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

Progress report for New Mexico Department of Wildlife
Share with Wildlife
Submitted for period ending 30th June 2026

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 Wildlife
1 Wildlife Way
Santa Fe, NM 87507

Table of Contents

Background

The Arkansas River Shiner (*Notropis girardi*) is an endemic pelagic broadcast-spawning (gametes are released into the water column) Leuciscid minnow 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 the Arkansas River Shiner is listed as threatened under the Endangered Species Act (U.S. Fish and Wildlife Service 1998). The recovery plan for the Arkansas River Shiner was published in 2024 (U.S. Fish and Wildlife Service 2024). Its current range includes two distinct fragments of the South Canadian River (275 kms between Ute Lake in New Mexico [NM] and Lake Meredith in Texas [TX]) and an 829-km reach from Lake Meredith to Lake Eufaula in Oklahoma [OK]). The Arkansas River Shiner also occupied the Cimarron River until around 1992, the Arkansas River and Salt Fork of the Arkansas until the mid to late 1980's, and the North Canadian River until approximately 1992 (U.S. Fish and Wildlife Service 2024). Like other members of the pelagic spawning reproductive guild, the Arkansas River Shiner is particularly sensistive to modifications to the flow regime, fragmentation imposed for diversion dams and reservoirs, and habitat changes that, in concert, reduce longitudinal and lateral connectivity (Wilde 2002; Durham and Wilde 2006). Likewise, groundwater depletions negatively impact streamflow in rivers occupied by the Arkansas River Shiner (Falke et al. 2010). The recovery plan for the Arkansas River Shiner includes reintroduction to formerly occupied sites and establishment/maintenance of a captive breeding program. Robust baseline genetic data can be used to inform these endeavours.

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 heavily relied on highly polymorphic microsatellite markers to obtain diversity and N_e estimates. Six temporal collections (2009, 2012, 2014, 2015, 2017, and 2019) of Arkansas River Shiner from the Canadian River were used to document the spatial distribution and amount of genetic diversity in this population and provided estimates of contemporary effective population size (N_e) using microsatellites and mitochondrial DNA markers (Osborne et al. 2021). Results included high and stable genetic diversity in the mtDNA ND4 gene (94 unique haplotypes identified across 941 individuals). Heterozygosity assessed with microsatellites was also moderate to high (H_E= 0.766-0.787; H_O=0.653-0.731) and across the time series, contemporary N_e was indistinguishable from infinity (indicative of very large N_e). In 2015 and 2017, N_e was estimated to be ~2400.

Rapidly-changing technology has led to a shift toward assaying genetic 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 sufficient SNPs in a sample to surpass the advantages 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, like 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 using multiplex PCR and NGS can be more advantageous than alternative methods. Genotyping-in-Thousands by sequencing (GT-seq) is a method of targeted SNP genotyping that uses multiplex PCR amplicon sequencing (Campbell et al. 2015). This method allows simultaneous amplification of hundreds of targeted genetic loci while barcoding of individuals allows up to a thousand 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. After a GT-seq panel has been 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, so the aim of this project was to develop these resources.

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 a total of 110 archived samples collected in 2009, 2012, 2015, and 2017. These samples represent multiple South Canadian River localities. These samples were sent to SNPsaurus for reduced representation sequencing using a nextRAD approach. A total of 5,033 loci including 8,990 SNPs (microhaplotypes) were identified from this data, representing 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. We used several subsets of 500 loci to test the power of

the loci to obtain estimates of genetic diversity that are comparable to the complete SNP dataset. A set of 500 loci was selected and sent to GTseek, LLC in Twin Falls, ID for PCR multiplex optimization. Following optimization, 272 loci were retained in the panel. Target SNPs in 14 loci were monomorphic or not scored in the samples used for optimization. We conducted a preliminary analysis retaining 258 loci using 1 SNP per locus to compare genetic diversity, LD-based N_e, and divergence estimates obtained to the complete nextRAD dataset. The complete panel of 272 loci will be used to genotype additional samples in order to identify truly monomorphic loci and to evaluate genotyping consistency between SNPs in the GT-seq panels and the same loci in the nextRAD data.

