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, NM 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

Final report submitted for period ending June 30th 2025.

Table of Contents

Introduction

Peppered chub (*Macrhybopsis tetranema*) is a small-bodied and short-lived fish species belonging to the pelagic-broadcast spawning reproductive guild (Perkin & Gido, 2011; Platania & Altenbach, 1998). Fishes belonging to this guild are found in streams of the North American Great Plains (Dodds et al., 2004; e.g., Dudley & Platania, 2007). Peppered chub and other species within this reproductive guild are negatively affected by anthropogenic changes to rivers. These impacts include fragmentation, altered flow regimes, and habitat degradation. Survival and reproductive success of peppered chub have been linked to river discharge (Wilde & Durham, 2008) and connectivity that facilitates source-sink dynamics (Luttrell et al., 1999; Perkin & 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 during this period (Pennock & Gido, 2017; Perkin, Gido, Cooper, et al., 2015). 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. The reservoir has caused a 49% reduction in mean annual discharge (Wilde & 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 genetic diversity and effective population size (N_e) 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 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. Single nucleotide polymorphisms (SNPs) are alternative molecular markers and can be used to gather the same type of information as microsatellites. Single nucleotide polymorphisms 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 the sequencing of smaller fragments of DNA, so it can be effective even when DNA is limited 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 necessary for genotyping is relatively small (e.g., a few hundred) and the number of samples is high (e.g., hundreds to thousands), methods based on multiplex PCR and NGS can be more advantageous. Genotyping-in-Thousands by sequencing (GT-seq) is a method of targeted SNP genotyping that uses multiplexed PCR amplicon sequencing (Campbell et al., 2015). This method allows simultaneous amplification of hundreds of targeted genetic loci while barcoding of individuals allows thousands of individual samples to be sequenced in a single lane with a compatible Illumina® sequencing instrument (Campbell et al., 2015). Once a GT-seq panel is developed for the target species, GT-seq provides a cost-effective and efficient means of monitoring genetic variation and effective population size estimated from hundreds of SNPs.

Here we report the (i) discovery of genetic variants and identification of SNPs using a new and more complete peppered chub draft genome, (ii) primer design to develop a GT-seq panel for peppered chub, (iii) loci selection; and (iv) the optimized GT-seq panel that can be used for annual genetic monitoring of the South Canadian River population and the recently-founded peppered chub population held at the Southwestern Native Aquatic Resource and Recovery Center.

Methods

Identification of SNPs

Prior to this project, 189 samples were provided by the New Mexico Department of Game and Fish. These samples were collected during routine population monitoring from five sites on 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 & Salzberg, 2012) using the “local alignment” and default “very sensitive” options. Successfully aligned reads were filtered with Samtools v. 1.16 (Danecek et al., 2021; Li et al., 2009) 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 & Marth, 2012) on genomic intervals with at least 150 bp 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 extensive computational filtering so that only high-quality SNPs were retained in the final dataset. Using VCFtools v. 0.1.16 (Danecek et al., 2011) we removed 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 insertions and deletions 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. The bash script *dDocent_filters* (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) was then 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 than 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 of 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 of 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 of 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 step 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. After filtering, the final dataset represents 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. Loci that would result in sequences longer than 150 bp were then 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, in a previous study we found that selecting half of the loci with higher F_{ST} and selecting another half at random was optimal for genetic monitoring (Caeiro-Dias et al., in review). As such, the goal was to select 500 loci that reflected F_{ST} values found in the complete dataset (2,804 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 (e.g., 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) and all of these loci were used for the optimization process. 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. The same metrics were estimated for the complete dataset (2,804 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 library preparation and PCR multiplex 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 47 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. Demultiplexing was

performed with BCLConvert v. 4.2.7 (Illumina®, Inc.; https://emea.support.illumina.com/sequencing/sequencing_software/bcl-convert.html) allowing one mismatch per index. Only read 1 was used for downstream analysis.

Demultiplexed data were used to estimate the number of reads containing the expected primer combination and the number of reads resulting from 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). The primary goal was to identify and remove primer pairs that disproportionally contributed to primer interactions. This increased the depth of coverage of target loci with a reduced number of primer interactions. Primers with excessive numbers 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}. Next, the GTscore pipeline v. 1.3 (https://github.com/gjmckinney/GTscore) was used to identify genotypes. Prior to the analysis of the first sequencing run results, in-silico probes were designed for each SNP to include eight nucleotides flanking for each SNP and to include variants when overlapping identified SNPs (see manual for details on probe design https://github.com/gjmckinney/GTscore/blob/master/GTScoreDocumentation%20V1.3.docx). *AmpliconRadCounter.pl* script was used to count the number of unique reads per individual to identify on-target reads and to count the number of reads containing each SNP allele for every individual. Then counts of reads containing a SNP allele for each individual were used for microhaplotype genotyping with the maximum likelihood algorithm described by McKinney et al. (2018) and implemented in *GTscore.R* script. Genotype accuracy between nextRAD_complete and GT-seq was estimated from samples genotyped for at least 70% of the loci.

