Development of Rio Grande Sucker (*Pantosteus plebeius*) Genomic Tools

Report for NMDOW-UNM Agreement #251124

30 June 2026

Submitted by:

Ashley Willis Mascareñas, Megan Osborne, and Thomas Turner
Department of Biology & Museum of Southwestern Biology
University of New Mexico
Albuquerque, NM 87131
505-277-3234
Emails: awillis5@unm.edu, mosborne@unm.edu, turnert@unm.edu

Submitted to:

Virginia Seamster, Ph.D.
Assistant Chief for Technical Guidance
Wildlife Management Division
New Mexico Department of Wildlife
1 Wildlife Way
Santa Fe, NM 87507

Project Background

Rio Grande Sucker (RGS) is classified as a species of immediate priority and Species of Greatest Conservation Need (SGCN) by the State of New Mexico (NMDOW 2025) and is listed as endangered and a Tier 1 SGCN in the State of Colorado. The species is sensitive to habitat alteration, fragmentation, and water extraction. Negative effects of these factors are likely to get worse with climate change and the species will require adaptive management to ensure its persistence in New Mexico waterways.

Genetic information is a linchpin of conservation and management because it can be used to evaluate current population status and future persistence probabilities. Baseline genetic information (2008-2022) from microsatellites and mitochondrial DNA documented the spatial distribution of genetic diversity of RGS across its entire geographic range in the United States (McPhee et al. 2008, Turner et al. 2019, 2022). Additional genetic monitoring that involved tracking genetic diversity and effective population size (N_e) across temporally spaced samples showed important changes in genetic diversity and abundance in response to catastrophic wildfire (Turner et al. 2015). Newer sequencing technology can characterize genetic variation at single nucleotide polymorphisms (SNPs) distributed across the entire genome (Brumfield et al. 2003), including loci that are presumably targets of natural selection or otherwise involved in determining organism performance. High-throughput Next Generation Sequencing (NGS) methods make it relatively straightforward to identify sufficient numbers of SNPs for robust assessment of population structure, biogeographic processes, and routine genetic monitoring (Hess et al. 2011).

Reduced representation sequencing methods, like the Nextera-tagmented reductively-amplified DNA sequencing (nextRAD-seq) approach that we used in this study, are cost-effective ways to identify thousands of SNPs across hundreds of samples (Russello et al. 2015). When a reference genome for the focal species is available, SNP identification is more accurate, especially compared to using a reference genome from a more distantly related organism (Huang et al. 2024). More importantly, a reference genome can be used to predict extinction risk and adaptation potential by providing information on genomic organization and variation in gene copies (i.e., genomic architecture), selection, and hybridization. This information is important for a more complete interpretation of spatial and temporal patterns of SNP diversity and provides a

valuable resource to inform conservation and management planning (Hudgell et al. 2026). A reference genome from the focal species is also important for development of future genetic tools like GT-seq (Genotyping-in-Thousands by sequencing; Campbell et al. 2015) panels that allow for parentage-based tagging and other applications by genotyping hundreds of SNPs at once. For these reasons, development of a high-quality reference genome for RGS was a high priority for this project.

Accordingly, we developed several key genomic resources for RGS. Project objectives were to:

(i) Isolate DNA from three Rio Grande Sucker (RGS) tissue samples, conduct quality-control assays prior to library preparation and genome sequencing.
Status: *Complete – samples from Museum of Southwestern Biology archives were used as templates for long-read PacBio* sequencing.

(ii) Use the resulting data from *PacBio* sequencing to generate a long-read genome and conduct chromosome conformation capture (Hi-C) sequencing to determine chromosomal structure.
Status: *In progress* – The *PacBio* sequencing data for the Jemez and Mimbres RGS samples have received, and we are awaiting data for the Gila sample. Data were received from University of California San Diego gCore for the Jemez lineage sample. Bioinformatic analysis is underway with preliminary results discussed below.

