Progress Report

Hybridization and Introgression in North American Box Turtles from the Southeastern United States

Submitted to:

John D. Groves

North American Box Turtle Conservation Committee

Principle Investigators:

Bradley T. Martin

Dr. Marlis R. Douglas

Dr. Michael E. Douglas

Arkansas Conservation and Molecular Ecology Laboratory

University of Arkansas

Department of Biological Sciences

850 W Dickson St.

Fayetteville, AR 72701

Respectfully submitted:

12 April 2019

Table of Contents

Overview

We performed a region-wide assessment of hybridization and introgression in the North American box turtles (*Terrapene*) using thousands of genome-wide single nucleotide polymorphisms (SNPs). The geographic region of interest, the southeastern United States, contains four distinct *Terrapene* taxa (*T. carolina carolina*, *T. c. major*, *T. c. bauri*, and *T. mexicana triunguis*) that have been known to hybridize (Milstead 1969; Dodd 2001). Herein we evaluated admixture, population structure, and estimates of hybrid generation (i.e., F1, F2, and backcross) within a hybrid zone. The presence of hybridization and introgression and the distribution of hybrid generations can have important conservation management-related consequences.

Study objectives

(1) **Detect admixture and population structure** between the four *Terrapene* taxa inhabiting the southeastern United States

(2) **Assess the distribution of hybrid generations in the region** to determine whether hybrids are primarily early-generation (F1) or if introgressive hybridization is occurring (F2 and backcross generations)

(3) Correlate landscape variables with the spatial occurrence of hybrids

Approach

We utilized a reduced-representation next-generation sequencing (NGS) approach to screen thousands of genome-wide SNPs, which were used to perform an assessment of population structure and admixture via a model-based maximum likelihood framework. We also used a similar Bayesian framework to assign hybrids to specific generations by comparing observed genotypic frequencies with that expected for F1, F2, and backcross generation hybrids.

This report includes accomplishments to date but does not include the planned landscape genomic analyses because they are still ongoing. We expect to complete the landscape genomic analyses in Summer 2019.

Step 1: Next-generation sequencing of remaining available tissues (~144 individuals).

When we submitted our proposal to the Box Turtle Conservation Committee we had sequenced ~220 individuals using reduced-representation next-generation sequencing. **We have now extracted genomic DNA from and sequenced an additional ~144 individuals.** The individuals were subjected to genomic library preparation according to the standard double digest restriction-associated DNA sequencing (ddRADseq) protocol (Peterson *et al.* 2012) and using the restriction enzymes *PstI* and *MspI*. The genomic libraries were sequenced on an Illumina Hi-Seq 4000 DNA sequencer at a core facility (U of Oregon, Genomics and Cell Characterization Core Facility).

Step 2: Sequence alignment and post-alignment filtering

Raw reads returned from the core facility were subjected to the ipyrad 0.7.28 pipeline (Eaton 2014). Ipyrad clusters and aligns the reads to create a consensus sequence for each sample, and includes various quality control filters to reduce or eliminate the impact of paralogs, low-quality reads, and missing data. The initial alignment contained 437 individuals. We further filtered the alignment by applying a 1.0% minor allele frequency filter and removing individuals with >90% missing data via a custom Perl script. Finally, we thinned the alignment to retain only one SNP per ddRAD locus to reduce the effect of linkage bias. The final filtered alignment contained 394 individuals from both within and outside the southeastern United States and 12,128 SNPs from independent ddRAD loci.

Step 3: Assessing population structure and admixture

We ran the ADMIXTURE software package (Alexander *et al.* 2009) to assess population structure and admixture for all 394 individuals. Twenty independent runs were conducted for each K-value (i.e., the number of *a priori* assigned clusters), which ranged from K=1 to K=20. Cross-validation (CV), which folds portions of the dataset as missing data and re-analyzes it to assess error, was then performed to select the most appropriate K (i.e., the K with the lowest CV score). We present the top three K-values per the three lowest CV scores (i.e., lowest error; Figure 1).

