**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 contract #A19-1844 between State Water Contractors and the Regents of the University of California.

# Prologue and Project Objectives

The abundance of delta smelt (*Hypomesus transpacificus*) has been in decline for decades1,2. Genetic tools have been useful for several management concerns: monitoring overall diversity and effective population size () in the wild population, genetic management in the captive population, developing assays to detect hybrids between delta smelt, wakasagi, and longfin smelt, and developing assays for eDNA sampling or species identification3–5. Attempts at quantifying in the wild population have been hampered by several interacting factors including a lack of power from using microsatellite markers and a very large historical 6,7. In order to improve estimates of , we assembled the delta smelt genome. We then leveraged the genome to estimate contemporary and the loss of genome-wide diversity between 1993 and 2020, interrogate domestication selection, and search for a sex marker. While is useful for monitoring purposes, we strongly advice against the use of alone for making management decisions.

For this project our tasks were to: 1) assemble a high-quality reference genome for delta smelt that is publicly available for all researchers, 2) estimate contemporary effective population sizes () for the wild 2017 to 2019 birth year cohorts of delta smelt, 3) search for genetic evidence of domestication selection across the genome, and 4) search for sex-specific markers and/or chromosomes in delta smelt. In order to obtain a more comprehensive understanding of through time, we have expanded Task 2 to include analysis of historical datasets of delta smelt dating back to 1995 and samples collected from 2020. The status of each deliverable is listed in Table 1. We have divided this final report into 4 sections, each with its own background, methods, results and discussion, followed by a final conclusion integrating our findings.

# Task 1: Genome assembly

* Background
  + Box1: Linked-read sequencing methods
  + Box 2: Long-read sequencing methods
  + Box 3:Hi-C chromatin confirmation capture
* Sample collection
* DNA isolation
* Long-read library prep & sequencing
* Linked-read library prep & sequencing
* Hi-C chromatic confirmation capture prep & sequencing
* Quality control box
* Long-read quality control
* Linked-read quality control
* Hi-C chromatic confirmation capture quality control
* Genome assembly
  + Box 4: Assembly statistics
* Cytogenetic (karyotype) analysis
* Genome assembly discussion

## Background

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 (e.g., analyses of population structure, genetic diversity, and local adaptation)8. Without a reference genome, DNA sequences 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 know where each segment of sequenced DNA lies within the genome and relative to other sequenced DNA. An assembled genome vastly increases the power to answer questions concerning demography, adaptation, fitness, and disease susceptibility.9,10 For example, previous work in Chinook salmon and steelhead found a small number of markers that were associated with run-timing in each species. In 2014, these markers were located on five different scaffolds and next to gaps in the highly fractured genome assembly. Reanalysis using long-reads to span gaps and reorient and link fragmented scaffolds demonstrated that all the associated markers were from a single locus (i.e., the GREB1L region)11. Since the 2017 study, continued improvements to salmonid reference genomes have enabled more detailed mapping of life-history variation and phenotypic traits in the same locus12, which may lead to significant changes in management. 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 a reliable way to achieve a chromosome-scale high-quality genome assembly13,14. 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 genomes14. Therefore, a main goal of this project is to develop a highly accurate chromosome-scale reference genome, hereafter called “reference genome”, using linked-reads (Box 1), long-reads (Box 2), hi-c chromatin confirmation capture (Box 3), a genetic linkage map15 and various assembly software programs (Figure 1).

**BOX 1: Linked-read sequencing methods**, (e.g., 10X Genomics) take large segments of extracted DNA (around 50kbp) from an organism’s genome and put 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 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 methods**, (e.g., PacBio HiFi) take long segments of extracted DNA (over 50kbp), shear them into a few segments (10-20kbp) and circularize each segment. 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 at which two segments of DNA are found linked together is inversely proportional to the distance between them 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/

## Sample collection (Figure 1A)

*Methods*

The first step in our genome assembly was to collect tissue for the extraction of high molecular weight (HMW) genomic DNA from both male and female fish. HMW DNA is extracted using a special protocol that produces longer DNA fragment lengths (>50 kilobases) compared to standard DNA extractions (<10 kilobases). HMW DNA is required for each of the three sequencing technologies (linked-read, long-read and hi-c) that we used to create our genome assembly. In order to examine potential sex-determining regions, we sequenced both male and female samples to create two sex-specific genome assemblies.

Though we had intended to make a single trip to the FCCL to collect delta smelt samples for sequencing, we found recovering sufficient HMW DNA difficult, and ended up conducting four separate trips to sample different tissue types from 600 days post hatch (dph) male and female delta smelt (Table 2) with the final goal of producing enough HMW DNA to sequence each sex with each of the three technologies. Additionally, at the start of this project long-read sequencing was costly and we sought to only incorporate the sequencing technology if absolutely necessary. Therefore, for our first and second sampling trips we 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. This development combined with our mixed results from Trips 1 and 2, prompted a third trip to acquire enough tissue to sequence a single male fish with all three of our chosen technologies, and enough tissue from a female fish to sequence with long-reads (as already we had enough HMW DNA for a female for hi-c and linked-reads). 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 for HMW DNA. 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 male fish produced sufficient quantities. Therefore we returned to the FCCL for a second trip and selected larger males. The extracted DNA lengths from Trip 2 males were also insufficient. A decision to take Trip 3 was made due to the discovery of a tissue sampling method that uses additional tissue types (not just back muscle tissue) and a new tissue preservation storage solution of cooled propylene glycol16 rather than flash freezing samples 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 4, we sampled back muscle, scales and internal organs from one male fish. All sampled tissue was flash frozen and stored on dry ice for transportation.

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

*Methods*

For 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)17. The size range of extracted DNA fragments were determined using a pulse field gel run for 24 hours. This run was conducted at a low frequency to not shear the DNA. Physical sampling and extractions were repeated until the mean distribution of extraction lengths was 50kbp 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). 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 from both males and females (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 where it was successfully extracted and sequenced.

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

*Methods*

The inclusion of long-read data dramatically assists in creating a more contiguous genome assembly because it spans repetitive elements and resolves chimeric sequences (erroneously joined artifact sequences) throughout the genome. This means that long-read data provides greater continuity of scaffolded **contigs** (a stretch of DNA sequence created from a consensus of reads). This is 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 a high error rate of up to 10%. During 2020, a new PacBio HiFi chemistry was developed which dramatically reduces the number of errors 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 fragment length of ~15kb. The input for the library prep was 5.6 ug of DNA and the library was size selected down to be 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). This amount of long-read sequencing data was sufficient to continue on to subsequent steps.

