**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, Alisha Goodbla, 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.

# Project Background

For this project we are assembling a reference genome for delta smelt in order to 1) calculate a more precise estimate 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. The status of each deliverable is 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 including analyses of population structure, genetic diversity and local adaptation.1 DNA sequences which are captured by the sequencer but not aligned 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 gain the knowledge of where each segment of sequenced DNA lies within the genome and relative to other sequenced DNA. Additionally, by knowing each marker’s distance from every other marker, and how it is or is not inherited with those other markers, we can survey a greater number of informative markers spread throughout the genome. This gives us the ability to better answer questions concerning demography, adaptation, fitness, and disease susceptibility.2,3 For example, previous work in Chinook salmon using the closely related rainbow trout reference genome found a small number 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” (assemblies that use multiple NGS technologies) are the most reliable way to achieve a chromosome-scale high-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 genomes.6 Therefore, a main goal of this project is to develop a highly accurate chromosome-scale reference genome, hereafter called “reference genome”, using linked-read (Box 1), long-read (Box 2), hi-c chromatin confirmation capture (Box 3), and a genetic linkage map (Figure 1).

**BOX 1: Linked-read Sequencing**, such as 10X Genomics used in this study, takes large segments of extracted DNA (around 50kbp) found in an organism’s genome and puts them into individual oil beads. Within each oil bead, the segment of DNA is sheared to a length that an Illumina sequencer can read (~300-500bp) and barcodes corresponding to the oil bead the fragments are found in are attached on either side. All of the oil beads are then pooled together and sequenced to get highly accurate base calls with pseudo-long reads.



Image taken from: https://genome.med.harvard.edu/services/singleCell/ViewOverview.action

**BOX 2:** Long-read Sequencing, such as PacBio HiFi used in this study, takes long segments of DNA (over 50 kbp), shears them into a few segments (10-20kbp) and circularizes each. The circular molecules are then run through a sequencer to get an accurate circular consensus sequence over 20kbp in length.

Image taken from: https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/

**BOX 3:** Hi-C Chromatin Confirmation Capture takes DNA in the nucleus and links all of the locations that are touching together. It then shears the linked DNA into lengths appropriate for short-read sequencing (300-500bp). The frequency two segments of DNA are found linked together is inversely proportional to their distance in the genome. That is, if two segments of DNA are close to one another, they will be sequenced together many times.



Image taken from: https://data.4dnucleome.org/experiment-types/dilution-hi-c/

# Task 1: Genome assembly

## Sample collection (Figure 1A)

*Methods*

The first step in our 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 (<10 kilobases). Because each of the three sequencing technologies (linked-read, long-read and hi-c) we utilized to create our genome assembly sequence fragments on the order of tens of thousands of base pairs, or rely on long range interactions throughout the genome, extracted DNA must be sufficiently long to generate a contiguous assembly.

Due to the difficulties of recovering sufficient HMW DNA, we conducted four separate trips to sample different tissue types from male and female delta smelt 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, during Trip 3 we sought to acquire enough tissue to sequence a single male fish with all three of our chosen technologies and a female fish with long-reads, as we had already generated female sequencing data for linked-reads and hi-c. An error committed by the sequencing center required us to make one additional trip, Trip 4, to sample one additional male specimen. 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 DNA Technologies & Expression Analysis Core Laboratory (UC Davis Sequencing Center).

*Results*

We took a total of four trips to sample tissue. On the Trip 1 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 and selected larger males on Trip 2. Extraction lengths from these male fish were also insufficient. A decision to take Trip 3 was made due to the availability of a new tissue sampling method that uses additional tissue types (not just back muscle tissue)<CITE> and new tissue preservation storage solution of samples in cooled propylene glycol<CITE> rather than solely flash freezing in liquid nitrogen. On Trip 3, 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 Trip 3 and preserved half of all sampled tissues in propylene glycol at 4°C, and half flash frozen and transported both on dry ice. On Trip 2, we sampled back muscle, scales and internal organs from one male fish. Sampled tissue was flash frozen and stored on dry ice for transportation.

