**Chromosome-scale genome assembly and investigation of *Hypomesus transpacificus* genome for estimates of effective population sizes and identification of sex-specific markers.**

By

SHANNON ERICA KENDAL JOSLIN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Dr. Andrea Schreier, Co-Chair

Dr. Michael R. Miller, Co-Chair

Dr. C. Titus Brown

Committee in Charge

2022

© 2022 Shannon Erica Kendal Joslin

© 2022 Shannon Erica Kendal Joslin

Preliminary Pages – Acknowledgements, Abstract, Table of Contents

Acknowledgements

tktkt.

Abstract

Delta smelt (*Hypomesus transpacificus* (delta smelt,McAllister, 1963)) is a federally Threatened and state endangered fish endemic to the San Francisco Estuary and Sacramento-San Joaquin Delta of North America (SFE). The species is a small, pelagic, mostly annually reproductive fish with freshwater resident, migratory, and semi-migratory life histories (Campbell et al., 2022; Hobbs et al., 2019). They have previously been thought of as an indicator species for the health of water quality in the SFE. The species has undergone a population collapse associated with drought and anthropogenic effects, and it is now believed stochastic processes may push the species to extinction (Fisch et al., 2011; Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, 2016). Meaningful conservation management of the species must encompass gaining a better understanding of the life history, ecology, and physiology of the species to effectively preserve biological components contributing to success in the wild. Because genetics, in combination with the environment, influence many aspects of individual and population level phenotypes, building a framework to better understand the species requires the development of genetic resources and monitoring of genetic diversity. Chapter one of this dissertation presents two chromosome-level genome assemblies –– one male and one female –– which are necessary resources for current and ongoing research within delta smelt, and evolutionary and conservation genetics. Chapter two investigates three novel methods for identifying sex marker(s) within the delta smelt genome. While no specific sequences diagnostic of sex were found, sex-specific discrepancies in kmer were identified. Chapter three uses restriction site-associated DNA sequencing (RADSeq) of \_\_tk\_\_ delta smelt from\_\_tk\_\_ generations to monitor contemporary and historical effective population size, a metric diagnostic of genetic diversity, through time. Results show a sweeping and significant decrease in genetic diversity for both measurements within wild delta smelt since 1993. This work contributes to the broad comparative biology and conservation genetic communities, and specifically to delta smelt monitoring, management, and research. To date, our assembled delta smelt genomes are the first to provide male and female specific assemblies, are the first chromosome-level and least fragmented publicly-available reference genomes within the *Osmeridae* (smelt) family. Examination of male and female sequencing data shows a discrete difference between sexes and establishes a framework for further investigation. Contemporary and historical declines in effective population size support the need for effective conservation management to prioritize preserving genetic diversity as a component of management within the cultivated refuge population. Together, these components provide necessary resources and research objectives for moving evolutionary biology and delta smelt management forward.

Table of Contents

[Preliminary Pages – Acknowledgements, Abstract, Table of Contents 3](#_Toc113273215)

[Acknowledgements 3](#_Toc113273216)

[Abstract 3](#_Toc113273217)

[Table of Contents 5](#_Toc113273218)

[Chapter 1 – Introduction 6](#_Toc113273219)

[Background of delta smelt & the SFE 6](#_Toc113273220)

[Chapter 2 – Genome assembly of Hypomesus transpacificus (delta smelt) 8](#_Toc113273221)

[Introduction 8](#_Toc113273222)

[Methods 11](#_Toc113273223)

[Sample collection & DNA extraction 11](#_Toc113273224)

[Linked-read library prep, sequencing & quality control 12](#_Toc113273225)

[Long-read library prep, sequencing & quality control 14](#_Toc113273226)

[Hi-C chromatin conformation capture library prep, sequencing & quality control 16](#_Toc113273227)

[Genome assembly 17](#_Toc113273228)

[Assembly quality assessment 18](#_Toc113273229)

[Cytogenic chromosome validation 19](#_Toc113273230)

[Results 20](#_Toc113273231)

[Sample collection & DNA extraction 20](#_Toc113273232)

[Linked-read library prep, sequencing & quality control 20](#_Toc113273233)

[Long-read library prep, sequencing & quality control 21](#_Toc113273234)

[Hi-C chromatin conformation capture library prep, sequencing & quality control 22](#_Toc113273235)

[Assembly quality assessment 22](#_Toc113273236)

[Cytogenic chromosome validation 23](#_Toc113273237)

[Discussion & Conclusion 24](#_Toc113273238)

[Works Cited 27](#_Toc113273239)

[Tables and Figures 35](#_Toc113273240)

Chapter 1 – Introduction

Background of delta smelt & the SFE

The SFE is a dynamic ecosystem encompassing 1,000 square miles of open water and wetlands in Northern California. Since the area became the primary distribution hub of California’s water supply in the 1960’s, the composition of SFE has been heavily altered by anthropogenic activity. Agriculture, water delivery, shipping, and urban development have spurred changes in the way water is distributed throughout the estuarine environment. State and federal agencies routinely monitor the relative abundance of fish in the SFE with initiatives like the California Department of Fish and Wildlife (CDFW) Fall Midwater Trawl Survey which began in 1967. Similar to many other estuarine ecosystems throughout the globe (Belarmino et al., 2021; Cottingham et al., n.d.; James et al., 2018), within the SFE many once abundant endemic pelagic fishes, such as delta smelt, have undergone broad declines in population size (Moyle, 2002; Moyle et al., 2018).

Delta smelt are a small (6 - 9cm), translucent, semi-anadromous species that migrates between fresh and saline water and reproduces annually (Sommer et al., 2011). Delta smelt are part of the Osmeridae family which represent a prosperous food source for human consumption in Japan, Europe, and North America, and have experienced declining populations worldwide (McAllister, 1963; Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, 2016; Rosenfield & Baxter, 2007). Because of their annual life cycle and relatively rapid response to the conditions of their habitat, delta smelt are considered an indicator of the overall health of the SFE ecosystem. Once one of the most abundant and widely distributed fish species in the SFE, delta smelt numbers have been declining since the 1980s (Moyle et al., 1992) (Figure 1.1). The species was listed as threatened under the Federal Endangered Species Act (ESA) in 1993 and endangered under the California ESA in 2009. Pelagic productivity and water temperature have been shown to be primary drivers of condition indices ultimately affecting the delta smelt’s fitness (Hammock et al., 2022) and it is now believed stochastic processes may push the species to extinction (Fisch et al., 2011; Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, 2016). As a result of their continued decline, resource management agencies, such as CDFW, continue to actively monitor the distribution and abundance of the wild population, and the Genomic Variation Laboratory genetically manages a genetically focused breeding program within a refuge population at the UC Davis Fish Conservation and Culture Laboratory (FCCL).

