Interim Progress Report for PRRIP project "Resolving Pallid Sturgeon Species Identification, Demographics and Hybridization using GT-Seq"

Edward J. Heist and Junman Huang
Center for Fisheries, Aquaculture & Aquatic Sciences
Southern Illinois University Carbondale
December 1, 2022

This report details progress made during the 2022 calendar year with expected progress for 2023-2026.

Student Recruiting and Training – Ph.D. student Junman Huang arrived in Carbondale and began his academic program on August 1, 2022. Mr. Huang is currently working on a dissertation proposal involving the use of GTseq (Campbell et al. 2015) for achieving goals related to conservation genetics of pallid sturgeon. Analyses of samples collected from the Platte River will be part of his work. A timeline for academic benchmarks is presented below.

Table 1. Academic benchmarks for Junman Huang's Ph.D. program.

1. Admitted to Graduate School – August 2022.
2. Signed committee form -- December 2022.
3. Plan of Study (list of formal courses) – May 2023.
4. Dissertation proposal –December 2024.
5. Preliminary examination –May 2025.
6. Dissertation defense/graduation – May 2027.

GTseq marker development and screening – We designed GTseq primers based on genomic sequences developed using double digest RADseq (Peterson et al. 2012) as part of Richard Flamio's dissertation research (Flamio et al. in press). We developed two separate GTseq panels for different purposes. The P-series of loci (where "P" stands for "polymorphic") are for population genetic analyses including testing for allele frequency differences among the current management units and estimates of effective population size. These markers were chosen based on high levels of heterozygosity in pallid sturgeon in the ddRAD study. The S-series of loci (where "S" stands for "species") had large allele frequency differences between pallid and shovelnose sturgeon in the ddRAD study. The S-series loci will be used for discriminating between pallid, shovelnose and hybrid sturgeon. Both panels will be genotyped simultaneously in each GTseq run.

We extracted sequences and identified target single nucleotide polymorphism (SNP) sites from the ddRAD sequence data and genotypes. Dr. Flamio's dissertation included a draft linkage map that included many but not all potential SNP sites. We removed loci that were identified on the linkage map such that no loci we retained were within 50 centi-Morgans of any other mapped SNP to avoid pseudoreplication (Waples et al. 2022). More than 400 sequences were sent to Nate Campbell of GTseek inc. for primer design using proprietary software. As part of the screening process, Dr. Campbell eliminated some potential primer sets because they lacked sufficient non-repetitive flanking sequence. Dr. Campbell ordered primers and performed an initial GTseq run to further identify and screen out problematic loci. Some loci produced an unusually large number of reads because they amplified multiple locations in the sturgeon genome. Some primers cross-amplified with primers from other loci to produce non target sequences. Problematic primer sets were removed from the panels prior to the next GTseq run, which is known as the "validation run."

Locus validation -- The validation run was performed on 96 sturgeon samples including 93 (38 pallid and 55 shovelnose) from the ddRAD study and 3 individuals that had borderline species assignments based on traditional microsatellite analyses (Schrey et al. 2007). Samples for the first run were chosen so that we could compare the genotypes for the same 93 individuals between ddRAD and GTSeq. The validation library was prepared by Nate Campbell of GTseek and sent to SIUC where we sequenced the instrument on our MiSeq. We compared the genotypes from the ddRAD study with those produced by GTseq. A small number of loci had discordant genotypes, or failed to amplify, and were discarded. Results from the comparisons with the remaining markers is presented as heat maps in Figures 1A and 1B. Both sets of loci (i.e., P and S) exhibited greater than 99.5% agreement between GTseq and ddRAD. At this time the panels are nearly set with 146 P-loci and 166 S-loci. Once more samples have been genotypes using GTseq, we will re-assess the reliability of amplification and test for Hardy Weinberg Equilibrium within populations to finalize the list of SNP loci included in each panel.

Analysis of the 3 individuals that had borderline species assignments demonstrated that GTseq provides much greater resolution than microsatellites (Table 2). The two potential broodstock fish had GTseq probabilities of >0.999 for the pure pallid sturgeon category and should be considered for broodstock. The larva that had a microsatellite-based p-value of 0.946 for the pure pallid sturgeon category had a GTseq p-value of 0.000 for the pure pallid sturgeon category and a p-value of 1.000 for the F_1 hybrid category. Examination of the raw genotypes (not shown) for this individual indicated that it was heterozygous for nearly all of the S-loci.

