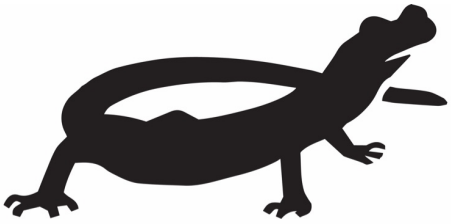
Identifying Conservation Units and Building A Northeast Specific Spatially Explicit Genetic Database For The Eastern Box Turtle (*Terrapene carolina carolina*)

An interim report submitted to the Mid Atlantic Center for Herpetology and Conservation

August 2022

Report prepared by:
Alex Krohn Ph.D.
and
JJ Apodaca Ph.D.
JJ@TBConservation.org
828-544-0581

Tangled Bank Conservation
192 E Chestnut St, Suite A
Asheville, NC 28801



TANGLED BANK

CONSERVATION™

Objective: To identify conservation units and build a genetic database that can assist in identifying the geographic origin of Eastern Box Turtles that have been confiscated.

Work completed to date: We successfully sequenced 580 million reads from 125 Eastern Box Turtles from WV, VA, DE, NJ, NY, RI, CT and MA using a 3RAD approach (Bayona-Vásquez et al. 2019). The quality of samples varied significantly, so we removed any samples that had fewer than 1 million raw reads, or fewer than 2,000 loci successfully sequenced. That left 83 individuals, including individuals from each of the above states. For these 83 individuals, we aligned reads to the Three-Toed Box Turtle reference genome (GenBank Accession number GCA_002925995.2), aligned raw reads the genome, assembled reads into RAD loci, and called SNPs using ipyrad (Eaton and Overcast 2020). The final dataset contained 1,425,877 SNPs, although not all of those SNPs were shared among all individuals.

We ran three preliminary analyses to quantify population structure. First, because previous work had shown a lack of population structure across the Eastern US, but a strong pattern of Isolation by Distance, we tested for Isolation by Distance (Wright 1943; Kimble et al. 2014). We included SNPs present in at least 50% of individuals, then randomly selected one SNP per RAD locus to remove the effect of linkage disequilibrium. The final dataset contained 83 individuals and 2,527 unlinked SNPs. We used the program SNPRelate (Zheng et al. 2012) to calculate the proportion of genetic differences (Nei 1987) between each individual, then plotted this pairwise genetic difference against pairwise geographic distance for each individual. We used a partial Mantel test to see if pairwise genetic distance increased linearly with pairwise geographic distance, as one would expect

under Isolation by Distance. Despite sampling over 1000 km, we did not find a significant effect of Isolation by Distance in this dataset (Figure 1; Mantel r = -0.085, P = 0.99, linear R^2 = 0.006).

Second, we decomposed all of the genetic variation into axes that contained the most variation in the dataset using a Principal Component Analysis (PCA). The PCA used all 83 individuals and 2,397 unlinked SNPs. The overall PCA showed WV, VA, and New England as distinct clusters (Figure 2). Within New England, RI surprisingly occupied more PC space than other states, indicating that it may harbor higher genetic diversity, or be genetically distinct. We ran a second PCA with just individuals from New England. This PCA contained 48 individuals and 48,826 SNPs. Here, states separate into distinct clusters, except for NJ, NY and MA, which remain a single cluster (Figure 3). This analysis shows that when using enough genetic markers, one can distinguish genetic groups among individuals from different states. Moreover, when colored by river drainage, it is apparent that river drainages generally cluster together, indicating that individuals can be identified at least to river drainage of origin.

