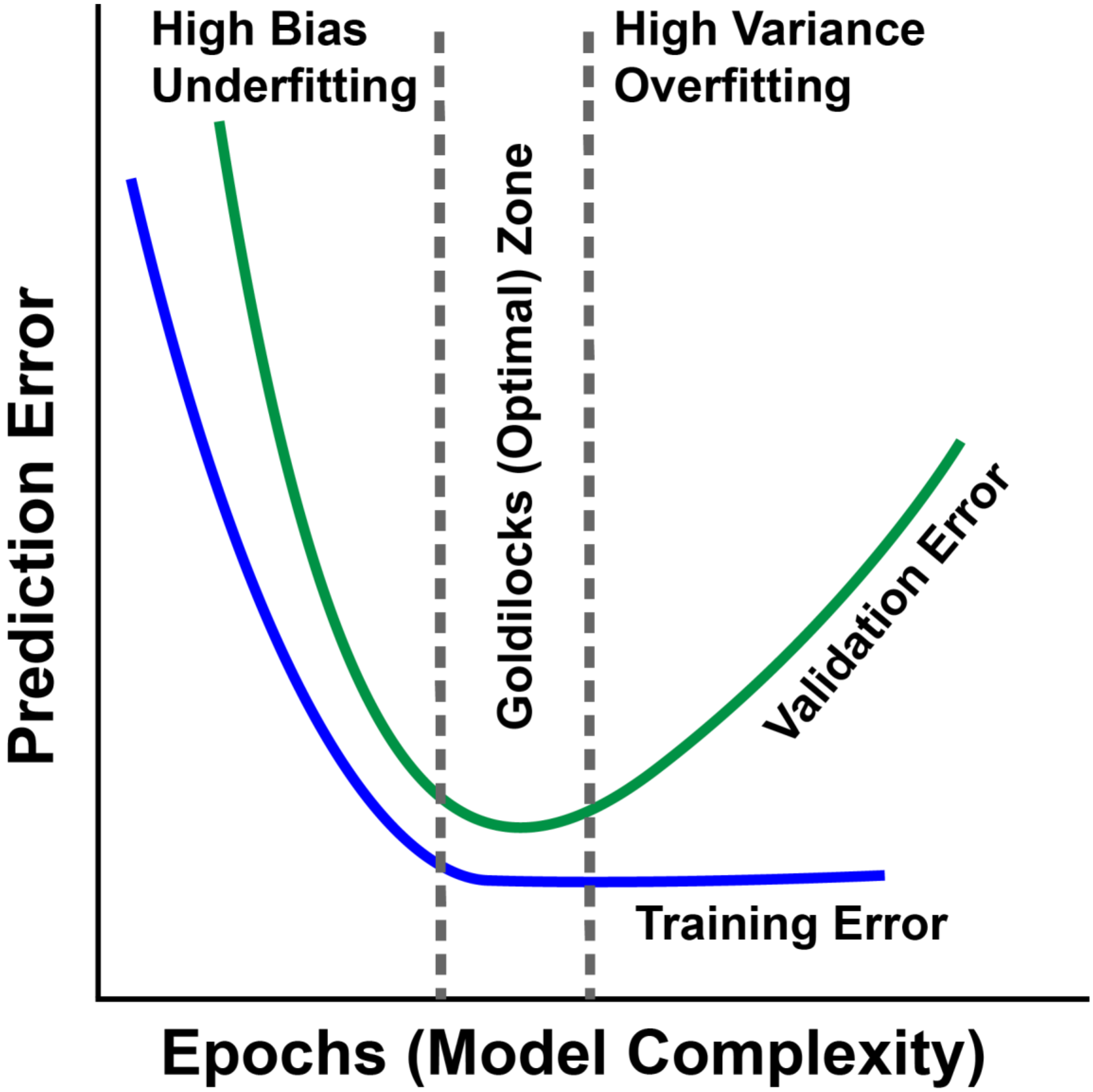


Figure S1: Prediction error *versus* model complexity for machine learning. Ideally, training should stop when the validation and training prediction error are at their lowest point [‘Goldilocks zone’, indicated by gray dashed lines; (Al’Aref *et al.* 2019)]. High bias and underfitting occur when training ends prior to reaching the Goldilocks zone, and high variance and overfitting occur when validation error begins to climb.

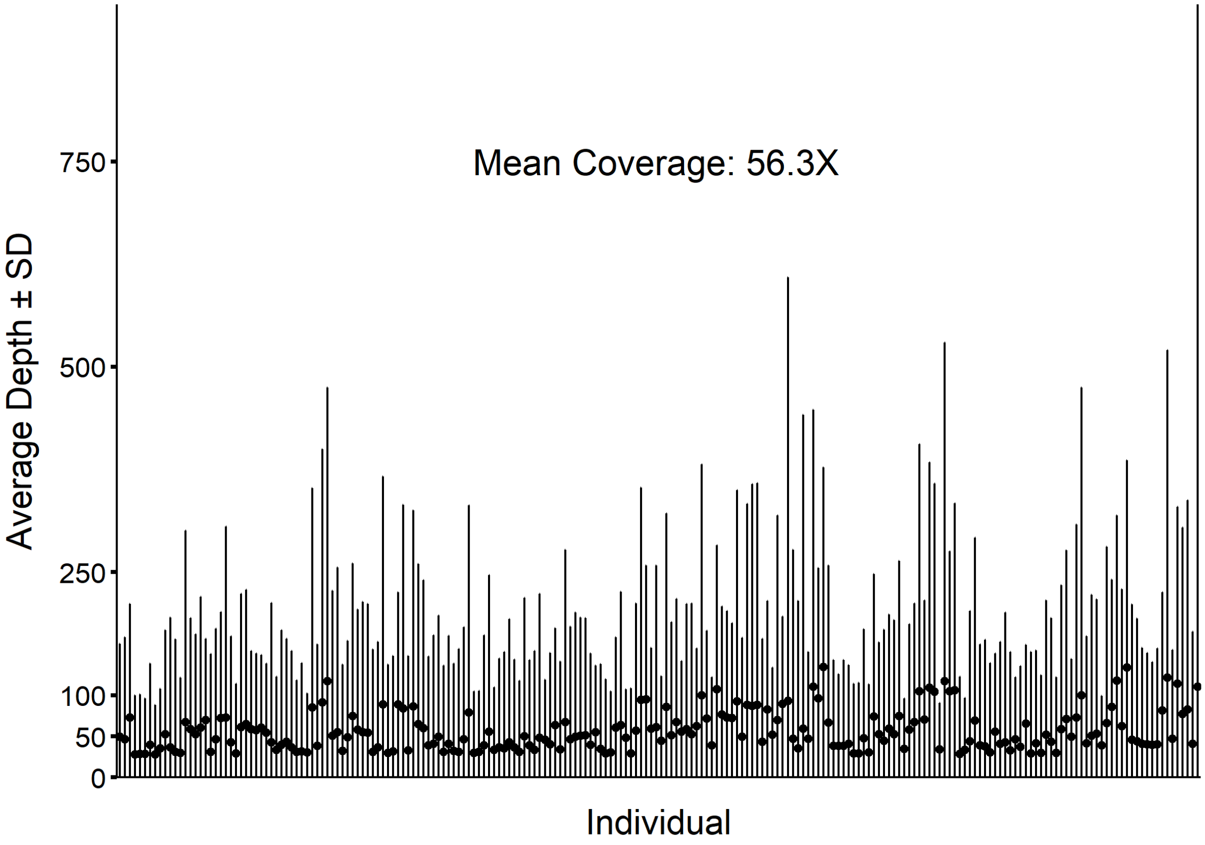


Figure S2: Mean per-individual read depth (points) ± standard deviation (gray bars) for 214 North American box turtle (*Terrapene* spp.) ddRAD sequencing samples.

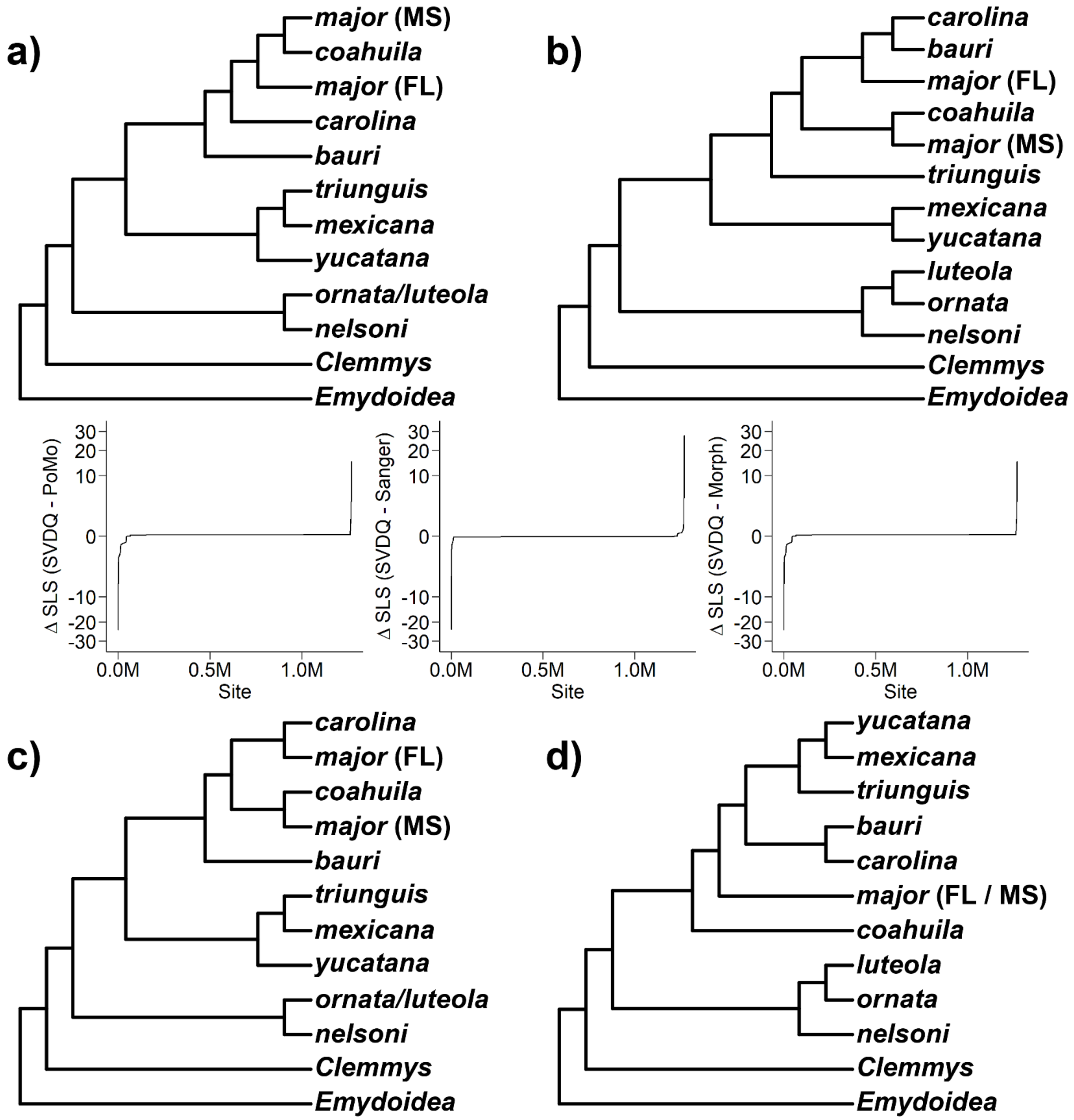


Figure S3: *Terrapene* constraint trees and respective changes in site-likelihood scores (**Δ**SLS) representing genome-wide support for the a) SVDQUARTETS, b) POMO, c) Sanger, and d) Morphological phylogenetic hypotheses. The SVDQUARTETS and POMO trees are derived in this study, whereas the Sanger and morphological hypotheses are results previously published (Minx 1996; Martin *et al.* 2013).

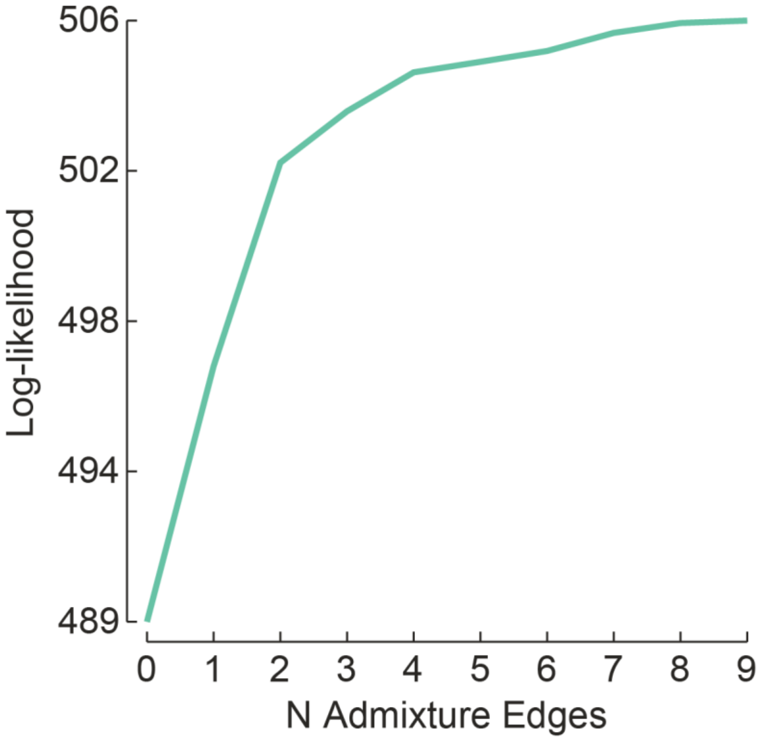


Figure S4: Log-likelihoods for the number (N) of TREEMIX admixture edges.

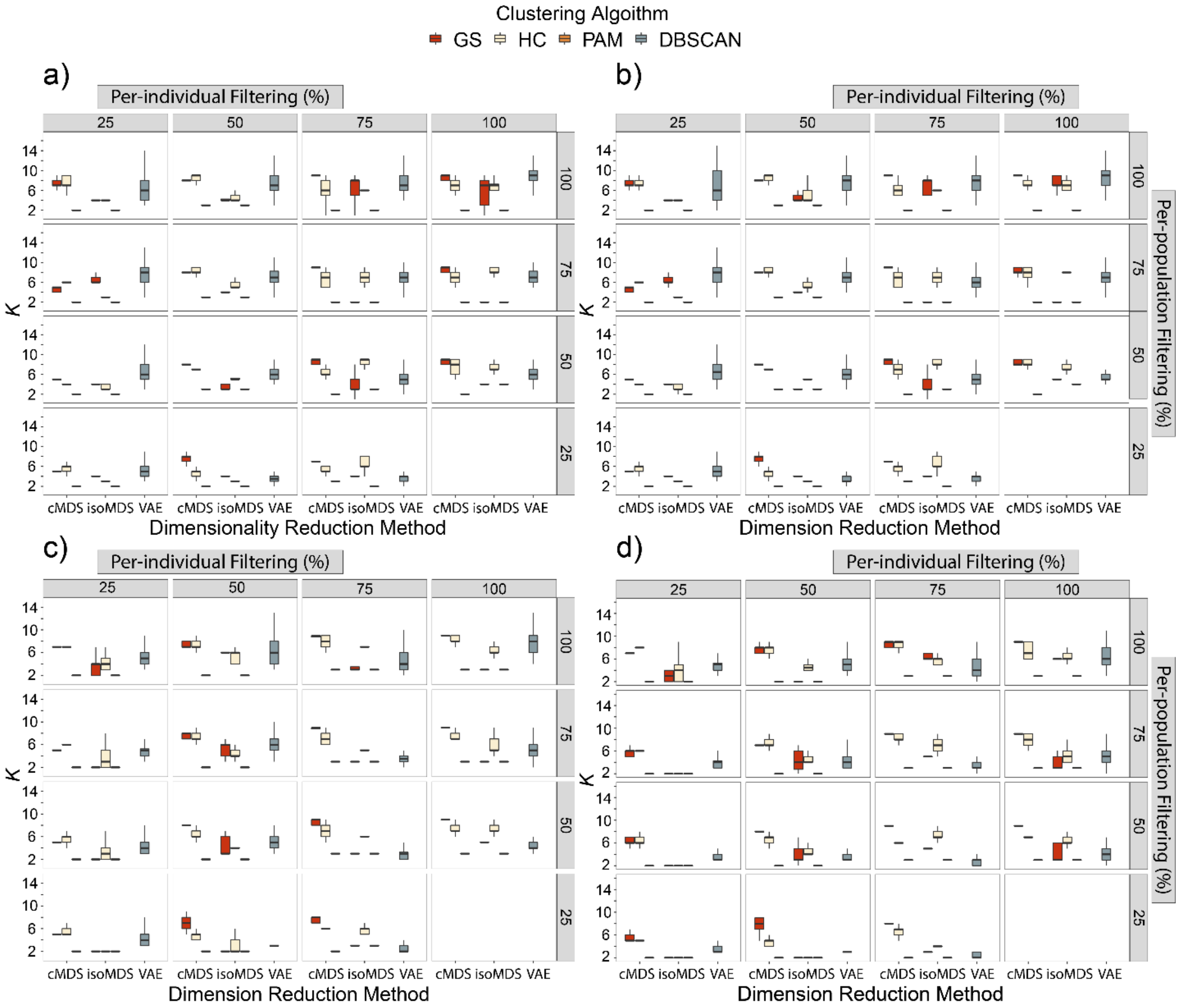


Figure S5: Tukey-style box and whiskers plots for 100 unsupervised machine learning (UML) species delimitation replicates. Data were filtered using per-individual (panel columns) and per-population (panel rows) missing data filters (25%=most stringent; 100%=no filtering) and a) 0%, b) 1%, c) 3%, and d) 5% minor allele frequency filters. Black bars on the boxplots indicate the median, lower and upper hinges represent the 25th and 75th percentiles, and the whisker range includes 1.5 times the inter-quartile range (IQR) past the hinges. cMDS and isoMDS = random forest classification visualized with classical and isotonic multidimensional scaling; VAE=variational autoencoder. Optimal *K* among cMDS and isoMDS was determined in three ways among two clustering algorithms: partition around medoids (PAM) with the gap statistic (GS), hierarchical clustering with the highest mean silhouette width (HMSW), and PAM with HMSW. Optimal *K* for VAE was determined using DBSCAN.

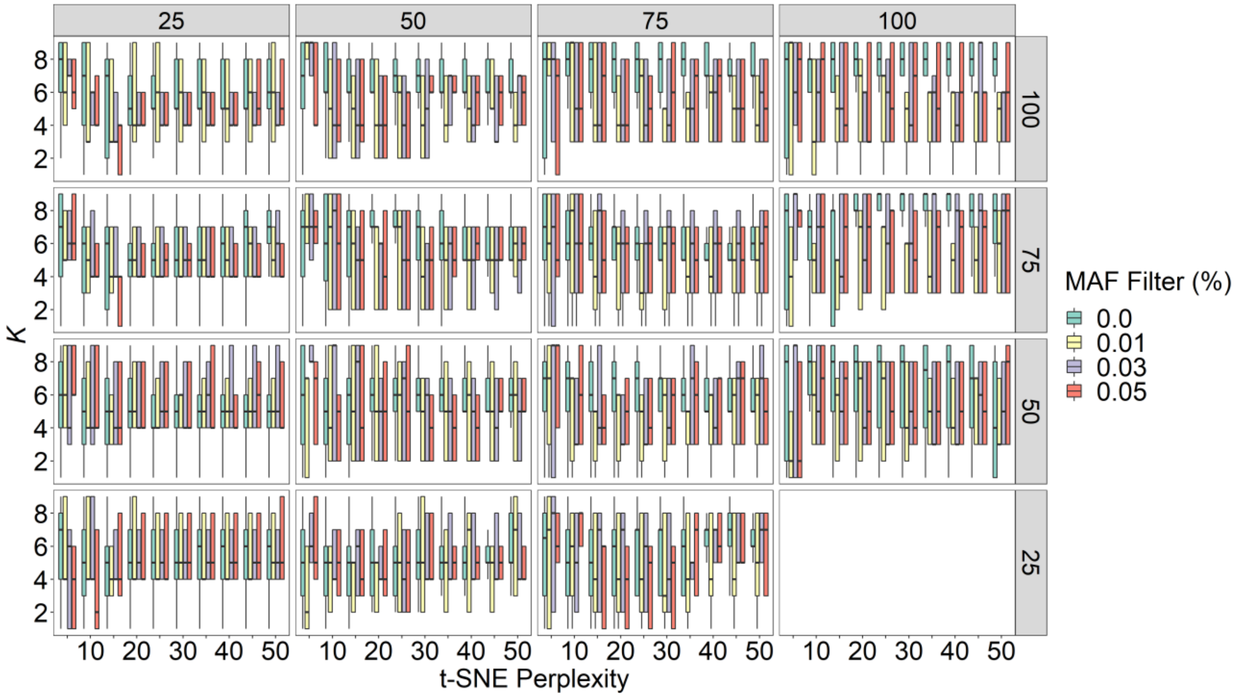


Figure S6: Tukey-style box and whiskers plots depicting variation in mean optimal K among a t-SNE perplexity grid search (x axes). Data were filtered with 25%, 50%, 75%, and 100% (no filtering) per-individual (panel columns) and per-population (panel rows) missing data filters and 0%, 1%, 3%, and 5% minor allele frequency (MAF) filters. Black bars on the boxplots indicate the median, lower and upper hinges represent the 25th and 75th percentiles, and the whisker range includes 1.5 times the inter-quartile range (IQR) past the hinges. t-SNE = t-distributed stochastic neighbor embedding.

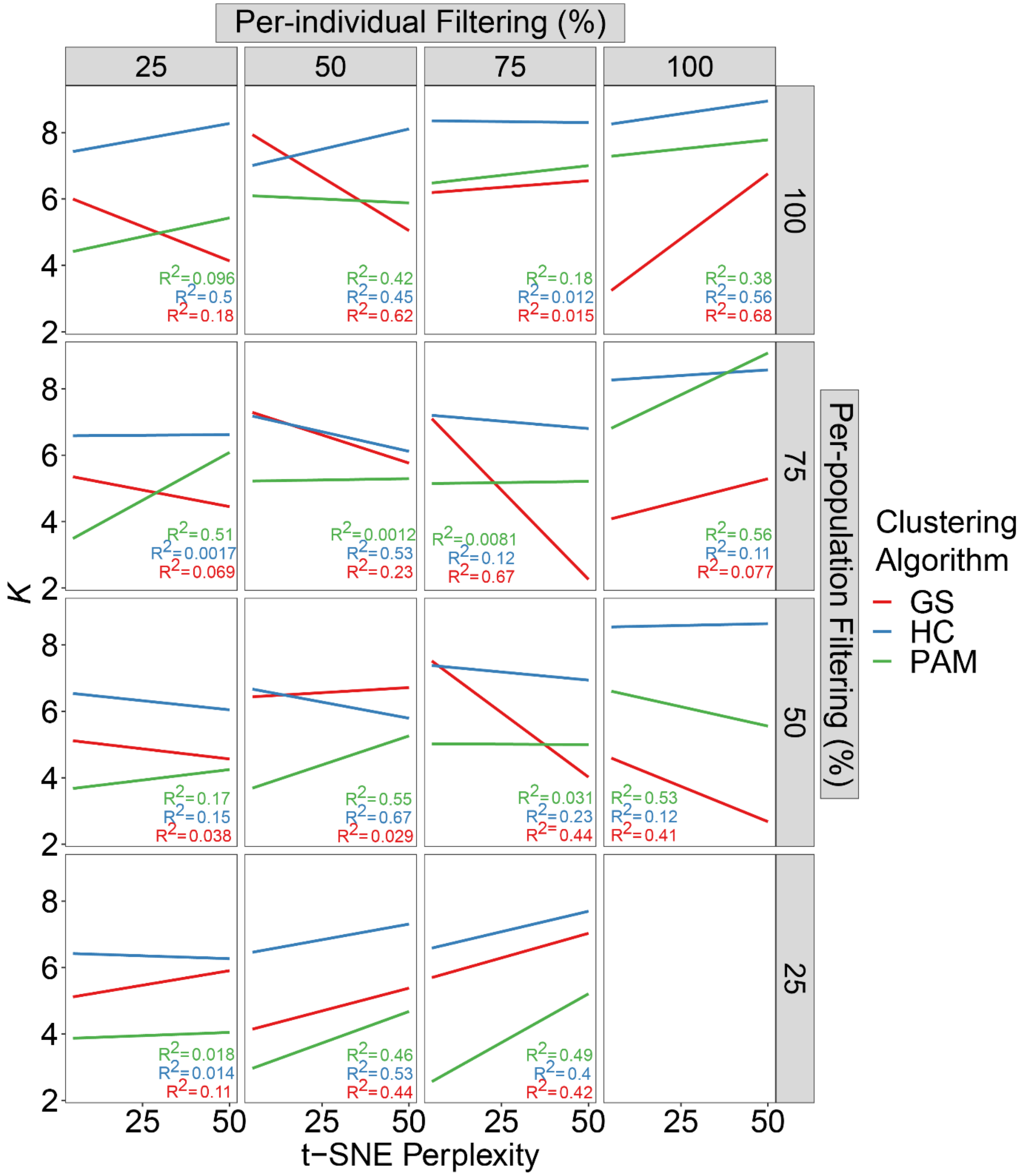


Figure S7: Regressions showing mean optimal *K* among ten t-SNE perplexity settings and three clustering algorithms for a minor allele frequency (MAF) filter=0% (no filtering). Panel columns and rows represent per-individual and per-population missing data filters, respectively. Mean optimal *K* was chosen using three clustering algorithms (fill colors): 1) Partition around medoids (PAM) with the gap statistic (GS), hierarchical clustering (HC) with the highest mean silhouette width (HMSW), and PAM with HMSW. R^2 values show correlations per clustering algorithm.

