Development of a Genotyping-in-Thousands by sequencing panel for genetic monitoring of peppered chub (*Macrhybopsis tetranema*)

Submitted by:

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

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

Progress Report submitted for period ending June 30th 2024

Introduction

Peppered chub (*Macrhybopsis tetranema*) is a small-bodied and short-lived species belonging to the pelagic-broadcast spawning reproductive guild (Platania and Altenbach, 1998; Perkin and Gido, 2011). Like all the members of this guild found across the North American Great Plains (Dodds *et al.*, 2004; e.g., Dudley and Platania, 2007), peppered chub is adversely affected by anthropogenic changes to rivers (e.g., fragmentation, altered flow regimes, habitat degradation). Survival and reproductive success of this species have been linked to river discharge (Wilde and Durham, 2008) and connectivity that preserves source-sink dynamics (Luttrell *et al.*, 1999; Perkin and Gido, 2011). Historically, peppered chub was found in the upper Arkansas River Basin in parts of Colorado (CO), Kansas (KS), New Mexico (NM), Oklahoma (OK), and Texas (TX). Intensive surveys in 2011 and 2013 recorded declines in the Ninnescah and Arkansas Rivers in KS, and sampling in 2015 documented probable extirpation of peppered chub from these rivers due to extensive regional drought from 2011 to 2013 (Perkin *et al.*, 2015; Pennock and Gido, 2017). Peppered chub is now extirpated from >94% of its historic range with only one remaining population inhabiting 218 km of the South Canadian River between Ute Lake (NM) and Lake Meredith (TX). The South Canadian River population of peppered chub has been in decline since impoundment of the South Canadian River by Ute Reservoir that resulted in a 49% reduction in mean annual discharge (Wilde and Durham, 2008). The restricted range of peppered chub makes the species extremely vulnerable to extinction through stochastic environmental events (e.g., drought) and/or demographic factors (e.g., recruitment failure, mortality caused by disease). As such, this species was listed as an endangered species in 2022 and 1,719 river kms were proposed as critical habitat (U. S. Fish and Wildlife Service, 2022).

Genetic monitoring is an important component of conservation and management efforts for imperiled species. This type of monitoring quantifies temporal changes in population genetic diversity and genetic estimates of effective population size over contemporary timescales (Schwartz *et al.*, 2007). These parameters are important to measure because they provide insight into the long-term adaptive potential and extinction risk of species that cannot be obtained solely with traditional demographic monitoring. Over the last ten years, nine neutrally evolving microsatellite loci were used to obtain empirical measurements of genetic diversity and contemporary effective population size (N_e) for peppered chub (Osborne *et al.*, 2021). Microsatellites were used due to their high variability and because they can be employed with minimal startup costs. Alternative molecular markers to microsatellites that can also be used to obtain that information are single nucleotide polymorphisms (SNPs). These are typically biallelic and have inherently lower resolution power when compared to the multi-allelic microsatellites. However, SNPs represent the most widespread source of variation within genomes (Brumfield *et al.*, 2003) and with the development of increasingly fast and inexpensive high-throughput Next Generation Sequencing (NGS) methods, it is now easy to identify enough SNPs to overcome the advantages of 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). Moreover, genotyping SNPs on large numbers of individuals is more cost- and labor-effective (after protocols are optimized for target species) and genotyping error rates are lower. In addition, SNP genotyping from reduced representation sequencing methods involves sequencing of smaller fragments of DNA, so it can be effective even when DNA is limiting or degraded. Finally, SNP genotyping is more easily standardized across laboratories compared to microsatellites and hence can be used by multiple facilities to produce comparable results. With current technology, these advantages make SNPs more powerful genetic markers for genetic monitoring as compared to microsatellites.

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, but the loci obtained from independent genomic library preparations may not always be consistent. When the number of loci needed to be genotyped is relatively small (a few hundred) and the number of samples is high (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). Once a GT-seq panel is developed for the target species, the method provides a cost-effective and efficient means of monitoring genetic variation and genetic effective population size estimated from hundreds of SNPs.

Here we report the progress on the discovery of genetic variants and identification of SNPs using as reference a new and more complete peppered chub draft genome; primer design to develop a GT-seq panel for peppered chub; loci selection; and optimization of the GT-seq panel that will be used for annual genetic monitoring of the wild and captive peppered chub populations.

Main tasks associated with this project are:

1. Discovery of genetic variants using the most updated and more complete draft genome as reference.
   Status: completed.