Third, we ran a Bayesian clustering analysis to assign individuals a percentage of admixture from the number of populations that best fit the data. We used the program fastSTRUCTURE (Raj et al. 2014) to delimit populations, both with the logistical prior for fine-scale population structure, and without, testing the fit of models with K = 1 populations to K = 10. We used the same dataset as for the IBD analysis. fastSTRUCTURE found that K = 1 best described the data, regardless of prior. We re-ran the analysis with just the New England subset (n = 48, 1,377 SNPs), and found that K = 1 also best described the data. This lack of structure was also found with mitochondrial data (Kimble et al. 2014).

fastSTRUCTURE analyses thus indicate that the clustering observed in the PCA is not the best explanation for the data. Alternatively, because PCA imputes for missing data, it thus has a higher tolerance for missing data and can run on more SNPs. It's possible that increasing the number of SNPs shared among individuals may increase our ability to detect subtle patterns of population structure.

Future Directions

Our current database is lacking in two main areas. First, it does not include samples across the range of the Eastern Box Turtle. By increasing our sampling across the range, we will better capture the range of genetic variation across the Eastern Box Turtle and be better able to geolocate confiscated samples. As of August 2022, we have successfully sequenced an additional 50 Box Turtles, although those data are not analyzed yet. Second, it is clear that a higher number of SNPs allows better resolution in visualizing very subtle population structure. Whenever possible, we will increase our sequencing depth, and increase sample extraction quality, in order to capture more SNPs shared across the range of Box Turtles sampled.

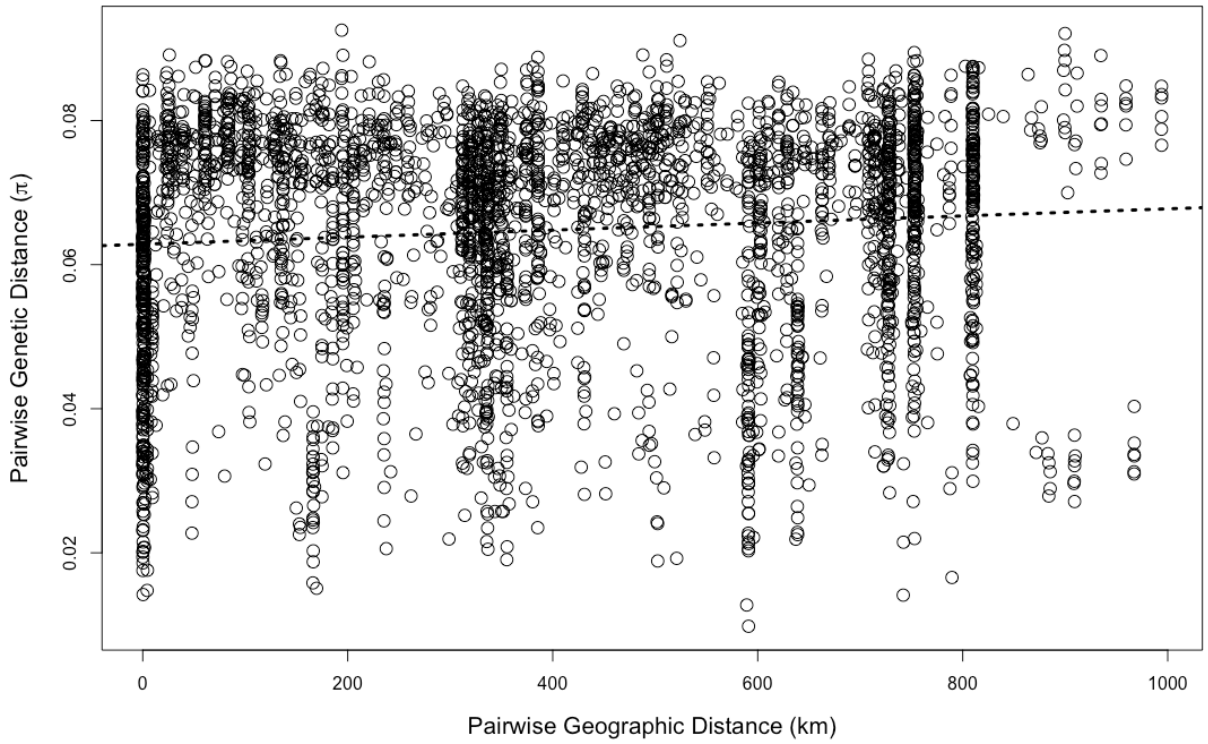


Figure 1: Pairwise genetic distance versus pairwise geographic distance for 83 Eastern Box Turtles. There is no significant relationship between pairwise genetic distance and pairwise geographic distance (Mantel R = -0.08, P = 0.99, linear R^2 = 0.006), indicating that Isolation by Distance does not play a significant role in this dataset.

