**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salmon reference genome found diagnostic markers located within the gap in the 2014 reference genome (Prince *et al.* unpublished).5 This example underscores how reference genomes can propel conservation research forward.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).

**Task 1: Genome assembly**

***Sample collection (Figure 1A)***

*Methods*

The first step in a genome assembly is to collect tissue for the extraction of high molecular weight (HMW) genomic DNA, which is extracted DNA that is longer in length (>50 kilobases) than standard DNA extractions. Because all of the sequencing technologies utilize sequence fragments on the order of tens of thousands of base pairs long or rely on long range interactions throughout the genome, extracted DNA for generating a contiguous genome assembly must be sufficiently long. Due to the difficulties in getting sufficient quantity of HMW DNA, we conducted four trips to sample different tissue types from fish 600 days post hatch at the FCCL (Table 2). At the start of this project long-read sequencing was costly and we sought to only incorporate the sequencing technology if absolutely necessary; because of this our first and second sampling trips sought to acquire enough tissue to provide sufficient quantity of HMW DNA for linked-read and hi-c sequencing. However, at the end of 2019 the price of long-read sequencing dropped dramatically and based on our results from trips 1 and 2, for trip 3 we sought to acquire enough tissue to sequence a single male fish with all three of the technologies and a female fish with the last long-read technology. However, prior to acquiring the last sequencing technology for the male fish, our samples were thrown out by the sequencing center and it was necessary to sample one additional male specimen on a fourth trip. On each occasion fish were euthanized according to the approved animal care protocols/standard operating procedures. Back muscle, internal organ and/or scale tissues were sampled onsite at the FCCL and cooled for transportation directly to the UC Davis Sequencing Center.

*Results*

We took a total of four trips to sample tissue. On the first trip we sampled back muscle from three males and three females (Table 2). These samples were flash frozen using liquid nitrogen and transported on dry ice. While we obtained sufficient HMW DNA from a female specimen, none of the sampled male fish from this trip produced enough HMW DNA. Therefore, we returned to the FCCL on a second trip and selected larger males. Extraction lengths from these male fish were also insufficient. A decision to take a third trip was made due to the availability of a new sampling method that uses additional tissue types (not just back muscle tissue) and new tissue preservation storage solution of samples in cooled propylene glycol rather than only flash freezing in liquid nitrogen. On the third trip we sampled back muscle tissue, internal organs, and scales from a total of two males and two females. Additionally, because DNA can be fragmented in the freeze-thaw process, we hedged our bets for sampling on third trip by preserving half of all sampled tissues in propylene glycol at 4°C, and half with the protocol used on trips 1 and 2 (flash freezing with liquid nitrogen and transporting on dry ice) which was successful for a prior female fish. Due to a sampling being thrown out we had to carry out one additional sampling in order to acquire tissue for hi-c sequencing. On the fourth trip we sampled back muscle, scales and internal organs from one male fish.

***Isolation of high molecular weight genomic DNA (Figure 1B)***

*Methods*

For all three sequencing types (linked-read, long-read and hi-c), HMW DNA is required. For the linked-read and long-read sequencing, HMW DNA extractions from the fish tissues occurred at the UC Davis Sequencing Center using the protocol described in Wasko et al. (2003)7. We acquired samples to extract DNA from both sexes to examine sex determining regions in delta smelt, which is currently unknown. The size range of extracted DNA fragments were determined using a pulse field gel run at a low frequency, to not shear the DNA, 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 the required sequencing technologies for each sex (Figure 2). For extractions and subsequent hi-c sequencing, a male and female sample was sent to Phase Genomics and the Vertebrate Genome Project, respectively.

*Results*

Trip 1: DNA from one male (T1M03) and one female (T1F02) DNA was extracted from the first sampling (See Table 2 for fish identifications corresponding to trips). HMW DNA from the T1F02 had DNA fragments of sufficient length for sequencing (Figure 2A).

Trip 2: We performed two rounds of extractions on two different males (T2M02 and T2M03). Yet despite multiple attempts at extraction, we did not obtain HMW DNA fragments of sufficient length (Figure 2B-D).

Trip 3: We performed two rounds of extractions on multiple different tissue types (Table 2). From these extractions we acquired DNA of sufficient length from flash frozen back muscle tissue from flash frozen back muscle tissue from one male (T03M02\_BM\_FF, Figure 2E) and flash frozen scales from one female (T3F02\_SC, Figure 2F). We did not observe that storing tissue in propylene glycol had an effect on increasing the fragment length of extracted DNA (Figure 2E & F).