The average proportion of primer interactions and average genotype accuracy previously estimated were used as PCR multiplex optimization criteria. A high proportion of reads resulting from primer interactions is expected to have a large negative impact on PCR multiplex performance. As such, primer pairs producing a high proportion of primer interactions are typically removed during optimization of GT-seq panels (Hayward et al., 2022; Schmidt et al., 2020). Another criterion commonly used to decide whether to exclude or retain loci in GT-seq panels is genotype accuracy (Bootsma et al., 2020; Schmidt et al., 2020). The optimization process (library preparation, sequencing, evaluation of primer interactions, and genotype accuracy estimation) was repeated four times, until the proportion of reads from primer interactions was lower than one third of the sum of total reads and genotype accuracy was higher than 95%. In the first optimization round, loci were discarded only based on primer interactions. Other criteria were not applied at this stage because a high proportion of primer interactions is expected to bias such criteria values (Caeiro-Dias et al., in review). In the second optimization round we removed loci with primer/probe proportion lower than 40%, single SNPs with

genotype rate lower than 40%, and over- and under-represented SNPs by discarding SNPs with disproportionally high or low numbers of reads. In the third optimization round, SNPs with genotype accuracy lower than 0.9 were also removed. For criteria applied to single SNPs, only those not passing the criteria were discarded while the other SNPs were retained. The criteria applied to single SNPs were estimated based on individuals with less than 30% missing data. The final (i.e., fourth) optimization run confirmed that the proportion of primer interactions was low and genotype accuracy was high.

GT-seq panel validation

Our set of testing samples for panel optimization included samples collected in 2015 (n=7), 2017 (n=7), 2019 (n=19), and 2020 (n=7) from five sites across NM. Between 1-3 sites were included for each annual sample with ~10 samples per locality. Samples collected in the same year were combined and considered as a single "population". The optimized GT-seq panel was used to prepare a library that included additional samples from 2019 (n=21) and 2020 (n=35). These samples were also used for nextRAD sequencing. In total, 96 samples (i.e., data from the fourth optimization round and the additional library) were sequenced using both nextRAD-seq and GT-seq to validate the accuracy of the optimized panel for measuring and evaluating changes in genetic diversity when compared to the complete set of loci. Allele reads were counted with *AmpliconRadCounter.pl* script and the *GTscore.R* script was run to identify microhaplotypes. Individuals with missing data higher than 30% were removed. Monomorphic loci across the 96 samples were removed.

Next, using the 387 loci in the optimized GT-seq panel, we estimated locus-specific allelic richness (A_R), observed heterozygosity (H_O), expected heterozygosity (H_E), inbreeding coefficient (F_{IS}), and F_{ST} with the diveRsity R package. As previously described, we also tested if locus-specific H_O and F_{ST} estimated between datasets were statistically different, using non-parametric Kruskal-Wallis tests followed by a pairwise Wilcoxon test if the Kruskal-Wallis test was significant. All tests were performed in R. Pairwise F_{ST} between temporal collections was calculated. Temporal variation in genetic diversity was assessed using a Discriminant Analysis of Principal Components (DAPC) performed with adegenet v. 1.3-1 (Jombart, 2008; Jombart & Ahmed, 2011) R package. Missing data within each year was replaced 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. An initial DAPC was performed using years as groups, without scaling allele frequencies, retaining all principal components (PCs) during the PCA step and all discriminant functions (DFs) during Discriminant Analysis step, and keeping other options as default. The *a-score* method was used to select the optimal number of PCs to retain. The final DAPC was performed using the optimal number of PCs, two DFs, and using the other default options. The same analyses were performed with the nextRAD dataset (2,804 loci) using the same samples for

which we had GT-seq data, and with a nextRAD and a GT-seq dataset selecting randomly one SNP per locus.

Results

Microhaplotype identification and primer design

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 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. Average depth per SNP was 46.1 (ranging from 20.4 to 98.5) and per individual was also 46.1 (ranging from 13.4 to 79.8).

From the 2,804 loci retained, 1,850 had sufficient size for sequencing and the flanking regions for primer design were adequate. 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), which suggested that the genetic diversity (measured as heterozygosity) contained in the reduced panel is lower than genome-wide heterozygosity. This result was re-evaluated after the PCR multiplex optimization was completed.