Chapter 2 – Genome assembly of *Hypomesus transpacificus* (delta smelt)

Introduction

From life-saving coagulation inhibitors isolated from bats, ticks, leeches, and hookworms to developing the theory of natural selection, understanding a diverse landscape of animal biology has led to scientific advances and innovation (Darwin, 1859; Gardell et al., 1991; Jacobs et al., 1990; Sawyer, 1991; Tanaka-Azevedo et al., 2010). Historically, molecular biology has focused on single genes to understand disease and fuel innovation. However, single gene studies fail to account for how genes work combination and coordination with other genes in a complex network of interactions within each organism’s genome. Just as many macroscopic observations led to innovation and understanding of the processes that affect organisms, such as the theory of evolution by natural selection, current molecular biology is in an age of compiling accurate microscopic observations for current and future scientific advancement unobtainable with previous technologies. Rapid development of high-throughput sequencing technologies over the past few decades has led to an era of genomic research, and much genomic research begins with reference genomes. Genomic resources contribute to two broad categories: medicine and biodiversity. Medicine has benefitted from genomic resources using comparative methods to identify conserved loci essential to life or different classes of organisms as well as identify genetic variants associated with disease, disease susceptibility and other phenotypic traits (Claussnitzer et al., 2020). Biodiversity relies on having genetically diverse organisms within and between species, as genetic diversity is related to the evolutionary capacity for adaptation to environmental change. Comparative methods allow for a broad investigation of genomic motifs in endangered animals (Zoonomia Consortium, 2020) and the identification of physical threats to endangered species (Wasser et al., 2015; Wright et al., 2015). In both categories, genomic studies involving reference genomes are limited by the completeness of the assembled resource. As such, the utilization of multiple sequencing technology types to carry out “hybrid” methods of genome assembly increases the capacity for capturing and assembling increasingly more of an organism’s actual genome.

Sequencing technologies have been in an era of rapid growth and next generation sequencing (NGS) and third generation sequencing (TGS) technologies have allowed for relatively easy to generate, low cost, high-throughput sequencing data. Of the numerous new methods to generate sequencing data for assembly, long-read sequencing, short-read sequencing, and interaction mapping have transformed the quality and contiguity of genome assemblies today. Current *de novo* genomes reach completeness standards which took decades for the first human genome to reach at a fraction of the cost and by relatively few individuals (International Human Genome Sequencing Consortium et al., 2001). A recent major publication to utilize and highlight current methods in the hybrid approach to *de novo* assembly was the domestic goat in 2017 (Bickhart et al., 2017) which used long-read sequencing, short-read sequencing, and interaction mapping to increase the previous goat genome’s contiguity by over two orders of magnitude and proposed it was the most continuous *de novo* mammalian assembly of its time. Since the publication of the *de novo* goat assembly, hybrid-assembly publications have become commonplace. As such, hybrid assembly is an accepted and reliable way to achieve a chromosome-scale high-quality reference genomes (Bickhart et al., 2017; Rhie et al., 2021). Since 2017, over half of all vertebrate chromosome-level assemblies submit to GenBank have implemented a hybrid assembly approach to genome assembly (Hotaling et al., 2021).

Each of the three technology types have different biases, errors, and uses. In general, hybrid assemblies use long, error-prone reads to generate scaffolds, correct base calling errors with short reads which have high accuracy but cannot span highly-repetitive sequences, and anchor scaffolds into chromosomes using interaction mapping which shows physical associations to span and link proximal scaffolds.

We used PacBio HiFi long-reads, 10X Genomics linked-reads, and Phase Genomics chromatin confirmation capture to assemble our reference-quality Delta smelt genome. In addition to these three sequencing technologies, we sought to further increase the contiguity of the genome by incorporating information from a previously published linkage map (Lew et al., 2015). Finally, we validated the number of expected chromosomes in the final delta smelt assembly by karyotyping the species.

Prior to this work, no endangered species and three highly fragmented genome assemblies from the Osmeridae(smelt) family were publicly available through GenBank. These assemblies came from three different genera (*Hypomesus, Thaleichthys*, and *Osmerus*) and the most closely related assembly publicly available was *Hypomesus nipponensis* (wakasagi smelt), a common species of smelt endemic to and used as a food commodity in Japan, introduced into the SFE and known to hybridize with delta smelt (Dill & Cordone, n.d.). While the wakasagi smelt genome, estimated to be 464 Mbp in size with 2n=26, was the most contiguous resource within the Smelt family, the assembly was not a chromosome-level assembly with an N50 of 0.46 Mbp and L50 of 477 (Table 2.1, Kitada et al., 1980; Xuan et al., 2021).

Methods

Sample collection & DNA extraction

To obtain high molecular weight (HMW) genomic DNA (gDNA), we collected tissue samples from male and female delta smelt reared within the refuge population at the FCCL at 600 days post hatch. Fish were euthanized according to IACUC protocol #21533. After euthanasia, we dissected the fish, sampled tissue including muscle, internal organs (heart, liver and spleen) and gills, immediately flash froze all tissue samples, and stored samples isolated or suspended in propylene glycol – two storage methods known to be conducive to HMW gDNA sequencing in different organisms (Patrick et al., 2016; Wasko et al., 2003).

HMW gDNA was isolated by the UC Davis DNA Technologies & Expression Analysis Core Laboratory (UC Davis Sequencing Center) following the protocol as described by Wasko et al.(Wasko et al., 2003). Briefly, ~25-50mg of flash frozen back muscle tissue and scales from a male and female individual were homogenized using liquid nitrogen grinding. Tissue was lysed in a buffer containing 10 mM Tris-HCl pH 8.0, 125 mM NaCl, 10 mM EDTA pH 8.0, 0.5 %SDS, 4M urea and 10mg/ml Proteinase K. The lysate was cleaned with equal volumes of phenol/chloroform using phase lock gels (Quantabio Cat # 2302830). The DNA was precipitated by adding NaCl to the final concentration of 0.3M and 2X volume of ice-cold ethanol. The DNA pellet was twice washed with 70% ethanol and resuspended in an elution buffer (10mM Tris, pH 8.0). The integrity of the high-molecular-weight DNA was verified on a Pippin Pulse gel electrophoresis system (Sage Sciences, Beverly, MA). Purity of the DNA was accessed by measuring 260/280 and 260/230 absorbance ratios on NanoDrop. Libraries of relevant sequencing technologies were created from extractions with an average read fragment length of 50kb with strong fluorescence.

