FINAL PERFORMANCE REPORT

Federal Aid Grant No. F18AP00228 (E-90-R-1)

Conservation Genomics of the Neosho Mucket
(Lampsilis rafinesqueana)

Oklahoma Department of Wildlife Conservation

January 1, 2018 – December 31, 2020

FINAL PERFORMANCE REPORT

State: Oklahoma **Grant Number:** F18AP00228 (E-90-R-1)

Grant Program: Endangered Species Act Traditional Section 6

Grant Title: Conservation Genomics of the Neosho Mucket *(Lampsilis rafinesqueana)*

Grant Period: January 1, 2018 to December 31, 2020

Principal Investigators:
David J. Berg, Professor, Department of Biology, Miami University, Oxford, OH 45056; bergdj@miamioh.edu
Co-PI: Steven R. Hein, Doctoral Student Department of Biology, Miami University, Oxford, OH

Executive Summary/Abstract:
The goal of this project is to gain an understanding of the population genomics of *L. rafinesqueana* by employing a next-generation sequencing method known as type IIb restriction site associated DNA sequencing (2bRAD). Samples from across the range of *L. rafinesqueana* were collected by a collaboration of federal and state biologist from Arkansas, Oklahoma, Missouri, and Kansas. All collected samples were shipped to Miami University overnight and were immediately placed in a -80°C ultra-cold freezer until DNA could be extracted for lab analyses. We have obtained a total of 164 *L. rafinesqueana* samples from seven of eight targeted rivers. Despite efforts made by state and federal biologists, no samples were recovered from the Neosho River. Currently, we have sequenced 79 individuals from these rivers (Shoal Creek, n=21; North Fork of Spring River, n=16; Fall River, n=6; Verdigris River, n=2; Illinois River, n=10; Spring River, n = 10; Elk River, n=14). Following our strict data filtering protocol, we were left with 2,405 SNPs in our data set. We calculated Tajima's estimator (π) for each population and found similar levels of nucleotide diversity. Pairwise FST values between populations showed low, but statistically significant, genetic differentiation among all pairs of populations. We believe gene-flow may have once been high amongst populations; however, in recent times all sites have become genetically isolated from each other. This presents a serious challenge to the future survival of each local population and the species as whole. Populations with low effective population sizes are at greater risk of extinction as they move into an "extinction vortex."

Objectives:
In cooperation with the states of Oklahoma, Kansas, Missouri, and Arkansas, this study will address the following goals:
1. To estimate within-river genetic variation for extant populations of Neosho mucket from the Illinois, Neosho, and Verdigris basins.
2. To estimate among-population genetic variation using a hierarchical approach that partitions variation into within-river, among-river, and among-basin components.

3. To use the results of these analyses to inform the development of recovery strategies that ensure maintenance of genetic variation into the future for this species, while also retaining significant geographic variation.

Summary of Progress:

Introduction:

Historically found throughout the Illinois, Neosho, and Verdigris basins, the federally endangered Neosho Mucket (Lampsilis rafinesqueana) has undergone a 62% reduction in occupied range. The extant populations are fragmented, and little is known about gene flow between them (USFWS, 2017). The goal of this project is to gain an understanding of the population genomics of *L. rafinesqueana* by employing a next-generation sequencing method known as type IIb restriction site associated DNA sequencing (2bRAD; Wang et al. 2012). This method allows the incorporation of thousands of single nucleotide polymorphisms (SNPs) found throughout the genome, increasing genetic loci sampled by orders of magnitude compared with microsatellites. We are currently using 2bRAD methods to understand genetic variation within and among geographic populations of *L. rafinesqueana*.