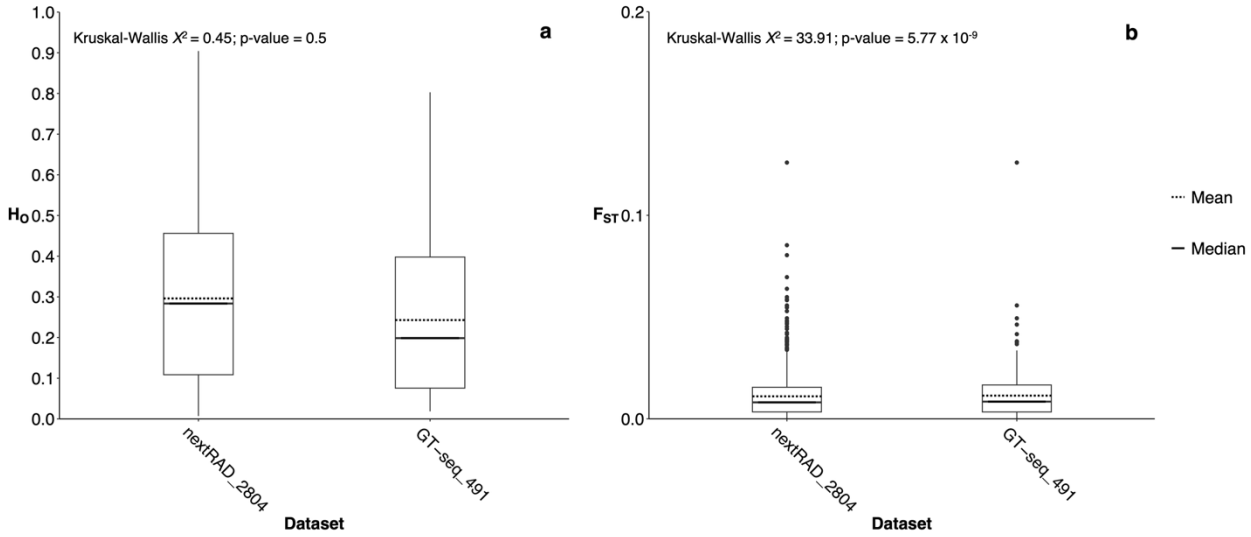

Figure 1 – Box plots of F_{ST} and observed heterozygosity (H_O) distributions across loci and the results of Kruskal-Wallis tests to assess if distributions were statistically different between datasets. (a) Comparison of the distribution of F_{ST} values between the complete microhaplotype dataset containing 2,804 loci (nextRAD_2804) and the microhaplotype dataset from 491 loci with suitable primer pairs for GT-seq (GT-seq_491). (b) Comparison of the distribution of H_O values between the datasets. The dashed line in the box plot represents the mean of the distribution and the solid line represents the median.

GT-seq library preparation and PCR multiplex optimization

Sequenced reads from the first optimization round were dominated by primer interactions, which constituted 79. 9% of demultiplexed reads (Table 1). From the initial 491 loci, 17 primer pairs disproportionally contributed to the majority of the primer interactions (71.3%) and were removed. Genotype accuracy was relatively high (91.8%), but loci were not discarded based on genotype accuracy at this stage because the genotype rate across SNPs was relatively low (77.6%), including 16 loci containing 33 SNPs that were not genotyped. Also, the genotype rate across individuals was relatively low (75.1%), including 14 individuals (29.8% of the individuals) with more than 30% missing data (these were not used for genotype accuracy estimates). Because the library was dominated by primer interactions, sequencing depth of coverage was reduced, and consequently the genotyping rate was relatively low. Indeed, the average read depth was relatively low for most SNPs, ranging from 0 to 278.1; the average across SNPs was 31.19 and the median was 24.4. The proportion of primer interactions in the second multiplex PCR decreased to 30.2%. However, the average genotype accuracy decreased to 81.3%. Twenty primer pairs with low genotyping rate (<0.5), overamplifying, with high contribution to primer interactions, and/or off-target amplification in the second optimization round were discarded prior to preparation of the third GT-seq library. Although the proportion of primer interactions only slightly decreased (28.9%), this further improved the overall genotype

accuracy to 94.01%. The third optimization round identified additional 58 loci where all SNPs (85) had genotyping accuracy below 90%. Another 65 SNPs across 43 loci with genotyping accuracy below 90% were removed but the other SNPs for those loci were retained. The last round of optimization containing 396 loci confirmed that the proportion of primer interactions among the demultiplexed reads was relatively small (27.2%) and that the individual SNP genotype accuracy when compared to genotypes obtained from nextRAD was high (97.4%). The optimized GT-seq panel was comprised of 396 neutral loci containing 614 SNPs; 145 loci contained two to five SNPs (microhaplotypes) and 251 were single SNP loci.

Table 1 – Summary results from GT-seq panel optimization. For each optimization round, the PCR multiplex performance was summarized with the overall proportion of reads resulting from primer interactions identified by *GTseq_Primer-Interactions.pl* script (Primer interactions), and the average SNP genotyping accuracy across individuals used in nextRAD-seq and for the GT-seq optimization with less than 30% missing data (Genotype accuracy). After the first optimization round, loci over- and under-amplifying and loci with genotype accuracy for all SNPs below 90% were discarded.