(iv) Use the GT-seq panel to run a subset of archived temporal samples and new material collected by NMDW personnel. This data will be used to estimate genetic diversity and effective population size using methods described in Osborne et al. (2021).

Status: *In Progress*.

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 individual that was 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 the transfer of individuals (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 was available for library preparation. A DNA sample was sent to University of California Davis 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 the Arkansas River Shiner individual were received from US 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 contig N50 value (i.e., the sequence length for the longest contigs that 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; 233 genes).

Sampling and reduced-representation sequencing

For nextRAD sequencing, we purified previously-isolated DNA using Zymo® DNA clean and concentrator kits following manufacturer's directions from samples collected in 2009 from the Pecos River (New Mexico) and 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) with enough high-quality DNA were selected for reduced-representation sequencing. Those samples were sent to SNPsaurus (http://snpsaurus.com). for sequencing. SNPsaurus employs a proprietary Nextera-tagmented reductively-amplified DNA sequencing method (nextRAD-seq) that reduces the proportion of genome to be sequenced by adding selective primers to DNA fragments as described by Russello et al. (2015). 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 series single-lane system to obtain 150 base pair (bp) long paired-end reads.

Identification of SNPs

Raw DNA sequences (i.e., 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 ambiguous (N) base calls. Then each read was screened with a 5-base sliding window and cut when the average quality per base dropped below 10. Resulting reads smaller than 60 bases 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 (sequence alignment map) files to BAM (binary alignment map) files. Picard Tools v. 3.1.0 (The Broad Institute 2023) was used to add reading groups (RG), 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.

To remove erroneous or potentially erroneous variants, we used VCFtools v. 0.1.16 (Danecek et al. 2011) to filter out variants with a mean depth of coverage lower than 20 and higher than 200, minor allele count of 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 at heterozygous positions, the alternative allele had a coverage lower than 20% or higher than 80% compared with the reference allele. This was done 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 than 0.99 were not removed to account for fixed alleles. Additionally, the locus was removed if the quality sum of the reference or the alternative allele was zero, which removes positions with spurious heterozygous genotype calls. Then, we removed loci if the ratio between the mean mapping quality of the alternative and reference alleles was lower than 0.25 or higher than 1.75 because loci from the same genomic location should not have large discrepancies between the mapping qualities of two alleles. Furthermore, loci with quality scores of less than half of the total depth were excluded because excessive depth inflates FreeBayes quality scores. Of the remaining loci, the average depth and standard deviation across all individuals were calculated. Loci with depths greater than the average depth plus one standard deviation were removed if the quality score was less than two times the depth. Finally, this script removed loci with a mean depth across individuals of greater than two times the mode (49) that corresponded approximately to the 95th percentile of mean depth. Subsequently, potential 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 than 50% of the populations (here, each year was considered a separate 'population') and with an HWE p-value 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 with the *stats* package implemented in R and the read depth was estimated with VCFtools. This 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 also eliminates possible paralogous loci while neutralizing physical linkage without losing data (Willis et al. 2017). Haplotyping was performed with the *rad_haplotyper.pl* pearl script (https://github.com/chollenbeck/rad_haplotyper), excluding microhaplotypes if they were considered paralogs in at least five individuals and if they were missing from more than 30% of individuals.

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 collection from 2017 was not included due to its small sample size. 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. 2026). Estimations of LD were performed on SNP data using the SNP of each microhaplotype that had the 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. 2026). If loci were found in multiple significant LD pairs, the loci that appeared in the highest number of comparisons were discarded in order to retain the maximum number of loci possible. 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 four temporal samples (p-value < 0.05). The resulting dataset should represent a robust genome-wide neutral SNP dataset that is suitable for primer design for the GT-seq panel development.

Genetic variation

Genetic variation between populations in the Canadian and Pecos Rivers and across time in the 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 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. This was preferred 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, allele frequencies centered but not scaled, retaining all principal components (PCs) and discriminant functions (DFs), and keeping other options at the default setting. The *a-score* method was used to select the optimal number of principal components to retain for the final DAPC; the maximum number of PCs and all DFs were used, and the other options were set as default. The final DAPC was performed using the optimal number of PCs and two DFs, and keeping the other default options.