(iii) Use appropriate software and filtering steps to map the Nextera-tagmented reductively amplified DNA (nextRAD) data obtained from 380 RGS tissue samples previously collected from 2009 to 2025 from fish in New Mexico and Arizona against the three long-read RGS genomes to identify genetic variants. Use appropriate tools and analyses to filter variants and loci and create a robust, genome-wide single nucleotide polymorphism (SNP) dataset that can ultimately be used for future genetic monitoring to support conservation and adaptive management of this species.

Status: *In progress.* DNA was isolated from samples listed in Table 1; samples were shipped and are being sequenced using nextRAD technology by Plasmidsaurus (https://plasmidsaurus.com/) .

Methods

Genome sequencing and assembly

We obtained high molecular weight (HMW) DNA from Rio Grande sucker (N=3) from three native lineages present in New Mexico waters: (1) Jemez River – Rio Grande lineage; (2) Mimbres River – Mimbres lineage; and (3) Centerfire Creek – Gila lineage. These samples were prepared and submitted for high-quality long-read genome sequencing. HMW DNA was isolated using Qiagen® genomic tips G/20 according to the manufacturer's directions. Double stranded DNA was quantified using Qubit® (Thermo Fisher Scientific) fluorometer assays to ensure that sufficient HMW DNA was available for library preparation. DNA samples were sent to the University of California Davis (UC Davis) Genomics Facility (https://genomecenter.ucdavis.edu/) for library preparation and sequencing. For each RGS sample, a genomic library was prepared using PacBio HiFi SMRTbell® and the whole genome was sequenced using one Revio SMRT cell on a PacBio Sequel II platform. Additional tissue samples from a different Jemez RGS were sent to the University of California San Diego (UC San Diego) Genomics Core (https://gcore.ucsd.edu/) for dual-ended chromosome conformation capture (Hi-C) library preparation and sequencing. The library was prepared using streptavidin beads and the genome was sequenced using a PE150(2 x 150 bp) flow cell.

PacBio long-reads sequenced from RGS (Rio Grande and Mimbres lineages) were received from UC Davis and a *de novo* genome was assembled using HiFiasm v. 0.25.1-r466 (Cheng et al. 2021) with haplotype number set to four as suckers are tetraploid. Genome contiguity was assessed using the contig N50 value (i.e., the sequence length for the longest contigs that contain 50% of the total genome length). Genome contiguity and completeness were assessed using Quast 5.2.0 (Mikheenko et al. 2018) and BUSCO v.6.1.0 (Manni et al. 2021) specifying the Actinopterygii Ortholog set containing 3,640 genes.

Genome sequences were received from UC San Diego and we are completing Hi-C assembly using HiFiasm v. 0.25.1-r466 (Cheng et al. 2021) with the same parameters used for the PacBio

long-reads (Hudgell et al. 2026). Genome quality and completeness will be assessed as described above. The assembly will be scaffolded using the software Juicer (Durand et al. 2016), YaHS (Zhou et al. 2023), and Juicebox (Durand et al. 2016) to create a chromosome-resolved, *de novo* genome of RGS from the Jemez.

Sampling and reduced-representation sequencing via nextRAD

For nextRAD sequencing, we purified previously isolated DNA using Zymo® DNA clean and concentrator kits following manufacturer's directions from samples collected throughout the range of RGS (Table 1). DNA was also isolated from samples collected in 2025 from the Upper San Francisco basin (New Mexico; NMDOW Scientific Collectors Permit 3015) using the Axygen genomic DNA isolation kit following manufacturer's instructions. Double-stranded DNA concentrations were quantified for all samples using Qubit® (Thermo Fisher Scientific) fluorometer with a 1X dsDNA HS Assay Kit. A total of 380 samples with sufficient high-quality DNA were selected for reduced-representation sequencing (Table 1) and sent to Plasmidsaurus (https://plasmidsaurus.com/) for sequencing. Plasmidsaurus employs a proprietary Nextera-tagmented reductively-amplified DNA sequencing method (nextRAD-seq) that reduces the proportion of a genome to be sequenced by adding selective primers to DNA fragments as described by Russello et al. (2015). Fragmentation of DNA is performed with a Nextera Reagent from Illumina® and adapter sequences are ligated to the fragments (Russello et al. 2015). Genomic libraries prepared by SNPsaurus were sequenced on a Novaseq 6000 with two S4 lanes of 150 base pair (bp) long paired-end reads.

