Range-wide genetic assessment of Loach Minnow (*Tiaroga cobitis*) and Spikedace (*Meda fulgida*)

Agreement Number: R18AC00069

Submitted by: Thomas Turner, Megan Osborne, and Alexander Cameron

Annual Report submitted for period ending 31 March 2020

Submitted to:

Bill Stewart
Gila River Basin Native Fish Program Manager
U.S. Bureau of Reclamation
6150 W. Thunderbird Rd.
Glendale, AZ 85306
Phone: 623-773-6252
Email: wstewart@usbr.gov

Jeff Anderson
U.S. Bureau of Reclamation
PO Box 61470
Boulder City, NV 89006
Phone: 702-293-8458
Email: jranderson@usbr.gov

Report submitted: 22 June 2020

Executive Summary

This project uses single-nucleotide polymorphisms from genomic DNA sequence data to characterize range-wide genetic variation in Loach Minnow (*Tiaroga cobitis*) and Spikedace (*Meda fulgida*), including broodstock held at the Arizona Aquatic Research and Conservation Center. Below are specific tasks outlined in the agreement and a brief description of progress to date. The text that follows provides more detail regarding project status and remaining tasks. The project is on schedule for completion on 9/30/2021.

1) Collect tissue samples from specified localities. **Status-Complete**. Localities for both species are listed in Table 1.

2) Isolate DNA from fin clips from up to 700 fish. **Status-Complete**. Genomic DNA was isolated from 316 Loach Minnow and 285 Spikedace. Sample sizes for each sampled population are listed in Table 1.

3) Conduct nextRAD DNA library preparation and next-generation DNA sequencing to discover and characterize single nucleotide polymorphism (SNP) loci. **Status-Complete**. Isolates were sent to SNPsaurus on 20 March 2019 and sequences were received on 11 August 2019.

4) Conduct bioinformatic analyses to verify and quality-check SNP genotypes and consistency across localities and samples. **Status-Ongoing**. Data sets for both species are in the final stages of bioinformatic QA/QC analyses that are necessary prior to data analysis.

5) Characterize genetic variation among wild, repatriated, and broodstock populations. **Status-Ongoing**. Exploratory analyses are currently being conducted for Spikedace. Population summary statistics will be reported once the final data set is assembled with the last iteration of the genome.

Sample Collection, DNA Isolation, and nextRAD Sequencing

Field sampling took place throughout the Gila River basin in Arizona and New Mexico during October 2018 through February 2019 (Table 1). At each site, a 1-km river reach was sampled at three distinct points (0 m, 500 m, and 1,000 m) by single-pass electrofishing, seining and shock-seining (seine dimensions = 6-m width X 2-m height and 9.5-mm mesh). Tissue samples were taken from the lower lobe of the caudal fin from up to 30 Loach Minnow and Spikedace encountered during sampling. These were preserved in 95% EtOH. Tissue holdings from the Museum of Southwestern Biology were also incorporated into the current study to represent an additional site that was not sampled in 2018, and to evaluate temporal stability of populations (Table 1). Finally, fin clips of hatchery broodstock from four genetic lineages of Loach Minnow and three genetic lineages of Spikedace maintained at the Arizona Aquatic Research and Conservation Center were collected in February 2019.

For both species, genomic DNA was isolated from up to 30 individuals per population using the E-Z96 Tissue DNA kit (Omega Bio-Tek) and E.Z.N.A. Tissue DNA kit (Omega Bio-Tek) following the manufacturer's instructions and included a RNAse_A treatment. Isolate quality was confirmed via visual inspection on a 1.5% agarose gel and quantified via QUBIT assays (ThermoFisher). A total of 316 Loach Minnow and 285 Spikedace DNA isolates were sent to the SNPsaurus facility for sequencing. Genomic DNA samples were fragmented with Nextera reagent (Illumina, Inc) and short adapter sequences were ligated to the ends of fragments. The Nextera reaction was scaled to accommodate 30 ng of genomic DNA, although 45 ng of genomic DNA was used for input to compensate for degraded DNA in the samples and to increase fragment sizes in sequencing reactions. Fragmented DNA was then amplified for 27 cycles at 74 degrees, with one of the primers matching the adapter and extending 10 nucleotides into the genomic DNA with the selective sequence GTGTAGAGCC. Thus, only fragments starting with a sequence that can be hybridized by the selective sequence of the primer are efficiently amplified. NextRAD DNA libraries were sequenced on an Illumina HiSeq 4000 with one lane of 150 base pair (bp) single-end reads (University of Oregon) per 106 samples for Loach Minnow and 95 samples for Spikedace.