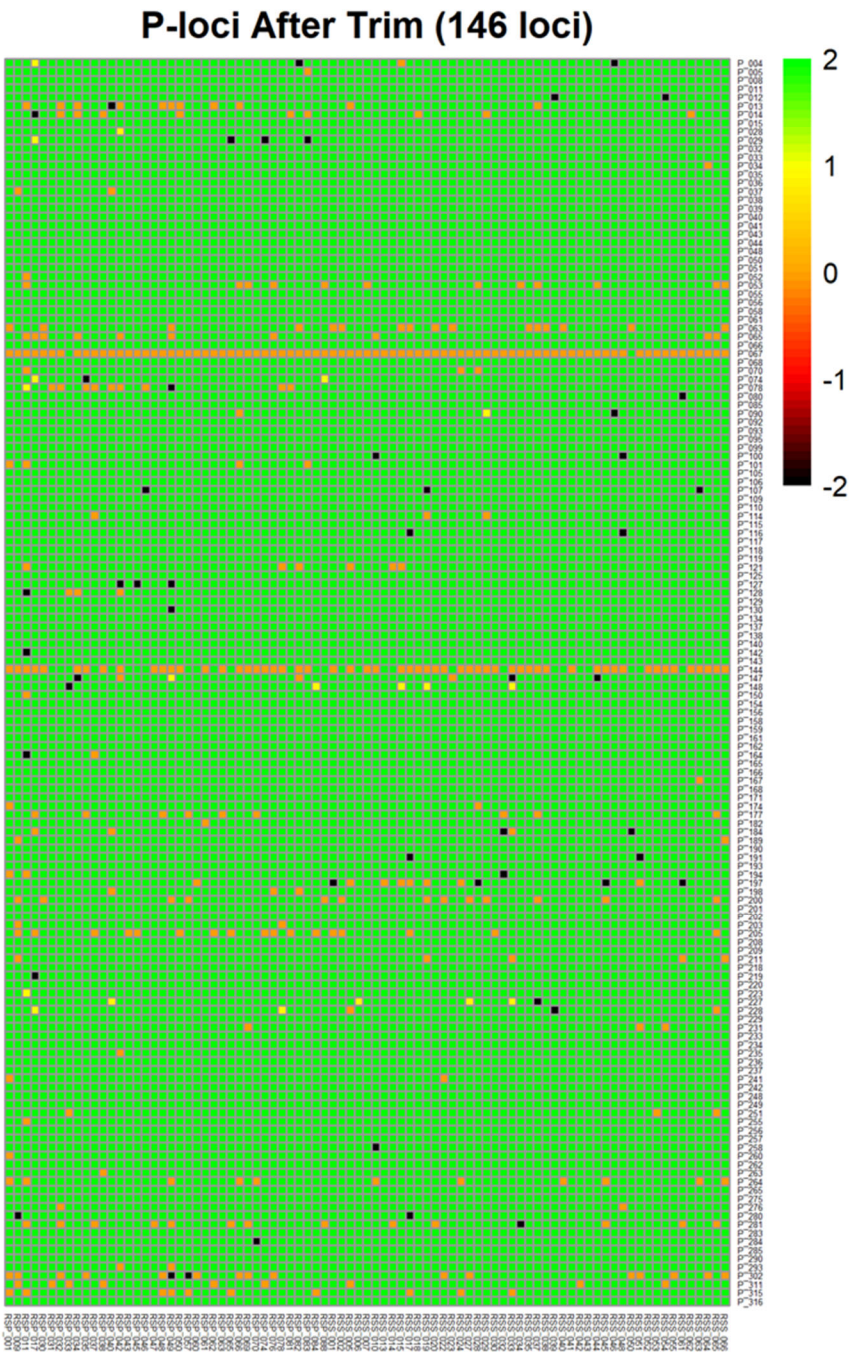

Figure 1A. heatmap of concordance between 146 GTseq P-loci and ddRAD for 93 sturgeon samples. Loci are on the long axis, individuals are on the short axis. Concordance = green, Missing ddRAD = yellow, Missing GTseq = orange, Discordance = black. Total concordance for the loci is 99.58%.

