Exploring genetic variation and population structure in a threatened species, *Noturus placidus*, with whole-genome sequence data

Lynsey K. Whitacre, 1,2 Mark L. Wildhaber, 3,*  Gary S. Johnson, 4 Harly J. Durbin, 2 Troy N. Rowan, 2 Peoria Tribe, 5 Robert D. Schnabel 2, 3,* Tendai Mhlanga-Mutangadura, 4 Vernon M. Tabor, 6 Daniel Fenner, 7 Jared E. Decker 3, 1, 2,*

1Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA,
2Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA,
3U.S. Geological Survey, Columbia Environmental Research Center, Columbia, MO 65201, USA,
4Department of Veterinary Pathobiology, College of Veterinary Medicine, University of Missouri, Columbia, MO 65211, USA,
5The Peoria Tribe of Indians of Oklahoma, Miami, OK 74354, USA,
6U.S. Fish and Wildlife Service, Kansas Ecological Services Field Office, Manhattan, KS 66502, USA,
7U.S. Fish and Wildlife Service, Oklahoma Ecological Services Field Office, Tulsa, OK 74129, USA

*Corresponding author: Informatics Institute, University of Missouri, Columbia, MO 65211, USA. Email: DeckerJE@missouri.edu; *Corresponding author: U.S. Geological Survey, Columbia Environmental Research Center, 4200 E New Haven Rd, Columbia, MO 65201, USA. Email: mwildhaber@usgs.gov

Abstract

The Neosho madtom (*Noturus placidus*) is a small catfish, generally less than 3 inches in length, unique to the Neosho-Spring River system within the Arkansas River Basin. It was federally listed as threatened in 1990, largely due to habitat loss. For conservation efforts, we generated whole-genome sequence data from 10 Neosho madtom individuals originating from 3 geographically separated populations to evaluate genetic diversity and population structure. A Neosho madtom genome was de novo assembled, and genome size and content were assessed. Single nucleotide polymorphisms were assessed from de Bruijn graphs, and via reference alignment with both the channel catfish (*Ictalurus punctatus*) reference genome and Neosho madtom reference genome. Principal component analysis and structure analysis indicated weak population structure, suggesting fish from the 3 locations represent a single population. Using a novel method, genome-wide conservation and divergence between the Neosho madtom, channel catfish, and zebrafish (*Danio rerio*) was assessed by pairwise contig alignment, which demonstrated that genes important to embryonic development frequently had conserved sequences. This research in a threatened species with no previously published genomic resources provides novel genetic information to guide current and future conservation efforts and demonstrates that using whole-genome sequencing provides detailed information of population structure and demography using only a limited number of rare and valuable samples.

Keywords: Genome assembly; demography; population structure

Introduction

A primary focus of management efforts to conserve the Earth’s biodiversity is decreasing the high rate of species loss (Segelbacher et al. 2021). Conservation genetics and genomics have advanced in recent years and are becoming more and more important to maintaining and improving biodiversity. Segelbacher et al. (2021) discuss cloning extinct species, controlling invasive species, gene editing, genetic rescue (e.g. translocation), genetic management (e.g. reproductive technology and genetic load) as possible tools for conservation and recovery of species. However, they emphasize the need to understand the genetic diversity of the species via genomic information prior to using these other approaches (Segelbacher et al. 2021).

The Neosho madtom (*Noturus placidus*) is a small ictalurid, generally less than 3 inches in length, unique to parts of Kansas, Missouri, and Oklahoma in the Neosho, Cottonwood, and Spring Rivers (Taylor 1969; Luttrell et al. 1992; Wilkinson et al. 1996; Wildhaber 2011, 2014) (Fig. 1). The small fish inhabit loosely compacted gravel in high to moderate velocity water that is generally associated with riffles (Wenke et al. 1992; Fuselier and Edds 1994). Due to dam construction, agricultural runoff, and lead-zinc mining, much of this species’ historical habitat has been destroyed (Wildhaber et al. 2000b; Allen et al. 2001; Tiemann et al. 2004). Neosho madtom was listed as threatened in 1990 by the U.S. Fish and Wildlife service (55 FR 21,148), and a recovery plan was enacted in 1991 (U.S. Fish and Wildlife Service 1991). Over the last 2 decades, research efforts have elucidated the fish’s habitat and behaviors to aid in conservation (Fuselier and Edds 1994; Pfingsten and Edds 1994; Wildhaber et al. 1999, Bryan et al. 2005). In addition, conservation efforts have focused on the effects of water quality, nutrients, and competition from other species on population density (Wildhaber et al. 1999, 2000b). However, the genetic diversity and population structure of Neosho madtoms is not known.

