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

**~~Introduction~~**

**~~Methods~~**

* ~~Tissue collection~~
* ~~Isolation of high molecular weight genomic DNA extraction~~
* ~~Library prep and sequencing~~
  + ~~Linked-read~~
  + ~~Long-read~~
  + ~~Hi-C~~
* ~~Post sequencing quality control~~
  + ~~Linked-read~~
  + ~~Long-read~~
  + ~~Hi-C~~
* ~~Assembly~~
  + ~~Long-read sequencing assembly to create Draft Assembly A using ipa~~
  + ~~Scaffolding Draft Assembly A with linked-reads to produce Draft Assembly B using Scaff10X~~
  + ~~Incorporation of hi-c sequencing data with Draft Assembly B to produce Draft Assembly C using SALSA2~~
  + ~~Anchoring Draft Assembly C scaffolds into chromosomes using chromonomer to produce a chromosome-level Final Assembly~~
  + ~~Genome assembly statistics~~
  + ~~Cytogenetic analysis~~

**Results**

* ~~Tissue collection~~
* ~~Isolation of high molecular weight genomic DNA extraction~~
* ~~Library prep and sequencing~~
  + ~~Linked-read~~
  + Long-read
  + Hi-C
* Post sequencing quality control
  + ~~Linked-read~~
  + Long-read
  + Hi-C
* Assembly
  + Long-read sequencing assembly to create Draft Assembly A using ipa
  + Scaffolding Draft Assembly A with linked-reads to produce Draft Assembly B using Scaff10X
  + Incorporation of hi-c sequencing data with Draft Assembly B to produce Draft Assembly C using SALSA2
  + Anchoring Draft Assembly C scaffolds into chromosomes using chromonomer to produce a chromosome-level Final Assembly
  + Genome assembly statistics
  + Cytogenetic analysis

**Discussion**

Introduction

From life-saving coagulation inhibitors isolated from bats, ticks, leeches, and hookworms, to developing the theory of natural selection, it is no secret that understanding a diverse landscape of animal biology has led to scientific advances and innovation. Historically, molecular biology has focused on single genes to understand disease and fuel innovation. However, single gene studies fail to account for how genes don’t exist in a vacuum, but rather work combination and coordination with other genes in a complex network of interactions. Recent advances in high-throughput sequencing technologies over the past few decades have led an era of genomic research and genomic research begins with reference genomes. Just as macroscopic observations led to innovation and understanding of the processes that effect organisms, such as the Theory of Evolution by Natural Selection, we are in an age where researchers are beginning to compile accurate microscopic observations for current and future research unthinkable with previous technologies. Genomic resources contribute to two broad categories: medicine and biodiversity. Medicine has benefitted from genomic resources by using comparative methods to identify conserved loci essential to life or different classes of organisms as well as identify genetic variants associated with disease or disease susceptibility. Biodiversity relies on having genetically diverse organisms within and between species, as the maintenance of genetic diversity is related to the evolutionary capacity for adaptation to environmental change. tktktk TRANSITION SENTENCE

Sequencing technologies have been in an era of rapid growth. From personalized medicine to population level studies, next generation sequencing (NGS) technologies have allowed for relatively easy to generate, low cost, high-throughput sequencing data. Of the numerous new methods to generate sequencing data three technologies have transformed the quality and contiguity of genome assemblies today: long-read sequencing, short-read sequencing, and interaction mapping. The first major publication to utilize a “hybrid” approach to *de novo* assembly was the domestic goat in 2017, which not only increased the previous genome’s contiguity by over two orders of magnitude but was also the most continuous *de novo* mammalian assembly of its time [tk Bickhart et al 2017]. As such, “hybrid” assemblies are an accepted and reliable way to achieve a chromosome-scale high-quality genome assembly.13,14 Since 2017, over half of all vertebrate chromosome-level assemblies submit to GenBank implemented a hybrid assembly approach to genome assembly [TK 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, short reads which have high accuracy but cannot span highly repetitive sequences to correct base calling errors, and interaction mapping which shows physical associations to span and link scaffolds with problematic motifs to fully sequence.