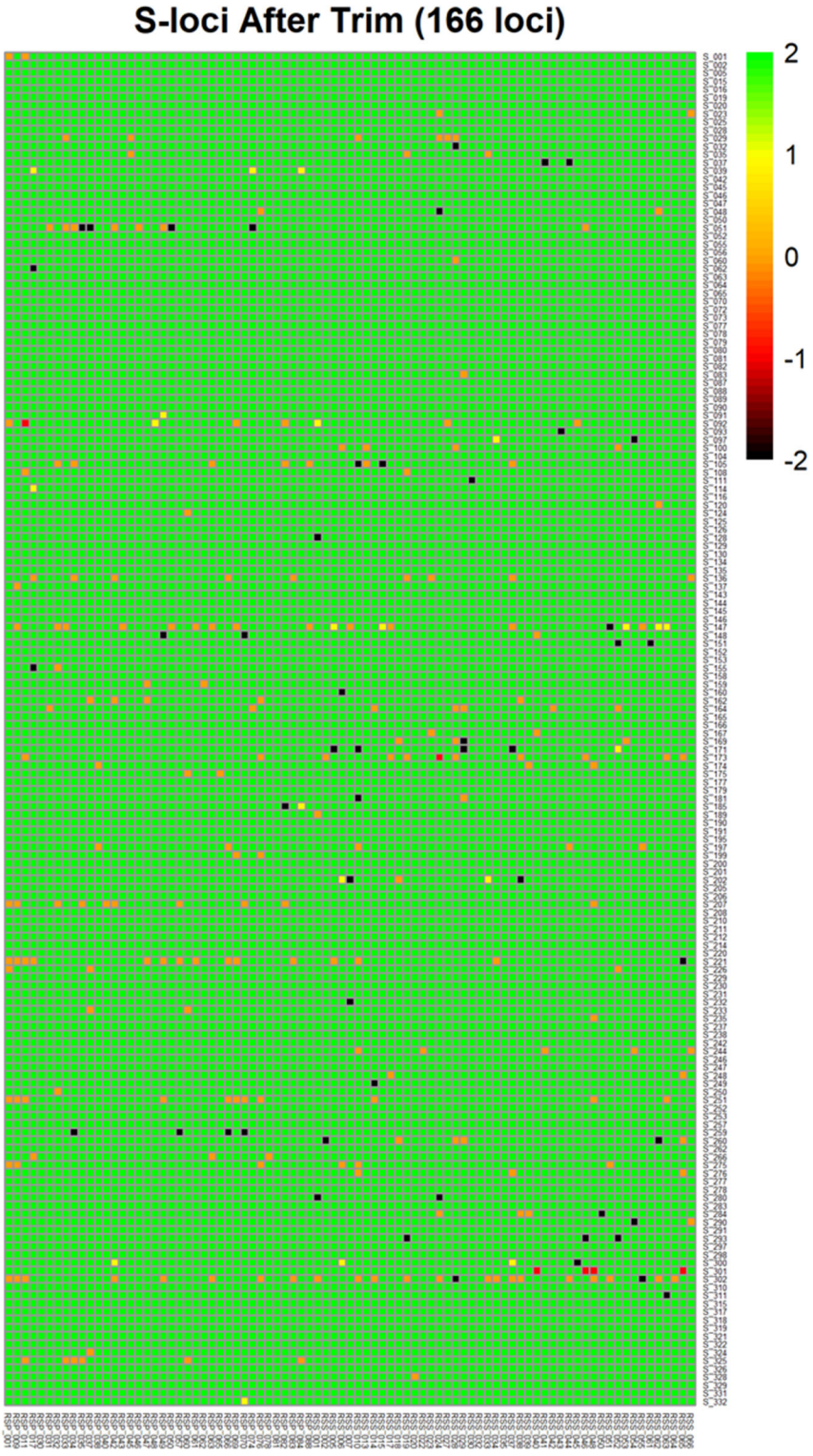

Figure 1B. heatmap of concordance between 166 GTseq S-loci and ddRAD for 93 sturgeon samples. Loci are on the long axis, individuals are on the short axis. Concordance = green, Missing ddRAD = yellow, Missing GTseq = orange, Missing both = red, Discordance = black. Total concordance for the loci is 99.64%.

