The Endangered White Sands pupfish (Cyprinodon tularosa) genome reveals low diversity and heterogenous patterns of differentiation

Andrew Black[1], Janna Willoughby[2], Anna Brüniche-Olsen[3], Brian Pierce[4], and Andrew DeWoody[1]

[1]Purdue University
[2]Auburn University
[3]University of Copenhagen
[4]Texas A and M University College Station

November 24, 2020

Abstract

The White Sands pupfish (Cyprinodon tularosa), endemic to New Mexico in Southwestern North America, is of conservation concern due in part to invasive species, chemical pollution, and groundwater withdrawal. Herein, we developed a high quality draft reference genome and use it to provide biological insights into the evolution and conservation of C. tularosa. Specifically, we localized microsatellite markers previously used to demarcate Evolutionary Significant Units, evaluated the possibility of introgression into the C. tularosa genome, and compared genomic diversity among related species. The de novo assembly of PacBio Sequel II error-corrected reads resulted in a 1.08Gb draft genome with a contig N50 of 1.4Mb and 25,260 annotated protein coding genes, including 95% of the expected Actinopterigii conserved orthologs. Many of the previously described C. tularosa microsatellite markers fell within or near genes and exhibited a pattern of increased heterozygosity near genic areas compared to those in intergenic regions. Genetic distances between C. tularosa and the widespread invasive species C. variegatus, which diverged ~1.6-4.7 MYA, were 0.027 (nuclear) and 0.022 (mitochondrial). Nuclear alignments revealed putative tracts of introgression that merit further investigation. Genome-wide heterozygosity was markedly lower in C. tularosa compared to estimates from related species, likely because of smaller long-term effective population sizes constrained by their isolated and limited habitat. These population inferences, generated from our new genome assembly, provide insights into the long term and contemporary White Sands pupfish populations that are integral to future management efforts.

Introduction

The integration of genomics into conservation biology has enabled the incorporation of genome-wide diversity metrics within the context of the demographic history of threatened and endangered populations (Leroy*et al.* 2018). For example, whole-genome sequence analysis of Grey wolf (*Canis lupus*) revealed patterns of low heterozygosity and high inbreeding depression, despite considerable amounts of detected historical introgression from domesticated dog (*C. lupus familiaris* ; Gómez-Sanchez *et al.* 2018). Such genome-wide assessments can provide key insights into historical population dynamics and are critical for guiding contemporary management actions. While overall genetic diversity is a key predictor of long-term population persistence, conservation geneticists have historically relied on tools that provide data from a limited number of anonymous, putatively neutral genetic markers (e.g. microsatellites) to quantify genetic diversity (Wan *et al* . 2004) and identify conservation units (Moritz, 1994). However, important differences in population and genome-wide patterns of variability can emerge when whole genome data are used compared to genotypes from a few microsatellite loci (Fischer *et al.* 2017). Specifically, genomics offers a more reliable

approach to achieve a high resolution image detailing patterns of diversity and population differentiation within and among endangered species (Lehmann *et al.* 2019; Shingate *et al.* 2020).

The pupfishes are a group of 10 genera of temperature and salinity-tolerant fishes (*Cyprinodontidae*) with a broad Nearctic and Neotropical distribution (Smith and Miller, 1986; Wildekamp, 1995). Many Nearctic pupfish species are restricted to small isolated desert springs (Miller, 1981), which makes them highly susceptible to groundwater withdrawal (Deacon, 1979), habitat loss (Black *et al.*2016) and interspecific hybridization with introduced congeners such as those used as bait by recreational fishers (e.g, *C .variegatus* ; Echelle and Echelle, 1997). Among the pupfish complex, there are currently 12 pupfish species listed as endangered or critically endangered and 14 others identified as threatened or vulnerable (IUCN 2020). Because of this, many pupfish species are commonly maintained in captive breeding facilities or artificial refugia to provide stock for reestablishing or augmenting declining wild populations, or as a hedge against the imminent extinction of natural populations (Koike *et al.* 2008; Martin *et al.* 2016; Black *et al.* 2017).

The present paper focuses on the White Sands pupfish (*C. tularosa*), which is endemic to the Tularosa Basin of Southern New Mexico. The species is listed as Endangered by IUCN, as threatened by the New Mexico Department of Game and Fish, and is under review as a possible endangered species by the U.S. Fish and Wildlife Service. Ancestral native populations of *C. tularosa* were first documented during the early 1900s at two ecologically distinct locations: Malpais Spring and Salt Creek. Using allozyme, mitochondrial, and microsatellite markers, Stockwell *et al.* (1998) found evidence for significant genetic differentiation between the two native populations and demarcated two Evolutionary Significant Units (ESUs). Furthermore, Collyer *et al.* (2005) demonstrated substantial adaptive body shape variation between individuals collected from the Malpais Spring and Salt Creek ESUs, putatively driven by salinity differences between these two habitats. Since recognition of the genetic and morphologic differences between these two native *C. tularosa* populations, nine additional microsatellite markers have been developed to aid in the classification and management of these ESUs (Iyengar *et al.*2004). Additionally, two refuge populations (Mound Spring and Lost River) were founded by the Salt Creek ESU and established to safeguard against any future and unforeseen population extirpations (Stockwell*et al.* 1998).

Herein, we present and evaluate a new draft genome assembly for a White Sands pupfish collected from the Upper Lost River population (a part of the Salt Creek ESU) and compare this assembly to genomes of other related fishes. We then leverage the new *C. tularosa* assembly to determine the genomic position of microsatellites previously assumed to be "neutral" markers in ESU classification. Next, we evaluate the*C. tularosa* genome for evidence of introgression via historical interspecific hybridization by estimating pairwise genetic distances at nuclear (nuDNA) and mitochondrial (mtDNA) sequences. Finally, we evaluate genome-wide diversity (Heterozygosity, *H*) for *C. tularosa* and compare this diversity estimate to several other fish genomes represented among the order *Cyprinodontiformes* . The overarching goal of this work is to lay the groundwork and genomic infrastructure to enable future high-resolution assessments of the conservation status and evolutionary potential of this imperiled species.

Methods and Methods

Sample Collection, Library Construction and Sequencing

In September 2018, a deceased heterogametic male (*wsp-4*) was collected from the upper reach of Lost River (32deg 54' 2.88" N, 106deg 6' 54" W), stored in 95% ethanol and deposited in a -80@C freezer until processing. For Illumina library preparations, a Qiagen DNeasy Blood and Tissue kit was used to extract DNA from fin tissue. For the PacBio library preparations, high molecular weight DNA was extracted from skeletal muscle tissue by Polar Genomics (Ithaca, NY) to meet the integrity requirements for PacBio long read sequencing.