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

*Methods*

All three of our chosen sequencing technologies (linked-read, long-read and hi-c), require HMW DNA. 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 continues to remain 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. For HMW DNA extractions and subsequent hi-c sequencing, a female tissue sample was sent to Phase Genomics and a male tissue sample was sent to the Vertebrate Genome Project for isolation of HMW DNA. Samples were sent to different locations due to COVID-19 delays at Phase Genomics during the time of sampling the male fish.

*Results*

Trip 1: DNA from one male (T1M03) and one female (T1F02) DNA was successfully 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 three separate rounds of extractions on tissue samples from two different males (T2M02 and T2M03). Yet despite multiple attempts at extracting HMW DNA, we did not obtain fragments of sufficient length for sequencing (Figure 2B-D).

Trip 3: We performed two separate 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 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 by 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%. During 2020, new PacBio HiFi chemistry has come out to dramatically reduce the number of errors from almost 10% to under 1%.

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 and sequencing data was collected as recorded “movies” of nucleotides. Each movie collection lasted for 30 hours per run. Additional movies were collected until the amount of sequencing data for each sex was ~25-30x coverage.

*Results*

A total of five movie collections (150 hours of sequencing data were collected). Two male movie collections generated roughly 25x coverage (data combined into Supplemental Data 1) and three female movie collections generated roughly 28x coverage (Run 1 & 2 combined into Supplemental Data 2, Run 3 data in Supplemental Data 3).

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

*Methods*

Once sufficient extracted HMW genomic DNA fragments were acquired, 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 on an Illumina NovaSeq6000 150bp PE lane (Illumina, San Diego, CA). We used a previous RAD-based estimate of a haploid genome size (0.6Gb) to sequence the first female sample to an estimated 80x coverage. We then used this sequencing data to better inform our estimate of delta smelt genome size by using a more accurate k-mer based estimation approach with the software Genomescope. We adjusted the amount of linked-read sequencing data collected for the male sample accordingly.

*Results*

The k-mer based haploid genome size was estimated to be 0.49Gb. We generated approximately 45 gigabytes of female linked-read sequencing data and 30 gigabytes of male linked-read sequencing data for a total of roughly 120x and 80x coverage, respectively (Table 3).

## Hi-C chromatin conformation capture prep & sequencing (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 used hi-c chromatin conformation capture (hi-c). We outsourced hi-c library prep and sequencing to Phase Genomics in Seattle, WA. Phased used their in-house proprietary library preparation and sequencing protocols. Raw sequencing data and an initial scaffolding report were received for the female sample.

*Results*

We received sequence files of 87,444,477 read pairs in total.

Quality Control: 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. Other sources of error may come from having too great of a proportion of poor-quality reads.

## Long-read post-sequencing quality control (Figure 1C)

*Methods*

Sequencing data stored in binary bam 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 to generate highly accurate consensus sequences with known base quality values and converted binary data to 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. Reads with quality scores over Q20, denoting an error probability of 0.01%, were used for subsequent assembly.

*Results*

A total of 3,095,133 male reads and 2,741,504 female reads representing 35,841,976,770 and 28,549,585,055 base pairs, respectively, passed quality control and will be used for subsequent assembly.

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

*Methods*

To quality control for contamination and sequencing bias errors, we conducted three computational quality control steps (kat hist, kat gcp and kat comp) using the software program KAT8<CITE>. 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).

We looked for signs of contamination using the kat hist and kat gcp functions within the software KAT<CITE>. First, we used the kat hist function to check for signs of contamination by plotting a histogram of the observed number of distinct k-mers at different frequencies for lengths k=21, 31, and 41 for female and male sequencing data. Uncontaminated samples are expected to have a single peak with a surplus of k-mers at a very low frequency due to sequencer errors. Second, we used the kat gcp function to plot the proportion of the k-mer comprised of the G and C nucleotides 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 k=21, 31, and 41 for the female and male sequencing data.