Table 2. Comparison of NewHybrid assignment probabilities using microsatellites (µSat) and GTseq for three sturgeon, one wild-caught larva (Larva) and two potential broodstock fish (6C00111938 and 4713130264). Assignment categories are pure pallid sturgeon (Pal), pure shovelnose sturgeon (Sho), F_1 hybrid (F1), F_2 hybrid (F2), F1 backcross to pallid (BxP), and F_1 backcross to shovelnose (BxS).

		Pal	Sho	F1	F2	BxP	BxS
Larva	µSat	0.946	0.000	0.037	0.005	0.012	0.000
	GTseq	0.000	0.000	1.000	0.000	0.000	0.000
6C00111938	µSat	0.976	0.000	0.001	0.006	0.017	0.000
	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
4713130264	µSat	0.951	0.000	0.004	0.003	0.043	0.000
	GTseq	1.000	0.000	0.000	0.000	0.000	0.000

Equipment and Supplies – The Illumina MiSeq instrument is working well. To date we have had two successful GTseq runs on the instrument, the first with 96 samples and the second with 192 samples. Based on past results and discussions with Nate Campbell of GTseq we believe that we can genotype 384 samples (i.e., 4 standard plates of 96 individuals each) on a single MiSeq run using a standard V3 MiSeq cartridge. Three cartridges appropriate for GTseq are available for the MiSeq. Capacities and current costs associated with each cartridge type is presented below (Table 3).

Table 3. MiSeq sequencing cartridges appropriate for GTseq, cost of cartridge, estimated number of samples that can be analyzed per cartridge, and sequencing cost per sample assuming the full set of samples is run. Costs do not include cost of DNA isolation and library preparation.

Cartridge	Cost	Samples	Cost/sample
V2 nano	$310	24	$12.92
V2 micro	$502	96	$5.23
V3 150 cycle	$1,035	384	$ 2.70

Platte River Sample Results

26 fin clips from sturgeon collected in the Platte River were provided by collaborators from the University of Nebraska Lincoln (UNL). We isolated DNA from the samples using standard Qiagen kits. We genotyped the fin clips first with 19 microsatellite loci then with GTseq. Species assignments of both methods were performed using NewHybrids software (Anderson and Thompson 2002) using previously identified pallid and shovelnose sturgeon as baselines. Twenty-five of the fin clips were identified as pallid sturgeon and one was identified as a hybrid (Table 4). We performed parentage analyses using Cervus software (Kalinowski et al. 2007) using genotypes from known broodstock fish. Twenty-four of the fin clips came from hatchery origin pallid sturgeon. One of the pallid sturgeon and the hybrid were of wild origin (Table 4). We also attempted to obtain GTseq genotypes on the fin clips and were successful for 23 of the 26 fin clips (Table 5). We expect to get successful GTseq genotypes for the remaining 3 Platte River fin clips on the next GTseq run.

We also received 3 acipenseriform eggs/embryos collected from the Platte River from a different field project from the one that collected the fin clips. One of the embryos had a visible developing embryonic fish attached while the other two appeared to be undifferentiated eggs. We isolated DNA from the embryos and first genotyped them using a single mitochondrial DNA marker (Kashiwagi et al. 2020) which identified all three as sturgeon and not paddlefish. The "egg" with the visible embryo was identified as a shovelnose sturgeon using both microsatellites and GTseq (individual PLT-082 in Table 5). The other two failed to amplify sufficient DNA for nuclear markers, which is typical of embryos earlier than stage 14 (Kashiwagi et al. 2020).

Comparing GTseq to microsatellites

Similar to the results for the three "borderline" samples in the validation run (Table 3), comparisons of NewHybrid assignments for the Platte River samples based on microsatellites and GTseq demonstrated that GTseq provides much higher resolution (Table 5). All pallid sturgeon had GTseq scores of 1.000 for the pure pallid sturgeon category while the p-values for microsatellites ranged from 0.982 to 1.000. The hybrid sturgeon had a microsatellite p-value of 0.879 for the F1 hybrid category while it's GTseq score was 1.000 for the F1 hybrid category and like the larva discussed in Table 3, it was heterozygous for nearly all S-loci.

