**Chromosome-scale genome assembly and investigation of the *Hypomesus transpacificus* genome for sex-specific markers, and association of the lactase persistence haplotype block with disease risk in populations of European descent.**

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, Chair

Dr. C. Titus Brown

Dr. Benjamin Sacks

Committee in Charge

2023

© 2022 Shannon Erica Kendal Joslin

© 2023 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 California 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 annual fish with freshwater resident, migratory, and semi-migratory life histories (Campbell et al., 2022; Hobbs et al., 2019). They have historically been considered an indicator species for water quality in the SFE. The species has undergone a population collapse associated with drought and anthropogenic disturbances, 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 so biological components contributing to success in the wild can be preserved. 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 evolutionary and conservation genetics research concerning delta smelt and other declining and vulnerable species in the *Osmeridae* family, such as longfin smelt. 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 kmers were identified. Chapter three is a first author paper I wrote titled “Association of the lactase persistence haplotype block with disease risk in populations of European descent” published in Frontiers in Genetics. This chapter switches organisms and investigates the potential for deleterious mutations to hitchhike in haplotype blocks which were heavily selected for in humans. This paper is a result of the work I completed in the first year and a half of my doctoral studies. Together this work contributes to the fields of evolutionary, comparative and conservation genomics. This work specifically contributes to delta smelt monitoring, management and research, and human disease risk studies.

In summary my doctoral work has provided a novel delta smelt genome assembly which is the first chromosome-level and least fragmented publicly available male and female reference genomes within the *Osmeridae* (smelt) family; an examination of male and female delta smelt sequencing data showing a discrete difference between sexes and establishes a framework for further investigation; and results suggesting that despite the fact that the human lactase persistence haplotype block harbors increased deleterious mutations compared to the rest of the genome, they seem to have little effect on prostate cancer, cardiovascular disease, and bone mineral density disease phenotypes.

Table of Contents

[Preliminary Pages – Acknowledgements, Abstract, Table of Contents 3](#_Toc144802721)

[Acknowledgements 3](#_Toc144802722)

[Abstract 3](#_Toc144802723)

[Table of Contents 4](#_Toc144802724)

[Project Background and Introduction 5](#_Toc144802725)

[The San Francisco Estuary 5](#_Toc144802726)

[Delta smelt 5](#_Toc144802727)

[Genomics and Conservation 5](#_Toc144802728)

[Lactase Persistence and Disease 7](#_Toc144802729)

[Chapter 1 – Genome assembly of Hypomesus transpacificus (delta smelt) 9](#_Toc144802730)

[Introduction 9](#_Toc144802731)

[Materials and Methods 10](#_Toc144802732)

[Sample collection & DNA extraction 10](#_Toc144802733)

[Linked-read library prep, sequencing & quality control 11](#_Toc144802734)

[Long-read library prep, sequencing & quality control 12](#_Toc144802735)

[Hi-C chromatin conformation capture library prep, sequencing & quality control 13](#_Toc144802736)

[Genome assembly 14](#_Toc144802737)

[Assembly quality assessment 14](#_Toc144802738)

[Cytogenic chromosome validation 15](#_Toc144802739)

[Results 16](#_Toc144802740)

[Sample collection & DNA extraction 16](#_Toc144802741)

[Linked-read library prep, sequencing & quality control 16](#_Toc144802742)

[Long-read library prep, sequencing & quality control 16](#_Toc144802743)

[Hi-C chromatin conformation capture library prep, sequencing & quality control 17](#_Toc144802744)

[Assembly quality assessment 17](#_Toc144802745)

[Cytogenic chromosome validation 17](#_Toc144802746)

[Discussion & Conclusion 18](#_Toc144802747)

[Works Cited 20](#_Toc144802748)

[Tables and Figures 27](#_Toc144802749)