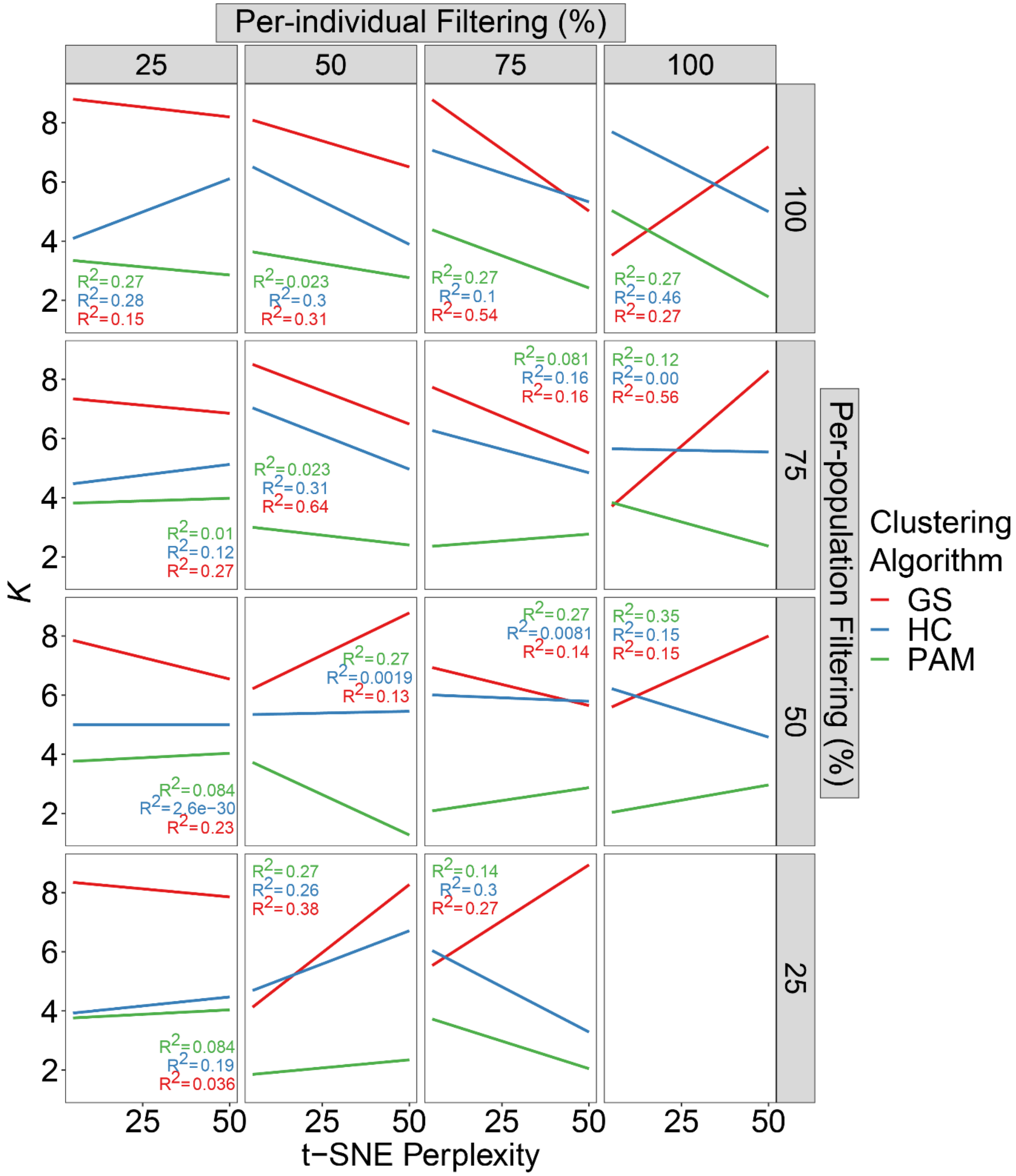


Figure S8: Regressions showing mean optimal *K* among ten t-SNE perplexity settings and three clustering algorithms for a minor allele frequency (MAF) filter=1%. Panel columns and rows represent per-individual and per-population missing data filters, respectively. Mean optimal *K* was chosen using three clustering algorithms (fill colors): 1) Partition around medoids (PAM) with the gap statistic (GS), hierarchical clustering (HC) with the highest mean silhouette width (HMSW), and PAM with HMSW. R^2 values show correlations per clustering algorithm.

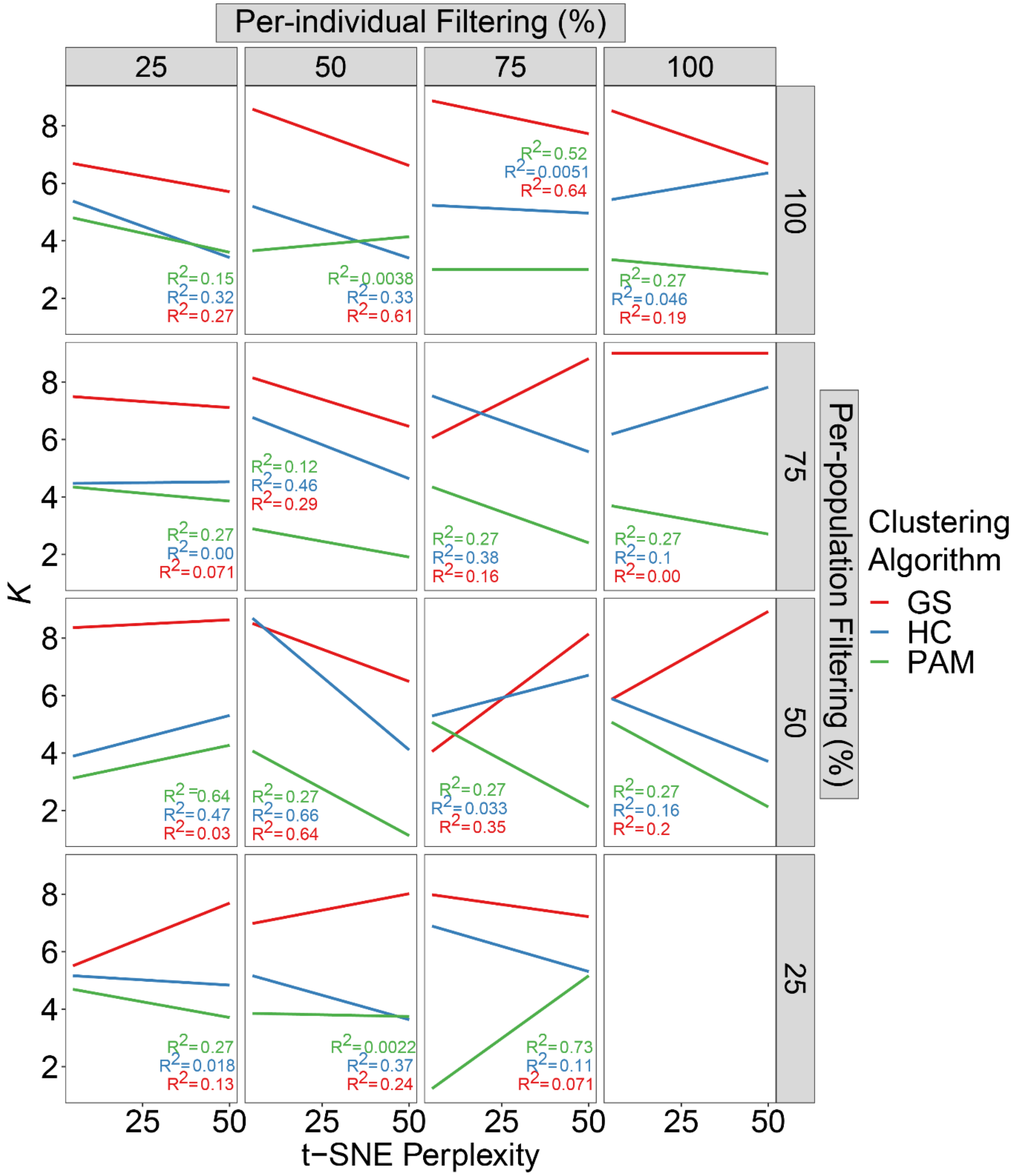


Figure S9: Regressions showing mean optimal *K* among ten t-SNE perplexity settings and three clustering algorithms for a minor allele frequency (MAF) filter=3%. Panel columns and rows represent per-individual and per-population missing data filters, respectively. Mean optimal *K* was chosen using three clustering algorithms (fill colors): 1) Partition around medoids (PAM) with the gap statistic (GS), hierarchical clustering (HC) with the highest mean silhouette width (HMSW), and PAM with HMSW. R^2 values show correlations per clustering algorithm.

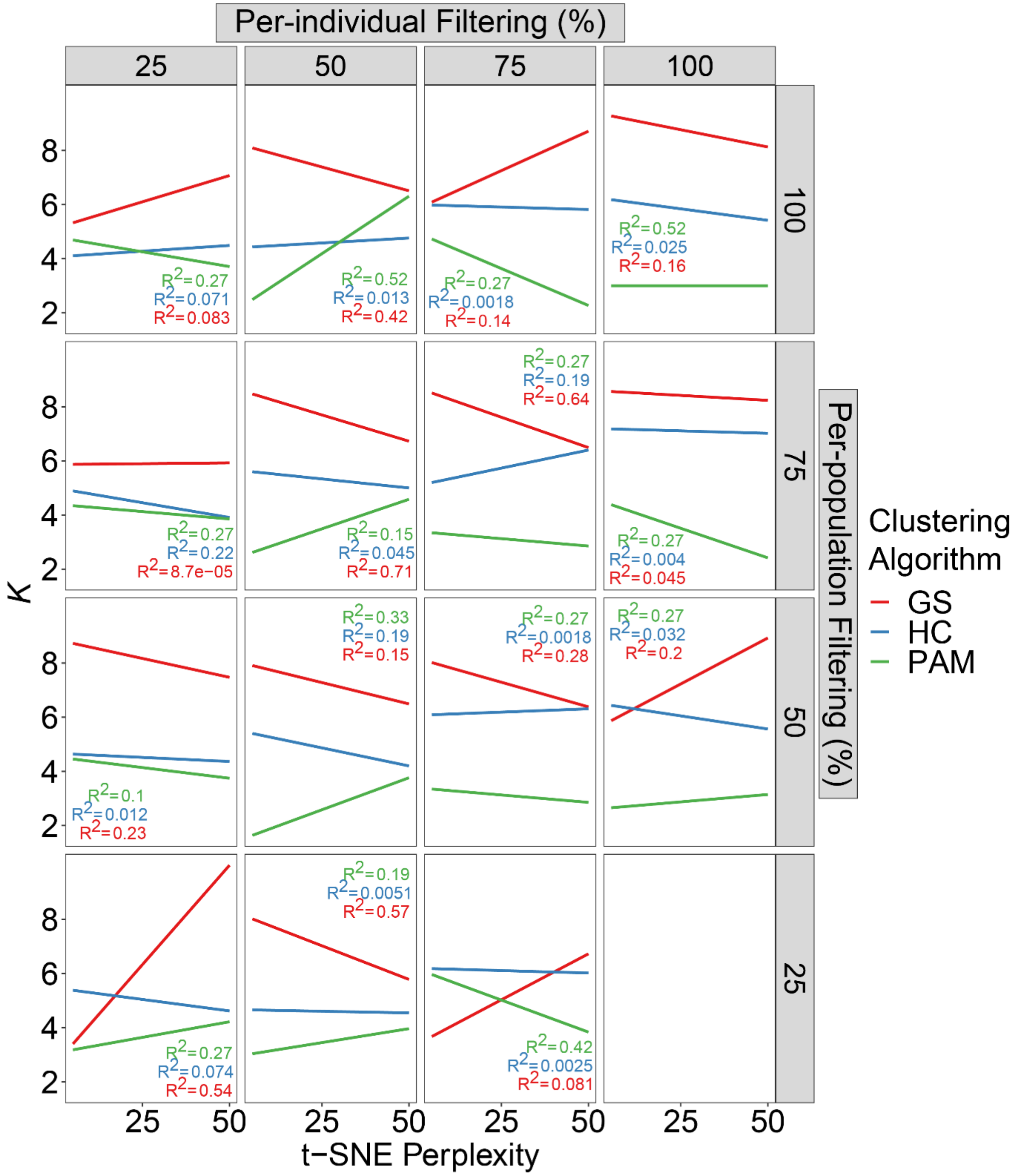


Figure S10: Regressions showing mean optimal *K* among ten t-SNE perplexity settings and three clustering algorithms for a minor allele frequency (MAF) filter=5%. Panel columns and rows represent per-individual and per-population missing data filters, respectively. Mean optimal *K* was chosen using three clustering algorithms (fill colors): 1) Partition around medoids (PAM) with the gap statistic (GS), hierarchical clustering (HC) with the highest mean silhouette width (HMSW), and PAM with HMSW. R^2 values show correlations per clustering algorithm

SUPPLEMENTARY INFORMATION APPENDIX A

1.1. Additional phylogenomic analyses

1.1.1. *Topology tests*

We performed IQ-TREE topology tests using for four *Terrapene* phylogenetic hypotheses: (a) The 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 per-partition substitution models (Kalyaanamoorthy *et al.* 2017), and nodal confidence of each tree was assessed using 1,000 ultrafast bootstrap (UFB) replicates (Hoang *et al.* 2017). We then compared support among the constraint trees using seven topological tests, each with 10,000 re-samplings: (a) Raw log-likelihoods; (b) bootstrap proportion test using the RELL approximation (bpRELL; Kishino *et al.* 1990); (c) Kishino-Hasegawa test (KH; Kishino & Hasegawa 1989); (d) Shimodaira-Hasegawa test (SH; Shimodaira & Hasegawa 1999); (e) Approximately Unbiased test (AU; Shimodaira 2002); and (f) Expected Likelihood Weights (ELW; Strimmer & Rambaut 2002). To visualize support for each topology across the genome, site-likelihood probabilities and pairwise site-likelihood score differences (Δ*SLS*) were calculated between the best-supported *versus* remaining trees.

2.1. Species delimitation analyses

2.1.1. *BFD* Prior Selection and Data Filtering*

Here, we derived appropriate priors following (Bangs *et al.* 2020). We first calculated a pairwise distance matrix using the DIVEIN web server (Deng *et al.* 2010). We did so with a random subset of the full concatenated alignment (N_{sites}=36,800, the maximum allowed by DIVEIN) derived using a custom Perl script, *nremover.pl* (https://github.com/tkchafin/scripts). Average within-species divergence was calculated from the DIVEIN pairwise distance matrix across all taxa to represent our prior for ancestral population size (θ=0.000730885) which served as the mean (α/β) for a gamma-distributed prior. The coalescent rate was set to 2/θ=2736.4086. The lineage birth-rate for the Yule process (λ=196.5038) was determined with *pyule* (https://github.com/joaks1/pyule), which invokes tree height and number of species to determine λ. Tree height was calculated as ½ the maximum among-group pairwise distance (=0.002549775), and the number of species was conservatively set to three to limit potential biases from over-splitting (a tendency for multi-species coalescent species delimitation approaches). The mutation rate priors were fixed to 1.0 per recommendations from the BFD* tutorial (Leaché & Bouckaert 2018).

Before running BFD*, we first removed loci containing >50% missing data, both globally and per-population, using a custom Perl script *phylipFilterPops.pl* (https://github.com/tkchafin/scripts). Thus, all retained sites contained SNP data in at least 50% of individuals from each population. We further filtered the alignment by removing non-binary SNPs and invariant sites via the *Phrynomics* R package (https://github.com/bbanbury/phrynomics). We then generated XML files for 20 models using BEAUTI v2.5.2 and ran BFD* via the SNAPP v1.4.2 plug-in for BEAST v2.5.2 (Bryant *et al.* 2012; Bouckaert *et al.* 2019).

2.1.2. *Machine learning data preparation*

Using R v3.5.1 (R Core R Development Core Team 3.0.1. 2013), we ran a slightly modified version of the R script developed by Derkarabetian *et al.* [(2019) (*PCA-DAPC-RF-tSNE_str.r;* https://github.com/shahanderkarabetian/uml_species_delim)] to load and prepare the input alignments, perform the random forest (RF) and t-distributed stochastic neighbor embedding (t-SNE) machine learning algorithms, and identify taxon clusters. The modified script, *PCA-DAPC-RF-tSNE_gridSearch_maf.r*, adds the capability of performing multiple independent runs among multiple datasets to assess RF and t-SNE variability and evaluate model performance among differently filtered datasets. It also performs a t-SNE grid search for the perplexity setting. Generally, the script used the R package *adegenet* v2.1.1 (Jombart & Ahmed 2011) to load the input alignments from STRUCTURE-formatted files. The data were scaled using the *scaleGen* function and subjected to dimensionality reduction via principle component analysis (PCA; *dudi.pca* function in *adegenet*). The full suite of PCA axes were assessed using DAPC (discriminant analysis of principle components; Jombart *et al.* 2010) cross-validation with 1,000 replicates (*xvalDapc* function in *adegenet*) to determine the optimal number of principle components and discriminant functions to retain. The scaled PCA data with the optimal number of axes were ultimately used as input for the RF and t-SNE analyses.

Input alignments for VAE (variational autoencoder; Kingma & Welling 2013) were generated from PHYLIP-formatted files using a custom python script, *phylip2onehotsnps.py*. VAE was then run via the Python3 script developed by Derkarabetian *et al.* (2019), with some minor modifications (*sp_deli_clust_commandline_noClust.py*; *vae_dbscan.py*). Changes included the implementation of a training/test data split to assess model performance, an early stopping callback to reduce overfitting, support for multiple independent runs to evaluate variability and cluster stability, and the DBSCAN clustering algorithm to determine the optimal number of clusters (*K*) in an unsupervised manner. Modified scripts can be found in a GitHub repository: https://github.com/btmartin721/mecr_boxturtle.

2.1.3. *Random forest*

A user-specified number of classification/ decision trees (i.e., a "forest") are created by the Random Forest (RF) algorithm (Breiman 2001), and classification trees (*N*=10,000) are then trained with random data subsets from which majority-vote class predictions are made. Nodes containing overlapping among-sample distances elevate a "proximity score" that is bootstrapped and aggregated (i.e., "bagged") over all classification trees, with higher proximity scores indicating similar individuals. The output proximity matrix was visualized using two dimensionality reduction algorithms, classic and isotonic multidimensional scaling (cMDS and isoMDS; Shepard *et al.* 1972; Kruskal & Wish 1978). CMDS utilizes the full dissimilarity matrix from the RF classifier, whereas ISOMDS forces a monotonic transformation that only uses the ranks from the proximity scores. Thus, CMDS preserves among-sample distances, whereas ISOMDS does not.

2.1.4. *T-SNE*

Similar to PCA, t-SNE is a dimensionality reduction algorithm (Maaten & Hinton 2008). Rather than using proximity scores to generate probability distributions representing similarities between samples in multidimensional space, it instead employs non-parametric, non-linear algorithms to estimate pairwise distances. It then attempts to minimize differences between high-dimensional space versus low-dimensional embedding. Samples with low similarity continue to repel each other as each iteration occurs, such that they become diffuse across parameter space. t-SNE was run for 20,000 iterations, within which the equilibria of the clusters were visually confirmed. Perplexity, which limits the effective number of neighbors, was tested at values ranging from 5-50 (incrementing by five), with the initial number of dimensions parameter set to five.

2.1.5. *Variational autoencoders*

SNPs were first converted from a PHYLIP file to a binary ‘one-hot’ format, from which two latent variables representing the sample mean (μ) and standard deviation (σ) can be inferred by VAE (as implemented by Derkarabetian *et al.* 2019). VAE reconstructs the SNP dataset using the latent variables and self-trains by minimizing the difference (i.e., model loss) between the input and reconstructed datasets. Following training, latent variables are predicted from the full dataset and represented in two-dimensional space.

VAE was run with three encoder and decoder layers, each containing 100 neurons subjected to a dropout rate of 0.5 to reduce overfitting. Encoded SNP data was normalized and scaled to reduce the impact of stochasticity, with input split into datasets representing 80% training/ 20% validation (as is standard with machine learning). Model loss was assessed using an early stopping callback function from the scikit-learn Python package (Pedregosa *et al.* 2011) to determine an optimal number of epochs (i.e. cycles through the training dataset). Ideally, this should terminate when loss (~error) has converged and is minimized among both the training and validation datasets [(i.e. the ‘Goldilocks zone’; Al’Aref *et al.* 2019) (Fig. S3)]. An escalating loss in the validation dataset indicates overfitting. On the other hand, losses that have not yet reached their minimum value suggest model underfitting (i.e. a lack of generalization for both training and unseen data). Other parameters were chosen following Derkarabetian *et. al.* (2019).

REFERENCES

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.

Bangs MR, Douglas MR, Chafin TK, and Douglas ME (2020) Gene flow and species delimitation in fishes of Western North America: Flannelmouth (*Catostomus latipinnis*) and Bluehead sucker (*C. Pantosteus discobolus*). *Ecology and Evolution*, **10**, 6477–6493.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I *et al.* (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis (M Pertea, Ed,). *PLOS Computational Biology*, **15**, e1006650.

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

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, and RoyChoudhury A (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, **29**, 1917–1932.

Deng W, Maust B, Nickle D, Learn G, Liu Y, Heath L, Kosakovsky Pond S, and Mullins J (2010) DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. *BioTechniques*, **48**, 405–408.

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.

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.

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

Jombart T, Devillard S, and Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

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.

Kingma DP and Welling M (2013) Auto-encoding variational bayes. *In: Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv:1312.6114 [stat.ML].

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 Publishing, Thousand Oaks, CA, USA.

Leaché A and Bouckaert R (2018) Species trees and species delimitation with SNAPP: a tutorial and worked example. http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2018/01/BFD-tutorial-1.pdf.

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

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

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, and Dubourg V (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

R Development Core Team 3.0.1. (2013) A language and environment for statistical computing. *R Foundation for Statistical Computing*, **2**, https://www.R-project.org.

Shepard RN, Romney AK, and Nerlove SB (1972) *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences: I. Theory.* Seminar Press, New York City, NY, USA.

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.

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.

CHAPTER III

ClinePlotR: Visualizing genomic clines and detecting outliers in R

ABSTRACT

Patterns of multi-locus differentiation (i.e., genomic clines) often extend broadly across hybrid zones and their quantification can help diagnose how species boundaries are shaped by adaptive processes, both intrinsic and extrinsic. In this sense, the transitioning of loci across admixed individuals can be contrasted as a function of the genome-wide trend, in turn allowing an expansion of clinal theory across a much wider array of biodiversity. However, computational tools that serve to interpret and consequently visualize 'genomic clines' are limited. Here, we introduce the CLINEPLOTR R-package for visualizing genomic clines and detecting outlier loci using output generated by two popular software packages, BGC and INTROGRESS. CLINEPLOTR bundles both input generation (i.e, filtering datasets and creating specialized file formats) and output processing (e.g., MCMC thinning and burn-in) with functions that directly facilitate interpretation and hypothesis testing. Tools are also provided for post-hoc analyses that interface with external packages such as ENMEVAL and RIDEOGRAM. Our package increases the reproducibility and accessibility of genomic cline methods, thus allowing an expanded user base and promoting these methods as mechanisms to address diverse evolutionary questions in both model and non-model organisms.

1. INTRODUCTION

Patterns of multi-locus differentiation, as distributed across admixture gradients, have long provided a window into divergence and speciation (e.g., Barton, 1983; Gompert, Mandeville, & Buerkle, 2017). Accordingly, they have been used to map loci associated with adaptation or reproductive isolation (Buerkle & Lexer, 2008; Martin et al., 2020), and as indicators of biotic responses to environmental change (Chafin, Douglas, Martin, & Douglas, 2019). Rather than relating these to patterns in the landscape, contemporary approaches have instead drawn conclusions based on genome-wide ancestries (Gompert & Buerkle, 2009; Fitzpatrick, 2013). The evolutionary processes that generate 'genomic clines' can be illuminated even when constituent taxa do not segregate geographically, but rather patchily (Bierne, Gagnaire, & David, 2013), or as a hybrid mosaic (Chafin et al., 2019).

Several programs are available specifically to investigate genomic clines. Of these, BGC (GOMPERT & BUERKLE, 2011, 2012) is the most robust to false positives and uses a Bayesian approach that accounts for genotype uncertainty (Gompert, Lucas, et al., 2012) and autocorrelation caused by physical linkage (Gompert, Parchman, & Buerkle, 2012) in next-generation sequencing datasets. Although a powerful tool for analyzing hybridization with molecular data, it lacks user-friendly output. Researchers must either develop custom scripts or build cumbersome, one-off pipelines, neither of which is parsimonious. A more direct approach is clearly necessary.

Here, we present a comprehensive R-package, CLINEPLOTR, that promotes the genomic cline methodology. The package includes functions that facilitate BGC input file generation and output visualization and extend the plotting functionality from another genomic cline software package, INTROGRESS (GOMPERT & BUERKLE, 2010). Locus-wise clinal patterns are visualized by accessing a suite of R-methods that interpret them as a function of the genome-wide average, genomic position along chromosomes, and in relation to spatial and environmental parameters.

2. DESCRIPTION

2.1. Overall package workflow

The CLINEPLOTR R-package incorporates an introduction to available functions and can be installed via provided instructions directly from the GitHub repository (github.com/btmartin721/ClinePlotR). CLINEPLOTR includes three primary pipelines, a summary of which can be visualized in Figure 1.

The workflow for our BGC PIPELINE includes functions to aggregate outputs from multiple independent runs, thin MCMC samples, and plot log-likelihood and BGC parameter traces. From these, CLINEPLOTR can both identify outlier loci using any of several user-defined options and plot locus-wise ancestry probabilities (ϕ) as a function of the hybrid index (Figure 2). Finally, users can examine the locus-wise relationship between cline center (α) and rate (β), with polygon hulls included to encapsulate 2D 'outlier space' for each parameter (Gauthier et al., 2020).

CLINEPLOTR additionally includes accessory functions that allow an examination of variation in clinal parameters across the genome. Although mapping loci to reference assemblies is outside the scope of this package, an example of a workflow using MINIMAP2 (Li, 2018) is in the documentation. If the user has access to physical SNP (single nucleotide polymorphism) coordinates and a closely-related chromosome-level assembly, CLINEPLOTR can integrate these data with the RIDEOGRAM package (Hao et al., 2020) to yield karyotype-style ideograms annotated with heatmaps for both BGC cline parameters (Figure 3).

Functions are also provided to facilitate an INTROGRESS workflow by generating input data frames as well as accessories that embellish the plotting functions already present in INTROGRESS. These accessory functions will visualize spatial patterns (e.g., latitude/ longitude) and environmental variables that are inherent to genomic clines (Figure 4), to include helper functions that invoke ecological niche models (MAXENT: Phillips, Anderson, & Schapire, 2006) as generated in the R-package ENMEVAL (Muscarella et al., 2014).

2.2. Input and file format

The primary purpose of CLINEPLOTR is to simplify the use of rather cumbersome software designed to estimate genomic clines. To facilitate this task, accessory scripts that prepare files for input into BGC and INTROGRESS are available in the GitHub repository, with a few variants. For example, *phylip2bgc.pl* script converts a PHYLIP-formatted alignment containing concatenated SNPs to the CUSTOM BGC input format. It can also subset populations and/ or individuals from a larger alignment. A similar script, *phylip2introgress.pl*, does likewise with INTROGRESS INPUT. Because BGC can additionally consider linkage among loci as well as genotype uncertainty, an input script (*vcf2bgc.py*) that employs the PYVCF Python library (https://pyvcf.readthedocs.io/) is also provided as a means to format an IPYRAD (Eaton & Overcast, 2020) VCF file containing annotations for physical position and genotype read counts. Finally, an additional script, *nremover.pl*, is provided to comprehensively filter a PHYLIP-formatted SNP file. The program includes the capacity to filter by matrix occupancy per individual and per SNP column, and by minor allele frequency. It will also remove non-biallelic or monomorphic SNPs, and can randomly subsample large datasets.