Table 4. Genetic identification of Platte River sturgeon fin clips collected in 2022.

UNL ID	Genetic ID	Species	Origin	Mother	Father
UNL-007	16275	Pallid	Hatchery 2007	471A2C1013	412C3D4B11
UNL-009	16276	Pallid	Hatchery 2016	4627201358	48683A3B7D
UNL-010	16280	Pallid	Hatchery 2008	423373582F	432A191F35
UNL-013	16281	Pallid	Hatchery 2008	1F497F1801	1F4849755B
UNL-014	16278	Pallid	Hatchery 2016	4627201358	48683A3B7D
UNL-029	16298	Pallid	Hatchery 2013	460E52494D	43497E1577
UNL-043	16299	Pallid	Hatchery 2015	4718485A15	4626111802
UNL-080	16279	Pallid	Wild	NA	NA
UNL-103	16284	Pallid	Hatchery 2019	4626D2971	48685D1608
UNL-159	16303	Pallid	Hatchery 2001	411D262C1F	411D0E2C5F
UNL-189	16289	Pallid	Hatchery 2005	115676635A	1F50072169
UNL-191	16286	Pallid	Hatchery 2018	4626641923	47191F7F39
UNL-231	16300	Pallid	Hatchery 2015	4718485A15	4626111802
UNL-232	16301	Pallid	Hatchery 2013	412C34470E	4627061C6F
UNL-236	16304	Pallid	Hatchery 2002	116224546A	116167123A
UNL-243	16293	Pallid	Hatchery 2018	462704502D	4715591B05
UNL-250	16294	Pallid	Hatchery 2018	4715674971	434A582F17
UNL-311	16291	Pallid	Hatchery 2002	116224546A	1F477B3A65
UNL-313	16288	Pallid	Hatchery 2018	4626711111	4626773563
UNL-320	16282	Pallid	Hatchery 2008	423373582F	435F60206E
UNL-335	16296	Pallid	Hatchery 2018	46270E6C3C	47041F697D
UNL-340	16302	Pallid	Hatchery 2018	4626641923	47191F7F39
UNL-341	16297	Pallid	Hatchery 2009	412C20001A	486762580F
UNL-342	16283	Hybrid	Wild	NA	NA
UNL-346	16292	Pallid	Hatchery 2018	46270E6C3C	47041F697D
UNL-347	16277	Pallid	Hatchery 2001	220E345E09	7F7D3C5708

Table 5. Comparison of species assignments for Plate River sturgeon using GTseq and microsatellites (µSat) using NewHybrids software. Species categories are pure pallid sturgeon (Pal), pure shovelnose sturgeon (Sho), F_1 hybrid (F1), F_2 hybrid (F2), F_1 backcross to pallid sturgeon (Bxp) and F_1 backcross to shovelnose sturgeon (BxS). Three sturgeon (UNL-029, UNL-043 and UNL-236) failed the first GTseq run and will be re-analyzed. Individual PLT-082 was a free embryo, the rest were fin clips.

Sample ID	Method	Pal	Sho	F1	F2	BxP	BxS
PLT-082	GTseq	0.000	1.000	0.000	0.000	0.000	0.000
PLT-082	µSat	0.000	0.999	0.000	0.000	0.000	0.000
UNL-007	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-007	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-009	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-009	µSat	0.998	0.000	0.000	0.000	0.001	0.000
UNL-010	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-010	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-013	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-013	µSat	0.998	0.000	0.000	0.000	0.002	0.000
UNL-014	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-014	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-029	GTseq	NA	NA	NA	NA	NA	NA
UNL-029	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-043	GTseq	NA	NA	NA	NA	NA	NA
UNL-043	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-080	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-080	µSat	0.996	0.000	0.001	0.001	0.003	0.000
UNL-103	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-103	µSat	0.999	0.000	0.000	0.000	0.000	0.000
UNL-159	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-159	µSat	0.995	0.000	0.001	0.001	0.003	0.000
UNL-189	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-189	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-191	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-191	µSat	0.999	0.000	0.000	0.000	0.000	0.000
UNL-231	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-231	µSat	0.999	0.000	0.000	0.000	0.000	0.000