Received: August 26, 2021. Accepted: January 06, 2022
Published by Oxford University Press on behalf of Genetics Society of America 2022. This work is written by US Government employees and is in the public domain in the US.
The largest extant populations of Neosho madtom reside in the Neosho and Cottonwood Rivers above John Redmond Reservoir (Fig. 1). The Spring River population (Fig. 1) is sparse, likely due to the presence of cadmium, lead, and zinc and limited food and habitats (Wilkinson et al. 1996; Wildhaber et al. 1998, 2000a). Sampling of Neosho madtom over the last 3 decades has indicated an overall steady decline in population density (fish per square meter) from 1991 to 2008 (Bryan et al. 2010); a similar decline occurred in all 3 rivers where the Neosho madtom is currently found (i.e., Neosho, Cottonwood, and Spring). Due to the overall declining Neosho madtom populations and the small size of the Spring River population, reintroduction efforts have been under consideration (Wildhaber 2014). Before reintroductions can occur, characterization of the genetic variability in Neosho madtoms is necessary to assess the degree of differentiation among geographically isolated populations. In addition, given the small sample sizes available in conservation genetic studies, whole-genome sequencing provides the most data for precious samples of imperiled species.

Populations upstream and downstream of the John Redmond Reservoir may have been geographically isolated since construction on the dam began in 1959 (Wildhaber 2011). Low-water dams found throughout the Neosho and Cottonwood Rivers may also be barriers between populations (Wildhaber 2011). These potential sources of population isolation and a generation interval of only 1 year prompted research to determine the extent of population divergence (Wildhaber 2011, 2014). Advances in DNA sequencing enable efficient genomic data production to reveal base-by-base level variation and aid in conservation efforts of threatened and endangered species such as the Iberian lynx, giant panda, scarlet macaw, and Tasmanian devil (Li et al. 2010; Miller et al. 2011; Seabury et al. 2013; Abascal et al. 2016). The work presented here demonstrates how genome sequencing provides additional information when biological samples are rare or valuable in conservation biology. The objective of this study was to determine if selective pressures or genetic drift has resulted in genetically unique populations within the current range of the Neosho madtom. We apply whole-genome sequencing to assess genetic variability and population structure in Neosho madtom from the upper and lower Neosho River and the Cottonwood River. These were assessed using variants from graph-based, reference alignment to a diverged species, and reference alignment to a de novo assembly for the species of interest. Further, we estimate historical effective population sizes to assess long-term demographic changes. Finally, with the draft de novo assembly of the Neosho madtom genome, we used a novel method to identify signatures of conserved and diverged DNA sequences.

**Methods**

**Sample collection**

Under federal and state scientific collection permits (TE-207526-0, TE207526-2), 3 Neosho madtoms were collected from the upper Neosho River above John Redmond Reservoir (UNR) and from the Cottonwood River (CR) and 4 were collected from the lower Neosho River below the John Redmond Reservoir (LNR) (Fig. 1; Supplementary Table 1). A stonecat (Noturus flavus) was also collected to serve as an outgroup. Given the small size of adult Neosho madtoms in the field (Wildhaber 2014), all fish were held and fed in the laboratory, following requirements of the USGS-CERC Institutional Animal Care and Use Committee (IACUC), with all applicable sections of the Final Rules of the Animal Welfare Act regulations (9 CFR), and with all CERC standard operating procedures for the humane treatment of test organisms during culture and experimentation, to grow to a size with enough tissue for genome sequencing. The fish were humanely euthanized using MS-222, and both lateral muscles were removed and stored at −80°C until DNA was extracted. Genomic DNA was extracted from tissue using phenol-chloroform extraction (Sambrook et al. 1989).