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

*Methods*

Once we had sufficient extracted HMW genomic DNA fragments, DNA was adjusted to a concentration of 0.91 ng/µl. We selected the 10X Genomics platform (<https://www.10xgenomics.com/technology/>) to generate our linked-read sequence data. 10X Genomics library preparation takes extracted HMW gDNA, shears it into 50kbp long segments, inserts each fragment into an oil coated gel emulsion bead (GEM), further shears DNA within each bead, and attaches one unique barcode 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-sequencing-based estimate of a haploid delta smelt genome size (0.6Gb) to sequence the first sample to an estimated 80x coverage.

Because we successfully extracted HMW DNA from a female individual first, we used the female linked-read data to improve our estimate of delta smelt genome size with a more accurate k-mer (where k is equal to a specified sequence length) based approach with the software Genomescope218. We then used the updated genome size estimate to adjust the amount of linked-read sequencing data collected for the male sample.

*Results*

The k-mer based haploid genome size from the female 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). This coverage was sufficient to continue on to subsequent assembly steps.

## 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 chromosome19. In order to resolve these gaps, we used hi-c chromatin conformation capture (hi-c). We outsourced the female hi-c library prep and sequencing to Phase Genomics in Seattle, WA in 2019. Phase used their in-house proprietary library preparation and sequencing protocols. Raw sequencing data and an initial scaffolding report were received for the female sample. Unfortunately, when we had male tissue available for hi-c sequencing the COVID-19 global pandemic response was underway. Due to the lack of availability of sequencing centers accepting new material, we outsourced the male hi-c sequencing to the Vertebrate Genome Project at Rockefeller. The data from these sequencing runs have yet to be returned to us at the GVL, and with no estimated return date. However, because hi-c links long-range interactions, we were able to use the female sequencing data for both male and female scaffolding.

*Results*

We received sequence files of 87,444,477 read pairs in total which is sufficient for resolving gaps and further scaffolding the linked and long-read combined assembly (Supplemental Data 4).

**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 were downloaded from Bioshare, the UC Davis Sequencing Center’s host service. We used PacBio’s CCS statistical model on raw reads (<https://github.com/PacificBiosciences/ccs>) to generate highly accurate consensus sequences with known base quality values and convert binary data to fastq format for downstream analysis. A base quality value assigns a score to each letter (nucleotide) 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 were sufficient to 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 KAT20. Each step splits sequencing data into sub-sequences of a given length, or k-mers, 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. 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 5kbp and 18.78% of the read pairs mapped to greater than 10 kilobases (kbp) apart. These data appear normal and indicate they will be useful in creating a more contiguous assembly.

## Genome assembly (Figure 1D-H)

*Methods*

We have broken this section into various steps for clarity and flow when describing the iterative process of assembling a eukaryotic genome.

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, <https://github.com/PacificBiosciences/pbipa>). 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 (<https://github.com/wtsi-hpag/Scaff10X>).

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 bwa21 and SAMtools22. We used the Arima Mapping pipeline perl scripts (<https://github.com/ArimaGenomics/mapping_pipeline>) 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 (<http://broadinstitute.github.io/picard/>). These data were then converted into mapped bed files using bedtools23. 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 pipeline24 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)15 with the output from the hi-c assembly step and the software chromonomer25.

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). BUSCO26 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.

## Genome assembly discussion

The diploid chromosome number of 56 for delta smelt aligns with that reported for other smelt species, 2n=54, 56 or 58 for European smelt27,28 and 2n=56 for the Japanese pond smelt29. As others have noted, Robertsonian fusions/fissions of chromosomes (acrocentrics fusing to form metacentrics or vice versa) may be the source of the karyotype variation, which is also the basis for karyotype variation observed within and among salmonid species28,30. No sex chromosomes have been reported to date for those smelts studied cytogenetically, and here we found no evidence for sex-specific chromosomes although a more detailed study is necessary to secure such a determination given the small sample size and low resolution of images. Similar to other reports, we note a preponderance of subtelocentric/acrocentric chromosome pairs over metacentric chromosome pair, as one might expect for closely related species. Chromosome composition is a descriptive metric, and our findings did not alter or affect our genome assembly process.

The primary objective of this study was to create a single highly contiguous genome assembly for use within and beyond the scope of this project. We assembled two independent delta smelt genomes (male assembly and female assembly) using gold-standard methods in addition to validating the number of chromosomes with an independent cytogenetic study. To assemble the delta smelt genomes, we combined PacBio long reads, 10X Chromium linked-reads, Phase hi-c chromatin conformation capture and a linkage map to create two sex-specific reference assemblies for male and female fish.

The final total lengths for the male and female assemblies were 0.47Gb and 0.44Gb, respectively which is similar to the wakasagi genome (*Hypomesus nipponensis*) which has a total length of 0.50Gb31. Our final male and female assemblies had 376 and 549 scaffolds with N50’s of 0.12Gb and 0.15Gb, respectively. The first 28 contigs, representing the number of haploid chromosomes confirmed by cytogenetic karyotyping contain 73.3% and 81.6% of the sequencing data of total assemblies in male and females, respectively. For comparison the 3.2Gb human genome––which is considered one of the leaders in genome assemblies and has been actively worked on for over a decade––has an N50 of 0.67Gb, has been assembled into 23 chromosomes and still contains 92 unplaced scaffolds for a total of 115 scaffolds. Thus, our reference genomes in a species that has a smaller genome with more chromosomes provides a highly successful foundation for the future of delta smelt research. Our final genome assemblies are roughly 25-30 times more contiguous that the newly published *H. niponnensis* assembly and our final assemblies contained 88.4% (in male reference) and 89.3% (in female reference) of core genes expected in the *Actinopterygii* BUSCO database.

In summation, our final reference genomes are highly contiguous and contain a high number of ultra-conserved genes. This indicates that this project was successful in assembling a useful resource for current and future studies involving delta smelt and closely related species.

# Task 2: *NE* Estimation

* Background
* Sample acquisition & estimation
* Read processing & alignment
* Principal component analysis for hybrid detection
* Contemporary temporal estimation
* Historical genetic diversity estimation
* discussion

## Background

The of a given population is defined as the size of a Wright-Fisher population that would have the same rate of change of a genetic parameter (i.e., genetic drift) as the population of interest32. A hypothetical Wright-Fisher population has a constant size, no migration, and random mating. Since wild populations inevitably violate these assumptions, is typically much lower than the true census size.