Lastly, we evaluated the data for sequencing bias between the forward (R1) and reverse (R2) files. We used the kat comp function to plot 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 k=21, 31, and 41 for the female and male sequencing data

*Results*

We saw clear peaks in the histograms for each sample and at each value of k (Figure 3). All GC count frequency plots show a normal distribution of distinct k-mers (Figure 4). Additionally, the number of distinct k-mers does not appear to be heavily skewed in the male or female sequencing (Figure 5). These data together indicate no observable signs of bacterial or organelle contamination or major sources of sequencing bias in our sequencing data.

## Hi-C chromatin conformation capture post-sequencing quality control (Figure 1C)

*Methods*

In order to assess if our sequencing data would be useful in linking scaffolds. We looked at a percentage of high-quality reads (minimum mapping quality of greater than or equal to 20, a maximum edit distance of less than or equal to 5, and no duplications) that mapped to our draft assembly created from the next step. We also observed the number of reads which aligned to each contig (>600 desired) and the number of reads that are sufficiently far apart (1-15% expected).

*Results*

Sequencing data reports from Phase Genomics indicate a successful library prep and sequencing (Supplemental Data 4). A total of 56.38% of reads were considered high quality. The data contained an average of 2,966.33 read pairs per contig greater than 5kb and 18.78% of the read pairs mapped to greater than 10 kilobases apart. These data appear normal and indicate they will be useful in creating a more contiguous assembly.

## Genome assembly (Figure 1D-H)

*Methods*

Step 1: Use long-read sequencing data to create Draft Assembly A

To begin, we generated two phased assemblies, one for each sex, by inputting the long-read sequencing data into the Improved Phased Assembly tool (IPA)<CITE>. The assembly product was polished of errors, purged of duplicate haplotigs, and phased into primary and alternative assembly files. The primary assembly file contains a contiguous haploid assembly, while the alternate assembly file contains the alternate haplotype of the diploid delta smelt.

Step 2: Incorporate linked-reads into Draft Assembly A to produce Draft Assembly B

After creating the initial draft assembly, we incorporated the linked-read data to first break the assembly at locations that were incorrectly joined, then to scaffold the assembly into larger, more contiguous chunks using the software scaff10x <CITE>.

Step 3: Incorporate hi-c data into Draft Assembly B to produce Draft Assembly C

In order to use long distance information, we indexed the assembly produced in the scaff10x step and mapped hi-c reads to the draft assembly using bwa<CITE> and samtools<CITE>. We used the Arima Mapping pipeline perl scripts<CITE> to pair reads, and quality filter the 5’ end and for mapping quality. Next, we added read group information, marked duplicated reads, and sorted the mapped read files with picard<CITE>. These data were then converted into mapped bed files using bedtools<CITE>. We then used the mapped bed files, scaffolded assembly and the initial alternative assembly as input to close gaps and further scaffold the assembly using the SALSA2 pipeline<CITE> with non-default parameters: -i 5 -x GATC -m yes. SALSA2 uses the frequency of how often reads pair (or are sequenced) together to determine how close any two locations in the genome are to one another, it finds the locations of the paired reads in the draft genome and links the two locations to close gaps and produce a more contiguous assembly.

Step 4: Use linkage map with Draft Assembly C to produce Final Assembly

Finally, we anchored our assembly into chromosomes by using a genetic linkage map produced in Lew et al (2015)<CITE> with the output from the hi-c assembly step and the software chromonomer<CITE>.

We used a combination of software and metrics to evaluate each draft assembly after every step of the assembly process (Table 4). The assembly length should be as close to the estimated genome size as possible. The N50 metric is the length of the scaffold where half of the assembly is held in scaffolds of that size or larger. The L50 metric tells the number of scaffolds that contain half of the assembly (Box 4). BUSCO scores were used to evaluate the completeness of each assembly as expected from a core set of highly conserved single copy genes in the Actinopterygii lineage.

*Results*

In each step the assembly length and N50 sizably increased and the L50 dropped precipitously. The final metrics for the male genome assembly were an N50 of 12,200,365 bp, an L50 =15, a total assembly length of 472,157,411 bp, with a total of 549 scaffolds. The final metrics for the female genome assembly were an N50 of 14,850,352 bp, an L50 =13, a total assembly length of 437,273,953 bp, with a total of 376 scaffolds. The final assemblies had BUSCO scores of 88.4% and 89.3% complete genes in the final male assembly female assembly, respectively (Table 4).