We used PacBio long-reads, 10X Genomics Linked-Reads, and Phase Genomics Chromatin Confirmation Capture. In addition to these three sequencing technologies, we further increased the contiguity of the genome by incorporating information from a previously published linkage map [tk CITE Lew et al.]. Finally, we independently 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* (Pond 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 [tk Dill and Cordone]. While the Pond 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 1) [tk CITE Kitada 1980 & Xuan 2021].

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Table 1.** Assembly metrics from Osmeridae genome assemblies publicly available prior to publication of the delta smelt genome (February 04, 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)** | **Reference** |  |
|  | 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 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

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 [tk Patrick 2016, Wasko 2018?].

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. [tk CITE]. 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 bar-coded DNA fragments were recovered for Illumina library construction. The amount and fragment size of post-GEM DNA was quantified prior using 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 in-house 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 first extracted HMW gDNA from a female, we used the female linked-read data to improve our estimate of delta smelt genome size through a more accurate k-mer based approach using Genomescope2 [tk CITE]. After, we used the updated genome size estimate to adjust the amount of all subsequent sequencing data generated for assembly.

To quality control for contamination and sequencing bias errors, we conducted three computational quality control steps (kat hist, kat gcp, and kat comp) using the software program KAT20. Each step splits sequencing data into sub-sequences of a given length, or k-mers, and plots out frequencies, or comparisons, to visually inspect the data for quality issues. 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. Sequencing bias in either of the two files would result in an irregular pattern in the number of distinct k-mers.

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 [tk CITE]. 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 data was generated by Phase Genomics (Seattle, WA) using Proximo Hi-C 2.0 Kit, a commercially available version of the Hi-C protocol [tk CITE], 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 were able to use the female sequencing data for both male and female scaffolding.

Male chromatin conformation capture data was generated by the Vertebrate Genome Project (VGP). tktktk

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 next step. We also observed the number of reads which aligned to each contig (>600 desired) and the number of reads that are sufficiently far apart (1-15% expected).

Genome Assembly

We generated an initial assembly (A1) purged of duplicate haplotigs using IPA v1.3.1 [tk CITE], which uses purge\_dups v1.2.3 [tk CITE] and Racon v1.4.13 [tk CITE] wrappers to generate phased primary (A1) and alternative assembly files polished of errors. To scaffold the initial A1 assembly using linked-read data, we used scaff10x [tk CITE] 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, then to scaffold the assembly into larger, more contiguous sequencing segments composing the linked- and long-read (A2) assembly. After linked-read scaffolding, we followed the Arima mapping protocol [tk CITE] to prepare the hi-c data to use interaction mapping information to further scaffold the A2 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-r118821 and samtools v1.7 22. Next, we retained the 5’ end of the read to eliminate chimeric reads using a custom Arima perl script [tk CITE?]. Then, we paired the hi-c reads to produce paired-end BAM files, and used PicardCommandTools [tk CITE] to add read groups and remove PCR duplicates. After filtering our hi-c data, we converted BAM files to sorted BED files with bedtools v2.29.223. Finally, we used SALSA224 with non-default parameters (-i 5 -x GATC -m yes) to scaffold the A2 assembly with our filtered hi-c data to produce a linked, long, and hi-c read (A3) assembly. To anchor the A3 assembly into chromosomes we used chromonomer v1.1325 in combination with a previously published linkage map9 to produce a chromosome-level reference genome (A4) assembly.

Assembly Quality Assessment

After each step we evaluated the contiguity, content, and composition of each draft (A1-A4) assembly. To assess each assembly’s completeness, we used the evolutionarily informed Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.626 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 [tk CITE]. Read length histograms were generated using jellyfish [tk CITE]. All assemblies within each sex, and between sexes were compared using QUAST [tk CITE].