Optimization round		**1st**	**2nd**	**3rd**	**4th**
PCR multiplex performance	**Primer/adapter interactions**	79.9%	30.2%	28.9%	27.2%
	Average genotype accuracy	91.8%	81.3%	94.1%	97.4%
Number of primer pairs (loci) excluded	**Primer interactions[a]**	17	4	0	0
	Primer/Probe <0.4[a]	-	13	0	0
	Genotype rate <0.4[b]	-	4	0	0
	Overamplification[b]	-	3	0	0
	Underamplification[b]	-	0	0	0
	Genotype accuracy <0.9[b]	-	-	58	0
	Total*	17	20	58	0
Number of retained primer pairs (loci)		474	454	396	396

[a] Criteria applied to entire locus.

[b] Criteria applied to single SNPs. Loci were discarded if all SNPs did not pass the criterion.

[*] Some loci fall within several categories and thus the total is not necessarily the sum of all numbers across categories listed above.

GT-seq panel validation

Our optimized GT-seq panel contained nine monomorphic loci across the 96 test samples, resulting in 387 loci used for panel validation. From those samples, two were removed due to missing data higher than 30%. All loci showed less than 30% of missing data. A total of 94 samples were genotyped with the nextRAD (2,804 loci) and the GT-seq (387 loci) approaches. Both datasets returned very similar estimates of F_{ST} (Figure 2b) and F_{IS} (Table 2). Pairwise F_{ST} values were very close to zero regardless of the dataset (Table 3). Conversely, A_R, H_O and H_E estimated with the GT-seq panel were slightly lower when compared to the nextRAD dataset (Table 2). However, when comparing the same metrics estimated with nextRAD using only single SNPs, and the GT-seq panel using both microhaplotypes and only one SNP per locus, the results were identical (Figure 2 and Table 2). In particular, H_O was not statistically different when estimated from the nextRAD with single SNPs only and the GT-seq panel with microhaplotypes (Wilcoxon p = 0.79 after Bonferroni correction).

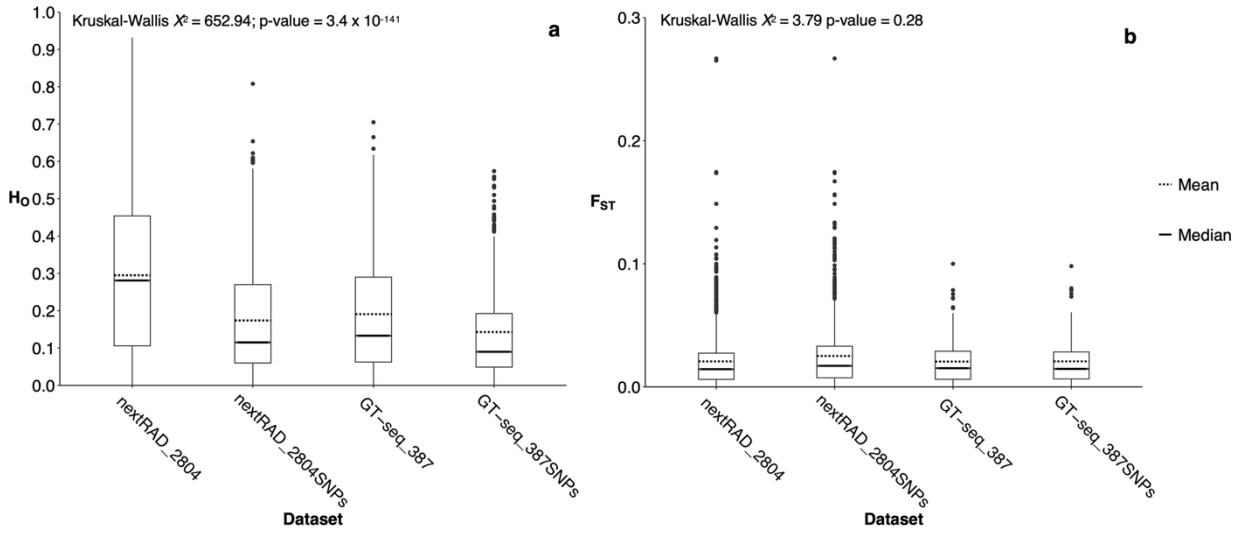

Figure 2 – Box plots of observed heterozygosity (H_O) and F_{ST} distributions across loci and the results of Kruskal-Wallis tests to assess if distributions were statistically different between datasets. (a) Comparison of the distribution of H_O values between the complete microhaplotype dataset containing 2,804 loci (nextRAD_2804), the complete dataset after randomly selecting one SNP per locus (nextRAD_2804SNPs), the optimized GT-seq panel with 387 (GT-seq_387), and the GT-seq panel with one SNP per locus selected at random. (b) Comparison of the distribution of F_{ST} values between datasets. The dashed line in the box plot represents the mean of the distribution and the solid line represents the median.