**BOX 4**: Assembly statistics

The N50 of an assembly is the length of contigs/scaffolds that contain 50% or more of the assembly.

The L50 of an assembly is the number of contigs/scaffolds that contain 50% or more of the assembly.



**N50 = 60**

**L50 = 3**

Figure taken from: https://www.molecularecologist.com/2017/03/29/whats-n50/

## Cytogenic (Karyotype) Chromosome Validation

*Methods*

We collaborated with Dr Mary Delaney to determine the diploid number of chromosomes in delta smelt, which has not been previously reported. This cytogenetic work does not use next generation sequencing-based methods and allows for an independent validation of the number of chromosomes expected in our final delta smelt genome assemblies.

Fish were anesthetized with Colchicine (10 microliters of 1 mg/ml stock) and returned to water after sampling. 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. Images of cells were captured and stored using the Cytovision Software and the number of chromosomes in the species were determined from those images.

*Results*

Seventy-five cell images were collected from the three pooled sample sets (mixed sex, males-only, females-only). Quality control steps discarded cells that were not intact, did not contain clearly defined chromosomes, or had significantly overlapping chromosomes. After quality control filtration, 18 cells were retained for counting/analysis. From these 18 cells, we determined the 2n (diploid) chromosome count for the delta smelt to be 56, with 15 cells exhibiting 2n=56 and 3 cells with hypomodal counts (1 cell with 2n=54, 2 cells with 2n=55, Table 5). Figure 6 shows a representative mitotic metaphase cell from a male spleen cell exhibiting 56 chromosomes.

# Task 2: NE Estimation

## Background

Two previous studies used microsatellite markers () to estimate the contemporary effective population size () of delta smelt, but found variable results.2,3 NGS technology presents more power to make precise estimates by increasing the number of putatively neutral markers.12–14

In combination with observing various population parameters, is a useful tool in monitoring endangered populations as it can inform how likely alleles are to be lost or fixed within a population.30,31 The of a given species is defined as the size of a Wright Fisher population that would have the same rate of change of a genetic parameter as the population under study.32 That is, an estimation of Ne is a single amount tied to the observation of many frequencies within a population based on the probability of called genotypes. Estimations can be highly variable depending on statistical priors and may not be useful as an absolute value taken alone and without further modeling. Therefore, the purpose of this study is to observe trends in the delta smelt population.

A number of methods exist to estimate both historical for evolutionary purposes, as well as contemporary, or short-term, to quantify more recent, modern day genetic diversity within a population.33 Contemporary methods use different genetic parameters such as linkage disequilibrium (LD) between two allele frequencies of individuals within on generation (inbreeding), and temporal allele frequency variance between generations (genetic drift).34–39 Today, LD and temporal methods are commonly used in conservation genetic studies and management as they can detect population declines up to as rapidly as one generation post- population decline40 and reliably detect population declines ten generations post decline.14 Early and reliable detection of diminished genetic diversity is important when protecting endangered species to seek to increase or maintain genetic diversity, as it is the raw component for natural selection to act on, and better undersand how management practices may be effecting the genetic diversity of delta smelt. 41

Two previous studies estimated contemporary in the wild population of delta smelt using 12-15 markers. Fisch et al. 2011 found the to be decreasing from 2003 to 2009 study period3. In contrast, Finger et al. 2017 found the could not be accurately estimated from 2011 to 2014 due to infinite confidence intervals2. Multiple factors may have contributed to this discrepancy: 1) different versions of NeEstimator42,43 were used between the two studies; 2) differences in the number of informative loci analyzed; or 3) the number of generations factored into the analysis. We sought to gain a broad understanding of the genetic diversity of delta smelt by estimating through a longer timeframe––from 1995 to 2019.

