**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**

THE NEED

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

CURRENT TECHNOLOIGES

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]. Since 2017, over half of all vertebrate chromosome-level assemblies submit to GenBank use two or more sequencing technologies [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 and three highly fragmented Osmeridae(smelt) genome assemblies from three different genera (*Hypomesus, Thaleichthys*, and *Osmerus*) were publicly available through GenBank. 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].

The use of next generation sequencing (NGS) technologies in conjunction with a highly contiguous and accurate reference genome increases the power and precision of inferences made in population genetic studies (e.g., analyses of population structure, genetic diversity, and local adaptation)8. Without a reference genome, DNA sequences captured by the sequencer but not aligned to a reference genome fail to account for how each piece of sequenced DNA interacts with all other sequences (i.e., linkage patterns). However, by aligning to a reference genome we know where each segment of sequenced DNA lies within the genome and relative to other sequenced DNA. An assembled genome vastly increases the power to answer questions concerning demography, adaptation, fitness, and disease susceptibility.9,10 For example, previous work in Chinook salmon and steelhead found a small number of markers that were associated with run-timing in each species. In 2014, these markers were located on five different scaffolds and next to gaps in the highly fractured genome assembly. Reanalysis using long-reads to span gaps and reorient and link fragmented scaffolds demonstrated that all the associated markers were from a single locus (i.e., the GREB1L region)11. Since the 2017 study, continued improvements to salmonid reference genomes have enabled more detailed mapping of life-history variation and phenotypic traits in the same locus12, which may lead to significant changes in management. This example underscores how reference genomes can propel conservation research forward.

Recently, genome sequencing technologies have become both more cost effective and efficient. “Hybrid assemblies” (assemblies that use multiple NGS technologies) are a reliable way to achieve a chromosome-scale high-quality genome assembly13,14. The Vertebrate Genomes Project, a consortium aimed towards developing an assembly pipeline and quality standards for genome assemblies of all vertebrates, established quality goal metrics for the continuity, completeness, and accuracy of reference genomes14. Therefore, a main goal of this project is to develop a highly accurate chromosome-scale reference genome, hereafter called “reference genome”, using linked-reads (Box 1), long-reads (Box 2), hi-c chromatin confirmation capture (Box 3), a genetic linkage map15 and various assembly software programs (Figure 1).

PREVIOUS EVOLUTIONARILY CLOSE GENOMES

**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 (Figure 1A)*

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

Though we had intended to make a single trip to the FCCL to collect delta smelt samples for sequencing, we found recovering sufficient HMW DNA difficult, and ended up conducting four separate trips to sample different tissue types from 600 days post hatch (dph) male and female delta smelt (Table 2) with the final goal of producing enough HMW DNA to sequence each sex with each of the three technologies. Additionally, at the start of this project long-read sequencing was costly and we sought to only incorporate the sequencing technology if absolutely necessary. Therefore, for our first and second sampling trips we sought to acquire enough tissue to provide sufficient quantity of HMW DNA for linked-read and hi-c sequencing. However, at the end of 2019 the price of long-read sequencing dropped dramatically. This development combined with our mixed results from Trips 1 and 2, prompted a third trip to acquire enough tissue to sequence a single male fish with all three of our chosen technologies, and enough tissue from a female fish to sequence with long-reads (as already we had enough HMW DNA for a female for hi-c and linked-reads). An error committed by the sequencing center required us to make one additional trip, Trip 4, to sample one additional male specimen. On each occasion fish were euthanized according to the approved animal care protocols/standard operating procedures. Back muscle, internal organ, and/or scale tissues were sampled onsite at the FCCL and cooled for transportation directly to the UC Davis DNA Technologies & Expression Analysis Core Laboratory (UC Davis Sequencing Center).