Table 2 - Summary statistics of genetic diversity (allelic richness [A_R], observed heterozygosity [H_O], expected heterozygosity [H_E], and inbreeding coefficient [F_{IS}]) estimated with the same individuals from two temporal collections (2019 and 2020) using the complete nextRAD dataset (2,804 loci), the optimized GT-seq panel (396 loci), and excluding four loci from the GT-seq panel with missing data higher than 30%. Average missing data (MD) per year for each dataset is also shown.

Year	Dataset	A_R	H_O	H_E	F_{IS}	MD
2015 (n=7)	nextRAD 2804 loci	2.01	0.29	0.29	0.02	5.7%
	nextRAD 2804 SNPs	1.51	0.17	0.17	0.01	4.5%
	GT-seq 387 loci	1.64	0.18	0.19	0.03	0.5%
	GT-seq 387 SNPs	1.45	0.13	0.14	0.05	0.3%
2017 (n=7)	nextRAD 2804 loci	2.03	0.30	0.29	-0.01	2.3%
	nextRAD 2804 SNPs	1.53	0.18	0.17	-0.02	1.5%
	GT-seq 387 loci	1.64	0.20	0.19	-0.05	0.8%
	GT-seq 387 SNPs	1.46	0.15	0.14	-0.06	0.6%
2019 (n=38)	nextRAD 2804 loci	2.21	0.29	0.32	0.07	3.8%
	nextRAD 2804 SNPs	1.60	0.17	0.18	0.06	2.6%
	GT-seq 387 loci	1.79	0.20	0.21	0.06	1.9%
	GT-seq 387 SNPs	1.56	0.15	0.16	0.05	1.6%
2020 (n=42)	nextRAD 2804 loci	2.20	0.30	0.31	0.04	6.4%
	nextRAD 2804 SNPs	1.61	0.18	0.18	0.03	4.4%
	GT-seq 387 loci	1.77	0.19	0.21	0.10	2.0%
	GT-seq 387 SNPs	1.54	0.14	0.15	0.09	1.8%

Table 3 – Pairwise F_{ST} estimated between temporal collections using the complete nextRAD dataset with 2,804 loci (nextRAD 2804 loci; top left quadrant), the nextRAD dataset with 1 SNP per locus randomly selected (nextRAD 2804 SNPs; top right quadrant); the GT-seq panel containing 387 loci (GT-seq 387 loci; bottom left quadrant); the GT-seq panel with 1 SNP per locus randomly selected (GT-seq 387 SNPs; bottom right quadrant).

nextRAD 2,804 loci	**2015**	**2017**	**2019**
2017	-0.002	-	-
2019	5.0×10^{-4}	-2.0×10^{-4}	-
2020	9.0×10^{-4}	$2.0\text{x}10^{-4}$	0.001

nextRAD 2,804 SNPs	**2015**	**2017**	**2019**
2017	0.002	-	-
2019	0.003	0.002	-
2020	0.002	7.0×10^{-4}	0.002

GT-seq 387 loci	**2015**	**2017**	**2019**
2017	-0.002	-	-
2019	0.002	-7.0×10^{-4}	-
2020	-0.003	-0.001	$8.0\text{x}10^{-4}$

GT-seq 387 SNPs	**2015**	**2017**	**2019**
2017	2.0×10^{-4}	-	-
2019	0.002	0.002	-
2020	-0.003	0	3.0×10^{-4}

The DAPC results from the complete set of 2,804 loci (including microhaplotypes) revealed small deviations in genetic variability between 2019 and 2020 when compared to previous temporal collections (Figure 3a). The same deviation for 2019 was not detected with both datasets using only one SNP per locus (Figure 3b and 3d). However, the GT-seq panel including microhaplotypes was able to detect similar deviations as detected by the complete set of 2,804 loci (Figure 3c), although the shifts detected with the GT-seq panel were more subtle.

