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

Year 1 Final Report for NMDGF-UNM Agreement #240911,
State Wildlife Grant #T-85-R-1
Submitted for period ending 15 December 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 (*Notropis girardi*) 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 single nucleotide polymorphisms (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 disadvantages of using microsatellites and to surmount the lower resolution power of small numbers of SNPs (Hess et al. 2011; Liu et al. 2005; Narum et al. 2008).

Reduced-representation sequencing methods, such as Nextera-tagmented reductively-amplified DNA sequencing (nextRAD-seq), are cost-effective ways to identify thousands of SNPs across several hundreds of samples (Russello et al. 2015). 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 enables the simultaneous amplification of hundreds of targeted genetic loci across thousands of individual samples due to genetic barcoding of individuals (Campbell et al. 2015). Sequencing the genome for the target taxon aids in identifying SNPs and facilitating locus-specific primer design (Caeiro-Dias et al. 2026). After a GT-seq panel has been developed for the target species, the panel 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: *Complete.*

(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 reduced representation sequencing using a nextRAD approach. From this data 5,033 loci including 8,990 SNPs (microhaplotypes) were identified and represent a putatively neutral dataset.

Status: *Complete.*

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

Status: *In Progress*. From the 5,033 loci identified in the previous step, 2,173 passed the filters for GT-seq compatibility. From those loci, several subsets of 500 loci were selected. Next, we tested the power of each subset to obtain estimates of genetic diversity comparable to the complete dataset. A set of 500 loci was selected. This reflects the progress accomplished until December 31st, 2025. Next, target loci information and DNA samples were sent to GTseek, LLC. in early January for PCR multiplex optimization.

Methods

Genome sequencing and assembly

To obtain a high-quality long-read genome, we isolated high molecular weight (HMW) DNA from one Arkansas River shiner collected by U.S. Fish and Wildlife Service personnel in 2025 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. 2014). DNA was isolated using Qiagen® genomic tips G/20 according to the manufacturer's directions. Double stranded DNA was quantified using Qubit® (Thermo Fisher Scientific) fluorometer assays to ensure that sufficient HMW DNA is available for library preparation. A DNA sample was sent to the University of California Davis (UC) Genomics Facility (https://genomecenter.ucdavis.edu/) for library preparation and sequencing. A genomic library was prepared using PACBio HiFi SMRTbell® and the whole genome was sequenced using one Revio SMRT cell on a PACBio Sequel II platform.

PacBio long-reads sequenced from Arkansas River shiner were received from UC Davis and a *de novo* genome assembly was performed with HiFiasm v. 0.18.1-r466 (Cheng et al. 2021) with default parameters. Genome contiguity was assessed using the continuous DNA sequence (contig) N50 value (i.e., the sequence length for the shortest contig that together with all other contigs of the same length or higher contain 50% of the total genome length). Genome completeness was measured using BUSCO scores v. 5.0 (Manni et al. 2021). Both genome contiguity and completeness were assessed using the online tool gVolante v. 2.0.0 (https://gvolante.riken.jp/analysis.html; Nishimura et al. 2017), specifying the Actinopterygii Ortholog set containing 3,640 genes. We also ran this analysis with the Core Vertebrate Genes (CVG) reference set of 233 genes.

Sampling and reduced-representation sequencing

For nextRAD sequencing, we used Zymo® DNA clean and concentrator kits to purify previously isolated DNA from samples collected by U.S. Fish and Wildlife Service personnel in in 2009 from the Pecos River (New Mexico) and by New Mexico Department of Game and Fish personnel in 2012, 2015, and 2017 from the South Canadian River (New Mexico). DNA was isolated from samples collected in 2024 from the South Canadian River (New Mexico) using an Axygen genomic DNA isolation kit following manufacturer's instructions. Double-stranded DNA concentrations were quantified for all samples using a Qubit® (Thermo Fisher Scientific) fluorometer with a 1X dsDNA HS Assay Kit. A total of 185 samples collected in 2009 (n=24), 2012 (n=32), 2015 (n=43), 2017 (n=4), and 2024 (n=82) containing enough high-quality DNA to sequence were then selected for reduced-representation sequencing and sent to SNPsaurus (http://snpsaurus.com) for sequencing (Table 1). SNPsaurus employs a proprietary nextRAD-seq method that reduces the proportion of genome to be sequenced by adding selective primers to DNA fragments as described by Russello et al. (2015); in this method, fragmentation of DNA is performed with a Nextera Reagent from Illumina® and adapter sequences are ligated to the fragments (Russello et al. 2015). Genomic libraries prepared by SNPsaurus were sequenced on an Illumina single-lane system to obtain 150 base pair (bp) long paired-end reads.