We then regionally partitioned the dataset into southeastern United States samples only because model-based methods such as ADMIXTURE often depict the uppermost hierarchical structure, and sub-structure can remain undetected (Evanno *et al.* 2005). However, some known pure individuals were retained to maintain appropriate sample sizes for the defined populations. A minor allele frequency filter of 1.0% was re-applied to the reduced-individual dataset, from which 11,142 distinct loci were analyzed. The ADMIXTURE analysis was repeated as above for K=1 to K=13 and is presented with the three lowest CV scores (Figure 2).

ADMIXTURE results

The Admixture analysis containing all the sequenced individuals (Figure 1) indicated five discrete populations at the best supported K: The outgroups (*Clemmys guttata* and *Emydoidea blandingii*), *T. ornata*, *T. c. carolina*, *T. c. major*, and *T. m. triunguis*. *T. c. bauri* was not well

resolved, likely due to its limited sample size (N=4). However, K=6 split *T. c. major* into two subgroups from Florida and Mississippi. These subgroups were explored in greater detail by subsetting the southeastern taxa and re-running ADMIXTURE.

Contrarily to the full dataset, the reduced southeastern dataset's (Figure 2) best supported analysis contained **four clusters** (K=4) that depicted **two distinct *T. c. major* subgroups** (FL and MS), similarly to K=6 in Figure 1. Interestingly, the Floridian *T. c. major* primarily displayed admixture with *T. c. carolina* whereas the Mississippians admixed with *T. m. triunguis*, with the two pure *T. c. major* populations being **highly coincident with the Alabama and Apalachicola River deltas**. The *T. c. carolina* populations from South Carolina and Georgia also contained admixture with *T. m. triunguis*. Finally, K=7 suggested *T. c. carolina* from South Carolina to be a distinct cluster, though it was not apparently biologically or geographically meaningful.

Step 4: Classifying hybrid generations

The hybrids were classified to a hybrid generation using the Bayesian model-based assignment test implemented in NEWHYBRIDS. The NEWHYBRIDS software works similarly to other assignment tests such as STRUCTURE and ADMIXTURE (Pritchard *et al.* 2000; Alexander *et al.* 2009), but assigns individuals to expected genotypic frequency classes reflective of parental groups (P1 and P2), and first (F1), second (F2), and backcross (B1 and B2; i.e., F1's backcrossing with parentals) generation hybrids.

We also ran a power analysis using HYBRIDDETECTIVE to validate the NEWHYBRIDS results (Wringe *et al.* 2017a), which was parallelized across multiple CPU cores using PARALLELNEWHYBRID (Wringe *et al.* 2017b). Both packages were loaded into and executed using R 3.5.1 (R Core Team 2018). Additionally, assignments were only reported if the posterior probability exceeded 80%, and known pure individuals were used to train the dataset.

NEWHYBRIDS primarily assigned hybrids to the backcross and F2 genotype frequency classes, respectively, with backcrosses representing the majority of cases (Figure 3). Mississippi displayed the highest frequency of both backcrossed and F2 hybrids, with the backcrosses being hybrid *T. c. major* X *T. m. triunguis* with parental *T. m. triunguis*. However, backcrosses with parental *T. c. major* were also apparent in these populations, albeit with low frequencies. Also of note were backcrosses between Floridian *T. c. carolina* X *T. c. major* hybrids with parental *T. c. major*, plus a low frequency of F2's. Finally, *T. c. carolina* X *T. m.*

triunguis backcrossed from both directions in Georgia, and *T. c. carolina* X *T. m. triunguis* backcrossed with parental *T. c. carolina* in South Carolina. Power analysis supported high accuracy (>90%) for most of the genotype classes with regards to the capability of classifying individuals.

Step 5: Landscape genomic analyses

The landscape genomics analyses are underway, but have not yet been completed. The expected completion date is by the end of Summer 2019. However, preliminary analyses suggest the presence of a latitudinal cline between *T. c. carolina* X *T. c. major*. Other geographic clines for *T. c. carolina* X *T. m. triunguis* and *T. c. major* X *T. m. triunguis* are possible, though our available samples within these geographic clines is more spatially limited than for *T. c. carolina* X *T. c. major*.