Trip 4: The tissue samples from the male specimen sampled on trip 4 were sent directly to the Vertebrate Genome Project for subsequent extraction and sequencing.

***Long-read library prep & sequencing (Figure 1C)***

*Methods*

The inclusion of long-read data dramatically assists in creating a more contiguous assembly by spanning repetitive elements and resolving chimeric sequences throughout the genome. Long-reads provide greater continuity of scaffolded contigs and resolving repetitive sequences within genome assemblies, because the reads (sequence fragments produced from the sequencer) are thousands rather than hundreds of base pairs long. Previously, long-read sequencing was both costly and had an error rate up to 10%. Within the past year, new PacBio HiFi chemistry has come out to dramatically reduce the number of errors from almost 10% to under 1%. Additionally, the price of PacBio long-read sequencing has dropped dramatically, making sequencing a male and female fish affordable for this project.

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.

*Results*

tktkt

***Long-Read Post-Sequencing QC (Figure 1C)***

*Methods and preliminary results*

Sequencing data stored in binary files were downloaded from Bioshare, the UC Davis Sequencing Center’s host service. We used CCS software’s (<https://github.com/PacificBiosciences/ccs>) statistical model on raw reads from male delta smelt to generate highly accurate consensus sequences with known base quality values and convert data to a fastq format for downstream analysis. A base quality value assigns a score to each letter of sequencing data to denote the level of confidence in each base called by the sequencer. Next, we will acquire sequence data from the female fish and perform the same QC steps on that sample.

***Linked-read library prep & sequencing (Figure 1C)***

*Methods*

Once sufficient extracted fragments were acquired, genomic DNA was adjusted to a concentration of 0.91 ng/µl. We selected the 10X Genomics platform (<https://www.10xgenomics.com/technology/>) for our linked-read sequence data. 10X Genomics library preparation takes extracted HMW gDNA, shears the DNA into 50kb long segments, inserts each fragment into an oil coated gel emulsion bead (GEM), further shears DNA within each bead, and attaches unique barcodes to all DNA fragments within each GEM for identification post-sequencing. Whole genome sequencing libraries were prepared using 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).

*Results*

Previous RAD-based estimates of genome size predicted the genome to be approximately 1.2Gb. We acquired XX bases for an estimated 80x. However, when we received the linked-read sequencing data we used a kmer based estimation technique to estimate a genome size of 0.9Gb. As such, we sequenced the male assembly to roughly 120x coverage. We amended the amount of data generated to XX and sequenced the female sample to roughly 80x coverage (Table 3).

***Linked-read post-sequencing quality control (Figure 1C)***

*Methods*

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-mers, (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).

*Preliminary results*

*Linked-read QC step 1 (contamination):*

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.

*Linked read QC step 2 (contamination):*

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.

*Linked read QC step 3 (sequencing bias).*

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).

In the past year we performed computational QC on the male sequencing data. 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).

***Hi-C Chromatin Conformation Capture Sequencing & Post-Seq QC (Figure 1C)***

*Methods*

Genomic DNA in eukaryotes has high levels of repetition, leading to unresolved gaps surrounding large repetitive elements such as in the middle and at the ends of each chromosome. In order to resolve these gaps, we are using hi-c chromatin conformation capture (hi-c). We outsourced hi-c library prep and sequencing to Phase Genomics in Seattle, WA. Tissue from the same female that was sequenced with linked-read sequencing was sent to Phase Genomics for hi-c library prep and subsequent sequencing. Raw sequencing data and an initial scaffolding report were received for the female sample.

*Preliminary results*

At present we have downloaded the raw sequencing data and associated files from Phase Genomics’ Google drive. Initial reports from Phase Genomics indicate a successful library prep and sequencing. A total of 87,444,477 read pairs were analyzed of which 56.38% where high quality with a minimum mapping quality of greater than or equal to 20, a maximum edit distance of less than or equal to 5, and no duplications. The data contained an average of 2,966.33 read pairs per contig greater than 5kb. These data appear normal and indicate they will be useful in creating a more contiguous assembly.

***Long-Read Draft Assembly (Figure 1D)***

*Methods and preliminary results*

After QC,an initial assembly with PacBio sequencing data strings together sequencing reads to create a draft assembly using the hicanu implementation of the canu version 2.0 software9 (preprint). This software processes reads to eliminate homopolymers (repeats of a single base pair) and then strings together reads with overlapping sequences. It then expands the sequences to include the information about the homopolymers and validates the assembly.

We created an initial draft assembly with our male long-read data using the hicanu implementation of canu version 2.0. Next we will process the female long-read sequence data similarly.

**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*

A 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