The TruSeq PCR-free protocol was used to prepare the Illumina library in the core genomics center at Purdue University (West Lafayette, IN) using an Illumina NovaSeq with 2x151bp chemistry. The SMRTBell Express Template Prep kit 2.0 was used on genomic DNA sheared to a mean fragment size of 30Kb and sequenced using two 8M single-molecule real-time cells (SMRTcells) of a PacBio Sequel II at the University

of Illinois (Roy J. Carver Biotechnology Center).

Quality Control and Genome Assemblies

Quality assessment of Illumina short-read sequences was conducted by visual examination of fastqc (v.0.11.7; Andrews, 2017) quality score distributions and the presence / absence of sequencing adapters. Following quality assessment, reads were filtered with trim galore (v.0.6.5; Krueger, 2015) by clipping identifiable Illumina adapters, removing low quality bases (Phred <20), and discarding any processed reads that were less than 30-nucleotides in length.

PacBio subread binary alignment mapping files from the two SMRTcells were converted over to *fastq* file format using the SMRTLink v8.0 command *bam2fastq* and concatenated. Reads <1Kb or >50Kb in length were removed using the program seqkit (v.0.12; Shen *et al.* 2016) prior to assembly. To generate a *de novo* assembly of both nuDNA and mtDNA genomes, the program wtdbg2 was used to create consensus contigs from a fuzzy *de Bruijn* graph (v.2.2; Ruan and Li, 2020). PacBio reads were sub-sampled at two different coverage levels (50x and 90x) to determine the optimal depth for assembly, as excessive reads can lead to subpar assemblies due to the accumulation of base-call errors. After each depth iteration, quast (v.3.2; Gurevich *et al.*2013) was used to generate summary statistics (e.g., N50) to identify the optimal depth.

PacBio subreads from both SMRTcells were then mapped back to the optimal assembly with minimap2 (v.2.11; Li, 2018), followed by consensus calling with arrow (v.2.3; Chin *et al* . 2013). Following long-read polishing, the quality controlled Illumina short-reads were used for additional error correction using the polca script compiled with the assembler MaSuRCA (v.3.41; Zimin *et al.* 2017). Changes to consensus assembly quality began to plateau after several iterations of short-read polishing, so three total error-correction iterations were conducted (Figure S1). High coverage heterozygous haplotypes that were assembled as separate contigs were identified and removed using the program purge_haplotigs (v.1.0; Roach *et al.* 2018). Primary haplotype sequences were then evaluated for contaminants using blobtoolkit2 (v.2.1; Challis *et al.* 2020) and those that were identified to have been derived from non-target species were removed. The curated genome assembly was then scanned for Benchmarking Universal Single-Copy Orthologs (busco; v.4.0.6; Seppey *et al.* 2019) to quantitatively assesses genome completeness by evaluating structural integrity (e.g., full:partial genes) and comprehensiveness (proportion of target genes in the assembly). Completeness was evaluated by performing local alignments between the assembly and the *Actinopterigii_ obd10* database, which contained a set of 3,640 core genes.

The complete mtDNA sequence of *C. tularosa* was reconstructed using the cleaned Illumina paired-end reads. To help prevent the inadvertent incorporation of Nuclear Translocations (NUMTs) into the mtDNA assembly, we first used coalqc (v.0.1; Patil *et al.* 2020) and samtools (v.1.17; Li *et al.* 2009) to extract Illumina reads that aligned to the Devils Hole pupfish (*C. diabolis*) mtDNA assembly (NC_030345.1; Lema *et al.* 2016). The resulting *bam* file was then converted back to *fastq* file format with bedtools (v.2.29.0; Quinlan and Hall, 2010) and used alongside the *C. diabolis* full mitogenome as a backbone for the *C. tularosa* mtDNA genome assembly with mitobim (v.1.8; Hahn *et al.* 2013).

Nuclear and Mitochondrial Genome Annotation

After draft genome curation, transposable element families were identified in the *C. tularosa* nuDNA assembly with repeatmodeler (v.1.09; Smit and Hubley, 2008) and regions matching known proteins were removed from the library (±50 nucleotides) with ProtExcluder (v.1.2; Campbell *et al.*2014). Using the custom *C. tularosa* transposable element library, along with the Zebrafish (*Danio rerio*) RepBase library, gene annotation were then accomplished by performing: 1) local alignments of translated protein and transcriptome sequences to the assembly and 2)*ab initio* gene predictions with the maker pipeline (v.2.31; Cantarel *et al.* 2008). To help annotate the nuDNA assembly, protein and transcriptome sequences were used from related species: Sheepshead Minnow (*C. variegatus*), Zebrafish (*D. rerio*) and the Southern Platyfish (*Xiphophorus maculatus*). Among all three species, 70,576 protein sequences were downloaded from*swiss-prot* (UniProt Consortium, 2019) and 54,466 cDNA sequences were obtained from *Ensembl* (Cunningham *et al.* 2019) prior to mapping these features to the assembly with blast (v.2.10; Gertz *et al.* 2006) and screening alignments with

exonerate (v.2.2.0; Slater and Birney, 2005). *Ab initio*gene predictions were made using Augustus (v.3.3.2; Hoff and Stanke, 2019) with the *D. rerio* and *X. maculatus* training sets as well as two rounds of snap (v.0.15; Korf, 2004). The gene models were then merged and redundancy removed, prior to functional annotation with the *UniProtKB* database and blastp. The assembled mtDNA sequence was annotated with geseq (v.1.84; Tillich *et al.* 2017), visualized with ogdraw (Lohse*et al.* 2013) and protein sequence similarity was compared to the RefSeq *C. tularosa* assembly (NC_028292.1) with blatn (v.35; Kent, 2002). For a visual representation of the methods used to assemble the *C. tularosa* nuDNA and mtDNA genomes, see Figure S2.

Assembly metrics

For perspective and to provide a comparative framework, we then sought to compare our new *C. tularosa* genome assembly to other published reference-enabled *Cyprinodontiformes* . Two scaffold-level pupfish reference genomes were available at the time of our analyses, *C. variegatus* (GCF_000732505.1) and the Amargosa pupfish (*C. nevadensis;* GCA_000776015.1). Three other*Cyprinodontiformes* were also included, Guppy (*Poecilia reticulata* ; GCF_000633615.1), *D. rerio* (GCF_002775205.1) and the Annual Killifish (*Austrofundulus limnaeus* ; GCF_001266775.1). For all six genomes (including the new *C. tularosa* reference), assembly statistics and annotation completeness were assessed by busco and quast using a minimum sequence length of 5Kb. The raw sequence data used for each assembly (i.e., Illumina*fastq* files for each Biosample) were obtained and cleaned with trim galore, using the same methods for *C. tularosa* (see above). Draft genome assemblies and raw shotgun sequence data for all species were downloaded from NCBI using the RefSeq (or GenBank) FTP site and Sequence Read Archive (SRA) files were obtained using the*sra-toolkit* on 15 August 2020.