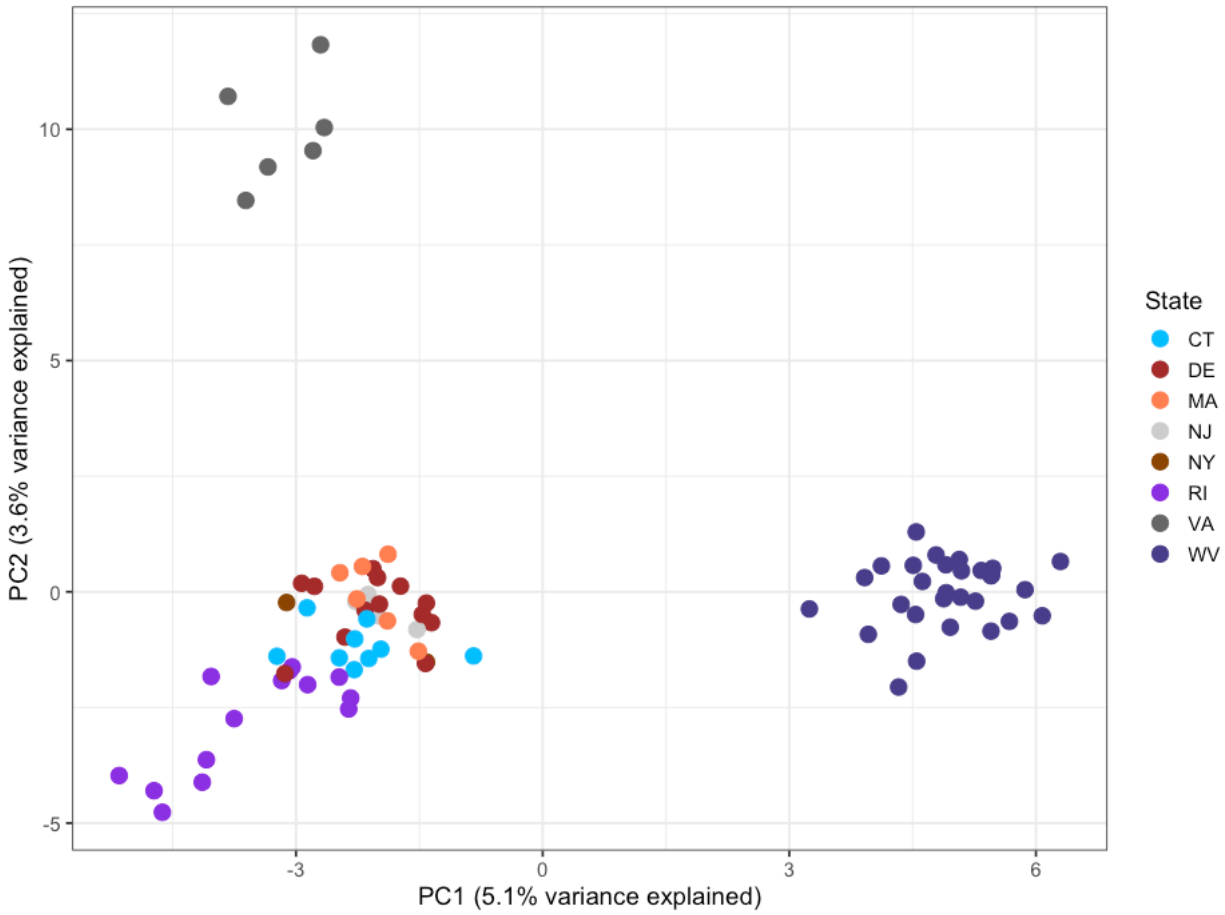


Figure 2: Principal components analysis of 83 Eastern Box Turtles. West Virginia, Virginia and New England each appear as distinct clusters.

