**Sequencing the delta smelt genome: improved annual monitoring of Ne and further understanding of the wild and hatchery delta smelt populations**

June 30, 2021 Final Report

Shannon E.K. Joslin, Ensieh Habibi, Mary Delaney, Justin Smith, Michael R. Miller and Amanda J. Finger

University of California, Davis

This document will serve as the final report for A19-1844 contract between State Water Contractors and the Regents of the University of California.

**Background**

For this project we are assembling a reference genome for delta smelt in order to 1) calculate a more precise estimation of contemporary effective population sizes (NE) for the 2017 to 2019 wild cohorts, 2) examine domestication selection at discrete locations in the genome, and 3) identify sex-specific markers and/or chromosomes in delta smelt. The creation of a reference genome for delta smelt will enable scientists to more fully take advantage of the decreasing samples available and use archival samples more effectively. All deliverable statuses are listed in Table 1.

The use of next generation sequencing (NGS) technologies in conjunction with a highly contiguous and accurate reference genome, increases the power and precision of inferences made in population genetic studies (such as structure, diversity, understanding local adaptation).1 DNA sequences unaligned to a reference genome fail to account for how each piece of sequenced DNA interacts with all other sequences (i.e. linkage patterns). However, by aligning to a reference genome we have knowledge of where each piece of sequenced DNA lies within the genome and relative to other sequenced DNA. Additionally, by knowing each markers distance from one another, and how it is or is not inherited with other markers we can survey a greater number of informative markers spread throughout the genome. This creates the ability to better answer questions involving demography, adaptation, fitness, and disease susceptibility.2,3 For example, previous work in Chinook salmon using the closely related rainbow trout reference genome found small amount of markers that were associated with runtime.4 In 2014, these markers were located next to a gap in the rainbow trout reference assembly. Reanalysis using a highly contiguous 2018 Chinook salmon reference genome found diagnostic markers located within the gap in the 2014 reference genome (Prince *et al.* unpublished).5 This example underscores how reference genomes can propel conservation research forward.

Recently, genome sequencing technologies have become both more cost effective and efficient. Hybrid assemblies, that is assemblies that use multiple NGS technologies, are the most reliable way to achieve a chromosome-scale hi-quality genome assembly. The Vertebrate Genomes Project, a consortium aimed towards developing an assembly pipeline and quality standards for genome assemblies of all vertebrates, established quality goal metrics for the continuity, completeness and accuracy of reference genomes6. Therefore, a main goal of this project is to develop a highly accurate chromosome-scale reference genome, hereafter called “reference genome”, using linked-read, long-read and hi-c chromatic conformation (Figure 1).

**Materials and Methods**

**Ethics Statement**

This study was approved by the UC Davis Review Committee for IUPAC Protocol #21533 and permitted under Department of Fish and Wildlife California Endangered Species Act Memorandum of Understanding (expiration date December 31, 2022).

**Sampling & DNA Extraction**

Tissue samples were collected from captive male and female fish 600 days post hatch (dph) located at the Fish Conservation and Culture Laboratory in Byron, CA. Muscle, gill, internal organ and scale tissues were sampled from each sex, the tissue was immediately flash frozen and stored on dry ice for transportation. High molecular weight (HMW) genomic DNA (gDNA) was extracted at the UC Davis DNA Technologies Core (Davis, CA) using the protocol described in Wasko et al. (2003).7 The size range of extracted gDNA fragments were determined using a pulse field gel run at a low frequency as to not shear the gDNA, for 24 hours. Physical sampling and extractions were repeated until the mean distribution of extraction lengths was 50kb or greater and there was sufficient quantity of HMW DNA to carry out each of the three required sequencing technologies (linked-read, long read and hi-c) for each sex (Supplemental Figure 1).

**Library Prep & Sequencing**

*Linked reads*

Once sufficient extracted fragments were acquired, genomic DNA was adjusted to a concentration of 0.91 ng/µl. Whole genome sequencing libraries were prepared using 10X Genomics Chromium Genome Library & Gel Bead Kit v.2 (10X Genomics, cat. 120258), Chromium Genome Chip Kit v.2 (10X Genomics, cat. 120257), Chromium i7 Multiplex Kit (10X Genomics, cat. 120262), and Chromium controller according to manufacturer’s instructions. After library preparation 1.14 ng of template gDNA was loaded on a Chromium Genome Chip and sequenced to roughly 80x coverage on an Illumina NovaSeq6000 150bp PE lane (Illumina, San Diego, CA).

*Long reads*

PacBio HiFi SMRTbell® Libraries (<https://www.pacb.com/>) were prepped following the SMRTbell Express Template Prep Kit 2.0 procedure. The UC Davis Sequencing Center used a Megaruptor to shear DNA to an average length of ~15kb. The input for the library prep was 5.6ug of DNA and the library was size selected to 11kb. Prepped DNA was run on a Sequel II machine for a total of 30 hours.