2. Variant filtering to remove erroneous or potentially erroneous variants and identification of suitable SNPs for population genomic analysis.
   Status: completed.

3. Design PCR primers for genomic loci containing the previously identified SNPs.
   Status: completed.

4. Selection of about 500 loci from those with primers succesfuly designed to start GT-seq panle optimization.
   Status: completed.

5. Optimization of PCR multiplex in order to retain in the final GT-seq panel of about 300 loci (i.e., PCR multiplex optimized for ~300 pairs of primers) from the previous pool of ~500.
   Status: in progress.

Methods

Identification of SNPs

Prior to this project, 189 samples were provided by New Mexico Department of Game and Fish and U.S. Fish and Wildlife Service from the South Canadian River between Ute Lake (NM) and Lake Meredith (TX) from 2015 to 2020. These were sequenced at SNPsaurus using a nextRAD sequencing protocol following Russello *et al.* (2015). The raw reads were mapped against the most updated version of a peppered chub draft genome developed in our laboratory (assembled to a scaffold level) with Bowtie version 2.4.2 (Langmead and Salzberg 2012) using the 'local alignment' and default 'very sensitive' options. Successfully aligned reads were filtered with Samtools v. 1.16 (Li *et al*., 2009; Danecek *et al*., 2021) to remove reads with mapping quality lower than 20. Before variant calling, we used Picard tools v. 2.20.8 (Broad Institute 2019; https://broadinstitute.github.io/picard/) to add read group (RG) flags to bam files. Genetic variants were identified using FreeBayes v. 1.3.6 (Garrison and Marth 2012) on genomic intervals with at least 150 bp of depth of coverage across all individuals. Raw variants were kept if base quality was at least 5 and a maximum of the best 10 from each alignment were kept, ranked by sum of base quality score.

To remove erroneous or potentially erroneous variants, we applied an extensive computational filtering. Using VCFtools v. 0.1.16 (Danecek *et al*., 2011) we started by filtering out variants with mean depth of coverage lower than 20 and higher than 100, with minor allele count less than three, with minor allele frequency lower than 2%, with genotype depth of coverage lower than five, and with genotype quality lower than 20. Multi-nucleotide states were decomposed into single variants with vcflib (https://github.com/ekg/vcflib) and VCFtools was used to filter out nucleotide insertion and deletion and to retain only the biallelic SNPs. The dataset was then filtered by missing data, keeping SNPs present at least in 80% of samples and removing individuals with more than 30% missing data. Afterward, the bash script *dDocent_filters* (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) was used to filter SNPs based on allelic balance at heterozygous genotypes, strand representation, quality vs depth. First, loci were removed if at heterozygous positions, the alternate allele had a coverage lower than 20% or higher than 80% compared with the reference allele, 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 to account for fixed alleles. Additionally, if the quality sum of the reference or alternate allele was zero, the locus was removed. This procedure removes positions with spurious heterozygous genotype calls. Loci with the ratio between the mean mapping quality of the alternate and reference allele lower than 0.9 or higher than 1.05 were also removed, because loci from the same genomic location should have large discrepancies between mapping qualities of two alleles. Furthermore, loci with quality scores less than half of the total depth were excluded because excessive depth inflates quality scores when using FreeBayes. Of the remaining loci, the average depth and standard deviation across all individuals was calculated. Loci with depth 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 greater than two times the mode (98) 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 ‘population’) and with an HWE p-value lower than 0.001 were removed. We further filtered out potentially incorrectly assembled paralogous loci that exhibit a large variation in read depth across all individuals. Standard deviation was estimated with package *stats* implemented in R v. 4.2.1 (R Core Team 2022) and read depth with VCFtools. An additional filtering based on missing data per locus (keeping loci present in 80% of individuals) was applied again at this point. The remaining SNPs were 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 was performed with the *rad_haplotyper.pl* pearl script (Willis et al. 2017; https://github.com/chollenbeck/rad_haplotyper). Microhaplotypes were then excluded when considered paralogs in at least five individuals and when 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 single ‘population’. 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 2021). Estimations of LD were performed on SNP data using the SNP of each microhaplotype with higher minimum allele frequency. If a SNP in LD was removed, 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. Loci were considered to be deviating from HWE and to be in LD if tests were significant across the six temporal samples (p-value < 0.05). In both cases, if loci with significant chi-square values appeared in multiple pairs, the loci that appeared in the highest number of comparisons were discarded to keep the maximum number of loci possible. In the remainder of instances, one locus from each pair was discarded randomly. The resulting dataset should represent a robust genome-wide neutral SNP dataset.