Microsatellite Identification

Earlier work to define *C. tularosa* ESUs was based on putatively neutral microsatellite data, along with allozyme and mtDNA D-loop control regions (Stockwell *et al.* 1998). To help gauge whether any of these 'neutral' microsatellite markers might be subject to genetic hitchhiking associated with selection on a functional gene (Maynard Smith and Haigh, 1974), we leveraged our genome assembly to determine the distances from each microsatellite locus to the nearest suspected functional gene(s). To this end, the full microsatellite sequences were obtained from Iyengar *et al.* (2004) and aligned to the *C. tularosa* genome with bwa (v.0.717; Li and Durbin, 2009). primersearch (v.6.31; Rice *et al.* 2000) was used to determine the location of each Stockwell *et al.* (1998) primer set in the *C. tularosa* assembly, requiring a zero mismatch rate for both forward and reverse primers. All mapped microsatellite sequences and amplimer lengths were then manually reviewed in Integrative Genome Viewer (IGV) to determine the proximity to neighboring genic regions. For each microsatellite marker, we report the distance (in Kb) to the nearest predicted protein coding gene (or if a microsatellite occurred within a gene, the feature that it was found in) and prior estimates of genetic diversity at each microsatellite marker. This includes estimates of observed (H_{obs}) and expected heterozygosity (H_{exp}) as well as the number of alleles (A) at each locus (Stockwell *et al.* 1998; Iyengar *et al.* 2004).

For each ESU, the relationship between these microsatellite loci (N=11), their observed homozygosity and the distance to the nearest predicted gene were then fit using a linear model (H_{obs}~ *Distance+ESU*) using the*lm* function in base R. Because of the non-normal distribution of distances between a given microsatellite and the nearest annotated gene (see Results), we conducted two regressions, one using all of our data associated with the microsatellite loci and another using only data associated with microsatellites found $<$ 50Kb from a gene. In both cases, significance was evaluated by permuting the model 1,000 times and plotting the results with *ggplot2 (* v.3.32; Wickham, 2016). To evaluate neutrality at the two microsatellite markers used in ESU demarcation (WSP-2 & WSP-11), genotypes were obtained from Stockwell *et al.* (1998). The 20 samples (N=10 / ESU) genotyped at these two loci were tested for departure from Hardy-Weinberg Expectations (HWE) as well as population differentiation. Default parameters were used for the HWE test (Guo and Thompson, 1992) and the log-likelihood exact G-test (Goudet *et al* . 1996), as implemented in genepop (v.4.6; Raymond, 1995).

Genomic diversity

Genome-wide heterozygosity (H) was estimated for six *Cyprinodontiformes* . First, to determine mappability at each site in each reference genome, genmap (v.1.3.0; Pockrandt *et al.* 2020) was run using *100-mers* and allowing up to two mismatches. repeatmasker (v.4.07; Smit *et al.* 2015) was then used to identify repeats using the *D. rerio* RepBase library as a reference. Sites with mappability <1 and repeated regions were excluded from downstream analyses (Table S1). Scaffolds >100Kb in length were removed and H was calculated using angsd (v.093; Korneliussen *et al.* 2014) using the unfolded site frequency spectrum (SFS), requiring a minimum mapping quality of 30 and base quality >20. The results were visualized with *ggplot2.*

Nuclear and mitochondrial genomic divergence

For related pupfish species with an available reference genome (*C. tularosa, C. variegatus* and *C. nevadensis),* and for context *X. maculatus* (GCF_002775205.1), homologous nuDNA regions were identified by first breaking each genome into 30Kb sections using the *splitter* function implemented in emboss (v.6.60; Rice *et al.* 2000). Pairwise local alignments were then performed against each congeneric repeat masked genome using blastn with the following non-default parameters (-perc_identity 3, -qcov_hsp_perc 3, -max_target_seqs 1, -evalue 50, -max_hsp 1, -culling_limit 1). Following local pairwise alignments, genetic distances were estimated between each aligned sequence (>2Kb in length) using the Kimura-2 parameter (K2P; Kimura, 1980) generated with the R package *ape* (v.5.4.1; Paradis *et al.* 2004). Results were visualized with *ggplot2* to evaluate patterns of pairwise genomic differentiation.

To compare the matrilineal relationship among pupfish species, all *Cyprinodontidae* mtDNA genomes available as of 15 August 2020 were downloaded from NCBI for the following species: *C. variegatus* (KT288182.1), *C. diabolis* (KX061747.1), *C. nevadensis amargosa* (KU883631.1), the Desert pupfish (*C. macularius* ; KM985373.1) and the Red River pupfish (C. *rubrofluviatilis* ; NC_009125). In addition to the new *C. tularosa* mtDNA assembly reported herein, the previously archived *C. tularosa* RefSeq mtDNA assembly was also included (NC_028292.1). The *X. maculatus* mtDNA assembly (NC_011379.1) was used as an outgroup. Therefore, eight full mitogenome sequences were used for phylogenomic analysis; two for *C. tularosa* , five from other *Cyprinodontidae* and an outgroup. Multiple sequence alignment of full length mitogenome sequences was then conducted using clustalw (v.2.1;Thompson *et al.* 1994) and a Neighbor Joining (NJ) distance matrix was generated using a Jukes-Cantor substitution model using 100 bootstrap replicates within megax (Kumar *et al.* 2018). Molecular time estimates were obtained from Timetree (Kumar *et al.*2017) and fixed rates were used as time constraints at each available node using the *reltime-ML* function (Tamura *et al.* 2018). Estimated mtDNA divergence times were then annotated to the NJ tree, exported in Nexus format and rendered using phlo.io (Robinson *et al.* 2016). In addition to creating a timetree, mean pairwise genetic distances (K2P) between species mtDNA sequences was estimated with megax. Our intention was not to perform a systematic review of *Cyprinodontidae* phylogeny, but to provide evolutionary context for our pairwise comparisons.

Results

Quality Control and Genome Assemblies

Two PacBio Sequel II SMRT cells yielded 165Gb worth of sequence data among 16 million reads with a mean N50 polymerase read length of 22Kb. After extracting reads between 5-50Kb, 100Gb remained among 14.3 million reads for use with genome assembly (~100x coverage). The llumina sequencing platform yielded 43Gb of raw sequence data among 286 million 151-bp paired-end reads. After quality filtering, there were 284.9 million quality controlled reads remaining (99.6% of total) with an estimated genome coverage of ~40x.