***Sample acquisition & sequencing***

*Methods*

We collected and sequenced archived samples collected in state and federal trawls from 1993-2024. Contemporary samples were collected and transferred into the custody of the Genomic Variation Laboratory as fin clips by Interagency Ecological Program surveys from 2015-2020. Genomic DNA was extracted using Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) according to the manufacturer’s protocol. In order to produce a large number of loci in a cost-effective manner, restriction site associated DNA (RAD) sequencing was carried out for all individuals. RAD libraries were prepared using the *Sbf1* restriction enzyme according to the ‘new RAD protocol’ described in Ali et al.,44 and sequenced 100 bp paired-end reads on an Illumina HiSeq 4000.

*Results*

A total of 2,976 samples from 24 years spanning 27 generations were sequenced.

## Read processing and alignment

*Methods*

In order to make demographic inferences, alleles spread across the genome of different individuals and contained within and across particular generations need to be compared. To do this, sequencing data must be split into individual files, aligned to the reference genome created in Task 1, and have genotypes called for each individual.

First, we used custom perl scripts to separate raw sequencing data into plates, then individuals. The separated (or split) individual sequencing files were then aligned to the genome using bwa <CITE> resulting in sequence alignment map (SAM) files. We then further processed the SAM files by sorting, filling in mate coordinates (samtools fixmate -m), removing duplicate reads (samtools markdup -r) and indexing the resulting files for create binary alignment map (BAM) files.

## Principle component analysis for hybrid detection

*Methods*

Delta smelt have been observed to hybridize with wakasagi smelt (*Hypomesus nipponensis*) in the SFE<CITE>. Due to the possibility of visual misidentification or technical error, we ran a principal component analysis to identify and exclude hybrid individuals or individuals with outlying genotypes.

To do this, we used the program ANGSD<CITE> to randomly sample a single read at all sites contained in at least half of the samples for each individual (angsd -doMajorMinor 1 -minMapQ 20 -minQ 20 -SNP\_pval 1e-12 -GL 1 -doMaf 1 -doCov 1 -doIBS 1 -doCounts 1). This creates a 0 to 1 matrix for each individual’s sampled allele at all locations in the form of a covMat file. We then calculated obtained eigenvalues using the program R<CITE>, calculated the observed variance for PC1 and PC2, and visualized the first and second principal components.

*Results*

We identified and removed individuals with a PC1 < 0.01 and/or PC2 > -0.01 from downstream analyses. A total of 19 individuals that were identified as hybrids or the result of a technical artifact (Figure 7).

## Contemporary Ne calculation

*Methods*

In order to standardize the number of gene copies contributing to the estimation of Ne, we subsampled to 50 individuals from each year. Years with fewer than 35 individuals were removed from subsequent analyses. We then called genotypes in the selected individuals using allele frequency as priors in ANGSD<CITE>. SNPs meeting the following criteria were accepted posterior probability greater than 0.85 (-postCutoff 0.85), a SNP p-value greater than 1e-6 (-SNP\_pval 1e-6), located in greater than 50% of individuals (-minInd 1472), minimum mapping quality of 20 of greater (-minMap 20), minimum base quality of 20 or greater (-minQ 20), and a minimum minor allele frequency of at least 0.05 (-minMaf 0.05), and genotypes were written as numbers (-doGeno 2) in a geno file. The geno file was read into R for further filtration using the snpR package<CITE>. Within snpR we filtered SNPs that violated Hardy-Weinberg Equilibrium (HWE=0.99) and did not have read coverage in at least 75% of individuals in each year (min\_ind=0.75).

We made one generation and all-by-all temporal method estimates of Ne using three different estimators, Nei & Tajima 1981, Pollak 1983, and Jorde & Ryman 2007. Temporal estimation gives the harmonic mean Ne between two different time points ( and ). For our first estimations we sought to set a specific time point by using years that were separated by one generation. One generation temporal method estimates of Ne were made between time points and in order to find the effective population size of generation . To do this we exported the snpR genotype data as a genepop file and imported it into NeEstimator V2 to make estimates in sequential years with sequencing data. Next, to get a broad understanding of how contemporary Ne is changing through time we performed all-by-all temporal method estimates. In these estimates and where increases to represent every other year in the dataset until estimates have been made for all possible combination pairs of generations.