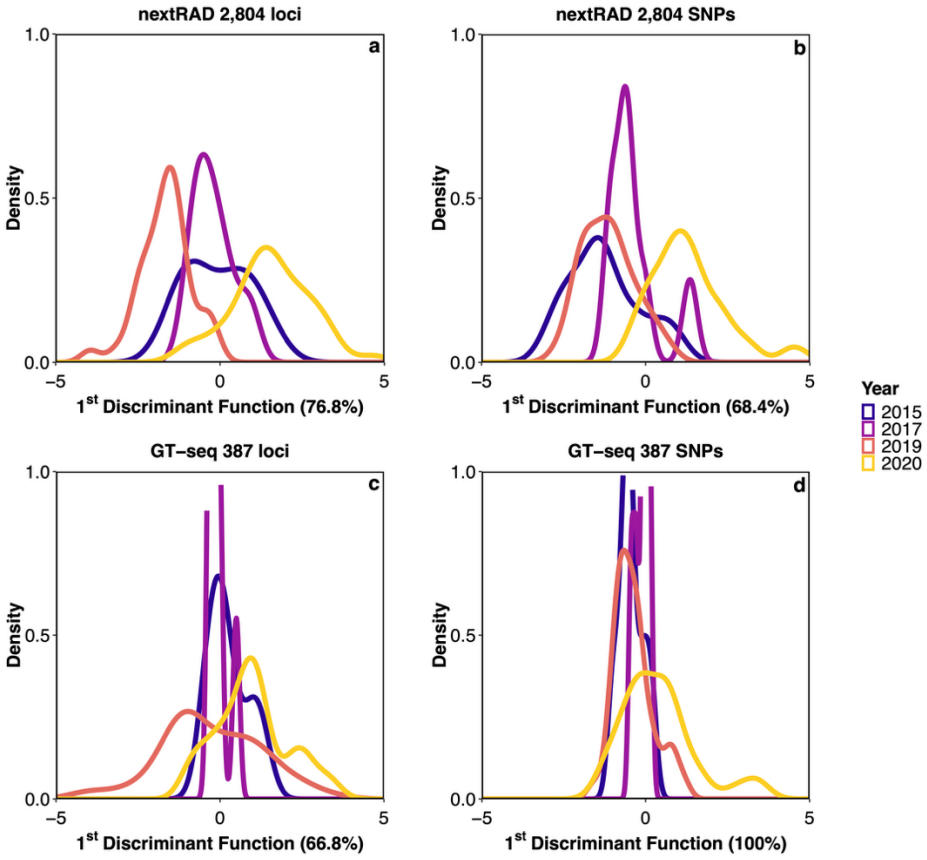

Figure 3 – Discriminant analysis of principal components (DAPC) results for four temporal collections obtained from **a**) the complete nextRAD dataset with 2,804 loci, **b**) the nextRAD dataset with 1 SNP per locus randomly selected, **c**) the GT-seq panel containing 387 loci, **d**) the GT-seq panel with 1 SNP per locus randomly selected. The percentages on the x-axis refer to the proportion of variation explained by the first discriminant function.

Discussion

In this study, we developed a GT-seq panel for genetic monitoring of the federally endangered peppered chub. We used temporal data from archived collections spanning six years to design and optimize a panel of 396 loci, including microhaplotypes and single biallelic SNPs. We showed that the loci included in the optimized GT-seq panel were consistent with those obtained from nextRAD-seq, with high genotype concordance (97.4%). Such results are comparable with other studies developing GT-seq panels from reduced representation sequencing methods (Hayward et al., 2022; Schmidt et al., 2020; Setzke et al., 2021). The loci in the nextRAD dataset were chosen to include only neutral genetic variation, and thus the GT-seq panel also includes only neutral genetic variation.

From the 396 loci included in the GT-seq panel, nine were monomorphic across the genotyped samples and were not included in the panel validation. These loci are likely monomorphic only

because of the relatively small number of samples. Increasing the sample size in future evaluations will likely identify other alleles, as suggested by nextRAD data. Once the panel is applied for genetic monitoring and more samples are genotyped it will be possible to evaluate whether those loci should be retained in the panel or discarded. Here we used the 387 polymorphic loci for panel validation.

F_{ST} and F_{IS} were virtually identical regardless of the dataset. However, A_R, H_O, and H_E estimates obtained with GT-seq panel (387 loci comprising 600 SNPs) were slightly smaller than those obtained from genome-wide microhaplotype data (2,804 loci comprising 6,725 SNPs). For example, the biggest differences in H_O between both datasets were detected in 2015 and 2020 (in both cases $H_{O[nextRAD]}$-$H_{O[GT\text{-}seq]}$ = 0.11; see Table 2). A similar range of differences were obtained in a GT-seq panel developed from RAD sequencing data (both including only biallelic SNPs) and were considered "comparable across sequencing methods" without further investigation (Garrett et al., 2024). In Garrett et al. (2024), discordances in genotype accuracy and F_{IS} estimations were attributed to sequencing methodologies and genotyping methods. In our case, the explanation for differences in heterozygosity and allelic richness is that the genome-wide data included a considerably higher proportion of microhaplotypes and consequently more variable loci (58.1% loci with more than two alleles) compared to the optimized GT-seq panel (31% of the loci had more than two alleles). Because microhaplotypes are intrinsically more variable than biallelic SNPs, metrics based on the number of alleles and heterozygosity can be directly impacted (Baetscher et al., 2018; Osborne et al., 2023). To evaluate the impact of the proportion of multi-allelic loci in our dataset, we compared the previous results with results from datasets where we retained a single SNP per locus (selected at random). Results showed no or only very small differences in genetic diversity estimated with the complete set of 2,804 single SNPs, the GT-seq panel including microhaplotypes, and the GT-seq panel retaining only single SNPs. In general, these results demonstrate that similar power to estimate genetic diversity is provided by the GT-seq panel and the genome-wide biallelic SNP dataset.