Identification of SNPs

NextRAD sequencing is nearly complete. Our plan is to analyze DNA sequences as described below. When raw DNA sequences (or reads) are received, adapters and DNA barcodes will be removed and demultiplexed (by individual) using a standard bioinformatic pipeline and workflow. Specifically, raw reads are trimmed with Trimmomatic v. 0.39 (Bolger et al. 2014) software using a quality threshold of 20 base pairs on leading and trailing ends to remove low-quality and ambiguous base pairs. Each read is screened with a 5-base sliding window and cut when the average quality per base drops below PHRED = 10. A PHRED score is a numerical measure of the accuracy of a DNA base call, indicating the likelihood that a nucleotide is called

incorrectly. It is often used to establish thresholds for quality in bioinformatic workflows. Raw reads smaller than 60 base pairs in length are also discarded.

Trimmed reads are aligned to the reference RGS genome with Bowtie v. 2.5.1 software (Langmead & Salzberg 2012) using paired-end input and a *very-sensitive-local* option. Next SAMtools v. 1.16.1 software (Danecek et al. 2021) is used to remove reads with mapping quality lower than PHRED = 20 and to convert SAM files to BAM files. Picard Tools v. 3.1.0 (The Broad Institute, 2025) is used to add reading groups (RG), sort BAM files, remove polymerase chain reaction (PCR) duplicates, and merge all BAM files into a single file containing all reads from all samples. Mapped regions with less than 150x depth of coverage across all samples are removed using BEDTools v. 2.30.0 (Quinlan et al. 2010). Coverage is the number of aligned sequencing reads overlapping a genomic region across all samples. Variants are identified with FreeBayes v. 1.3.6 (Garrison and Marth 2012) following the parallel scheme implemented in dDocent pipeline version 2.7.8 (Puritz et al. 2014). FreeBayes uses base quality scores to estimate a probability for each allele. We will retain a maximum of 10 raw variants from each alignment with higher probabilities and at least a base quality of five.

To remove erroneous or potentially erroneous variants, VCFtools v. 0.1.16 (Danecek et al. 2011) is used to filter out (i.e., remove from the dataset) variants with mean depth of coverage lower than 20 and higher than 200, minor allele count less than three, minor allele frequency lower than 5%, genotype depth of coverage lower than five, and with genotype quality lower than 20. Multi-nucleotide states are decomposed into single variants with VCFtools (Danecek et al. 2011) and VCFtools is used to filter out nucleotide insertions and deletions and to retain only the bi-allelic SNPs. The dataset is then filtered by missing data, keeping SNPs present in at least 80% of samples and removing individuals with more than 30% missing data.

Next, SNPs are filtered using the bash script *dDocent_filters* (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) that uses vcflib v. 1.0.9 (Garrison et al. 2022) and VCFtools to filter loci based on allelic balance at heterozygous genotypes, strand representation, and quality versus depth of coverage. First, loci are removed if, at heterozygous positions, the alternative allele has a depth of 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 are not removed to account for fixed alleles. Additionally, if the quality sum of the reference or alternative allele is zero, the locus is removed. This removes positions with spurious heterozygous genotype calls. Then loci with the ratio between the mean mapping quality of the alternative and reference allele lower than 0.25 or higher than 1.75 are removed, because loci from the same genomic location should not have large discrepancy between mapping qualities of two alleles. Furthermore, loci with quality scores less than half of the total depth are excluded because excessive depth inflates FreeBayes quality scores. Of the remaining loci, the average depth and standard deviation across all individuals are calculated. Loci with depth greater than the average depth plus one standard deviation are removed if the quality score is less than two times the depth. Finally, this script removes loci with a mean depth across individuals greater than two times the mode that corresponds approximately to the 95th percentile of mean depth. This removes loci with unusually high sequencing depth. Subsequently potential erroneous SNPs are then 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). SNPs present in more than 50% of the populations (here each sampling locality is considered a 'population') and with a deviation from HWE p-value lower than 0.001 are removed. Potential incorrectly assembled paralogous loci that exhibited a large variation in read depth across all individuals are then removed. Standard deviation is estimated with the R package *stats* and read depth with VCFtools. R v. 4.4.3 (R Core Team 2025). RStudio 2024.12.1.563 (Posit Team 2025) is used to conduct these analyses. Additional filtering based on missing data per locus (keeping loci present in 80% of individuals) is then reapplied.