Identification of SNPs

Raw DNA sequences (or reads) were received from SNPsaurus with adapters and barcodes removed and demultiplexed by individual. Raw reads were trimmed with Trimmomatic v. 0.39 (Bolger et al. 2014) using a quality threshold of 20 base pairs on leading and trailing ends to remove low-quality and 'N' base calls (i.e., ambiguous bases). Each read was then screened with a 5-base sliding window and cut when the average quality per base dropped below 10, i.e., when average error probability is 10% or more. Resulting reads smaller than 60 bases long were discarded. Trimmed reads were aligned to the draft genome with Bowtie v. 2.5.1 (Langmead & Salzberg 2012) using paired-end input and a very-sensitive-local option. Next SAMtools v. 1.16.1 (Danecek et al. 2021) was used to remove reads with mapping quality lower than 20 and to convert SAM files to BAM files. Picard Tools v. 3.1.0 (The Broad Institute 2023) was used to

add reading groups, sort bam files, remove PCR duplicates and merge all BAM files into a single file containing all reads from all samples. Mapped regions with less than 150 of depth of coverage across all samples were filtered out using BEDTools v. 2.30.0 (Quinlan et al. 2010). Variants were identified with FreeBayes v. 1.3.6 (Garrison & Marth 2012) following the parallel scheme implemented in dDocent pipeline version 2.7.8 (Puritz et al. 2014).
FreeBayes uses base quality scores to estimate a probability for each allele. We kept a maximum of 10 raw variants from each alignment with higher probabilities and at least a base quality of five.

Table 1. Arkansas River shiner collections from the South Canadian River (SCR) and Pecos River (PR) and corresponding sample sizes used for nextRAD sequencing. Collectors include U.S. Fish and Wildlife Service (USFWS) and New Mexico Department of Game and Fish (NMDGF).

Drainage	Year	Collector	Number of Samples
PR	2009	USFSW	24
SCR	2012	NMDGF	32
SCR	2015	NMDGF	43
SCR	2017	NMDGF	4
SCR	2024	NMDGF	82
Total			185

To remove erroneous or potentially erroneous variants, we used VCFtools v. 0.1.16 (Danecek et al., 2011) to filter out variants with average depth of coverage (i.e., the average number of times a nucleotide was sequenced) lower than 20 and higher than 200, minor allele count less than three, minor allele frequency lower than 5%, genotype depth of coverage lower than five, and with genotype quality lower than 20. Multi-nucleotide states were decomposed into single variants with BCFtools (Danecek et al. 2021) and VCFtools was used to filter out nucleotide insertions and deletions and to retain only the bi-allelic SNPs. The dataset was then filtered by missing data, keeping SNPs present in at least 80% of samples and removing individuals with more than 30% missing data.

Next, SNPs were filtered using the bash script *dDocent_filters* (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) that used vcflib v. 1.0.9 (Garrison et al. 2022) and VCFtools to filter loci based on allelic balance at heterozygous genotypes, strand representation, and quality vs depth of coverage. First, loci were removed if heterozygous positions the alternative allele had a coverage lower than 20% or higher than 80% compared with the reference allele. This filter was applied because reads with alleles from heterozygous positions are expected to have similar frequencies in the same individual. Alleles with frequencies smaller than 0.01 and higher that 0.99 were not removed in order to account for

fixed alleles. Additionally, if the quality sum of the reference or the alternative allele was zero, the locus was removed. This removes positions with spurious heterozygous genotype calls. Loci with ratios between the mean mapping quality of the alternative and the reference allele of lower than 0.25 or higher than 1.75 were removed, because loci from the same genomic location should not have large discrepancies between the mapping qualities of the two alleles. Furthermore, loci with quality scores less than half of the total depth were excluded because excessive depth inflates FreeBayes quality scores. The average depth and standard deviation across all individuals were then calculated for the remaining loci. Loci with depths greater than the average depth plus one standard deviation were removed if their quality score was less than two times the depth. Finally, this script removed loci with an average depth across individuals greater than two times the mode (49) that corresponded approximately to the 95th percentile of average depth. Subsequently, potentially erroneous SNPs were 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). Typically, errors would have a low p-value and would be present in many populations; SNPs present in more that 50% of the populations (here, each year was considered as a discrete 'population') and with HWE p-values lower than 0.001 were removed. We filtered out potential incorrectly assembled paralogous loci that exhibited large variations in read depth across all individuals. Standard deviation was estimated using the R package *stats* and read depth was estimated with VCFtools. The previous and the following analyses conducted in R were performed in R v. 4.4.3 (R Core Team, 2025) in RStudio 2024.12.1.563 (Posit Team, 2025). Additional filtering based on missing data per locus (keeping loci present in 80% of individuals) was then reapplied.

Remaining SNPs were used to identify haplotypes within loci (referred to as microhaplotypes). Haplotyping SNPs within a locus eliminates possible paralogous loci while neutralizing physical linkage without losing data (Willis et al., 2017). This identification was performed with the *rad_haplotyper.pl* pearl script (https://github.com/chollenbeck/rad_haplotyper; Willis et al., 2017), excluding microhaplotypes if they were considered to be paralogs at least in five individuals and if they missing from more than 30% of individuals.