2.3. Outlier detection for Bayesian genomic clines

BGC output (extracted from HDF5 format using BGC's *estpost* function) must be named as *prefix*_bgc_stat_*param_replicate*, where *prefix* is shared across all independent BGC replicates, *param* is an individual output parameter (e.g., LnL), and *replicate* is an integer. Outputs from any number of replicates can then be parsed, thinned, and combined via the *combine_bgc_output* function in ClinePlotR. The *combine_bgc_output* function provides arguments for the number of MCMC samples to be removed as burn-in, and for a sampling frequency with which to thin samples. Following BGC run aggregation, the MCMC samples can be visually inspected for mixing and convergence using a trace plotting function, *plot_traces*. Adjustments can then be made to thinning or burn-in parameters by re-running the *combine_bgc_output* function or, if necessary, by re-running BGC with altered parameters or increased MCMC length.

A primary goal of genomic cline analysis is to identify loci that possess either excess ancestry or exceptionally steep transitions relative to the genome-wide average. Here, we provide the function *get_bgc_outliers* that offers two outlier detection methods [described in Gompert & Buerkle (2011, 2012)]. Briefly, the first simply queries if the credibility intervals for the posterior probability distribution of cline parameters α or β (i.e., cline center and rate, respectively) exclude the neutral expectation (i.e., α or $\beta = 0$). If this interval excludes zero for either parameter, a locus can be flagged as either an α-outlier, β-outlier, or both.

The second method considers if per-locus parameter estimates are statistically unlikely, given the distribution of values across all loci. This is accomplished by classifying outliers as those for which posterior median α and ß estimates are not encapsulated by the $(\frac{n}{2})$ and $(\frac{1-n}{2})$ quantiles from a conditional α and ß prior distribution (Gaussian with a mean of zero), where *n* represents a user-specified threshold (e.g., 95%, 97.5%). Users can choose whether to classify outliers using any combination of the above methods, but all require the zeta and gamma quantile estimates from the BGC output.

We additionally track whether parameter values are significantly positive or negative. This indicates either an increase ($\alpha > 0$) or decrease ($\alpha < 0$) in the probability of parental population ancestry among hybrids for a given locus, or deviation in the rate of transition in probabilities of locus-specific ancestries towards either very steep ($\beta > 0$) or wide ($\beta < 0$) shapes (Gompert & Buerkle, 2011).

2.4. Visualization options

We attempted to tailor available visualizations in CLINEPLOTR towards common applications of Bayesian genomic clines found in the literature, and we will continue to add additional ones as need arises. Many applications seek to identify loci subject to various selective processes (Parchman et al., 2013) by comparing how ancestries transition among loci with respect to the genome wide average. To facilitate this, the *phiPlot* function computes ϕ_{ijn}, the probability of parental population1 ancestry for each locus (*i*) and individual (*n*) within each admixed population (*j*) [Eqn. 3 and 4; Gompert & Buerkle (2011)]. It then produces a plot of ϕ (per locus) on the y-axis against posterior estimates of hybrid index on the x-axis (*sensu* Gompert et al., 2012), with a user-specified color scheme that designates statistical outliers (Figure 2).

Other applications have specifically examined relationships among cline rate and center parameters (Gauthier et al., 2020), and we also do so by implementing the *alphaBetaPlot* function. A 2-D density contour plot of α and ß parameters is produced, with values for individual loci optionally mapped, and with the potential to calculate and plot polygon hulls that encapsulate positive and negative outliers with respect to each parameter (Figure 2).

2.5. Extended functions and helper scripts

We also provide several additional functions (see Figure 1) that have considerable use cases, although some seemingly deviate from the 'core' BGC workflow. The first of several can be used to map parameter values of BGC clines onto a chromosomal ideogram via the function *plot_outlier_ideogram* (e.g., Figure 3), Here, BGC results are depicted for a case study examining hybridization between Woodland (*Terrapene carolina carolina*) and Three-toed box turtles (*Terrapene mexicana triunguis*) (Martin et al., 2020). However, some external user-steps are required to use the function.

Briefly, we mapped the *Terrapene* ddRAD sequencing alignment against the available *Terrapene mexicana triunguis* scaffold-level assembly (GenBank Accession: GCA_002925995.2). Scaffold coordinates were then converted to chromosome coordinates by mapping *Terrapene* scaffolds against the closely related chromosome-level *Trachemys scripta* assembly [(Simison, Parham, Papenfuss, Lam, & Henderson, 2020); GenBank accession: GCA_013100865.1]. This was accomplished by employing MINIMAP2 (Li, 2018) and PAFSCAFF (github.com/slimsuite/pafscaff). The output from *get_bgc_outliers* and PAFSCAFF, plus a GFF file read/ parsed via the provided functions *parseGFF* and *join_bgc_gff*, were used to plot a heatmap of BGC α- and ß-values on an ideogram. Essentially, the ideogram plot (generated using the RIDEOGRAM R-package) allows the chromosomal locations of each outlier to be visualized (Figure 3). It also provides a distinction between transcriptomic SNPs falling within known genes *versus* loci from surrounding scaffolds. For additional details, a more in-depth tutorial is provided at github.com/btmartin721/ClinePlotR.

Other extended functions include a wrapper to simplify running INTROGRESS (*runIntrogress*), and a function that allows genomic clines (Figure 4A) and hybrid indices (Figure 4B) from INTROGRESS to be correlated with spatial and environmental variables. To access this functionality, one can run *clinesXenvironment* using the object returned from *runIntrogress* and raster values extracted from each sample locality. Multiple rasters can be included (e.g., the 19 BioClim layers; https://worldclim.org/), and users can run the included ENMEVAL wrapper functions (*runENMeval* and *summarize_ENMeval*) to identify uninformative layers that may subsequently be excluded from *clinesXenvironment*. These latter functions access MAXENT using the ENMEVAL pipeline (Muscarella et al., 2014), whereby the most informative raster layers are designated with the 'permutation importance' statistic.

3. CONCLUSIONS

Genomic clines are useful for assessing patterns of introgression in hybrid zones. Unfortunately, parsing and plotting results from the available genomic cline software can be difficult. Given that genomic clines have a variety of applications, to include conservation genetics, evolutionary biology, and speciation research, it is clearly important that they be accessible for use by researchers. Here, we present an R-package that greatly simplifies the parsing of output from available genomic cline software, as well as the production of publication-quality figures. Our R-functions are intended to be user-friendly, and to this end employ a variety of parameters that can be altered by users to suit specific research needs. Furthermore, CLINEPLOTR allows outlier SNPs to be visualized across the genome, while also distinguishing known genes and surrounding loci. In addition, the environmental and spatial effects on genomic clines can be assayed. This extended functionality enhances the interpretation of genomic clines and provides greater insight into those underlying processes that potentially contribute to the observed patterns. Hopefully, future iterations of genomic cline software can act to extend chromosomal and environmental associations, particularly as whole genome sequencing becomes less expensive and more common.

FUNDING

No external funding was provided to the authors in support of this work.

ACKNOWLEDGEMENTS

We thank S.M. Mussmann and M.R. Bangs for contribution to the *nremover.pl* script. We also thank staff of the Arkansas High Performance Compute Center (AHPCC) for access to computational resources that were used to test this package on the University of Arkansas Trestles cluster (funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission).

DATA ACCESSIBILITY

ClinePlotR is available as a GitHub repository: https://github.com/btmartin721/ClinePlotR. We also plan to submit to CRAN prior to publication. The data used herein will be available as an example dataset in a Dryad Digital Repository [DOI]. During review, the data will also be temporarily accessible from Box Drive (https://uark.box.com/s/ei21v3unvxfxczbfnfflwhrk536bsd5h)

4. REFERENCES

Barton, N. H. (1983). Multilocus Clines. *Evolution, 37*(3), 454–471. doi:10.2307/2408260

Bierne, N., Gagnaire, P. A., & David, P. (2013). The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Current Zoology*, *59*(1), 72–86. doi:https://doi.org/10.1093/czoolo/59.1.72

Buerkle, C. A., & Lexer, C. (2008). Admixture as the basis for genetic mapping. *Trends in Ecology and Evolution*, *23*(12), 686–694. doi:10.1016/j.tree.2008.07.008

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*, *123*, 759–773. doi:10.1038/s41437-019-0259-2

Eaton, D. A. R., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, *36*(8), 2592–2594. doi:10.1093/bioinformatics/btz966

Fitzpatrick, B. M. (2013). Alternative forms for genomic clines. *Ecology and Evolution*, *3*(7), 1951–1966. doi:https://doi.org/10.1002/ece3.609

Gauthier, J., de Silva, D. L., Gompert, Z., Whibley, A., Houssin, C., Le Poul, Y., … Elias, M. (2020). Contrasting genomic and phenotypic outcomes of hybridization between pairs of mimetic butterfly taxa across a suture zone. *Molecular Ecology*, *29*(7), 1328–1343. doi:https://doi.org/10.1111/mec.15403

Gompert, Z., & Buerkle, C. A. (2009). A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, *18*(6), 1207–1224. doi:10.1111/j.1365-294X.2009.04098.x

Gompert, Z., & Buerkle, C. A. (2010). INTROGRESS: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, *10*(2), 378–384. doi:10.1111/j.1755-0998.2009.02733.x

Gompert, Z., & Buerkle, C. A. (2011). Bayesian estimation of genomic clines. *Molecular Ecology*, *20*(10), 2111–2127. doi:10.1111/j.1365-294X.2011.05074.x

Gompert, Z., & Buerkle, C. A. (2012). BGC: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, *12*(6), 1168–1176. doi:10.1111/1755-0998.12009.x

Gompert, Z., Lucas, L. K., Nice, C. C., Fordyce, J. A., Forister, M. L., & Buerkle, C. A. (2012). Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, *66*(7), 2167–2181. doi:10.1111/j.1558-5646.2012.01587.x

Gompert, Z., Mandeville, E. G., & Buerkle, C. A. (2017). Analysis of population genomic data from hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, *48*, 207–229. doi:10.1146/annurev-ecolsys-110316-022652

Gompert, Z., Parchman, T. L., & Buerkle, C. A. (2012). Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1587), 439–450. doi:10.1098/rstb.2011.0196

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., & Chen, J. (2020). RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the ideograms. *PeerJ Computer Science*, *6*, e251. doi:10.7717/peerj-cs.251

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094–3100. doi:10.1093/bioinformatics/bty191

Martin, B. T., Douglas, M. R., Chafin, T. K., Placyk, J. S., Birkhead, R. D., Phillips, C. A., & Douglas, M. E. (2020). Contrasting signatures of introgression in North American box turtle (*Terrapene* spp.) contact zones. *Molecular Ecology*. doi:10.1111/mec.15622

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models . *Methods in Ecology and Evolution*, *5*(11), 1198–1205. doi:10.1111/2041-210x.12261

Parchman, T. L., Gompert, Z., Braun, M. J., Brumfield, R. T., McDonald, D. B., Uy, J. a C., … Buerkle, C. a. (2013). The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, *22*(12), 3304–3317. doi:10.1111/mec.12201

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modeling*, *190*(3), 231–259. doi:10.1016/j.ecolmodel.2005.03.026

Simison, W., Parham, J., Papenfuss, T., Lam, A., & Henderson, J. (2020). Annotated chromosome-level reference genome of the red-eared slider turtle (*Trachemys scripta elegans*). *Genome Biology and Evolution*, *12*(4), 456–462. doi:10.1093/gbe/evaa063

TABLES AND FIGURES

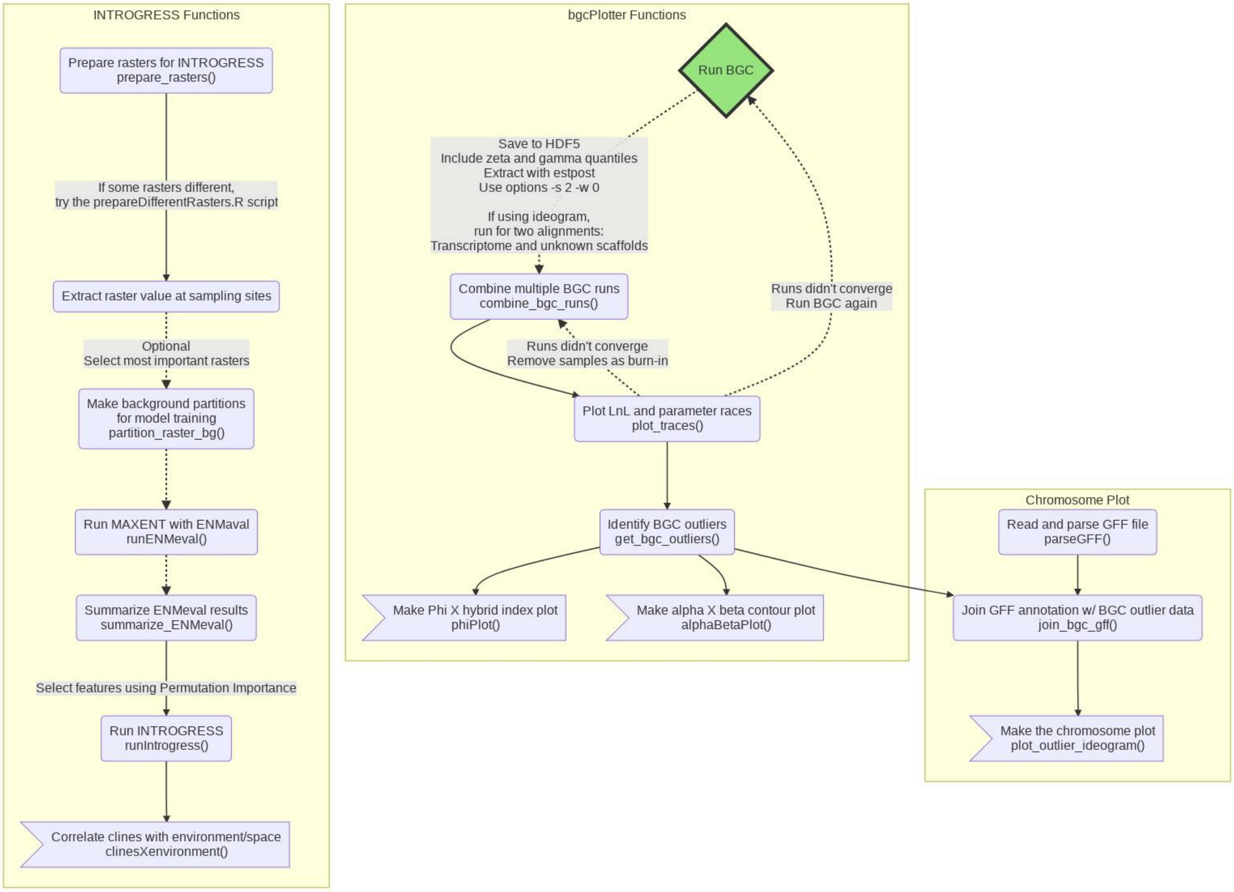
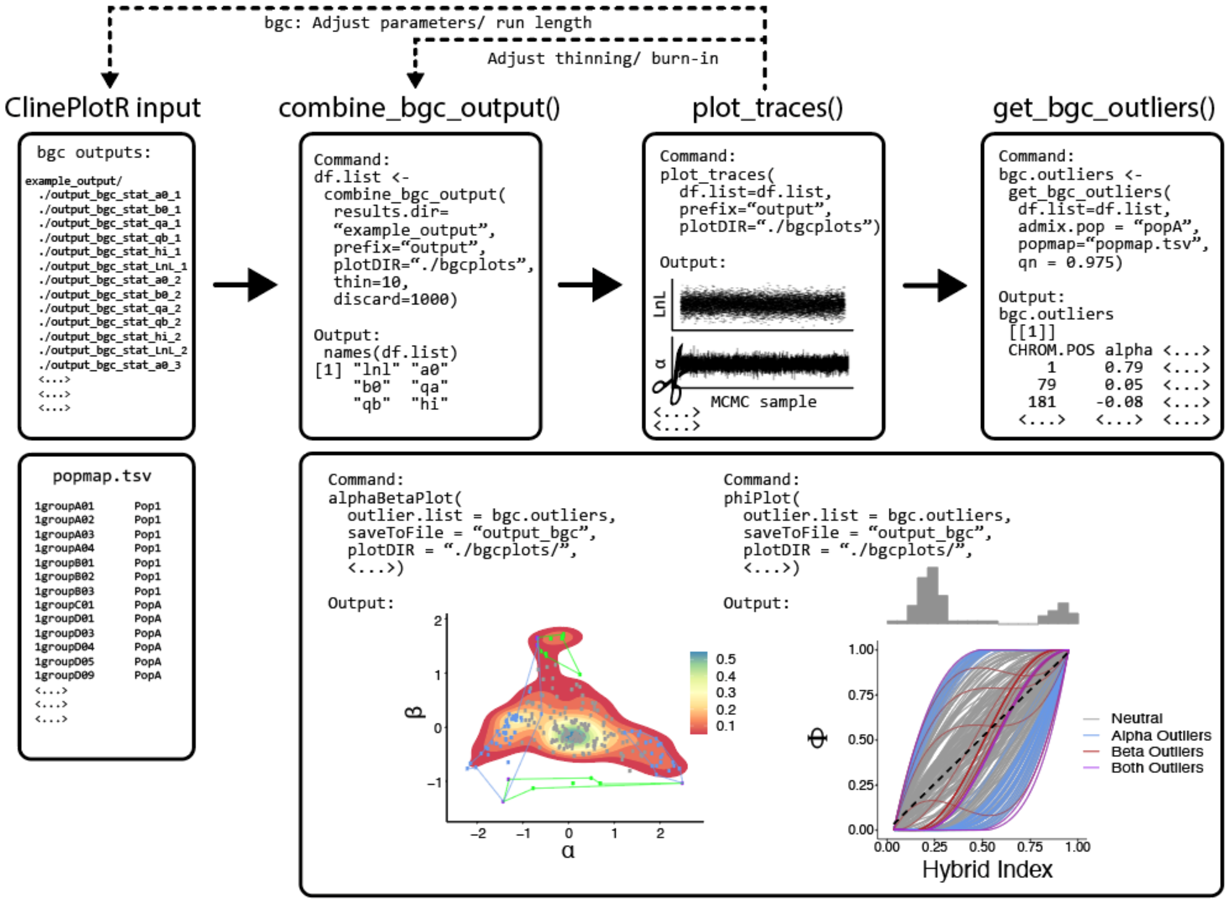


Figure 1: Simplified example workflow listing all available ClinePlotR functions. Yellow boxes group inter-dependent functions working towards producing one or two particular plots (terminal plotting steps depicted as flags). Connecting arrows indicate a pipeline where each step is dependent on the returned R objects. The green 'Run BGC' diamond identifies BGC as an external *a priori* step for the *bgcPlotter* and *chromosome plot* functions. The dotted lines indicate optional steps.



```
bgc outputs:

example_output/
  ./output_bgc_stat_a0_1
  ./output_bgc_stat_b0_1
  ./output_bgc_stat_qa_1
  ./output_bgc_stat_qb_1
  ./output_bgc_stat_hi_1
  ./output_bgc_stat_LnL_1
  ./output_bgc_stat_a0_2
  ./output_bgc_stat_b0_2
  ./output_bgc_stat_qa_2
  ./output_bgc_stat_qb_2
  ./output_bgc_stat_hi_2
  ./output_bgc_stat_LnL_2
  ./output_bgc_stat_a0_3
  <...>
  <...>
  <...>
```

```
popmap.tsv

1groupA01      Pop1
1groupA02      Pop1
1groupA03      Pop1
1groupA04      Pop1
1groupB01      Pop1
1groupB02      Pop1
1groupB03      Pop1
1groupC01      PopA
1groupD01      PopA
1groupD03      PopA
1groupD04      PopA
1groupD05      PopA
1groupD09      PopA
<...>
<...>
<...>
```

```
Command:
df.list <-
 combine_bgc_output(
  results.dir=
  “example_output”,
  prefix=“output”,
  plotDIR=“./bgcplots”,
  thin=10,
  discard=1000)

Output:
 names(df.list)
[1] "lnl" "a0"
    "b0"  "qa"
    "qb"  "hi"
```

```
Command:
plot_traces(
  df.list=df.list,
  prefix=“output”,
  plotDIR=“./bgcplots”)

Output:
<...>
<...>
```

```
Command:
bgc.outliers <-
 get_bgc_outliers(
  df.list=df.list,
  admix.pop = “popA”,
  popmap=“popmap.tsv”,
  qn = 0.975)

Output:
bgc.outliers
 [[1]]
 CHROM.POS alpha <...>
     1      0.79 <...>
    79      0.05 <...>
   181     -0.08 <...>
  <...>    <...> <...>
```

```
Command:
alphaBetaPlot(
  outlier.list = bgc.outliers,
  saveToFile = “output_bgc”,
  plotDIR = “./bgcplots/”,
  <...>)

Output:
```

```
Command:
phiPlot(
  outlier.list = bgc.outliers,
  saveToFile = “output_bgc”,
  plotDIR = “./bgcplots/”,
  <...>)

Output:
```

Figure 2. Example workflow for parsing Bayesian genomic cline (BGC) output, visualizing MCMC traces, detecting outliers, and plotting results. The ‘phiPlot’ (right-side, lower right box) shows hybrid indices (x-axis) and probability of parental population1 alleles (y-axis), plus a histogram of hybrid indices in the admixed population. The ‘alphaBetaPlot’ (left-side, lower right box) shows 2-D density of cline width/ rate representing the cline center (i.e., bias in SNP ancestry; α; x-axis) and steepness of clines (ß; y-axis). Outliers are additionally encapsulated using polygon hulls.

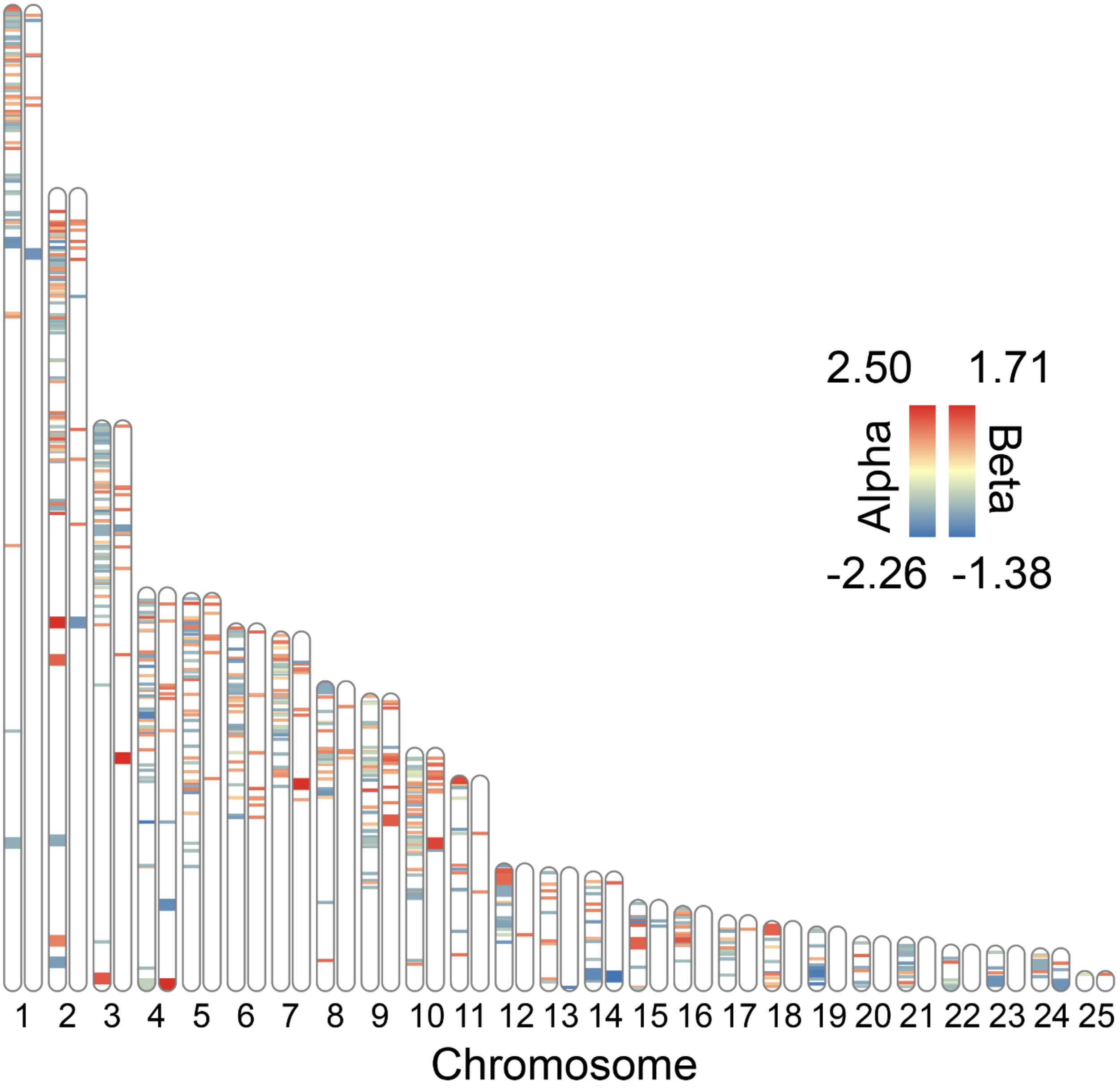


Figure 3. Example ideogram plot using Bayesian genomic cline (BGC) outliers for *Terrapene* ddRAD SNPs (y-axis), plotted onto *Trachemys scripta* chromosomes (x-axis). Chromosomes are duplicated, with alternative heatmaps for cline center (α; left) and rate (ß; right). Larger heatmap bands correspond to SNPs located within known genes, whereas smaller bands were found in unknown scaffolds.