Remaining SNPs are used to identify haplotypes within loci (referred to as microhaplotypes). Haplotyping SNPs within a locus also eliminates possible paralogous loci by neutralizing physical linkage without losing data (Willis et al. 2017). We use the *rad_haplotyper.pl* pearl script (https://github.com/chollenbeck/rad_haplotyper) excluding microhaplotypes if considered paralogs at least in five individuals and if missing from more than 30% of individuals.

Retained loci are tested for deviations from HWE and for linkage disequilibrium (LD) considering individuals captured in each locality as a different 'population'. Departures from HWE are 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 2025), as implemented in the R function multi_HWE_tests (https://github.com/gcaeirodias/multi HWE tests; Caeiro-Dias et al. 2026). Estimations of LD are performed on SNP data using the SNP of each microhaplotype with higher minimum allele frequency. If a SNP is found in LD, then the entire locus is removed. Tests for LD are 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 as implemented in the R pipeline significantLD (https://github.com/gcaeirodias/significantLD; Caeiro-Dias et al. In Press). If loci are found in multiple significant LD pairs, the loci that appeared in the highest number of comparisons are discarded to keep the maximum number of loci possible. In the remainder of instances, one locus from each pair is discarded randomly. Loci will be considered as deviating from HWE and to be in LD if tests are significant across all sampled localities (p-value < 0.05). After all these filtering steps, the resulting data should represent a robust, genome-wide, neutral SNP dataset that can be used to reliably estimate genetic diversity, effective population size, and population structure.

Results

Genome Sequencing

The Jemez RGS produced a total of 5,724,254 raw HiFi reads, containing 102 Giga bases (Gb) of sequence data. The contig assembly resulted in 160 sequences ranging from a least 5,000 to 94,104,479 bp (mean = 13.7 Mb) with a total length of 2,186,257,691 bp and an N50 of 43,388,746 bp. This corresponds to a 46.4x genome coverage if we assume that the total assembly size is similar to the true genome size, which is expected to be about 2.2 Gb based on knowledge of genomes of other related species (Yang et al. 2024). The assembly contained 99.8% complete Benchmarking Universal Single-Copy Orthologs (BUSCOs), which are a set of single-copy, highly conserved genes, that are expected to be found in the genome of Actinopterygians. Of these, 51.5% were single-copy and 48.3% duplicated (expected for

polyploid genomes), with 0.2% missing. These results indicate a highly complete and good quality reference genome.

Table 1. Number of samples by locality chosen for nextRAD sequencing.

Locality	**N**
Jemez River	20
Bluewater Creek	20
Hot Creek	20
Rio Bonito	22
Sapillo Creek	56
Alamosa Creek	29
Mimbres	23
Rio Embudo	37
Rio Tusas	31
Trout Creek	32
Seco Creek	23
Centerfire Creek	30
Upper San Francisco River	37
Total	380

Conclusions

Despite the short performance period of this project (e.g., February through June 2026), we were able to complete a significant portion of the proposed work and obtain sufficient genomic resources to support future understanding of spatial patterns of genomic variation across the entire known range of RGS. Furthermore, these data meet all requirements necessary for future work aimed at designing and implementing a GT-seq panel of SNP loci. GT-seq coupled with ecological sampling is a powerful tool for demographic monitoring, parentage and pedigree estimation, and close-kin mark-recapture applications to estimate abundance and dispersal. The data generated in this project will represent a major portion of a doctoral dissertation for a

participating graduate student, Ashley Willis Mascareñas, and will form the basis for much future work focused on this SGCN.