The retained loci were tested for deviations from HWE and for linkage disequilibrium (LD), considering individuals captured in each year as a different 'population' (including the Pecos River collection from 2009). The samples from fish collected in 2017 were not included due to small sample size (n = 4). Departures from HWE were assessed using a chi-square test on microhaplotype data with R package *pegas* v. 1.0 (Paradis 2010) and using the Bonferroni correction for multiple comparisons implemented in the R package rcompanion v. 2.4.0 (Mangiafico 2025), as implemented in the R function multi_HWE_tests (https://github.com/gcaeirodias/multi HWE tests; Caeiro-Dias et al. In Press). Estimations of LD were performed on SNP data using the SNP of each microhaplotype with higher minimum allele frequency. If a SNP was found in LD, then the entire locus was removed. Tests for LD were

performed using the chi-square test implemented in the R package *GUSLD* v. 1.0.1 (Bilton et al. 2018) and the Bonferroni correction to account for multiple simultaneous tests as implemented in the R pipeline significantLD (https://github.com/gcaeirodias/significantLD; Caeiro-Dias et al. In Press). If loci were found in multiple significant LD pairs, the loci that appeared in the highest number of comparisons were discarded in order to keep the maximum possible number of loci. In the remainder of instances, one locus from each pair was discarded randomly. Loci were considered to be deviating from HWE and to be in LD if tests were significant across the four temporal samples (p-value < 0.05). The resulting dataset should represent a robust genome-wide neutral SNP dataset and be suitable for primer design for the GT-seq panel development.

Genetic variation

Genetic variation between populations in the South Canadian and Pecos Rivers and across time in the South Canadian River was visualized using a discriminant analysis of principal components (DAPC), which summarizes genotypes in principal components to construct linear functions that maximize among-group variation while minimizing within-group variation. The analysis was performed using the R package adegenet v. 1.3–1 (Jombart 2008; Jombart & Ahmed 2011). Prior to performing the DAPC, we replaced missing data within each temporal sample using the Breiman's regression random forest algorithm (Breiman 2001) implemented in R package randomForest v. 4.6-14 (Liaw & Wiener 2002). Values of missing data were predicted from 1,000 independently constructed regression trees and 100 bootstrap iterations with default bootstrap sample size. We preferred using this method over the default "mean method" (i.e., missing genotypes are replaced by the average estimated across the data set) implemented in adegenet to ensure that we did not artificially increase similarity of allele frequencies across temporal samples. An initial DAPC was performed using temporal samples as groups, centering but not scaling allele frequencies, retaining all principal components (PCs) and discriminant functions (DFs), and keeping other options as default. The *a-score* method was used to select the optimal number of principal components to retain for the final DAPC, using the maximum number of PCs; all DFs and other options were set as default. The final DAPC was performed using the optimal number of PCs and two DFs while using the other default options.

Next, we assessed genetic differences between all collections by estimating pairwise Fixation index (F_{ST}) and p-values using 1,000 bootstrap iterations over loci as implemented in GenoDive v. 3.06 (Meirmans 2020).

Genetic diversity

For each temporal collection, we estimated standard genetic diversity and inbreeding metrics. The R package poppr v. 2.9.8 (Kamvar et al. 2014; 2015) was used to estimate observed heterozygosity (H_O), expected heterozygosity (H_E), and the corresponding 95% confidence intervals (CIs) using 1,000 bootstrap iterations over loci. Allelic richness (A_R), inbreeding

coefficient (F_{IS}) and the 95% CIs were estimated with R package diverRsity v. 1.9.90 (Keenan et al. 2013) using 1,000 bootstrap iterations over loci.

Contemporary effective population size (N_e)

SNP-containing loci (5,033) were used to estimate effective population size (N_e) for Arkansas River shiner using the single sample method based on linkage disequilibrium (LD N_e) implemented in NeEstimator. We also estimated variance effective population size (N_{eV}) for Arkansas River shiner using methods of Nei and Tajima (1981). N_{eV} measures the change in allele frequencies between two population samples that is caused by genetic drift. Both N_e estimates were calculated after excluding alleles occurring at frequencies of less than 2% (P_{CRIT}=0.02). The 95% confidence intervals for N_{eV} were calculated using the parametric approach (Waples 1989). The 2017 sample from the South Canadian River was not used for N_{eV} estimates because of small sample size. We also did not estimate N_{eV} for the Pecos River sample collected in 2009 because this method requires at least two temporal samples and we only had one sample.