Next, we assessed genetic differences between all collections by estimating pairwise 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

SNP-containing loci (5,033) were used to estimate N_e for the Arkansas River Shiner population using the single sample method based on linkage disequilibrium (LD N_e) implemented in NeEstimator. We also estimated N_{eV} for the Arkansas River Shiner using methods of Nei and Tajima (1981). 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). N_{eV} measures the change in allele frequencies between two samples caused by genetic drift. The 2017 sample was not used for N_{eV} estimates because of its small sample size. We also did not estimate N_{eV} for the one Pecos River sample because this method requires at least two temporal samples.

Loci selection for GT-seq panel optimization

To select loci for GT-seq panel optimization, we used the approach that was described in Caeiro-Dias et al. (2026). 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. 2026). 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 (T_m) of 60ºC, GC content of 50%, and fewer than four consecutive repeat motifs (PolyX). When possible, we

allowed design of up to 5 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 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. 2026). 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. 2026) 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 retained loci, we selected a subset of 500 for panel optimization. To select those 500 loci, we followed a strategy similar to that used by Caeiro-Dias et al. (2026). Based on preliminary results and results from Caeiro-Dias et al. (2026), 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); and 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}), 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 CIs overlapped, we considered that estimates were not significantly different between datasets.

GT-seq panel optimization and testing

A list of 500 loci was sent to GTseek, LLC for panel optimization. GTseek designed primers then tested the efficacy of those primers on a set of Arkansas River Shiner samples (n=96). These samples included 4 temporal collections (2009, 2012-2015, 2017, and 2024), and also included a subset of 32 samples used for nextRAD-seq and SNP discovery to allow evaluation of genotyping consistency across methods (nextRAD & GT-seq).

After primer pool optimization, the primers for the optimized GT-seq panel were received from GTseek. Monomorphic loci or loci that were not scored in at least 70% of the samples were not used for further analysis. We then evaluated genotype consistency across methods and calculated genetic diversity metrics, LD N_e and pairwise F_{ST} as described above using the retained loci.

Results

Genome sequencing

A total of 8,040,297 raw HiFi reads were generated, containing 133.7 Giga bases (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 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 current understanding of the genomes of other related species (e.g., Alexandre et al. 2023).

Microhaplotype dataset - nextRAD

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. 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 - nextRAD

Most samples from the Canadian River 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) while samples from Pecos River (2009) were separated along the second DF. Estimated pairwise F_{ST} values were very low (≤0.003) but the Pecos River 2009 and the Canadian River 2024 collections were significantly different from zero. Overall, pairwise F_{ST} and DAPC suggest small but significant changes in allele frequencies.

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.

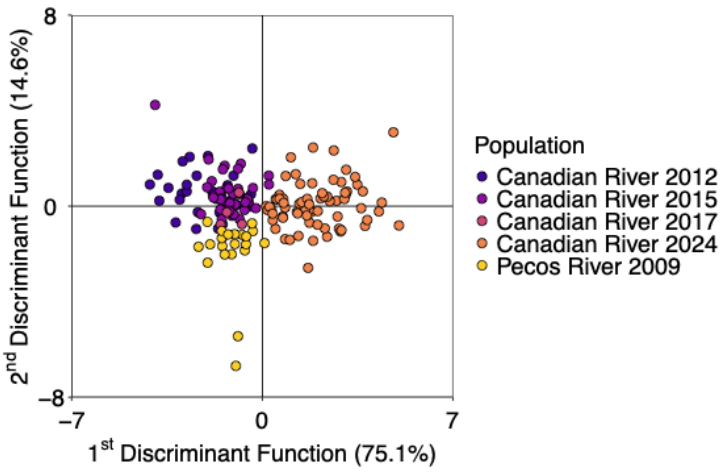

Table 1. Pairwise F_{ST} comparisons between Arkansas River Shiner collections estimated with complete nextRAD dataset. Top triangular matrix contains F_{ST} estimates and lower matrix contains the p-values. Significant F_{ST} values are highlighted with an asterisk (*). The Pecos River sample is highlighted in grey.