*Hi-C Chromatin Confirmation Capture*

Hi-C library prep and subsequent sequencing were carried out at Phase Genomics in Seattle, WA.

**Quality Control & Genome Size Estimation**

*Illumina based reads (linked reads & hi-c)*

We used the software program kat6 and fastqc <cite> to perform quality control on raw Illumina files. First, to see if there is evidence of contamination, we observed the number of distinct k-mers at different frequencies for k-mers of length 16, 21, 27 and 33.

After sequencing, bioinformatics quality control (QC) steps are necessary at multiple stages in the workflow of assembling of a reference genome. Immediate sources of error in sequencing data can result from sequenced DNA being contaminated with off-target DNA in the wet lab during preparation for sequencing, or biases in base calls from the sequencing machine. To QC for these two types of error, we conducted three computational quality control steps (kat hist, kat gcp and kat comp) using the software program KAT8. Each step splits sequencing data into sub-sequences of a given length, or k-mer, (where k is equal to a specified sequence length), and plots out frequencies, or comparisons, to visually inspect the data for quality issues. All bioinformatics work was conducted on the UC Davis farm compute cluster (the farm).

*Long reads*

asdf

**Assembly**

*Initial assembly from long reads*

We performed an initial phased genome assembly using the PacBio long reads with the Improved Phased Assembly (IPA) HiFi Genome Assembly software <cite>. Duplicated haplotigs were purged and the assembly was polished within the default parameters of the software.

*Scaffolding with linked-reads*

The initial phased assembly was then scaffolded using the 10X genomics Illumina linked-reads and the software scaff10x <cite>.

*Scaffolding with hi-c data*

We then mapped and formatted our Hi-C reads onto the scaffolded assembly using bwa-mem <CITE>, samtools <CITE>, picard <CITE>, bedtools <CITE> and custom perl scripts following the arima mapping pipeline <CITE>. After mapping the hi-c reads we performed one round of scaffoldind with SALSA2 <CITE>.

*Chromosome-level genome assembly using linkage map*

In order to achieve a chromosome level assembly we used linkage map data from Lew et al (2015)<CITE> to anchor scaffolds onto chromosomes using the Chromonomer software.

**Ctyogenic Chromosome Validation**

*Karyotype*

Fish were anesthetized with Colchicine [(10 microliters of 1 mg/ml stock) and returned to water. Individual fish lengths were measured. In both cases organs were rinsed, stored in PBS at ambient temperature of the CABA environment. We established two spleen pools, one of male (n = 15), and one of female (n = 13) specimens and one additional male gonad was harvested. Organ pools were gently aspirated into single cell suspensions by pipetting in hypotonic solution (0.56% KCl) for 15-20 min. Cells were centrifuged at ~1000 rpm for 10 min, supernatant hypotonic solution was removed and a 3:1 fixative (methanol:glacial acetic acid) was added. Cell pellets were resuspended and stored at 4ºC. Two to three more fixative washes (centrifugation, resuspension in new fixative) were conducted, and cells were applied to slides one week later. Slides were stained using the DNA staining fluorescent dye (DAPI) and cells were examined using an Olympus BX-40 Microscope. We then captured and stored images using Cytovision Software and determine chromosome number of this species from those images.

**Results**

**Sampling & DNA Extraction**

A total of eight male and five female delta smelt were sampled from the FCCL during three different sampling trips. In order to acquire sufficiently long HMW gDNA 23 extractions were carried out.

**Library Prep & Sequencing**

One male fish was used for all sequencing used for the male assembly. Due to the low concentration of HMW DNA two separate female fish were used to acquire enough genomic DNA for all the three different sequencing types.

*Linked reads*

Tktktkt report coverage

*Long reads*

tktktktk give coverage

*Hi-C Chromatin Confirmation Capture*

Tktktk give coverage

**Quality Control & Genome Size Estimation**

*Illumina based reads (linked reads & hi-c)*

For the first QC step to detect potential contamination, within KAT, we used kat hist and observed the number of distinct k-mers at different frequencies for lengths 16, 21, 27 and 33. (Figure 3A-D) in the female sample and k-mers of length 19, 23, 27 and 31 (Figure 3E-H) in the male sample. Uncontaminated samples are expected to have a single peak with a lot of k-mers at a very low frequency due to sequencer errors.

The second QC contamination step looks for signs of contamination by plotting the proportion of the k-mer comprised of the G and C nucleotide against the frequency of the that k-mer in the sequencing data and the number of distinct k-mers for a given GC count vs. frequency. Contaminated samples are expected to have a non-normal distribution. For example, samples contaminated with bacteria will have more k-mers with GC counts above 50%. We plotted GC counts against the frequency of k-mers of length 16, 21, 27 and 33 in the female sequencing data (Figure 4A-D) and k-mers of length 19, 23, 27 and 31 in the male sequencing data.