~~We used a LD method to calculate Ne by generating info and option NeEstimator V2-formatted files through snpR. We acquired LDNe<CITE> estimates of effective population size for each generation and plotted these estimates through time. Since estimation of Ne through a temporal method is not currently supported in snpR, we exported the genotype data as a genepop file and manually created info and option files to carry out a temporal method of estimating Ne. Because rare alleles can upwardly bias estimates of Ne, we used three different “Pcrit” cutoffs (pcrit=005, 0.02 and 0.01) to discard values of estimates made from alleles or allele pairs have an observed frequency less than the cutoff in all estimates of Ne<CITE WAPLES 2006 AND WAPLES&DO2010>.~~

*Results*

Birth year 1993, 2000, 2007 were all removed from analyses due to having insufficient individuals (Table 6). A total of 27,809 loci were read into R for snpR filtration. After Hardy-Weinberg and minimum yearly individual purging a total of 911 loci remained. One generation yearly estimates were made from 1995 to 1999 and from 2008 to 2019 due to the availability of samples (Figure 8). All-by-all estimates were made, and data was plotted as a point located at the mean generation between and (Figure 9).

## Long-term genetic diversity calculation

*Methods*

Since the value of effective population size is directly related to genetic diversity through ,where =a measure of genetic diversity and is the organism’s mutation rate, we wanted to see if the size of Ne translated to genetic diversity loss. To quantify genetic diversity in the through time we used two different estimates of theta––the normalized number of segregating sites () and average pairwise nucleotide differences ().

First, we obtained a global estimate of the folded site frequency spectrum. To do this we obtained site allele frequency likelihoods using the reference genome as an ancestral state in ANGSD (angsd -GL 2 -doSaf 1 -minMapQ 10 -minQ 20), and then estimated the maximum likelihood of the site frequency spectrum (SFS) in realSFS (realSFS -maxIter 100 -fold 1). Next, we calculated thetas for each site by using realSFS (realSFS saf2theta) and converted those data into logscale per-site thetas with thetaStat (thetaStat print).

*Results*

Both and appear to be in decline. Estimates of range from TKTKT to tTKTKTK and estimates of range from TKTKT to TKTKT.

***Potential Problems***

Both temporal and LD methods of estimating account for the expected contribution from random sampling error () to the observed genetic characteristic of or , respectively. As such, estimates of can be negative with infinite confidence intervals due to the expected contribution of random sampling being larger than the actual contribution 57. If I observe negative estimates of , I will run replicate subsets of samples from each year to estimate and obtain an overall estimate of by taking the harmonic mean of all the estimates as prescribed in Waples and Do (2010).12 Additionally, a new method of jointly estimating and pedigrees as a single-sample method shows promise to give accurate estimates of contemporary .61 However, this method is under review.

# Task 4: Sex Marker

## Background

Sex determination in fish is a highly variable trait62 and understanding its mechanisms is crucial not only for understanding the biology of the individual species of fish but also for gaining insight into the evolution of sex chromosomes and genetic mechanisms underlying sex determination.63 Fish represent the most diverse group of vertebrates on earth with over 30,000 described species.64 With this diversity and constant exposure to variable environments comes a vast array of morphological, physiological, behavioral, developmental and sexual mechanisms.65–68 In teleost fishes, sex determination can be genetic or environmental and varies even between closely related species.62,69–71 Delta smelt are a unisexual species that do not appear to have environmental regulation of sex determination which suggests sex may be determined genetically. In teleost fishes, endogenous genetic sex determination mechanisms can occur at the chromosomal level, where heterogametic males (XY) or females (ZW) exist, or mechanisms can be at the genic level where single or multiple genes influence sex determination.70 While clarifying the mechanism of sex determination in delta smelt will increase our biological knowledge, it will also allow us to identify and develop diagnostic markers for the practical management of the species.

