Development of Arkansas River Shiner (*Notropis girardi*) Genomic Tools

Progress report for NMDGF-UNM Agreement #240911, State Wildlife Grant #T-85-R-1
Submitted for period ending 30th June 2025

Submitted by:

Megan Osborne and Guilherme Caeiro-Dias
Department of Biology & Museum of Southwestern Biology
University of New Mexico
Albuquerque, NM 87131
505-277-3234
Email: mosborne@unm.edu

Submitted to:
Karen Gaines
Share with Wildlife Program Coordinator
Wildlife Management Division
New Mexico Department of Game and Fish
1 Wildlife Way
Santa Fe, NM 87507

Background

The Arkansas River Shiner is an endemic pelagic broadcast-spawning fish restricted to the Arkansas River basin. This species has been extirpated from much of its native range including the Ninnescah and Arkansas Rivers in Kansas (Perkin et al. 2014) and is listed as threatened under the Endangered Species Act (U.S. Fish and Wildlife Service 1998). Hence, its sole stronghold is two distinct fragments (separated by Lake Meredith in Texas [TX]) of the South Canadian River (between Ute Lake in New Mexico [NM] and Lake Eufaula in Oklahoma [OK]). Baseline genetic information (collected in 2009, 2012, 2014, 2015, 2017 and 2019) from microsatellites and mitochondrial DNA documented the spatial distribution and amount of diversity and provided estimates of contemporary effective population size (Osborne et al. 2021).

Genetic monitoring involves tracking genetic diversity and effective population size (N_e) across temporally spaced samples from the same population using neutral genetic markers (Schwartz et al. 2007). Until recently, genetic monitoring programs relied on highly polymorphic microsatellite markers to obtain diversity and N_e estimates, but rapidly changing technology has led to a shift toward assaying variation using SNPs that represent the most widespread source of variation within genomes (Brumfield et al. 2003). With the development of increasingly fast and inexpensive high-throughput Next Generation Sequencing (NGS) methods, it is now easy to identify a sufficient number of SNPs in a sample to overcome the advantages of using microsatellites and to surmount the lower resolution power of small numbers of SNPs (Hess et al. 2011).

Reduced representation sequencing methods, like Nextera-tagmented reductively-amplified DNA sequencing (nextRAD-seq; Russello et al. 2015), are cost-effective ways to identify thousands of SNPs across several hundreds of samples. When the number of loci to be genotyped is relatively small (e.g., a few hundred) and the number of samples is high (e.g., hundreds to thousands), methods based on multiplex PCR and NGS can be more advantageous. Genotyping-in-Thousands by sequencing (GT-seq) is a method of targeted SNP genotyping that uses multiplexed PCR amplicon sequencing (Campbell et al. 2015). This method allows simultaneous amplification of hundreds of targeted genetic loci while barcoding of individuals allows thousands of individual samples to be sequenced in a single lane with a compatible Illumina® sequencing instrument (Campbell et al. 2015). Sequencing the genome for the target taxon aids in identification of SNPs and facilitates locus-specific primer design. Once a GT-seq panel is developed for the target species, the method provides an efficient means of monitoring genetic variation and N_e estimated from hundreds of SNPs. There are currently no genomic resources available for the Arkansas River Shiner.

The project objectives are to:

(i) Sequence the genome of an Arkansas River Shiner individual.

Status: *In progress (see below for details).*

(ii) Use a representative subset of archived DNA samples to discover SNPs in the genome. DNA was isolated from samples collected in 2024 (n=82), and DNA was purified from 110 archived samples collected in 2009, 2012, 2015, 2017. These samples represent multiple South Canadian River localities. These samples were sent to SNPsaurus for library preparation and DNA sequencing.

Status: *In progress*. Sequencing data has been received from SNPsaurus and have commenced bioinformatic analyses to identify genome-wide neutrally evolving SNPs.

(iii) Develop and optimize PCR primers to characterize variation in ~300 variable loci (GT-seq panel) distributed across the genome.

Status: *Pending*- PCR primers will be identified and optimized once filtering of the SNP data is completed.

Molecular Methods: To obtain a high-quality long-read genome, we isolated high molecular weight (HMW) DNA from one Arkansas River Shiner collected from the Pecos River population where the species is not native. Prior genetic evaluation of the Pecos River population showed that this population was most likely established by transfer of individuals by bait bucket release from multiple source populations including the South Canadian River (Osborne et al. 2013). DNA was isolated using Qiagen® genomic tips. DNA was quantified using Qubit® assays to ensure that sufficient HMW DNA is available for library preparation. DNA was sent to UC Davis Genomics Facility (https://genomecenter.ucdavis.edu/) for library preparation and sequencing using PacBio technology.