Linked-read library prep, sequencing & quality control

Genomic DNA was adjusted to a concentration of 0.91 ng/µl and 1.14 ng of template gDNA was loaded on a Chromium Genome Chip. 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 with one modification. Briefly, gDNA was combined with Master Mix, a library of Genome Gel Beads, and partitioning oil to create Gel Bead-in-Emulsions (GEMs) on a Chromium Genome Chip. The GEMs were isothermally amplified with primers containing an Illumina Read 1 sequencing primer, a unique 16-bp 10X barcode and a 6-bp random primer sequence, and barcoded DNA fragments were recovered for Illumina library construction. The amount and fragment size of post-GEM DNA was quantified by running 1µl of sample on a Bioanalyzer 2100 with an Agilent High sensitivity DNA kit (Agilent, cat. 5067-4626). Prior to Illumina library construction, the GEM amplification product was sheared on an E220 Focused Ultrasonicator (Covaris, Woburn, MA) to approximately 350bp (50 seconds at peak power = 175, duty factor = 10, and cycle/burst = 200). Then, the sheared GEMs were converted to a sequencing library following the 10X standard operating procedure.

The sequencing library was quantified by qPCR with a Kapa Library Quant kit (Kapa Biosystems-Roche) and sequenced on NovaSeq6000 sequencer (Illumina, San Diego, CA) with paired-end 150 bp reads. We used a previous inhouse RAD-sequencing-based estimate of a haploid delta smelt genome size of 0.6 Gb to sequence the first sample to an estimated 80x coverage. Because we successfully extracted HMW gDNA from a female first, we used the female linked-read sequencing data to improve our estimate of delta smelt genome size through a more accurate k-mer based approach using Genomescope2 (Vurture et al., 2017). After, we used the updated genome size estimate to adjust the amount of all subsequent sequencing data generated for assembly.

We evaluated linked-read sequencing data by looking at base quality metrics, and for signs of contamination and sequencing bias. For basic metric quality assessment, we used FastQC v0.11.9 (Andrews, Simon, 2010) to obtain raw sequencing data metrics such as per sequence quality scores, GC content, total number of reads, average read length and number of bases. To quality control linked-read files for contamination and sequencing bias errors, we conducted three computational quality control steps (kat hist, kat gcp, and kat comp) using the software program KAT (Mapleson et al., 2017). Each step splits sequencing data into sub-sequences of a given length (k-mers) to plot out frequencies or comparisons to visually inspect the data for quality issues. We looked for signs of bacterial and organelle contamination using kat hist and kat gcp within the male and female sequencing data. First, we used kat hist to plot a histogram of the observed number of distinct k-mers at different frequencies for lengths k=21 and 31. Second, we used kat gcp to plot the proportion of the k-mers comprised of the G and C nucleotides against the frequency of that k-mer in the sequencing data and the number of distinct k-mers for a given GC count vs. frequency. We plotted GC counts against the frequency of k-mers of length k=21 and 31. Lastly, we evaluated the data for sequencing bias between the forward (R1) and reverse (R2) sequencing data. We used kat comp to plot the frequency of a given k-mer in each of the paired-end sequence data files (R1 and R2) for k-mers of length k=21 and 31.

Long-read library prep, sequencing & quality control

Genomic DNA was sheared to roughly 17kb using Diagenode's Megaruptor's (Diagenode, cat B06010001) long hydropores (Diagenode hydropores, cat E07010002). Sheared DNA was quantified by Quantus Fluorometer (Promega, cat #E6150) using a QuantiFluor® ONE dsDNA Dye assay (Promega, cat #E4871) and size distribution was checked by Agilent Femto Pulse (Agilent Technologies, cat P-0003-0817). Sheared gDNA was then concentrated using AMPure PB beads (Pacific Biosciences, cat 100-265-900). Concentrated, sheared gDNA was quantified by Quantus Fluorometer (Promega, cat #E6150) using a QuantiFluor® ONE dsDNA Dye assay (Promega, cat #E4871). Low- or high-input PacBio HiFi library construction was carried out based on the amount of available concentrated sheared gDNA present for each sample.

High-input HiFi libraries were constructed using the SMRTbell® Express Template Prep Kit v2.0 (Pacific Biosciences, cat #100-938-900) with protocol “Procedure & Checklist - Preparing HiFi SMRTbell® libraries using SMRTbell® Express Template Prep Kit 2.0 v3, January 2020”. We used sheared DNA as input for removal of single-strand overhangs at 37°C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37°C for 30 minutes, End Repair and A-tailing at 20°C for 10 minutes and 65°C for 30 minutes, ligation of overhang adapter v3 at 20°C for 1 hour and 65°C for 10 minutes, and nuclease treatment of SMRTbell® library at 37°C for 1 hour to remove damaged or non-intact SMRTbell® templates (SMRTbell® Enzyme Cleanup Kit, Pacific Biosciences, cat #107-746-400). The resulting SMRTbell® libraries were purified and concentrated by 0.45X AMPure PB beads (Pacific Biosciences, cat #100-265-900) then pooled for size selection using the SageELF system (Sage Science, cat #ELF0001). Input of the purified SMRTbell® library was used to load into the SageELF 0.75% Agarose Cassette (Sage Science, cat ELD7510) using cassette definition 0.75% 1-18 kb v2 for the run protocol. Fragments roughly 16 kb to 18 kb were collected from elution wells and the size-selected SMRTbell® library was purified and concentrated with 0.5X AMPure beads (Pacific Biosciences, cat 100-265-900).

Low-input HiFi libraries were constructed using the SMRTbell® Express Template Prep Kit v2.0 (Pacific Biosciences, cat #100-938-900) with protocol “Procedure & Checklist - Preparing HiFi SMRTbell® libraries from Low DNA Input using SMRTbell® Express Template Prep Kit 2.0 v6, June 2020”. We used sheared DNA as input for removal of single-strand overhangs at 37°C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37°C for 30 minutes, End Repair and A-tailing at 20°C for 10 minutes and 65°C for 30 minutes, ligation of overhang adapters v3 at 20°C for 1 hour. Low Input HiFi SMRTbell® library was purified and concentrated twice first by 1.8X AMPure PB beads (Pacific Biosciences, cat #100-265-900) and 40% diluted AMPure beads to remove < 3 kb SMRTbell® templates. Both high and low-input libraries were subsequently loaded onto 8M SMRT Cells and sequenced using a Sequel II sequencing plate 2.0 on Pacific Biosciences Sequel II sequencer.

We used PacBio’s CCS v3.3.0 statistical model on raw reads (<https://github.com/PacificBiosciences/ccs>) to generate base quality called circular consensus (ccs) reads and convert binary data to fastq format for downstream analysis. Reads with quality scores over Q20, denoting an error probability of 0.01% or less, were accepted and used for subsequent assembly.