Using the optimal 90x coverage level, the assembler wtdbg2 yielded 1.09Gb among 2,606 contigs, with a N50 of 1,357Mb and a GC content of 39%. The assembly was then polished using PacBio subreads from the two SMRTcells and error corrected using three iterations of short-read polishing. The program purge_haplotigs discarded 332 contigs from the assembly as haplotigs (total length 3.5Mb) and 12 as artifactual (total length 596Kb), resulting in 2,009 contigs. blobtools2 was then used to identify non-target DNA sequenced in the

pupfish libraries. Overall, five contigs (68Kb) were not assigned as *Actinopterigii* sequences (classified as 'no-hits',*Arthropoda* , or *Chytribiomycota*) and were subsequently filtered out of the assembly. The curated *fasta* file containing 2,004 contigs was then used for downstream analyses.

The complete mtDNA sequence of *C. tularosa* was reconstructed using congeneric (C. *diabolis*) aligned Illumina short-reads. The assembly of the *C. tularosa* mitogenome produced a 16,508 nt contiguous sequence similar in length to other published*Cyprinodon* mtDNA assemblies (16,499-16501 nt). The *C. tularosa* mtDNA assembly had a 100% identity match with contig1125 within the primary assembly and was subsequently removed from the nuDNA sequence file prior to annotation.

Nuclear and Mitochondrial Genome Annotation

maker identified the genomic coordinates of 25,260 protein coding genes in the curated *C. tularosa* assembly, which is generally congruent with the 23,019 coding genes annotated in the current genome release of *C. variegatus* (GCF_000732505.1). Functional annotation of the new *C. tularosa* mtDNA assembly illustrated that these annotations were also consistent with other published *Cyprinodon* mitogenomes, with the correct ordering of 13 polypeptides, 22 tRNA genes, 2 rRNA genes, and one control region. A sequence similarity search between the two *C. tularosa*mitogenomes revealed *pls scores* of 97-100% among protein coding genes.

Assembly metrics

For the White Sands pupfish draft genome, assembly statistics on 1,995 contigs (those greater than 5Kb in length) showed a total length of 1.086Mb, a max contig size of 8.12Mb and a N50 of 1.36Mb (Table 1). Thus, the new *C. tularosa* assembly is more contiguous than the scaffold-level *C. variegatus* assembly (contigs=4,093; max contig size=4.51Mb; N50=0.0843Mb) but less contiguous than the chromosome-level*P. reticulata* assembly (contigs=843; max contig size=46.3Mb; N50=31.4Mb) or the *X. maculatus* assembly (contigs=97; max contig size=35.3Mb; N50=31.5Mb; Table 1). Using the *Actinopterigii*database of 3,640 orthologs, busco identified 3,452 complete single-copy genes (C 95%) in the *C. tularosa* assembly with little inherent fragmentation (F 1%). Genome completeness, based upon these values, was moderately lower than the *X. maculatus* assembly (C 96.6%; F 0.3%), equivalent to the chromosome-level *P. reticulata* assembly (C 95%; F 1%), similar to *C. variegatus* (C 94%; F 1%) and *A. limnaeus* (C 92%; F 2%), but markedly improved over the *C. nevadansis* assembly (C 70%; F 13%; Figure 1).

Microsatellite Identification

Each of the 16 microsatellites previously used to characterize *C. tularosa* populations were successfully located in the *C. tularosa* genome assembly. Three were found within an intron of a gene, five were located within 10Kb of a gene(s), five were located between 10-100Kb of one or more genes and three were mapped to a location >100Kb from any gene (Table 2). Using 11/16 loci which had previously reported heterozygosity estimates, our comprehensive regression suggested no significant relationship between microsatellite heterozygosity and distance to the nearest gene or ESU (Figure 2A; Intercept: Estimate=0.265, SE=0.082, p< 0.001;*Distance* : Estimate=0.000, SE=0.000, p-value=0.780; *ESU* : Estimate=0.020, SE=0.110, p-value=0.850; R^2=0.005; F=0.050; DF=2,19). Upon limiting our regression to microsatellite loci that were < 50Kb, we found that genic distance was negatively related to H_{OBS} (Figure 2B; *Intercept* : Estimate=0.414, SE=0.092, p-value< 0.001;*Distance* : Estimate=-0.008, SE=0.003, p-value=0.022; *ESU* : Estimate=0.003, SE=0.103, p-value=0.953; R^2=0.302; F=3.3249; DF=2,15). Both microsatellite markers used to inform ESU demarcation (Stockwell *et al.* 1998) conformed to Hardy-Weinberg Expectations within their respective ESU (p-values=0.195-1.00). An exact G-test at these loci illustrated significant (p-values=0) levels of differentiation (WSP-2, F_{ST} =0.350; WSP-11, F_{ST} =0.624) and were located nearby (<2Kb) predicted genes in the *C. tularosa* genome (Table 2).

Genomic diversity

Genome-wide heterozygosity (H) was then estimated for all six reference enabled *Cyprinodontiformes* (Table S1). Diversity values among all six species illustrated that H estimates ranged from 0.01225 (*C. variegatus*) to 0.00003049 (*X. maculatus*), which we note was derived from a highly inbred line. Estimated $H$for *C.*

tularosa (0.00053) was found at the low end of these estimates, near *C. nevadensis* (0.00116) and the inbred *X. maculatus* line. According to the IUCN, both vulnerable (*C. nevadensis*) and endangered (*C. tularosa*) pupfish species were found at the low end of our diversity estimates (Figure 3).

Nuclear and mitochondrial genomic divergence

Pairwise local alignments of [?]30Kb sequences were performed between reference enabled species to: 1) compare levels of nuDNA genome sequence divergence between related species; 2) compare levels of nucDNA and mtDNA divergence; and 3) assay the *C. tularosa* genomic landscape for signatures of a) potential introgression with *C. variegatus*and / or b) localized windows within the genome that may be involved with selective sweeps. We found that nuDNA sequence divergence generally mirrored the mtDNA genome tree (Figs. 4 and 5). Genetic distances were approximately 2x greater for mtDNA sequences compared to nuDNA data, with mean (+-SD) pairwise nuDNA K2P values of 0.0309+-0.010 (*C. tularosa vs C. nevadensis*) and 0.0304+-0.009 (*C. variegatus vs C. nevadensis*). Across 20,000 alignments averaging 5,528 nt in length, mean nuDNA K2P values were lower between *C. tularosa* and *C. variegatus* (0.0269) with highly variable distances (K2P range =0.0004-0.265; Figure 4A).

Following multiple sequence alignment of full mitogenomes, mean Kimura 2 values (K2P; Kimura, 1980) and divergence times (MYA) were estimated for all species pairs. As expected, the greatest genetic distances were observed between the outgroup *X. maculatus* and other pupfish species (K2P range=0.2655-0.2712). Pairwise mean K2P values using the new *C. tularosa* mtDNA assembly ranged from 0.0009 (RefSeq,*C. tularosa*) to 0.0691 (*C. diabolis* ; Figure 4B). Mean genetic distance between the White Sands pupfish and the Sheepshead minnow (*C. variegatus* vs *C. tularosa;* K2P=0.0222; timetree estimates an estimated divergence time of 1.6-4.7 MYA) were ˜3x lower than other congener comparisons (K2P range=0.053-0.069; Figure 4B). The clades recovered in our phylogenomic assessment (Figure 5) were largely congruent with previous assessments of *Cyprinodontidae* evolutionary history (e.g., Echelle *et al.* 2005).