Primer design for GT-seq and loci used for panel optimization

To facilitate primer design, the loci containing the filtered SNPs were filtered based on the SNP positions within each locus sequence. Only loci with at least 33 bp before the first and after the last SNP were retained. The first and last 25 bp allow for sufficient flanking regions free of variable positions for primer design, while the other 8 bp ensures that primers were not designed in close proximity to the first and last SNP on the sequence. For loci with multiple SNPs that were discarded after applying these initial filters, we removed the SNP closer to the edge of the locus and applied the same filters to potentially retain the remaining SNPs. This step was performed iteratively, until all of the remaining SNPs were either discarded or retained. Then, loci that would result in sequences longer than 150 bp were removed because this is the maximum length permitted by the sequencing approach employed for GT-seq. Resulting loci were used for primer design. Using the draft peppered chub 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%, and fewer than four consecutive repeat motifs (PolyX). When possible, we allowed design of up to 5 primer pairs for each locus. For each locus, the best pair was mapped against the peppered chub draft genome using the *blastn* program (Altschul *et al.*, 1997) with the *blastn-short* task implemented in BLAST+ v.

2.9.0 (Camacho *et al.*, 2009). If at least one primer matched one or more off-target sites with 100% coverage and identity, that pair was discarded. For those cases, the next best pair was mapped on the draft genome as previously described and the process was repeated until a primer pair mapped only to the target locus or until no primer pairs remained.

Research has shown that approximately 300 amplicons is a reasonable number to optimize panel performance during library construction (Beacham *et al.*, 2018; McKinney *et al.*, 2018). Previous studies also suggest that choosing loci with greater genetic differentiation (e.g., F_{ST}) should maximize accuracy for genetic assignment analysis (Ackerman *et al.*, 2011; Storer *et al.*, 2012). Furthermore, we found in a previous GT-seq panel optimization for another species that selecting half of the loci with higher F_{ST} and selecting another half at random performed better for genetic monitoring (unpublished data). As such, the goal was to select 500 loci that reflected F_{ST} values found in the complete dataset (2804 loci) from the pool of loci with successful primers designed to start the GT-seq library optimization. Such optimization consists of several rounds of library preparation and sequencing to assess the performance of primers. Primers identified as problematic in the PCR multiplex (primers involved in high proportion of primer interactions, primers over- or under-amplifying, and primers amplifying off-target products) would then be removed from the next library preparation and sequencing until we reached an optimized PCR multiplex performance to produce the GT-seq library. However, we were able to retain only 491 loci with adequate primer pairs (see Preliminary Results). As such, we could not select loci from a bigger pool and all primers for t 491 loci were used for the optimization process. Nevertheless, the SNPs present in those loci were haplotyped using *rad_haplotyper.pl* script with default parameters and the resulting microhaplotypes were used to estimate F_{ST} and observed heterozygosity (H_o) with R package *diveRsity* v. 1.9.90 (Keenan *et al.*, 2013) to evaluate the levels of diversity discriminated by that dataset. Also, the same metrics were estimated for the complete dataset (2804 loci), and we then tested if the distributions of values were similar between dataset, using non-parametric Kruskal-Wallis tests (because none of the metrics follow a normal distribution) implemented in R.

GT-seq panel optimization

To test the efficacy of designed primers to amplify the target loci, an initial GT-seq library was prepared using the 491 primer pairs (see Preliminary Results) with 48 samples previously used for nextRAD-seq and SNP discovery (to compare genotyping accuracy). The library was sequenced following Campbell *et al.* (2015) with two minor modifications. First, the read 1 primer that allows sequencing of our target fragment was used without the last adenine base (A), as advised by the authors. Second, to facilitate sequencing on an Illumina® NextSeq 2000 platform, we designed a custom index 2 primer to read the i5 index. This primer was the reverse-complement of the read 1 primer. Single-end sequencing was performed on an Illumina® NextSeq 2000 at the University of New Mexico Health Sciences Center. After sequencing, raw data was used to estimate the number of several types of primer interactions using the script *GTseq_Primer-Interactions.pl* from GTseq-Pipeline (Campbell *et al.* 2015; https://github.com/GTseq/GTseq-Pipeline). Primers with excessive number of interactions with primers from other pairs were discarded from PCR multiplex to prepare the subsequent genomic library. In cases where a primer interacted mostly with a single other primer, we kept the pair that sequenced the locus with higher F_{ST}. This process will be repeated until the proportion of reads in the library corresponding to primer interactions is relatively low compared to target reads.