Methods:

Samples from across the range of *L. rafinesqueana* were collected by a collaboration of federal and state biologist from Arkansas, Oklahoma, Missouri, and Kansas. We have received DNA samples from seven of eight occupied rivers: Fall River, Verdigris River, Illinois River, Elk River, Spring River, the North Fork of the Spring River (considered a distinct geographic population segment), and Shoal Creek. Ironically, sampling of the Neosho River yielded no individuals. NOTE: Sampling in 2019 was limited due to weather, while sampling in 2020 was delayed due to the pandemic. Final samples were not sent to us until November 2020. This late date, along with pandemic restrictions in terms of laboratory access, delayed DNA extraction and library preparation of samples from 2020. We finished all laboratory work in December 2020. Samples will be sequenced by the University of Oregon and data transmitted to us in early January. A revised final report will be sent once analysis of the complete data set is finished.

All collected samples were shipped to Miami University overnight and were immediately placed in a -80°C ultra-cold freezer until DNA could be extracted. We extracted DNA from swabs using a Qiagen QIAamp DNA mini kit (Qiagen Inc., Hilden, Germany) with slight modification to the manufacture's recommended protocol. Extracted DNA was further purified and concentrated using a Promega ReliaPrep DNA Clean-Up and Concentration kit (Promega Corporation, Madison, WI, USA) following the manufacture's recommended protocol. Gel electrophoresis was performed on the samples using a 0.7 or 1% agarose gel in 1X sodium borate EDTA buffer (SBE) alongside a high-range DNA ladder to assess the quality of the extracted DNA. High quality or high molecular weight double-stranded DNA (dsDNA) produces a sharp band on the gel at around 26 kilo-bases. Degraded or low-quality dsDNA produces a lighter band or a long smear with no distinct bands but still some material in the ~26kb range. Highly degraded DNA with very little or no dsDNA will produce no band at ~26kb range and only a light smear toward the end of the gel lane. All of the 2019 swab samples from the Verdigris and Fall rivers were

degraded with little or no banding at the 26kb region. All the other samples (including 2020 samples from the Verdigris and Fall rivers) displayed a solid firm band at the correct region of the gel, indicating a high concentration of high-quality dsDNA. We fquantified dsDNA concentration using a Qubit 4 fluorometer (Thermo Fisher Scientific, Waltham MA, USA). All samples displayed high concentrations of dsDNA (>90ng/μL), with the exception of the 2019 Verdigris and Fall samples, which were much lower. We suspect that the DNA on these swabs did not properly preserve. A number of possibilities could lead to failed DNA preservation including an issue with the preservation fluid, an accidental exposure to extend period of heat or direct sunlight, or some other issue.

Following DNA extraction and quality control, we created 2bRAD libraries of the samples following protocols outlined by Wang et al. (2012). In brief, 8ul of sample DNA was digested with AlfI restriction enzyme at 37°C for 16 hours to produce 36 base-pair (bp) DNA fragments. We then ligated 1/8th reduction scheme adaptors to both the 5' and 3' ends of DNA fragments using T4 ligase at 16°C for 16 hours. Following ligation, we performed a polymerase chain reaction (PCR) to attach custom made oligonucleotide barcodes onto either end of the ligation construct and amplify the number of constructs. Double-stranded DNA was then quantified for each of the barcoded DNA constructs using a Qubit 4 fluorometer. Once DNA concentrations were determined, we pooled equal quantities from each sample into 2bRAD libraries. The pooled libraries were once again quantified to ensure dsDNA concentration fell within the required sequencing concentration of 2–20nM. Libraries were then sent to The University of Oregon's Genomics & Cell Characterization Core Facility (GC3F) for sequencing on an Illumina HiSeq 4000 (Illumina Inc., Sane Diego, CA, USA). Individual samples from the combined pool were demultiplexed by the GC3F, based on unique barcodes.