The ability to non-invasively identify sex in delta smelt will assist in management of the captive colony and develop knowledge of the biology of wild delta smelt. Currently, wild fish can only be sexed by the expression of gametes from ripe adult fish or through dissection, both sacrifice the life of the fish or gametes. When sex is identified by expression, pressure is put on the abdomen of fish until eggs are excreted (in females) or running milt (in males).72 Because sexual identification of wild fish primarily relies on the physiological status of an individual fish, only about two-thirds of wild adult delta smelt sampled can be sexed (Hammock pers. comm.). Knowledge of the genetic underpinnings of sex determination in fishes is a vital asset to the better management of captive populations, basic knowledge of life history characteristics of the species, ecological surveys and management regarding population metrics, species modeling, demographic inference, and sex-based survival.73

In order to non-invasively identify the sex of wild and captive delta smelt, we sought to identify potential candidate allele(s) which could be used as genetic diagnostics for classifications of sex.

***DNA sampling & sequencing***

*Methods*

To identify a sex specific marker or markers for delta smelt, we sampled adipose fin clips from 24 female and 24 male captive-bred individuals taken from the FCCL and sexually identified through either dissection or gametic expression. DNA was extracted using the Qiagen DNEasy 96 Blood & Tissue Kit with a modification of elution in100uL of H2O rather than the proprietary AE Buffer included with the kit.

Prior analyses that attempted to determine sex markers in delta smelt used the *Sbf1* restriction enzyme, which cuts DNA approximately every 65,000 base pairs, but no sex markers were identified. For this library preparation we sought to maximize the number of restriction enzyme cut sites and acquire reads from more locations throughout the genome. To do this we digested extracted DNA using the *Pst1* restriction enzyme, which shears DNA sixteen times as often as the *Sbf1* restriction enzyme, or roughly once every 4,100 base pairs. RAD sequencing libraries were prepared at the GVL according to Ali et al (2016)<CITE>, and sequenced at the UC Davis Sequencing Center with 150 bp paired-end reads on an Illumina HiSeq.

## Genome wide association study

*Methods*

We performed two rounds of a genome wide association study (GWAS) using the new male and female reference genomes. For each GWAS, we first aligned raw RAD sequencing data from 24 male and 24 female sexed fish to the reference genome. Next, we looked for the association of an allele at any location in the genome with sex classification by carrying out a dominant model case-control GWAS using males as controls (0) and female as cases (1) in the program ANGSD12. If we found alleles that associate with a particular sex that locus could be used as a diagnostic for genetically identifying sex. The association of a particular allele with sex category was reported as likelihood ratio test (LRT) statistic and is chi square distributed with one degree of freedom. We applied a conservative significance cutoff with a Bonferroni corrected p-value of 0.05 using the formula where is the number of loci analyzed, is the desired p-value or significance threshold (0.05), and is the adjusted p-value given the number of loci used in the analysis.

*Results*

We analyzed 848,444 and 922,975 loci spread across the male and female genome assemblies, respectively. These loci correspond to a Bonferroni corrected p-value cutoff of 5.893141e-08 in loci found within the male reference genome and 5.417265e-08 in loci analyzed within the female reference genome. No significant association in sex was found using the female reference genome. Two loci located in the male reference genome on Chromosome 5 were significantly associated with sex in delta smelt (Figure 7). The two SNPs most associated with sex in delta smelt had an LRT scores of 37.854854 and 35.802804 which correspond to p-values of 7.621e-10 and 2.183e-9, respectively.

## Depth analysis

*Methods*

If sex determination in delta smelt were caused by chromosomal differences, it would be expected that the heterogametic sex (i.e. XY) would have roughly half the sequencing depth of the homogametic sex (i.e. XX) at the large region that determines sex in the fish. Additionally, the heterogametic sex would be expected to have novel sequence content. To look for signs of sex specific sequencing depth differences, we looked for male and female differences in the presence and depth of RAD markers spread throughout the genome. To do this we performed two experiments, one using the male genome as a reference and another using the female genome as a reference. Each experiment used the 24 male and 24 female alignment files from the previous GWAS. First, we acquired the depth of aligned reads at each location in the reference genome using samtools depth. Next, using custom bash and perl scripts we threw out all locations where no male or female RAD sequencing data aligned, and we totaled the number of male alignments and gathered the total coverage for each sex (<https://github.com/shannonekj/DS_sex-marker/blob/master/analysis/>). Finally, we sorted the coverage difference files and looked for locations in the genome where one sex had high and consistent coverage and the other sex had shallow or no coverage.