Although locus variability can be increased by using microhaplotypes due to the higher number of SNPs per locus (Baetscher et al., 2018; McKinney et al., 2017), the ability to retain some SNPs in the GT-seq panel can be restricted by technical limitations (e.g., SNP position on the locus; genotype rate/errors) that result in a decreased number of SNPs per locus and consequently a potential decrease in some diversity metric estimations. In the present study, we were also constrained by a restricted number of initial loci that were compatible with GT-seq method, most likely due to reduced genetic diversity as consequence of recent population declines experienced by peppered chub (Caeiro-Dias et al., unpublished; Osborne et al., 2021). As such, it was not possible to include other loci to correct for the small bias in heterozygosity estimates. A way to circumvent similar issues in cases where genetic diversity is too low and the number of SNPs is predicted to be small, is to consider sequencing methods for SNP discovery the cover a higher proportion of the genome (e.g., Beemelmanns et al., 2024). Nevertheless, the

differences between nextRAD and GT-seq results (both with microhaplotypes) are small and consistent across temporal collections. Therefore, the optimized GT-sq panel can be considered a good representation of the genetic diversity obtained from the genome-wide nextRAD data.

The results from the DAPC showed that the GT-seq panel including microhaplotypes had the power to discriminate small shifts in genetic variations across years, similarly to the complete nextRAD dataset. Those differences are not driven by small sample sizes from 2015 and 2017, as similar shifts were detected when using higher sample sizes for those years (Caeiro-Dias et al., unpublished). This confirms that the optimized GT-sq panel was also able to track known temporal genetic changes in peppered chub, making previous temporal collections directly comparable to results obtained from future collections. Similar conclusions were reached by applying the same methodology to develop a GT-seq panel for the Rio Grande silvery minnow (*Hybognathus amarus*; Caeiro-Dias et al., in review). Overall, our results show that the optimized GT-seq panel successfully captures the same signal as the complete nextRADseq data.

Utility of the GT-seq panel for conservation of peppered chub

The GT-seq panel developed here for peppered chub offers a low-cost and efficient genotyping tool for regular genetic monitoring. We plan to publish this research in one of the American Fisheries Society peer reviewed journals, including a file with primer sequences and another file with the *in-silico* probe sequences needed for SNP genotyping. As such, all information needed to use the GT-seq panel developed here will be publicly available. Peppered chub was recently listed as endangered (U. S. Fish and Wildlife Service, 2022) under the Endangered Species Act. A broodstock was established recently for developing captive rearing and breeding methods. Individuals produced in captivity could support future augmentation of the extant wild population and provide individuals for reestablishing the species where it has been recently extirpated. Understanding the effect of demographic changes on effective population size and genetic diversity is essential for developing management actions that aim to maintain and promote genetic diversity and avoid further losses. Maximizing effective population size and hence genetic diversity is critical for recovery of imperiled species. This approach is currently being employed in other species facing similar management actions (Osborne et al., in review). Furthermore, this GT-seq panel will be an important tool in future attempts to reestablish extirpated populations (e.g., in the Ninnescah and Arkansas rivers; Pennock et al., 2017; Perkin, Gido, Costigan et al., 2015), because it will allow the diversity of the founding population to be evaluated prior to introducing new individuals. Also, future genetic monitoring of those populations will be directly comparable with the source population.

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. We thank Karen H. Gaines for editing support.

References

Ackerman, M. W., Habicht, C., & Seeb, L. W. (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. *Transactions of the American Fisheries Society*, *140*(3), 865–881.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Baetscher, D. S., Clemento, A. J., Ng, T. C., Anderson, E. C., & Garza, J. C. (2018). Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources*, *18*(2), 296–305.

Beacham, T. D., Wallace, C., MacConnachie, C., Jonsen, K., McIntosh, B., Candy, J. R., & Withler, R. E. (2018). Population and individual identification of Chinook salmon in British Columbia through parentage-based tagging and genetic stock identification with single nucleotide polymorphisms. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(7), 1096–1105.

Beemelmanns, A., Bouchard, R., Michaelides, S., Normandeau, E., Jeon, H., Chamlian, B., Babin, C., Hénault, P., Perrot, O., & Harris, L. N. (2024). Development of SNP Panels from Low-Coverage Whole Genome Sequencing (lcWGS) to Support Indigenous Fisheries for Three Salmonid Species in Northern Canada. *Molecular Ecology Resources*, e14040.

Bilton, T. P., McEwan, J. C., Clarke, S. M., Brauning, R., van Stijn, T. C., Rowe, S. J., & Dodds, K. G. (2018). Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, *209*(2), 389–400.

Bootsma, M. L., Gruenthal, K. M., McKinney, G. J., Simmons, L., Miller, L., Sass, G. G., & Larson, W. A. (2020). A GT-seq panel for walleye (*Sander vitreus*) provides important insights for efficient development and implementation of amplicon panels in non-model organisms. *Molecular Ecology Resources*, *20*(6), 1706–1722.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*(5), 249–256.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 1–9.

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

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., & Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008.

Dodds, W. K., Gido, K., Whiles, M. R., Fritz, K. M., & Matthews, W. J. (2004). Life on the edge: The ecology of Great Plains prairie streams. *BioScience*, *54*(3), 205–216.