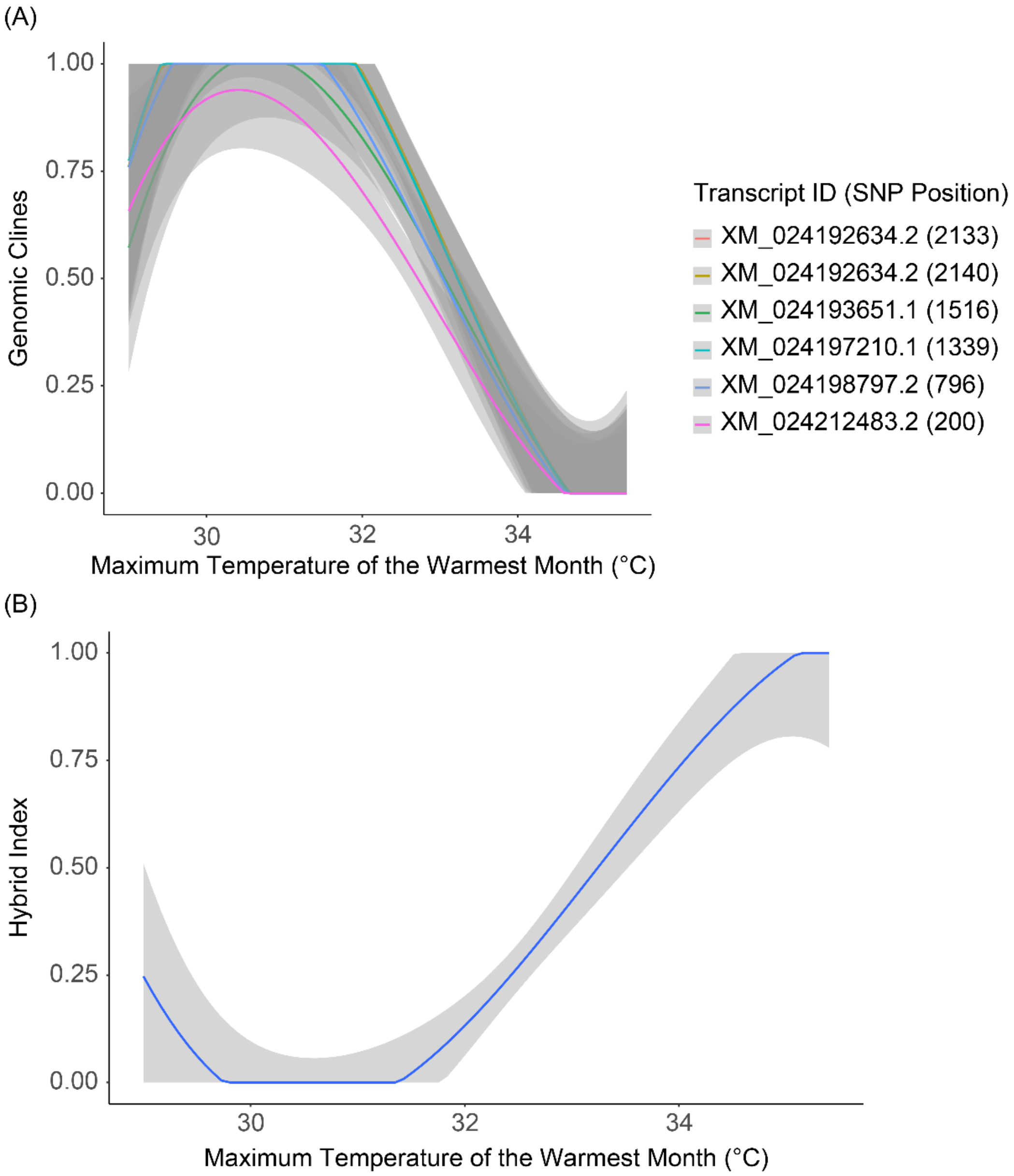


Figure 4: Example plots that can be made using the Introgress pipeline in ClinePlotR. The includes climatic variable on the X-axis corresponds to BioClim raster layer 5 (https://worldclim.org). The gray shading indicates confidence intervals for each regression line. (A) Genomic clines for six outlier SNPs mapped to the *Terrapene mexicana triunguis* transcriptome. Transcript IDs correspond to GenBank accession numbers and the position of each SNP (in base pairs) on the locus. (B) Hybrid index output from Introgress *versus* an environmental variable.

CHAPTER IV

The impacts of relictual hybridity and primary divergence on phylogeographic patterns in Massasauga rattlesnakes (*Sistrurus* spp.)

ABSTRACT

Broad geomorphic processes interspersed with climatic fluctuations have not only yielded major North American biomes but also driven the phylogenies and phylogeographies of biodiversity so contained. Cycles of demographic expansion/contraction reticulate evolution of historical biogeography in driving patterns in the Massasauga rattlesnakes (*Sistrurus spp.*). We utilized ddRAD sequencing to sample thousands of genome-wide SNPs necessary to assess demographic and phylogeographic patterns at a fine temporal and spatial grain. In so doing, we observed paraphyly in the Prairie Massasauga, *S. t. tergeminus*, with respect to the Desert Massasauga, *S. t. edwardii*, and fine-scale intraspecific structure within *S. t. tergeminus* and *S. t. edwardsii* that was previously unsupported with single-gene and microsatellite markers. We also detected directional re-expansion following a period of secondary contact in the Midwest, which we posit is associated with historical fluctuation of the Mississippi River during altithermal Pleistocene periods that act in synergy with glacial-interglacial cycles to define alternating periods of refugial contact and vicariance. We found hybridization as resulting in the origin of a relictual lineage in Missouri lying between the respective eastern and western edges of *S. t. tergeminus* and Eastern Massasauga (*S. catenatus*) ranges. We also detected contemporary intergradation between *S. t. tergeminus* and *S. t. edwardsii* in a contact zone in western Texas and eastern New Mexico, but with demographic model selection indicating primary divergence. Our results

contribute to a broader understanding of the role of historical biogeography in driving hybridization and range dynamism and clarify units for management within *Sistrurus*.

1. INTRODUCTION

Species distributions in North America exhibit widespread dynamism in response to historic climate fluctuations that rendered biogeographic barriers into semi-permeable zones of transition (Antonelli 2017). Major periods of Quaternary range expansion and contraction (Hewitt 1996, 1999, 2001) were punctuated by variable temperature regimes largely modulated by glacial cycles that varied over time and by region (Clark *et al.* 1995; Adams *et al.* 1999). Some species were largely displaced by glaciation, followed by retreat into and/or emergence from isolated refugia such that complex demographic patterns developed. This process often occurred iteratively as the climate waxed and waned, with ample opportunities for secondary contact as inhospitable habitat re-opened via brief interglacials (Douglas *et al.* 2009). Cataloguing biotic responses to dramatic climatic change, and contrasting the resulting histories of biodiversity, is now much more accessible and interpretable given our capacity to assay the adaptations so produced and recorded in genomes of constituent species. This, in turn promotes a predictive capacity going forward, particularly with regards to the ongoing (and rapid) changes being elicited by the Anthropocene.

The Eastern United States represents a good example of historic climate change and its impacts. Recurring glaciations drove biodiversity southward, forcing previously isolated taxa and overlapping refugia to intersect (Swenson & Howard 2005). Vegetation south of the glacial maximum consisted primarily of mesic broadleaf forests (Pound *et al.* 2012) that has remained relatively stable since the Miocene (~16-5 million years ago, Mya). Many non-glaciated geological features serve as phylogeographic breakpoints, to include Ozark and Appalachian mountains, the Apalachicola and Tombigbee rivers, and intermittent oceanic incursions that served to bifurcate the Florida peninsula (Avise *et al.* 1987; Walker & Avise 1998; Soltis *et al.* 2006).

In contrast, the western U.S. was far less affected by the physical presence of glaciers, but rather by climatic oscillations driving drastic changes in precipitation and temperature (Axelrod 1948, 1979; Owen *et al.* 2003; Licciardi *et al.* 2004). The major river basins therein served to segregate biogeography over longer timescales, but with periodic alterations in flow and stream captures driving distributional fluctuations (Cole & Armentrout 1979; Trimble 1980; Wood *et al.* 2013; Graham *et al.* 2015; O'Connell *et al.* 2017). As a result, speciation in southwestern deserts and nearby shortgrass prairies was often driven by environmental heterogeneity and 'soft allopatry' (i.e., unstable isolation over time provoked by physiographic transition zones) (Douglas *et al.* 2006; Myers *et al.* 2019). Regional vegetation was driven by Plio-Pleistocene fluctuations, with the reemergence of mesic habitat from montane areas and highlands during altithermal periods, whereas arid refugia persisted in lowland patches (Axelrod 1979; Van Devender *et al.* 1987). These patterns were imprinted upon endemic biota as a result of fluctuations in genetic structure as populations 'tracked' shifting habitats through time (Douglas *et al.* 2006).

Finally, the Great Plains is characterized by sparse endemism, which reflects a relatively recent formation near the Miocene-Pliocene boundary, as large-scale shifts occurred from forested habitat to tallgrass prairie (Axelrod 1985). The Mississippi River basin was a major phylogeographic barrier dividing the Great Plains from the eastern deciduous forest (Braun 1950; Soltis *et al.* 2006), with periodic reductions in discharge allowing northeastward dispersal via a 'prairie corridor' (Cook 1993). Deciduous habitat encroached into some areas during the Plio-Pleistocene, with the re-emergence of prairie habitat during drier periods (Axelrod 1985).

However, much of the region lacks elevation such that post-glacial vegetation remained relatively continuous from west of the Mississippi to northwest Texas, where the habitat transitions into xeric shrubland (Greene & Oliver 1965; Wells 1970).

Regional biogeography is often reflected in distributions and divergences of species therein (Avise 2000; Soltis *et al.* 2006). Likewise, contact zones are often elicited by common barriers and climatic processes, and as such, serve as reservoirs for genetic transitions, intergradations, and hybridizations (Anderson 1949; Remington 1968). The hybrid zones form via a range of scenarios, to include secondary contact via glacially-induced range contractions and expansions (Hewitt 1996, 1999), and porous species boundaries as a product of interdigitated habitats (Rhymer & Simberloff 1996). As such, they are rather ubiquitous and are manifested in all three continental regions (Remington 1968; Swenson & Howard 2005), although with distinctly different origins and variable genomic consequences (Martin *et al.* 2020). Herein we examine the historic biogeography its genomic corollaries in a group of viperid snakes that span all three regions.

1.1. Study species - *Sistrurus*

The Massasauga and Pygmy rattlesnakes (Viperidae: Crotalinae: *Sistrurus*; Garman 1883) form a sister clade respective to the remaining North American crotalids, *Crotalus* and *Agkistrodon* (Murphy *et al.* 2002). *Sistrurus* inhabit much of the eastern United States, the Great Plains from Iowa to southernmost Texas, and the shortgrass prairie in central Texas, New Mexico, and southeastern Arizona (Fig. 1). It contains three currently recognized species: Eastern (*S. catenatus*) and Western *(S. tergeminus*) Massasaugas, and the Pygmy rattlesnake (*S. miliarius*).

Sistrurus catenatus is monotypic, whereas *S. miliarius* and *S. tergeminus* include subspecies: The Carolina (*S. m. miliarius*), Dusky (*S. m. barbouri*), and Western Pygmy (*S. m. streckeri*) rattlesnakes, and the Prairie (*S. t. tergeminus*) and Desert *(S. t. edwardsii*) massasaugas. The latter two are of interest as they contain several highly disjunct populations in Missouri, Colorado, New Mexico, and southeastern Arizona (Greene 1997). Herein our focus is on *S. catenatus* and *S. tergeminus* sensu lato.

Sistrurus catenatus typically inhabit mesic areas that include upland pastures and agricultural areas, open grasslands, coniferous and deciduous forests, lowland swampy areas, and edge habitat (Wright 1941; Reinert & Kodrich 1982; Weatherhead & Prior 1992; Johnson 2000; Harvey & Weatherhead 2006). The spatial ecology of *S. t. tergeminus* is far less studied but is characterized by tallgrass prairie, adjacent woodlands, and their edge habitats (Patten *et al.* 2016). In contrast, *S. t. edwardsii* occupies xeric shortgrass prairie and sandhills (Hobert *et al.* 2004; Wastell & Mackessy 2011). All three exhibit marked seasonal habitat preferences, with *S. catenatus* (Harvey & Weatherhead 2006) and *S. t. tergeminus* (Patten *et al.* 2016) preferring open summer habitats and forested winter hibernacula. Similarly, *S. t. edwardsii* overwinters in lowland shortgrass prairie with summers in upland sandhills (Wastell & Mackessy 2011, 2016). As such, *S. t. edwardsii* displays starkly different habitat preferences than do *S. catenatus* and *S. t. tergeminus*.

Sistrurus catenatus is recognized as threatened in multiple midwestern and northeastern U.S. states and Canadian provinces (Szymanski *et al.* 2016), as well as via the U.S. Endangered Species Act (ESA). While not federally-listed, *S. tergeminus* is considered Vulnerable, whereas *S. t. tergeminus* is Critically Imperiled in Nebraska and Iowa, and Vulnerable in Kansas and

Texas (NatureServe 2021). *Sistrurus t. edwardsii* is a Vulnerable Subspecies that is Critically Imperiled in Arizona, Imperiled in Colorado, and Vulnerable in Texas and New Mexico.

Anthropogenically-induced habitat loss and fragmentation, road-based mortality, and prescribed fires have been the impetus for overall conservation status (Szymanski 1998; Szymanski *et al.* 2016), although fragmented habitat prior to anthropogenic impacts is also acknowledged (Chiucchi & Gibbs 2010; Sovic *et al.* 2019; Ochoa *et al.* 2020). Nevertheless, small effective population sizes and population bottlenecks in *S. catenatus* are consistent with anthropogenic timescales (Sovic *et al.* 2019). *Sistrurus t. edwardsii* has also experienced desertification and a loss of shortgrass prairie habitat, with several populations in New Mexico, Arizona, and Colorado now disjunct (Greene 1997; Samson *et al.* 2004). Despite this, and in contrast to *S. catenatus*, genetic diversity has been seemingly retained (Anderson *et al.* 2009). In addition the southeastern Colorado population may also be relatively large, despite being highly isolated (Mackessy 2005; Wastell & Mackessy 2011). In comparison, Great Plains *S. t. tergeminus* also displays stable population sizes, but with less habitat fragmentation than the other taxa (McCluskey & Bender 2015).

Recent phylogenetic assessments (e.g., Kubatko *et al.* 2011) indicate *S. catenatus* as being monophyletic and *S. t. tergeminus* as paraphyletic with respect to *S. t. edwardsii*. Other studies also documented population structure within both *S. catenatus* (Chiucchi & Gibbs 2010; Sovic *et al.* 2016) and *S. t. edwardsii* (Anderson *et al.* 2009). However, while the only study to date that included both *S. t. edwardsii* and *S. t. tergeminus* did observe regional population structure, it spanned subspecific boundaries (Ryberg *et al.* 2015). Therein, *S. t. tergeminus*-only populations occurred in Missouri and Oklahoma, *S. t. edwardsii*-only were in New Mexico west of the Pecos River/ Arizona and the southern tip of Texas, and subspecies-transcending structure was found in Oklahoma/ western Texas/ eastern New Mexico and Colorado/ Kansas. Notably, the Missouri and western Texas/ eastern New Mexico populations may also overlie contact zones (Minton 1983), although these may involve different chronologies.

Hybridization in Missouri *Sistrurus* is controversial, in that individuals in transitional habitat were morphologically identified as intermediate between *S. catenatus* and *S. t. tergeminus* (Evans & Gloyd 1948). However, two recent genetic studies identified these individuals as pure *S. t. tergeminus* (Gerard *et al.* 2011; Gibbs *et al.* 2011). However, in neighboring Iowa where individuals are also disjunct from the larger *S. tergeminus* and *S. catenatus* populations (as in Missouri), introgression between the two subspecies has been identified (Sovic *et al.* 2016). Given the contemporary isolation of this Iowa population, it is assumed the introgression is historical in nature.

In addition, there is anecdotal evidence for potential admixture between the two *S. tergeminus* subspecies, given the presence of intermediate habitat in the contact zone separating western Texas/ eastern New Mexico. Few studies have assessed genetic population structure in *S. tergeminus* ssp., and those that have employed only single-gene markers and/ or microsatellites. The utilization of next generation sequencing (NGS) may elucidate fine-scale population structure and admixture in these clades and allow a much closer examination of their phylogeography. Herein, we do so by performing double digest RAD sequencing (ddRADseq) to ascertain a large-scale single nucleotide polymorphism (SNP) dataset with strong sampling across a broad geographic range.

2. METHODS

2.1. Study system, tissue collection, and DNA extraction

We acquired a wide geographic sampling of *Sistrurus* (Fig. 1), to include every recognized species and subspecies across the majority of states within which they occur. Additional sampling was focused on three disjunct Missouri populations, as well as a contact zone located in the southwestern United States. Blood and skin sheds were either collected or provided by colleagues, volunteers, museums, and agencies (Table S1), with genomic DNA extracted upon receipt using PureGene® or DNeasy Blood and Tissue Kits (Qiagen) and subsequent storage at -20ºC. DNA quantification was performed with broad-range DNA fluorometry (Qubit; Thermo Fisher Scientific), with each sample subsequently tested for high molecular weight DNA via 2% agarose gel electrophoresis.

2.2. Library preparation and bioinformatics

Genomic DNA (~500-1,000ng) was digested at 37ºC using *PstI* (5'-CTGCA|G-3') and *MspI* (5'-C|CGG-3'). Another 2% agarose gel electrophoresis was run to confirm fragmentation of the genomic DNA, followed by a 1.5X AMPure XP (Beckman Coulter) purification. DNA was then standardized to 100ng, ligated with unique barcoded adapters and pooled into 48-individual libraries. Following ligation, a size-selection of 323-423bp was performed using a Pippin Prep (Sage Science), followed by a 12-cycle polymerase chain reaction (PCR) using high-fidelity Phusion DNA Polymerase (New England BioLabs). DNA amplification was confirmed using Qubit fluorometry, and quality control steps (qPCR and fragment visualization) were performed at the Genomics and Cell Characterization Core Facility (University of Oregon/Eugene). Three libraries (N=144 individuals each) were pooled per lane of 1x100 single-end ddRADseq on an Illumina Hi-Seq 4000.

Sequence quality was first assessed (FASTQC v0.11.5) and raw reads subsequently demultiplexed, clustered, and aligned (IPYRAD v0.7.30; Eaton & Overcast 2020). We removed barcodes and adapters and assembled our alignments *de novo* with a minimum coverage depth requirement of ≥20 reads/ fragment, and a 50% missing data threshold. Sites with >75% heterozygosity were removed as potential paralogs, and the last five bases trimmed from each locus to mitigate potential sequencing error due to primer degradation. Various clustering thresholds, ranging from 70% to 97%, were tested and the optimal threshold determined (CLUSTOPT; McCartney-Melstad *et al.* 2019) via R v3.6.3 (R Development Core Team 2018). In doing so, we evaluated multiple clustering thresholds using three metrics: 1) Cumulative variance [principal component analysis (PCA)], 2) Pearson's correlation coefficients between genetic distance versus percent missing data, and 3) isolation-by-distance (IBD). The optimal threshold is interpreted as the inflection point in the resulting slopes. Finally, to remove poorly sequenced individuals (>90% missing data), a post-hoc filtering of alignments was applied via a custom script (*nremover.pl*; github.com/tkchafin/scripts).

2.3. Phylogenomic analysis of *Sistrurus*

We first inferred a maximum-likelihood phylogenomic tree for *Sistrurus* (IQ-TREE v2.0.6; Minh *et al.* 2020), using the edge-linked partition model (Chernomor *et al.* 2016) with sister genera *Crotalus* and *Agkistrodon* as outgroups. Loci without parsimoniously informative sites were removed and the remaining full loci were concatenated into a NEXUS file containing a per-locus partition block, as executed via three custom scripts [*filterLoci.py*, *concatenateNexus.py*, and *filterUninformative.py* (github.com/tkchafin/scripts, and github.com/btmartin721/ddrad_scripts)]. Those partitions (=ddRAD loci) exhibiting overlapping substitution models were merged using the fast-relaxed clustering method ('--rclusterf' option), with the top 10% of clusters considered, and the optimal nucleotide substitution model for each super-partition selected (MODELFINDER; Kalyaanamoorthy *et al.* 2017). A consensus tree was then constructed from the merged partitions, 1,000 ultrafast bootstrap (=UFBOOT) replicates (Hoang *et al.* 2017), and the '--bnni' and '--safe' options that mitigated overestimation of UFBOOT-support and prevented numerical underflow, respectively.

2.4. Divergence dating

We estimated divergence times from the IQ-TREE species tree using the LSD2 (least square dating v2) approach (IQ-TREE v2.1.2; To *et al.* 2016). Molecular clock calibrations included fossil-based minimum/ maximum age constraints for four nodes, with lower bounds supplied for the MRCA (most recent common ancestor). These are: the *Sistrurus* genus [(9.0 Mya) (Parmley & Holman 2007)], *Agkistrodon contortrix* + *A. piscivorous* [(4.5 Mya) (Holman 2000; Guiher & Burbrink 2008; Douglas *et al.* 2009)], *Crotalus* + *Sistrurus* [(15.5 Mya) (Parmley & Holman 2007)], and *Agkistrodon* + *Crotalus* [(7.0 Mya) (Douglas *et al.* 2006, 2009)]. A maximum root constraint of 22.0 Mya represented the earliest known pitviper fossil (Holman 2000; Parmley & Holman 2007; Douglas *et al.* 2009). To calculate confidence intervals for the divergence times, LSD2 was run on 1,000 bootstrapped trees with branch lengths simulated from a relaxed-clock Poisson distribution. The resulting phylogeny was then plotted [R packages: PHYTOOLS (Revell 2012); GGTREE (Yu *et al.* 2017)].

2.5. Phylogeographic signal within *Sistrurus*

We explored our time-calibrated phylogenetic signal using a Brownian motion evolutionary model corresponding to the observed spatial signal (PICANTE R package; Kembel *et al.* 2010). Phylogenetic signal was tested using the *multiPhylosignal* function for all combinations of adjacent Massasauga taxa (i.e., *S. catenatus, S. t. tergeminus,* and *S. t. edwardsii,* plus the whole tree). Signal was significantly influenced by latitude and longitude when the *K*-statistic (Blomberg *et al.* 2003) was >1.0 and the phylogenetic signal non-random and greater than expected ($P < 0.05$).

We also estimated the distance of each phylogeny tip from the most ancestral node per subspecies (PALEOTREE R package; Bapst 2012). Here, the time-calibrated IQ-TREE phylogram was input into the dateNodes function, and the number of tips that occur between the current one and the root of the clade was calculated. Nodal distances were standardized as proportions to account for differences in clade depth.

2.6. Admixture and population structure

We assessed population structure and admixture [ADMIXTURE v1.3.0 (Alexander & Lange 2011) via ADMIXPIPE (Mussmann *et al.* 2020)]. Sites that were either monomorphic or had a minor allele frequency <1% were *a priori* removed from the unlinked IPYRAD SNP output (Linck & Battey 2019) using ADMIXPIPE filtering options. We ran ADMIXTURE for *K*=1-10 clusters with 20 replicates per *K* and 20-fold cross-validation (CV). Optimal *K* represented the lowest CV score and the ADMIXTURE results were visualized as stacked bar-plots (ADMIXPIPE scripts *cvSum.py* and *distructRerun.py*). ADMIXTURE proportions were plotted as pie charts on a range map of

Sistrurus so as to visualize spatial distributions of derived clusters (QGIS v3.16; QGIS Development Team 2009).

2.7. Tests of hybridization and deep-time reticulation

We also evaluated if our observed admixture was statistically supported [4-taxon *D*-statistic (Green *et al.* 2010; Durand *et al.* 2011) via COMP-D (Mussmann *et al.* 2019)]. As with ADMIXTURE, singletons and monomorphic sites were removed. We employed a custom script to generate the input COMP-D files (*makeCompD.py*; https://github.com/tkchafin/makeCompD). The MPI (message passing interface) version of COMP-D was run with 1,000 bootstrap replicates and heterozygous sites included ('--hinclude' option). Benjamini and Yekutieli (B-Y) corrections were independently applied to adjust *P*-values for multiple comparisons (Bonferroni 1936; Benjamini & Yekutieli 2001). Positive *D*-statistic tests were then summarized per individual and population.

We inferred phylogenetic networks and computed ancestry components (SNAQ method in PHYLONETWORKS; Solís-Lemus & Ané 2016; Solís-Lemus *et al.* 2017)as a means to refine hypothesized reticulations previously identified using ADMIXTURE and the *D*-statistic. We first computed locus-wise concordance factors (CFs) as our input, using a Bayesian concordance analysis (BUCKy; Larget *et al.* 2010), run in parallel across quartets (TICR pipeline; Stenz *et al.* 2015). Given the computational overhead imposed by network inference, we computed our CFs across a reduced dataset of 3,988 loci, chosen as a subset of non-monomorphic loci containing at least five parsimony-informative sites, and for which at least a single diploid genotype could be sampled per targeted tip. To maximize the number of loci employed, we reduced taxa to eight sampled tips, to include the major taxa and intraspecific clades identified in earlier analyses. Posterior distributions were generated for each gene-tree (MRBAYES v3.2.6; Ronquist *et al.* 2012) with 4 independent chains each and 100,000,000 iterations, 50% of which were discarded as burn-in, with sampling every 10,000 generations so as to reduce autocorrelation among samples. Quartet concordance factors (CFs) were then run across all possible four-taxon combinations, using a chain length of 10,000,000 with 50% burn-in. As input to PHYLONETWORKS, a starting tree was first generated (QUARTETMAXCUT; Snir & Rao 2012), followed by network estimation under models of 0-6 hybrid nodes (*h*). Models were evaluated using 100 independent replicates each, with the best-fit model maximizing first-order change in pseudo-likelihood (ΔL).

To corroborate our PHYLONETWORKS analysis, we also estimated admixture edges (*m*) along a maximum-likelihood phylogeny [TREEMIX v1.1 (Pickrell & Pritchard 2012) via the IPYRAD analysis toolkit (Eaton & Overcast 2020)]. The input SNP alignment was randomly subsampled to one SNP per ddRAD locus, replicated nine times. Optimal *m* was determined for each subsample replicate by observing the inflection point of the log-likelihood scores, using 1,000 bootstrap replicates and the more thorough 'global' search option.

2.8. Demographic modeling using GADMA

We assessed the pattern of introgression events in Missouri *Sistrurus* by employing demographic modeling [(MOMENTS; Jouganous *et al.* 2017) via the GADMA pipeline (Noskova *et al.* 2020)]. A joint site-frequency spectrum (jSFS) was applied to user-defined demographic models so as to estimate the size (N_E), divergence time (τ), and migration (m) parameters of each population. an iterative MOMENTS model search was automated in GADMA to select the optimal model per the Akaike Information Criterion (AIC). Model search terminates when log-likelihood values fail to improve over 100 genetic algorithm iterations. GADMA then further optimizes the parameters of the selected model using local optimization searches in MOMENTS.