Project Background and Introduction

The San Francisco Estuary

The San Francisco Estuary (SFE) drains 40% of California’s surface area and is a dynamic, radically anthropogenically altered ecosystem currently encompassing 1,240 square kilometers of open water and wetlands in Northern California (Conomos, 1979). Large-scale abiotic and biotic human-induced alteration of the SFE has been documented since the mid 1800s and has transformed the way water is distributed throughout the estuarine environment. Abiotic drivers of change include hydraulic mining, watershed modification, subsidence, and direct sediment removal resulting in increased pollutants such as mercury and alterations in the abundance and timing of freshwater inflow (Barnard et al., 2013). Biotic drivers of change include introduced species affecting competition for resources such as habitat and food availability (Glibert et al., 2011; Nichols et al., 1986). 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; Sommer et al., 2007).

Delta smelt

Delta smelt are a small (6 – 9 cm), 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 represents an abundant 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 at the University of California Davis (UC Davis) genetically manages a captive breeding program to maintain a refuge population at the UC Davis Fish Conservation and Culture Laboratory (FCCL).

Genomics and Conservation

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 locations of the genome (loci) essential to life for 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 to adapt 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. If a genome is highly fragmented or contains a large number of gaps, research cannot accurately investigate the extent of conserved loci and may fail to probe significant regions of the genome. As such, using multiple sequencing technologies to carry out “hybrid” methods of genome assembly increases the capacity for capturing and assembling increasingly more of an organism’s actual genome.

Over the past few decades sequencing technologies have been in an era of rapid growth. Previously, first-generation, or Sanger, sequencing only allowed for capturing DNA segments on the order of hundreds of base pairs, was limited to one individual and one locus per well, and required time consuming primer design to amplify each region of the genome under investigation. Recently developed sequencing methods generate millions to billions of base pairs of sequencing data. Massively parallel next generation sequencing technology (NGS) generates millions to billions of short reads (50 to 400 bases each) per sequencing run and third generation sequencing technology (TGS) generates millions of long reads (thousands to hundreds of thousands of base pairs in length) per sequencing run.

NGS and TGS technologies have allowed biologists to generate, low cost, high-throughput sequencing data with relative ease. Of the numerous new methods to generate sequencing data, a few (long-read sequencing, linked-read sequencing, and hi-c chromatin confirmation capture) have been found to be extremely useful for assembling highly contiguous reference genomes. PacBio HiFi generates long reads tens of thousands of base pairs in length; 10X Genomics linked-reads are sequences of Illumina short reads made into pseudo long-reads through barcoding short segments of DNA contained on the same a long fragment; and Phase Genomics hi-c chromatin confirmation capture (hi-c) are sequences Illumina short reads generated from crosslinked physically interacting DNA. There are limitations to the newer sequencing technologies: TGS requires high molecular weight (HMW) DNA, which can be difficult to generate in sufficient quantities, especially for exceedingly small or rare species and each of the three technologies (long-read linked-read and hi-c) 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 crosslinks physically interacting segments of DNA to map physical associations to span and link proximal scaffolds.

Current *de novo* genome assemblies 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, linked-read sequencing, and interaction mapping to increase the previous goat genome’s contiguity by over two orders of magnitude. At the time of publication, it proposed their hybrid genome assembly 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).

A high-quality delta smelt reference genome provides a valuable resource for conservation research. At the broadest level, a chromosome level genome will provide a resource for evolutionary studies across the tree of life. Within the *Osmeridae* family, it can be used as a reference genome for other listed, vulnerable, or declining species, such as longfin smelt (*Spirinchus thaleichthys*). And finally the delta smelt reference genome can be used to identify associations between non-neutral genetic variants and life history phenotypes in delta smelt (Campbell et al., 2022) and used to investigate domestication occurring in the propagated refuge population (Finger et al. 2018). Since the captive population of delta smelt is being used as a source for supplementing the wild population and currently represents the majority of all living individuals of the species, this genome provides a powerful resource to develop hatchery management strategies to best support adaptive genetic variation contributing to survival within the delta smelt species.