Loci selection for GT-seq panel optimization

To select loci for GT-seq panel optimization, we used the same approach described in Caeiro-Dias et al. (In Press). Loci containing the filtered SNPs were filtered based on the SNP positions within each locus sequence using the bash script *identify_GT-seq_loci.sh* (https://github.com/gcaeirodias/GT-seq_filters; Caeiro-Dias et al. In Press). Using the draft Arkansas River shiner genome as a template, Primer3 command line version 2.5.0 (Untergasser et al. 2012) was used to design primers for those loci. Primer design parameters were defined as primer length of 18 to 25 bp, product size of 100 to 150 bp, melting temperature (Tm) of 60ºC, GC content of 50% (this ensures strong and stable binding to DNA template while reduces primer-dimer and poor annealing issues), and fewer than four consecutive repeat motifs (PolyX). When possible, we allowed design of up to five primer pairs for each locus. Finally, we mapped all primers to the reference genome using *blastn* program (Altschul et al. 1997) implemented in BLAST+ v. 2.9.0 (Camacho et al. 2009) and retained those primers that both forward and reverse mapped only once to the reference genome with 100% coverage and identity, as implemented in the bash script *blast_primers.sh* (https://github.com/gcaeirodias/GT-seq_filters; Caeiro-Dias et al. In Press). If a primer pair was discarded, the next best pair was selected with the bash script *alternative_primers.sh* (https://github.com/gcaeirodias/GT-seq_filters; Caeiro-Dias et al. In Press) and mapped on the draft genome as previously described. The process was repeated until a primer pair mapped only to the target locus or until no primer pairs remained.

From the loci with primer pairs that were potentially compatible with the GT-seq protocol, we selected a subset of 500 loci for panel optimization. To select those 500 loci, we followed a similar strategy used by Caeiro-Dias et al. (In Press). Based on preliminary results and results from Caeiro-Dias et al. (In Press), we compared three datasets. One dataset included the loci with

higher F_{ST} across collections (F_{ST}500); another dataset included 500 loci selected randomly (Rdm500); the third dataset included 250 loci with higher F_{ST} across collections and 250 loci selected randomly (F_{ST}250+Rdm250). For each subset of 500 loci, locus-specific F_{ST} was estimated with R package diveRsity. Inbreeding coefficient (G_{IS}) analogous to F_{IS}, H_O, and H_E, were estimated with GenoDive. Those metrics were compared to the complete nextRAD dataset with all filtered loci. Next, we assessed the power of each subset to obtain population-level metrics comparable to the complete nextRAD dataset. For each subset of 500 loci, pairwise F_{ST} was estimated for each collection with R package diveRsity using 1,000 bootstraps to estimate 95% CIs. For each temporal collection, we estimated standard genetic diversity and inbreeding metrics. The R package poppr was used to estimate H_O, H_E, and the corresponding 95% confidence intervals (CIs) using 1,000 bootstrap iterations over loci. Allelic richness (A_R), inbreeding coefficient (F_{IS}) and the 95% CIs were estimated with R package diverRsity using 1,000 bootstrap iterations over loci. The same metrics were estimated with the complete nextRAD dataset. If the CIs overlapped, we considered that estimates were not significantly different between datasets.

Results

Genome sequencing

A total of 8,040,297 raw HiFi reads were generated, containing 133.7 gigabases (Gb) of sequence data. Contig assembly resulted in 1,067 sequences, ranging from 1,032 to 74,822,736 bp (mean = 1,079,293; median = 34,322) with a total length of 1,151,605,706 bp (1.15 Gb) and an N50 of 38,127,334 bp. This N50 includes 12 contigs, meaning that 50% of the genome is covered by those 12 contigs. The total length assembled corresponds to a 106.7x genome coverage if we assume that the total assembly size is similar to the true genome size, which is expected to be about 1.2 Gb based on published genome sizes of related species (e.g., Alexandre et al. 2023).

Microhaplotype dataset

After trimming, read alignment to the draft genome yielded an average of 4.9 million (M) reads per individual (minimum = 14,045; maximum = 8.6 M). FreeBayes identified 3.4 M raw variants (including SNPs, multi-nucleotide polymorphisms, indels and other complex variants) across the 185 individuals. After filtering, the dataset consisted of a total of 5,033 loci containing 8,990 SNPs across 168 individuals with a maximum of 30% missing data. After thoroughly filtering, this dataset should represent a robust putatively neutral dataset. Average depth per locus and per individual was 23.1 (ranging from 13.2 to 50.2 and 8.6 to 36.2, respectively). Total missing data was 7.5%.

Genetic variation

Most samples from the South Canadian River that were collected in 2012, 2015, and 2017 overlapped across the DAPC space (Figure 1). However, the 2024 collection was segregated along the first DF (x-axis). In contrast, two from the Pecos River (2009) were separated along the second DF. Estimated pairwise F_{ST} values were very low (≤0.003); these values for the Pecos River 2009 and South Canadian River 2024 samples were significantly different from zero.