Two GADMA runs were conducted. First, the *Sistrurus* alignment was subset to three populations (the maximum allowed). These were: *Sistrurus t. tergeminus* from Missouri (=TEMO), *S. t. tergeminus* from all other localities (=TERG), and *S. catenatus* (=CAT). Second, GADMA was run with *S. t. tergeminus* and *S. t. edwardsii*. Sites in each input alignment were filtered to a minimum of 50% site-wise coverage in any given population, and SNPs thinned to one random bi-allelic per ddRAD locus. This random selection was then repeated 100 times to provide non-parametric bootstrap replicates for estimating GADMA parameter confidence intervals (*easySFS.py* script; https://github.com/isaacovercast/easySFS). Alleles in each resulting SFS were down-projected (*easySFS.py*) to yield counts that maximized the number of segregating sites (per the MOMENTS manual). The GADMA model search was permitted to explore up to two divergence times prior to and subsequent to each population split. GADMA requires the population size of the reference population (N_{ref}) as a means of scaling the moments parameters in actual time (years), rather than genetic units. This was calculated as:

$$N_{ref} = \frac{\theta}{\theta_0} = \frac{4 \times Ne \times u \times L}{4 \times \mu \times L},$$

where N_e is the effective population size, μ is the mutation rate per generation and L is the effective sequence length after filtering. The effective sequence length was calculated as:

$$L = \frac{total\ sequence\ analyzed\ \times\ SNPs\ retained\ for\ use\ in\ moments}{total\ SNPs\ in\ analyzed\ sequence}.$$

To calculate a mutation rate [per Gutenkunst et al. (2009) and the GADMA manual], we first derived the following: Average sequence divergence (D_{xy}) from the input alignment (DNASP v6.12.03; Rozas *et al.* 2017), a *Sistrurus* generation time (G) of ~4.0 years (Sovic *et al.* 2016), and a divergence time for *S. catenatus* X *S. t. tergeminus*, as estimated from the time-calibrated phylogeny. Given these, we then calculated the mutation rate as:

$$\mu = \frac{Dxy \times G}{2 \times \tau},$$

3. RESULTS

3.1. Data assembly and filtering

The IPYRAD clustering threshold was set to 0.82 to maintain a weak correlation ($r \leq 0.3$) between genetic distance and percent missing data in the alignment (Fig. S1A). IBD and the cumulative variance from PCA were largely unaffected by clustering thresholds (Figs. S1B, S1C). The final IPYRAD alignment included 49,879 parsimoniously informative sites across 10,190 ddRAD loci from 226 *Sistrurus* individuals (Fig. 1; Table S2).

3.2. Phylogeographic and demographic analyses

The ML phylogeny found reciprocal monophyly for *S. miliarius*, *S. catenatus*, and *S. t. tergeminus* + *S. t. edwardsii*. However, *S. t. tergeminus* was paraphyletic with respect to *S. t. edwardsii* (Figs. 2A, 3). The *S. tergeminus* clade was largely pectinate from the northeast to the southwest, with the most ancestral individuals in Missouri and the most derived in Texas and New Mexico. This longitudinal signal was statistically corroborated (PICANTE; P=0.001), with the *K*-statistic >1.0 for all but the whole-tree analysis. The longitudinal signal in the analysis was consistently higher than latitudinal (Table 1), except when *S. t. edwardsii* was independently considered. Finally, PALEOTREE indicated disparate directional patterns of increasingly derived lineages (Fig. 2B), with *S. catenatus* extending northeast and *S. tergiminus* southwest. Individuals with the least nodal distance to the root node in *S. t. tergeminus* and *S. t. edwardsii* were in Kansas/ Missouri/ Iowa, and Texas/ New Mexico.

The Midwestern GADMA SFS was down projected to yield 30 x 30 x 28 alleles for *S. catenatus, S. t. tergeminus* (Missouri), and all other *S. t. tergeminus* (hereafter referred to as the 'CAT x TERG x TEMO' model). Southwestern *S. t. edwardsii*/ *S. t. tergeminus* ('ED X TERG' model) were down projected to 40 x 56 alleles. The optimal demographic model for CAT x TERG x TEMO included a post-divergence period of weak migration followed by considerable recent migration within the last 7 Kya (Figs. 4A, 4B). Every supported migration edge and population bottleneck involved *S. catenatus* and *S. t. tergeminus* (Missouri). In contrast, a strong continuous migration since divergence was found in ED x TERG, as well as a recent (<7 Kya) *S. t. edwardsii* population expansion (Figs. 4C, 4D).

3.3. Hybridization in two North American contact zones

The CV scores from ADMIXTURE (Fig. S2) suggested an optimal *K*=5, with *K*=4 reasonably close. The five identified clusters included *S. catenatus* and two populations for each of *S. t. tergeminus* and *S. t. edwardsii* (Fig. 2C). The two *S. t. tergeminus* populations were generally (albeit porously) geographically localized along a gradient from Northwest (darker blue shade primarily in northcentral MO/ eastern Kansas/ southeastern Nebraska) to Southwest (lighter blue in Kansas/ Texas/ Oklahoma/ New Mexico) clusters (Figs. 2C, 2D). The dark blue population also aligns with the point-of-origin from PALEOTREE analysis (Fig. 2B). Similarly, *S. t. edwardsii* was partitioned into Texas/ New Mexico/ Arizona (darker red) and Colorado (lighter red) populations (Figs. 2C, 2D). Admixture across species/ subspecies was localized regionally in the midwestern and southwestern contact zones. Results from the *K*=4 admixture analysis (Fig. S3) were very similar, with the exception that *S. t. edwardsii* (CO) did not represent its own cluster but instead displayed ~50% ancestry between southern *S. t. tergeminus* and *S. t. edwardsii*.

The four-taxon *D*-statistic tests only supported introgression between *S. t. tergeminus* (MO) and *S. catenatus* (Figs. 5A, 5B, 6), with significant tests negligible among southwestern *S. t. tergeminus* and *S. t. edwardsii* (Fig. S4). Results from PHYLONETWORKS and TREEMIX agreed (Fig. 6), with a single admixture edge (Fig. S5) between *S. t. tergeminus* (MO) + *S. catenatus.* The ancestry proportions (α) from PHYLONETWORKS assigned *S. t. tergeminus* (MO) as 33.8% *S. catenatus*. TREEMIX results differed slightly from *D*-statistics and PHYLONETWORKS by only connecting *S. catenatus* (IA) to *S. t. tergeminus* (MO), whereas the latter two did so with all *S. catenatus*.

4. DISCUSSION

4.1. Population structure in *Sistrurus tergeminus*

Kubatko et al. (2011) identified *S. catenatus* as being monophyletic, and *S. t. tergeminus* paraphyletic with respect to *S. t. edwardsii*. They also partitioned *S. t. tergeminus* into MO and KS clades, with the latter sister to *S. t. edwardsii*. Our ADMIXTURE results broadened and extended their findings by recognizing northeastern (Missouri/ eastern Kansas) and southwestern (south-central Kansas/ Nebraska/ Oklahoma/ Texas) populations of *S. t. tergeminus* (Figs. 2C, 2D). Although reproductive boundaries between the two populations are seemingly porous, their primary separation corresponds to the Arkansas River, a barrier likewise significant in other studies (Fontanella *et al.* 2008; Ruane *et al.* 2014; Herman & Bouzat 2016). Differentiation of central Texas and Oklahoma has also been previously observed (Ryberg *et al.* 2015), with our results again in agreement. Our ADMIXTURE results also delineated two *S. t. edwardsii* populations in Texas/ New Mexico/ Arizona and Colorado, in concordance with their topograpic separation. The initial separation of northern and southern *S. t. edwardii* populations may have been due to the occurrence of glaciation within the southern Rocky Mountains (Hafner & Sullivan 1995; Arbogast *et al.* 2001).

However, our results also contrast with previous studies that found minimal structure within *S. t. tergeminus* (McCluskey & Bender 2015) or failed to support subspecific boundaries (Ryberg *et al.* 2015). The former study was geographically constrained to KS and MO, and thus may have missed the southern population. Additionally, both studies were based on microsatellite or mitochondrial DNA/ introns, and thus had a reduced capacity to discriminate relative to our large SNP dataset (Rašić *et al.* 2014; Vendrami *et al.* 2017), particularly given the complications introduced by admixture (Haasl & Payseur 2011).

4.2. Hybridization and phylogeography in *Sistrurus tergeminus* and *S. catenatus*

We demonstrated that admixture occurred regionally in two contact zones, although contrasting evolutionary processes may be involved. In Missouri, 22/24 *S. t. tergeminus* (92%) were identified as hybrids, with the network analysis attributing 33.8% ancestry to *S. catenatus*. Our results concur with a previous morphological assessment that portrayed these Missouri populations as intermediate (Evans & Gloyd 1948), yet contradict more recent molecular studies utilizing microsatellites and nuclear introns (Gerard *et al.* 2011; Gibbs *et al.* 2011). Given that the Missouri populations are disjunct from the larger (and contiguous) populations of *S. catenatus* and *S. t. tergeminus*, the introgression would seemingly be defined as both historical and discontinued. In this sense, SNP data have greater capacity to discern evolutionary and demographic processes that have occurred at deeper time scales (Morin *et al.* 2004; Haasl & Payseur 2011; Lee *et al.* 2018; Camacho-Sanchez *et al.* 2020; Chafin *et al.* 2021).

GADMA further supported periods of intermittent secondary contact and isolation, likely due to glacial cycles and the vicariant barrier created by the Mississippi or Missouri River (Szymanski 1998; Sovic *et al.* 2016). Indeed, the Mississippi River has impacted phylogeographic patterns in other viperid snakes, such as *Agkistrodon* (Douglas *et al.* 2009) as well as Iowa *Sistrurus* (Sovic *et al.* 2016) where demographic modeling supported secondary contact consistent with our GADMA analysis, and at similarly-estimated times (11 *versus* 7 Kya) (Figs. 4A, 4B) with both models.

Secondary contact and northeastern expansion in *S. catenatus* (Fig. 2B; Table 1) may stem from periodic vicariance. Likely candidates include the Mississippi and Missouri Rivers, which saw drastic increases in glacial discharge from the Appalachian and Rocky Mountains during late Miocene [(~4 Mya) (Bentley Sr *et al.* 2016)]. Furthermore, the late Miocene/ early Pliocene transition is consistent with the estimated divergence time between *S. catenatus* and *S. t. tergeminus* (~4.86 Mya; 95% CI=4.09 – 6.25 Mya). The earliest known fossils attributable to *S. catenatus* or *S. tergeminus* date to the Pliocene in Kansas, Texas, and Nebraska, whereas the earliest identifiable *Sistrurus* dates to the Miocene (~9 Mya) in Nebraska. It is likely the MRCA to *S. catenatus/ S. tergeminus* originated in the Great Plains and dispersed both northeastward and southwestward, per our PALEOTREE results. However, northeastern *S. catenatus* could alternatively reflect newer, post-glacial populations that expanded from refugia. The more 'ancestral' *S. t. tergeminus* lineages are Missouri and Iowa, in agreement with the fossil record (Holman 2000; Parmley & Holman 2007).

It is also possible that the observed introgression in Missouri individuals has yielded artificially short nodal-root distances (Bangs *et al.* 2018). A similar pattern is found in *S. t. edwardsii* where individuals with short nodal distances in Texas/ New Mexico may represent artifacts of recent admixture in the southwestern contact zone. Although admixture seemingly affects diversification patterns in PALEOTREE, eastern Kansas does not contain admixed individuals and may thus represents the origin of both northeastern and southwestern diversifications, a prospect consistent with the fossil record.

Despite ADMIXTURE results depicting mixed ancestry in the southwestern contact zone, our *D*-statistics, PHYLONETWORKS, or TREEMIX results failed to statistically support introgression between *S. t. tergeminus* and *S. t. edwardsii* (Fig. 6). The paraphyletic relationship of the two *S. tergeminus* subspecies and their more recent divergence time may indicate ongoing primary divergence. Our GADMA analyses concur (Figs. 4C, 4D), with strong, persistent migration occurring since divergence. The southwestern contact zone seemingly lacks an apparent vicariant barrier (Kubatko *et al.* 2011), and this would conform with ongoing primary divergence.

Distinct ecological and physiological differences between *S. t. tergeminus* and *S. t. edwardsii* may also contribute towards divergence. For example, *S. t. edwardsii* prefers ectothermic prey (Holycross & Mackessy 2002) and is found in xeric grasslands and dunes

(Hammerson 1999; Stebbins 2003; Degenhardt *et al.* 2005; Wastell & Mackessy 2011, 2016), whereas *S. t. tergeminus* prefers mammalian prey (Holycross & Mackessy 2002) and mesic prairies and grasslands (Seigel 1986). These differences are reinforced by ecological niche modeling (Wooten & Gibbs 2012), with temperature and precipitation regimes delineating the two subspecies. They also differ in venom composition which may also support their dietary preferences (Sanz *et al.* 2006; Gibbs & Mackessy 2009; but see Gibbs *et al.* 2013).

Our more recent divergence estimated for *S. t. tergeminus* and *S. t. edwardsii* (1.44 Mya; 95% CI=1.26-1.73 Mya) coincides with the Pliocene-Pleistocene transition (~1.5-2.0 Mya). The southwestern United States has undergone many climate and vegetational fluctuations since (Savage 1960; Findley 1969; Morafka 1977; Axelrod 1983), oscillating between aridification/ increasing xeric shrub vegetation (Axelrod 1979; Wilson & Pitts 2010) versus cooler climates and mesic habitat more suitable for *S. t. tergeminus* (MacKay & Elias 1992; Pendall *et al.* 1999; Holycross 2002; Holmgren *et al.* 2003; Wilson & Pitts 2010). The region also seemingly arrived at its current state ~8-4 Kya (Van Devender 1977; MacKay & Elias 1992; Hunter *et al.* 2001; Holmgren *et al.* 2003), and is consistent with population expansion from ~7 Kya, as modeled by GADMA (Fig. 4). Moreover, mesic versus xeric habitat fluctuation may have contributed to a gradual divergence by promoting long-term mosaic zones of contact between the two *S. tergeminus* ecotypes. A contemporary example that may resemble Pleistocene conditions can be seen in northwestern Texas, where intermediate habitat forms a contact zone that spans numerous taxa (Swenson & Howard 2005). We cannot confirm if *S. tergeminus* was present in the southwest at end-of-Pliocene, or if it gradually encroached afterwards, given a weak fossil record (Holman 2000; Parmley & Holman 2007). Nevertheless, our observed diversification patterns and a recent but highly labile habitat suggest a strong ecological component to divergence.

4.3. Conservation implications

Our analyses clearly delineate *S. catenatus* and *S. tergeminus* and support previous assessments of genetic diversity. First, GADMA showed a population bottleneck in *S. catenatus* consistent with recent estimates of small effective population size (Sovic *et al.* 2019; Ochoa *et al.* 2020), and with major contributing factors to their decline being habitat fragmentation and loss (Szymanski *et al.* 2016). Second, the higher genetic diversity found in *S. t. tergeminus* and *S. t. edwardsii* also supports other recent studies, with each having greater effective population size compared to *S. catenatus* (Sovic *et al.* 2016) and recent population growth (Anderson *et al.* 2009). The *S. t. tergeminus* habitat in at least parts of its range is less affected by anthropogenic fragmentation (Greene 1997; Szymanski 1998; McCluskey & Bender 2015), although previously undescribed population structure (herein) warrants further investigation.

Our analyses also identified population structure within *S. t. edwardsii*, as an addendum to previous research (e.g., Anderson *et al.* 2009), but with the recognition that population structure also transcended subspecific boundaries (Ryberg *et al.* 2015). Although *S. t. edwardsii* has elevated genetic diversity, its habitat is either disappearing or being seriously fragmented (Lowe *et al.* 1986; Greene 1997; Werler & Dixon 2010), which in turn implies an impending 'drift debt' (per *S. catenatus*; Ochoa *et al.* 2020) that stems from the time lag between habitat fragmentation/ loss and a decline in genetic diversity. Given the observed genetic differentiation as well as the aforementioned ecological and physiological disparities, we posit that *S. t.*

tergeminus and *S. t. edwardsii* should be recognized as evolutionarily significant units (ESUs). The population structure within each also reflects a cryptic genetic diversity potentially warranting consideration as management units (MUs).

4.4. Conclusions

Herein, we have expanded upon and contextualized the population structure and phylogeography of *S. catenatus* and *S. tergeminus* ssp. In doing so, we demonstrate these taxa have a more complex and diversified evolutionary history than previously understood. In the northeastern *S. tergeminus* range, speciation has been strongly influenced by vicariant barriers and secondary contact, whereas a primary divergence event may be currently underway in the southwest, as facilitated by ecological differences and a habitat whose availability has fluctuated since Pliocene.

ACKNOWLEDGEMENTS

This manuscript represents one component of materials submitted in partial fulfillment of the Ph.D. degree in Biological Sciences at University of Arkansas/ Fayetteville. I thank the countless colleagues, volunteers, museums, and agencies who have either collected or contributed samples for this project (Table S1). I also thank several members of the Douglas Lab and University of Arkansas faculty for their invaluable guidance, to include: A. Alverson, W. Anthonysamy, M. Bangs, J. Beaulieu, T. Chafin, S. Mackessy, S. Mussmann, J. Placyk, J. Pummill, and Z. Zbinden. Funding sources included University of Arkansas endowments to MRD and MED. The

Arkansas High Performance Computing Cluster (AHPCC) and Jetstream cloud (XSEDE #TG-BIO160065) provided computational resources.

5. REFERENCES

Adams J, Maslin M, and Thomas E (1999) Sudden climate transitions during the Quaternary. *Progress in Physical Geography*, **23**, 1–36.

Alexander DH and Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**, 246.

Anderson E (1949) *Introgressive Hybridization*. John Wiley and Sons, New York City, NY, USA.

Anderson CD, Gibbs HL, Douglas ME, and Holycross AT (2009) Conservation genetics of the desert massasauga rattlesnake (*Sistrurus catenatus edwardsii*). *Copeia*, **2009**, 740–747.

Antonelli A (2017) Biogeography: drivers of bioregionalization. *Nature Ecology & Evolution*, **1**, 1–2.

Arbogast BS, Browne RA, and Weigl PD (2001) Evolutionary genetics and Pleistocene biogeography of North American tree squirrels (*Tamiasciurus*). *Journal of Mammalogy*, **82**, 302–319.

Avise JC (2000) *Phylogeography: the history and formation of species*. Harvard University Press, Cambridge, MA, USA.

Avise JC, J. Arnold, R.M. Ball, E. Bermingham, T. Lamb, J.E. Neigel, Reeb CA, Saunders NC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC, J. Arnold, R.M. Ball, E. Bermingham, T. Lamb *et al.* (1987) Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Sytematics*, **18**, 489–522.

Axelrod DI (1948) Climate and evolution in western North America during middle Pliocene time. *Evolution*, **2**, 127–144.

Axelrod DI (1979) Age and origin of Sonoran Desert vegetation. *Occasional Papers of the California Academy of Sciences*, **132**, 1–74.

Axelrod DI (1983) Paleobotanical history of the western deserts. In: *Origin and Evolution of Deserts* (eds Wells SG and Haragan DR), pp. 113–129. University of New Mexico Press, Albuquerque, New Mexico, USA.

Axelrod DI (1985) Rise of the grassland biome, central North America. *Bot. Rev.*, **51**, 163–201.

Bangs MR, Douglas MR, Mussmann SM, and Douglas ME (2018) Unraveling historical introgression and resolving phylogenetic discord within *Catostomus* (Osteichthys: Catostomidae). *BMC Evolutionary Biology*, **18**, 86.

Bapst DW (2012) paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology and Evolution*, **3**, 803–807.

Benjamini Y and Yekutieli D (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, **29**, 1165–1188.

Bentley Sr SJ, Blum MD, Maloney J, Pond L, and Paulsell R (2016) The Mississippi River source-to-sink system: Perspectives on tectonic, climatic, and anthropogenic influences, Miocene to Anthropocene. *Earth-Science Reviews*, **153**, 139–174.

Blomberg SP, Garland Jr T, and Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita'. *Pubbl. R Ist. Sup. Sci. Econ. Commer. Fir*, **8**, 3–62.

Braun EL (1950) Deciduous Forests of Eastern North America. *Soil Science*, **71**, 155.

Camacho-Sanchez M, Velo-Antón G, Hanson JO, Veríssimo A, Martínez-Solano Í, Marques A, Moritz C, and Carvalho SB (2020) Comparative assessment of range-wide patterns of genetic diversity and structure with SNPs and microsatellites: A case study with Iberian amphibians. *Ecology and Evolution*, **10**, 10353–10363.

Chafin TK, Zbinden ZD, Douglas MR, Martin BT, Middaugh CR, Gray MC, Ballard JR, and Douglas ME (2021) Spatial population genetics in heavily managed species: Separating patterns of historical translocation from contemporary gene flow in white-tailed deer. *bioRxiv, https://doi.org/10.1101/2020.09.22.308825*.

Chernomor O, Von Haeseler A, and Minh BQ (2016) Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology*, **65**, 997–1008.

Chiucchi JE and Gibbs HL (2010) Similarity of contemporary and historical gene flow among highly fragmented populations of an endangered rattlesnake. *Molecular Ecology*, **19**, 5345–5358.

Clark PU, MacAyeal DR, Andrews JT, and Bartlein PJ (1995) Ice sheets play important role in climate change. *Eos, Transactions American Geophysical Union*, **76**, 265–270.

Cole MR and Armentrout JM (1979) Neogene paleogeography of the western United States. In: *Pacific Section Society of Economic Paleontologist and Mineralogist Pacific Coast Paleogeography Symposium, 3*, pp. 297–323. Pacific Section SEPM.

Cook FR (1993) After an Ice Age: Zoogeography of the Massasauga within a Canadian Herpetofaunal Perspective. In: *Rattlesnake Symposium*, pp. 19–25. Metro Toronto Zoo, Toronto, CA.

Degenhardt WG, Painter CW, and Price AH (2005) *Amphibians and Reptiles of New Mexico*. UNM Press.

Van Devender TR (1977) Holocene woodlands in the southwestern deserts. *Science*, **198**, 189–192.

Van Devender TR, Thompson RS, and Betancourt JL (1987) Vegetation history of the deserts of southwestern North America: the nature and timing of the late Wisconsin-Holocene transition. In: *North America and Adjacent Oceans During the Last Glaciation: The Geology of North America* (eds Ruddiman WF and Wright Jr. HE), pp. 323–352. Geological Society of America, Boulder, CO, USA.

Douglas ME, Douglas MR, Schuett GW, and Porras LW (2006) Evolution of rattlesnakes (Viperidae; *Crotalus*) in the warm deserts of western North America shaped by Neogene vicariance and Quaternary climate change. *Molecular Ecology*, **15**, 3353–3374.

Douglas ME, Douglas MR, Schuett GW, and Porras LW (2009) Climate change and evolution of the New World pitviper genus *Agkistrodon* (Viperidae). *Journal of Biogeography*, **36**, 1164–1180.

Durand EY, Patterson N, Reich D, and Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**, 2239–2252.

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

Evans PD and Gloyd HK (1948) The subspecies of the massasauga, *Sistrurus catenatus*, in Missouri. *Bulletin of the Chicago Academy of Sciences*, **8**, 225–232.

Findley JS (1969) Biogeography of southwestern boreal and desert animals. In: *Contributions in Mammology: A Volume Honering Professor E. Raymond Hall* (ed Jones Jr. JK), pp. 113–128. University of Kansas Press, Lawrence, Kansas, USA.

Fontanella FM, Feldman CR, Siddall ME, and Burbrink FT (2008) Phylogeography of *Diadophis punctatus*: extensive lineage diversity and repeated patterns of historical demography in a trans-continental snake. *Molecular Phylogenetics and Evolution*, **46**, 1049–1070.

Gerard D, Gibbs HL, and Kubatko L (2011) Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evolutionary Biology*, **11**, 291.

Gibbs HL and Mackessy SP (2009) Functional basis of a molecular adaptation: prey-specific toxic effects of venom from *Sistrurus* rattlesnakes. *Toxicon*, **53**, 672–679.

Gibbs HL, Murphy M, and Chiucchi JE (2011) Genetic identity of endangered massasauga rattlesnakes (*Sistrurus* sp.) in Missouri. *Conservation Genetics*, **12**, 433–439.

Gibbs HL, Sanz L, Sovic MG, and Calvete JJ (2013) Phylogeny-based comparative analysis of venom proteome variation in a clade of rattlesnakes (*Sistrurus* sp.). *PloS one*, **8**, e67220.

Graham MR, Hendrixson BE, Hamilton CA, and Bond JE (2015) Miocene extensional tectonics explain ancient patterns of diversification among turret-building tarantulas (*Aphonopelma mojave* group) in the Mojave and Sonoran deserts. *Journal of Biogeography*, **42**, 1052–1065.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, and Fritz MH-Y (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Greene HW (1997) *Snakes: The Evolution of Mystery in Nature*. Univ of California Press, Berkeley and Los Angeles, California, USA.

Greene HW and Oliver G V (1965) Notes on the natural history of the western massasauga. *Herpetologica*, **21**, 225–228.

Guiher TJ and Burbrink FT (2008) Demographic and phylogeographic histories of two venomous North American snakes of the genus *Agkistrodon*. *Molecular Phylogenetics and Evolution*, **48**, 543–553.

Gutenkunst RN, Hernandez RD, Williamson SH, and Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, **5**, e1000695.

Haasl RJ and Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, **106**, 158–171.

Hafner DJ and Sullivan RM (1995) Historical and ecological biogeography of Nearctic pikas (Lagomorpha: Ochotonidae). *Journal of Mammalogy*, **76**, 302–321.

Hammerson GA (1999) *Amphibians and Reptiles in Colorado*. University Press of Colorado.

Harvey DS and Weatherhead PJ (2006) A test of the hierarchical model of habitat selection using eastern massasauga rattlesnakes (*Sistrurus c. catenatus*). *Biological Conservation*, **130**, 206–216.

Herman TA and Bouzat JL (2016) Range-wide phylogeography of the four-toed salamander: out of Appalachia and into the glacial aftermath. *Journal of Biogeography*, **43**, 666–678.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnaean Society*, **58**, 247–276.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnaean Society*, **68**, 87–112.

Hewitt GM (2001) Speciation, hybrid zones and phylogeography—or seeing genes in space and time. *Molecular Ecology*, **10**, 537–549.

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.

Hobert JP, Montgomery CE, and Mackessy SP (2004) Natural history of the massasauga, *Sistrurus catenatus edwardsii*, in southeastern Colorado. *The Southwestern Naturalist*, **49**, 321–326.

Holman JA (2000) *Fossil snakes of North America: origin, evolution, distribution, paleoecology*. Indiana University Press, Bloomington, IN, USA.

Holmgren CA, Penalba MC, Rylander KA, and Betancourt JL (2003) A 16,000 14C yr BP packrat midden series from the USA–Mexico Borderlands. *Quaternary Research*, **60**, 319–329.

Holycross AT (2002) Conservation biology of two rattlesnakes, Crotalus willardi obscurus and Sistrurus catenatus edwardsii. Arizona State University.

Holycross AT and Mackessy SP (2002) Variation in the diet of *Sistrurus catenatus* (Massasauga), with emphasis on *Sistrurus catenatus edwardsii* (Desert Massasauga). *Journal of Herpetology*, **36**, 454–464.

Hunter KL, Betancourt JL, Riddle BR, Van Devender TR, Cole KL, and Spaulding WG (2001) Ploidy race distributions since the Last Glacial Maximum in the North American desert shrub, *Larrea tridentata*. *Global Ecology and Biogeography*, **10**, 521–533.

Johnson G (2000) Spatial ecology of the eastern massasauga (*Sistrurus c. catenatus*) in a New York peatland. *Journal of Herpetology*, **34**, 186–192.

Jouganous J, Long W, Ragsdale AP, and Gravel S (2017) Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, **206**, 1549–1567.