Lactase Persistence and Disease

Lactose is the main carbohydrate found in milk. The enzyme lactase, encoded by the LCT gene, allows for the breakdown of lactose in infant mammals. Various human populations continue to express LCT post weaning and can digest lactose into adulthood, a trait known as lactase persistence (LP) (Swallow, 2003). The rs4988235 (−13910 C > T) transition variant, or LP allele, in the promoter of the LCT gene allows for LP in populations of European descent. The allele frequency of this advantageous mutation rapidly rose in groups with milk and dairy production and consumption relatively recently, as seen by the signature of a relatively large haplotype block surrounding the LP allele (Bersaglieri et al., 2004; Itan et al., 2009; Ségurel & Bon, 2017). Therefore, lactose tolerant and intolerant individuals’ genetic backgrounds differ in the alleles surrounding the LP allele. Positive selection for the trait of LP can hold slightly deleterious alleles that are in linkage disequilibrium (LD) with the LP allele at a higher frequency than expected under balancing selection alone (Chun & Fay, 2011; Cutter & Payseur, 2013; Fay & Wu, 2000; Smith & Haigh, 1974). Prior work by Chun and Fay (2011) found European samples harbored multiple deleterious or neutral non-synonymous SNPs within the LCT gene and two other genes in the surrounding region. However, it is unclear whether mutations found within the LP haplotype block give rise to unfavorable phenotypes. Determining the differential risk of disease based on individual genetic backgrounds with the indirect phenotype of lactase persistence may help resolve contrasting epidemiological findings and improve public health. Therefore, the objective this study was to determine the size of the LP haplotype block and its impact on disease risk in humans with and without the LP allele.

Chapter 1 – Genome assembly of *Hypomesus transpacificus* (delta smelt)

*Formatted for submission to Journal of Heredity*

Introduction

Delta smelt (*Hypomesus* transpacificus) (phylum: Chordata, class: Actinopterygii, order: Osmeriforme, family: Osmeridae) is a small (6 – 9 cm), translucent, annually reproductive species endemic to the Sacramento-San Joaquin River Delta of the San Francisco Estuary in California (SFE). Historically thought to be solely semi-anadromous (Sommer et al., 2011), recent otolith analyses have shown the species exhibits complex migratory phenotypes of three primary life-history strategies: semi-anadromous, freshwater resident, and low-salinity brackish-water (Hobbs et al., 2019). Delta smelt are part of the Osmeridae family which represents an abundant food source for human consumption in Japan, Europe, and North America, and has 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 (CESA) in 2009. Despite protections under the ESA and CESA the species has continued to decline throughout the SFE (insert SMT and FMWT). 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 decline, resource management agencies, such as CDFW, continue to actively monitor the distribution and abundance of the wild population, and the Genomic Variation Laboratory at the University of California Davis (UC Davis) genetically manages a captive breeding program to maintain a refuge population at the UC Davis Fish Conservation and Culture Laboratory (FCCL).

How about including a transition to introductory material on Osmeridae genomes by describing how a reference genome could be used to inform genetic management at the FCCL and then follow with other uses relevant to conservation.

Prior to our work, GenBank contained three highly fragmented draft genomes from the *Osemeridae* family: 1) *Hypomesus nipponensis* (wakasagi, N50 = 0.46 Mb); 2) *Osmerus eperlanus* (European smelt, N50 = 6.8 Kb); and 3) *Thaleichthys pacificus* (euchlon, N50 = 3.1 Kb) (Table 2.1). Understanding the differences between the delta smelt genome and wakasagi genome, estimated to be 474 Mb in size and 2n=56 chromosomes, may be an important factor in conservation efforts as wakasagi were introduced into the SFE in 1959 and are known to hybridize with delta smelt in the wild (Dill & Cordone, 1997; Kitada et al., 1980; Xuan et al., 2021). Despite more contiguous metrics listed in the Xuan et al. (2021) publication, the associated reference genome hosted on GenBank presents a highly fragmented contig assembly (L50 = 477) and is less than 7.5% (34.4 Mb) of the total sequence length for an estimated genome size of 464 Mb (Table 2.1). This discrepancy poses limitations when using the wakasagi assembly as a reference genome in analyses, as it only a partial assembly.