Conclusions

Introgressive hybridization between *T. c. carolina, T. c. major,* and *T. m. triunguis* are apparent in both the ADMIXTURE and NEWHYBRIDS results. The introgression also appears to be **localized to certain geographic areas**, including Mississippi, the Florida Panhandle, western Georgia, and the southern portion tip of South Carolina, though additional landscape genomic analyses are underway. Importantly, **two distinct *T. c. major* populations** may exist in the Mississippi and Florida panhandles, of which pure individuals appear to be coincident with the Alabama and Apalachicola river basins. Though these results conflict with previous morphological analyses (Butler *et al.* 2011), it is possible that these two *T. c. major* populations represent **cryptic genetic variation**. In such a case, they would certainly warrant consideration of unique conservation management strategies.

Furthermore**, all hybrids were of later generations (backcrosses or F2's)**, with no F1 hybrids being detected in the three southeastern *Terrapene* taxa that were analyzed. The high frequency of backcrosses and F2's and corresponding paucity of early-generation hybrids suggest either a degree of **reproductive isolation between the three taxa or selection against hybrids within the hybrid zone** (Scordato *et al.* 2017). In the case of the former, reproductive isolation can be maintained by a select few loci despite introgression occurring porously across other parts of the genome. Furthermore, the introgression that is occurring may be adaptive in nature, which can have evolutionary consequences and thus needs to be explored with further genomic analyses. Finally, the latter case of selection against hybrids would suggest that the

hybrid zone is maintained by a balance between the dispersal of parental types into the hybrid zone and selection against hybrids (i.e., **a tension zone**) (Barton & Hewitt 1985).

The presence and distribution of hybrids, both geographically and with regard to the identified genotypic classes, can potentially be informative towards conservation efforts. Further geographic and landscape genomic analyses may illuminate ecological and environmental factors contributing to the observed introgression patterns. In the meantime, we have demonstrated introgression in southeastern *Terrapene* that may be inhabitants of a tension zone. Given that tension zones can move spatially with environmental conditions (Barton & Hewitt 1985), the expected changes due to oncoming climate change will potentially play a role in shaping both the evolutionary trajectories of *Terrapene* in the region and the boundaries of the hybrid zone.