Discussion

The conservation of endangered species is increasingly informed by genomic data. To that end, we have generated a high quality draft genome assembly for the endangered White Sands pupfish (*C. Tularosa*). Our 1.08Gb assembly is largely complete (˜25k protein coding genes including 95% of conserved *Actinopterygii* genes) with very little fragmentation (1%). Furthermore, 95% of the quality controlled Illumina paired-end reads aligned to our *C. tularosa*assembly (data not shown), demonstrating the utility that this new draft genome will have in future conspecific, as well as many heterospecific, re-sequencing and/or transcriptomic studies. Finally, our ˜16Kb *C. tularosa* assembled mitogenome was consistent with other pupfish mtDNA annotations (13 polypeptides, 22 tRNA genes, 2 rRNA genes and one control region). Below, we briefly discuss the results and implications of our work.

Intraspecific evaluation

Heilveil and Stockwell (2017) demonstrated that the Lost River refuge population showed reduced genetic diversity and the current work validates and extends their insight to the genomic level. Heterozygosity estimates revealed low levels of genomic diversity in the single *C. tularosa* specimen (*wsp-4*) collected from the upper reach of Lost River and used for sequencing and assembly. Compared to the other *Cyprinodontoformes* examined, *C. tularosa* showed a genome-wide heterozygosity that was one (or two) orders of magnitude less than other species diversity estimates (Figure 3). If this *C. tularosa* sample is representative of the heterozygosity of the refuge population, this may be of conservation concern as low genetic diversity can constrain adaptive potential in light of future environmental change (e.g., Finger *et al.* 2013). Lost River was founded by 30 individuals translocated from Salt Creek *circa* 1970, with approximately ˜bi-annual translocations of 40 fish (from Salt Creek) beginning in 2008 (Carman, 2010). It is currently unknown what, if any, affect these translocations have on the genetic diversity of the Lost River refuge population but the low level of heterozygosity of our study suggests that there may have been a substantial reduction in genomic diversity due to the founder effect and/or drift.

From a genetic management perspective, refuge populations should strive to replicate their source populations and maintain original levels of genetic diversity. If the other *C. tularosa* populations are as homogenous at the whole genome level, facilitated local gene flow (i.e., genetic rescue) may be one approach to help increase heterozygosity (Petit and Excoffier, 2009). However, to do so effectively, the source population needs considerably more genetic variation than the recipient population (Rails *et al.* 2020). More extensive population genomic surveys of the Lost River and Salt Creek populations are required for certainty, but this refuge population may be similar to an insurance policy with insufficient coverage for future losses. Furthermore, it is unclear if the potential cost of homogenizing locally adapted / differentiated gene-pools, through the act of small reciprocal inoculations between ESUs (Arnold, 2016), would outweigh the need to preserve and protect the two separate *C. tularosa* ESUs. These are conservation issues that will require the careful planning and examination of individual- and population-level diversity metrics in future whole genome re-sequencing studies.

Microsatellite Identification

Our intention was to identify the genomic coordinates of microsatellites previously used to define and monitor ESU in the White Sands pupfish. We successfully identified the location of all 16 microsatellite markers within the *C. tularosa* draft genome. Many of these loci were either within 10Kb of predicted genes (5/16) or inside an intron of a gene (3/16), suggesting they may not have evolved in a strictly neutral fashion. Indeed, we found a negative correlation between heterozygosity and genic proximity, suggesting that selective processes differ between genic and intergenic regions within the *C. tularosa* genome. However, for the two loci where genotype data were publicly available (WSP-02 & WSP-11; Stockwell *et al.* 1998), we found no deviation from Hardy-Weinburg Expectations. This is only a weak test of natural selection and our assessments of genomic provenance are not formal tests of microsatellite neutrality, but the data provided herein add considerable context to earlier microsatellite studies that helped define the two ESUs. High resolution neutrality tests may provide future genomic insights into local adaptation associated with the two *C. tularosa* ESUs.

Nuclear and mitochondrial genomic divergence

Our mtDNA based phylogeny provided 100% bootstrap support at each node (Figure 5) and was largely in agreement with the evolutionary history of*Cyprinodontidae* (e.g., Echelle *et al.* 2005; Sağlam*et al.* 2016). In addition to the phylogenomic tree, we also used our mtDNA assembly to quantify genetic distance with a previously assembled *C. tularosa* mitogenome that was sourced from the Salt Creek ESU on December 1983 (http://fishnet2.net/; MSB:Fish:96664). The genetic distance estimate between these two *C. tularosa*mitogenomes was essentially zero (K2P=0.0009; Figure 4B), which is what one would expect to observe between individuals collected from the same ESU and sampled less than 50 years apart. As translocation of 40 individuals have occurred on a semi-annual basis since 2008 (Carman, 2010), and there was little difference between mitochondrial genomes, we expect little genomic divergence between these two populations (Salt Creek and Upper Lost River). However, Heilveil and Stockwell (2017) demonstrated moderate population structure between the Upper and Lower Lost River sub-population ($F_{ST} = 0.024$) which suggests that future research is required to evaluate levels of genomic differentiation between these two waterways within the Lost River refuge population.

Using both nuDNA and mtDNA sequences, pairwise comparisons between related species demonstrated that genomic distance (Figure 4B) increased with taxonomic divergence (Figure 5) with a range in pupfish divergence times of 0-10 MYA. For the most part, differentiation values reported herein were 4-5x higher in mtDNA divergence (relative to nuDNA) with genetic distances roughly equivalent to those reported in Willoughby*et al.* (2020) at both nuDNA (*C. variegatus vs C. nevadensis* =0.04) and mtDNA (*C. variegatus* vs *C. nevadensis* =0.06) regions. We note that K2P variability was substantially lower in nuDNA alignments (K2P=0.01-0.05) relative to mtDNA (K2P=0.006-0.137) alignments (Willoughby *et al.* 2020). This disparity in variance is likely a consequence of sequence mappability; the entire mtDNA assembly is well-known and annotated so sequence mappability is relatively high and the K2P variance likely represents the biological variation inherent to hypervariabile regions (e.g., the D-loop) compared to constrained regions (e.g., tRNA sequences and ND4). In contrast, nuDNA genomes are still incomplete even when considering high-quality

chromosome-level assemblies. Furthermore, our strategy was designed to reduce spurious matches, but this means we almost certainly homogenized mappable regions and necessarily underestimated genetic distances in the nuDNA data. Thus, absolute K2P estimates of genomic divergence are provisional for nuDNA data relative to mtDNA sequences.