Hi-C chromatin conformation capture library prep, sequencing & quality control

Female chromatin conformation capture sequencing data was generated by Phase Genomics (Seattle, WA) using Proximo Hi-C 2.0 Kit, a commercially available version of the Hi-C protocol, in 2019. Following the manufacturer's instructions for the kit, intact cells from two samples were crosslinked using a formaldehyde solution, digested using the SAUIII restriction enzyme (cut site GATC), end repaired with biotinylated nucleotides, and proximity ligated to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal in vivo, but not necessarily genomically proximal. Continuing with the manufacturer's protocol, molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Finally, 150bp paired-end reads were generated on an Illumina HiSeq sequencer. Raw sequencing data and an initial scaffolding report were received for the female sample. Due to the lack of availability of sequencing centers accepting new material during the COVID-19 pandemic, we outsourced the male hi-c sequencing to the Vertebrate Genome Project (VGP) at Rockefeller. The data from these sequencing runs had yet to be returned upon the completion of the research for this dissertation. Because hi-c links long-range interactions, we used the female hi-c sequencing data for both male and female scaffolding.

We evaluated hi-c sequencing data by looking at base quality and mapping metrics. For basic metric quality assessment, we used FastQC v0.11.9 (Andrews, Simon, 2010) to obtain raw sequencing data metrics such as per sequence quality scores, GC content, total number of reads, average read length and number of bases. To assess if the hi-c 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 prior step. We also observed the number of reads which aligned to each contig (>600 desired) and the number of high-quality reads greater than 10Kbp apart (1-15% expected).

Genome assembly

We generated an initial draft assembly (A1) purged of duplicate haplotigs using the IPA HiFi Genome Assembler (ipa) v1.3.1 (https://github.com/PacificBiosciences/pbipa), with purge\_dups v1.2.3 (Guan et al., 2020) and Racon v1.4.13 (Vaser et al., 2017) wrappers enabled to generate phased primary and alternative assembly files polished of errors. To scaffold the initial A1 assembly using linked-read data, we used scaff10x (Ning, n.d.) with the following parameters: -longread 1 -gap 100 -matrix 2000 -reads 10 -link 8 -score 20 -edge 50000 -block 50000 which first breaks the assembly at locations that were incorrectly joined, scaffolds the assembly into larger, more contiguous sequencing segments composing the linked- and long-read (A2) draft assembly. After linked-read scaffolding, we followed the Arima mapping protocol (https://github.com/ArimaGenomics/mapping\_pipeline) to prepare the hi-c data to use interaction mapping information to further scaffold the A2 draft assembly. To prep the sequencing data, we independently aligned paired-end hi-c reads as single-ended reads to the A2 assembly using BWA v0.7.17-r1188 (Li & Durbin, 2009) and samtools v1.7 (Li et al., 2009). Next, we retained the 5’ end of the read to eliminate chimeric reads using a custom Arima perl script. Then, we paired the hi-c reads to produce paired-end BAM files and used PicardCommandTools (https://github.com/broadinstitute/picard) to add read groups and remove PCR duplicates. After filtering the hi-c sequencing data, we converted BAM files to sorted BED files with bedtools v2.29.2 (Quinlan & Hall, 2010). The A2 draft assembly and BED files were input into SALSA2 (Ghurye et al., 2019) with non-default parameters (-i 5 -x GATC -m yes) to scaffold the A2 assembly with the filtered hi-c data to produce a linked, long, and hi-c read (A3) assembly. Finally, to anchor the A3 assembly into chromosomes-scale scaffolds we used chromonomer v1.13 (Catchen et al., 2020) in combination with a previously published delta smelt linkage map (Lew et al., 2015) to produce a chromosome-level reference genome (A4) assembly.

Assembly quality assessment

After each step generated a draft assembly (A1 - A4, we evaluated the contiguity, content, and composition of the resulting fasta file. To assess each assembly’s completeness, we used the evolutionarily informed Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.6 (Simão et al., 2015) Actinopterygii lineage (actinopterygii\_odb10) dataset in genome mode.

To acquire assembly metrics, such as N50, L50, number of contigs, and assembly length, we used GenomeTools v1.5.10 (Gremme et al., 2013). Read length histograms were generated using jellyfish (Marçais & Kingsford, 2011). All assemblies within each sex, and between sexes were compared using QUAST v5.2.0 (Gurevich et al., 2013).

Cytogenic chromosome validation

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

Sample collection & DNA extraction

After mixed results in length and quantity yield of HMW gDNA from back muscle tissue flash frozen and stored unsuspended in liquid, we expanded our sampling and storage methods through the additional collection of scale and internal organ tissue, and by storing samples of back muscle tissue in propylene glycol. However, we did not find that suspending flash frozen back muscle in propylene glycol provided more success in the yield of HMW gDNA. We used tissue samples from two female individuals and two male individuals to extract roughly 3.4µg of HMW gDNA at a concentration of 87ng/µL for subsequent sequencing (Figure 2.1). Back muscle tissue from one female individual (T1F02\_BM\_FF) was used for linked-read, long-read, and hi-c sequencing and gill tissue from one female (T3F02\_SC\_FF) was used for further PacBio sequencing. One male individual (T3M02\_BM\_FF) was used for linked and long read sequencing, and internal organ tissue from the second male (T4M01\_IO\_FF) was used for generating hi-c sequencing data (Table 2.2). Suspending tissues in propylene glycol did not increase the success rate or yield of extracted HMW gDNA in our experience, and all successful extractions were not suspended in any kind of solution after flash freezing. NanoDrop 260/280 absorbance ratios of male and female extractions were 1.91 and 1.90, respectively. NanoDrop 260/230 ratios for male and female extractions were 2.02 and 1.79, respectively.

Linked-read library prep, sequencing & quality control

Post-GEM DNA quantified library electropherograms showed expected distributions (Figure 2.2). We generated 94,825,601,818 bp of paired-end linked-read sequencing data from the female specimen. Using the Genomescope2 k-mer based haploid genome size estimation of the female 10X sequencing data we estimated the delta smelt genome size to be 0.49 Gb. We used the updated k-mer based estimate to reduce the amount of male data generated in sequencing. In total, we generated 94,825,601,818 bp and 65,806,680,934 bp of female and male linked-read sequencing data, respectively. The average per sequence base quality was 33 and 32 in the female R1 and R2 fastq files, respectively and 34 and 32 in the male R1 and R2 fastq files, respectively (Table 2.3).

Mapped k-mer histograms for each sample and at each value of k showed discrete, single peaks indicating no sign of contamination (Figure 2.3). All GC count frequency plots show roughly normal circular distributions of distinct k-mers with no aberrant spotting (Figure 2.4). Additionally, the number of distinct k-mers does not appear to be heavily skewed, indicating no sequencing bias, in the male or female sequencing (Figure 2.5). These data together indicate no observable signs of bacterial or organelle contamination or major sources of sequencing bias in the linked-read sequencing data.