Several methods exist to estimate and examine which can be broadly separated into two categories: contemporary and historical :

**Contemporary** (temporal) describes the effective population size of recent generations as few as one generation back. We used a temporal method of estimating contemporary by examining genetic drift between two time points, which yields the harmonic mean of the years between the two time points. By looking at contemporary we can understand the extent of genetic drift occurring between two sample years, which tells us how quickly a population may have lost genetic diversity between two time points33,34. Contemporary/temporal methods are widely used and a reliable way of measuring in multi-generational datasets.35–40. They are often used in conservation genetic studies and management as they can detect population declines as rapidly as one generation post-population decline41 and can detect population declines ten generations post decline42. Here we estimated contemporary *NE* using more markers and a longer time frame than previous studies6,7.

**Historical** represents the effective population size of the historical population that led to the amount of genetic diversity currently observed in the modern-day population and can be derived from the equation where (theta) is a measure of genetic diversity and is the organism’s mutation rate. The mutation rate in delta smelt is unknown, but it is a constant rate, so by observing genetic diversity over time, we can observe the relative trend in over time. Historical *NE* can be used to understand population fluctuations going back further in time.

For isolated panmictic populations, the “50/500 rule” was a widely referenced rule proposed by Franklin in 1980 for conservation biology to establish parameters for a minimal viable population. The rule denotes that of a given population needs to be above 500 for long-term population persistence and cannot drop below 50 for short-term population persistence43–45. Above an of 500, variation is thought to be sufficiently replenished by new mutations46. In the past few decades, this rule has come under debate and more recent empirical modeling suggests a larger of 100 and 1,000 for short and long-term persistence, respectively47,48. However, even the notion of using empirical models to glean a threshold for conservation management has been rejected by some49,50.

In combination with information on various population parameters, is a useful tool for monitoring endangered populations because it gives us some insight into risk of extinction. However is by no means predictive of recovery potential49,51. So, although there have been several recommendations regarding how large needs to be in a “healthy” wild population, there is no way to predict how is affecting any one given population or how it relates to the census size; should not be taken as a threshold value for population persistence or recovery potential43,47,50. Here we take advantage of the highly contiguous female genome assembly to observe trends in genetic diversity and contemporary *NE* in the wild delta smelt population through time.

***Sample acquisition & sequencing***

*Methods*

We collected and sequenced archived samples obtained in state and federal trawls from 1993-2014. 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.*52. Sample years 1993-2017 and 2018-2020 were sequenced with 100 bp and 150 bp paired-end reads, respectively, on an Illumina HiSeq 4000.

*Results*

A total of 2,976 samples from 24 generations spanning 27 years were sequenced creating a rich dataset for further analysis.

## Read processing and alignment

*Methods*

In order to make demographic inferences, alleles spread across the genome of different individuals and contained within and across generations need to be compared. To do this, we split sequencing data into files corresponding to single individuals, using previously described perl scripts to separate raw sequencing data into plates, then individuals52. We then aligned the separated (or split) individual sequencing files to the assembled genome using bwa, which resulted in sequence alignment map (SAM) files. We then further processed the SAM files by sorting according to read name (samtools sort), filling in mate coordinates (samtools fixmate -m), removing duplicate reads (samtools markdup -r), and indexing the resulting files to create binary alignment map (BAM) files (samtools index) for downstream analyses.

## Principal component analysis for hybrid detection

*Methods*

When conducting population genetic analyses on a species, it is important to make sure that erroneous individuals are excluded. Delta smelt have been observed to hybridize with wakasagi smelt (*Hypomesus nipponensis*) in the San Francisco estuary4. 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 ANGSD53 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 R54, calculated the observed variance for PC1 and PC2, and visualized the first and second principal components (PC) and removed outlier individuals.

*Results*

The first principal component (PC1) showed 30.7% variance across the x-axis while the principal component 2 (PC2) had 24.5% variance across the y-axis. These large proportions of variance indicate the presence of hybrids or technical artifacts from sequencings. We identified and removed a total of 19 individuals with a PC1 > 0.01 and/or PC2 < -0.01 from downstream analyses (Table 6, Figure 7) as these individuals are likely hybrids or the result of a technical artifact of sequencing.

## Contemporary temporal NE estimation

*Methods*

In order to standardize the number of gene copies contributing to the estimation of , we subsampled to 50 individuals from each birth 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. 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), found 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 package55. 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 also used the reference assembly to select SNPs that were not within 1Mbp of each other to reduce bias in from linkage.

We made one-generation and all-by-all temporal method estimates of using three different estimators: Nei & Tajima56, Pollak57, and Jorde & Ryman58. Temporal estimations give the harmonic mean 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. Thus, one-generation temporal method estimates of 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 NeEstimator59 to make estimates in sequential years with sequencing data. Next, to get a broad understanding of how contemporary is changing through time, we performed all-by-all generations 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.

*Results*

Birth years 1993, 2000, 2007 were all removed from subsequent analyses due to having an insufficient number of individuals (Table 7). 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 birth year estimates were made from 1995 to 1999 and from 2008 to 2019 due to the availability of samples. All temporal-method estimations from 1995 to 1997, 1999, and 2008 to 2010 show an above 5,000 or infinite. For birth year 1998, the Nei & Tajima, and Pollak estimators show below 1,000, however, the Jorde & Ryman method estimates to be above 5,000. More recent estimates show consistently lower with estimations from 2011 to 2017, and 2019 having an of 4,000 or less for all temporal-method estimators. Additionally, most estimates within these years are below 1,500 with multiple years lower than 150, although birth year 2018 shows the Nei and Tajima estimation of to be above 5,000 but both the Jorde & Ryman, and Pollak estimators infer an below 2,500 (Figure 8).

All-by-all generation estimates were made, and data was plotted as a point located at the mean generation between and (Figure 9). Similar to the one-generation experiment, estimates made further back in time are more consistently infinite or above 10,000. The more recent estimations, notably from 2013 to 2020, are both decreasing and below 2,500 with many comparisons below 100 (Supplemental Table 1).

## Long-term genetic diversity estimation

*Methods*

Since the value of is directly related to genetic diversity through the equation , where (theta) is a measure of genetic diversity and is the organism’s mutation rate, we wanted to see if the size of translated to genetic diversity loss. To quantify genetic diversity through time we used two different estimates of : 1) the normalized number of segregating sites (Watterson, )60; and 2) the average pairwise nucleotide differences between gene copies (Pi, )61.