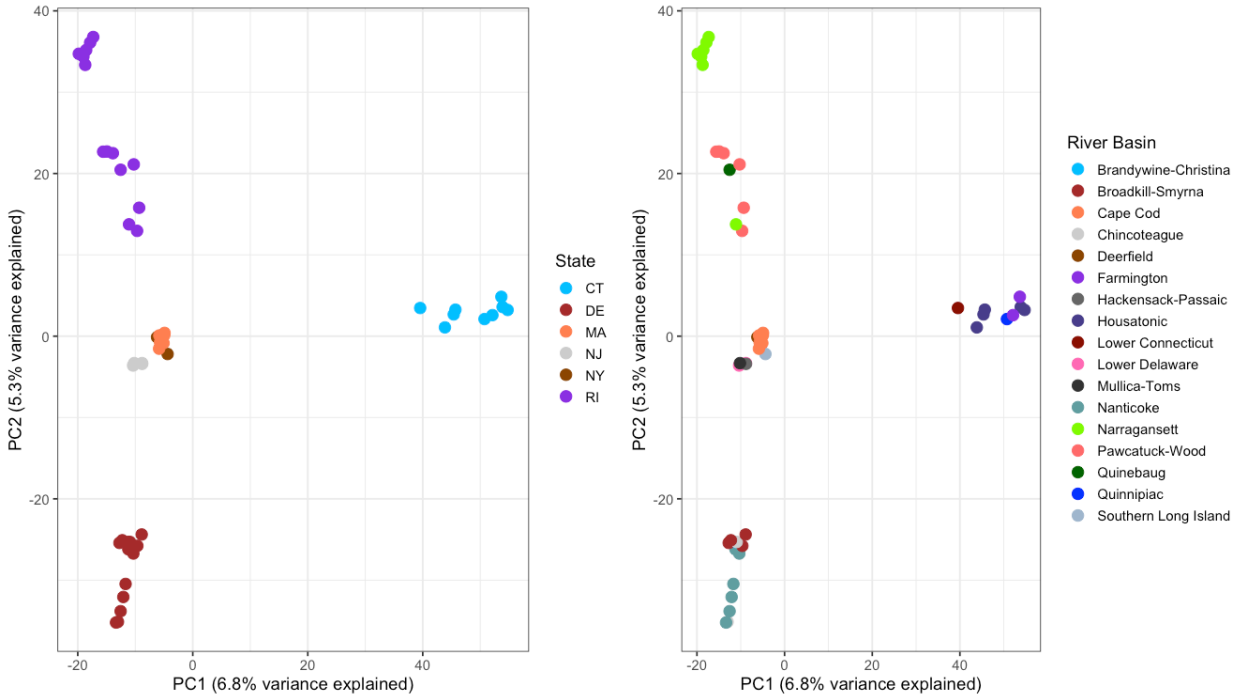


Figure 3: Principal component analysis of 48 Eastern Box Turtles from New England, colored by state and by river drainage.

Literature Cited

Bayona-Vásquez, N. J., T. C. Glenn, T. J. Kieran, T. W. Pierson, S. L. Hoffberg, P. A. Scott, K. E. Bentley, et al. 2019. Adapterama III: Quadruple-indexed, double/triple-enzyme RADseq libraries (2RAD/3RAD). PeerJ 7:e7724.

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

Kimble, S. J. A., O. E. Rhodes Jr., and R. N. Williams. 2014. Unexpectedly Low Rangewide Population Genetic Structure of the Imperiled Eastern Box Turtle Terrapene c. carolina. (A. Stow, ed.)PLoS ONE 9:e92274.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press.

Raj, A., M. Stephens, and J. K. Pritchard. 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics 197:573–589.

Wright, S. 1943. Isolation by Distance. Genetics 28:114–138.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328.