Long-read library prep, sequencing & quality control

We constructed one high-input library from the male HMW gDNA extractions. To reach sufficient coverage sequencing data with the available HMW gDNA from the female individual, we constructed one high-input and one low-input library. Starting gDNA inputs ranged from 6.5ug to 20ug of gDNA. The sheared gDNA input for the removal of single strand overhangs ranged from 1000ng to 7ug, and the average length of gDNA for sequencing ranged from 14-18.4kb.

Five movie collections (150 hours of sequencing data) from two male and one female high-input library, and two low-input female library runs were collected. 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 used for subsequent assembly (Table 2.3).

Hi-C chromatin conformation capture library prep, sequencing & quality control

Hi-C sequencing files contained 87,444,477 read pairs in total (Table 2.3). The data contained an average of 2,966.33 read pairs per contig greater than 5kbp, 18.78% of the read pairs mapped to greater than 10 kilobases (kbp) apart and 56.38% of reads were considered high quality indicating successful library prep and sequencing. The average per sequence base quality was 38 and 36 in the R1 and R2 fastq files, respectively (Table 2.3).

Assembly quality assessment

We searched raw data and each iteration of the assemblies for 3,640 conserved single-copy orthologs contained within the 05 August 2020 Actinopterygii lineage dataset using BUSCO. The quality filtered female and male HiFi data contained whole genes or sequence fragments of 95.6% (3.3% complete single copy, 89.3% complete double copy, and 3.0% fragmented) and 94.4% (3.4% complete single copy, 87.0% complete double copy, and 4.0% fragmented) of the conserved Actinopterygii gene dataset were present in the quality filtered female and male HiFi data, respectively (Table 2.4).

After each step, the total length and N50 increased, and the L50 and total number of contigs decreased (Figure 2.6 & 2.7). Female HiFi sequencing data had an N50 of 15,048 and an L50 of 771,808, while the male HiFi data had an N50 of 11,604 and an L50 of 1,276,120. The final female assembly contained 89.3% complete (87.7% single copy and 1.6% double copy) genes and fragments of an additional 0.8% of conserved genes, had an N50 of 14,850,352, L50 of 13, and was a total of 437,273,953 bp long with a total of 376 contigs. The final male assembly contained 88.4% complete (81.2% single copy and 7.2% double copy) genes and fragments of an additional 1.0% of conserved genes, had an N50 of 12,200,365, L50 of 15 and was a total of 472,157,411 bp long with a total of 549 contigs (Table 2.4).

Cytogenic chromosome validation

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 2.5). Figure 2.8 shows a representative mitotic metaphase cell from a male spleen cell exhibiting 56 chromosomes.

Discussion & Conclusion

The diploid chromosome number of 56 for delta smelt from our karyotyping aligns with those reported for other smelt species, 2n=54, 56 or 58 for European smelt (Nygren et al., 1971; Ocalewicz et al., 2007) and 2n=56 for pond smelt (Kitada et al., 1980). 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 species (Hartley, 1987; Ocalewicz et al., 2007). 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.

Long-read sequencing files contained a high frequency of double-copy genes. This is inherent in generating high-coverage sequence data covering the same genes multiple times. As expected, upon assembling long-reads, the number of single-copy orthologs found in A1 dropped precipitously and the number of single-copy genes found in the A1 assembly is roughly equal to the numbers of double-copy genes in the sequencing data files used for input. The female assembly A1 contained more single-copy genes than double-copy genes in the raw sequencing data. This could result from having high coverage over fragmented genes contained within the raw sequencing data, which could then be assembled into complete genes in A1.

The final total lengths for the male and female assemblies were 0.47Gb and 0.44Gb, respectively. These total lengths are similar to the wakasagi smelt genome (*Hypomesus nipponensis*) which has a total length of 0.50Gb(Xuan et al., 2021). 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. The delta smelt reference assemblies are roughly 25-30 times more contiguous that the previously published *H. niponnensis* assembly and our final male assembly contained 88.4% and final female contained 89.3% of core genes expected in the *Actinopterygii* BUSCO database. As such, our reference genomes provide a strong foundation for the future of delta smelt and evolutionary genomic research.

The male assembly is roughly 0.03Gb, or 8.0% longer than the female assembly and has a 5.6% increase of double-copy genes. These double-copy genes may account for the longer assembly length. Alternatively, or perhaps additionally, the male genome may have male specific sequences, such as a sex chromosome, which we could not determine the presence of absences of in our cytogenetic work.

The primary objective of this chapter was to create a highly contiguous chromosome-scale *de novo* genome assembly for use within and beyond the scope of this dissertation. We have achieved two chromosome-scale reference genomes, one for a female fish and one for a male fish, where a chromosome counts of 2n=56 for the species was independently validated by sequencing-free cytogenetic methods. Our two reference genomes were published to NCBI with GenBank assembly accession numbers GCA\_021917145.1 (female) and GCA\_021870715.1 (male) on February 02, 2022 and February 03, 2022, respectively and the more contiguous female genome has subsequently annotated by the NCBI Eukaryotic Genome Annotation Pipeline.

Works Cited

Andrews, Simon. (2010). *FastQC: a quality control tool for high throughput sequence data.* (0.11.9). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Belarmino, E., Nóbrega, M. F. de, Grimm, A. M., Copertino, M. da S., Vieira, J. P., & Garcia, A. M. (2021). Long-term trends in the abundance of an estuarine fish and relationships with El Niño climatic impacts and seagrass meadows reduction. *Estuarine, Coastal and Shelf Science*, *261*, 107565. https://doi.org/10.1016/j.ecss.2021.107565

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., Ponce de León, F. A., … Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*(4), 643–650. https://doi.org/10.1038/ng.3802

Campbell, M. A., Joslin, S. E. K., Goodbla, A. M., Willmes, M., Hobbs, J. A., Lewis, L. S., & Finger, A. J. (2022). Polygenic discrimination of migratory phenotypes in an estuarine forage fish. *G3 Genes|Genomes|Genetics*, *12*(8), jkac133. https://doi.org/10.1093/g3journal/jkac133

Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: A Tool Set for Repairing and Enhancing Assembled Genomes Through Integration of Genetic Maps and Conserved Synteny. *G3: Genes|Genomes|Genetics*, *10*(11), 4115–4128. https://doi.org/10.1534/g3.120.401485

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., North, K. N., Plon, S. E., Rehm, H. L., Risch, N., Rotimi, C. N., Shendure, J., Soranzo, N., & McCarthy, M. I. (2020). A brief history of human disease genetics. *Nature*, *577*(7789), 179–189. https://doi.org/10.1038/s41586-019-1879-7

Cottingham, A., Huang, P., Hipsey, M. R., Hall, N. G., Ashworth, E., Williams, J., & Potter, I. C. (n.d.). *Growth, condition, and maturity schedules of an estuarine fish species change in estuaries following increased hypoxia due to climate change*. 20.

Darwin, C. (1859). *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life.* London: J. Murray.