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.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, and Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

Kubatko LS, Gibbs HL, and Bloomquist EW (2011) Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in Sistrurus rattlesnakes. *Systematic Biology*, **60**, 393–409.

Larget BR, Kotha SK, Dewey CN, and Ané C (2010) BUCKy: Gene tree/species tree reconciliation with concordance analysis. *Bioinformatics*, **26**, 2910–2911.

Lee KM, Kivelä SM, Ivanov V, Hausmann A, Kaila L, Wahlberg N, and Mutanen M (2018) Information dropout patterns in RAD phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus. *Systematic Biology*, **67**, 925–939.

Licciardi JM, Clark PU, Brook EJ, Elmore D, and Sharma P (2004) Variable responses of western US glaciers during the last deglaciation. *Geology*, **32**, 81–84.

Linck EB and Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. *Molecular Ecology Resources*, **19**, 639–647.

Lowe CH, Schwalbe CR, and Johnson TB (1986) *The venomous reptiles of Arizona*. Phoenix, AZ.

MacKay WP and Elias SA (1992) Late Quaternary ant fossils from packrat middens (Hymenoptera: Formicidae): implications for climatic change in the Chihuahuan Desert. *Psyche*, **99**, 169–184.

Mackessy S (2005) Desert Massasauga Rattlesnake (Sistrurus catenatus edwardsii): a technical conservation assessment. [Online].

Martin BT, Douglas MR, Chafin TK, Placyk JS, Birkhead RD, Phillips CA, and Douglas ME (2020) Contrasting signatures of introgression in North American box turtle (*Terrapene* spp.) contact zones. *Molecular Ecology*, **29**, 4186–4202.

McCartney-Melstad E, Gidiş M, and Shaffer HB (2019) An empirical pipeline for choosing the optimal clustering threshold in RADseq studies. *Molecular Ecology Resources*, **19**, 1195–1204.

McCluskey EM and Bender D (2015) Genetic structure of western massasauga rattlesnakes (Sistrurus catenatus tergeminus). *Journal of Herpetology*, **49**, 343–348.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, and Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, **37**, 1530–1534.

Minton SA (1983) *Sistrurus catenatus*. *Catalogue of American Amphibians and Reptiles (CAAR), Society for the Study of Amphibians and Reptiles*.

Morafka DJ (1977) A biogeographical analysis of the Chihuahuan desert through its herpetofauna. *Biogeographica*, **9**, 1–313.

Morin PA, Luikart G, and Wayne RK (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Murphy RW, Fu J, Lathrop A, Feltham J V, and Kovac V (2002) Phylogeny of the rattlesnakes (Crotalus and Sistrurus) inferred from sequences of five mitochondrial DNA genes. In: *Biology of the Vipers* (eds Schuett GW, Höggren M, Douglas ME and Greene HW), pp. 69–92. Eagle Mountain Press LC, Salt Lake City, UT, USA.

Mussmann SM, Douglas MR, Bangs MR, and Douglas ME (2019) Comp-D: a program for comprehensive computation of D-statistics and population summaries of reticulated evolution. *Conservation Genetics Resources*, 1–5.

Mussmann SM, Douglas MR, Chafin TK, and Douglas ME (2020) AdmixPipe: population analyses in Admixture for non-model organisms. *BMC Bioinformatics*, **21**, 1–9.

Myers EA, Xue AT, Gehara M, Cox CL, Davis Rabosky AR, Lemos-Espinal J, Martínez-Gómez JE, and Burbrink FT (2019) Environmental heterogeneity and not vicariant biogeographic barriers generate community-wide population structure in desert-adapted snakes. *Molecular Ecology*, **28**, 4535–4548.

NatureServe (2021) NatureServe Explorer [web application]. NatureServe, Arlington, Virginia. Available https://explorer.natureserve.org/. (Accessed: January 10, 2021).

Noskova E, Ulyantsev V, Koepfli K-P, O'Brien SJ, and Dobrynin P (2020) GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*, **9**, giaa005.

O'Connell KA, Streicher JW, Smith EN, and Fujita MK (2017) Geographical features are the predominant driver of molecular diversification in widely distributed North American whipsnakes. *Molecular Ecology*, **26**, 5729–5751.

Ochoa A, Broe M, Moriarty Lemmon E, Lemmon AR, Rokyta DR, and Gibbs HL (2020) Drift, selection and adaptive variation in small populations of a threatened rattlesnake. *Molecular Ecology*, **29**, 2612–2625.

Owen LA, Finkel RC, Minnich RA, and Perez AE (2003) Extreme southwestern margin of late Quaternary glaciation in North America: timing and controls. *Geology*, **31**, 729–732.

Parmley D and Holman JA (2007) Earliest fossil record of a pigmy rattlesnake (Viperidae: *Sistrurus* Garman). *Journal of Herpetology*, **41**, 141–144.

Patten TJ, Fogell DD, and Fawcett JD (2016) Spatial ecology and habitat use of the western massasauga (*Sistrurus tergeminus*) in Nebraska. *Journal of North American Herpetology*, **1**, 31–38.

Pendall E, Betancourt JL, and Leavitt SW (1999) Paleoclimatic significance of δD and δ13C values in piñon pine needles from packrat middens spanning the last 40,000 years. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **147**, 53–72.

Pickrell JK and Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Pound MJ, Haywood AM, Salzmann U, and Riding JB (2012) Global vegetation dynamics and latitudinal temperature gradients during the Mid to Late Miocene (15.97–5.33 Ma). *Earth-Science Reviews*, **112**, 1–22.

QGIS Development Team (2009) QGIS Geographic Information System.

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

Rašić G, Filipović I, Weeks AR, and Hoffmann AA (2014) Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, Aedes aegypti. *BMC Genomics*, **15**, 275.

Reinert HK and Kodrich WR (1982) Movements and habitat utilization by the massasauga, Sistrurus catenatus catenatus. *Journal of Herpetology*, **16**, 162–171.

Remington CL (1968) Suture-zones of hybrid interaction between recently joined biotas. In: *Evolutionary Biology* (ed Dobzhansky T), pp. 321–428. Springer, New York, NY, USA.

Revell LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Rhymer JM and Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, and Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539–542.

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, and Sánchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, **34**, 3299–3302.

Ruane S, Bryson Jr RW, Pyron RA, and Burbrink FT (2014) Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Systematic Biology*, **63**, 231–250.

Ryberg WA, Harvey JA, Blick A, Hibbitts TJ, and Voelker G (2015) Genetic structure is inconsistent with subspecies designations in the western massasauga *Sistrurus tergeminus*. *Journal of Fish and Wildlife Management*, **6**, 350–359.

Samson FB, Knopf FL, and Ostlie WR (2004) Great Plains ecosystems: past, present, and future. *Wildlife Society Bulletin*, **32**, 6–15.

Sanz L, Gibbs HL, Mackessy SP, and Calvete JJ (2006) Venom proteomes of closely related *Sistrurus* rattlesnakes with divergent diets. *Journal of Proteome Research*, **5**, 2098–2112.

Savage JM (1960) Evolution of a peninsular herpetofauna. *Systematic Zoology*, **9**, 184–212.

Seigel RA (1986) Ecology and conservation of an endangered rattlesnake, *Sistrurus catenatus*, in Missouri, USA. *Biological Conservation*, **35**, 333–346.

Snir S and Rao S (2012) Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution*, **61**, 1–8.

Solís-Lemus C and Ané C (2016) Inferring phylogenetic networks with Maximum Pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, **12**, 1–21.

Solís-Lemus C, Bastide P, and Ané C (2017) PhyloNetworks: A package for phylogenetic networks. *Molecular Biology and Evolution*, **34**, 3292–3298.

Soltis DE, Morris AB, McLachlan JS, Manos PS, and Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Sovic MG, Fries AC, and Gibbs HL (2016) Origin of a cryptic lineage in a threatened reptile through isolation and historical hybridization. *Heredity*, **117**, 358–366.

Sovic M, Fries A, Martin SA, and Lisle Gibbs H (2019) Genetic signatures of small effective population sizes and demographic declines in an endangered rattlesnake, *Sistrurus* catenatus. *Evolutionary Applications*, **12**, 664–678.

Stebbins RC (2003) *A Field Guide to Western Reptiles and Amphibians*. Houghton Mifflin Harcourt.

Stenz NWM, Larget B, Baum DA, and Ané C (2015) Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species *Arabidopsis thaliana* (L.) heynh. *Systematic Biology*, **64**, 809–823.

Swenson NG and Howard DJ (2005) Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American Naturalist*, **166**, 581–591.

Szymanski J (1998) *Status assessment for the eastern massasauga (Sistrurus c. catenatus)*. U.S. Fish and Wildlife Service. Fort Snelling, MN.

Szymanski J, Pollack C, Ragan L, Redmer M, Clemency L, Voorhies K, and Jaka J (2016) Species status assessment for the eastern massasauga rattlesnake (*Sistrurus catenatus*). *US Fish and Wildlife Service*.

To T-H, Jung M, Lycett S, and Gascuel O (2016) Fast dating using least-squares criteria and algorithms. *Systematic Biology*, **65**, 82–97.

Trimble DE (1980) The geologic story of the Great Plains. *US Geological Survey Bulletin, 1493,* **30**, 55.

Vendrami DLJ, Telesca L, Weigand H, Weiss M, Fawcett K, Lehman K, Clark MS, Leese F, McMinn C, and Moore H (2017) RAD sequencing resolves fine-scale population structure in a benthic invertebrate: implications for understanding phenotypic plasticity. *Royal Society Open Science*, **4**, 160548.

Walker DE and Avise JC (1998) Principles of phylogeography as illustrated by freshwater and terrestrial turtles in the southeastern United States. *Annual Review of Ecology and Systematics*, **29**, 23–58.

Wastell AR and Mackessy SP (2011) Spatial ecology and factors influencing movement patterns of desert massasauga rattlesnakes (*Sistrurus catenatus edwardsii*) in southeastern Colorado. *Copeia*, **2011**, 29–37.

Wastell AR and Mackessy SP (2016) Desert Massasauga Rattlesnakes (*Sistrurus catenatus edwardsii*) in southeastern Colorado: life history, reproduction, and communal hibernation. *Journal of Herpetology*, **50**, 594–603.

Weatherhead PJ and Prior KA (1992) Preliminary observations of habitat use and movements of the eastern massasauga rattlesnake (*Sistrurus c. catenatus*). *Journal of Herpetology*, **26**, 447–452.

Wells P V (1970) Postglacial vegetational history of the Great Plains. *Science*, **167**, 1574–1582.

Werler JE and Dixon JR (2010) *Texas Snakes: Identification, Distribution, and Natural History*. University of Texas Press, Austin, TX.

Wilson JS and Pitts JP (2010) Illuminating the lack of consensus among descriptions of earth history data in the North American deserts: a resource for biologists. *Progress in Physical Geography*, **34**, 419–441.

Wood DA, Vandergast AG, Barr KR, Inman RD, Esque TC, Nussear KE, and Fisher RN (2013) Comparative phylogeography reveals deep lineages and regional evolutionary hotspots in the Mojave and Sonoran Deserts. *Diversity and Distributions*, **19**, 722–737.

Wooten JA and Gibbs HL (2012) Niche divergence and lineage diversification among closely related *Sistrurus* rattlesnakes. *Journal of Evolutionary Biology*, **25**, 317–328.

Wright BA (1941) Habit and habitat studies of the massasauga rattlesnake (*Sistrurus catenatus catenatus* Raf.) in northeastern Illinois. *American Midland Naturalist*, **25**, 659–672.

Yu G, Smith DK, Zhu H, Guan Y, and Lam TT (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, **8**, 28–36.

TABLES AND FIGURES

Table 1: PICANTE analysis assessing spatial signals along a time-calibrated phylogeny (Fig. 3).

Included Taxa	Trait	K-statistic	Variance (Obs.)	Variance (Null)	P-value	Z
Whole Tree	latitude	0.35	2684.63	26211.12	0.001	-5.56
	longitude	0.52	4900.14	104569.21	0.001	-5.90
S. catenatus +	latitude	1.67	1981.66	21012.10	0.001	-5.15
all S. tergeminus	longitude	3.51	4032.37	88579.73	0.001	-5.39
S. t. tergeminus +	latitude	1.91	2078.13	10941.77	0.001	-3.40
S. t. edwardsii	longitude	3.59	3622.37	23196.68	0.001	-3.94
S. catenatus	latitude	1.90	1704.21	12610.29	0.001	-4.26
	longitude	3.97	5598.68	66123.00	0.001	-4.33
S. t. tergeminus	latitude	1.44	2094.76	9206.51	0.001	-2.41
	longitude	2.21	1368.13	8067.78	0.001	-2.95
S. t. edwardsii	latitude	4.07	1558.06	10628.07	0.001	-2.94
	longitude	3.50	955.64	6605.08	0.001	-2.94

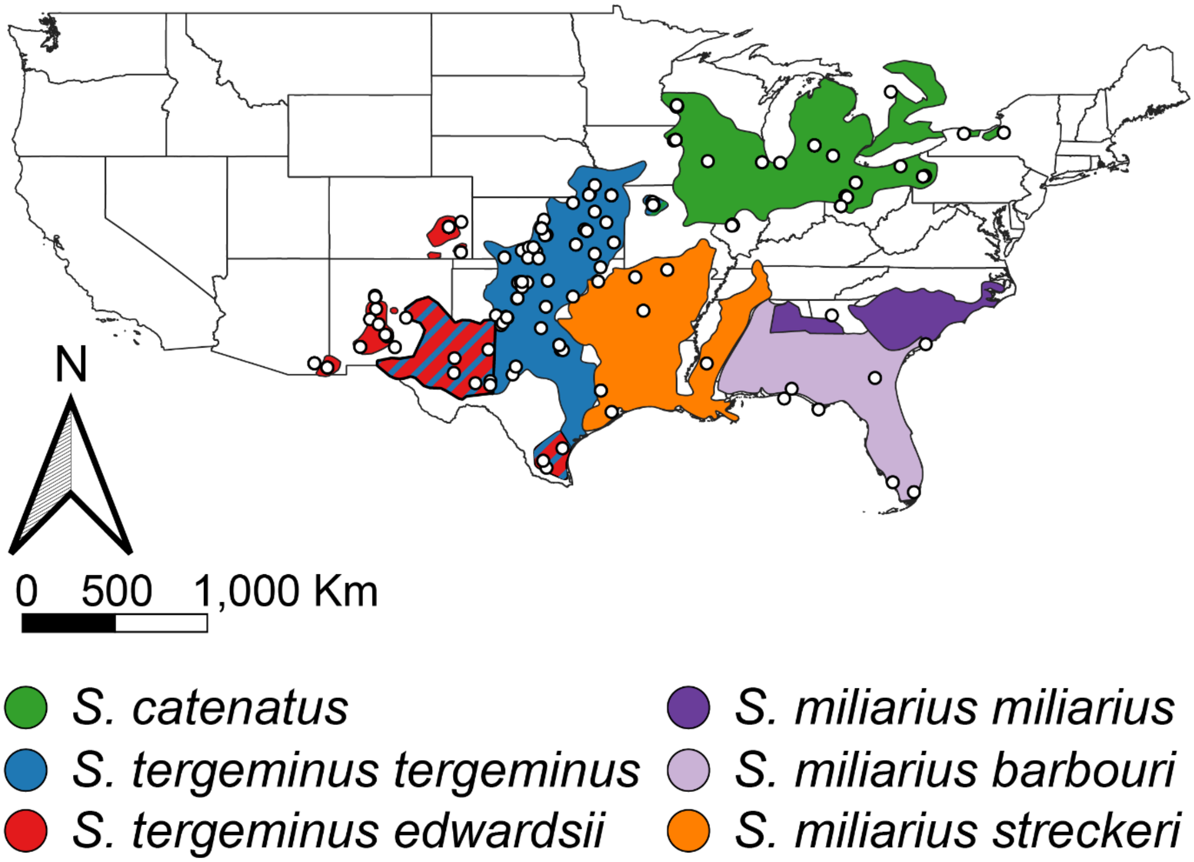


Figure 1: *Sistrurus* range map depicting subspecies distributions and ddRAD sequencing samples. Cross-hatched areas in Texas, New Mexico, and Missouri indicate overlapping ranges with known contact zones. White circles indicate samples sequenced for this study. Adapted from Sanz *et al.* (2006).

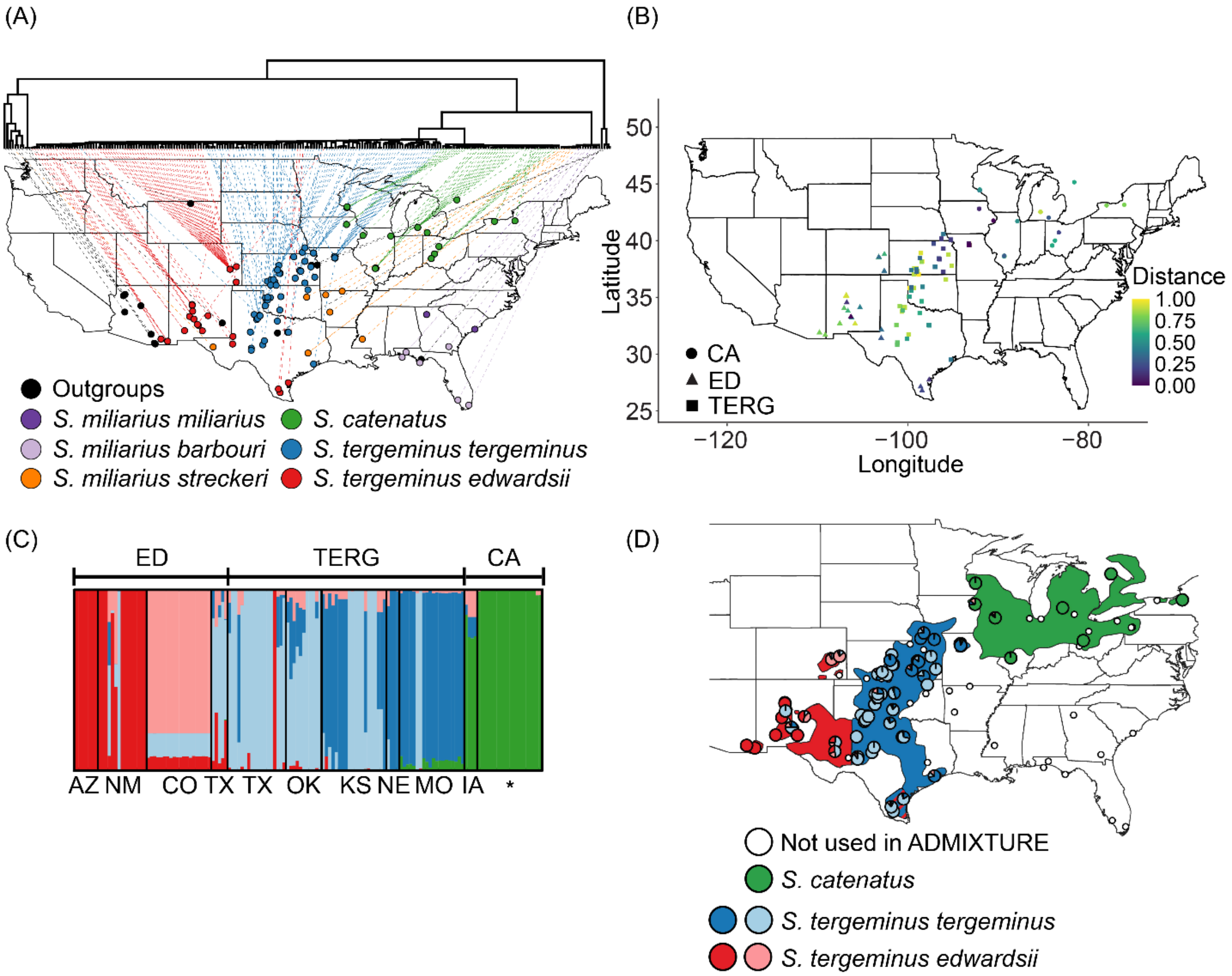


Figure 2: Phylogenetic and ADMIXTURE results for *Sistrurus*. Colors correspond to each currently recognized *Sistrurus* species and subspecies (per field identification). (A) IQ-TREE phylogeny depicting the geographic localities per tip. Outgroups included several species of *Crotalus* and *Agkistrodon*. Dashed lines link the phylogeny tips with the corresponding sample on the map. (B) PALEOTREE diversification distances, with nodal distances corresponding to the proportion of nodes between the tip and the most ancestral node per species (*S. catenatus*=CA, *S. tergeminus edwardsii*=ED, *S. tergeminus tergeminus*=TERG). (C) Barplot depicting the ADMIXTURE proportions for *K*=5, with each bar representing one individual and bars with mixed colors indicating mixed ancestry. The barplot is partitioned into sub-populations based on subspecies field identification (upper guide) and U.S. state locality (bottom guide; IA=Iowa, MO=Missouri, NE=Nebraska, KS=Kansas, OK=Oklahoma, TX=Texas, CO=Colorado, NM=New Mexico, AZ=Arizona). The asterisk for *S. catenatus* indicates that individuals were included from multiple state localities. (D) Pie charts on the map geographically display the ADMIXTURE proportions.

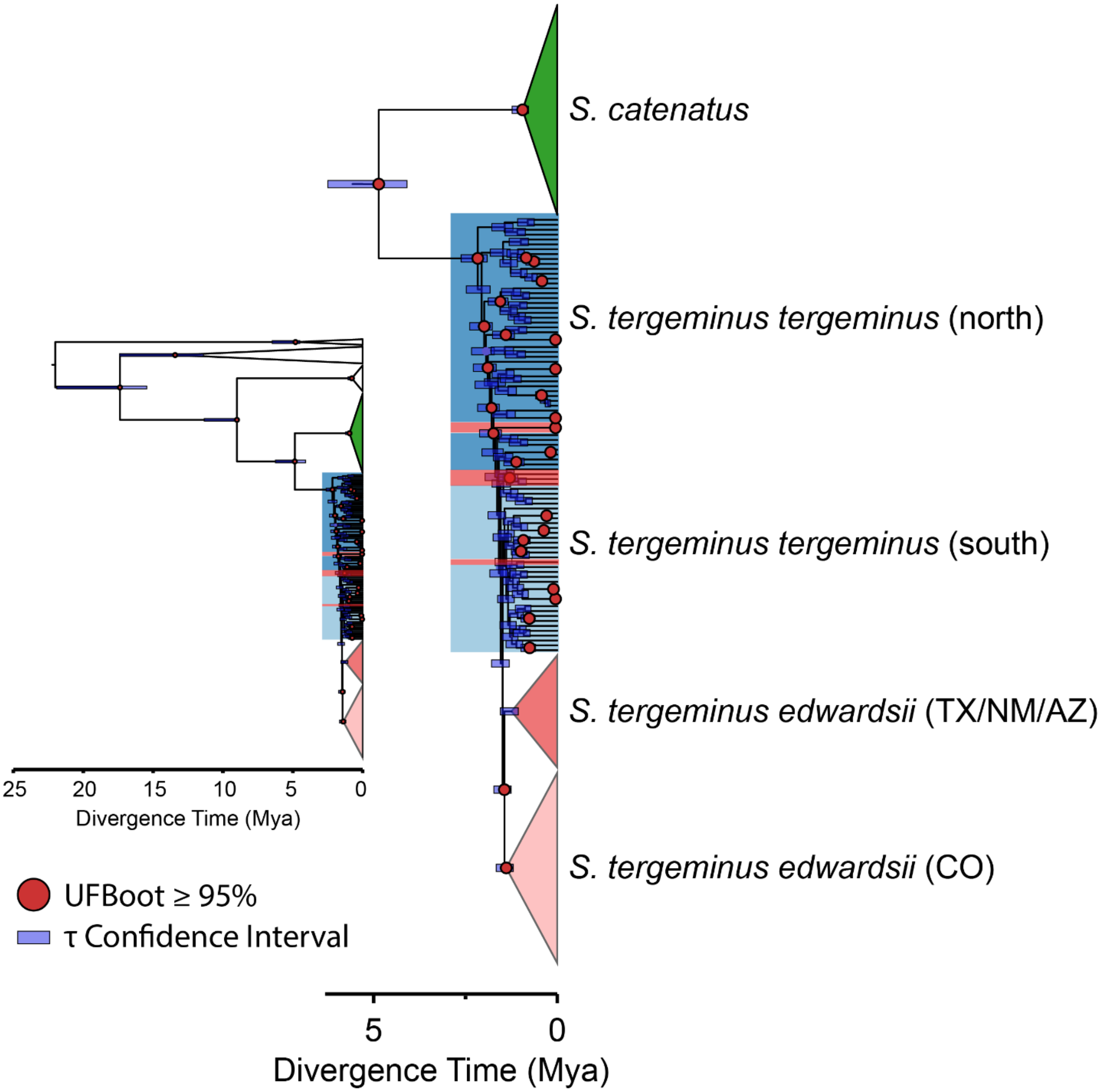


Figure 3: *Sistrurus* molecular clock, with the phylogeny inferred using the edge-linked partition model from IQ-TREE v2.1.2 and scaled to actual time (millions of years ago, Mya) using the IQ-TREE implementation of least square dating (LSD). Populations are color-coded according to the ADMIXTURE analysis in Figures 1C and 1D, with *S. tergeminus tergeminus* subdivided into northern and southern clusters and *S. t. edwardsii* represented by Texas/ New Mexico/ Arizona and Colorado populations. Red circles indicate nodes with ultrafast bootstrap (UFBoot) support ≥ 95%, and the node bars correspond to the divergence time (τ) confidence interval as calculated from 1,000 input trees with branch lengths simulated from a relaxed-clock Poisson distribution. The main plot is zoomed to *S. catenatus* and *S. tergeminus* ssp., whereas the inset plot shows the full tree including *S. miliarius* and the outgroup taxa (several species of *Agkistrodon* and *Crotalus*).