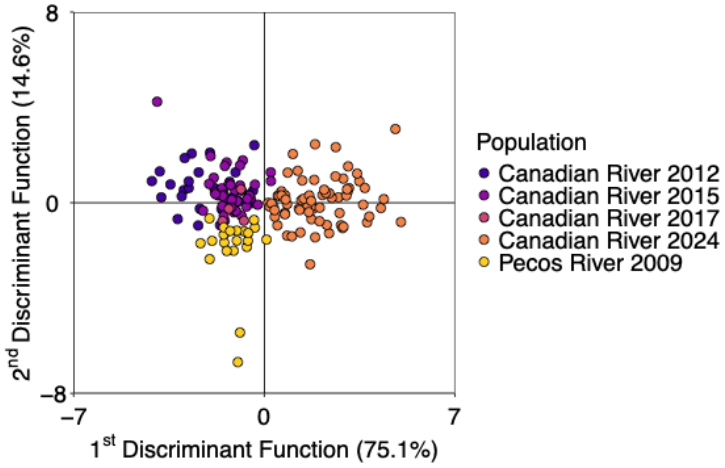

Figure 1. Results of discriminant analysis of principal components (DAPC). The percentage values on the axis labels refers to the variance explained by each discriminant function.

Table 2. Pairwise F_{ST} comparisons between Arkansas River shiner collections (South Canadian River [SCR] and Pecos River [PR]) estimated with complete nextRAD dataset. Top triangular matrix contains F_{ST} estimates and lower matrix contains the p-values. Significant F_{ST} values are highlighted with an asterisk (*). The Pecos River sample is highlighted in grey.

	SCR 2012	**SCR 2015**	**SCR 2017**	**SCR 2024**	**PR 2009**
SCR 2012	-	0	-0.004	0.001*	0.002*
SCR 2015	0.068	-	-0.003	0.001*	0.003*
SCR 2017	0.946	0.898	-	-0.005	-0.003
SCR 2024	0.002	0.006	0.967	-	0.002*
PR 2009	0.001	0.001	0.794	0.001	-

Genetic Diversity

Genetic diversity and inbreeding metrics (A_R, H_O, H_E, and F_{IS}) were very similar across time and not significantly different across collections (Table 3).

Contemporary effective population size

No finite estimates of LD N_e were obtained for the temporal South Canadian River samples. For the Pecos River sample collected in 2009, LD N_e was 1,599 (95% CI 1,400 – 1,863). N_{eV} estimates for the 2012 – 2015 and for the 2015 – 2024 temporal comparisons were 485 (95% CI 367 - 703) and 758 (95% CI 602 – 1,007), respectively.

Table 3. Genetic diversity and inbreeding metrics estimated from Arkansas River shiner collections (South Canadian River [SCR] and Pecos River [PR]) with complete nextRAD dataset, and each subset of 500 loci used to select loci for optimization. Allelic richness (A_R) and inbreeding coefficient (F_{IS}) were estimated with R package diversity. Observed heterozygosity (H_O) and expected heterozygosity (H_E) were estimated with R package poppr. Values between parentheses correspond to 95% confidence intervals (CIs) estimated using 1,000 bootstrap iterations over loci. All CIs for the same metrics estimated from each collection overlap across all datasets. The Pecos River sample is highlighted in grey.

Dataset	Collection	A_R	H_O	H_E	F_{IS}
nextRAD complete (5,033 loci)	SCR 2012	1.91 (1.756 – 1.974)	0.204 (0.12 – 0.238)	0.221 (0.14 – 0.257)	0.069 (0.055 – 0.08)
	SCR 2015	1.97 (1.83 – 2.015)	0.216 (0.142 – 0.243)	0.226 (0.162 – 0.251)	0.035 (0.016 – 0.0.48)
	SCR 2017	1.79 (1.623 – 2.012)	0.238 (0 – 0.313)	0.204 (-0.055 – 0.286)	-0.169 (-0.429 – -0.019)
	SCR 2024	1.92 (1.752 – 1.992)	0.209 (0.177 – 0.245)	0.226 (0.192 – 0.26)	0.076 (0.063 – 0.087)
	PR 2009	1.93 (1.708 – 2.005)	0.206 (0.136 – 0.26)	0.221 (0.15 – 0.275)	0.019 (0.019 – 0.12)
F_{ST}500	SCR 2012	1.71 (1.598 – 1.77)	0.212 (0.201 – 0.295)	0.232 (0.181 – 0.294)	0.076 (0.047 – 0.102)
	SCR 2015	1.77 (1.654 – 1.834)	0.24 (0.231 – 0.314)	0.244 (0.213 – 0.301)	0.031 (-0.018 – 0.035)
	SCR 2017	1.71 (1.536 – 1.866)	0.298 (-0.198 – 0.31)	0.247 (-0.105 – 0.331)	-0.19 (-0.462 – -0.045)
	SCR 2024	1.73 (1.59–1.816)	0.229 (0.141–0.3)	0.249 (0.244–0.31)	0.072 (0.052–0.09)
	PR 2009	1.76 (1.606 – 1.83)	0.227 (0.185 – 0.297)	0.249 (0.194 – 0.302)	0.072 (0.021 – 0.142)
Rdm500	SCR 2012	1.79 (1.658 – 1.858)	0.220 (0.147 – 0.251)	0.239 (0.198 – 0.289)	0.07 (0.052 – 0.085)
	SCR 2015	1.83 (1.7 – 1.904)	0.237 (0.184 – 0.283)	0.242 (0.205 – 0.294)	0.01 (-0.013 – 0.03)
	SCR 2017	1.72 (1.544 – 1.906)	0.253 (-0.139 – 0.38)	0.218 (-0.132 – 0.34)	-0.177 (-0.44 – -0.033)
	SCR 2024	1.79 (1.65 – 1.88)	0.226 (0.192 – 0.256)	0.245 (0.221 – 0.282)	0.071 (0.049 – 0.089)
	PR 2009	1.8 (1.612 – 1.882)	0.224 (0.14 – 0.296)	0.239 (0.183 – 0.316)	0.049 (0 – 0.111)
F_{ST}250+Rdm 250	SCR 2012	1.74 (1.618 – 1.806)	0.214 (0.192 – 0.285)	0.229 (0.162 – 0.272)	0.063 (0.041 – 0.082)
	SCR 2015	1.8 (1.676 – 1.878)	0.237 (0.224 – 0.309)	0.241 (0.207 – 0.299)	0.008 (-0.019 – 0.03)
	SCR 2017	1.72 (1.544 – 1.884)	0.29 (-0.214 – 0.303)	0.24 (-0.111 – 0.316)	-0.185 (-0.443 – -0.045)
	SCR 2024	1.75 (1.616 – 1.828)	0.223 (0.231 – 0.287)	0.241 (0.236 – 0.298)	0.073 (0.053 – 0.09)
	PR 2009	1.77 (1.602 – 1.848)	0.225 (0.191 – 0.309)	0.241 (0.196 – 0.308)	0.06 (0.012 – 0.128)