The third QC step (sequencing bias between forward (R1) and reverse (R2) sequences) plots the frequency of a given k-mer in each of the paired-end sequence data files. Sequencing bias in either of the two files would result in an irregular pattern in the number of distinct k-mers. We plotted the number of distinct k-mers at different frequencies for the R1 and R2 files against one another for k-mers of length 16, 21, 27 and 33 in the female sample (Figure 5A-D) and k-mers of length 19, 23, 27 and 31 in the male sample (Figure 5E-H).

We have included female plots to provide a comprehensive view of all quality metrics in the linked-read sequencing for this project. We observed single peaks for each histogram indicating contamination is unlikely (Figure 3A-H). All GC count frequency plots show a single peak of distinct k-mers, meaning there are no signs of bacterial or organelle contamination (Figure 4A-H). The number of distinct k-mers does not appear to be skewed indicating that the raw data does not appear to have major sources of sequencing bias (Figure 5A-H).

*Long reads*

Tktktk

**Assembly**

*Initial assembly from long reads*

tktktk

*Scaffolding with linked-reads*

tktkt

*Scaffolding with hi-c data*

tktktk

*Chromosome-level genome assembly using linkage map*

tktktk

**Ctyogenic Chromosome Validation**

*Karyotype*

Tktktk

**Discussion**

Tktktk

**Tables & Figures**

**Table 1**. List and description of deliverables with the status of each task.

|  |  |  |
| --- | --- | --- |
| **Deliverable** | **Description** | **Status** |
| **Task 1. Sequence and assemble a draft delta smelt genome + linkage map** | | |
| **Task 1.1 Linked-read sequencing (10X Genomics)** | | |
| a. Sample acquisition | Sample tissue of male (M) and female (F) fish. | Completed |
| b. Sequencing | 10X Chromium Library Prep | F Completed,  M Completed |
| Illumina NovaSeq Sequencing | F Completed,  M completed |
| c. Assembly | Supernova assembly | F Completed,  M Completed |
| Assembly QC | F Completed,  M Completed |
| tigmint assembly correction | In progress |
| **Task 1.2 Long-read sequencing (PacBio)** | |  |
| a. Sample preparation | Sample tissue and organs of male and female fish. | F completed,  M completed |
| b. Sequencing | PacBio Prep | F completed,  M completed |
| HiFi Sequencing | F in progress,  M completed |
| c. Assembly | Canu assembly | F in progress,  M completed |
| Assembly polishing | In progress |
| **Task 1.3 Hi-C Chromatin Conformation Capture & Sequencing (Phase Genomics)** | | |
| a. Sample preparation | Sample tissue and organs of male and female fish. | F completed,  M completed |
| b. Sequencing | Hi-C Library Prep | F completed,  M in progress |
| Deep Sequencing | F completed,  M in progress |
| c. Assembly | Salsa2 scaffolding | In progress |
| Assembly curation | In progress |
| **Task 1.4 High-density linkage map** | | |
| a. SNP Discovery | Sample tissue and organs of male and female fish. | In progress |
| b. Construction | Construction of a linkage map using HighMap | In progress |
| **Task 1.5 Map linked regions to reference genome** | | |
| a. Map regions | Align linked regions to genome | In progress |
| **Task 2. Estimate contemporary Ne using RAD-seq data & draft genome** | | |
| **Task 2.1 DNA collection and extraction** | | |
| a. Sample Collection | Collect samples from BY2017, BY2018 and BY2019 cohorts | Completed |
| b. DNA Extraction | Extract DNA from BY2017, BY2018 and BY2019 cohorts | BY2017 completed  BY2018 & BY2019 in progress |
| c. RAD-sequencing | Sequence BY2017, BY2018 and BY2019 cohorts | BY2017 completed  BY2018 & BY2019 in progress |
| **Task 2.2 Estimate Ne in BY2017, BY2018 and BY2019 cohorts** | | |
| a. RAD-sequencing data processing | Split, align and quality control all sequencing data | In progress |
| b. Estimation of Ne | Temporal Ne  LD Ne | In progress |
| **Task 3. Domestication selection** | | |
| **Task 3.1 Select individuals for RAD-seq analysis** | | |
| a. Determine domestication index (DI) | Calculate domestication index for each individual from each generation | Completed |
| b. Group individuals based on their DI | Assign individuals to three DI group: Low, medium and high | Completed |
| c. Determine recovery rate of each individual | Calculate recovery rate of each individual | Completed |
| d. Group individuals based on their recovery rate | Assign individuals to two groups: Low and high recovery rate | Completed |
| **Task 3.2 Preparation and sequencing of individuals** | | |
| a. Sample Collection | Collect samples from BY2008 to BY2015 cohorts | Completed |
| b. DNA Extraction | Extract DNA from BY2008 to BY2015 cohorts | Completed |
| c. RAD-sequencing | Sequence BY2008 to BY2015 cohorts | Completed |
| **Task 3.3 Population Genetic Analysis** | | |
| a. RAD-sequencing data processing | De-multiplexing  Quality filtering | Completed |
| a. Identifying SNPs of domestication selection | Genotype calling  Fst analysis between groups with completed reference genome assembly | In progress |
| d. Association study | GWAS between DI and recovery rate | In progress |
| **Task 4. Identification of sex specific markers** | | |
| **Task 4.1 Investigation of delta smelt genome** | | |
| a. RAD-sequencing data processing | De-multiplex | Completed |
| QC Sequencing data | Completed |
| b. Identification of sex marker(s) | GWAS in M vs F | Completed |
| Depth comparison | Completed |
| **Task 4.2 Fluidigm assay creation** | | |
| a. Submit sequences | Submit sex specific sequences to Fluidigm for SNP Type Assay design | In progress |
| b. Validation | Validation of SNP Assay | In progress |