Bioinformatics

A major challenge in population genomic studies of non-model organisms like Spikedace and Loach Minnow is that bioinformatic computational steps (called pipelines) have the potential to influence downstream results. Therefore, best practice strategies include generating summary statistics through multiple pipelines and compare results to confirm that population genetic inferences are robust (Shafer et al. 2017). Differences in genomic resources available for Loach Minnow (a well-annotated genome is not yet available, but pending) and Spikedace (a well-annotated genome is available) requires that different approaches are taken for each species. Details regarding bioinformatic processing and project status for each species is outlined in the following sections.

Loach Minnow

Currently, a full genome sequence is not available for Loach Minnow or for a closely-related species whose genome could serve as a surrogate reference. To address this issue, we prepared paired end libraries (150 base pair; bp) for two individuals from the Gila Forks broodstock that were sequenced across four lanes on an Illumina NextSeq 500 (University of New Mexico). Estimated sequencing depth was roughly 29x per individual assuming a genome length of 1.1 gigabases (*i.e.*, each nucleotide within the genome was sequenced 29 times). Cleaned paired reads for both individuals were assembled into contiguous sequences (contigs) using the multiple *k-mer* strategy via MEGAHIT (Li et al. 2015). The resulting assembly consisted of 403,616 contigs that ranged in size from 200 to 50,859 bp. We then mapped trimmed reads to for all Loach Minnow to reference contigs using BWA-MEM (Li and Durbin 2009; Li 2013). Binary alignment/MAP files were then fed into the *dDocent* (Puritz et al. 2014) bioinformatic pipeline utilizing the *ddocent*.FB module (*FreeBayes*; Garrison and Marth 2012). The resulting variant call format (VCF) file was subjected to an eight-step quality filtering process. First, sites with a genotype with a call rate < 50%, minimum allele count < 3, minimum read depth < 5 and quality score below 30 were removed. Individuals with >35% missing data for were removed from the dataset, followed by removing loci that had > 10% missing data in any given population. Complex variants were decomposed with remaining indels and multiallelic sites subsequently removed by computational filtering. A series of filters described in O'Leary et al. (2018) for identifying false SNPs, multicopy loci and paralogs were applied (e.g., allele balance, locus quality/depth ratio, strand bias and mapping quality ratio). Finally, loci were filtered for significant deviations from Hardy-Weinberg Equilibrium (*p*-value = 0.001) within each population and sites with a minor allele frequency of less than 0.05 across all population were removed. The post-filtering vcf contained 1,500 bi-alleic SNPs across 298 individuals with 5% missing data. However, these reference contigs represent a very rough draft of the Loach Minnow genome, as reference assemblies typically are sequenced to a depth of at least 50x and incorporate long-read sequences. Thus, we explored additional approaches toward maximizing the accuracy of assembly, alignment, and genotyping.

We are currently in the process of building a *de novo* assembly using the software *stacks* (Catchen et al. 2013) that is regarded as the industry standard for short-read assembly in non-model organisms lacking genomic resources (Shafer et al. 2017). Briefly, alleles are identified in identical sequences (or 'stacks') within individuals, which are then merged into loci within individuals and then across individuals. (Catchen et al. 2011; Paris et al. 2017). Assembly parameters are currently being optimized across 3 individuals from all 2018 collections (n=36) following the procedure outlined in Paris et al. (2017). The goal of optimization is to select the suite of parameters that maximizes the number of polymorphic loci retained in 80% of samples. The rational being that these loci are highly replicated across individuals thus are unlikely to be derived from paralogous sequences, repetitive sequences, or sequences that contain a high degree of genotyping error (Paris et al. 2017). Once an optimal set of parameters is determined, all individuals will be incorporated into assembly. Finally, once a set of *de novo* loci are constructed, we will implement a *de novo-integrated* approach whereby *stacks* consensus sequences are mapped to the independently generated reference contigs. This approach will aid in identification and removal of erroneous loci.