Loci selection for GT-seq panel optimization

Summary statistics of genetic diversity were estimated per locus using the three subsets of 500 loci selected and compared to the nextRAD_complete dataset to identify a set of markers that tracked temporal changes in genetic diversity (Figure 2). Regardless of the subset, the distribution of values for each metric (A_R, H_O, H_E, and G_{IS}) were essentially the same across all temporal collections (Figure 2A to 2D). Similarly, the same metrics estimated for each temporal collection did not show significant differences among subsets or when compared to the complete nextRAD dataset (Table 3).

Locus-specific F_{ST} distribution estimated from the F_{ST}500 and F_{ST}250+Rdm250 subsets were higher than the F_{ST} estimated from the complete nextRAD dataset (Figure 2E). Locus-specific F_{ST} distribution estimated from the Rdm500 subset was the most similar to the distribution from the complete nextRAD dataset (Figure 2E). Population-level pairwise F_{ST} estimates between all years were relatively small across datasets, but some differences were detected (Table 4). The pairwise F_{ST} estimated from the Rdm500 subset was the most similar to the complete nextRAD

dataset (i.e., all CIs overlapped with the complete nextRAD dataset). The same pattern was not observed for the other two subsets. F_{ST}500 had all CIs higher than the complete nextRAD dataset while for F_{ST}250+Rdm250 only half of the comparisons had overlapping CIs.