Initial data filtering steps were performed using custom Perl scripts written by Eli Meyer of Oregon State University that are freely available on Github (https://github.com/EliMeyer/2bRAD_utilities). We first truncated the raw reads for each DNA fragment sequence (RADtag) to 36 base-pairs (the length of the Alf1 fragment) to remove adaptors and barcode sequences using the script "Truncate.pl". We then used "QualFilterFastq.pl" to filter each RADtag by the Phred score (a measure of nucleotide sequencing quality) assigned to each nucleotide. Any RADtags containing more than two nucleotides with a Phred score of 25 or below were removed because this score represents only a 0.5% chance of sequencing error. We then applied the "AdaptorFilterFastq.pl" script, which utilizes the software cross-match (Gordon 2003), to filter and remove RADtags for adaptor sequence artifacts created during library construction. The high-quality truncated reads were then further processed using the software Stacks v 2.4.1 (Rochette et al. 2019). Each RAD data set is unique and therefore requires specific parameterizations. We first determined optimal Stacks parameters using the "r80" method described in Paris et al. (2017). Based on our optimizations, we used the following parameters for genotyping and SNP discovery: minimum number of raw reads required to form a putative allele or stack (m) was set to 3, number of mismatches allowed between stacks to merge them into a putative locus (M) was set to 2, and number of mismatches allowed between stacks during construction of the catalog (n) was set to 2. Once we determined our "m, M, n" parameters, we ran the "denovo_map.pl" program in Stacks 2 using those parameters while keeping all other options at default. We further filtered the dataset using the "Populations" program in Stacks 2. We set the minimum minor allele frequency to 0.05 and

maximum observed heterozygosity to 0.50. To reduce the possibility of linkage disequilibrium, we restricted our data set to retain only a single SNP per locus. We retained a maximum of only 20% missing data for a given sample by setting the minimum percentage of individuals across populations and minimum percentage of individuals in a population required to process a locus for that population, “R” and “r” respectively, both to 0.8. We further reduced missing data by requiring a locus to be present at 6 of the 7 sampling localities in order to be processed.

We used the program Populations, implemented in Stacks 2, to generate standard summary statistics within-and-between each population. In this initial analysis, we considered each sampling location as a distinct geographic population. We used a combination of the R packages Stampp (Pembleton et al. 2013), adegenet (Jombart 2008), and HIERFSTAT (Goudet 2004) to generate a fixation index (FST) between each pair of populations. This is a standard measurement to describe among-population lgenetic differentiation. The value of FST ranges from 0 to 1, with 0 meaning no genetic differentiation between populations and 1 meaning completely distinct populations, i.e. fixation of alternate alleles between populations. We assessed statistical significance of FST by generating 95% confidence intervals using a bootstrap analysis with 1000 replicates. We assessed genetically effective population size (Ne) following the linkage disequilibrium (LDNe) method for each of the sites using the software NeEstimator V2 (Do and Waples 2014). Following the recommendations of Waples and Do (2010) for small sample sizes, we excluded singleton loci to increase overall accuracy of the analysis. We then used adegenet to assess population structure using a K-means clustering algorithm and a discriminant analysis of principle components (DAPC). Finally, we used the Bayesian clustering program ADMIXTURE (Alexander et al. 2009) to determine number of genetic populations and look for signs of population admixture. Number of populations (k) was determined using a 10-fold cross-validation. The optimal k value is determined by the lowest cross-validation score. We assessed contemporary migration rate (mc) using the BayesAss algorithm (Wilson and Rannala 2003) optimized for genomic scale data and implemented in BA3-SNP (Mussman et al. 2019). We first determined optimal model mixing parameters using the python script BA3-SNPautotune.py (github.com/stevemussmann/BA3-SNPS-autotune, Mussman et al. 2019). Once the parameters were determined, we ran BA3-SNP for one million generations with a 100,000 burnin period, and a Markov chain Monte Carlo sampling interval of 100 generations. We set our previously optimized mixing parameters for migration rate (deltaM), allele frequency (deltaA), and inbreeding coefficient (deltaF) to 0.2125, 0.5500, and 0.050, respectively. Statistical significance of the contemporary migration rate was determined by generating a 95% confidence interval around the mean mc value. If the confidence interval included zero, we concluded that there was no evidence of contemporary migration between populations.