Here I describe the development of a chromosome level delta smelt reference genome. Maybe follow this with a “high level” statement about using a hybrid approach (what technologies used?) to get a contiguous genome (etc). is a valuable tool for active genetic research aiding in the conservation and management of delta smelt and in global efforts to preserve biodiversity and understand genomic commonalities and differences in the branches on the tree of life. It has been immediately useable for identifying polygenic adaptive genetic variants such as the species’ complex migratory phenotypes (Campbell et al., 2022) to inform hatchery management strategies. Further this genomic resource will be useful for investigating hybridization with wakasagi, the genetic basis for sex determination, and domestication effects of the refuge population. On a broad level, this genomic resource allows for future evolutionary theory development by preserving a record of an imperiled species’ genome and current diversity in light of all time high rates of species loss and the persistence of the principal drivers of current and future biodiversity loss: increasingly severe drought, temperatures, and habitat loss (Caro et al., 2022; East & Sankey, 2020; Strona & Bradshaw, 2022; Ullrich et al., 2018).

Materials and Methods

Sample collection & DNA extraction

To obtain high molecular weight (HMW) genomic DNA (gDNA) for long read sequencing, we collected tissue samples from 2 male and 2 female adult delta smelt 600 days post hatch reared within the refuge population at the FCCL. Fish were euthanized according to UC Davis 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 following the protocol described by Wasko et al.(2003). Briefly, ~25-50 mg of flash frozen back muscle tissue and scales from a male and female 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, 4 M urea and 10 mg/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 0.3 M and 2X volume of ice-cold ethanol. The DNA pellet was twice washed with 70% ethanol and resuspended in an elution buffer (10 mM 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 determined by measuring 260/280 and 260/230 absorbance ratios on a NanoDrop. Libraries were created for relevant sequencing technologies were created from extractions with an average read fragment length of 50 kb.

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 a second female (T3F02\_SC\_FF) was used for generating additional long-read sequencing data to increase the total depth of coverage. One male individual (T3M02\_BM\_FF) was used for linked and long read sequencing, and internal organ tissue from a second male (T4M01\_IO\_FF) was used for generating Hi-C sequencing data (Table 2.2).

Linked-read library prep, sequencing & quality control

The linked-read sequencing technology provided by 10X Genomics is a useful technology to incorporate into my hybrid genome assembly, as it requires small amounts (nanograms) of DNA and creates highly accurate pseudo-long reads. Linked-read technology combines the benefits of long and short reads through encapsulating individual long DNA fragments (50-200 kb) into droplets of water-in-oil emulsion (GEMs), shearing long fragments into short fragments (400-600 bp), attaching a unique barcode to all DNA fragments in one GEM, sequencing paired end (PE) reads on an Illumina sequencer, and using bioinformatics to create pseudo-long reads by reassociating short reads previously contained in one GEM. Linked-read library prep was carried out by the UC Davis DNA Technologies & Expression Analysis Core Laboratory. Genomic DNA for the male and female extractions were 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 1 µl 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 350 bp (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) to generate 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. We used FastQC v0.11.9 (Andrews, 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 organellar DNA 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) reads. 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

Long-read library prep was carried out by the UC Davis DNA Technologies & Expression Analysis Core Laboratory. Genomic DNA from the same male and female used for linked-read seuqencing was sheared to roughly 17 kb 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. Each high and low-input library was subsequently loaded onto a single 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 (<https://github.com/PacificBiosciences/ccs>) statistical model on raw reads 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. If the first library prep and sequencing run of an individual did not obtain sufficient coverage for each sex, we selected a second high or low-input library prep based on the amount of HMW gDNA remaining for each individual. If no further HMW gDNA was available, we resampled tissue, extracted HMW gDNA, library-prepped and sequenced a new individual of the same sex.