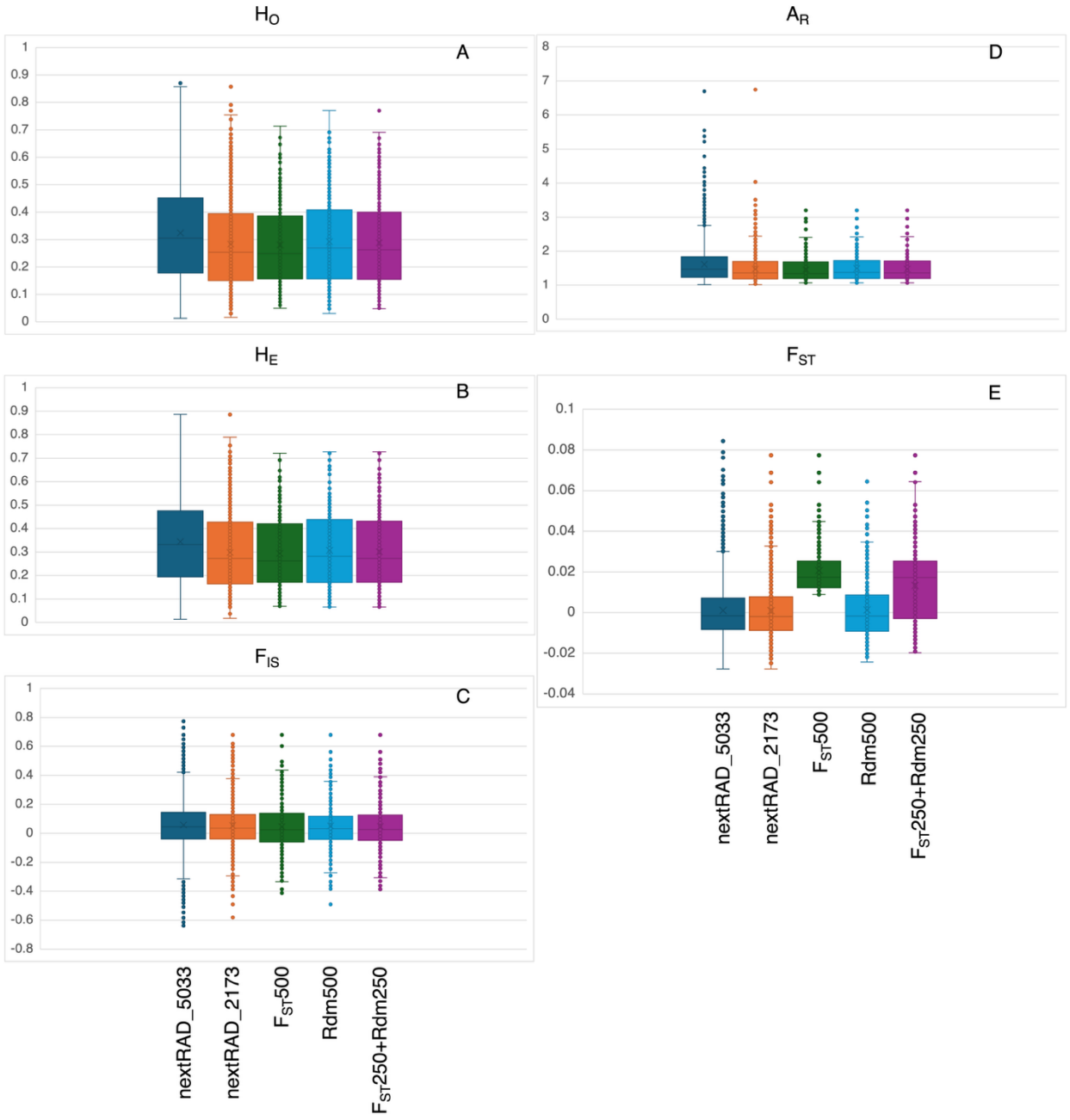

Figure 2. Locus-specific distributions of each diversity metric estimated with the complete nextRAD dataset (nextRAD_5033), the nextRAD dataset including only the loci compatible with the GT-seq protocol (nextRAD_2173), and with each of the subsets of 500 loci used to select a GT-seq panel for optimization (F_{ST}500, Rdm500, and F_{ST}250+Rdm250). A) Observed heterozygosity (H_O); B) Expected heterozygosity (H_E); C) Inbreeding coefficient (F_{IS}); D) Allelic richness (A_R); E) Fixation index (F_{ST}).

Table 4. Pairwise F_{ST} comparisons between Arkansas River shiner collections estimated (South Canadian River [SCR] and Pecos River [PR]) with each of the subsets of 500 loci used to select a GT-seq panel for optimization (F_{ST}500, Rdm500, and F_{ST}250+Rdm250). Confidence intervals (CIs) are between parentheses. For each comparison, the results from F_{ST}500 are on the top row (white background), Rdm500 on the middle row (light gray background), and F_{ST}250+Rdm250 are on the bottom row (dark gray background). CIs that overlap with complete nextRAD dataset are highlighted with an asterisk (*). All CIs estimated with Rdm50 (middle row from each comparison overlap with CIs estimated with complete nextRAD dataset.

	SCR 2012	**SCR 2015**	**SCR 2017**	**SCR River 2024**
SCR 2015	0.017 (0.011 – 0.025)	-	-	-
	0.001* (-0.005 – 0.008)			
	0.011 (0.004 – 0.018)			
SCR 2017	0.049 (0.006 – 0.11)	0.042 (0.004 – 0.103)	-	-
	-0.016* (-0.062 – 0.051)	-0.012* (-0.055 – 0.05)		
	0.018* (-0.025 – 0.078)	0.02* (-0.02 – 0.077)		
SCR 2024	0.019 (0.014 – 0.025)	0.013 (0.009 – 0.017)	0.044 (0.007 – 0.105)	-
	0.002* (-0.003 – 0.008)	0.001* (-0.003 – 0.005)	-0.01* (-0.053 – 0.051)	
	0.012 (0.007 – 0.019)	0.008 (0.004 – 0.013)	0.02* (-0.019 – 0.078)	
PC 2009	0.031 (0.022 – 0.043)	0.024 (0.016 – 0.035)	0.042 (0.001 – 0.11)	0.02 (0.013 – 0.03)
	0.005* (-0.004 – 0.017)	0.004* (-0.004 – 0.015)	-0.008* (-0.054 – 0.059)	0.003* (-0.004 – 0.013)
	0.019 (0.011 – 0.031)	0.016 (0.009 – 0.027)	0.019* (-0.024 – 0.082)	0.012* (0.005 – 0.022)

Discussion

Characterization of genetic diversity

Overall, pairwise F_{ST} and DAPC suggest small but significant changes in allele frequencies over time in the South Canadian River, likely driven by genetic drift. Genetic drift has previously been identified as a major factor influencing temporal genetic variation in several pelagophilic species in southwestern USA (Osborne at al. 2021; Osborne et al. 2023), including Arkansas River shiver (Osborne at al. 2021). Despite those changes in allele frequencies across time, average genome-wide diversity remained stable. Similar results were obtained in Rio Grande silvery minnow (Osborne et al., 2023) and peppered chub (Caeiro Dias et al., unpublish data).