Tables and Figures

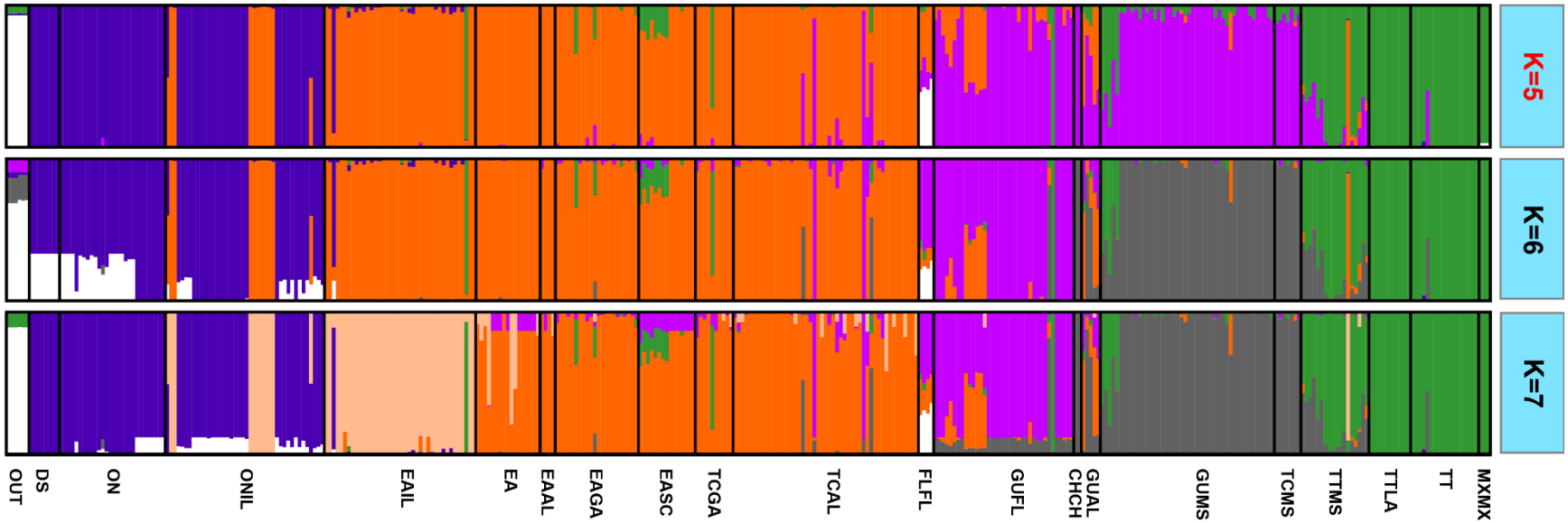

Figure 1: ADMIXTURE plot for K=5, K=6, and K=7 representing 12,128 unlinked SNPs across all sampled populations. The cross-validation score was lowest for K=5 ($\bar{x}$=0.18065, SD=0.00217; depicted in red), followed by K=6 ($\bar{x}$ =0.18119, SD=0.00333, and then K=7 ($\bar{x}$ =0.18321, SD=0.00301). Each bar represents a unique individual, and bars with mixed colors represent mixed ancestry. The first two letters of the population codes correspond to subspecific field identification (OUT=outgroups, *Clemmys guttata* and *Emydoidea blandingii*; DS=Desert box turtle, *T. o. luteola*; ON=Ornate, *T. o. ornata*; EA=Eastern, *T. c. carolina*; FL=Florida, *T. c. bauri*; GU=Gulf Coast, *T. c. major*; TT=Three-toed, *T. m. triunguis*; MX=Mexican, *T. m. mexicana*; TC=*Terrapene carolina*, with subspecies unidentified in the field). The second two letters (if present) represent locality codes by U.S. or Mexican state (IL=Illinois; AL=Alabama; GA=Georgia; SC=South Carolina; FL=Florida; CH=Coahuila, Mexico; MS=Mississippi; LA=Louisiana; MX=Tamaulipas, Mexico). Populations lacking the state locality code consisted of multiple localities sampled outside the hybrid zone.

