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

**~~Introduction~~**

**Methods**

* Tissue collection
* Isolation of high molecular weight genomics DNA extraction
* Library prep and sequencing
  + Linked-read
  + Long-read
  + Hi-C
* Post sequencing quality control
* 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**

**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. In addition to various sequencing technology, genome assembly contiguity can be further increase by incorporating linkage maps, which uses the frequency of recombination between markers to generate relative distances throughout the genome.

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 had an N50 of 0.46 Mbp and an L50 of 477 [tk CITE Kitada 1980 & Xuan 2021].

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



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

**BOX 2: Long-read sequencing methods**, (e.g., PacBio HiFi) take long segments of extracted DNA (over 50kbp), shear them into a few segments (10-20kbp) and circularize each segment. The circular molecules are then run through a sequencer to get an accurate circular consensus sequence over 20kbp in length.

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

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



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

**Methods**

*Sample Collection*

To obtain high molecular weight (HMW) genomic DNA, we collected the tissue of male and female fish 600 days post hatch reared within the refuge population at the FCCL. One female individual and two male individuals used for subsequent sequencing and genome assembly. The first male individual was used for linked and long-read sequencing, and the second individual was used for chromatin conformation capture. 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, and immediately flash froze all tissue samples until DNA extraction occurred.

*Isolation of high molecular weight genomic DNA*

High molecular weight (HMW) genomic DNA (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.

*Linked-read library prep & sequencing*

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 our previous RAD-sequencing-based estimate of a haploid delta smelt genome size (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.

*Long-read library prep & sequencing*

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).

Two PacBio HiFi library types (high- and low-input) were created. 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).

A low-input HiFi library was 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 0.45X 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.

*Hi-C chromatin conformation capture library prep & sequencing*

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

*Long-read post-sequencing quality control*

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%, were accepted and used for subsequent assembly.

*Linked-read post-sequencing quality control*

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 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, 31, and 41. Uncontaminated samples were expected to have a single peak with a surplus of k-mers at a very low frequency due to sequencer errors. 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. Contaminated samples are expected to have a non-normal distribution. For example, samples contaminated with bacteria will have more k-mers with GC counts above 50%. We plotted GC counts against the frequency of k-mers of length k=21, 31, and 41.

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, 31, and 41. Sequencing bias in either of the two files would result in an irregular pattern in the number of distinct k-mers.

*Hi-C chromatin conformation capture post-sequencing quality control*

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 evaluate each assembly’s gene content using an evolutionarily informed number of conserved single-copy orthologs in the Actinopterygii lineage we used BUSCO v4.0.626 to obtain BUSCO metrics. To acquire assembly metrics, such as N50, L50, number of contigs, and assembly length, we used GenomeTools v1.5.10 [tk CITE] (Table TK). Read length histograms were generated using jellyfish [tk CITE]. All assemblies within each sex, and between sexes were compared using QUAST [tk CITE].

*Cytogenic chromosome validation*

Since cytogenetic work does not use sequencing-based methods and allows for an independent validation of the number of chromosomes expected in the final delta smelt genome assemblies we karyotyped male and female delta smelt.

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*

We extracted HMW gDNA from the tissue samples of one female individual and two male individuals (Supplemental Figure 1). One male individual (T3M02\_BM\_FF) was used for linked and long read sequencing, and the second was used for generating hi-c sequencing data.

*Linked-read library prep & sequencing*

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). The library was quantified by qPCR with a Kapa Library Quant kit (Kapa Biosystems-Roche). [NEED THESE FROM DIANA]. The k-mer based haploid genome size from the female was estimated to be 0.49 Gb. We generated approximately 45 gigabytes of female linked-read sequencing data and 30 gigabytes of male linked-read sequencing data for a total of roughly 120x and 80x coverage, respectively (Table 1).

*Long-read library prep & sequencing*

A total of five movie collections (150 hours of sequencing data) were collected. Two male movie collections generated roughly 25x coverage (data combined into Supplemental Data 1) and three female movie collections generated roughly 28x coverage (Run 1 & 2 combined into Supplemental Data 2, Run 3 data in Supplemental Data 3). This amount of long-read sequencing data was sufficient to continue on to subsequent steps.

*Linked-read library prep & sequencing*

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

*Hi-C chromatin conformation capture prep & sequencing*

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

**Quality Control**: After sequencing, bioinformatics quality control (QC) steps are necessary at multiple stages in the workflow of assembling of a reference genome. Immediate sources of error in sequencing data can result from sequenced DNA being contaminated with off-target DNA in the wet lab during preparation for sequencing, or biases in base calls from the sequencing machine. Other sources of error may come from having too great of a proportion of poor-quality reads.

*Long-read post-sequencing quality control*

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

*Linked-read post-sequencing quality control*

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

*Hi-C chromatin conformation capture post-sequencing quality control*

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

*Genome Assembly*

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

**BOX 4: Assembly statistics**

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

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



**N50 = 60**

**L50 = 3**

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

*Cytogenic 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.