Both mtDNA and nuDNA divergence rates showed a positive correlation (R^2=0.99) among all six pairwise comparisons (Figure 4C). For 5/6 pairwise comparisons, results demonstrated a 1.9-2.2x increase in pairwise mtDNA sequence divergence compared to nuDNA divergence, in agreement with the well-recognized higher levels of mitochondrial mutation rates first reported by Brown *et al.*(1979). However, we found similar estimates of genetic distances between the *C. tularosa* and *C. variegatus* mtDNA (K2P distance = 0.022) and nuDNA (K2P distance = 0.0269) sequences, suggesting that our nuDNA estimates may not have fully captured the true patterns of sequence divergence (Figure 4A). Alternatively, this raises the possibility of contemporary introgression from *C. variegatus* into a *C. tularosa* population(s), which could have partially homogenized the *C. tularosa* gene pool and led to this reduction in nuDNA (and perhaps mtDNA) divergence. Furthermore, evaluating the distribution of all pairwise sequence alignments between *C. tularosa* and *C. variegatus* illustrated a highly variant mosaic of homologous blocks between *C. tularosa* and *C. variegatus*(K2P range =0.0004-0.264). Here, we suggest that contigs with near zero K2P values may represent *C. variegatus* introgressed regions (Figure 4A). However, extensive sampling and re-sequence data will be better suited to confirm (or reject) these potential tracts of introgression within the *C. tularosa* genome. Nevertheless, identification of the genomic patterns of introgressions, and how this influenced evolution of *C. tularosa* , is important because it may alter how we manage this endangered desert fish. For example, an extreme but often utilized approach in managed desert fish populations is to cull any 'genetically compromised' fish and reintroduce or translocate 'representative' samples back into the original population (e.g., Echelle *et al.* 1997). Thus, we think the data and analyses presented herein provide a significant resource to the broader community of conservation biologists, especially those working with the pupfishes.

Conclusions

We present the first draft of an annotated genome for *C. tularosa*and find that it is among the best assemblies available for this group of fishes. To add context to our assembly and to more broadly illustrate its potential as a community resource, we performed a number of exploratory analyses related to the diversity and divergence of *C. tularosa* . We found that heterozygosity in our assembly is strikingly low relative to related species, including other pupfish, and that population genomic surveys are warranted to determine if the lack of heterozygosity extends to individuals from other populations and/or encompass large runs of homozygosity (e.g., due to recent inbreeding). Our retrospective analyses of microsatellite heterozygosity indicates that population genetic diversity at this suite of "neutral" markers is not distributed randomly, but is partitioned relative to genomic distance from annotated functional genes. We also used our assembly to evaluate genome-wide divergence from related species to help identify outlying contigs that could represent regions homogenized due to introgression (a major conservation concern) or divergent regions that could represent targets of selection (e.g., adaptation to different salinity or parasite regimes). Overall, we think these new resources and analyses will benefit future ecological and evolutionary studies of the *Cyprinodontidae* and ultimately, hope they contribute to pupfish conservation.

Acknowledgements

This research was funded by the U.S. Corps of Engineers (W9126G-12-2-0019). JAD was supported in part by the National Institute for Food and Agriculture. We thank members of the DeWoody lab group for constructive feedback throughout this project.

References

Andrews, S. (2017). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arnold, M.L. (2016). Divergence With Genetic Exchange. Oxford University Press

Black, A. N., Seears, H. A., Hollenbeck, C. M., Samollow, P. B. (2017). Rapid genetic and morphologic divergence between captive and wild populations of the endangered Leon Springs pupfish, *Cyprinodon bovinus* . *Molecular Ecology* **26** :2237-2256.

Black, A. N., Snekser, J. L., Al-Shaer, L., Paciorek, T., Bloch, A., Little, K., Itzkowitz, M. (2016). A review of the Leon springs pupfish (*Cyprinodon bovinus*) long-term conservation strategy and response to habitat restoration. *Aquatic Conservation: Marine and Freshwater Ecosystems* **26** :410-416

Brown, W.M., George, M., Wilson, A.C., (1979). Rapid evolution of animal mitochondrial DNA. *Proceedings of the Royal Society B***76** : 967–1971

Campbell, M.S., Law, M., Holt, C., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology* **164** :513-524

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B.,*et al.,* Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* , **18** :188-196.

Carman, S. (2010) White Sands pupfish status report, 2009. New Mexico Game and Fish, Santa Fe, NM.

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit–Interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics* **10** :1361-1374

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., *et al* ., Turner, S. W. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10** :563-569

Collyer, M.L., Novak, J.M., Stockwell, C.A. (2005). Morphological divergence of native and recently established populations of White Sands pupfish (*Cyprinodon tularosa*). *Copeia* **2005** :1-11

Cunningham, F., Achuthan, P., Akanni, W. *et al.* (2019). Ensembl 2019. *Nucleic Acids Research* **47** :D745–51

Deacon, J. E. (1979). Endangered and threatened fishes of the West.*Great Basin Naturalist Memoirs* **3** :41-64

Echelle, A. A., Echelle, A. F. (1997). Genetic Introgression of Endemic Taxa by Non-natives: A Case Study with Leon Springs Pupfish and Sheepshead Minnow. *Conservation Biology* **11** :153-161

Echelle, A.A., Carson, E.W., Echelle, A.F., Van Den Bussche, R.A., Dowling, T.E., Meyer, A. (2005). Historical biogeography of the new-world pupfish genus Cyprinodon (Teleostei: Cyprinodontidae). *Copeia* **2** :320-339

Finger, A. J., Parmenter, S., May, B. P. (2013). Conservation of the Owens pupfish: genetic effects of multiple translocations and extirpations. *Transactions of the American Fisheries Society***142** :1430-1443

Fischer, M.C., Rellstab, C., Leuzinger, M. *et al.* (2017). Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri* . *BMC Genomics* **18:** 69 https://doi.org/10.1186/s12864-016-3459-7

Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A., Altschul, S. F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology***4** :1-14

Gómez-Sanchez, D., Olalde, I., Sastre, N., Ensenat, C., Carrasco, R., Marques-Bonet, T., *et al.* (2018). On the path to extinction: inbreeding and admixture in a declining grey wolf population. *Molecular Ecology* **27** :3599-3612

Goudet, J., Raymond, M., deMeeus, T., Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics***144** :1933-1940

Guo, S. W.,Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*361-372

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29** :1072-1075

Hahn, C., Bachmann, L., Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. *Nucleic acids research* **41** : e129-e129

Heilveil, J.S., Stockwell, C.A. (2017). Genetic signatures of translocations and habitat fragmentation for two evolutionarily significant units of a protected fish species. *Environmental Biology of Fishes* **100:** 631-638 https://doi.org/10.1007/s10641-017-0591-4