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 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 proximal in DNA sequence. Continuing with the manufacturer's protocol, molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Finally, 150 bp 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 COVID-19 pandemic, we were unable to acquire male Hi-C sequencing reads, so 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. We used FastQC v0.11.9 (Andrews, 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 1) 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 using only PacBio reads; and 2) observed the number of reads which aligned to each contig (> 600 desired) and the number of high-quality reads greater than 10 kb 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. A linked- and long read- (A2) draft assembly was created using scaff10x (Ning, n.d.) with the following parameters: -longread 1 -gap 100 -matrix 2000 -reads 10 -link 8 -score 20 -edge 50000 -block 50000 to first break the assembly at locations that were incorrectly joined and scaffold the assembly into larger, more contiguous sequencing segments. After linked-read scaffolding, we prepared the Hi-C data following the Arima mapping protocol (https://github.com/ArimaGenomics/mapping\_pipeline) so interaction mapping information could be used to further scaffold the A2 draft assembly. To prep the sequencing data for further scaffolding, 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 following parameters described in the sequencing and quality control sections, 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 chromosome-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

We carried out two rounds of organ harvesting for cytogenic chromosome evaluation by karyotyping. All delta smelt were obtained from the FCCL and were transported and handled according to a UC Davis IACUC-approved animal care protocol (#21533) and standard operating procedures at the UC Davis Center for Aquatic Biology and Aquaculture (CABA). The first harvest involved fifteen 240 day post hatch subadult delta smelt and the second harvest using phenotypically sexed female (N = 13) and male (N = 15) adults (360 day post hatch). Prior to colchicine injections, fish were held in 140 L aerated tanks at 12 °C. Delta smelt (N = ?) were removed from aerated tanks at CABA, anesthetized, measured in length, injected (i.p.) with colchicine (10 µL of 1 mg/mL stock) to arrest cells in metaphase, and immediately returned to saline water (X ppt) in 5-gallon temperature-controlled buckets held at 12°C. Adult fish were separated according to sex. Fish were euthanized and a pool of organs (spleen, head kidney, heart and/or gonads) were collected post colchicine injection. For the first harvest, where sex could not be visually determined, organs were collected 4 hours post colchicine injection. For the second harvest, organs were sampled 2 hours post colchicine injection. All organs were rinsed, stored in PBS, and processed within 2 hours of dissection at ambient temperature of the CABA environment.

We established three pools of organs: a subadult mixed sex and mixed organ pool, a female spleen pool, and a male spleen and gonad pool. Organ pools were taken out of PBS, gently minced, and pipette-aspirated into single cell suspensions in a hypotonic solution (0.56% KCl) for 15-20 minutes. The cell suspensions 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 4°C. Two to three more fresh fixative treatments (centrifugation, resuspension in new fixative) were conducted, and cells were applied to slides using an air-dry method one week later. Slides were stained using the DAPI-Vectashield fluorescent dye and cell nuclei were examined using an Olympus BX-40 Microscope. Images of mitotic metaphase cells were captured and stored using the CytoVision Software (v7.4) and the number of chromosomes in the species was determined from images with intact cells with clearly defined and nonoverlapping chromosomes.

Results

Sample collection & DNA extraction

HMW gDNA extraction from back muscle tissue flash frozen and stored unsuspended in liquid yielded mixed results, so 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 3.4 µg of HMW gDNA at a concentration of 87 ng/µL for subsequent sequencing (Figure 2.1). 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 from tissues not suspended in any kind of solution after flash freezing. NanoDrop 260/280 absorbance ratios of HMW extractions were between 1.80 to 1.91, and NanoDrop 260/230 ratios ranged from 1.73 to 2.22 (Table 2.2)

Linked-read sequencing & quality control

Post-GEM DNA 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 linked- read 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 organellar DNA contamination or major sources of sequencing bias in the linked-read sequencing data.