**Genome sequencing**

Genomic DNA was used to construct 1 small insert paired-end library per fish with an approximate insert size of 300bp. The library was constructed following the manufacturer’s protocol.
with reagents supplied in Illumina’s TruSeq DNA PCR-Free sample preparation kit (#FC-121-3001). Briefly, 1 microgram of genomic DNA was sheared using standard Covaris methods to generate average fragmented sizes of 300 bp. The resulting 3’ and 5’ overhangs were converted to blunt ends by an end repair reaction which uses a 3’ to 5’ exonuclease and polymerase activity. The desired size of fragment (about 300 bp) was selected by sample purification beads (AMPure XP). A single adenosine nucleotide was added to the 3’ ends of the blunt fragment followed by the ligation of Illumina indexed paired-end adapters. The adapter ligated library was purified twice with sample purification beads. The purified library was quantified with a Qubit assay and library fragment size confirmed by Fragment Analyzer (Advanced Analytical Technologies, Inc., Ames, Iowa). Library was diluted and sequenced using a paired-end 100 base run according to Illumina’s standard sequencing protocol for the Illumina HiSeq 2000 system (Supplementary Table 2). Two mate-pair libraries with insert sizes of 2,000 and 3,000 bp were constructed from 1 fish to improve the scaffolding of the de novo assembly. All sequences have been deposited in NCBI’s SRA database (Supplementary Table 2); associated descriptive data provided by Whitacre et al. (2022).

**De novo variant calling and filtering**

The cortex_var algorithm was used for de novo variant discovery (Iqbal et al. 2012). Cortex calls variants by constructing multicolor de Bruijn graphs from the DNA sequence reads from all samples. We constructed graphs using k-mer sizes of 31 and 63 bp from individual graphs corrected for low coverage supernodes. Variants were called as the algorithm searched for motifs within the graph, referred to as bubbles, that are created by polymorphisms or repeats. K-mers of 31 and 63 were used jointly in variant calling to maximize the likelihood of variant detection.

Variants were first filtered based on variant type such that only biallelic SNPs were retained. These were then filtered based on genotype confidence, which is defined as the natural log probability of the maximum likelihood genotype minus the natural log probability of the second most likely genotype. Finally, SNPs with a genotype confidence < 5.54 (meaning the remaining genotypes called are less than $e^{-5.54}$, or 254.5, times more likely than the alternative) were filtered from the data set.

**Genome size and de novo assembly**

Two paired-end and 2 mate-pair libraries were sequenced from a single Neosho madtom and assembled using MaSuRCA-2.3.2 (Zimin et al. 2013). MaSuRCA uses the QuorUM error correction method (Marçais et al. 2015) and automatically chooses an appropriate K-mer size. It relies on both de Bruijn graph and overlap layout consensus (OLC) methods for accurate assembly. Parameters were adjusted to limit jump coverage to 150, set the maximum error rate for the scaffolder to 0.10, and use a Jellyfish (Marçais and Kingsford 2011) hash size of 1010. Scaffolds and contigs smaller than 200 base pairs were removed from the assembly using seqkit (Shen et al. 2016). The resulting assembly is available via NCBI’s Genome database (accession number JAFFST000000000, named MU_Nplac_3.0).

We used K-mers from sequence reads to estimate the genome size of the Neosho madtom. Assuming K-mers are uniquely mapped to the genome, their frequency reflects the depth of coverage (the sequencing of each base in the genome multiple times). Therefore, the genome size can be estimated as the total number of K-mers divided by the average frequency or coverage. We determined the K-mer distribution using Jellyfish (Marçais and Kingsford 2011) to count k-mers using $k = 31$. Counts were summed to determine the total number of K-mers analyzed and visualized as a histogram. K-mers with coverage < 5 were truncated as they likely represented sequencing errors. Using the peak of the histogram, genome size was estimated as the total number of K-mers analyzed divided by the value of the most frequent K-mer (coverage). We also estimated the portion of the genome that is single copy by calculating the total number of K-mers represented by the bell curve without considering the tails of the distribution (in this case, $5 < K$-mer count < 45).

**Reference variant calling and filtering**

The channel catfish (*Ictalurus punctatus*) reference genome (Liu et al. 2016) was also used to call variants. Raw sequences were trimmed for adaptors and base quality using Trimomatic-0.33 (Bolger et al. 2014). The reads were then aligned to both the IpCoco_1.2 channel catfish reference assembly (GCA_001660625.1) and the MU_Nplac_3.0 Neosho madtom catfish reference assembly using the BWA-MEM algorithm, version 0.7.10-r789 (Li and Durbin 2009). We called genotypes according to GATK Best Practices from the alignment files (McKenna et al. 2010; Depristo et al. 2011; Van der Auwera et al. 2013). Variant calling included modules for duplicate removal using Picard (http://broadinstitute.github.io/picard, last accessed 2/28/22), INDEL realignment, SNP and INDEL discovery using HaplotypeCaller, and genotype calling with GenotypeGVCFs. This resulted in 2 sets of variants, those from alignment to IpCoCo_1.1 and those from alignment to MU_Nplac_3.0.