Results and Discussion:

We have obtained a total of 164 *L. rafinesqueana* samples from seven of eight targeted rivers. Despite efforts made by state and federal biologists, no samples were recovered from the Neosho River. Currently, we have sequenced 79 individuals from these rivers (Shoal Creek, n=21; North Fork of Spring River, n=16; Fall River, n=6; Verdigris River, n=2; Illinois River, n=10; Spring River, n = 10; Elk River, n=14). Following our strict data filtering protocol, we were left with 2,405 SNPs in our data set. The additional library to be sequenced includes 64 samples from the Spring (n = 18), Fall (n =6), Verdigris (n = 18), and Illinois (n = 20) rivers. NOTE: The analyses

below are based on the sequences completed so-far and should be considered preliminary. We found very few private alleles within any of the seven analyzed populations (n = 0–3), suggesting little genetic differentiation between populations. Populations in the Illinois River, the North Fork of the Spring River, and Shoal Creek showed heterozygote deficits (Ho< He) and significant inbreeding coefficients (FIS); these results may be associated with a reduction in population size and genetic drift (Table 1). The Verdigris River showed a higher rate of observed heterozygosity than expected under Hardy-Weinberg equilibrium. However, the sample size was very small and therefore this result should be viewed with caution.

We calculated Tajima's estimator (π) for each population and found similar levels of nucleotide diversity (Table 1). Pairwise FST values between populations showed low, but statistically significant, genetic differentiation among all pairs of populations (Table 2). We are currently exploring tests of isolation-by-distance to further explain these patterns. The K-means clustering analysis indicated that the most likely number of clusters within the data is 1. This shows that without any prior population identity, the data do not cluster into specific groups or populations. The results from our ADMIXTURE analysis displayed the same pattern by again indicating a most likely k value of 1. This is an indication of long-term, large-scale admixture or migration amongst all sites. A k-value of one is an indication that gene flow amongst the rivers has historically been high enough to prevent any one population from developing its own genetic signature. In genetic terms, there is only a single population shared amongst all rivers. Our tests of genetically effective population sizes (Ne) are relatively consistent with estimation of total population sizes (USFWS, 2017). The measurements are also equal to the order of magnitude expected of an endangered species (Table 1). The Spring River, which is considered to have one of the largest "populations", has Ne = 2315 (95% CI= 524.2–∞). The Illinois River is on the other end of the spectrum, with Ne = 12.3 (95% CI= 12.1–12.6).

Effective population size is essentially a measurement of genetic drift in a population. When a population declines, heterozygosity is lost (via inbreeding) and drift increases. Therefore, a small value of Ne is an indication of increased genetic drift. Values of infinity do not indicate that there is no drift but that we lack power to detect it, likely due to small samples sizes (Do et al. 2014, Waples and Do 2010). In the case of the Spring River, where the upper bound is infinity but the lower bound is still a finite number, we lack resolution of the upper confidence bound but we still have a good estimate of the lower bound (Waples and Do 2010). We found that there is no detectable current migration amongst any pair of sites, including the two portions of the Spring River. All of the 95% confidence intervals overlap zero, indicating that there no statistical differences between our values and zero (Table 3). This result, along with the lack of population structure between each pair of sites, is evidence for the disruption of contemporary gene flow due to dams and other anthropogenic obstructions throughout the range of *L. rafinesqueana*. We believe gene-flow may have once been high amongst populations; however, in recent times all sites have become genetically isolated from each other. This presents a serious challenge to the future survival of each local population and the species as whole. Populations with low effective population sizes are at greater risk of extinction as they move into an "extinction vortex (Fagan and Holmes, 2006)." Genetic drift will continue to increase in small populations leading to an increasing loss of genetic diversity and increased inbreeding. An increase in inbreeding can lead to a decrease in adaptive potential and increased vulnerability to disease, hereditary condition, and environmental change (Pavlova et al. 2017). This causes population size to decrease which,

in turn increases genetic drift. As a result, Ne decreases and the population spirals further down the vortex. We are currently working on forward-time models to project genetic drift and Ne into the future.

Recommendations:

We believe gene-flow may have once been high amongst populations; however, in recent times all sites have become genetically isolated from each other. This presents a serious challenge to the future survival of each local population and the species as whole. Populations with low effective population sizes are at greater risk of extinction as they move into an "extinction vortex". Conservation of these local populations is paramount. Any regulatory protections available should be enacted to protect these local populations and their habitats.