To make estimates of genetic diversity ( estimates), we obtained a global estimate of the folded site frequency spectrum. First, we acquired site allele frequency likelihoods from the reference genome using ANGSD (angsd -GL 2 -doSaf 1 -minMapQ 10 -minQ 20). Next, we estimated the maximum likelihood of the site frequency spectrum (SFS) in realSFS (realSFS -maxIter 100 -fold 1). Finally, we calculated for each site by using the realSFS (realSFS saf2theta) and converted those data into logscale per-site with thetaStat (thetaStat print).

*Results*

Genetic diversity measures significantly declined from 1995 to 2020 (p = 7.77e-7 and 1.17e-6 for and , respectively). Per-base estimates of range from 2.55e-4 to 7.12e-5 and estimates of range from 1.50e-4 to 6.13e-5 (Figure 10).

## NE Discussion

Two previous studies estimated contemporary in the wild population of delta smelt using 12-15 microsatellite markers. Fisch *et al.* 20117 found the to be decreasing during the 2003 to 2009 study period7. In contrast, Finger *et al.* 20176 found the could not be accurately estimated from 2011 to 2014 due to infinite confidence intervals. Multiple factors may have contributed to this discrepancy: 1) different versions of NeEstimator59,62 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.

In this study we removed non-neutral loci in order to make unbiased contemporary/temporal estimates. Our results show a broad decline of delta smelt *NE* taking place between 1995 and 2020. In general, estimates made with birth year cohorts from the 1990’s and early 2000’s had higher variance and were more frequently infinite, negative (indicating is too large to accurately be estimated), or greater than 5,000. More recent estimates starting around birth year 2013 show less variation, narrower confidence intervals and lower estimates of (Figure 8 and 9). Though the methods were different, compared to Finger *et al.* (2017) we observed lower estimates of contemporary from 2011 to 2013 (Table 8).

Our genetic diversity estimates corroborated our temporal estimates. We quantified genetic diversity through two statistics––average pairwise nucleotide difference (pi, ) and the normalized number of segregating sites (Watterson, ). We found both parameters to be significantly declining within the surveyed generations with pi to be slightly more stable (Figure 10). This stronger decrease in Watterson’s theta is expected as low frequency alleles are more rapidly lost as populations decline, which has a greater effect on Watterson’s theta because it leads to fewer segregating sites. In both statistics, we see similar trends in our estimates––large decreases in genetic diversity followed by periods of slight stability followed by continued loss of genetic diversity over the past 25 generations (Figure 10). Thus, our diversity estimates are consistent with our *NE* estimates that show an overall decline over the past quarter century.

# Task 3 Examining domestication selection in the conservation hatchery for delta smelt

* Background
* Experimental design
* Summary of prior work
* Methods
* Domestication results and discussion
* Next steps

## Background

Delta smelt have been cultured in a conservation hatchery system at the Fish Conservation and Culture Lab at UC Davis (FCCL). The FCCL uses a pedigree-based breeding system that applies parentage analysis and the addition of wild fish to keep the hatchery fish genetically similar to the wild population63,64. The pedigree-based management applied at the FCCL aims to both minimize average co-ancestry (i.e., kinship) in the refuge population and maintain their genetic diversity. Despite the intense management at the FCCL, Finger *et al.*(2018)65 showed strong evidence of genetic adaptation to captivity in the hatchery population. For example, offspring from crosses with two cultured parents with high levels of hatchery ancestry (e.g., a high domestication index; DI) are statistically more likely to survive to adulthood than those with wild parents or cultured-born parents with low hatchery ancestry (low DI), and this trend has increased over time. At present, the number of wild fish that are available to incorporate as broodstock in the FCCL refuge population is nearly zero, and without this wild input the rate of domestication will increase. Understanding the mechanism behind hatchery adaptation will allow for the FCCL to adjust their protocol to minimize the rate of hatchery adaptation, and thus improve the likelihood that future wild population supplementation will be successful.

We hypothesized: 1) there are heritable genetic changes associated with domestication selection in the FCCL; 2) domestication adaptation happens during the first few generations in the hatchery; and 3) the locus or loci under domestication selection contribute to the reproductive success of delta smelt in the hatchery.

## Experimental design

At the UC Davis Genomic Variation Lab, we accessed archived FCCL delta smelt samples (fin clips) from eight generations from 2008 to 2015. We grouped 240 individuals (not related at least at F1) for sequencing based on their domestication index (DI) and offspring survival rate (Figure 11). DI is calculated based on the number of generations that an individual's genome spent in captivity and is equal to the average of parents DI plus 1 (because another year has passed in captivity). Based on this, smelt are grouped into low (0 < DI ≤ 2), medium (3 ≤ DI ≤ 4), and high (DI≥7) DI groups. We created this study design in part to examine how quickly adaptation is happening in captivity. At the FCCL, the rate of survival of offspring from a family (offspring survival rate of a family) serves as a proxy for fitness in captivity and is calculated based on the total number of an individual offspring from a family that survive to be tagged at adulthood. The relative reproductive success was normalized by taking the ratio of each individual’s offspring survival rate to the mean of that in the corresponding generation and grouped as low (≤ 0.3) and high (≥ 0.7).

## Summary of prior work

Previously, we reported some results based on the incomplete genome assembly. To do that, we aligned and conducted population genetic analyses from the four groups of delta smelt with varying DI level: wild, low, medium, and high. RAD sequencing by *PstI* enzyme was used based on the protocol described in Ali *et al.*52 to collect genomic data. In order to detect outlier SNPs associated with domestication selection, genome-wide SNP FST (FST calculated per each SNP across genome) was calculated between wild and the different DI groups, FST is a measure of genetic differences between groups.

We first produced a RAD sequencing derived partial assembly (RAD assembly) and reported some outlier SNPs (90 SNPs) as candidate loci associated with domestication selection. Next, we produced a linked-read draft assembly that was orders of magnitude more contiguous than the previous RAD sequencing derived assembly. We found several FST peaks across the genome as candidate loci under domestication selection. Although with the increase in the reference genome contiguity, candidate loci appeared as elevated peaks, we were still observing multiple peaks across the genome. Because the draft assembly was not yet at chromosome level contiguity, we argued that with the chromosome-level reference genome, we may find that the peaks currently located on different scaffolds may belong to the same region. If the peaks located on different scaffolds come from one region, it would signify domestication at the hatchery is a monogenic trait (a trait that is determined by alleles of a single gene) under selection. If the FST peaks continue to be distributed at different locations spread across the genome, it would signify that domestication is a polygenic trait (a trait that is caused by effects of many genes) and multiple genes or loci throughout the genome are under selection.