We first assessed the completeness of ‘LA1670’ and ‘LA1673’ genome assemblies using the Benchmarking Universal Single-Copy Orthologs (BUSCO) database (v.4.1.1) in the genome mode to search for genes conserved in Embryophyta species. The embryophyta\_odb10 data set, created on 20 November 2019, comprises 57 species and 425 genes.

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 1). Back muscle tissue from one female individual (T1F02\_BM\_FF) was used for linked, long, and hi-c sequencing and scale 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. All successful extractions were not suspended in solution after flash freezing (Table 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  | **Table 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. | | | | | |  |
|  | **Trip Number** | **Sex** | **Sample ID** | **Tissue Type** | **Storage Solution** | **Used for Sequencing** |  |
|  | 1 | 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 | - |  |
|  | 2 | M | T2M01\_BM\_FF | back muscle | no solution | - |  |
|  | T2M02\_BM\_FF | back muscle | no solution | - |  |
|  | T2M03\_BM\_FF | back muscle | no solution | - |  |
|  | 3 | 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 | - |  |
|  | 4 | M | T4M01\_BM\_FF | back muscle | no solution | - |  |
|  | T4M01\_IO\_FF | internal organ | no solution | yes |  |
|  | T4M01\_SC\_FF | scales | no solution | - |  |
|  |  |  |  |  |  |  |  |

**A picture containing text, electronics, computer

Description automatically generated**

**Figure 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) No usable samples, D) 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 scale tissue sample (T3F02\_SC\_FF).

Linked-read library prep, sequencing & quality control

Post-GEM DNA quantified libraries showed normal distributions and were accepted for sequencing. [NEED THESE FROM DIANA]. We generated 94,825,601,818 bp of paired-end linked-read sequencing data from the female speciemen. Using the Genomescope 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 generated 94,825,601,818 bp (97x coverage) and 65,806,680,934 bp (67x coverage) of female and male linked-read sequencing data, respectively. (Table TK).

Mapped histograms for each sample and at each value of k showed discrete, single peaks indicating no sign of contamination (Figure TK). All GC count frequency plots show roughly normal circular distributions of distinct k-mers (Figure TK). 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 TK). These data together indicate no observable signs of bacterial or organelle contamination or major sources of sequencing bias in our sequencing data.



**Figure TK.** 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 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 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 and no major sources of sequencing bias appear to occur.

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 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 (Supplemental Table 1). 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 (Supplemental Table 1).

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 sufficient to be used for subsequent assembly.

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

From Phase Genomics, we received sequence files with 87,444,477 read pairs in total. Associated sequencing data reports indicate successful library prep and sequencing. The data contained an average of 2,966.33 read pairs per contig greater than 5kbp and 18.78% of the read pairs mapped to greater than 10 kilobases (kbp) apart and 56.38% of reads were considered high quality.

Assembly Quality Assessment

We searched raw data and each draft assembly for 3,640 conserved single-copy orthologs withing the 05 August 2020 Actinopterygii lineage dataset using BUSCO. Whole genes or sequence fragments of 95.6% (3.3% complete single copy, 89.3% complete double copy, and 3.0% fragmented) and TKTK% (TK% complete single copy, TK% complete double copy, and TK% fragmented) of the conserved gene dataset were present in the female and male HiFi data, respectively.

After each step, the assembly length and N50 increased, and the L50 and total number of contigs decreased. 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 (Figure TK). 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 (Figure TK[cumulative read length & quasty thingz], Table TK).

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

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 5). Figure 6 shows a representative mitotic metaphase cell from a male spleen cell exhibiting 56 chromosomes.

Discussion

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

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

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

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

* its expected that the raw data would have a bunch of double copies
* also expected BUSCO score doesn’t go up much from the # of double copies in the raw data… this is due to not having overlap for high a high confidence in assembly
* surprising that also the male asm is much long than female, it doesn’t contain more BUSCOs