SNPs were filtered based on the number of detected alleles $< 3$ (biallelic), QD (variant confidence/quality by depth) $< 2.0$, FS (Phred-scaled P-value using Fisher’s exact test to detect strand bias) $> 60.0$, SOR (symmetric odds ratio of 2 contingency table to detect strand bias) $> 4.0$, MQ (RMS mapping quality) $< 40.0$, MQRankSum (Z-score from Wilcoxon rank sum test of Alt vs Ref read mapping qualities) $< -12.5$, or ReadPosRankSum (Z-score from Wilcoxon rank sum test of Alt vs Ref read position bias) $< -8.0$.

**Principal component and structure analysis**

Principal component analysis (PCA) was conducted using the smart pca module from the EIGENSOFT 5.0 package (Patterson et al. 2006) using the 3 genotype sets from de Bruijn graph, channel catfish reference, and Neosho madtom reference variant calling for all samples (including the outgroup) and for only Neosho madtom. PCA excluding the outgroup was conducted after removing variants that were fixed for different alleles between the Neosho madtom and the outgroup. Tracy-Widom statistics were calculated to determine the significance of eigenvalues.

Structure analysis was performed using fastSTRUCTURE, an algorithm that uses high-density SNP data and a variational Bayesian framework for posterior inference to infer the ancestry of individuals (Raj et al. 2014). FastSTRUCTURE was run using $K = 1$ to $K = 4$ with simple priors on SNP data from all individuals (Neosho madtom and stonecat) and from $K = 1$ to $K = 3$ on SNP data from Neosho madtom only. These $K$ values were tested as they represent the total number of possible madtom populations and outgroup species. Both analyses were run using the 3 variant sets separately.

**Estimation of historical effective population size**

The historical effective population size of Neosho madtom and stonecat populations was estimated using SMC++ (Terhorst et al. 2017) and genotypes from unphased whole genome sequence
data. SMC++ uses variant calls and genome position information to estimate recombination rates across varying block sizes. Therefore, variants from the more contiguous channel catfish reference alignment were used with SMC++. The divergence time between Neosho madtom and stonecat was also inferred using SMC++ using the "split" command. All estimates were calculated and interpreted using a generation time of 1 year (Bulger and Edds 2001) and a Watterson estimation of mutation rate.

**Whole genome analysis of conservation and divergence**

Scaffolds from the de novo assembly of Neosho madtom were aligned to the channel catfish (GenBank accession GCA_001660625.1) and zebrafish (GenBank accession GCA_000002035.3) reference genomes using NUCmer 3.1 (Kurtz et al. 2004) with a break length of 200 bp, a minimum length for a maximal exact match equal to 20, and a minimum cluster length of 65. We used the alignment to perform whole-genome analyses of conservation and divergence. Orthologous regions of the channel catfish and zebrafish genomes to the Neosho madtom genome were determined by observing the best alignment and selecting additional secondary, nonoverlapping alignments on the same chromosome. All other spurious alignments to a chromosome or region that differed from the identified orthologous region for each contig were filtered from the alignment data set.

For each orthologous alignment, the number of identical aligned bases was calculated as the % identity times alignment length. A generalized linear model was then fit analyzing the number of identical aligned bases/alignment length using a quasi-binomial distribution weighted by the alignment length. A generalized linear model was then fit analyzing the number of aligned bases/alignment length as the % identity times alignment length. The model was corrected for alignment length analysis reported by Seabury et al. (2013). Regions determined to be significantly conserved or diverged were next considered, the primary source of variation was, as expected, between the 2 species of madtom (Fig. 2a, Supplementary Figs. 2a, and 3a). For the alignment to the Neosho madtom genome, eigenvalue 1 was 19.32 and accounted for 96% of the variation (Supplementary Fig. 4). Eigenvector 1 for the PCA of the alignment to the channel catfish variants accounted for 89.21% of the variation with an eigenvalue of 8.92 (Supplementary Table 6). For the de Bruijn graph and MU_Nplac_3.0 alignment SNPs, eigenvector 2 appeared to separate the Neosho madtom populations (Fig. 2a and Supplementary Fig. 1a), but the clusters are not well defined and the eigenvalue of 0.10 is small. Tracy-Widom statistics are not reported due to the apparent difficulty of estimating n', the effective number of columns (Patterson et al. 2006), given the small number of fish sequenced and the small eigenvalues for principal components 2 through 10. However, principal component 1, with eigenvalues ranging from 8.92 to 19.32 was clearly substantial, and the others, which are all smaller than 0.15, were not significant.