Loci selection for GT-seq panel optimization

In this study, we are developing a GT-seq panel for genetic monitoring of Arkansas River shiner. We used temporal data from archived collections from South Canadian River spanning five years and an additional collection from the introduced population in Pecos from which we identified 3,055 variable loci across the genome, including microhaplotypes and single biallelic SNPs. By comparing the results of genetic diversity and inbreeding, including H_E, H_O, A_R, F_{IS}, and F_{ST} from three subsets of 500 loci to the complete nextRAD dataset, we identified the subset Rdm500 as the dataset providing more consistent results with those obtained from the complete nextRAD dataset. While we did not find significant differences in H_E, H_O, A_R, and F_{IS} between any subset of 500 loci and the complete nextRAD dataset, the same was not true for F_{ST}. The

subset Rdm500 was the only one without significant differences across temporal collections when compared to the results from the complete nextRAD dataset. As such, this was the subset selected for GT-seq optimization.

Future work

The reference genome in fasta format, the VCF file containing the 8,990 filtered SNPs, and a BED file containing the location of each nextRAD locus included in the Rdm500 subset were sent to GTseek LLC (https://gtseek.com/) for locus-specific primer design and panel optimization. Panel optimization is currently on-going. We expect to receive the primers from loci included in the optimized panel in the next three months. After that we will use the optimized panel to genotype Arkansas River shiner archived collections, including other temporal samples not included in the nextRAD sequencing and other samples from the same years to increase sample size where possible. Once all archived collections are genotyped, the same metrics estimated here will be re-estimated to better understand temporal changes in genetic diversity and effective population size. If justifiable, additional analysis may be performed including additional inbreeding metrics (Osborne et al. 2023; Osborne et al. 2025).

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 contract management and editing assistance.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389-3402.

Bilton TP, McEwan JC, Clarke SM, Brauning R, van Stijn TC, Rowe SJ, Dodds KG (2018) Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics* 209(2):389-400.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114-2120.

Breiman L (2001) Random Forests. *Machine Learning* 45:5-32.

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.

Caeiro-Dias G, Osborne MJ, Turner TF (In Press) Time is of the essence: using archived samples in the development of a GT-seq panel to preserve continuity of ongoing genetic monitoring. *PeerJ* 13:e20726

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10(1):421.

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.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 18(2):170-175.

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.

Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P (2022) A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Computational Biology* 18(5):e1009123.

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.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24(11):1403-1405.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27(21):3070-3071.

Kamvar ZN, Brooks JC, Grünwald, NJ (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics* 6:208.

Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:281.

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

Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News*, 2(3):18-22.

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.

Mangiafico SS (2025) rcompanion: Functions to Support Extension Education Program Evaluation. Rutgers Cooperative Extension, New Brunswick, New Jersey. version 2.5.0.

Manni M, Berkeley MR, Seppey M, Zdobnov EM (2021) BUSCO: Assessing genomic data quality and beyond. *Current Protocols* 1:323.

Meirmans PG (2020) GENODIVE version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. *Molecular Ecology Resources* 20:1126-1131.

Nishimura O, Hara Y, Kuraku S (2017) gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* 33(22):3635-3637.
Osborne MJ, Caeiro-Dias G, Turner TF (2023) Transitioning from microsatellites to SNP-based microhaplotypes in genetic monitoring programmes: Lessons from paired data spanning 20 years. *Molecular Ecology* 32(2):316-334.
Osborne MJ, Caeiro-Dias G, Turner TF (2025) Manmade barriers drive temporal and spatial trends of genetic diversity and effective population size in a riverine fish. *Authorea.* DOI: 10.22541/au.173640622.26831605/v1
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.
Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 26(3):419-420.
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.
Posit team (2022). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/
Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2:431.
R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www/R-project.org
Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features, *Bioinformatics* 26(6):841-842.
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.
The Broad Institute (2025) Picard Tools. https://broadinstitute.github.io/picard/
Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Research* 40(15):e115-e115.
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.
Waples, R.S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121(2):379-391.

Waples, R.S., Do, C.H. (2010). Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications* 3(3):244-62.

Waples, R.S., Do, C.H. (2008). LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* 8(4):753-756.

Willis SC, Hollenbeck CM, Puritz JB, Gold JR, Portnoy DS (2017) Haplotyping RAD loci: an efficient method to filter paralogs an account for physical linkage. *Molecular Ecology Resources* 17(5):955-965.