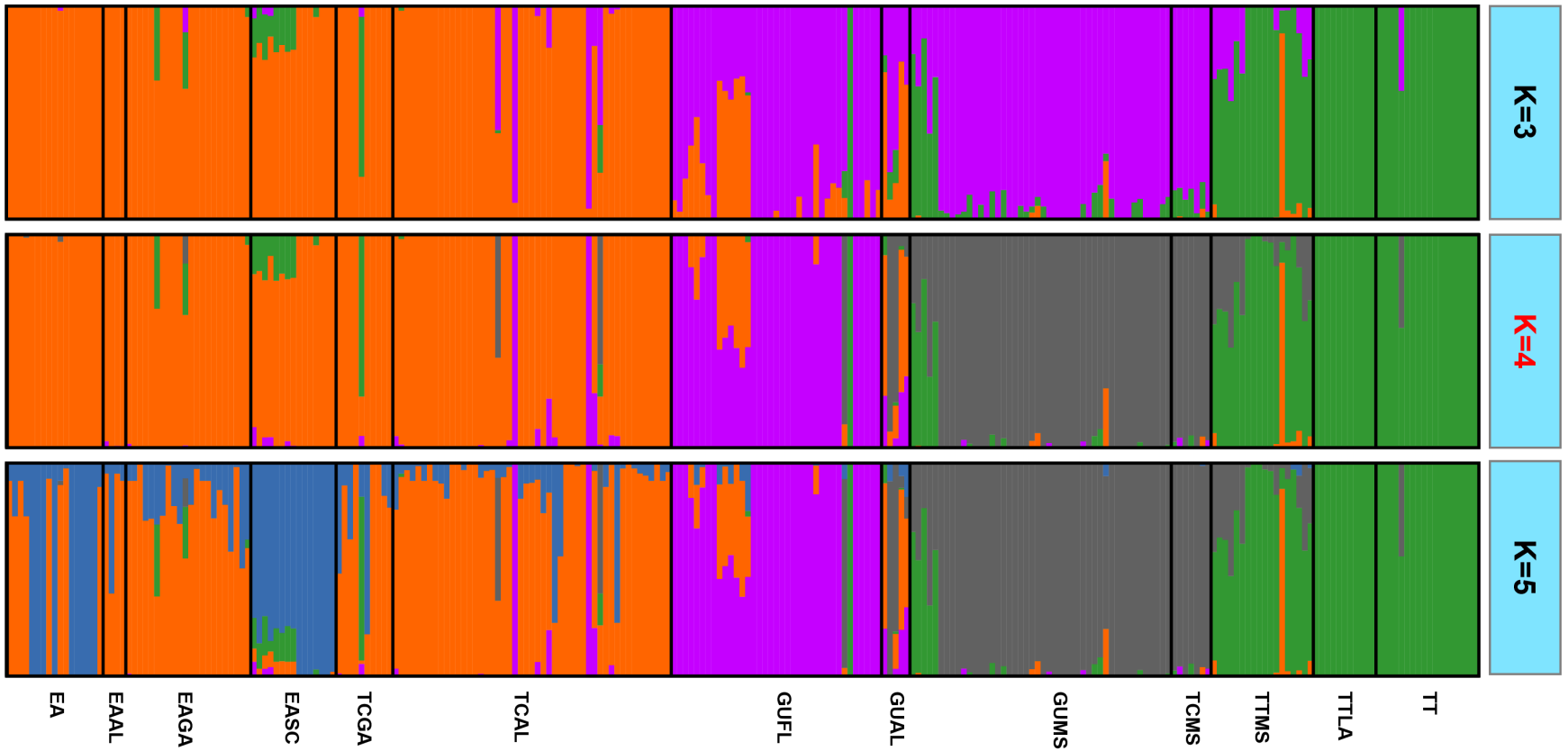

Figure 2: ADMIXTURE plot for K=3, K=4, and K=5 representing 11,142 unlinked SNPs across southeastern taxa. The cross-validation score was lowest for K=4 ($\bar{x}$= 0.21851, SD=0.00016; depicted in red), followed by K=3 ($\bar{x}$=0.22134, SD=0.00015), and then K=5 ($\bar{x}$= 0.22519, SD=0.00082). Each bar represents a unique individual, and bars with mixed colors depict mixed ancestry. The first two letters of the population codes correspond to subspecific field identification (EA=Eastern, *T. c. carolina* GU=Gulf Coast, *T. c. major*; TT=Three-toed, *T. m. triunguis*; TC=*Terrapene carolina*, with subspecies unidentified in the field). The second two letters (if present) represent locality codes by U.S. state (AL=Alabama; GA=Georgia; SC=South Carolina; FL=Florida; MS=Mississippi; LA=Louisiana). Populations lacking the state locality code consisted of multiple localities sampled outside the hybrid zone.