When only considering Neosho madtom, the populations, and some individuals, separated from one another (Fig. 2b, Supplementary Figs. 2b, and 3b), but, unlike the first PCA, their eigenvalues were nearly equal (Supplementary Fig. 4). For the IpCoco_1.2 called SNPs, eigenvector 1 only explained 13.56% of the variance and has a Tracy-Widom P-value = 0.9847 (Supplementary Table 6). For the MU_Nplac_3.0 called SNPs, eigenvector 1 only explained 13.45% of the variance. This indicates that there was no evidence of population structure among the different populations of Neosho madtom. With Neosho madtom catfish reference aligned calls, de Bruijn graph variant calls, and channel catfish reference aligned calls, the ancestry analysis recapitulated this result, giving an optimal value of K (the number
of population clusters) of 2 if Neosho madtom and stonecat SNP genotypes were provided (Supplementary Figs. 5 and 6) or $K = 1$ if only Neosho SNP genotypes were provided. This suggests that the stonecat and Neosho madtom populations have differentiable ancestry, but the UNR, CR, and LNR Neosho madtom populations do not have significant population structure.

**Estimation of historical effective population size**

Estimation of the historical effective population size ($N_e$) of Neosho madtom and stonecat suggests that the stonecat population has historically been a larger population than the Neosho madtom (Fig. 3a). This was also the case for estimates of most recent $N_e$. As expected, the $N_e$ estimates for the 3 Neosho madtom sampling sites were similar (Fig. 3a). The most recent $N_e$ estimates for Neosho madtom show the largest $N_e$ in the LNR population and the smallest $N_e$ in the UNR population. It is important to note that, due to the large parameter space associated with the estimation of $N_e$, the precision of the inferences is low. We also estimated the divergence time between Neosho madtom and stonecat to be approximately 10,000 years ago (Fig. 3b) assuming a clean split and a generation interval of 1 year.

**Whole-genome analysis of sequence conservation and divergence**

Contigs produced from the de novo assembly of the Neosho madtom were compared to the channel catfish and zebrafish reference assemblies to assess conservation and divergence on a genome-wide scale. Using a novel regression method, contigs were aligned to each reference assembly, and the alignment lengths and % identity were used to determine significantly diverged and conserved regions (see Methods). NUCmer alignments against the channel catfish reference genome produced an alignment for 60.1% of the de novo assembled contigs. The average % identity and length of all alignments was 84.68% and 2,477 bases, respectively (Fig. 4, a and b). Of these alignments, 32 were highly conserved statistical outliers and 5 were highly diverged statistical outliers between the species (Fig. 4c). Highly conserved regions contained 47 unique genes annotated in the channel catfish reference (Supplementary Table 7). This included 19 homeobox or homeobox-like genes, known to encode transcription factors involved in the development of the vertebrate body plan (Lewis 1978). As expected, highly diverged regions were mostly found in noncoding regions of the genome and overlapped with only 3 genes (Supplementary Table 8).

Alignments of Neosho madtom contigs against the zebrafish reference genome only resulted in an alignment for 14.4% of the de novo assembled contigs. The average % identity of alignments was 87.74%, and the average length of alignments was 3,766 bases (Fig. 5, a and b). No significantly conserved regions were identified using our method. However, 26 significantly diverged regions were identified (Fig. 5c). Those regions overlapped 12 unique genes annotated in the zebrafish reference (Supplementary Table 9).

**Discussion**

A primary need identified in conservation and recovery efforts of Neosho madtom has been population genetic information to
support reintroduction efforts into areas of their historic range from which they have been extirpated (e.g. sections of the Spring River in Oklahoma) (Wildhaber 2011, 2014). To address genetic variation and population structure in 3 of the remaining populations, we collected and sequenced DNA from 11 fish (10 Neosho madtom and 1 stonecat). The whole-genome sequencing used in this research provides more lines of investigation than the historically used microsatellite analyses, which yield only a small fraction of the genetic data generated by sequencing projects. Using sequencing data, we investigated population structure, estimated historical effective population sizes, assembled a genome, and looked for genome-wide signatures of conservation and divergence.