Acknowledgements

Funding for the project was provided by the Land of Enchantment Legacy Fund through the New Mexico Department of Wildlife agreement #251124. We thank Virginia Seamster for contract management and editing assistance. We thank the Genome Center at UC Davis and the Stem Cell Genomics Core at the Sanford Stem Cell Institute at UC San Diego for providing sequencing services.

References

Bilton TP, McEwan JC, Clarke SM, Brauning R, van Stijn TC, Rowe SJ, Dodds KG. (2018). Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*, 209(2):389-400.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114-2120.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, 18(5):249-56.

Campbell NR, Harmon SA, Narum SR. (2015). Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, 15(4):855-867.

Caeiro-Dias G, Osborne MJ, Turner TF. (2026). Time is of the essence: using archived samples in the development of a GT-seq panel to preserve continuity of ongoing genetic monitoring. *PeerJ*, 13:e20726

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2):170-175.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo, MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin K, 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, 3(1):95-98.

Durand N C, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems*, 3(1):99-101.

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

Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. (2022). A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Computational Biology*, 18(5):e1009123.

Hess JE, Matala AP, S. R. Narum. (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

Huang PH, Wang TR, Li M, Fang OY, Su RP, Meng HH, ... Li J. (2024). Different reference genomes determine different results: comparing SNP calling in RAD-seq of *Engelhardia roxburghian*a using different reference genomes. *Plant Science*, 344:112109.

Hudgell MAB, Osborne MJ, Salinas I, Turner TF. (2026). A chromosome-level genome for Gila trout (*Oncorhynchus gilae*) provides a novel resource for conservation and management. *BMC Genomics.*

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357-359.

Mangiafico SS. (2025). rcompanion: functions to support extension education program evaluation, version 2.5.0. Rutgers Cooperative Extension, New Brunswick, New Jersey, USA.

Manni M, Berkeley MR, Seppey M, Zdobnov EM. (2021). BUSCO: assessing genomic data quality and beyond. *Current Protocols*, 1:323.

McPhee MV, Osborne MJ, Turner TF. (2008). Genetic diversity, population structure and demographic history of the Rio Grande sucker, *Catostomus plebeius*, in New Mexico. *Copeia*, 2008:189-197.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 34(13):i142-i150.

New Mexico Department of Wildlife. (2025). State Wildlife Action Plan for New Mexico. New Mexico Department of Wildlife, Santa Fe, New Mexico, USA.

Paradis E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3):419-420.

Posit team. (2025). RStudio: integrated development environment for R. Posit Software, PBC, Boston, Massachusetts, USA. <http://www.posit.co/>.

Puritz JB, Hollenbeck CM, Gold JR. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2:431.

Quinlan AR, Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841-842.

R Core Team. (2025). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, AUT. <https://www/R-project.org>.

Russello MA, Waterhouse MD, Etter PD, Johnson EA. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ*, 3:e1106.

Broad Institute. (2019). *Picard toolkit* (Version 3.1.0) [Computer software]. https://broadinstitute.github.io/picard/Turner TF, Osborne MJ, McPhee MV, and Kruse CG. (2015). High and dry: intermittent watersheds provide a test case for genetic response of desert fishes to climate change. *Conservation Genetics*, 16:399-410.

Turner TF, Pilger TJ, Osborne MJ, and Propst D L. (2019). Rio Grande sucker *Pantosteus plebeius* is native to the Gila River basin. *Ichthyology and Herpetology,* 107:393-403.

Turner TF, Cameron AC, Osborne MJ, Propst DL. (2022) Origins and diversity of peripheral populations of Rio Grande sucker *Pantosteus plebeius* in the southwestern United States. *Southwestern Naturalist*, 66:25-34.

Willis SC, Hollenbeck CM, Puritz JB, Gold JR, Portnoy DS. (2017). Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*, 17(5):955-965.

Yang L, Mayden RL, Naylor GJ. (2024). Phylogeny and polyploidy evolution of the suckers (Teleostei: Catostomidae). *Biology*, 13(12):1072.

Zhou C, McCarthy SA, Durbin R. (2023). YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*, 39(1):btac808.