*Results*

We found no large areas that corresponded to an inflation of any one particular sex having higher depth of coverage compared to the sex and could not identify any sex specific loci in this analysis.

## K-mer analysis

*Methods*

In our k-mer analysis we sought to identify unique differences of sequence content in males versus females. To do this, we used 10X Genomics linked-read data from one male and one female. First, we created a signature of all k-mers belonging to each sex with sourmash compute -scaled 1000 to sample one k-mer from ever 1000 base pairs. The signature compute step was followed by sourmash signature merge to incorporate all data from the R1 and R2 files for each sex. We purged the signature files of low abundance k-mers (abundance < 5) to eliminate k-mers that are more likely sequencing errors and threw out k-mers that were shared between male and females to only leave sex-specific k-mers. The resulting high abundance, single sex k-mers were used in our analysis.

First, we plotted and compared k-mer abundance for each sex to determine if there were observable differences between sexes. A difference in k-mer abundance could correspond to a sex determining regions within a sex (i.e. sex chromosome). Next, we extracted contigs which contain five or more k-mers within one contig which corresponds to a contig length of roughly 5,000bp. We then compared the abundance of male and female k-mers found within those contigs. After, we took the median abundance of k-mers within a contig to find the contigs abundance in each sex. We compared the male contig abundance to the female contig abundance and isolated male-only contigs deemed “putative Y” sequences for further validation.

To validate our results, we mapped RAD sequencing data to the putative Y sequences and ran a depth analysis. In order to ensure the putative Y reads were indeed mapping to one location in the genome, we performed a stringent end-to-end alignment of the putative Y sequencing data using bowtie2. Only reads that entirely aligned to regions in the genome were used in the subsequent analysis. To find depth locations, we aligned the RAD sequencing data to the male reference genome using samtools <CITE>. We then pulled RAD alignment depth information from all of the locations where the putative Y sequencing data had also aligned using the software bedtools<CITE> and custom bash and perl scripts. After we obtained depth information across all of the putative Y regions, we ran the same depth analysis as above.

*Results*

After abundance filtration, there were approximately 118,191,000 male-only k-mers and 494,251,000 female-only k-mers. There was a clear distinction between the distribution of male and female abundances, where males had more high abundance k-mers compared to females (Figure 8). Upon filtering k-mers for those found on long contigs (contigs containing 5 k-mers or more) there was a clear increase of male-specific k-mers at half the abundance of the female and male peak on the right (Figure 9). This is consistent with the male sequencing data potentially having heterogametic regions in its genome. We found 44 contigs with k-mer mean abundance in the male sequencing data that had zero abundance in the female sequencing data (Figure 10).

We mapped the putative Y data back to the male reference genome and found the reads mapped to multiple regions within the genome (Table 6). We did not find a significant difference in male versus female read depth at locations across the putative Y regions.

# Glossary

**BUSCO score** – the percent of highly conserved universal single copy orthologs found within an assembly. Percentage derived from assigned lineage (e.g. Actinopterygii)

**CABA** – Center for Aquatic Biology and Aquaculture

**contig** – a continuous stretch of DNA sequence created from a consensus of reads

**DNA** – deoxyribonucleic acid

**FCCL** – Fish Conservation & Culture Laboratory

**genetic** **marker** – a location in the genome with a SNP that can be used for analysis

**HMW** – high molecular weight

**movie** – the time specified for collecting sequencing data from a SMRT Cell

**PBS** – phosphate-buffered saline

**QC** – quality control

**scaffold** – a string of DNA sequences with potential gaps created from chaining contigs together. Created by using sequence data and relative position and orientation data. May contain gaps, which are denoted by the letter N.

**SNP** – single nucleotide polymorphism

**UC Davis Sequencing Center** – UC Davis DNA Technologies & Expression Analysis Core

# Software Versions

BUSCO

samtools=1.10

bwa=0.7.17-r1188

bedtools

chromonomer

htslib=1.10.2

SALSA2

scaff10x

kat

ccs

picard

IPA

Genomescope

**Works Cited**