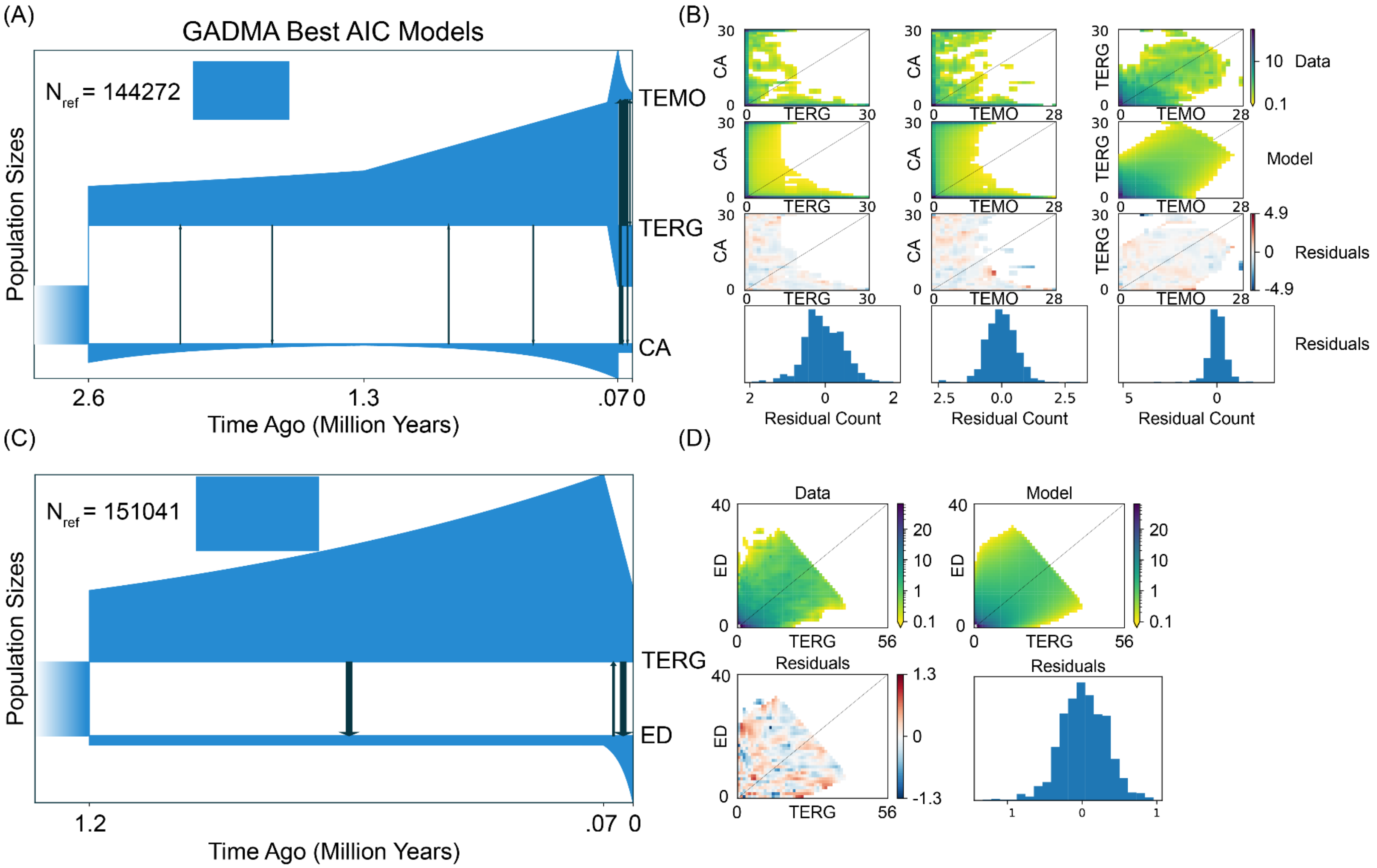


Figure 4: Demographic models chosen by GADMA. Arrows depict migration events with sizes proportional to estimated migration parameters. Branch size on the Y-axis is scaled to the reference population size (N_{ref}). (A) Optimal model for *Sistrurus catenatus* (CA) X *S. tergeminus tergeminus* from Missouri (TEMO) and from all other sample localities (TERG). (B) Allele frequency spectra for the data, model, and residuals as heatmaps and histograms. (C) Optimal model for *S. t. tergeminus* from all sample localities except Missouri (TERG) and *S. t. edwardsii* (ED). (D) Allele frequency spectra and residuals for the TERG X ED GADMA analysis.

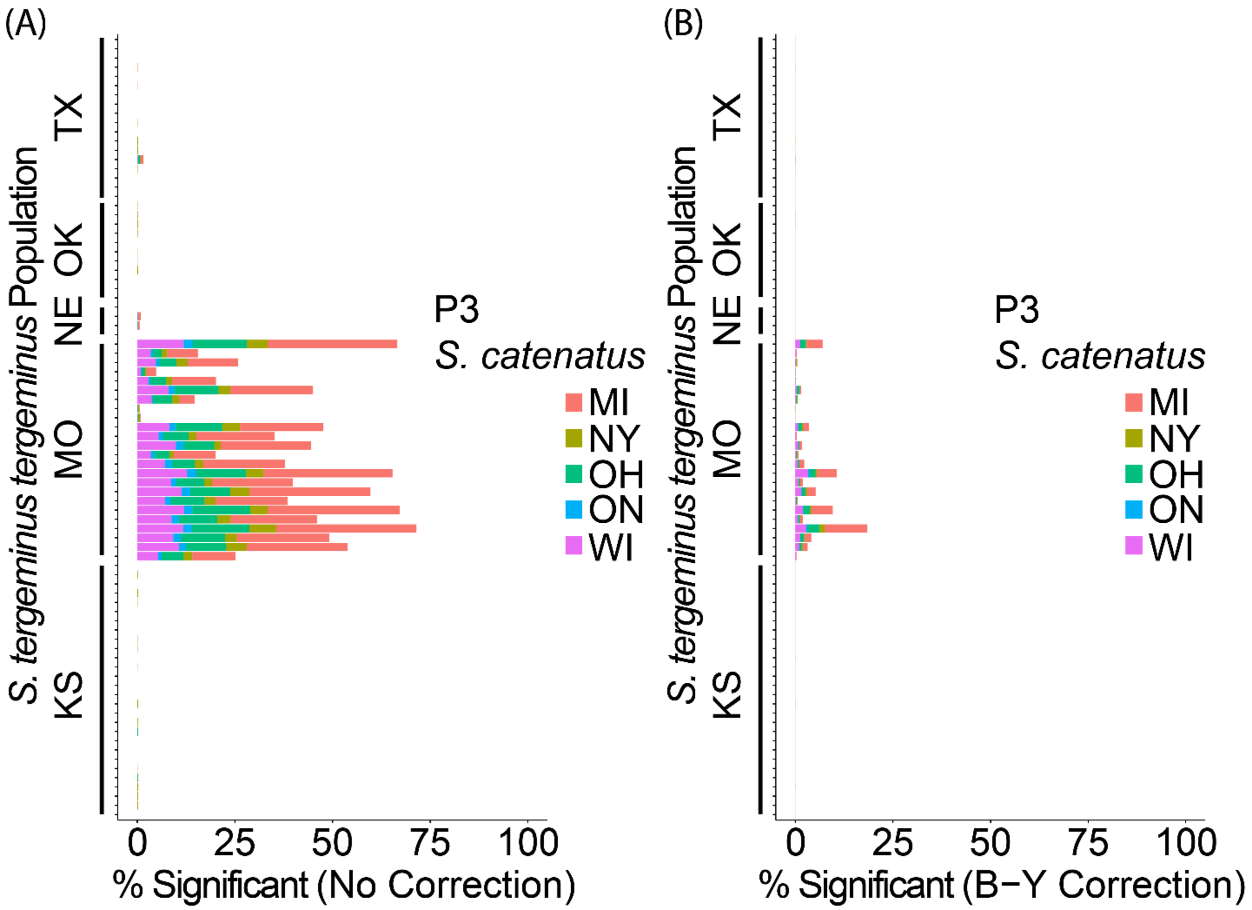


Figure 5: Four-taxon *D*-statistic tests for *S. catenatus* (P3) and *S. tergeminus tergeminus* (P1 and P2). Stacked bars indicate the percentage of significant *D*-statistic tests per P3 sample locality with (A) no *P*-value correction for multiple tests and (B) a Benjamini–Yekutieli (B-Y) correction for controlling the false discovery rate. State locality abbreviations include: TX=Texas, OK=Oklahoma, NE=Nebraska, MO=Missouri, KS=Kansas, MI=Michigan, NY=New York, OH=Ohio, WI=Wisconsin, and ON=Ontario (Canada). The outgroups (P4) included *S. miliarius streckeri* and *S. m. barbouri*.

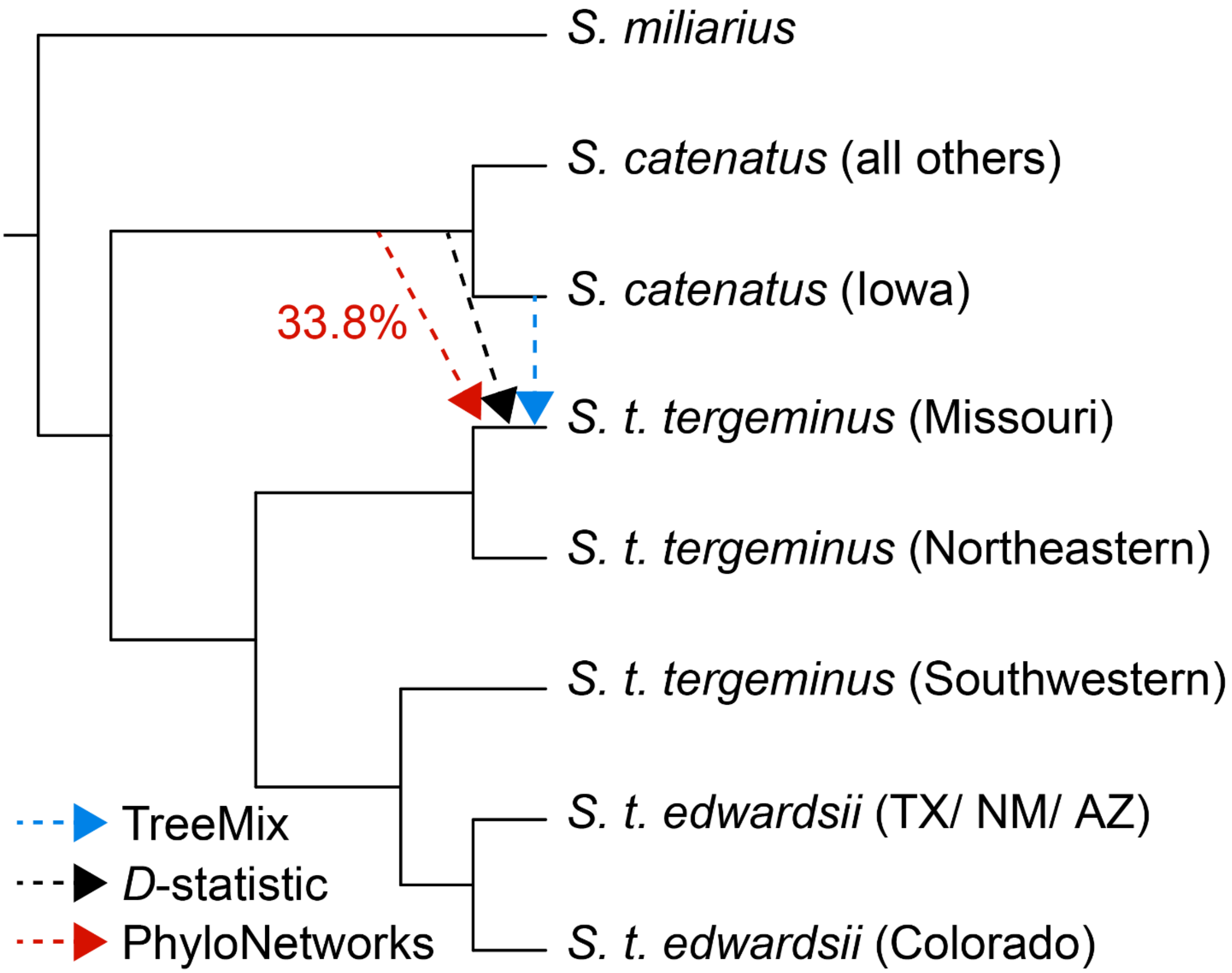

S. miliarius
S. catenatus (all others)
S. catenatus (Iowa)
33.8%
S. t. tergeminus (Missouri)
S. t. tergeminus (Northeastern)
S. t. tergeminus (Southwestern)
S. t. edwardsii (TX/ NM/ AZ)
S. t. edwardsii (Colorado)
TreeMix
D-statistic
PhyloNetworks

Figure 6: Supported *Sistrurus* admixture edges among TREEMIX, four-taxon *D*-statistic tests, and PHYLONETWORKS. The percentage corresponding to the PHYLONETWORKS arrow illustrates the estimated *S. catenatus* ancestry in Missouri *S. t. tergeminus*. Scientific names are followed by state or regional locality. Northwestern and Southwestern *S. t. tergeminus* correspond to sub-populations identified in the ADMIXTURE analysis (Fig. 2). TX=Texas, NM=New Mexico, AZ=Arizona.

SUPPLEMENTAL TABLES AND FIGURES

Table S1: Sample and collector information for the *N*=226 sequenced *Sistrurus*, *Crotalus*, and *Agkistrodon* samples.

ID	Date	Sex	Collector(s)/ Source(s)	State	County
SMAR1	X-00	F	C. Montgomery	AR	Washington
SMAR4	X-00	M	K. Irwin	AR	Montgomery
SMAR5	8-VII-00	F	K. Irwin	AR	Marion
CM2		M	A. Holycross	AZ	Coconino
CMO72	04-VII-02	M	G. Carpenter	AZ	Mohave
CS8	07-VIII-00		G. Schuett	AZ	Yavapai
CVC32	VIII-98	-	C. Meachum	AZ	Pima
CVC6		M	G. Schuett	AZ	Coconino
CWW6	VII-93		T. LaDuc	AZ	Cochise
CWW7	VIII-02		G.W. Schuett	AZ	Cochise
SCE10		F	Phx. Zoo	AZ	Cochise
SCE11		F	Phx. Zoo	AZ	Cochise
SCE12		F	Phx. Zoo	AZ	Cochise
SCE9		M	Phx. Zoo	AZ	Cochise
SICA52G			L. Gibbs; A. Holycross	AZ	Cochise
SICA53G			L. Gibbs; A. Holycross	AZ	Cochise
SICA54G			L. Gibbs; A. Holycross	AZ	Cochise
SCCO1		F	S. Mackessy	CO	Lincoln
SCCO10			S. Mackessy	CO	Lincoln
SCCO12			S. Mackessy	CO	Lincoln
SCCO13			S. Mackessy	CO	Lincoln
SCCO14			S. Mackessy	CO	Lincoln
SCCO15			S. Mackessy	CO	Lincoln
SCCO16			S. Mackessy	CO	Lincoln
SCCO17			S. Mackessy	CO	Lincoln
SCCO18			S. Mackessy	CO	Lincoln
SCCO19			S. Mackessy	CO	Lincoln
SCCO2		F	S. Mackessy	CO	Lincoln
SCCO20			S. Mackessy	CO	Lincoln
SCCO21			S. Mackessy	CO	Lincoln
SCCO22			S. Mackessy	CO	Lincoln
SCCO23			S. Mackessy	CO	Lincoln
SCCO24			S. Mackessy	CO	Lincoln
SCCO25			S. Mackessy	CO	Lincoln
SCCO26			S. Mackessy	CO	Lincoln
SCCO27			S. Mackessy	CO	Lincoln

Table S1 (Cont.)

ID	Date	Sex	Collector(s)/ Source(s)	State	County
SCCO28			S. Mackessy	CO	Lincoln
SCCO29			S. Mackessy	CO	Lincoln
SCCO3		M	S. Mackessy	CO	Lincoln
SCCO35			S. Mackessy	CO	Lincoln
SCCO37			S. Mackessy	CO	Lincoln
SCCO38			S. Mackessy	CO	Lincoln
SCCO39			S. Mackessy	CO	Lincoln
SCCO4			S. Mackessy	CO	Lincoln
SCCO41			S. Mackessy	CO	Lincoln
SCCO42			S. Mackessy	CO	Lincoln
SCCO43			S. Mackessy	CO	Lincoln
SCCO44			S. Mackessy	CO	Lincoln
SCCO45			S. Mackessy	CO	Lincoln
SCCO47			S. Mackessy	CO	
SCCO48			S. Mackessy	CO	Baca
SCCO49			S. Mackessy	CO	Baca
SCCO5			S. Mackessy	CO	Lincoln
SCCO50			S. Mackessy	CO	Cheyenne
SCCO6			S. Mackessy	CO	Lincoln
SCCO9			S. Mackessy	CO	Lincoln
APFL1			W. Hayes	FL	Liberty
SMFL12	21-VIII-03		P. Moler	FL	Wakulla
SMFL13	21-VIII-03		P. Moler	FL	Okaloosa
SMFL5			S. Conners	FL	Collier
SMFL6	23-XI-01		S. Conners	FL	Miami-Dade
SMFL8	27-XII-02		G. Pyron	FL	Franklin
SMGA1	16-IV-03		G. Schuett	GA	Cherokee
SMGA2	15-VI-03	M	C. Ponder	GA	Ware
SCIA3	20-IV-04	F	T.VanDeWalle	IA	Bremer
SCIA4	3-V-04	M	T.VanDeWalle	IA	Bremer
SCIA5	3-V-04	F	T.VanDeWalle	IA	Bremer
SCIA7	3-V-04	F	T.VanDeWalle	IA	Bremer
SCIA8	VIII-90		J. Christiansen	IA	Scott
SCIL1		-	C. Phillips	IL	Clinton
SCIL2		-	C. Phillips	IL	Clinton
SCIL3		-	C. Phillips	IL	Clinton
SCIL4	08-VI-01	M	T. Anton	IL	Cook
SCIL6		F		IL	Carlisle
SCIL7		F		IL	Carlisle
SCIL8		F		IL	Carlisle
ACKS3	VI-00		H. Alamillo	KS	Johnson

Table S1 (Cont.)

ID	Date	Sex	Collector(s)/ Source(s)	State	County
ACKS5	10-V-00		C. Shiel	KS	Douglas
SCKS1		F	S. Mackessy	KS	Barton
SCKS11	19-IX-04		T. Taggart	KS	Clark
SCKS12	20-IX-04		T. Taggart	KS	Comanche
SCKS13	20-IX-04		C.J. Schmidt	KS	Kiowa
SCKS14				KS	Chase
SCKS15				KS	Chase
SCKS16				KS	Barber
SCKS17				KS	Chase
SCKS18				KS	Kiowa
SCKS19				KS	Comanche
SCKS2		M	S. Mackessy	KS	Barton
SCKS20				KS	Allen
SCKS21				KS	Barber
SCKS22				KS	Reno
SCKS23				KS	Russell
SCKS24				KS	Kiowa
SCKS25				KS	Douglas
SCKS26				KS	Stafford
SCKS3		M	S. Mackessy	KS	Barton
SCKS4		M	S. Mackessy	KS	Barton
SCKS5		M	S. Mackessy	KS	Barton
SCKS6	28-IV-01		D. Fogell	KS	Pottawatomie
SCKS7	VI-02	F	D. Fogell	KS	Chase
SCKS8			D. Shepard	KS	Butler
SCKS9	23-VI-04		J. Voelkler	KS	Washington
SICA18G	25-IV-09		R. Brown, KU, P. Ingram	KS	Chase
SICA21G			B. coyner SNOMNH	KS	Butler
SICA27G	15-X-04		C.J. Schmidt SMNH, M. Washburne	KS	Chautauque
E058			J. Moore	MI	
E114			J. Moore	MI	
E270			J. Moore	MI	
E294			J. Moore	MI	
E301			J. Moore	MI	
E356			J. Moore	MI	
E513			J. Moore	MI	
E545			J. Moore	MI	
E577			J. Moore	MI	
E635			J. Moore	MI	
E794			J. Moore	MI	
E819			J. Moore	MI	

Table S1 (Cont.)

ID	Date	Sex	Collector(s)/ Source(s)	State	County
SCMI2	VI-01		K. Schuett	MI	Hillsdale
SICA46G			J. Moore	MI	
SCMO1	26-IX-01	F	F. Durbian	MO	Holt
SCMO11	VI-02	F		MO	Chariton
SCMO12	VI-02	M		MO	Chariton
SCMO13	VI-02			MO	Chariton
SCMO14	02-IV-05	F	R.Seigel/T.Crabill	MO	Linn
SCMO15	02-IV-05	M	R.Seigel/T.Crabill	MO	Linn
SCMO16	02-IV-05	F	R.Seigel/T.Crabill	MO	Linn
SCMO17	03-IV-05	M	R.Seigel/T.Crabill	MO	Linn
SCMO18	03-IV-05	M	R.Seigel/T.Crabill	MO	Linn
SCMO19	03-IV-05	F	R.Seigel/T.Crabill	MO	Linn
SCMO2	27-IX-01	M	F. Durbian	MO	Holt
SCMO20	04-IV-05	F	R.Seigel/T.Crabill	MO	Linn
SCMO21	04-IV-05	M	R.Seigel/T.Crabill	MO	Linn
SCMO22	04-IV-05		R.Seigel/T.Crabill	MO	Linn
SCMO23	04-IV-05		R.Seigel/T.Crabill	MO	Linn
SCMO24				MO	Chariton
SCMO25				MO	Linn
SCMO3	27-IX-01	M	F. Durbian	MO	Holt
SCMO4	27-IX-01	M	F. Durbian	MO	Holt
SCMO5	27-IX-01	F	F. Durbian	MO	Holt
SCMO6	27-IX-01	M	F. Durbian	MO	Holt
SCMO7	02-X-01	F	F. Durbian	MO	Holt
SCMO8	02-X-01	F	F. Durbian	MO	Holt
SCMO9	VI-02	J		MO	Chariton
SMMS1	VI-01	F	VanDevender	MS	Copiah
SCNE1	VIII-00	F	D. Fogell	NE	Pawnee
SCNE2	VIII-00	M	D. Fogell	NE	Pawnee
SCNE3	VIII-00	F	D. Fogell	NE	Pawnee
SCNE4	V-99		M. Ingrasci	NE	Russell
CVV12	09-IV-00	F	C. Painter	NM	Chaves
SCE1	22-VII-00	J	D.&K.Salceies	NM	Bernalillo
SCE2	14-VII-00	J	D.&K.Salceies	NM	Bernalillo
SCE3	7-VII-00	J	D.&K.Salceies	NM	Bernalillo
SCE4	3-VII-00	J	D.&K.Salceies	NM	Bernalillo
SCE5	22-VFII-00	J	D.&K.Salceies	NM	Bernalillo
SCE6	4-VII-00	J	D.&K.Salceies	NM	Bernalillo
SCE7	14-VII-00	J	D.&K.Salceies	NM	Bernalillo
SCE8	IX-00	M	C. Painter	NM	Socorro
SCNM1	29-VI-00	M	T. LaDuc	NM	Valencia

Table S1 (Cont.)

ID	Date	Sex	Collector(s)/ Source(s)	State	County
SCNM2	06-X-96		T.R. Jones	NM	Lincoln
SCNM3	18-V-98		L. Kamees	NM	Socorro
SCNM4			B. Christman	NM	Otero
SCNM5	VII-01		B. Christman	NM	Otero
SCNM6			B. Mackin	NM	Grant
SCNM7			B. Mackin	NM	Grant
SICA67G			Elda Sanchez NNTRC	NM	Otero
SCNY1	09-VI-92		G. Johnson	NY	Monroe
SCNY2	09-VI-92		G. Johnson	NY	Monroe
SCNY3	4-V-92		G. Johnson	NY	Onandaga
SCNY4	18-VI-92	F	G. Johnson	NY	Onandaga
SCOH1	25-VIII-00	-	M. Spille	OH	Greene
SCOH11			M. Spille	OH	Greene
SCOH12			M. Spille	OH	Greene
SCOH16	29-V-92	F	G. Johnson	OH	Clark
SCOH17	29-V-92	F	G. Johnson	OH	Clark
SCOH3	25-VIII-00	-	M. Spille	OH	Greene
SCOH8	06-IX-00	-	D. Wynn	OH	Wyandot
SCOK08			B. coyner SNOMNH	OK	Roger Mills
SCOK1	26VIII05	F	D.Shepard/L.Vitt	OK	Beckham
SCOK10				OK	Dewey
SCOK11				OK	Beckham
SCOK12				OK	Rogers
SCOK2	26VIII05	J	D.Shepard/L.Vitt	OK	Beckham
SCOK3			D.Shepard/L.Vitt	OK	Roger Mills
SCOK4			D.Shepard/L.Vitt	OK	Roger Mills
SCOK5			D.Shepard/L.Vitt	OK	Ellis
SCOK6			D.Shepard/L.Vitt	OK	Blaine
SICA65G			Elda Sanchez NNTRC	OK	Comanche
SMOK3	7-VII-05		C. Whitney	OK	Tulsa
SICA57G			L. Gibbs	ON	Dorcas Bay
SMSC1	15-XI-00	F	B. Starrett	SC	Charleston
APTX1			W. Hayes	TX	Eastland
CSTX6	2003		B. Mackin	TX	Culberson
CSTX6	2003		B. Mackin	TX	Culberson
SCT1	9-V-00	-	K. McCoy	TX	Concho
SCT2	24-VI-00	M	T. Hibbitts	TX	Parker
SCT3	24-VI-00	M	T. Hibbitts	TX	Hood
SCTX1		M	A. Price	TX	Cottle
SCTX10				TX	Cottle
SCTX11				TX	Cottle

Table S1 (Cont.)

ID	Date	Sex	Collector(s)/ Source(s)	State	County
SCTX14				TX	Runnels
SCTX15				TX	Motley
SCTX16				TX	Dickens
SCTX17				TX	Borden
SCTX18				TX	Cottle
SCTX2	29-V-01		T. LaDuc	TX	Andrews
SCTX3	06-V-01		T. LaDuc	TX	Crockett
SCTX4	26-V-01		T. LaDuc	TX	Motley
SCTX5	26-VI-01		T. LaDuc	TX	Crockett
SCTX8	114-IX-99	J	Tamu-K	TX	Galveston
SCTX9			GladysPorterZoo	TX	Starr
SICA3G	9-IX-14		Matador WMA, S. Hein, M. Barazowski	TX	Cottle
SICA48G			Matador WMA, S. Hein, M. Barazowski	TX	Cottle
SICA61G	8-VI-15		R. Couvillian	TX	Jim Hogg
SICA62G	26-VI-15		S. Hein and S. Pitts	TX	Ward
SICA66G			Elda Sanchez NNTRC	TX	Nueces
SICA70G			T. Hibbitts	TX	Archer
SMTX1	31-VII-06		T. Sinclair	TX	Montgomery
SMTX2	31-VII-06		T. Sinclair	TX	Montgomery
SCWI10	10-X-01	M	E. McCumber	WI	Buffalo
SCWI2	01-VIII-00	F	via ATH	WI	Buffalo
SCWI6	30-VIII-01	F	E. McCumber	WI	Buffalo
SCWI7	02-IX-01	F	E. McCumber	WI	Buffalo
SCWI8	02-IX-01	M	E. McCumber	WI	Buffalo
SCWI9	12-VII-01	M	E. McCumber	WI	Buffalo
CVV56			W. Hayes	WY	Sheridan

Table S2: Number of sequenced individuals per *Sistrurus* taxon.