A second library was prepared based on the results from the sequencing of the first library (see Preliminary Results) and is currently in the queue for sequencing.

Preliminary Results

Identification of SNPs

After sequencing nextRAD libraries, demultiplexing the raw reads (i.e., DNA sequences prior to any filtering) and trimming approximately 661.8 million (M) reads were retained with a mean of 3.5 M sequences per individual (minimum = 1.6 thousand; maximum = 5.1 M). From these reads, 98.9% (minimum = 78.8%; maximum = 99.6%) were aligned to the peppered chub draft genome.

FreeBayes identified 1.6 M raw variants (including SNPs, multi-nucleotide polymorphisms, indels, and other complex variants) across the 189 individuals. A total of 2,804 loci containing 6,725 SNPs across 187 individuals with less than 30% missing data passed all filtering steps and were used for primer design. Depth per SNP was on average 46.1 (ranging from 20.4 to 98.5) and per individual was also 46.1 (ranging from 13.4 to 79.8).

Primer design for GT-seq and loci used for panel optimization

From the 2,804 loci retained, 1,850 had sufficient size for sequencing and flanking regions for primer design. However, we were able to retain only 491 loci with suitable primer pairs that followed the primer design parameters and without off-target matches across the draft genome.

When comparing the complete dataset with 2,804 loci and the reduced dataset for GT-seq panel optimization with 491 loci, we found the distribution of F_{ST} values was similar between both datasets (Kruskal-Wallis X^2 = 0.45; p-value = 0.5), suggesting that this panel should be adequate to evaluate changes in allelic frequencies that reflect those genome-wide changes. On the other hand, H_O was significantly lower in the reduced dataset (Kruskal-Wallis X^2 = 33.91; p-value = 5.77 x 10-9). While this suggests that the genetic diversity (measured as heterozygosity) contained in the reduced panel is lower than genome-wide heterozygosity, this can change by discarding loci during optimization process. Once the panel is optimized, the loci included in the final panel may still be adjusted to better reflect genome-wide diversity. However, we are constrained by the number of loci that we had at the beginning of the optimization process.

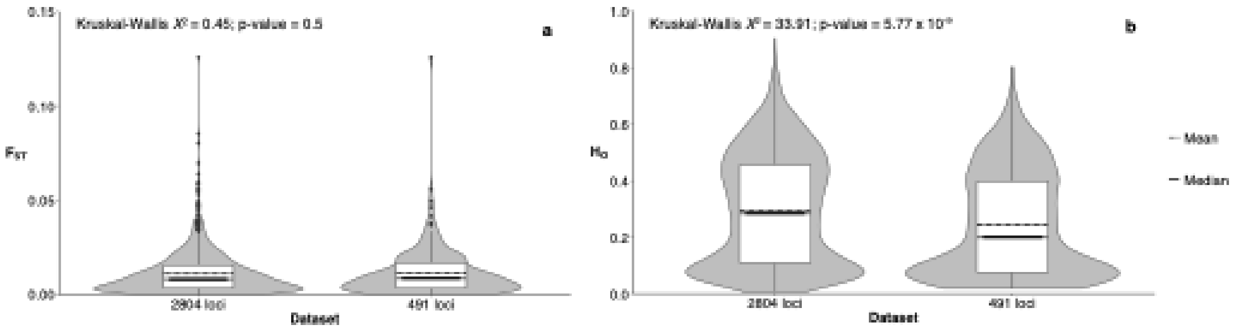

Figure 1 – Violin and box plots of F_{ST} and observed heterozygosity (H_O) distributions across loci and the results of Kruskal-Wallis tests to assess if distributions were statistically different between datasets. (a) Comparison of the distribution of F_{ST} values between the complete microhaplotype dataset containing 2,804 loci and the microhaplotype dataset from 491 loci with suitable primer pairs for GT-seq. (b) Comparison of the distribution of H_O values between the datasets. The shaded areas represent the density of loci across the spectrum of values for each statistic. The dashed line in the box plot represents the mean of the distribution and the solid line the median.