Dudley, R. K., & Platania, S. P. (2007). Flow regulation and fragmentation imperil pelagic-spawning riverine fishes. *Ecological Applications*, *17*(7), 2074–2086.

Garrett, M. J., Nerkowski, S. A., Kieran, S., Campbell, N. R., Barbosa, S., Conway, C. J., Hohenlohe, P. A., & Waits, L. P. (2024). Development and validation of a GT-seq panel for genetic monitoring in a threatened species using minimally invasive sampling. *Ecology and Evolution*, *14*(5), e11321.

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

Hayward, K. M., Clemente-Carvalho, R. B. G., Jensen, E. L., de Groot, P. V. C., Branigan, M., Dyck, M., Tschritter, C., Sun, Z., & Lougheed, S. C. (2022). Genotyping-in-thousands by sequencing (GT-seq) of noninvasive faecal and degraded samples: A new panel to enable ongoing monitoring of Canadian polar bear populations. *Molecular Ecology Resources*, *22*(5), 1906–1918.

Hess, J. E., Matala, A. P., & Narum, S. R. (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.

Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W., & Prodöhl, P. A. (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, *4*(8), 782–788.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357.

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

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22.

Liu, N., Chen, L., Wang, S., Oh, C., & Zhao, H. (2005). Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics*, *6*(1), S26.

Luttrell, G. R., Echelle, A. A., Fisher, W. L., & Eisenhour, D. J. (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*, *1999*(4), 981–989.

McKinney, G. J., Seeb, J. E., & Seeb, L. W. (2017). Managing mixed-stock fisheries: Genotyping multi-SNP haplotypes increases power for genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(4), 429–434.

McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources*, *18*(3), 570–579.

Narum, S. R., Banks, M., Beacham, T. D., Bellinger, M. R., Campbell, M. R., Dekoning, J., Elz, A., Guthrie III, C. M., Kozfkay, C., & Miller, K. M. (2008). Differentiating salmon

populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, *17*(15), 3464–3477.
Osborne, M. J., Caeiro-Dias, G., & Turner, T. F. (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, M. J., Hatt, J. L., Gilbert, E. I., & Davenport, S. R. (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.
Paradis, E. (2010). pegas: An R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420.
Pennock, C. A., Gido, K. B., Perkin, J. S., Weaver, V. D., Devenport, S. R., & Caldwell, J. M. (2017). Collapsing range of an endemic Great Plains minnow, Peppered Chub *Macrhybopsis tetranema*. *The American Midland Naturalist*, *177*(1), 57–68.
Perkin, J. S., & Gido, K. B. (2011). Stream fragmentation thresholds for a reproductive guild of Great Plains fishes. *Fisheries*, *36*(8), 371–383.
Perkin, J. S., Gido, K. B., Cooper, A. R., Turner, T. F., Osborne, M. J., Johnson, E. R., & Mayes, K. B. (2015). Fragmentation and dewatering transform Great Plains stream fish communities. *Ecological Monographs*, *85*(1), 73–92.
Perkin, J. S., Gido, K. B., Costigan, K. H., Daniels, M. D., & Johnson, E. R. (2015). Fragmentation and drying ratchet down Great Plains stream fish diversity. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *25*(5), 639–655.
Platania, S. P., & Altenbach, C. S. (1998). Reproductive strategies and egg types of seven Rio Grande basin cyprinids. *Copeia*, *3*, 559–569.
Russello, M. A., Waterhouse, M. D., Etter, P. D., & Johnson, E. A. (2015). From promise to practice: Pairing non-invasive sampling with genomics in conservation. *PeerJ*, *3*, e1106.
Schmidt, D. A., Campbell, N. R., Govindarajulu, P., Larsen, K. W., & Russello, M. A. (2020). Genotyping-in-Thousands by sequencing (GT-seq) panel development and application to minimally invasive DNA samples to support studies in molecular ecology. *Molecular Ecology Resources*, *20*(1), 114–124.
Schwartz, M. K., Luikart, G., & Waples, R. S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology & Evolution*, *22*(1), 25–33.
Setzke, C., Wong, C., & Russello, M. A. (2021). Genotyping-in-Thousands by sequencing of archival fish scales reveals maintenance of genetic variation following a severe demographic contraction in kokanee salmon. *Scientific Reports*, *11*(1), 22798.
Storer, C. G., Pascal, C. E., Roberts, S. B., Templin, W. D., Seeb, L. W., & Seeb, J. E. (2012). Rank and order: Evaluating the performance of SNPs for individual assignment in a non-model organism. *PLoS One*, *7*(11), e49018.
Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—New capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115–e115.
Wilde, G. R., & Durham, B. W. (2008). A life history model for peppered chub, a broadcast-spawning cyprinid. *Transactions of the American Fisheries Society*, *137*(6), 1657–1666.
Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD loci: An efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*, *17*(5), 955–965.

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.