Taxonomic ID	**Count**
S. catenatus	44
S. tergeminus tergeminus	85
S. tergeminus edwardsii	68
S. miliarius miliarius	2
S. miliarius barbouri	6
S. miliarius streckeri	7
Total	212

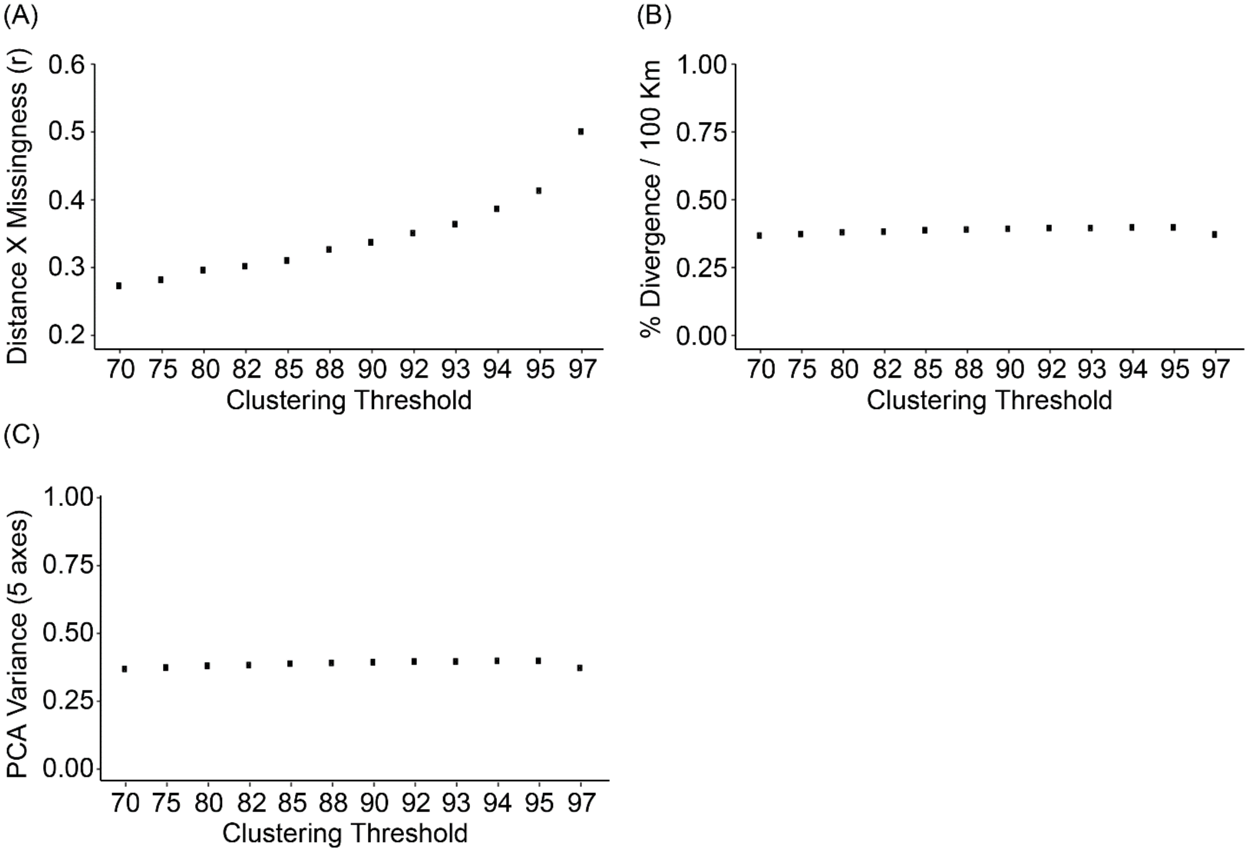


Figure S1: Assessments of varying IPYRAD clustering threshold parameters on *Sistrurus* ddRAD data. (A) Pearson's correlation coefficient (r) between genetic distance and the percentage of missing data in the alignment. (B) Percent sequence divergence per 100 kilometers (Km) as a measure of isolation by distance; (C) principal component analysis (PCA) variance across five principal component axes.

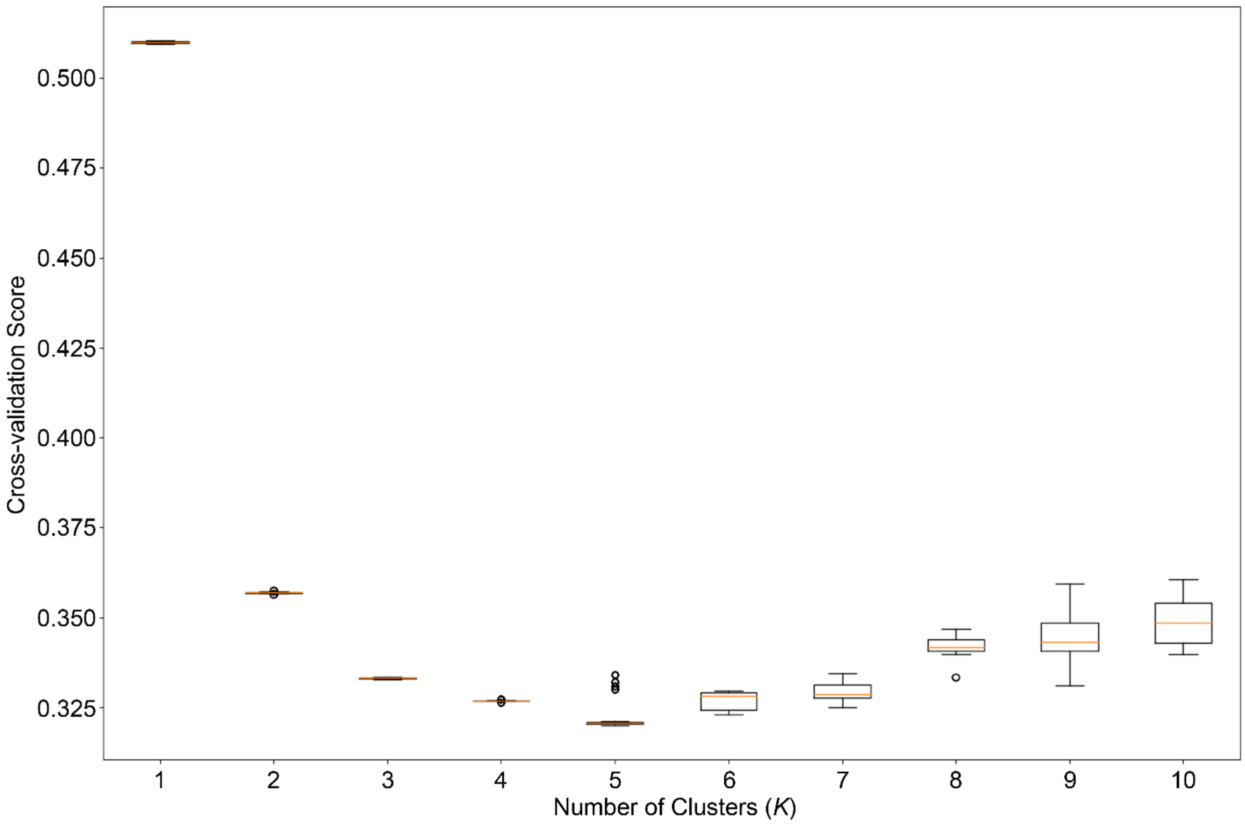


Figure S2: Boxplots for cross-validation scores from an ADMIXTURE analysis between *Sistrurus catenatus*, *S. tergeminus tergeminus*, and *S. t. edwardsii*. ADMIXTURE was run with 20 replicates per *K* and 20-fold cross-validation.

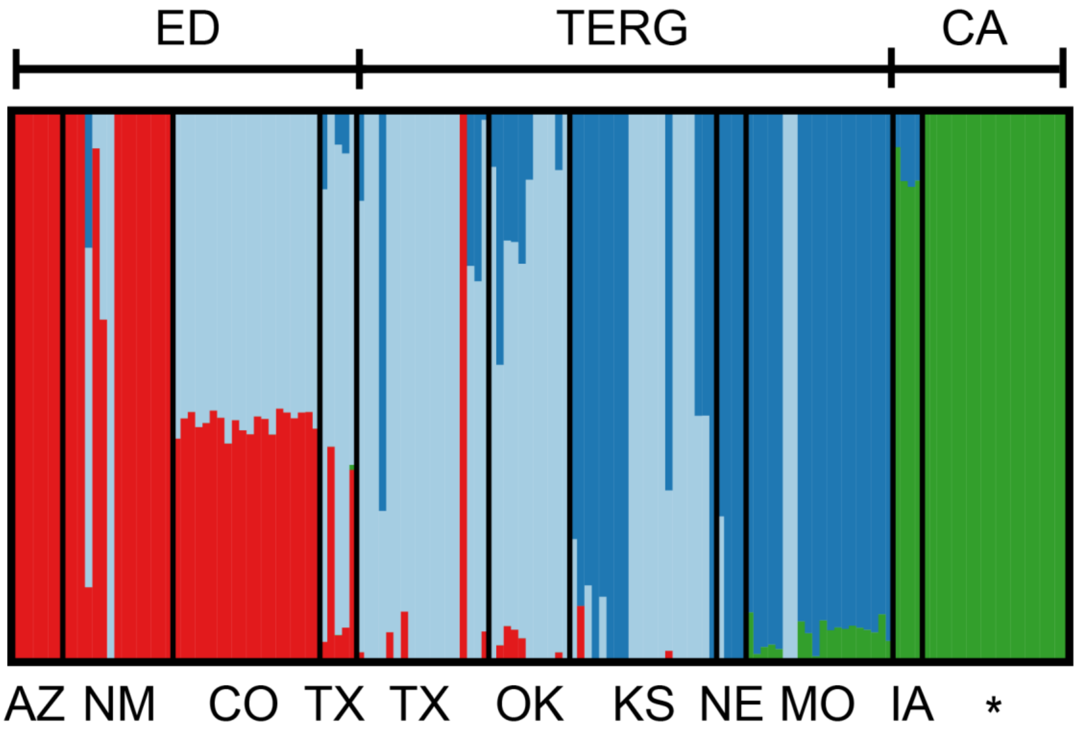


Figure S3: Admixture barplot depicting ADMIXTURE proportions among *Sistrurus* populations. Each bar represents one individual and bars with mixed colors indicating mixed ancestry. The top guide indicates subspecies designation as identified in the field: ED=*S. tergeminus edwardsii*, TERG=*S. tergeminus tergeminus*, CA=*S. catenatus*. The barplot is further partitioned into sub-populations based on U.S. state locality (IA=Iowa, MO=Missouri, NE=Nebraska, KS=Kansas, OK=Oklahoma, TX=Texas, CO=Colorado, NM=New Mexico, AZ=Arizona). The asterisk for *S. catenatus* indicates that individuals were included from multiple state localities.

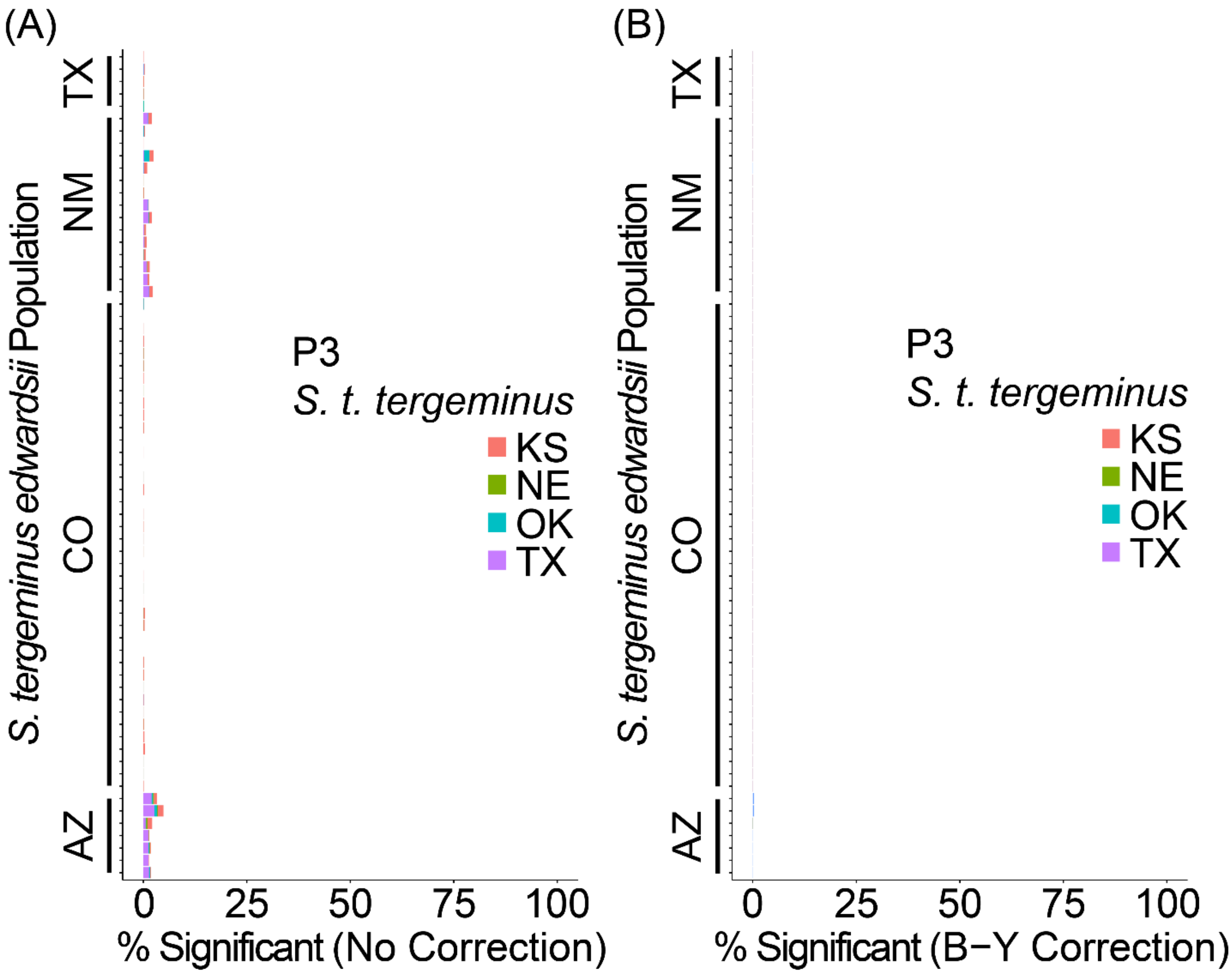


Figure S4: Four-taxon *D*-statistic tests for *Sistrurus tergeminus tergeminus* (P3) and *S. tergeminus edwardsii* (P1 and P2). Stacked bars indicate the percentage of significant *D*-statistic tests per P3 sample locality with (A) no *P*-value correction for multiple tests and (B) a Benjamini–Yekutieli (B-Y) correction for controlling the false discovery rate. State locality abbreviations include: TX=Texas, NM=New Mexico, CO=Colorado, AZ=Arizona, KS=Kansas, NE=Nebraska, OK=New Oklahoma. The outgroups (P4) included *S. miliarius streckeri* and *S. m. barbouri*.

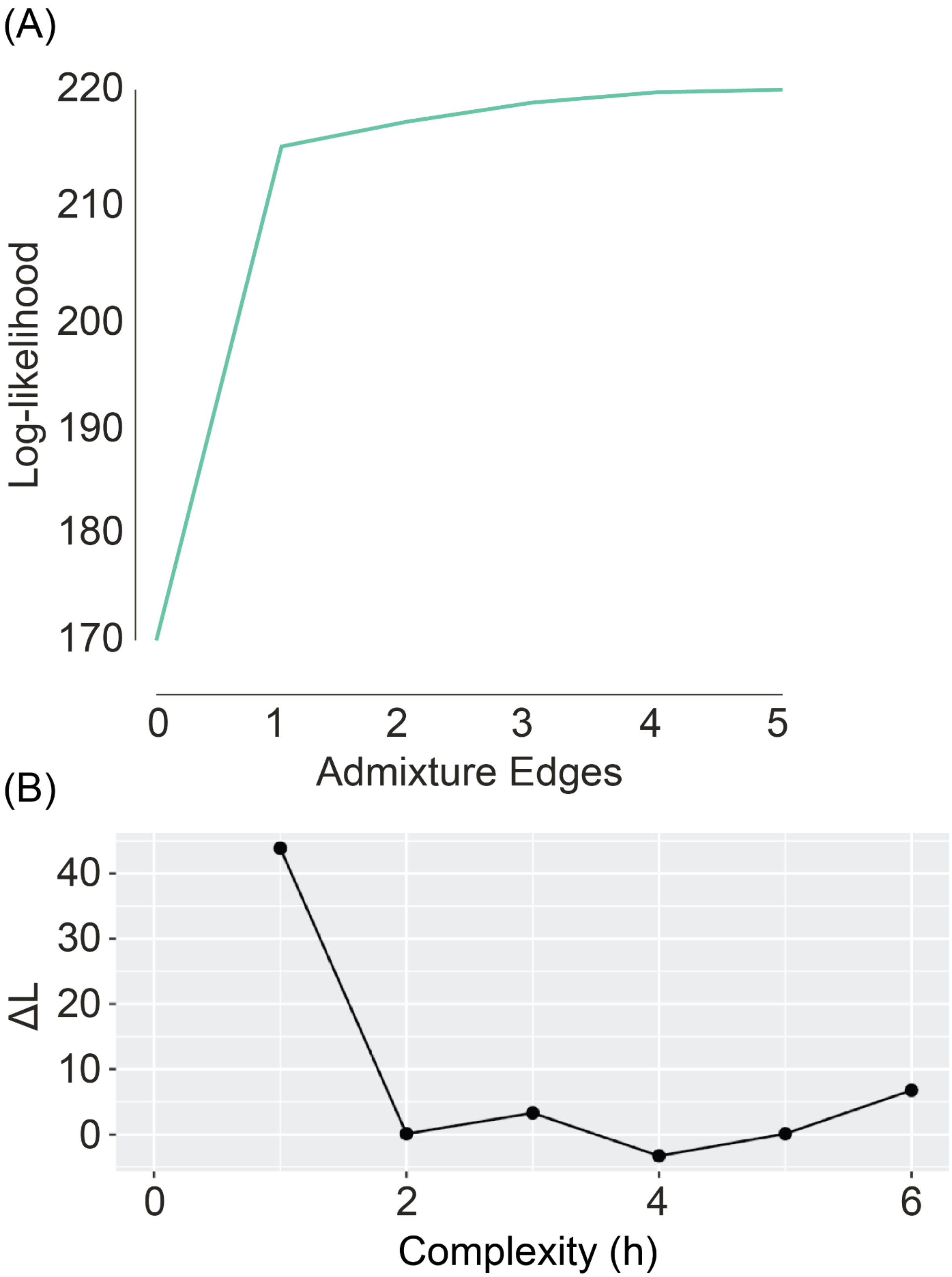


Figure S5: (A) TREEMIX log-likelihoods for 1-10 migration edges. (B) PHYLONETWORKS first-order change in pseudo-likelihood (ΔL) for 0-6 hybrid nodes (h).

CONCLUSIONS

Herein, I have contextualized the genome-wide impacts of historical and contemporary hybridization in several North American regions across two reptilian groups of conservation concern. The formation and maintenance of these regional hybrid zones were demonstrated to involve a variety of evolutionary processes, including secondary contact of previously allopatric lineages, primary divergence pertaining to ecological speciation, and selective processes occurring differentially with respect to species boundaries. They also provide insight into the current and future status of species from conservation management, taxonomic, and evolutionary standpoints. Below, I summarize my findings for each hybrid zone and their associated study species.

Secondary contact in historical and contemporary hybrid zones

Southeastern United States

Species in the southeastern United States have been subjected to several glacially-induced historical and contemporary processes. First, the glacial expansion pushed distributions southward, thereby forcing secondary contact of previously isolated taxa (Hewitt 1996, 2000). Taken together with the abundance of phylogeographic breaks and the juxtaposition of multiple habitat types in the unglaciated Southeast (Walker & Avise 1998; Avise 2000; Soltis *et al.* 2006), the region clearly involves a complex array of evolutionary processes that extend into contemporary times. The Woodland (*Terrapene carolina carolina*/ *T. c. major*) and Three-toed (*T. mexicana triunguis*) Box Turtles, for example, have introgressed on anthropogenic timescales, with secondary contact being a promoter for the formation of this hybrid zone.

Genome-wide investigation also revealed that while selection against hybrids was evident across inter-specific boundaries (as supported by species delimitation analyses; Chapter III), intra-specific introgression may be adaptive in nature (Chapter I). This disparity was most prominent at functional loci potentially governed by thermal adaptation, indicating that introgression patterns in these ectothermic turtles may exhibit marked susceptibility to oncoming climate change.

Midwestern region

On the other hand, introgression between the Eastern (*Sistrurus catenatus*) and and Prairie (*S. tergeminus tergeminus*) Massasaugas occurred on historical timescales in highly disjunct Missouri and Iowa populations. Given the fossil record (Holman 2000; Parmley & Holman 2007), *Sistrurus* likely originated in nearby Kansas with subsequent diversification extending northeastward across a 'Prairie Corridor' (Cook 1993). Importantly, this corridor was likely controlled by glacial-interglacial cycles that modulated Mississippi River discharge, with secondary contact occurring among Missouri and Iowa individuals that are presently isolated from the larger contiguous *S. t. tergeminus* range. Indeed, the Missouri individuals, at least, may manifest local adaptation as a consequence of this historical introgression with intermediate morphologies and inhabitance of habitat graded between the western and eastern sides of the Mississippi River (Evans & Gloyd 1948). Thus, despite their potentially 'hybrid' status, they may represent an evolutionarily significant or management units that represent unique genetic diversity. As with *S. catenatus* (Sovic *et al.* 2019), the Missouri *S. t. tergeminus* population may have undergone bottlenecks, with current levels of adaptive variation perhaps being overestimated due to a time lag with genetic drift (Ochoa *et al.* 2020). However, *S. t. tergeminus* are not currently listed as of conservation concern in Missouri, a status that may warrant reconsideration following further investigation into whether local adaptation exists.

Ecological refugia in the southwest

As with *S. catenatus*, *S. t. tergeminus* also appears to have undergone a southwestward diversification concomitant with the Mississippi River as a dispersal barrier. During an interglacial period high discharge volumes of the Mississippi River may have prompted *S. tergeminus* to disperse into the southwestern Great Plains in search of new habitat. The region largely lacks physical influence from glaciers, but has been subject to climate change effected by glacial-interglacial cycles (Axelrod 1948, 1979), which may have induced ecological divergence between *S. t. tergeminus* and the Desert Massasauga (*S. t. edwardsii*). The data presented herein indicate that these two taxa are currently undergoing primary divergence, which I postulate was maintained by persistent gene flow occurring between individuals inhabiting lowland arid and highland mesic refugia. A contemporary example may include a *S. t. tergeminus* and *S. t. edwardsii* contact zone in western Texas and eastern New Mexico where the habitat grades from tallgrass to shortgrass prairies. Essentially, *S. t. edwardsii* and *S. t. tergeminus* may represent taxa in the process of diverging, and the markedly differing habitat preferences of *S. t. edwardsii* should preclude dissolution of their subspecies status or at least warrant their consideration as an evolutionarily significant or management unit.

Overall conclusions

I have presented data describing the historical and contemporary effects of hybridization and introgression on a variety of evolutionary, biogeographic, and ecological processes. The data involve two ectothermic genera that are susceptible to impending climate change and anthropogenic effects. However, these genera each inhabit regions where numerous other co-distributed species also hybridize (Remington 1968; Swenson & Howard 2005), and thus the data herein may serve as a proxy for other taxa that are also of conservation concern. For example, the ecological divergence observed between *S. t. tergeminus* and *S. t. edwardsii* is hardly unique, with other co-distributed species often following similar patterns (Douglas *et al.* 2006). Likewise, in the southeastern United States other ectotherms are subject to hybrid zones coinciding with similar phylogeographic breaks, habitat juxtapositions, and ecological processes (Walker & Avise 1998; Soltis *et al.* 2006; Rissler & Smith 2010). Finally, the Great Plains region has been strongly influenced by the Mississippi River as a physiographic barrier (Braun 1950; Soltis *et al.* 2006), which may have left relictual locally adapted populations following intermittent periods of secondary contact and isolation. On a continental scale, hybrid zones are clearly important for creating and sustaining adaptive variation across and within species. Thus the underlying evolutionary, physiographic, and ecological processes should be strongly considered to facilitate successful conservation management strategies.

REFERENCES

Avise JC (2000) *Phylogeography: the history and formation of species*. Harvard University Press, Cambridge, MA, USA.

Axelrod DI (1948) Climate and evolution in western North America during middle Pliocene time. *Evolution*, **2**, 127–144.

Axelrod DI (1979) Age and origin of Sonoran Desert vegetation. *Occasional Papers of the California Academy of Sciences*, **132**, 1–74.

Braun EL (1950) Deciduous Forests of Eastern North America. *Soil Science*, **71**, 155.

Cook FR (1993) After an Ice Age: Zoogeography of the Massasauga within a Canadian Herpetofaunal Perspective. In: *Rattlesnake Symposium*, pp. 19–25. Metro Toronto Zoo, Toronto, CA.

Douglas ME, Douglas MR, Schuett GW, and Porras LW (2006) Evolution of rattlesnakes (Viperidae; *Crotalus*) in the warm deserts of western North America shaped by Neogene vicariance and Quaternary climate change. *Molecular Ecology*, **15**, 3353–3374.

Evans PD and Gloyd HK (1948) The subspecies of the massasauga, *Sistrurus catenatus*, in Missouri. *Bulletin of the Chicago Academy of Sciences*, **8**, 225–232.

Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnaean Society*, **58**, 247–276.

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Holman JA (2000) *Fossil snakes of North America: origin, evolution, distribution, paleoecology*. Indiana University Press, Bloomington, IN, USA.

Ochoa A, Broe M, Moriarty Lemmon E, Lemmon AR, Rokyta DR, and Gibbs HL (2020) Drift, selection and adaptive variation in small populations of a threatened rattlesnake. *Molecular Ecology*, **29**, 2612–2625.

Parmley D and Holman JA (2007) Earliest fossil record of a pigmy rattlesnake (Viperidae: *Sistrurus* Garman). *Journal of Herpetology*, **41**, 141–144.

Remington CL (1968) Suture-zones of hybrid interaction between recently joined biotas. In: *Evolutionary Biology* (ed Dobzhansky T), pp. 321–428. Springer, New York, NY, USA.

Rissler LJ and Smith WH (2010) Mapping amphibian contact zones and phylogeographical break hotspots across the United States. *Molecular Ecology,* **19**, 5404–5416.

Soltis DE, Morris AB, McLachlan JS, Manos PS, and Soltis PS (2006) Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology*, **15**, 4261–4293.

Sovic M, Fries A, Martin SA, and Lisle Gibbs H (2019) Genetic signatures of small effective population sizes and demographic declines in an endangered rattlesnake, *Sistrurus* catenatus. *Evolutionary Applications*, **12**, 664–678.

Swenson NG and Howard DJ (2005) Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. *The American Naturalist*, **166**, 581–591.

Walker DE and Avise JC (1998) Principles of phylogeography as illustrated by freshwater and terrestrial turtles in the southeastern United States. *Annual Review of Ecology and Systematics*, **29**, 23–58.