Mullen et al. (2010) describe 4 models of population structure in organisms inhabiting streams, namely (1) the Null Model (panmictic population), (2) the Death Valley Model, (3) the Stream Hierarchy Model, and (4) the Headwater Model (Mullen et al. 2010). We hypothesized that due to the dams in this river system (Fig. 1) and the small size of the Neosho madtom we would identify significant population structure. However, analysis of population structure based on genome-wide SNPs indicates that Neosho madtom from the upper and lower Neosho River and the Cottonwood River represent a single population lacking population structure. Although these populations were not thought to interbreed due to geographical isolation, the genomic data do not support this. Possibly, eggs or fish from the UNR and CR populations are moving downstream and contributing to the LNR population. This would explain the increased genetic variation and effective population size seen in the LNR Neosho madtom. However, denser geographical sampling is needed to investigate this hypothesis further. The absence of population structure is useful for future conservation efforts, which may use translocation.

Historic \( N_e \) was estimated using SNPs and provides insight into the demographics in both Neosho madtom and stonecat. Neosho madtom populations experienced steep declines since approximately 10,000 years ago. The stonecat has also shown an overall decline in effective population size. Although contemporary estimates of \( N_e \) are not estimated in our research, the SMC++ estimates do suggest that for the last 1,000 years the \( N_e \) of the madtom species has been less than 1,000.

Using these data, we demonstrate that although de Bruijn graph variant detection is a viable option, additional variation is detected by aligning to a reference genome of a closely related species. Including an outgroup in the cortex_yar analysis dramatically decreased the number of variants called (Supplementary Note 1). However, results using variants from de Bruijn graph and Neosho reference alignment were qualitatively more similar to
each other compared to results using variants from the channel catfish reference alignment. Thus, if reference bias is a concern, de Bruijn graph variant calling avoids this pitfall. Alignment to the channel catfish reference genome revealed genetic variation at a density of approximately 1 SNP per 5 kb of sequence in the Neosho madtom genome. This is much lower than the extent of common polymorphisms in humans who have approximately 4 SNPs per 5 kb of sequence (Zhao et al. 2003). Even greater genetic diversity is found in domesticated livestock (Wong et al. 2004; Gibbs et al. 2009; Bianco et al. 2015). In channel catfish, there are 43 SNPs per 5 kb of sequence (Sun et al. 2014). The SNP density in Neosho madtom catfish is even lower than the 5 SNPs per 5 kb in scarlet macaw (Seabury et al. 2013).

Comparison of the Neosho de novo assembly to the channel catfish genome identified signatures of conservation and divergence that reflect the unique evolutionary history of the Neosho madtom. These sequences provide a novel resource for future catfish genomics research. Many conserved genes encode transcription factors, which are conserved across species (Vaquerizas et al. 2009). Extreme conservation of developmental regulatory genes such as the homeobox clusters emphasizes the importance of the sequence and, subsequently, structure of these proteins. The homeobox genes and their regulatory elements are highly conserved across vertebrates (McGinnis et al. 1992; Gehring 1993; Lutz et al. 1996; Lee et al. 2006), indicating the importance of developing an apt body plan. Genes identified as significantly diverged between Neosho madtom and either channel catfish or zebrafish do not appear to have any coordinating theme. However, both genes identified as significantly diverged between Neosho madtom and channel catfish (PLCH1 and PFTPDR) have functions related to metabolism, perhaps reflecting the ecological niche of the Neosho madtom. Interestingly, several genes identified as conserved between Neosho madtom and channel catfish were identified as diverged between Neosho madtom and zebrafish.

It is important to note that our linear regression diverged sequence method inherently fails to identify some significantly diverged regions of the genome when contigs cannot be aligned at the required stringency due to the high divergence. However, high divergence is not the only reason why a contig might not align to a reference genome. Other reasons include misassemblies or regions that are absent from one or both genomes and low complexity or repetitive sequence. Contigs falling into each of these classes is difficult to identify with any confidence.

Overall, our research characterizes the genetic variability in Neosho madtom and suggests that the geographically isolated populations have no significant population structure. This information and the resources we have developed can be used to advise future conservation efforts to minimize accumulating detrimental variants (translocation of Neosho madtom) (Tailmon et al. 2004). A detailed genetic history and robust analysis of population structure are also important prerequisites for management strategies such as genetic restoration, a more comprehensive approach that aims to eliminate the effects of detrimental variation while maintaining advantageous variation and neutral variation that may be adaptive in the future (Hedrick 2005; Hansen et al. 2006; Bouzat et al. 2009; Johnson et al. 2010; Hostetler et al. 2013; Segelbacher et al. 2021). We have demonstrated the feasibility of using whole-genome sequencing in the population genetic analysis of a threatened or endangered species that does not have genomic resources. Whole-genome comparisons were crucial to reinforcing the validity of our conclusions due to the small sample size, which is common in conservation genetic studies where samples are often difficult to obtain. Collecting and sequencing samples from populations in the Spring River could help determine if Neosho and Cottonwood River populations are sufficiently genetically similar to use for reintroduction into sections of the Spring River where they are currently not found.