	Canadian River 2012	**Canadian River 2015**	**Canadian River 2017**	**Canadian River 2024**	**Pecos River 2009**
Canadian River 2012	-	0	-0.004	0.001*	0.002*
Canadian River 2015	0.068	-	-0.003	0.001*	0.003*
Canadian River 2017	0.946	0.898	-	-0.005	-0.003
Canadian River 2024	0.002	0.006	0.967	-	0.002*
Pecos River 2009	0.001	0.001	0.794	0.001*	-

Table 2. Genetic diversity and inbreeding metrics estimated with complete nextRAD dataset and each of the subsets of 500 loci used to select a GT-seq panel for optimization. Allelic richness (A_R) and inbreeding coefficient (F_{IS}) were estimated with the R package diversity. Observed heterozygosity (H_O) and expected heterozygosity (H_E) were estimated with the 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 (5033 loci)	Canadian River 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)
	Canadian River 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)
	Canadian River 2017	1.79 (1.623 – 2.012)	0.238 (0 – 0.313)	0.204 (-0.055 – 0.286)	-0.169 (-0.429 – -0.019)
	Canadian River 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)
	Pecos River 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	Canadian River 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)
	Canadian River 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)
	Canadian River 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)
	Canadian River 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)
	Pecos River 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	Canadian River 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)
	Canadian River 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)
	Canadian River 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)
	Canadian River 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)
	Pecos River 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 + Rdm250	Canadian River 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)
	Canadian River 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)
	Canadian River 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)

	Canadian River 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)
	Pecos River 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)

Genetic diversity and contemporary effective population size - nextRAD dataset

Genetic diversity metrics based on microhaplotypes (A_R, H_O, H_E, and F_{IS}) were very similar across time and not significantly different across collections. No finite estimates of LD N_e were obtained for the temporal Canadian River samples. For the Pecos River sample collected in 2009, LD N_e was 1599 (95% CI 1400 - 1863). 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 - 1007) respectively.

Loci selection for GT-seq panel optimization

Summary statistics of genetic diversity per locus were estimated from three subsets of 500 loci with primers and compared to the nextRAD_complete dataset to identify a set of markers that tracked temporal changes in genetic diversity. Regardless of the subset, the distribution of values for each metric (A_R, H_O, H_E, and F_{IS}) were essentially the same across all temporal collections (Figure 2). When compared to the nextRAD_complete dataset, the distributions of genetic diversity values across time were similar to the complete dataset (Table 2). 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 2).

Locus-specific F_{ST} estimated from F_{ST}500 and F_{ST}250+Rdm250 showed higher values compared to the complete nextRAD dataset, but the Rdm500 provided similar results to the complete nextRAD dataset (Figure 2b). Pairwise F_{ST} estimates between all years were relatively small across datasets, but some differences were detected. Similarly, the Rdm500 provided the most similar results to the complete nextRAD dataset (i.e., all CIs overlapped with complete nextRAD dataset). There were differences between the other two subsets and the complete nextRAD datasets. Specifically, F_{ST}500 had all CIs higher than the complete nextRAD dataset; in contrast, only half of the comparisons had overlapping CIs for FST250+Rdm250.

Figure 2. Locus-specific distributions of each diversity metric (allelic richness [A_R], observed heterozygosity [H_O], expected heterozygosity [H_E], and inbreeding coefficient [F_{IS}]) and F_{ST} 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).