Hoff, K. J., Stanke, M. (2019). Predicting genes in single genomes with augustus. *Current Protocols in Bioinformatics* **65** :e57

IUCN 2020. The IUCN Red List of Threatened Species. Version 2020-2. https://www.iucnredlist.org Downloaded on 09 July 2020

Iyengar, A., Stockwell, C. A., Layfield, D., Morin, P.A. (2004). Characterization of microsatellite markers in a threatened species, the White Sands pupfish (*Cyprinodon tularosa*). *Molecular Ecology Notes* **4** :191-193

Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res* earch **12** :656-664

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16** :111-120

Koike H., Echelle A.A., Loftis D., Van den Bussche, P.A. (2008) Microsatellite DNA analysis of success in conserving genetic diversity after 33 years of refuge management for the desert pupfish complex.*Animal Conservation* **11:** 321-329

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59

Korneliussen, T.S., Albrechtsen, A., Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15** :356

Krueger, F. (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files 516-517

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms.*Molecular Biology and Evolution* **35** :1547-1549

Kumar, S., Stecher, G., Suleski, M., Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution* doi:10.1093/molbev/msx116

Lema, S.C., Kevin, P., Wilson, B., Senger, L., Lee H. Simons (2016). Sequencing and characterization of the complete mitochondrial genome of the endangered Devils Hole pupfish *Cyprinodon diabolis* (Cyprinodontiformes: Cyprinodontidae)*Mitochondrial DNA Part B* **1** :705-707

Lehmann, R., Lightfoot, D.J., Schunter, C., Michell, C.T., Ohyanagi, H., Mineta, K., Foret, S., Berumen, M.L., Miller, D.J., Aranda, M., Gojobori, T. (2019). Finding Nemo's Genes: A chromosome-scale reference assembly of the genome of the orange clownfish *Amphiprion percula* . *Molecular Ecology Resources* **19** :570-585

Leroy, G., Carroll, E. L., Bruford, M. W., DeWoody, J. A., Strand, A., Waits, L.,Wang, J. (2018). Next-generation metrics for monitoring genetic erosion within populations of conservation concern. *Evolutionary Applications* **11** :1066-1083

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34** :3094-3100

Li, H., Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25** :1754-60

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N,. Marth, G,. Abecasis ,G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools, *Bioinformatics* **25** :2078-9

Lohse, M., Drechsel, O., Kahlau, S., Bock, R. (2013). OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets.*Nucleic Acids Research* **41**doi: 10.1093/nar/gkt289

Martin, C. H., Crawford, J. E., Turner, B. J., & Simons, L. H. (2016). Diabolical survival in Death Valley: recent pupfish colonization, gene flow and genetic assimilation in the smallest species range on earth. *Proceedings of the Royal Society B: Biological Sciences* 283 : doi:https://doi.org/10.1098/rspb.2015.2334

Miller, M. L., Miller, R. R. (1986). The evolution of the Rio Grande Basin as inferred from its fish fauna, p. 457–485. In: The Zoogeography of North American Freshwater Fishes. C. H. Hocutt and E. O. Wiley (eds.). John Wiley and Sons, New York

Miller, R. R. (1981). Coevolution of deserts and pupfishes (genus Cyprinodon) in the American Southwest (pp. 94-101). In Fishes in North American deserts, Naiman R.J., Soltz D. J. (Eds) Wiley and Sons, New York

Moritz, C. (1994). Defining 'evolutionarily significant units' for conservation. *Trends in Ecology & Evolution* **9** :373-375

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** :289–290

Patil, A.B., Shinde, S.S., Raghavendra, S., Satish, B.N., Kushalappa, C.G., Vijay, N. (2020). CoalQC - Quality control while inferring demographic histories from genomic data: Application to forest tree genomes. *bioRxiv*

Petit, R. J., Excoffier, L. (2009). Gene flow and species delimitation. *Trends in Ecology & Evolution* **24** :386-393

Pockrandt, C., Alzamel, M., Iliopoulos, C.S., Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability.*Bioinformatics* **36** :3687-3692

Quinlan, A.R., Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics***26** :841-2

Raymond, M. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86** :248-249

Rice, P., Longden, I., Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite *Trends in Genetics***16** :276-277

Roach, M.J., Schmidt, S.A. & Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19:** 460

Robinson, O., Dylus, D., Dessimoz, C. (2016). Phylo.io : Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web*Molecular Biology and Evolution* **33** :2163-2166

Ruan, J., Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature methods* **17** :155-158. https://doi.org/10.1038/s41592-019-0669-3

Sağlam, İ.K., Baumsteiger, J., Smith, M.J., Linares-Casenave, J., Nichols, A.L., O'Rourke, S.M., Miller, M.R. (2016). Phylogenetics support an ancient common origin of two scientific icons: Devils Hole and Devils Hole pupfish. *Molecular Ecology* **25** :3962-3973

Seppey, M., Manni, M., Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar M. (eds) Gene Prediction. *Methods in Molecular Biology* , vol 1962. Humana, New York, NY

Shen**,** W., Le, S., Li, Y., Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation.*PLoS ONE* doi:10.1371/journal.pone.0163962

Shingate, P., Ravi, V., Prasad, A., Tay, B.H., Venkatesh, B. (2020). Chromosome-level genome assembly of the coastal horseshoe crab (*Tachypleus gigas*). *Molecular Ecology Resources* . DOI: 10.1111/1755-0998.13233

Slater, G. S. C., Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics***6** :31

Smit, A., Hubley, R. (2008). RepeatModeler Open-1.0. Available at www.repeatmasker.org

Smith, J. M., Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research* **23** :23-35

Stockwell, C., Mulvey, M., Jones, A. (1998). Genetic evidence for two evolutionarily significant units of White Sands pupfish. *Animal Conservation* **1** :213-225

Tamura, K., Tao, Q., Kumar, S. (2018). Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. *Molecular Biology and Evolution***35** :1770-1782

Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice *Nucleic Acids Research* **11** :4673-80 https://doi.org/10.1093/nar/22.22.4673

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R. and Greiner, S. (2017). GeSeq -versatile and accurate annotation of organelle genomes. *Nucleic Acids Research***45** :W6-W11

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47** :D506-D515

Wan, Q. H., Wu, H., Fujihara, T., Fang, S. G. (2004). Which genetic marker for which conservation genetics issue? *Electrophoresis***25** :2165-2176

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* . Springer-Verlag New York

Wildekamp, R. H. (1995). A world of killies: Atlas of the oviparous Cyprinodontiform fishes of the world. Vol II. American Killifish Association, Mishawaka, Indiana

Willoughby, J.R., Harder, A.M., Sundaram, M., Mathur, S., Bylsma, R., DeWoody J.A. (2020). Predictors of nuclear and mitochondrial genome divergence in congeneric vertebrates. Submitted

Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Yorke, J.A., Dvorak, J., Salzberg, S.(2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the mega-reads algorithm. *Genome Research* **1** :066100

Competing Interests

All authors declare that they have no competing interests.