For nextRAD sequencing, we purified previously isolated DNA from 110 samples collected from the South Canadian River, New Mexico, using Zymo® DNA clean and concentrator kits. DNA was isolated from samples collected in 2024 (n=82) using an Axygen genomic DNA isolation kit following the manufacturer's directions. DNA concentrations were quantified for all samples. Double-stranded DNA was quantified using Qubit® assays, and samples with high-quality DNA were selected for sequencing by SNPsaurus (http://snpsaurus.com). DNA sequence data were received from SNPsaurus on June 10th and we have begun bioinformatics analysis to identify SNPs. A brief description is provided below.

Data filtering
Initially, raw sequence reads will be mapped against a *de novo* genome assembly for Arkansas River Shiner using Bowtie version 2.4.2 (Langmead & Salzberg, 2012). Successfully aligned

reads will be filtered with Samtools v. 1.16 (Danecek et al., 2021; Li et al., 2009) to remove reads with low mapping quality. Genetic variants will be identified using FreeBayes v. 1.3.6 (Garrison & Marth, 2012). We will then apply extensive computational filtering so that the final dataset only contains high quality SNPs. VCFtools v. 0.1.16 (Danecek et al., 2011) will be used to remove variants with low coverage, to remove insertions and deletions, to retain only the biallelic SNPs, and to remove individuals with a high proportion of missing data. The bash script *dDocent_filters* (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) will be used to filter SNPs based on allelic balance at heterozygous genotypes, strand representation, and quality vs depth. Potential erroneous SNPs will be filtered based on Hardy-Weinberg equilibrium (HWE) expectations with the pearl script *filter_hwe_by_pop.pl* (https://github.com/jpuritz/dDocent/blob/master/scripts/filter_hwe_by_pop.pl). SNPs present in more than 50% of the populations (here each year was considered a "population") and with an HWE p-value lower than 0.001 will be removed. Potentially incorrectly-assembled paralogous loci that exhibit a large variation in read depth across all individuals will also be removed. The remaining SNPs will be used to identify haplotypes within genetic loci (referred to as "microhaplotypes"). Haplotyping SNPs within a locus also eliminates possible paralogous loci while neutralizing physical linkage without losing data (Willis et al., 2017). This step will be performed with the *rad_haplotyper.pl* pearl script (Willis et al. 2017; https://github.com/chollenbeck/rad_haplotyper). Retained loci will be tested for deviations from HWE and for linkage disequilibrium (LD), considering individuals captured in each year as a single "population." After filtering, the final dataset will represent a robust genome-wide neutral SNP dataset suitable for primer design for the GT-seq panel optimization steps.

Acknowledgements

Funding for this project was provided by the Share with Wildlife program of the New Mexico Department of Game and Fish (State Wildlife Grant #T-85-R-1; NMDGF-UNM Agreement # 240911). We thank Karen H. Gaines for editing assistance.

References

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, 18(5):249-256.

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, 15(4):855-867.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo, MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin K, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:*1207.3907.

Hess JE, Matala AP, S. R. Narum (2011) Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. *Molecular Ecology Resources*, 11:137-149.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357-359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Osborne MJ, Hatt JL, Gilbert EI, Davenport SR (2021) Still time for action: genetic conservation of imperiled South Canadian River fishes, Arkansas River Shiner (*Notropis girardi*), Peppered Chub (*Macrhybopsis tetranema*) and Plains Minnow (*Hybognathus placitus*). *Conservation Genetics*, 22(6):927-945.

Osborne MJ, Diver TA, Turner TF (2014) Introduced populations as genetic reservoirs for imperiled species: a case study of the Arkansas River Shiner (*Notropis girardi*). *Conservation Genetics* 14:637-47.

Perkin JS, Gido KB, Cooper AR, Turner TF, Osborne MJ, Johnson ER, Mayes KB (2014) Fragmentation and dewatering transform Great Plains stream fish communities. *Ecological Monographs,* 85(1):73-92.

Russello MA, Waterhouse MD, Etter PD, Johnson EA (2015) From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ*, *3*:e1106.

Schwartz, MK, Luikart, G, Waples, RS (2007) Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology and Evolution,* 22:11-16.

U. S. Fish and Wildlife Service (1998) Endangered and Threatened Wildlife and Plants; Final Rule to List the Arkansas River Basin Population of the Arkansas River Shiner (*Notropis girardi*) as Threatened. *Federal Register*, 63:64772-64799.