Since then, we have produced a high-quality genome assembly to explore genomic signature of domestication selection and whether it is a monogenic or polygenic trait. The following is our updated analysis and results, and interpretation.

## Methods

First, we aligned sequences from our four groups to the reference genome using BWA software. The output files from the alignments were SAM files, which were then converted to BAM files using SAMtools22. SAMtools was then used to sort reads by name, remove reads that did not have a paired end, remove PCR duplicates, and index the BAM files. After aligning to the draft assembly, we began our analyses. All population genetic analyses were conducted using ANGSD which analyzes raw RAD sequencing data based on a probabilistic framework in the form of genotype likelihoods (GL). For the analyses, we used ANGSD’s implementation of a SAMtools genotype likelihood model (-GL 1) with a minimum base quality of 20 (-minQ 20) and minimum mapping quality of 20 (-minMapQ 20).

We wanted to standardize aligned read counts across individuals to account for variability that could bias downstream analysis. To do this, we used SAMtools-view to examine alignment quality, and observed high variance in quality and number of aligned reads. Due to the high variability in the number of aligned reads we observed (~20,000 – ~15,000,000) within individuals from the four groups, we evaluated the distribution of aligned-read counts and genotype call counts for individuals in each of the four groups (Figure 12). We then used this distribution to choose an aligned-read count per individual ranging between 3,000,000bp to 6,000,000bp. Accordingly, individuals with aligned-read counts below our preferred range were removed, and individuals with aligned-read counts above our range had their reads subsampled using SAMtools -view.

As a result of the aligned-read count filtration, the number of individuals in each group was reduced (Table 9). This resulted in relatively low sample numbers in the low and high DI groups compared to medium and wild. Because substantial differences in group sample sizes will affect the significance in FST pairwise comparisons, we decided to group individuals and make two comparisons: 1) wild/low DI group vs. medium/high DI group, and 2) wild group vs. hatchery (low/medium/high DI) group. With the former, we expected to detect loci that changed gradually in captivity while with the latter we expected to detect loci that change rapidly in captivity.

Next, we performed a pairwise FST sliding window analysis to detect SNPs associated with domestication selection.Sliding window FST is the FST analysis performed interactively over a certain length along a sequence or chromosome66. Since our SNP FST analyses were very noisy, and also because when selection affects a single SNP it also affects region where it is located, we decided to conduct a sliding window analysis which examines SNP outliers within their genomic regions. To conduct the sliding window analysis, we used snpR package6, which is a software package designed for whole-genome analysis with SNP data. In snpR, we performed the FST sliding window function (calc\_smoothed\_averages) using 200kbp window length (sigma=200, step = 50) to find differentiated regions on the genome. We then applied the bootstrapping function (do\_bootstraps) to test for the significance of those regions (at p≤0.001).

## Domestication results

The resultsof our twopairwise FST sliding window analyses are shown in Figure 13. In both comparisons (1 and 2), there are highly and moderately elevated regions with significant FST windows. Those elevated regions have high genetic differences between the two groups in each comparison, and are candidate regions associated with domestication selection. In comparison 1, wild/low DI group vs. medium/high DI group, two elevated regions are significant: one on Chromosome 15 and one on Chromosome 22 (lg 15 and lg 22). In comparison 2, wild vs. low/medium/high DI group, there are multiple significant peaks on Chromosomes 4, 10, 13, 15, 16, 19, 22, 23, and 26. Interestingly, the two elevated regions on Chromosome 15 and Chromosome 22 are present in both comparisons which makes them more likely as candidate regions under the selection.

To further explore those candidate regions, we looked at the distribution of the windows across chromosomes 15 and 22. We observed that outlier windows clustered in one region on each chromosome (Figure 14).This strengthens the probability of the region being under selection (i.e. the existence of peak of windows around significant windows indicates regional selection). Moreover, the high resolution sliding window analyses in both comparisons (Figure 14a and 14b) show the same pattern of peak FST distribution, which is strong evidence that these regions are under selective pressure.

***Domestication discussion***

The mechanism(s) underlying domestication can be either or both genetic evolution or epigenetics (i.e., heritable changes not linked to the DNA sequence). Our analyses have provided strong evidence that heritable genetic changes (i.e., genetic evolution) are occurring in captivity. The fact that we identified at least two candidate regions on two different chromosomes that may be under domestication selection suggests that domestication is a polygenic trait (i.e., it is caused by more than one region in the genome). In addition, since each of the two candidate regions are present in both comparisons, it seems domestication selection is happening in the early generations. This is because our two candidate regions have elevated FST values both in comparison 1, which was designed to identify regions under domestication selection in early (wild and low ID) vs. late (med and high DI) generations, and comparison 2, which was designed to identify candidate regions differentiating wild vs. hatchery delta smelt (low, med and high DI). Comparison 2 also revealed other elevated regions besides lg 15 and 22 showing significant FST windows. Although these regions may potentially be under domestication selection, more investigation is required.

The management benefit of identifying regions under domestication selection is that it provides the foundation for future research to elucidate the genes and biological pathways involved in the process of adaptation to captivity. This knowledge can provide insight into, and thus the potential to modify, environmental conditions that are causing the domestication. For example, if we find that our candidate regions are involved in salinity regulation, the FCCL can modify the salinity condition in the hatchery. Or, if the genes are involved in immune response, fish density may be a factor that can be modified.

## Future research

For further research, we recommend exploring the genes and molecular pathways associated with the two candidate regions discovered in comparison 1 and we will do more investigation of the other regions discovered in comparison 2 that may potentially be under domestication selection. Using the reference genome, we will also run a genome-wide association study (GWAS) to assess association of our candidate loci with the recovery rate at the FCCL. If there is an association between domestication selection and recovery rate, we expect to observe the same loci as the domestication selection candidate loci.

# Task 4: Sex Marker

* Background
* DNA sampling & sequencing
* Genome wide association study
* Depth analysis
* K-mer analysis
* Sex marker discussion

## Background

Sex determination in fish is a highly variable trait67 and understanding its mechanisms is crucial for understanding both the biology of a species and for gaining insight into the evolution of sex chromosomes and genetic mechanisms underlying sex determination68. Fish represent the most diverse group of vertebrates on earth with over 30,000 described species69. With this diversity and constant exposure to variable environments comes a vast array of morphological, physiological, behavioral, developmental and sexual mechanisms70–73. In teleost fishes, sex determination can be genetic or environmental and varies even between closely related species67,74–76. Delta smelt are born the sex they will be throughout their entire life cycle and 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 determination75. 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 of which sacrifice the life of the fish or gametes. When sex is identified by expression, pressure is put on the abdomen of fish until eggs (in females) or running milt (in males) are excreted63. Because sexual identification of wild fish depends 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 survival77.