Dill, W. A., & Cordone, A. J. (n.d.). *History And Status of Introduced Fishes In California, 1871 – 1996*. 415.

Fisch, K. M., Henderson, J. M., Burton, R. S., & May, B. (2011). Population genetics and conservation implications for the endangered delta smelt in the San Francisco Bay-Delta. *Conservation Genetics*, *12*(6), 1421–1434. https://doi.org/10.1007/s10592-011-0240-y

Gardell, S. J., Ramjit, D. R., Stabilito, I. I., Fujita, T., Lynch, J. J., Cuca, G. C., Jain, D., Wang, S. P., Tung, J. S., & Mark, G. E. (1991). Effective thrombolysis without marked plasminemia after bolus intravenous administration of vampire bat salivary plasminogen activator in rabbits. *Circulation*, *84*(1), 244–253. https://doi.org/10.1161/01.CIR.84.1.244

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, *15*(8), e1007273. https://doi.org/10.1371/journal.pcbi.1007273

Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *10*, 645–656. https://doi.org/10.1109/TCBB.2013.68

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hammock, B. G., Hartman, R., Dahlgren, R. A., Johnston, C., Kurobe, T., Lehman, P. W., Lewis, L. S., Van Nieuwenhuyse, E., Ramírez-Duarte, W. F., Schultz, A. A., & Teh, S. J. (2022). Patterns and predictors of condition indices in a critically endangered fish. *Hydrobiologia*, *849*(3), 675–695. https://doi.org/10.1007/s10750-021-04738-z

Hartley, S. E. (1987). THE CHROMOSOMES OF SALMONID FISHES. *Biological Reviews*, *62*(3), 197–214. https://doi.org/10.1111/j.1469-185X.1987.tb00663.x

Hobbs, J. A., Lewis, L. S., Willmes, M., Denney, C., & Bush, E. (2019). Complex life histories discovered in a critically endangered fish. *Scientific Reports*, *9*(1), 16772. https://doi.org/10.1038/s41598-019-52273-8

Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences*, *118*(52), e2109019118. https://doi.org/10.1073/pnas.2109019118

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jacobs, J. W., Cupp, E. W., Sardana, M., & Friedman, P. A. (1990). Isolation and characterization of a coagulation factor Xa inhibitor from black fly salivary glands. *Thrombosis and Haemostasis*, *64*(2), 235–238.

James, N. C., Cowley, P. D., & Whitfield, A. K. (2018). The marine fish assemblage of the East Kleinemonde Estuary over 20 years: Declining abundance and nursery function? *Estuarine, Coastal and Shelf Science*, *214*, 64–71. https://doi.org/10.1016/j.ecss.2018.09.010

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

Lew, R. M., Finger, A. J., Baerwald, M. R., Goodbla, A., May, B., & Meek, M. H. (2015). Using Next-Generation Sequencing to Assist a Conservation Hatchery: A Single-Nucleotide Polymorphism Panel for the Genetic Management of Endangered Delta Smelt. *Transactions of the American Fisheries Society*, *144*(4), 767–779. https://doi.org/10.1080/00028487.2015.1037016

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Mapleson, D., Accinelli, G. G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). *Sequence analysis KAT : a K-mer analysis toolkit to quality control NGS datasets and genome assemblies*. *33*(November 2016), 574–576. https://doi.org/10.1093/bioinformatics/btw663

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

McAllister, D. E. (1963). *Revision of the smelt family, Osmeridae.* National Museum of Canada, Biological Series 191.

Moyle, P. B. (2002). *Inland Fishes of California* (2nd ed.). University of California Press.

Moyle, P. B., Herbold, B., Stevens, D. E., & Miller, L. W. (1992). Transactions of the American Fisheries Society Life History and Status of Delta Smelt in the Sacramento-San Joaquin Estuary , California. *Transactions of the American Fisheries Society*, *121*(1), 67–77. https://doi.org/10.1016/j.fertnstert.2012.04.042

Moyle, P. B., Hobbs, J. A., & Durand, J. R. (2018). Delta Smelt and Water Politics in California. *Fisheries*, *43*(1), 42–50. https://doi.org/10.1002/fsh.10014

Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, J. A. (2016). Delta Smelt: LIfe History and Decline of a Once-Adundant Species in the San Francisco Estuary. *San Francisco Estuary and Watershed Science*, *14*(2), 1–40.

Ning, Z. (n.d.). *Scaff10X: Pipeline for scaffolding and breaking a genome assembly using 10x genomics linked-reads*.

Nygren, A., Nilsson, B., & Jahnke, M. (1971). Cytological studies in the smelt (Osmerus eperlanus L.). *Hereditas*, *67*(2), 283–286. https://doi.org/10.1111/j.1601-5223.1971.tb02381.x

Ocalewicz, K., Hliwa, P., Krol, J., Rábová, M., Stabinski, R., & Ráb, P. (2007). Karyotype and chromosomal characteristics of Ag–NOR sites and 5S rDNA in European smelt, Osmerus eperlanus. *Genetica*, *131*(1), 29–35. https://doi.org/10.1007/s10709-006-9110-9

Patrick, H. J. H., Chomič, A., & Armstrong, K. F. (2016). Cooled Propylene Glycol as a Pragmatic Choice for Preservation of DNA From Remote Field-Collected Diptera for Next-Generation Sequence Analysis. *Journal of Economic Entomology*, *109*(3), 1469–1473. https://doi.org/10.1093/jee/tow047

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rosenfield, J. A., & Baxter, R. D. (2007). Population Dynamics and Distribution Patterns of Longfin Smelt in the San Francisco Estuary. *Transactions of the American Fisheries Society*, *136*(6), 1577–1592. https://doi.org/10.1577/T06-148.1

Sawyer, R. T. (1991). Thrombolytics and anti-coagulants from leeches. *Bio/Technology (Nature Publishing Company)*, *9*(6), 513–515, 518. https://doi.org/10.1038/nbt0691-513

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sommer, T., Mejia, F., Nobriga, M., Feyrer, F., & Grimaldo, L. (2011). The Spawning Migration of Delta Smelt in the Upper San Francisco Estuary. *San Francisco Estuary and Watershed Science*, *9*(2), 1–44.