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

For linked-read and long-read sequencing, HMW DNA extractions from the fish tissues occurred at the UC Davis Sequencing Center using the protocol described in Wasko *et al.* (2003)17. The size range of extracted DNA fragments were determined using a pulse field gel run for 24 hours. This run was conducted at a low frequency to not shear the DNA. Physical sampling and extractions were repeated until the mean distribution of extraction lengths was 50kbp or greater and there was sufficient quantity of HMW DNA to carry out the required sequencing technologies for each sex. For HMW DNA extractions and subsequent hi-c sequencing, a female tissue sample was sent to Phase Genomics and a male tissue sample was sent to the Vertebrate Genome Project for isolation of HMW DNA. Samples were sent to different locations due to COVID-19 delays at Phase Genomics during the time of sampling the male fish.

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

The inclusion of long-read data dramatically assists in creating a more contiguous genome assembly because it spans repetitive elements and resolves chimeric sequences (erroneously joined artifact sequences) throughout the genome. This means that long-read data provides greater continuity of scaffolded **contigs** (a stretch of DNA sequence created from a consensus of reads). This is because the reads (sequence fragments produced by the sequencer) are thousands rather than hundreds of base pairs long. Previously, long-read sequencing was both costly and had a high error rate of up to 10%. During 2020, a new PacBio HiFi chemistry was developed which dramatically reduces the number of errors to under 1%.

PacBio HiFi SMRTbell® Libraries (<https://www.pacb.com/>) were prepped following the SMRTbell Express Template Prep Kit 2.0 procedure. The UC Davis Sequencing Center used a Megaruptor to shear DNA to an average fragment length of ~15kb. The input for the library prep was 5.6 ug of DNA and the library was size selected down to be 11kb. Prepped DNA was run on a Sequel II machine and sequencing data was collected as recorded “movies” of nucleotides. Each movie collection lasted for 30 hours per run. Additional movies were collected until the amount of sequencing data for each sex was ~25-30x coverage.

*Methods:* *Linked-read library prep & sequencing (Figure 1C)*

Once we had sufficient extracted HMW genomic DNA fragments, DNA was adjusted to a concentration of 0.91 ng/µl. We selected the 10X Genomics platform (<https://www.10xgenomics.com/technology/>) to generate our linked-read sequence data. 10X Genomics library preparation takes extracted HMW gDNA, shears it into 50kbp long segments, inserts each fragment into an oil coated gel emulsion bead (GEM), further shears DNA within each bead, and attaches one unique barcode to all DNA fragments within each GEM for identification post-sequencing. Whole genome sequencing libraries were prepared using Chromium Genome Library & Gel Bead Kit v.2 (10X Genomics, cat. 120258), Chromium Genome Chip Kit v.2 (10X Genomics, cat. 120257), Chromium i7 Multiplex Kit (10X Genomics, cat. 120262), and Chromium controller according to manufacturer’s instructions. After library preparation, 1.14 ng of template gDNA was loaded on a Chromium Genome Chip and sequenced on an Illumina NovaSeq6000 150bp PE lane (Illumina, San Diego, CA). We used a previous RAD-sequencing-based estimate of a haploid delta smelt genome size (0.6Gb) to sequence the first sample to an estimated 80x coverage.

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

*Methods* *Hi-C chromatin conformation capture prep & sequencing (Figure 1C)*

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

*Methods* *Long-read post-sequencing quality control (Figure 1C)*

Sequencing data were downloaded from Bioshare, the UC Davis Sequencing Center’s host service. We used PacBio’s CCS statistical model on raw reads (<https://github.com/PacificBiosciences/ccs>) to generate highly accurate consensus sequences with known base quality values and convert binary data to fastq format for downstream analysis. A base quality value assigns a score to each letter (nucleotide) of sequencing data to denote the level of confidence in each base called by the sequencer. Reads with quality scores over Q20, denoting an error probability of 0.01%, were used for subsequent assembly.

*Methods* *Linked-read post-sequencing quality control (Figure 1C)*

To quality control for contamination and sequencing bias errors, we conducted three computational quality control steps (kat hist, kat gcp and kat comp) using the software program KAT20. Each step splits sequencing data into sub-sequences of a given length, or k-mers, and plots out frequencies, or comparisons, to visually inspect the data for quality issues. All bioinformatics work was conducted on the UC Davis farm compute cluster (the farm).