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 each fish through either dissection or gametic expression. DNA was extracted using the Qiagen DNEasy 96 Blood & Tissue Kit with a modification of elution in 100uL 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 more often than the *Sbf1* restriction enzyme, or roughly once every 4,100 base pairs, providing more coverage of the genome than Sbf1. RAD sequencing libraries were prepared at the GVL according to Ali *et al.* (2016)52 and sequenced at the UC Davis Sequencing Center with 150 bp paired-end reads on an Illumina HiSeq 4000 sequencer.

## Genome wide association study

*Methods*

We performed two rounds of a 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 in the program ANGSD53 using males or females as controls (0) and the opposite sex as cases (1). The goal for this GWAS was to find alleles associated with a particular sex. The association of a particular allele with sex category was reported as a 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 15). The two SNPs most associated with sex in delta smelt had LRT scores of 37.854854 and 35.802804 which correspond to p-values of 7.621e-10 and 2.183e-9, respectively. Despite being highly associated with sex the genotypes at these loci were not diagnostic of sex (Table 10).

## Depth analysis

*Methods*

If sex determination in delta smelt is caused by chromosomal differences, it would be expected that the heterogametic sex (e.g., XY) would have roughly half the sequencing depth (the number of sequences that cover a given locus) of the homogametic sex (e.g., 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 aforementioned 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 discarded all genomic 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. Finally, we totaled the difference of coverage between male and females and sorted the output to look for locations in the genome where one sex had high and consistent coverage and the other sex had low or no coverage.

*Results*

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

## K-mer analysis

*Methods*

In our k-mer (where k is equal to a specified sequence length) 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 used the software sourmash78 to create 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 discarded k-mers that were shared between male and females to leave only 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 one sex (i.e., sex chromosome). Next, we extracted contigs containing five or more k-mers, which corresponds to a contig length of roughly 5,000 bp. 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 its 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. We then extracted RAD alignment depth information from all of the locations where the putative Y sequencing data had also aligned using the software bedtools 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. 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 main distribution of female and male k-mers. We found 44 contigs with k-mer mean abundance above 5 in the male sequencing data that had zero abundance in the female sequencing data.

We mapped the putative Y data back to the male reference genome and found the reads mapped to multiple regions within the genome (Table 11). However, we did not find a significant difference in male versus female read depth at locations across the putative Y regions.

## Sex Marker Discussion

Our study thoroughly probed the genome and two RAD sequencing data sets in multiple ways to identify sex-specific markers, but we did not find diagnostic SNPs. This means that delta smelt may not have straightforward chromosomal sex-determination, though we cannot yet completely rule it out. While we did not find diagnostic SNPs, we did find paths forward for further analysis that may result in understanding delta smelt sex determination. For example, we found candidate loci via GWAS using our RAD sequencing dataset and also via k-mer analysis using the linked-read sequencing data generated for the genome assembly. The GWAS found two markers on Chromosome 5 that were significantly associated with sex but did not have alleles diagnostic of sex. Interestingly, k-mer analysis detected DNA sequences only found within the male individual’s linked-read sequencing––one or more of these loci could contain a sex determining region or SNPs diagnostic of sex. The post k-mer analysis depth analysis showed that the observed increase in male specific k-mers at roughly 50% abundance of the normally distributed peak of the female k-mer abundance is consistent with the male sequencing data potentially having heterogametic (male sex-specific) regions in its genome (such as the 50:50 ratio between Y chromosomes paired with X chromosomes in human males). However, we could not identify or confirm sex-specific markers with the RAD sequencing data generated for this project.

Additionally, many contigs containing male-specific k-mers were located on Chromosome 9. While there is a clear increase in associated SNPs on Chromosome 9, none met the significance threshold or were found to be diagnostic of sex. An additional important observation is that the k-mer analysis revealed male-specific linked-read sequencing data from an individual male aligned to multiple regions throughout the genome. This may indicate that sex determination in delta smelt is polygenic (determined by several genes or locations scattered throughout the genome rather than a single region as in the human Y chromosome) but further sequencing and analysis is needed to test this hypothesis.

Our work shows a need for further investigation using high-coverage whole-genome resequencing (WGS) data from a large cohort of male and female fish to more evenly survey the genome in hopes of identifying sex-specific markers. While RAD sequencing data provide an adequate distribution of discrete locations throughout the genome of individuals, high-coverage WGS data more comprehensively survey the entire genome of individuals, rather than just a fraction. Furthermore, including a large number of individuals (e.g., 500) in this analysis would provide more statistical power to detect loci with a modest effect on sex, as would be expected with polygenic sex determination.

# Conclusion

As the refuge population of delta smelt become increasingly important for the preservation of the species it is essential that genetic resources and sequencing data are available to inform population management and improve planning for supplementation. In this study, we utilized next- and third-generation sequencing technologies to generate two chromosomal-level genome assemblies. These assemblies will be invaluable for the preservation and management of delta smelt. To exemplify this point, we used the female assembly with multi-generational RAD sequencing data to make demographic inferences of effective population size, to observe genetic diversity through time, to observe the genetic effects of domestication selection, and we used both assemblies to interrogate the genome for sex-specific markers.

We showed that the contemporary/temporal *NE* and genetic diversity of delta smelt is decreasing through time and caution that no single point estimate of *NE* should be used to make management decisions. Instead, we urge managers to observe the overall trend in *NE* and genetic diversity estimates in delta smelt––one of decline.

Based on our sex marker results, we recommend high coverage WGS of a large cohort of male and female delta smelt from the conservation hatchery in order to further probe the genome to identify a potential sex-determining region or regions. For our domestication study, we identified two genomic regions that are associated with domestication index, and will use this information to probe the genome further for molecular pathways associated with time spent in the hatchery.

# Glossary & Acronyms

**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

**coverage** – (or depth) the number of sequences that cover a given locus

**depth** – (or coverage) the number of sequences that cover a given locus

**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

**k-mer** – a sequence of length k

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

**PBS** – phosphate-buffered saline

**QC** – quality control

**RAD sequencing** – restriction site associated DNA sequencing that produces genetic markers spread throughout the genome. Markers are located at each location that contains the cut site of the specific restriction enzyme used to shear DNA for sequencing.