Tanaka-Azevedo, A. M., Morais-Zani, K., Torquato, R. J. S., & Tanaka, A. S. (2010). Thrombin inhibitors from different animals. *Journal of Biomedicine & Biotechnology*, *2010*, 641025. https://doi.org/10.1155/2010/641025

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wasko, A. P., Martins, C., Oliveira, C., & Foresti, F. (2003). Non-destructive genetic sampling in fish. An improved method for DNA extraction from fish fins and scales. *Hereditas*, *138*(3), 161–165. https://doi.org/10.1034/j.1601-5223.2003.01503.x

Wasser, S. K., Brown, L., Mailand, C., Mondol, S., Clark, W., Laurie, C., & Weir, B. S. (2015). Genetic assignment of large seizures of elephant ivory reveals Africa’s major poaching hotspots. *Science*, *349*(6243), 84–87. https://doi.org/10.1126/science.aaa2457

Wright, B., Morris, K., Grueber, C. E., Willet, C. E., Gooley, R., Hogg, C. J., O’Meally, D., Hamede, R., Jones, M., Wade, C., & Belov, K. (2015). Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics*, *16*(1), 791. https://doi.org/10.1186/s12864-015-2020-4

Xuan, B., Kim, E. M., Song, M.-Y., Shin, Y., Jeon, J.-H., & Bae, E. (2021). Draft Genome of the Korean smelt Hypomesus nipponensis and its transcriptomic. *G3: Genes|Genomes|Genetics*, 35. https://doi.org/10.1093/g3journal/jkab147

Zoonomia Consortium. (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, *587*(7833), 240–245. https://doi.org/10.1038/s41586-020-2876-6

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Table 2.1.** Assembly metrics from Osmeridae genome assemblies publicly available prior to public release of the delta smelt genome on February 03, 2022. | | | | | | | | | | | | | | |  |
|  | **Species** | **Sex** | **Release Date** | **Accesstion  Number** | **Assembly  Level** | **Coverage** | **Total sequence length** | **Number of scaffolds** | **Scaffold N50** | **Scaffold L50** | **Number of contigs** | **Contig N50** | **Contig L50** | **Number of chr. and plasmids** | **Number of component sequences (WGS or clone)** |  |
|  | Hypomesus transpacificus (female) | F | 2022.02.04 | GCA\_021917145.1 | chromosome | 120x | 437,273,953 | 376 | 14,850,352 | 13 | 1,850 | 412,669 | 267 | 26 | 376 |  |
|  | Hypomesus transpacificus (male) | M | 2022.02.04 | GCA\_021870715.1 | chromosome | 137x | 471,985,164 | 548 | 12,200,365 | 15 | 2,127 | 347,532 | 333 | 26 | 548 |  |
|  | Thaleichthys pacificus (eulachon) | N/A | 2021.03.09 | GCA\_017311245.1 | scaffold | 210x | 416,131,685 | 324,311 | 3,050 | 34,112 | 330,739 | 2,918 | 35,367 | 0 | 324311 |  |
|  | Osmerus eperlanus (European smelt) | N/A | 2018.03.18 | GCA\_900302275.1 | scaffold | 19x | 342,758,722 | 73,274 | 6,820 | 13,139 | 99,348 | 4,524 | 21,105 | 0 | 73274 |  |
|  | Hypomesus nipponensis (Pond smelt) | N/A | 2021.05.12 | GCA\_018346875.1 | contig | 126x | 34,375,595 | N/A | 460,000 | 477 | 20,639 | 2,124 | 4,887 | 0 | 20639 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Tables and Figures

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | **Table 2.2.** Table of tissue type and storage method of sampled delta smelt from four sampling events. Included are the names referred to in the text. T= trip, F= female, M=male, BM = back muscle, SC = scales, IO = internal organ, FF = flash frozen, and PG = propylene glycol. | | | | |  |
|  | **Sex** | **Sample ID** | **Tissue Type** | **Storage Solution** | **Used for Sequencing** |  |
|  | F | T1F01\_BM\_FF | back muscle | no solution | - |  |
|  | T1F02\_BM\_FF | back muscle | no solution | yes |  |
|  | T1F03\_BM\_FF | back muscle | no solution | - |  |
|  | M | T1M01\_BM\_FF | back muscle | no solution | - |  |
|  | T1M02\_BM\_FF | back muscle | no solution | - |  |
|  | T1M03\_BM\_FF | back muscle | no solution | - |  |
|  | M | T2M01\_BM\_FF | back muscle | no solution | - |  |
|  | T2M02\_BM\_FF | back muscle | no solution | - |  |
|  | T2M03\_BM\_FF | back muscle | no solution | - |  |
|  | F | T3F01\_BM\_PG | back muscle | propylene glycol | - |  |
|  | T3F01\_BM\_FF | back muscle | no solution | - |  |
|  | T3F01\_IO\_FF | internal organ | no solution | - |  |
|  | T3F01\_SC\_FF | scales | no solution | - |  |
|  | T3F02\_BM\_PG | back muscle | propylene glycol | - |  |
|  | T3F02\_BM\_FF | back muscle | no solution | - |  |
|  | T3F02\_IO\_FF | internal organ | no solution | - |  |
|  | T3F02\_SC\_FF | scales | no solution | yes |  |
|  | M | T3M01\_BM\_PG | back muscle | propylene glycol | - |  |
|  | T3M01\_BM\_FF | back muscle | no solution | - |  |
|  | T3M01\_SC\_FF | scales | no solution | - |  |
|  | T3M02\_BM\_PG | back muscle | propylene glycol | - |  |
|  | T3M02\_BM\_FF | back muscle | no solution | yes |  |
|  | T3M02\_SC\_FF | scales | no solution | - |  |
|  | M | T4M01\_BM\_FF | back muscle | no solution | - |  |
|  | T4M01\_IO\_FF | internal organ | no solution | yes |  |
|  | T4M01\_SC\_FF | scales | no solution | - |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | | |  |
|  | **Table 2.3.** Raw sequencing data metrics. | | |  |  |  |  |  |  |
|  | **Sex** | **Sequening Company** | **Run ID** | **Number of Bases** | **Number of Reads** | **Mean Read Length** | **GC%** | **Average Read Quality\*** |  |
|  | F | PacBio HiFi | m64069\_201002\_215024 | 7,617,422,156 | 1,275,836 | 5,970 | 45% | 36 |  |
|  | F | PacBio HiFi | m64069\_200830\_055940 | 6,404,937,329 | 624,944 | 10,248 | 45% | 33 |  |
|  | F | PacBio HiFi | m64069\_200603\_183739 | 13,962,511,851 | 840,724 | 16,607 | 45% | 30 |  |
|  | M | PacBio HiFi | m64069\_200220\_045555 | 23,993,220,246 | 2,054,534 | 11,678 |  | 35 |  |
|  | M | PacBio HiFi | m64069\_200211\_020731 | 11,151,984,598 | 1,040,599 | 10,716 | 44% | 33 |  |
|  | F | 10X Illumina | 10X\_R1\_F | 94,825,601,818 | 627,984,118 | 151 | 51% | 33 |  |
|  | F | 10X Illumina | 10X\_R2\_F | 94,825,601,818 | 627,984,118 | 151 | 51% | 32 |  |
|  | M | 10X Illumina | 10X\_R1\_M | 65,806,680,934 | 435,805,834 | 151 | 49% | 34 |  |
|  | M | 10X Illumina | 10X\_R2\_M | 65,806,680,934 | 435,805,834 | 151 | 49% | 32 |  |
|  | F | Phase Genomics Hi-C | hic\_R1\_F | 13,116,671,550 | 87,444,477 | 150 | 46% | 38 |  |
|  | F | Phase Genomics Hi-C | hic\_R2\_F | 13,116,671,550 | 87,444,477 | 150 | 46% | 36 |  |
|  | \* Illumina based sequencing (10X and hi-c) calculated from the output of fastqc (Supplemental Table 2.1). | | | | | | | |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Table 2.4.** Table of assembly steps with corresponding metrics. A0 = Metrics for unassembled, filtered PacBio HiFi reads; A1 = draft resulting from initial long-read assembly step; A2 = draft resulting from scaffolding A1 assembly using linked-reads; A3 = draft resulting from scaffolding A2 assembly using hi-c data; A4 = final assembly metrics resulting from anchoring chromosomes with a linkage map. Continuity metrics created from genometools, BUSCO scores from comparison to August 05, 2020 Actinopterygii gene (n=3640) dataset. | | | | | | | | | | | |  |
|  | Metrics | | Male | | | | | Female | | | | |  |
|  | A0 | A1 | A2 | A3 | A4 | A0 | A1 | A2 | A3 | A4 |  |
|  | Continuity Metrics | N50 (bp) | 11,604 | 353,581 | 1,188,596 | 2,749,144 | 12,200,365 | 15,048 | 418,614 | 1,392,224 | 4,383,157 | 14,850,352 |  |
|  | L50 | 1,276,120 | 324 | 106 | 38 | 15 | 771,808 | 264 | 80 | 26 | 13 |  |
|  | # contigs (bp) | 3,095,133 | 2,086 | 1,106 | 705 | 549 | 2,741,504 | 1,805 | 1,012 | 515 | 376 |  |
|  | total length | 35,145,204,844 | 471,831,811 | 471,929,811 | 472,145,811 | 472,157,411 | 27,984,871,336 | 436,920,153 | 436,999,453 | 437,264,453 | 437,273,953 |  |
|  | BUSCO Scores | complete | 90.4% | 88.0% | 88.5% | 88.2% | 88.4% | 92.6% | 89.0% | 85.9% | 89.5% | 89.3% |  |
|  | single | 3.4% | 79.5% | 80.5% | 80.5% | 81.2% | 3.3% | 87.4% | 84.4% | 88.0% | 87.7% |  |
|  | double | 87.0% | 8.5% | 8.0% | 7.7% | 7.2% | 89.3% | 1.6% | 1.5% | 1.5% | 1.6% |  |
|  | fragmented | 4.0% | 1.5% | 1.1% | 1.1% | 1.0% | 3.0% | 1.1% | 3.1% | 0.8% | 0.8% |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | **Table 2.5.** Chromosome counts for delta smelt. | | | | |  |
|  | 2n Count | 54 | 55 | 56 | Total # of Cells |  |
|  | # Cells | 1 | 2 | 15 | 18 |  |
|  |  |  |  |  |  |  |