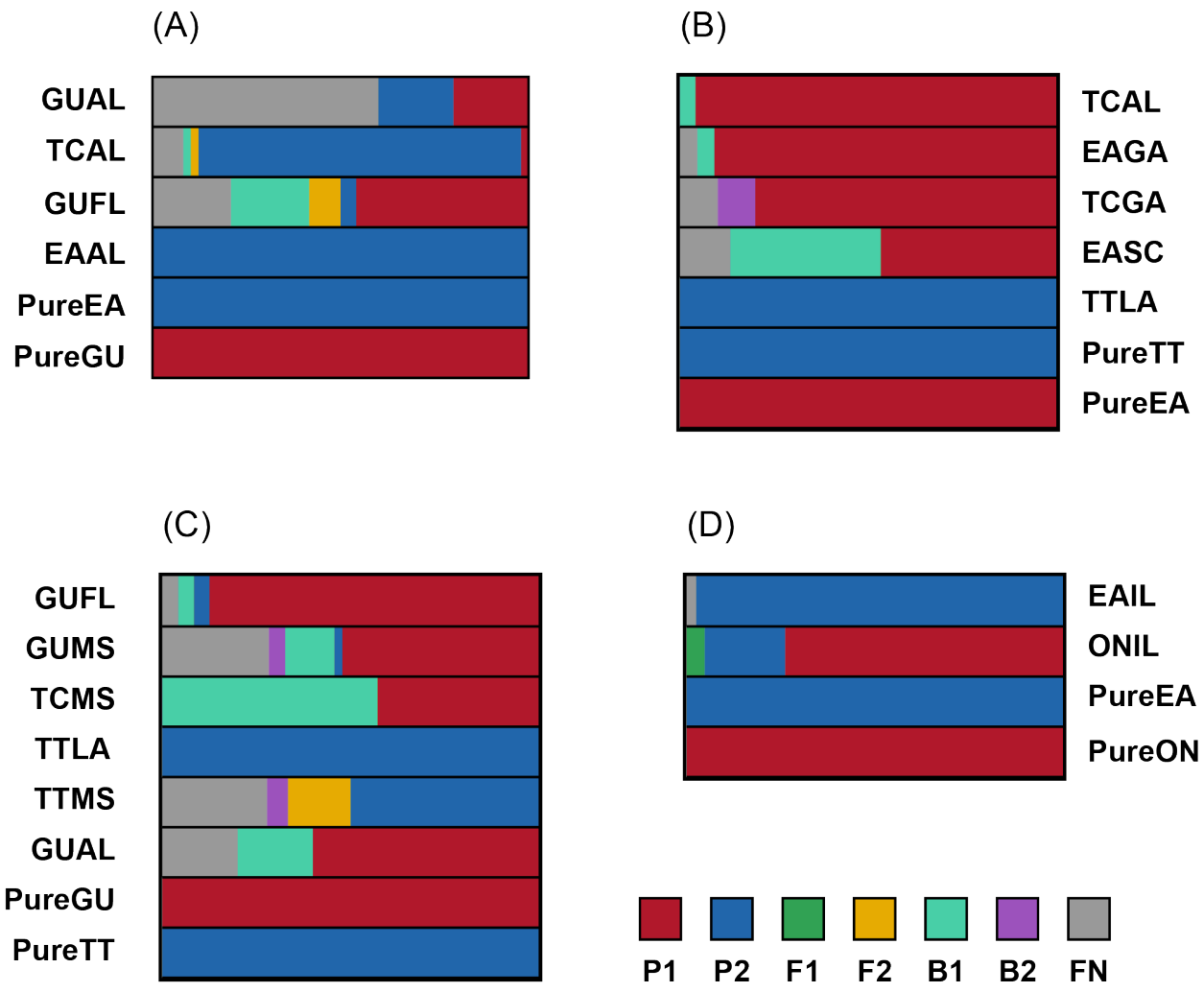

Figure 3: Population-level NEWHYBRIDS plots for four pairs of southeastern and midwestern *Terrapene* taxa. Individuals were collapsed into populations based on field identification at the subspecific level (first two characters; GU=*T. c. major*, EA = *T. c. carolina*, TT = *T. m. triunguis*, ON=*T. o. ornata*, TC=*T. carolina* only identified to species-level) and by U.S. state (last two characters; AL=Alabama, FL=Florida, MS=Mississippi, SC=South Carolina, GA=Georgia, LA=Louisiana, and IL=Illinois). Each plot corresponds to tests between parental groups (A) EA X GU (N=109), (B) EA X TT (N=135), (C) GU X TT (N=139), (D) EA X ON (N=112). A posterior probability threshold of at least 80% was required for genotype frequency class assignments. The genotype classes included P1 and P2 (parental types), F1 and F2 (first and second-generation hybrids), backcrosses (B1 and B2), and FN (unclassified).

Literature Cited

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

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

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.

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

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

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.

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.

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.

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

Scordato ESC, Wilkins MR, Semenov G, Rubtsov AS, Kane NC, and Safran RJ (2017) Genomic variation across two barn swallow hybrid zones reveals traits associated with divergence in sympatry and allopatry. *Molecular Ecology*, **26**, 5676–5691.

Team RC (2018) R: A Language and Environment for Statistical Computing.

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.