Spikedace

Trimmed reads were mapped to the second iteration of the Spikedace genome using BWA-MEM (Li and Durbin 2009; Li 2013). Binary alignment/MAP files were then fed into the *dDocent* (Puritz et al. 2014) pipeline and genotypes were called with the *ddocent*.FB module (*FreeBayes*; Garrison and Marth 2012). An initial 3,831,877 variant positions were identified across all 286 individuals with *FreeBayes*. The eight-step filtering process described for Loach Minnow was applied and resulted in the retention of 5,892 bi-allelic SNPs across 260 individuals with 5% missing data post-filtering. We tested for presence of linkage disequilibrium among filtered SNPs using the R package *GUSLD* (Bilton et al. 2018) to calculate r^2 values between pairs of SNPs with default parameter settings. A total of 62 pairs of SNPs exhibited an r^2 value $\geq$ 0.65 and one SNP per pair was pruned from the data set leaving 5,836 SNPs. Exploratory population genetic analyses are currently ongoing with this data set. However, the third iteration of the Spikedace genome is in the final stages of chromosome-level assembly and when complete, the analyses described above will be re-run. The third iteration of these data will provide greatest resolution and will allow for identification of SNPs residing in functional genes.

We are also exploring a second approach for generating genotype data for Spikedace using the software ANGSD (Korneliussen 2014). This approach differs from conventional pipelines in that population genetic analyses can be performed directly on genotype likelihoods (GL), where GL refers to the marginal probability of the sequencing data given a genotype in a particular individual, at a particular site (Korneliussen 2014). Working directly with GLs can offer a statistical advantage for low and medium coverage data sets, such as Spikedace. Currently, we have conducted the first quality control step for this software and are determining the appropriate method for estimating GLs.

Table 1. Spikedace and Loach Minnow locality information. Sample size reflects the number of individuals sequenced for each population.

Species	N	Locality	County	State	Collection Date
Meda fulgida	30	Spring Creek	Yavapai	AZ	Oct 2018
	14	Fossil Creek	Yavapai	AZ	Oct 2018
	27	Aravaipa Creek	Graham	AZ	Oct 2018
	30	Blue River	Greenlee	AZ	Oct 2018
	30	TNC Riverside - Gila Mainstem	Grant	NM	Oct 2018
	30	West Fork Gila River	Catron	NM	Oct 2018
	30	Bird Area - Gila Mainstem	Grant	NM	Oct 2017
	4	West Fork Gila River 2012	Catron	NM	Oct 2012
	30	Brood stock-Gila Mainstem			Feb 2019
	30	Brood stock-Gila Forks			Feb 2019
	30	Brood stock-Aravaipa Creek			Feb 2019
Tiaroga cobitis	28	Aravaipa Creek	Graham	AZ	Oct 2018
	12	Blue River	Greenlee	AZ	Oct 2018
	11	Campbell Blue River	Greenlee	AZ	Oct 2018
	30	TNC Riverside – Gila Mainstem	Grant	NM	Oct 2018
	30	West Fork Gila River	Catron	NM	Oct 2018
	1	Little Creek	Grant	NM	Oct 2018
	19	Bird Area – Gila Mainstem	Grant	NM	Oct 2017
	10	San Francisco River	Catron	NM	Nov 2018
	30	Tularosa River	Catron	NM	Feb 2019
	13	San Francisco River 2009	Catron	NM	June 2009
	15	TNC Riverside – Gila Mainstem 2009	Grant	NM	June 2009
	30	Brood stock - Aravaipa Creek			Feb 2019
	27	Brood stock - Blue River			Feb 2019
	30	Brood stock - Gila Forks			Feb 2019
	30	Brood stock – San Francisco River			Feb 2019

Literature Cited

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), 289-400. doi:10.1534/genetics.118.300831

Catchen, Julian, et al. "Stacks: an analysis tool set for population genomics." *Molecular ecology* 22.11 (2013): 3124-3140.

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

Garrison E. 2016. Vcflib, a simple C++ library for parsing and manipulating VCF fileshttps://github.com/vcflib/vcflib.

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

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, *15*(1), 356.

Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25: 1754-1760.

Li D, Liu CM, Luo R, Sadakane K, and Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31: 1674-1676.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, and Portnoy DS. 2018. These aren’t the loci you’re looking for: Principles of effective SNP filtering for molecular ecologists. Molecular Ecology, 27: 3193-3206.

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution*, *8*(10), 1360-1373.

Puritz JB, Hollenbeck CM, and Gold JR. 2014. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ, 2, e431.

Shafer, A. B., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution, 8*(8), 907-917.

Willis SC, Hollenbeck CM, Puritz JB, Gold JR, and Portnoy DS. 2017. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. Molecular Ecology Resources, 17: 955-965.