Long-read sequencing & quality control

In order to obtain sufficient sequencing data we created two high-input libraries from one male individual (T3M02\_BM\_FF), one high-input library from one female individual (T1F02\_BM\_FF), and one high-input and one low-input library from a second female individual (T3F02\_SC\_FF) (Table 2.2). Starting gDNA inputs ranged from 6.5 ug to 20 ug of gDNA. The sheared gDNA input for the removal of single strand overhangs ranged from 1000 ng to 7 ug, and the average length of gDNA for sequencing ranged from 14-18.4 kb.

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 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 5 kb, 18.78% of the read pairs mapped to greater than 10 kb 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, respectively (Table 2.4).

After each step of the assembly 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

Organs from a total of 43 fish, comprised of subadults (n=15) and adults (n=28), were sampled for cytogenic analyses. Subadult and adult body lengths ranged from 2 to 3 inches (mean=2.5 inches) and 2.75 to 4 inches (mean=3.5 inches), respectively. Eight and fifteen cell slides were prepared from the first and second harvest, respectively. Seventy-five cell images were collected from the three pooled sample sets (mixed sex, males-only, females-only). After quality control filtration to retain only images with intact cells with clearly defined and nonoverlapping chromosomes, 18 cell images were kept for counting analysis. We determined the diploid (2n) 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) (Figure 2.8, Table 2.5).

Discussion & Conclusion

The primary objective of this chapter was to create a highly contiguous chromosome-scale *de novo* delta smelt genome assembly for use within and beyond the scope of this dissertation. We have achieved two chromosome-scale reference genomes, one for a female and one for a male delta smelt, and a chromosome count 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 was annotated by the NCBI Eukaryotic Genome Annotation Pipeline.

The final total lengths for the male and female assemblies were 0.47 Gb and 0.44 Gb, respectively. These total lengths are similar to the wakasagi smelt genome (*Hypomesus nipponensis*) which has a total length of 0.50 Gb (Xuan et al., 2021). Our final male and female assemblies had 376 and 549 scaffolds with N50’s of 0.12 Gb and 0.15 Gb, respectively. The first 28 contigs, representing the number of haploid chromosomes confirmed by karyotyping, contained 73.3% and 81.6% of the total assemblies in male and females, respectively. The delta smelt reference assemblies are roughly 25-30 times more contiguous than 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 0.03 Gb, 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 detect in our cytogenetic work.

The diploid chromosome number of 56 for delta smelt revealed by 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 a source of the karyotype variation with the Osmeridae, as observed within and among salmonid species (Hartley, 1987; Ocalewicz et al., 2007). No sex chromosomes have been reported to date for any smelts studied cytogenetically, and here we found no evidence for sex-specific chromosomes although a more detailed study is necessary given the small sample size and low resolution of images. Similar to reports from other smelt species, we note a preponderance of subtelocentric/acrocentric chromosome pairs over metacentric chromosome pairs. Chromosome composition is a descriptive metric, and our findings did not alter or affect our genome assembly process.

This high-quality reference genome provides a valuable resource for smelt conservation research. It is the most contiguous assembled smelt genome publicly available, may be broadly and specifically. At the broadest level, this chromosome level genome provides a resource for evolutionary studies across the tree of life. Within the *Osmeridae* family, it can used as a reference genome for other listed, vulnerable, or declining species, such as longfin smelt (*Spirinchus thaleichthys*). Specifically, this genome tool has already been used to identify associations between non-neutral genetic variants and life history phenotypes in delta smelt (Campbell et al., 2022). It has also been used to investigate mechanisms driving domestication of the FCCL population (Habibi et al. 2022; Habibi et al., in prep). Since the captive population of delta smelt are being used as a source for supplementing the wild population and currently represents the majority of all living individuals of the species, this genome provides a powerful resource to develop hatchery management strategies to best support conservation of adaptive genetic variation.