Significant Deviations:
None

Equipment Purchased During Grant (Cumulative):
No equipment was purchased.

Prepared by: David J. Berg, Professor, Department of Biology, Miami University, and Steven R. Hein, Doctoral Student, Department of Biology, Miami University

Date Prepared: March 4, 2021

Approved by: Russ Horton, Assistant Chief of Wildlife
Oklahoma Department of Wildlife Conservation

Andrea K. Crews, Federal Aid Coordinator
Oklahoma Department of Wildlife Conservation

Tables

Table 1. Summary statistics for the 79 genotyped samples, H_o = observed heterozygosity, H_e = expected heterozygosity, F_{IS} = inbreeding coefficient, π = Tajima's estimator or the average number of differences between individuals in a population, N_e = genetically effective population size.

	H_o	H_e	π	F_{IS}	N_e
Elk	**0.245** (0.238 - 0.252)	**0.256** (0.249 - 0.262)	**0.265** (0.26 - 0.272)	**0.060** (0.033 - 0.125)	∞ (1572.2 - ∞)
Falls	**0.220** (0.211 - 0.229)	**0.217** (0.209 - 0.224)	**0.237** (0.229 - 0.246)	**0.000** (0.024 - 0.087)	∞ (∞ - ∞)
Illinois	**0.229** (0.223 - 0.236)	**0.261** (0.255 - 0.268)	**0.276** (0.269 - 0.282)	**0.138** (0.113 - 0.358)	**12.3** (12.1 – 12.6)
N. Fork Spring	**0.238** (0.231 - 0.245)	**0.258** (0.252 - 0.264)	**0.267** (0.261 - 0.273)	**0.092** (0.056 - 0.203)	**102.9** (96.7 – 109.8)
Shoal	**0.242** (0.236 - 0.249)	**0.257** (0.250 - 0.262)	**0.263** (0.257 - 0.27)	**0.066** (0.027 - 0.119)	**269.8** (244.7 – 300.4
Spring	**0.238** (0.230 - 0.246)	**0.245** (0.238 - 0.251)	**0.260** (0.252 - 0.267)	**0.055** (0.037 - 0.127)	**2315** (524.2 - ∞)
Verdigris	**0.230** (0.216 - 0.244)	**0.164** (0.156 - 0.172)	**0.219** (0.207 - 0.23)	**-0.018** (-0.018 - -0.052)	∞ (∞ - ∞)

Table 2. Pairwise F_{ST} values between populations.

	Illinois	Falls	Verdigris	Shoal	N. Fork	Elk
Illinois						
Falls	**0.0594**					
Verdigris	**0.0690**	**0.0843**				
Shoal	**0.0341**	**0.0418**	**0.0423**			
N. Fork	**0.0314**	**0.0418**	**0.0447**	**0.0270**		
Elk	**0.0343**	**0.0468**	**0.0523**	**0.0296**	**0.0255**	
Spring	**0.0402**	**0.0584**	**0.0733**	**0.0293**	**0.0255**	**0.0314**

Table 3. Contemporary migration rate (m_c) between pairs of populations, calculated by BAE-SNP; numbers in parentheses are 95% confidence intervals. Migration is from row population to column population.