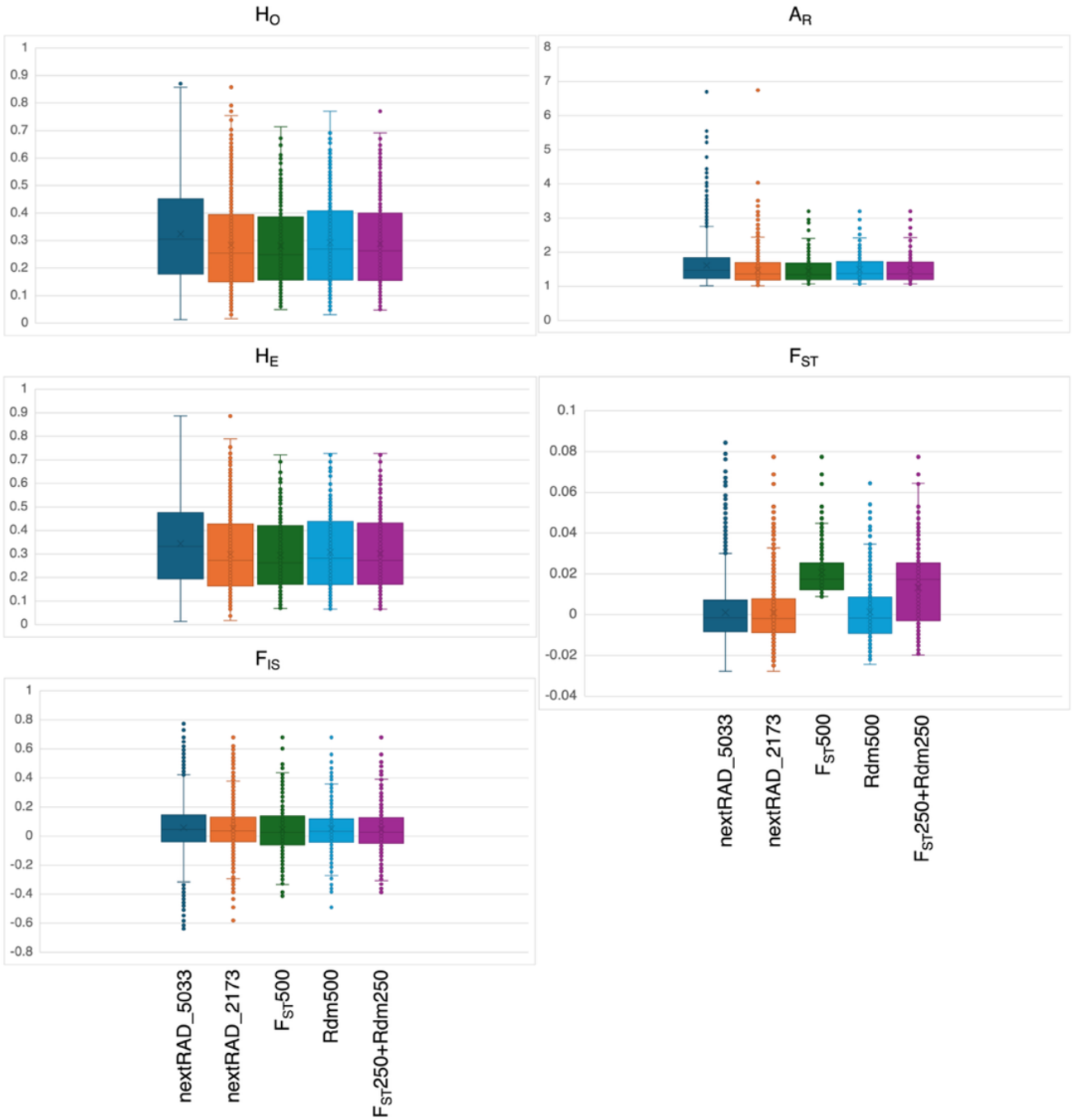

Table 3. Pairwise F_{ST} comparisons between Arkansas River Shiner collections estimated with each of the subsets of 500 loci used to select a GT-seq panel for optimization (F_{ST}500, Rdm500, and F_{ST}250+Rdm250). Confidence intervals (CIs) are between parentheses. For each comparison, the results from F_{ST}500 are on the top row, Rdm500 on the middle row, and F_{ST}250+Rdm250 are on the bottom row. CIs that overlap with the complete nextRAD dataset are highlighted with an asterisk (*). All CIs estimated with Rdm50 (middle row) from each comparison, overlap with the CIs estimated with the complete nextRAD dataset.

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

GT-seq panel optimization and testing

After primer pool optimization, 272 primer pairs for the optimized GT-seq panel were received from GTseek, in addition to the genotypes from the 96 testing samples. Across those samples, we excluded 14 loci due to lack of variation (i.e., they were monomorphic) or were not scored in at least 70% of the samples. This resulted in 258 loci containing 338 SNP. One SNP was selected at random and used for estimation of genetic diversity metrics, LD N_e and pairwise F_{ST}. Genotype accuracy between methods was estimated for each SNP; on average, accuracy was 80% which is a relatively low accuracy. Because we used only 32 individuals that were common to both nextRAD and GT-seq datasets, genotyping errors in few individuals will result in relatively high proportions of inaccuracies between datasets. This is especially true if some individuals have some level of missing data. To improve these estimates, we will use the panel to genotype additional samples included in the original nextRAD experiment.