References

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

Barnard, P. L., Schoellhamer, D. H., Jaffe, B. E., & McKee, L. J. (2013). Sediment transport in the San Francisco Bay Coastal System: An overview. *Marine Geology*, *345*, 3–17. https://doi.org/10.1016/j.margeo.2013.04.005

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

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, *74*(6), 1111–1120. https://doi.org/10.1086/421051

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

Caro, T., Rowe, Z., Berger, J., Wholey, P., & Dobson, A. (2022). An inconvenient misconception: Climate change is not the principal driver of biodiversity loss. *Conservation Letters*, *15*(3), e12868. https://doi.org/10.1111/conl.12868

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

Chun, S., & Fay, J. C. (2011). Evidence for Hitchhiking of Deleterious Mutations within the Human Genome. *PLoS Genetics*, *7*(8), e1002240. https://doi.org/10.1371/journal.pgen.1002240

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

Conomos, T. J. (1979). *San Francisco Bay—The urbanized estuary*. American Association for the Advancement of Science- Pacific Division.

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.

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nature Reviews Genetics*, *14*(4), 262–274. https://doi.org/10.1038/nrg3425

Dill, W. A., & Cordone, A. J. (1997). *STATE OF CALIFORNIA THE RESOURCES AGENCY DEPARTMENT OF FISH AND GAME FISH BULLETIN 178 History And Status of Introduced Fishes In California, 1871 – 1996*. 415.

East, A. E., & Sankey, J. B. (2020). Geomorphic and Sedimentary Effects of Modern Climate Change: Current and Anticipated Future Conditions in the Western United States. *Reviews of Geophysics*, *58*(4), e2019RG000692. https://doi.org/10.1029/2019RG000692

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics*, *155*(3), 1405–1413.

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

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

Glibert, P. M., Fullerton, D., Burkholder, J. M., Cornwell, J. C., & Kana, T. M. (2011). Ecological Stoichiometry, Biogeochemical Cycling, Invasive Species, and Aquatic Food Webs: San Francisco Estuary and Comparative Systems. *Reviews in Fisheries Science*, *19*(4), 358–417. https://doi.org/10.1080/10641262.2011.611916

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

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The Origins of Lactase Persistence in Europe. *PLoS Computational Biology*, *5*(8), e1000491. https://doi.org/10.1371/journal.pcbi.1000491

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.

Nichols, F. H., Cloern, J. E., Luoma, S. N., & Peterson, D. H. (1986). The Modification of an Estuary. *Science*, *231*(4738), 567–573. https://doi.org/10.1126/science.231.4738.567

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

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

Ségurel, L., & Bon, C. (2017). On the Evolution of Lactase Persistence in Humans. *Annual Review of Genomics and Human Genetics*, *18*(1), 297–319. https://doi.org/10.1146/annurev-genom-091416-035340

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

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35. https://doi.org/10.1017/S0016672300014634

Sommer, T., Armor, C., Baxter, R., Breuer, R., Brown, L., Chotkowski, M., Culberson, S., Feyrer, F., Gingras, M., Herbold, B., Kimmerer, W., Mueller-Solger, A., Nobriga, M., & Souza, K. (2007). The Collapse of Pelagic Fishes in the Upper San Francisco Estuary: El Colapso de los Peces Pelagicos en La Cabecera Del Estuario San Francisco. *Fisheries*, *32*(6), 270–277. https://doi.org/10.1577/1548-8446(2007)32[270:TCOPFI]2.0.CO;2

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.