Table 5. (continued)

Sample ID	Method	Pal	Sho	F1	F2	BxP	BxS
UNL-232	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-232	µSat	1.000	0.000	0.000	0.000	0.000	0.000
UNL-236	GTseq	NA	NA	NA	NA	NA	NA
UNL-236	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-243	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-243	µSat	0.982	0.000	0.006	0.001	0.010	0.000
UNL-250	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-250	µSat	0.999	0.000	0.000	0.000	0.000	0.000
UNL-311	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-311	µSat	0.996	0.000	0.000	0.001	0.003	0.000
UNL-313	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-313	µSat	0.998	0.000	0.000	0.000	0.001	0.000
UNL-320	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-320	µSat	0.997	0.000	0.000	0.000	0.002	0.000
UNL-335	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-335	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-340	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-340	µSat	0.999	0.000	0.000	0.000	0.001	0.000
UNL-341	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-341	µSat	0.999	0.000	0.000	0.000	0.000	0.000
UNL-342	GTseq	0.000	0.000	1.000	0.000	0.000	0.000
UNL-342	µSat	0.000	0.068	0.879	0.030	0.019	0.004
UNL-346	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-346	µSat	0.999	0.000	0.000	0.000	0.000	0.000
UNL-347	GTseq	1.000	0.000	0.000	0.000	0.000	0.000
UNL-347	µSat	0.999	0.000	0.000	0.001	0.000	0.000

Prospects for future work – We now have a working panel of GTseq markers that provide much greater resolution of species than has previously been available. In future seasons we will work closely with the UNL field crews to provide timely information to aid their research. When fin clips are received, we will determine the microsatellite genotypes within 2 working days, which will provide a tentative species identification and identification hatchery/wild status. We believe that we can produce GTseq genotypes and analyses within 5 working days. To fill out our sequencing runs for the Platte River samples, we will use Missouri River samples that have already been collected and whose genotyping is funded by the Army Corps of Engineers. The Platte and Missouri River samples will be analyzed to fulfill the project objectives following the timeline described below (Table 6).

Table 6. Anticipated timeline for completion of project goals

1. Marker development/validation – December 2022
2. Refined species ID and baselines – December 2023
3. Population structure/Redefine management units – June 2024
4. Population composition by species/hybrid – June 2025
5. Demographics – pallid sturgeon N_e by population – December 2025
6. Final report to PRRIP – June 2026

Literature cited

Anderson, E. C., and E. A. Thompson. 2002. A model-based method for identifying species hybrids using multilocus genetic data Genetics 160:1217-1229.

Campbell, N. R., S. A. Harmon, and S. R. Narum. 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.

Flamio, R., D. G. Swift, D. Portnoy, A. DeLonay, K. A. Chojnacki, J. Powell, P. J. Braaten, and E. J. Heist. in press. Disomic marker development in the paleotetraploid sturgeons, *Scaphirhynchus albus* and *S. platorynchus*, facilitated by sequencing functional haploids. Molecular Ecology Resources.

Kalinowski, S. T., M. L. Taper, and T. C. Marshall. 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology 16(5):1099-1106.

Kashiwagi, T., A. J. DeLonay, P. J. Braaten, K. A. Chojnacki, R. M. Gocker, and E. J. Heist. 2020. Improved genetic identification of acipenseriform embryos with application to the endangered pallid sturgeon *Scaphirhynchus albus*. Journal of Fish Biology 96(2):10.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5):11.

Schrey, A. W., B. L. Sloss, R. J. Sheehan, R. C. Heidinger, and E. J. Heist. 2007. Genetic discrimination of middle Mississippi River *Scaphirhynchus* sturgeon into pallid, shovelnose, and putative hybrids with multiple microsatellite loci. Conservation Genetics 8(3):683-693.

Waples, R. S., R. K. Waples, and E. J. Ward. 2022. Pseudoreplication in genomic-scale data sets. Molecular Ecology Resources 22(2):503-518.