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

**SNP** – single nucleotide polymorphism

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

# Software Versions

ANGSD=0.934

bedtools=2.29.2

bwa=0.7.17-r1188

BUSCO=4.0.6

ccs=3.3.0

chromonomer=1.13

Genomescope2=2.0

htslib=1.10.2

IPA=1.3.1

kat=2.4.2

picard=2.23.3

SALSA2=2.0

samtools=1.10

scaff10x=4.2

# Works Cited

1. Moyle, P. B., Brown, Larry R. ,. Durand, John R. ,. Hobbs, James A. Delta Smelt: LIfe History and Decline of a Once-Adundant Species in the San Francisco Estuary. *San Franc. Estuary Watershed Sci.* **14**, 1–40 (2016).

2. Moyle, P. B. *Inland Fishes of California*. 682 (2003). doi:10.1643/OT-03-100.

3. Fisch, K. M., Mahardja, B., Burton, R. S. & May, B. Hybridization between delta smelt and two other species within the family Osmeridae in the San Francisco Bay-Delta. *Conserv. Genet.* **15**, 489–494 (2014).

4. Benjamin, A. *et al.* Use of single nucleotide polymorphisms identifies backcrossing and species misidentifications among three San Francisco estuary osmerids Collection Sites for Unknown Fish Yolo Bypass sites Lower Sacramento River site. *Conserv. Genet.* **0**, 0–0 (2018).

5. Brandl, S. *et al.* Ten real-time PCR assays for detection of fish predation at the community level in the San Francisco Estuary–Delta. *Mol. Ecol. Resour.* **15**, 278–284 (2015).

6. Finger, A. J., Schumer, G., Benjamin, A. & Blankenship, S. Evaluation and Interpretation of Genetic Effective Population Size of Delta Smelt from 2011–2014. *San Franc. Estuary Watershed Sci.* **15**, (2017).

7. Fisch, K. M., Henderson, J. M., Burton, R. S. & May, B. Population genetics and conservation implications for the endangered delta smelt in the San Francisco Bay-Delta. *Conserv. Genet.* **12**, 1421–1434 (2011).

8. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nat. Rev. Genet.* (2010) doi:10.1038/nrg2844.

9. Steiner, C. C., Putnam, A. S., Hoeck, P. E. A. & Ryder, O. A. Conservation Genomics of Threatened Animal Species. *Annu. Rev. Anim. Biosci.* **1**, 261–281 (2013).

10. Waples, R. S. Making sense of genetic estimates of effective population size. *Mol. Ecol.* **25**, 4689–4691 (2016).

11. Prince, D. J. *et al.* The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *Sci. Adv.* **3**, (2017).

12. Narum, S. R., Genova, A. Di, Micheletti, S. J. & Maass, A. Genomic variation underlying complex life-history traits revealed by genome sequencing in Chinook salmon. *Proc. R. Soc. B Biol. Sci.* **285**, (2018).

13. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).

14. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).

15. Lew, R. M. *et al.* Using Next-Generation Sequencing to Assist a Conservation Hatchery: a Single-Nucleotide Polymorphism Panel for the Genetic Management of Endangered Delta Smelt. *Trans. Am. Fish. Soc.* **144**, 767–779 (2015).

16. Patrick, H. J. H., Chomič, A. & Armstrong, K. F. Cooled Propylene Glycol as a Pragmatic Choice for Preservation of DNA From Remote Field-Collected Diptera for Next-Generation Sequence Analysis. *J. Econ. Entomol.* **109**, 1469–1473 (2016).

17. Wasko, A. P., Martins, C., Oliveira, C. & Foresti, F. Non-destructive genetic sampling in fish. An improved method for DNA extraction from fish fins and scales. *Hereditas* **138**, 161–165 (2003).

18. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).

19. López-Flores, I. & Garrido-Ramos, M. A. The Repetitive DNA Content of Eukaryotic Genomes. in *Genome Dynamics* (ed. Garrido-Ramos, M. A.) vol. 7 1–28 (S. KARGER AG, 2012).

20. Mapleson, D., Accinelli, G. G., Kettleborough, G., Wright, J. & Clavijo, B. J. Sequence analysis KAT : a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. **33**, 574–576 (2017).

21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

23. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

24. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Comput. Biol.* **15**, e1007273 (2019).

25. Catchen, J., Amores, A. & Bassham, S. Chromonomer: A Tool Set for Repairing and Enhancing Assembled Genomes Through Integration of Genetic Maps and Conserved Synteny. *G3 GenesGenomesGenetics* **10**, 4115–4128 (2020).

26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

27. Nygren, A., Nilsson, B. & Jahnke, M. Cytological studies in the smelt (Osmerus eperlanus L.). *Hereditas* **67**, 283–286 (1971).

28. Ocalewicz, K. *et al.* Karyotype and chromosomal characteristics of Ag–NOR sites and 5S rDNA in European smelt, Osmerus eperlanus. *Genetica* **131**, 29–35 (2007).

29. Kitada, J., Tatewaki, R. & Tagawa, M. Chromosomes of the pond smelt Hypomesus transpacificus nipponensis. *Chrom Inf. Serv* **28**, 8–9 (1980).

30. Hartley, S. E. THE CHROMOSOMES OF SALMONID FISHES. *Biol. Rev.* **62**, 197–214 (1987).

31. Xuan, B. *et al.* Draft Genome of the Korean smelt Hypomesus nipponensis and its transcriptomic. *G3 GenesGenomesGenetics* 35 (2021) doi:10.1093/g3journal/jkab147.

32. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).

33. Frankham, R. Genetics and extinction. *Biol. Conserv.* **126**, 131–140 (2005).

34. Luikart, G., Ryman, N., Tallmon, D. A., Schwartz, M. K. & Allendorf, F. W. Estimation of census and effective population sizes: The increasing usefulness of DNA-based approaches. *Conserv. Genet.* **11**, 355–373 (2010).

35. Hill, W. G. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* **38**, 209 (1981).

36. Krimbas, C. B. & Tsakas, S. THE GENETICS OF DACUS OLEAE. V. CHANGES OF ESTERASE POLYMORPHISM IN A NATURAL POPULA- TION FOLLOWING INSECTICIDE CONTROL-SELECTION OR DRIFT? *Evolution* 454–460 (1971).

37. Watterson, G. A. On the Number of Segragating Sites in Genetic Models without Recombination. *Theor. Popul. Biol.* **276**, 256–276 (1975).

38. Caballero, A. Developments in the prediction of effective population size. *Heredity* **73**, 657–679 (1994).

39. Robertson, A. The interpretation of genotypic ratios in domestic animal populations. *Anim. Prod.* **7**, 319–324 (1965).