Data availability
Sequence data are available via NCBI SRA, under BioProject PRJNA437130. The genome assembly is available via NCBI’s Genome database (accession number JAFFST0000000, named MU_Nplac_3.0). Data describing original tissue sources is provided by Whitacre et al. (2022). Supplementary Tables 1–9, Supplementary Figs. 1–6, and Supplementary Note 1 are available online. Supplemental material is available at G3 online.

Acknowledgments
Funding for data fish collection and maintenance was provided by the Peoria Tribe of Oklahoma through grants from the U.S. Fish and Wildlife. The findings and conclusions in this study are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service. Thanks go to Dr. Diana Papoulia of CERC for assistance in tissue recovery from study specimens and Janice Albers and numerous students who helped in maintenance of the data and sample storage and tracking. In addition, we would like to acknowledge the critical nature of the scientific contributions and his importance as a colleague of Vernon Tabor in helping to make this work possible who passed away just before final publication of this manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Funding
JED was supported by grants MO-HAAS0027 and 2016-68004-24827 from the USDA National Institute of Food and Agriculture.

Conflicts of interest
The authors declare no conflicting interests.

Literature cited
Abascal F, Corvelo A, Cruz F, Villanueva—Cañas JL, Vlasova A, Marcet-Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol. 2016;17:251. https://doi.org/10.1186/s13059-016-1090-1
Allen GT, Blackford SH, Tabor VM, Cringan MS. Metals, boron, and selenium in Neosho madtom habitats in the Neosho river in Kansas, U.S.A. Environ Monit Assess. 2001;66(1):1–21.
Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Rozen D, Thibault J, et al. From FastQ to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Prot Bioinformatics. 2013; 43(1):11.10.1–11.10.33.
Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M. A deep catalog of autosomal single nucleotide variation in the pig. PLoS One. 2015;10(3):e0118867.
Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):214–2120.

Bouzat JL, Johnson JA, Toepfer JE, Simpson SA, Estler TL, Westemeier RL. Beyond the beneficial effects of translocations as an effective tool for the genetic restoration of isolated populations. Conserv Genet. 2009;10(1):191–201.

Bryan JL, Wildhaber ML, Leeds WB, Dey R. Neosho madtom and other Ictalurid populations in relation to hydrological characteristics of an impounded midwestern warmwater stream: update. U.S. Geological Survey Open-File Report 2010-1109. 2010. https://pubs.usgs.gov/of/2010/1109/pdfs/OFR2010-1109.pdf.

Bryan J, Wildhaber M, Noltie D. Examining Neosho madtom reproductive biology using ultrasound and artificial photothermal cycles. North Am J Aquacul. 2005;67(3):221–230.

Bulger AG, Edds DR. Population structure and habitat use in Neosho madtom (Noturus placidus). Southwestern Nat. 2001;46(1):8.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Phillippakis AA, Del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–498.

Fuselier L, Edds D. Seasonal variation in habitat use by the Neosho madtom (Teleostei, Ictaluridae, Noturus placidus). Southwestern Nat. 1994;39(3):217–222.

Gehring WJ. Exploring the homeobox. Gene. 1993;135(1–2):215–221.

Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, Green RD, Hamernik DL, Kappes SM, Lien S, et al. Bosvime HapMap Consortium. Genome-wide SNP survey uncovers the genetic structure of cattle breeds. Science. 2009;324(5926):528–532.

Hansen MM, Bekkevold D, Jensen LF, Mensberg K-LD, Nielsen EE. The sequence and de novo assembly of the giant panda genome. Nature. 2010;463(7279):311–317.

Liu Z, Liu S, Yao J, Bao L, Zhang J, Li Y, Jiang C, Sun L, Wang R, Zhang Y, et al. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nat Commun. 2016;7:11757.

Luttrell GR, Larson RD, Stark WJ, Ashbaugh NA, Echelle AA, et al. Status and distribution of the Neosho madtom (Noturus placidus) in Oklahoma. Proc Okla Acad Sci. 1992;72:5–6.