	Elk	Falls	Illinois	N. Fork	Shoal	Spring	Verdigris
Elk		**0.016** (-0.013 - 0.045)	**0.015** (-0.012 - 0.0432)	**0.017** (-0.013 - 0.046)	**0.015** (-0.015 - 0.045)	**0.016** (-0.015 - 0.047)	**0.019** (-0.011 - 0.049)
Falls	**0.025** (-0.018 - 0.069)		**0.025** (-0.02 - 0.07)	**0.028** (-0.017 - 0.072)	**0.025** (-0.02 - 0.071)	**0.024** (-0.02 - 0.067)	**0.024** (-0.019 - 0.067)
Illinois	**0.016** (-0.017 - 0.05)	**0.019** (-0.017 - 0.055)		**0.020** (-0.015 - 0.0549)	**0.021** (-0.016 - 0.059)	**0.02** (-0.015 - 0.054)	**0.019** (-0.016 - 0.053)
N. Fork	**0.014** (-0.011 - 0.039)	**0.015** (-0.012 - 0.041)	**0.015** (-0.012 - 0.042)		**0.013** (-0.012 - 0.039)	**0.014** (-0.013 - 0.041)	**0.015** (-0.011 - 0.041)
Shoal	**0.012** (-0.01 - 0.033)	**0.013** (-0.012 - 0.038)	**0.011** (-0.01 - 0.032)	**0.011** (-0.008 - 0.03)		**0.011** (-0.008 - 0.03)	**0.015** (-0.011 - 0.04)
Spring	**0.018** (-0.016 - 0.052)	**0.018** (-0.014 - 0.05)	**0.021** (-0.018 - 0.06)	**0.020** (-0.017 - 0.056)	**0.020** (-0.015 - 0.054)		**0.021** (-0.016 - 0.058)
Verdigris	**0.037** (-0.026 - 0.100)	**0.037** (-0.028 - 0.101)	**0.039** (-0.029 - 0.107)	**0.040** (-0.032 - 0.112)	**0.036** (-0.023 - 0.095)	**0.036** (-0.03 - 0.101)	

Literature Cited

- Alexander, DH, Novembre, J, & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19, 1655-1664.
- Do, C, Waples, RS, Peel, D, Macbeth, GM, Tillett, BJ, & Ovenden, JR. (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular Ecology Resources, 14, 209-214.
- Fagan, WF, & Holmes, EE. (2006) Quantifying the extinction vortex. Ecology Letters, 9, 51-60.
- Gruber, B, Unmack, PJ, Berry, OF, & Georges, A. (2018) dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Molecular Ecology Resources, 18, 691-699.
- Gordon, D. (2003) Viewing and editing assembled sequences using Consed, chapter 11, unit 11.2. In Current Protocols in Bioinformatics, John Wiley and Sons, Somerset, NJ.
- Goudet, J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes, 5, 184-186.
- Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24, 1403-1405.
- Kardos, M, Taylor, HR, Ellegren, H, Luikart, G, & Allendorf, FW. (2016) Genomics advances the study of inbreeding depression in the wild. Evolutionary Applications, 9, 1205–1218.
- Mussmann, SM, Douglas, MR, Chafin, TK, & Douglas, ME. (2019) BA3-SNPs: Contemporary migration reconfigured in BayesAss for next-generation sequence data. Methods in Ecology and Evolution, 10, 1808-1813.
- Nomura, T. (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. Evolutionary Applications, 1, 462-474.
- Paris, JR, Stevens, JR, & Catchen, JM. (2017) Lost in parameter space: a road map for stacks. Methods in Ecology and Evolution, 8, 1360-1373.
- Pavlova, A, Beheregaray, LB, Coleman, R, Gilligan, D, Harrisson, KA, Ingram, BA, ... & Nguyen, TT. (2017) Severe consequences of habitat fragmentation on genetic diversity of an endangered Australian freshwater fish: A call for assisted gene flow. Evolutionary Applications, 10, 531-550.
- Pembleton, LW, Cogan, NO, & Forster, JW. (2013) St AMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Molecular Ecology Resources, 13, 946-952.
- Rochette N, Rivera-Colón A, and Catchen J. (2019) Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Molecular Ecology, 28, 4737-4754.
- US Fish and Wildlife Service. (2017) Species Biological Report Neosho Mucket (Lampsilis rafinesqueana). Southeast Regional Office, Atlanta, GA.
- Wang, S, Meyer, E, McKay, JK and Matz, MV. (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature Methods, 9, 808.
- Waples, RS, & Do, CHI. (2010) Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. Evolutionary Applications, 3, 244-262.
- Wilson, GA, & Rannala, B. (2003) Bayesian inference of recent migration rates using multilocus genotypes. Genetics, 163, 1177-1191.