We looked for signs of contamination using the kat hist and kat gcp functions within the software KAT. First, we used the kat hist function to check for signs of contamination by plotting a histogram of the observed number of distinct k-mers at different frequencies for lengths k=21, 31, and 41 for female and male sequencing data. Uncontaminated samples are expected to have a single peak with a surplus of k-mers at a very low frequency due to sequencer errors. Second, we used the kat gcp function to plot the proportion of the k-mer comprised of the G and C nucleotides against the frequency of the that k-mer in the sequencing data and the number of distinct k-mers for a given GC count vs. frequency. Contaminated samples are expected to have a non-normal distribution. For example, samples contaminated with bacteria will have more k-mers with GC counts above 50%. We plotted GC counts against the frequency of k-mers of length k=21, 31, and 41 for the female and male sequencing data.

Lastly, we evaluated the data for sequencing bias between the forward (R1) and reverse (R2) files. We used the kat comp function to plot the frequency of a given k-mer in each of the paired-end sequence data files. Sequencing bias in either of the two files would result in an irregular pattern in the number of distinct k-mers. We plotted the number of distinct k-mers at different frequencies for the R1 and R2 files against one another for k-mers of length k=21, 31, and 41 for the female and male sequencing data.

*Methods* *Hi-C chromatin conformation capture post-sequencing quality control (Figure 1C)*

In order to assess if our sequencing data would be useful in linking scaffolds, we looked at a percentage of high-quality reads (minimum mapping quality of greater than or equal to 20, a maximum edit distance of less than or equal to 5, and no duplications) that mapped to our draft assembly created from the next step. We also observed the number of reads which aligned to each contig (>600 desired) and the number of reads that are sufficiently far apart (1-15% expected).

*Methods Genome assembly (Figure 1D-H)*

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

Step 1: Use long-read sequencing data to create Draft Assembly A

To begin, we generated two phased assemblies, one for each sex, by inputting the long-read sequencing data into the Improved Phased Assembly tool (IPA, <https://github.com/PacificBiosciences/pbipa>). The assembly product was polished of errors, purged of duplicate haplotigs, and phased into primary and alternative assembly files. The primary assembly file contains a contiguous haploid assembly, while the alternate assembly file contains the alternate haplotype of the diploid delta smelt.

Step 2: Incorporate linked-reads into Draft Assembly A to produce Draft Assembly B

After creating the initial draft assembly, we incorporated the linked-read data to first break the assembly at locations that were incorrectly joined, then to scaffold the assembly into larger, more contiguous chunks using the software scaff10x (<https://github.com/wtsi-hpag/Scaff10X>).

Step 3: Incorporate hi-c data into Draft Assembly B to produce Draft Assembly C

In order to use long distance information, we indexed the assembly produced in the scaff10x step and mapped hi-c reads to the draft assembly using bwa21 and SAMtools22. We used the Arima Mapping pipeline perl scripts (<https://github.com/ArimaGenomics/mapping_pipeline>) to pair reads, and quality filter the 5’ end and for mapping quality. Next, we added read group information, marked duplicated reads, and sorted the mapped read files with picard (<http://broadinstitute.github.io/picard/>). These data were then converted into mapped bed files using bedtools23. We then used the mapped bed files, scaffolded assembly and the initial alternative assembly as input to close gaps and further scaffold the assembly using the SALSA2 pipeline24 with non-default parameters: -i 5 -x GATC -m yes. SALSA2 uses the frequency of how often reads pair (or are sequenced) together to determine how close any two locations in the genome are to one another; it finds the locations of the paired reads in the draft genome and links the two locations to close gaps and produce a more contiguous assembly.

Step 4: Use linkage map with Draft Assembly C to produce Final Assembly

Finally, we anchored our assembly into chromosomes by using a genetic linkage map produced in Lew *et al.* (2015)15 with the output from the hi-c assembly step and the software chromonomer25.