Data Accessibility

DNA sequences and genome assemblies: https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA670474

Scripts used to perform assembly and comparative genomic analyses:*https://doi.org/10.5061/dryad.p8cz8w9nt*

Author Contributions

A.N.B, J.R.W, B.L.P. and JAD contributed to the sample collection and design of the study. A.N.B and A.B.O analyzed the data. A.N.B and J.A.D. wrote the paper with input from all authors.

Tables:

Table 1. Quast assembly statistics for the six *Cyprinodontidae* assemblies, reported using contigs [?]5,000 nucleotides in length. The *C. tularosa* assembly is new, the other five assemblies were downloaded from NCBI.

	C. tularosa	*C. variegatus*	*C. nevadensis*	*A. limneaus*	*P. reticulata*	*X.*
Accession	*Pending*	GCF_000732505.1	GCA_000776015.1	GCF_001266775.1	GCF_000633615.1	GC
Level	Contig	Scaffold	Scaffold	Scaffold	Chromosome	Ch
# contigs	1,995	4,093	19,582	3,687	853	97
Contig N50	1.363Mb	0.843Mb	0.0087Mb	1.13Mb	31.4Mb	31.
Contig L50	236	360	3,062	175	11	11
Max contig size	8.12Mb	4.51Mb	0.750Mb	11.19Mb	46.3Mb	35.
Total Size	1,086Mb	1,026Mb	0.957Mb	848Mb	728Mb	704
***GC (%)**	39	38	38	39	39	39

Table 2 . Genic proximity to previously published microsatellite markers within the newly annotated *C. tularosa*genome. Each row corresponds to one microsatellite marker, the observed (H_{obs}), expected Heterozygosity (H_{exp}), and number of Alleles (A) at that locus within the Malpais Spring (N=10) or Salt Creek (N=10) population and the genic distance (Kb) downstream or upstream to the closest gene. Markers that fell within a genic area are classified according to the associated feature (i.e.,*intron*). Protein coding genes of unknown function are labeled as "UKP".

$^{+}$ Microsatellite markers used in *C. tularosa*ESU classification calculated from genotypes in Stockwell *et al.*1998

$^{++}H_{obs}$, H_{exp} and$A$ (only available for both populations) obtained from Iyengar*et al.* 2004

							Up
Microsatellite	**Malpais Spring** H_{obs}	**Malpais Spring** H_{exp}	**Malpais Spring** A	**Salt Creek** H_{obs}	**Salt Creek** H_{exp}	**Salt Creek** A	**Ge**
WSP-02 +	0.13	0.12	3	0.53	0.49	4	-
WSP-11 +	0.70	0.81	6	0.53	0.49	3	-
WSP-20	0.00	0.00	2	0.00	0.00	2[a]	
WSP-34^{++}	0.50	0.66	5	0.50	0.39	5	EN
WSP-33^{++}	0	0	2	0.60	0.51	2	TF
WSP-30^{++}	0.20	0.34	2	0	0	2	CI
WSP-31	-	-	-	-	-	-	AT
WSP-22	-	-	-	-	-	-	TF
WSP-24^{++}	0.50	0.39	3	0.40	0.51	3	TF
WSP-32^{++}	0.20	0.19	2	0	0	2	EI
WSP-23^{++}	0	0	4	0.50	0.51	4	NF
WSP-26^{++}	0.40	0.53	2	0	0	2	CF
WSP-29	-	-	-	-	-	-	RI
WSP-25^{++}	0.20	0.19	3	0	0.19	3	-
WSP-35	-	-	-	-	-	-	PC
WSP-21	-	-	-	-	-	-	FA
WSP-27	-	-	-	-	-	-	UF

Figures Legends:

Figure 1 . BUSCO results for analysis of genome completeness among six *Cyprinodontiformes* , using the*Actinopterigii* database containing 3,640 core orthologs. A star and bold font are used to designate the new *C. tularosa* assembly.

Figure 2 . Genetic diversity (H_{OBS}) at each microsatellite marker (N=11) and the distance (Kb) to the closest predicted *C. tularosa* gene. The shape of each microsatellite locus depicts the source ESU (Malpais Spring or Salt Creek). The relationship between H_{OBS}, ESU and the distance to closest gene were tested using a linear model (lm= H_{OBS}~Distance+ESU). A) For all available microsatellite markers, the regression analysis showed a non-significant (p-value=0.780) relationship (R^2=0.005) between heterozygosity and genic distance. B) Using only markers located [?]50Kb from neighboring genes revealed a significant (p-value=0.022) negative correlation (R^2=0.30) between $H_{OBS}$and genic distance. H_{obs} estimates were obtained from Iyengar *et al.* (2004) and Stockwell *et al.* (1998).

Figure 3 . Genomic estimates of heterozygosity (H) for the six *Cyprinodontiformes* examined. H was estimated for each species using the unfolded site frequency spectrum, as implemented in angsd (Korneliussen *et al.* 2014). A red star is used to designate the new *C. tularosa* assembly. With the exception of the highly inbred *X. maculatus* line, heterozygosity in *C. tularosa* is greatly reduced compared to related species.

Figure 4 . Nuclear and mitochondrial mean pairwise genetic distances among *Cyprinodontiformes* : a) Pairwise Kimura 2 distances (K2P) between aligned nuDNA reference sequences. Dashed red line signifies the mean nuDNA K2P value across >30kb windows for each comparison. Points colored red signify distances that were >0.99 or <0.01 percentiles; b) Upper left triangle matrix, mean genetic distances between aligned nuDNA sequences. Lower right triangle matrix, mean K2P genetic distances between aligned mtDNA sequences. Black boxes signify pairwise comparisons which had both nuDNA and mtDNA distances; c) scatterplot illustrating the correlation between mean K2P nuDNA and mtDNA divergence for the six pairwise comparisons. The dashed red line depicts a 1:1 ratio of nuDNA to mtDNA divergence, and the solid grey line is the K2P correlation among taxa. A star is used to designate the new *C. tularosa* assembly.

Figure 5. Neighbor joining tree constructed using the Jukes-Cantor model, following multiple sequence alignment of full mtDNA sequences. Significance of 100% bootstrap support are signified by black circles and divergence times (MYA) are labeled under the tree. Estimated divergence times were generated using the *reltime-ML*function (Tamura *et al.* 2018) from constrained time estimates (min/max) obtained from TimeTree (Kumar *et al.* 2017). A red star is used to designate the new *C. tularosa* assembly.