GT-seq panel optimization

After sequencing the first library prepared with all 491 primer pairs, we found that reads from on-target loci constituted only 20.1% of the total sequencing data and the remaining 79.9% were the result of primer interactions. Seventeen primer pairs contributed to 71.3% of the total number of primer interactions and were discarded from the second optimization round. However, in the absence of those 17 primer pairs, it is possible that the remaining primers form new interactions. Thus, removing those primers does not guarantee an optimized PCR multiplex yet. On the other hand, the other 474 primer pairs still contribute to primer interactions and those existing interactions might increase in proportion per our previous experience on GT-seq panels optimization; a relatively small fraction of interactions can be allowed in the final library as long as it does not exceed the proportion of on-target loci.

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-84-R-1.

References

Ackerman MW, Habicht C, Seeb LW (2011). Single-nucleotide polymorphisms (SNPs) under diversifying selection provide increased accuracy and precision in mixed-stock analyses of sockeye salmon from the Copper River, Alaska. *Trans Am Fish Soc* **140**: 865–881.

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

Beacham TD, Wallace C, MacConnachie C, Jonsen K, McIntosh B, Candy JR, *et al.* (2018). Population and individual identification of Chinook salmon in British Columbia through parentage-based tagging and genetic stock identification with single nucleotide polymorphisms. *Can J Fish Aquat Sci* **75**: 1096–1105.

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

Brumfield RT, Beerli P, Nickerson DA, Edwards S V (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* **18**: 249–256.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 1–9.

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. *Mol Ecol Resour* **15**: 855–867.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, *et al.* (2021). Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008.

Dodds WK, Gido K, Whiles MR, Fritz KM, Matthews WJ (2004). Life on the edge: the ecology of Great Plains prairie streams. *Bioscience* **54**: 205–216.

Dudley RK, Platania SP (2007). Flow regulation and fragmentation imperil pelagic-spawning riverine fishes. *Ecol Appl* **17**: 2074–2086.

Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. *arXiv Prepr arXiv12073907*.

Hess JE, Matala AP, Narum SR (2011). Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. *Mol Ecol Resour* **11**: 137–149.

Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöhl PA (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods Ecol Evol* **4**: 782–788.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005). Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. In: *Bmc Genetics*, BioMed Central Vol 6, p S26.

Luttrell GR, Echelle AA, Fisher WL, Eisenhour DJ (1999). Declining status of two species of the Macrhybopsis aestivalis complex (Teleostei: Cyprinidae) in the Arkansas River basin and related effects of reservoirs as barriers to dispersal. *Copeia*: 981–989.

Mangiafico S (2021). rcompanion: Functions to Support Extension Education Program Evaluation. R package version 2.4.0. https://CRAN.R-project.org/package=rcompanion

McKinney GJ, Waples RK, Pascal CE, Seeb LW, Seeb JE (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Mol Ecol Resour* **18**: 570–579.

Narum SR, Banks M, Beacham TD, Bellinger MR, Campbell MR, Dekoning J, *et al.* (2008). Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Mol Ecol* **17**: 3464–3477.

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*). *Conserv Genet* **22**: 927–945.

Paradis E (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**: 419–420.

Pennock CA, Gido KB (2017). Collapsing range of an endemic Great Plains minnow, Peppered

Chub *Macrhybopsis tetranema*. *Am Midl Nat* **177**: 57–68.

Perkin JS, Gido KB (2011). Stream fragmentation thresholds for a reproductive guild of Great Plains fishes. *Fisheries* **36**: 371–383.

Perkin JS, Gido KB, Cooper AR, Turner TF, Osborne MJ, Johnson ER, *et al.* (2015). Fragmentation and dewatering transform Great Plains stream fish communities. *Ecol Monogr* **85**: 73–92.

Platania SP, Altenbach CS (1998). Reproductive strategies and egg types of seven Rio Grande basin cyprinids. *Copeia*: **3**: 559–569.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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 Ecol Evol* **22**: 25–33.

Storer CG, Pascal CE, Roberts SB, Templin WD, Seeb LW, Seeb JE (2012). Rank and order: evaluating the performance of SNPs for individual assignment in a non-model organism. *PLoS One* **7**: e49018.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, *et al.* (2012). Primer3 – new capabilities and interfaces. *Nucleic Acids Res* **40**: e115–e115.

U. S. Fish and Wildlife Service (2022) Endangered and threatened wildlife and plants; endangered species status for the peppered chub and designation of critical habitat. Federal Register 87:11188–11220.

Wilde GR, Durham BW (2008). A life history model for peppered chub, a broadcast-spawning cyprinid. *Trans Am Fish Soc* **137**: 1657–1666.

Willis SC, Hollenbeck CM, Puritz JB, Gold JR, Portnoy DS (2017). Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Mol Ecol Resour* **17**: 955–965.