We used a combination of software and metrics to evaluate each draft assembly after every step of the assembly process (Table 4). The assembly length should be as close to the estimated genome size as possible. The N50 metric is the length of the scaffold where half of the assembly is held in scaffolds of that size or larger. The L50 metric tells the number of scaffolds that contain half of the assembly (Box 4). BUSCO26 scores were used to evaluate the completeness of each assembly as expected from a core set of highly conserved single copy genes in the *Actinopterygii* lineage.

*Methods* *Cytogenic (karyotype) chromosome validation*

We collaborated with Dr. Mary Delaney to determine the diploid number of chromosomes in delta smelt, which has not been previously reported. This cytogenetic work does not use next generation sequencing-based methods and allows for an independent validation of the number of chromosomes expected in our final delta smelt genome assemblies.

Fish were anesthetized with Colchicine (10 microliters of 1 mg/ml stock) and returned to water after sampling. Individual fish lengths were measured. In both cases organs were rinsed, stored in PBS at ambient temperature of the CABA environment. We established two spleen pools, one of male (n = 15), and one of female (n = 13) specimens and one additional male gonad was harvested. Organ pools were gently aspirated into single cell suspensions by pipetting in hypotonic solution (0.56% KCl) for 15-20 min. Cells were centrifuged at ~1000 rpm for 10 min, supernatant hypotonic solution was removed and a 3:1 fixative (methanol:glacial acetic acid) was added. Cell pellets were resuspended and stored at 4ºC. Two to three more fixative washes (centrifugation, resuspension in new fixative) were conducted, and cells were applied to slides one week later. Slides were stained using the DNA staining fluorescent dye (DAPI) and cells were examined using an Olympus BX-40 Microscope. Images of cells were captured and stored using the Cytovision Software and the number of chromosomes in the species were determined from those images.

**Results**

*Sample Collection*

We took a total of four trips to sample tissue for HMW DNA. On the Trip 1 we sampled back muscle from three males and three females (Table 2). These samples were flash frozen using liquid nitrogen and transported on dry ice. While we obtained sufficient HMW DNA from a female specimen, none of the male fish produced sufficient quantities. Therefore we returned to the FCCL for a second trip and selected larger males. The extracted DNA lengths from Trip 2 males were also insufficient. A decision to take Trip 3 was made due to the discovery of a tissue sampling method that uses additional tissue types (not just back muscle tissue) and a new tissue preservation storage solution of cooled propylene glycol16 rather than flash freezing samples in liquid nitrogen. On Trip 3, we sampled back muscle tissue, internal organs, and scales from a total of two males and two females. Additionally, because DNA can be fragmented in the freeze-thaw process, we hedged our bets for sampling on Trip 3 and preserved half of all sampled tissues in propylene glycol at 4°C, and half flash frozen and transported both on dry ice. On Trip 4, we sampled back muscle, scales and internal organs from one male fish. All sampled tissue was flash frozen and stored on dry ice for transportation.

*Isolation of high molecular weight genomic DNA*

Trip 1: DNA from one male (T1M03) and one female (T1F02) DNA was successfully extracted from the first sampling (See Table 2 for fish identifications corresponding to trips). HMW DNA from the T1F02 had DNA fragments of sufficient length for sequencing (Figure 2A).

Trip 2: We performed three separate rounds of extractions on tissue samples from two different males (T2M02 and T2M03). Despite multiple attempts at extracting HMW DNA, we did not obtain fragments of sufficient length for sequencing (Figure 2B-D).

Trip 3: We performed two separate rounds of extractions on multiple different tissue types from both males and females (Table 2). From these extractions we acquired DNA of sufficient length from flash frozen back muscle tissue from one male (T03M02\_BM\_FF, Figure 2E) and flash frozen scales from one female (T3F02\_SC, Figure 2F). We did not observe that storing tissue in propylene glycol had an effect on increasing the fragment length of extracted DNA (Figure 2E & F).

Trip 4: The tissue samples from the male specimen sampled on trip 4 were sent directly to the Vertebrate Genome Project for subsequent extraction and sequencing where it was successfully extracted and sequenced.

*Long-read library prep & sequencing*

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.