Chart, histogram

Description automatically generated

**Figure 1.1.** CDFW annual Fall Midwater Trawl delta smelt catch numbers (indices) from 1967 to 2021. CDFW did not sample in 1974 and 1979. Indices taken from CDFW publicly hosted dataset (https://www.dfg.ca.gov/delta/data/fmwt/indices.asp).

**A picture containing text, electronics, computer

Description automatically generated**

**Figure 2.1.** Pulse field gel images of extracted HMW gDNA from six rounds (A-F) of extractions. Green boxes surround lanes from extracted samples usable for long-read and linked-read sequencing (extraction distribution centered ~ 50kb) by the UC Davis Sequencing Center, yellow boxes surround lanes from samples with insufficient extraction lengths or concentration. A) Extraction #1: usable HMW gDNA from female back muscle tissue sample (T1F02\_BM\_FF), B) Extraction #2: no usable samples, C) Extraction #3: no usable samples, D) Extraction #4: No usable samples, E) Extraction #5: usable HMW gDNA from male back muscle tissue sample (T3M02\_BM\_FF); F) Extraction #6: usable HMW gDNA from female gill tissue sample (T3F02\_SC\_FF).

Chart, histogram

Description automatically generated

**Figure 2.2.** Electropherograms of male and female samples used for linked-read library prep.

Chart, line chart

Description automatically generated

**Figure 2.3.** Linked-read k-mer spectra histograms created from kat hist function. Each plot shows the number of distinct k-mers at different frequencies from male (A & B) and female (C & D) sequencing data. Histograms using k=21 (A & C), and k=31 (B & D). The high abundance of low frequency k-mers are expected as a product of sequencing and base calling errors.

A picture containing text, printer, screenshot

Description automatically generated

**Figure 2.4.** Heatmaps of k-mer frequency (x-axis) vs GC count (y-axis) colored by the number of distinct k-mers created by with the kat gcp function. Blue indicates fewer distinct k-mers with a given GC count and frequency, while yellow indicates more distinct k-mers. Plots using k=21 (A & C), and k=31 (B & D). No indication of contamination was detected in female (A & B) and male (C & D) sequencing data. Low frequency k-mers with a broad distribution of GC content is expected from sequencing and base calling errors.

Graphical user interface

Description automatically generated with medium confidence

**Figure 2.5.** K-mer comparison plot of the number of distinct k-mers at different frequencies in linked-read sequence data using the kat comp function with female (A & B) or male (C & D) samples. Plots using k=21 (A & C), and k=31 (B & D). For all plots the R1 (x-axis) and R2 (y-axis) captures a slightly different information, but no major sources of sequencing bias appear to occur. Sequencing bias in either of the two files would result in an irregular pattern in the number of distinct k-mers

Chart

Description automatically generated

**Figure 2.6.** Metrics from each assembly step. A) N50 and L50 metrics and B) The total number of contigs resulting from each assembly step. Metrics show an increase in contiguity after each step of the assembly pipeline. Large raw data (a0) metrics (L50=771,808 and total contig count= 2,741,504 for the female data, and L50=1,276,120 and total contig count=3,095,133 for the male data) have been omitted to better visualize the difference of metrics in the post assembly steps.

Chart, line chart

Description automatically generated

**Figure 2.7.** Cumulative read length plots of each iteration (A1 – A4) of the female and male genome assemblies.

Black dots on a white background

Description automatically generated with medium confidence

**Figure 2.7.** Karyotype of metaphase stage mitotic cell from a male delta smelt showing 2n = 56 chromosomes. A) unmodified image, no scale bar; B) Adobe Photoshop modified image, plus scale bar; C) Adobe Photoshop focused image, plus scale bar.