**Table 2.** Table of tissue type and storage method of sampled delta smelt from three different trips. Included are the names referred to in the text. T= trip, F= female, M=male, BM = back muscle, SC = scales, IO = internal organ, FF = flash frozen, and PG = propylene glycol.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Trip Number** | **Sex** | **Sample Name** | **Tissue Type** | **Storage Method** | **Used for Sequencing** |
| 1 | F | T1F01 | Back muscle | Flash frozen | - |
| T1F02 | yes |
| T1F03 | - |
| M | T1M01 | - |
| T1M02 | - |
| T1M03 | - |
| 2 | M | T2M01 | Back muscle | Flash frozen | - |
| T2M02 | - |
| T2M03 | - |
| 3 | F | T3F01\_BM\_PG | back muscle | propylene glycol | - |
| T3F01\_BM\_FF | back muscle | flash frozen | - |
| T3F01\_IO\_FF | internal organ | flash frozen | - |
| T3F01\_SC\_FF | scales | flash frozen | - |
| T3F02\_BM\_PG | back muscle | propylene glycol | - |
| T3F02\_BM\_FF | back muscle | flash frozen | - |
| T3F02\_IO\_FF | internal organ | flash frozen | - |
| T3F02\_SC\_FF | scales | flash frozen | yes |
| M | T3M01\_BM\_PG | back muscle | propylene glycol | - |
| T3M01\_BM\_FF | back muscle | flash frozen | - |
| T3M01\_SC\_FF | scales | flash frozen | - |
| T3M02\_BM\_PG | back muscle | propylene glycol | - |
| T3M02\_BM\_FF | back muscle | flash frozen | yes |
| T3M02\_SC\_FF | scales | flash frozen | - |

Diagram

Description automatically generated

**Figure 1.** Workflow diagram for completion of assembling reference genome for delta smelt (Task 1). Both male and female delta smelt will undergo the pipeline separately to produce two reference genomes––one male and one female. All work was completed at the GVL unless otherwise noted. A.) Tissue from male and female delta smelt are sampled from the captive colony at the FCCL. B.) HMW DNA is extracted from the fish tissue samples. C.) Extracted DNA of sufficient length undergoes a library prep, subsequent sequencing and quality control for each of the three NGS technologies (linked-read, long-read and hi-c). D.) Linked- and long-read sequencing data are assembled into individual draft genomes. The draft genomes are quality assessed by the software BUSCO13. E.) The linked- and long-read assemblies are scaffolded together to produce a consensus assembly. The quality of the consensus assembly is then assessed by the software BUSCO13. F.) Hi-c data connects long-range gaps in the linked- & long-read consensus assembly. The quality of the consensus assembly is then assessed by the software BUSCO13. G.) Final polishing to produce a reference genome.

A picture containing text, electronics, computer

Description automatically generated

**Figure 2.** Pulse field gel images of extracted HMW gDNA from three sampling trips (“T”) and six rounds of extractions (“E”). Green boxes surround lanes from extracted samples usable for long-read and linked-read sequencing (extraction distribution centered ~ 50kb), yellow boxes surround lanes from samples with insufficient extraction lengths. A) T1, E1: one usable female sample (T1\_F02) B) T2, E2: no usable samples C) T2, E3: no usable samples D) T2, E4: no usable samples (E) T3, E5: one usable male sample (T3M02\_BM\_FF) F) T3, E6: one usable female (T3F02\_SC\_FF).