Strona, G., & Bradshaw, C. J. A. (2022). Coextinctions dominate future vertebrate losses from climate and land use change. *Science Advances*, *8*(50), eabn4345. https://doi.org/10.1126/sciadv.abn4345

Swallow, D. M. (2003). Genetics of Lactase Persistence and Lactose Intolerance. *Annual Review of Genetics*, *37*(1), 197–219. https://doi.org/10.1146/annurev.genet.37.110801.143820

Ullrich, P. A., Xu, Z., Rhoades, A. M., Dettinger, M. D., Mount, J. F., Jones, A. D., & Vahmani, P. (2018). California’s Drought of the Future: A Midcentury Recreation of the Exceptional Conditions of 2012–2017. *Earth’s Future*, *6*(11), 1568–1587. https://doi.org/10.1029/2018EF001007

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 listed in peer reviewed papers or publicly available on NCBI’s GenBank prior to public release of the delta smelt genome on February 04, 2022. | | | | | | | | | | | | | | |  |
|  | **Species** | **Sex** | **Release Date**  **(YYYY.MM.DD)** | **Accession  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** | **Final number of component sequences** |  |
|  | *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 | 324,311 |  |
|  | *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 | 73,274 |  |
|  | *Hypomesus nipponensis*  (Japanese smelt/wakasagi) | 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 | 20,639 |  |
|  | *Hypomesus nipponensis*  (Japanese smelt/wakasagi)\* | N/A | 2021.09.06 | N/A | contig | 51x | 498,930,205 | 1,987 | 464,585 | 300 | 4,106 | 316,684 | 477 | N/A | 1,987 |  |
|  | \* metrics taken from publication |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

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= sampling 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** | **NanoDrop**  **260/280** | **NanoDrop**  **260/230** | **Used for Sequencing** |  |
|  | F | T1F01\_BM\_FF | back muscle | no solution |  |  | - |  |
|  | T1F02\_BM\_FF | back muscle | no solution | 1.8 | 1.73 | 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 | 1.84 | 2.22 | 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 | 1.91 | 2.02 | yes |  |
|  | T3M02\_SC\_FF | scales | no solution |  |  | - |  |
|  | M | T4M01\_BM\_FF | back muscle | no solution |  |  | - |  |
|  | T4M01\_IO\_FF | internal organ | no solution | 1.84 | 1.8 | yes |  |
|  | T4M01\_SC\_FF | scales | no solution |  |  | - |  |
|  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | | |  |
|  | **Table 2.3.** Raw sequencing data metrics for the delta smelt genome assembly. | | | | |  |  |  |  |  |
|  | **Sex** | **Sequencing Platform** | **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 |  |  |
|  | M | VGP Hi-C | hic\_R1\_M | 77,212,809,900 | 514,752,066 | 150 | 46% | 34 |  |  |
|  | M | VGP Hi-C | hic\_R2\_M | 77,212,809,900 | 514,752,066 | 150 | 46% | 34 |  |  |
|  | \*Illumina based sequencing (linked read and hi-c) calculated from the output of fastqc (Supplemental Table 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 lineage 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 ~ 50 kb) by the UC Davis DNA Technologies and Expression Analysis Core, yellow boxes surround lanes from samples with insufficient extraction lengths or concentration. A) Extraction #1: usable HMW gDNA from female back muscle tissue (T1F02\_BM\_FF), B) Extraction #2: no usable DNA, C) Extraction #3: no usable DNA, D) Extraction #4: No usable DNA, E) Extraction #5: usable HMW gDNA from male back muscle tissue (T3M02\_BM\_FF); F) Extraction #6: usable HMW gDNA from female gill tissue (T3F02\_SC\_FF).

Chart, histogram

Description automatically generated

**Figure 2.2.** Electropherograms of male and female samples used for linked-read library prep. What do the purple and green lines on the lane diagram indicate?

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 female (A & B) and male (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, not observed here, would be expected from sequencing and base calling errors.

Graphical user interface

Description automatically generated with medium confidence

**Figure 2.5.** Plot comparing 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) capture slightly different information, but no major sources of sequencing bias are observed. 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. Metrics from a0 (Table 2.4) have been omitted to better visualize assembly improvement as sequence data from different platforms are added.

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.