40. Nei, M. & Tajima, F. Genetic drift and estimation of effective population size. *Genetics* **98**, 625–640 (1981).

41. Antao, T., Pérez-Figueroa, A. & Luikart, G. Early detection of population declines: High power of genetic monitoring using effective population size estimators. *Evol. Appl.* **4**, 144–154 (2011).

42. Nunziata, S. O. & Weisrock, D. W. Estimation of contemporary effective population size and population declines using RAD sequence data. *Heredity* **120**, 196–207 (2018).

43. Franklin, I. R. EVOLUTIONARY CHANGE IN SMALL POPULATIONS. in *Conservation Biology - An evolutionary-ecological perspective* 135–149 (Sinauer Associates, U.S.A., 1980).

44. Allendorf, F. W., Luikart, G. & Aitken, S. N. *Conservation and the genetics of populations*. (Hoboken : John Wiley & Sons, 2013).

45. Hössjer, O., Laikre, L. & Ryman, N. Effective sizes and time to migration–drift equilibrium in geographically subdivided populations. *Theor. Popul. Biol.* **112**, 139–156 (2016).

46. Allendorf, F. W. & Ryman, N. The role of genetics in population viability analysis. in *Population Viability Analysis* vol. 1 50–85 (The Univeristy of Chicago Press, 2002).

47. Frankham, R., Bradshaw, C. J. A. & Brook, B. W. Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biol. Conserv.* **170**, 56–63 (2014).

48. Franklin, I. R. & Frankham, R. How large must populations be to retain evolutionary potential? *Anim. Conserv.* **1**, 69–73 (1998).

49. Jamieson, I. G. & Allendorf, F. W. How does the 50/500 rule apply to MVPs? *Cell* **27**, 7 (2012).

50. Ryman, N., Laikre, L. & Hössjer, O. Do estimates of contemporary effective population size tell us what we want to know? *Mol. Ecol.* **28**, 1904–1918 (2019).

51. Frankham, R. Genetics and extinction. *Biol. Conserv.* **126**, 131–140 (2005).

52. Ali, O. A. *et al.* Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics* **202**, 389–400 (2016).

53. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 1–13 (2014).

54. R Core Team. R: A language and environment for statistical computing. (2013).

55. Hemstrom, W. & Jones, M. *snpR: user friendly population genomics for SNP datasets with categorical metadata*. https://www.authorea.com/users/394171/articles/507615-snpr-user-friendly-population-genomics-for-snp-datasets-with-categorical-metadata?commit=919fb47b3a2e64b824457249aeef488d2c31099a (2021) doi:10.22541/au.161264719.94032617/v1.

56. Nei, M. & Tajima, F. Genetic drift and estimation of effective population size. *Genetics* **98**, 625–640 (1981).

57. Pollak, E. A New Method for Estimating the Effective Population Size from Allele Frequency Changes. *Genetics* **104**, 531–548 (1983).

58. Jorde, P. E. & Ryman, N. Unbiased estimator for genetic drift and effective population size. *Genetics* **177**, 927–935 (2007).

59. Do, C. *et al.* Ne ESTIMATOR v2 : re - implementation of software for the estimation of contemporary effective population size ( N e ) from genetic data. *Mol. Ecol. Resour.* **14**, 209–214 (2014).

60. Watterson, G. A. On the Number of Segragating Sites in Genetic Models without Recombination. *Theor. Popul. Biol.* **276**, 256–276 (1975).

61. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**, 5269–5273 (1979).

62. Peel, D., Ovenden, J. R. & Peel, S. L. NeEstimator: software for estimating effective population size (version 1.3). *Dep. Prim. Ind. Fish. Qld. Gov. Brisb.* (2004).

63. Lindberg, J. C. *et al.* Aquaculture methods for a genetically managed population of endangered delta smelt. *North Am. J. Aquac.* **75**, 186–196 (2013).

64. Fisch, K. M., Ivy, J. A., Burton, R. S. & May, B. Evaluating the Performance of Captive Breeding Techniques for Conservation Hatcheries: A Case Study of the Delta Smelt Captive Breeding Program. *J. Hered.* **104**, 92–104 (2012).

65. Finger, A. J. *et al.* A Conservation Hatchery Population of Delta Smelt Shows Evidence of Genetic Adaptation to Captivity After 9 Generations. *J. Hered.* **109**, 11 (2018).

66. Zhu, Z. *et al.* SWAV: a web-based visualization browser for sliding window analysis. *Sci. Rep.* **10**, 4 (2020).

67. Volff, J. N. Genome evolution and biodiversity in teleost fish. *Heredity* **94**, 280–294 (2005).

68. Mei, J. & Gui, J. F. Genetic basis and biotechnological manipulation of sexual dimorphism and sex determination in fish. *Sci. China Life Sci.* **58**, 124–136 (2015).

69. Nelson, J. S., Grande, T. C. & Wilson, M. V. H. *Fishes of the World*. (John Wiley & Sons, 2016).

70. Nagahama, Y. Molecular mechanisms of sex determination and gonadal sex differentiation in fish. *Fish Physiol. Biochem.* **31**, 105–109 (2005).

71. Baroiller, J.-F., Guiguen, Y. & Fostier, A. Endocrine and environmental aspects of sex differentiation in fish. *Cell. Mol. Life Sci.* **55**, 910–931 (1999).

72. Kikuchi, K. & Hamaguchi, S. Novel sex-determining genes in fish and sex chromosome evolution. *Dev. Dyn.* **242**, 339–353 (2013).

73. Nakamura, M., Kobayashi, T. & Chang, X. Gonadal sex differentiation in teleost fish. *J. Exp. Zool.* **281**, 362–372 (1998).

74. Conover, D. O. & Kynard, B. E. Environmental Sex Determination : Interaction of Temperature and Genotype in a Fish Environmental Sex Determinaffon : Interaction of Temperature and Genotype in a Fish. **213**, 577–579 (2013).

75. Devlin, R. H. & Nagahama, Y. Sex determination and sex differentiation in fish: An overview of genetic, physiological, and environmental influences. *Aquaculture* **208**, 191–364 (2002).

76. Mank, J. E. & Avise, J. C. Evolutionary diversity and turn-over of sex determination in teleost fishes. *Sex. Dev.* **3**, 60–67 (2009).

77. Martínez, P. *et al.* Genetic architecture of sex determination in fish: Applications to sex ratio control in aquaculture. *Front. Genet.* **5**, 1–13 (2014).

78. Titus Brown, C. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).