Table 5. Genetic diversity calculated using the GT-seq panel containing 258 loci. Sample size (n), expected heterozygosity (H_E), observed heterozygosity (H_O), average inbreeding co-efficent (F_{IS}), allelic richness (A_R) and linkage disequilibrium effective population size (LD N_e) estimates are provided. Lower and upper 95% confidence intervals for LD N_e are provided in the parentheses below point estimates.

Sample	n	H_E	H_O	F_{IS}	A_R	LD N_e
Pecos River 2009	24	0.213	0.195	0.059	1.79	1,538 (212 - ∞)
Canadian River 2012-2015	14	0.233	0.210	0.058	1.80	∞ (108 - ∞)
Canadian River 2017	32	0.235	0.215	0.070	1.85	856 (298 - ∞)
Canadian River 2024	26	0.237	0.215	0.068	1.83	∞ (128 - ∞)

Table 6. Pairwise F_{ST} comparisons between Arkansas River Shiner collections estimated with GT-seq panel containing 258 loci. Top triangular matrix contains F_{ST} estimates and lower matrix contains 95% confidence intervals. Following Bonferroni correction, no values of F_{ST} were significantly differed from zero. The Pecos River sample is highlighted in grey.

	2009 Pecos	**2012-2015 Canadian**	**2017 Canadian**	**2024 Canadian**
2009 Pecos	-	-0.0015	0.0038	0.0054
2012-2015 Canadian	-0.0203 - 0.0231	-	-0.0054	-0.0014
2017 Canadian	-0.006 - 0.0164	-0.0222 - 0.0197	-	-0.0025
2024 Canadian	-0.008 - 0.0197	-0.021 - 0.0238	-0.0132 - 0.0106	-

Preliminary genetic diversity and effective size using the optimized GT-seq panel
Expected heterozygosity ranged from 0.213 to 0.237, observed heterozygosity ranged from 0.195 to 0.215 and allelic richness varied from 1.76 to 1.85 (Table 5). Across metrics, values were very similar to the nextRAD complete dataset. Using the GT-seq panel, two finite values (LD N_e = 856, LD N_e = 1,538; Table 5) of LD N_e were obtained for 2017 Canadian River and 2009 Pecos River sample respectively. Pairwise values of F_{ST} among the temporal Canadian River and Pecos River samples were small and not significantly different zero (Table 6).

The final component of this project is to genotype a subset of archived samples and recently collected samples. This data will be used to assess current genetic diversity and effective population size of the Arkansas River Shiner and to compare those results to the broader time series using larger sample sizes.

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 (2026) 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.

Durham BW, Wilde G (2006) Influence of stream discharge on reproductive success of a prairie stream fish assemblage. *Transactions of the American Fisheries Society*, 135:1644–1653.

Falke JA, Fausch KD, Magelky R, Aldred A, Durnford DS, Riley LK, Oad (2010) The role of groundwater pumping and drought in shaping ecological futures for stream fishes in a dryland river basin of the western Great Plains, USA: Ecohydrology 4:682–697.

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.

Narum SR, Banks M, Beacham TD, Bellinger MR, Campbell MR, Dekoning J, ... & Garza JC (2008). Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, 17(15):3464-3477.

Nishimura O, Hara Y, Kuraku S (2017) gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* 33(22):3635-3637.

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.

U. S. Fish and Wildlife Service (2024) Recovery Plan for the Arkansas River Shiner (*Notropis girardi*). U.S. Fish and Wildlife Service, Oklahoma Ecological Services Field Office, Tulsa, OK. 41 pages.

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121(2):379-391.

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

Wilde GR (2002) Threatened fishes of the world—*Notropis girardi* Hubbs and Ortenburger, 1929 (Cyprinidae): *Environmental Biology of Fishes*, 65: 98–98.

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.