Lutz B, Lu HC, Eichele G, Miller D, Kaufman TC. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. Genes Dev. 1996;10(2):176–184.

Marcäis G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–770.

Marcäis G, Yorke JA, Zimin A. QuORUM: an error corrector for Illumina reads. PLoS One. 2015;10(6):e0130821.

McGinnis W, Krumlauf R, Gehring WJ, Paro R, DeRobertis EM, et al. Homeobox genes and axial patterning. Cell. 1992;68(2):283–302.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–1303.

Miller W, Hayes VM, Ratan A, Petersen DC, Wittekindt NE, Miller J, Walenz B, Knight J, Qi J, Zhao F, et al. Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). Proc Natl Acad Sci U S A. 2011;108(30):12348–12353.

Mullen LB, Woods HA, Schwartz MK, Sepulveda AJ, Lowe WH. Scale-dependent genetic structure of the Idaho giant salamander (Dicamptodon aterrimus) in stream networks. Mol Ecol. 2010;19(5):988–909.

Patterson NJ, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.

Pingstgen DG, Edds DR. Reproductive Traits of the Neosho madtom, Noturus placidus (Pisces: Ictaluridae). Trans Kansas Acad Sci. 1994;97(3/4):82.

Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–589.

Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press; 1989.

Seabury CM, Dowd SE, Seabury PM, Rausch M, Brightsmith DJ, Seabury PM, Raudsepp T, Brightsmith DJ, et al. Identification and analysis of Hox-dependent genetic structure of the Idaho giant salamander (Dicamptodon aterrimus) in stream networks. Mol Ecol. 2010;19(5):988–909.

Tallmon DA, Luikart G, Waples RS. The alluring simplicity and complex reality of genetic rescue. Trends Ecol Evol. 2004;19(9):489–496.
Taylor WR. A Revision of the Catfish Genus Noturus Rafinesque, with an Analysis of Higher Groups in the Ictaluridae. Washington (DC): Smithsonian Institution Press; 1969.

Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49(2):303–309.

Tiemann JS, Gillette DP, Wildhaber ML, Edds DR. Effects of lowhead dams on riffle-dwelling fishes and macroinvertebrates in a Midwestern River. Trans Am Fish Soc. 2004;133(3):705–717.

Tiersch TR, Goudie CA. Inheritance and variation of genome size in half-sib families of hybrid catfishes. J Hered. 1993;84(2):122–125.

U.S. Fish and Wildlife Service. Neosho madtom recovery plan. 1991. U.S. Fish and Wildlife Service, Denver, Colorado. pp. 42.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10(4):252–263.

Wenke TL, Eberle ME, Ernsting GW, Stark WJ. Winter collections of the Neosho madtom (Noturus placidus). Southwestern Nat. 1992;37(3):330.

Whitacre LK, Wildhaber ML, Johnson GS, Schnabel RD, Mhlanger-utangadura T, Tabor VM, Fenner D, Decker JE. Neosho Madtom (Noturus placidus) short read archive and whole genome sequence data: U.S. Geological Survey data release. 2022. https://doi.org/10.5066/P9MAP T9T.

Wildhaber ML. Neosho madtom, Noturus placidus. In: Distler D, Eberle M, Edds D, Gido K, Haslouer S, Huggins D, Mosher T, Stark B, Tomelleri J, Triplett J, Wiley E, editors. Kansas Fishes Committee. Kansas, Lawrence: University Press of Kansas; 2014. p. 303–305. https://kansaspress.ku.edu/978-0-7006-1961-0.html.

Wildhaber ML and Altman AL. Neosho madtom (Noturus placidus) and the multifaceted nature of population limiting factors. In: PH Michaletz, VH Travinich, editors. Conservation, Ecology, and Management of Catfish: The Second International Symposium. Bethesda (MD): American Fisheries Society; 2011. p. 281–294.

Wildhaber ML, Allert AL, Schmitt CJ. Both contaminants and habitat limit Neosho madtom (Noturus placidus) numbers in the Spring River, a midwestern warm water stream affected by runoff from historic zinc and lead mining. In: C Kennedy, D MacKinlay, editors. Fish Response to Toxic Environments. Baltimore (MD): International Congress on the Biology of Fish, 1998. p. 9–14.

Wilkinson C, Edds D, Dorlac J, Wildhaber ML, Schmitt CJ, Allert A. Neosho madtom distribution and abundance in the Spring River. Southwestern Nat. 1996;41:79–81.

Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 2003;312:207–213.

Zimin AV, Marcais G,Puig D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–2677.