**Tables & Figures**

**Table 1**. List and description of deliverables with the status of each task. Tasks were re-ordered according to the chronological steps taken to carry out the primary goal of each task.

|  |  |  |
| --- | --- | --- |
| **Deliverable** | **Description** | **Status** |
| **Task 1. Sequence and assemble a draft delta smelt genome + linkage map** | | |
| **Task 1.1 Long-read sequencing & assembly (PacBio)** | |  |
| a. Sample preparation | Sample tissue and organs of male and female fish. | Completed |
| b. Sequencing | PacBio Library Prep | Completed |
| HiFi Sequencing | Completed |
| c. Assembly | IPA assembly | Completed |
| BUSCO & Summary Statistics | Completed |
| **Task 1.2 Linked-read sequencing (10X Genomics)** | | |
| a. Sample acquisition | Sample tissue of male (M) and female (F) fish. | Completed |
| b. Sequencing | 10X Chromium Library Prep | Completed |
| Illumina NovaSeq Sequencing | Completed |
| c. Assembly | Scaff10X assembly | Completed |
| BUSCO & Summary Statistics | Completed |
| **Task 1.3 Hi-C Chromatin Conformation Capture & Sequencing (Phase Genomics)** | | |
| a. Sample preparation | Sample tissue and organs of male and female fish. | Completed |
| b. Sequencing | Hi-C Library Prep | Completed |
| Deep Sequencing | Completed |
| c. Assembly | Salsa2 scaffolding | Completed |
| BUSCO & Summary Statistics | Completed |
| **Task 1.5 Map linked regions to reference genome** | | |
| a. Map regions | Chromonomer chromosome anchoring | Completed |
| **Task 2. Estimate contemporary Ne using RAD-seq data & completed genome** | | |
| **Task 2.1 DNA collection and extraction** | | |
| a. Sample Collection | Collect samples from BY2017, BY2018 and BY2019 cohorts | Completed |
| b. DNA Extraction | Extract DNA from BY2017, BY2018 and BY2019 cohorts | Completed |
| c. RAD-sequencing | Sequence BY2017, BY2018 and BY2019 cohorts | Completed |
| **Task 2.2 Estimate Ne in BY2017, BY2018 and BY2019 cohorts** | | |
| a. RAD-sequencing data processing | Split, align and quality control all sequencing data | Completed |
| b. Estimation of Ne | Temporal Ne | Completed |
| c. Quantify Genetic Diversity | Theta estimates | Completed |
| **Task 3. Domestication selection** | | |
| **Task 3.1 Select individuals for RAD-seq analysis** | | |
| a. Determine domestication index (DI) | Calculate domestication index for each individual from each generation | Completed |
| b. Group individuals based on their DI | Assign individuals to three DI group: Low, medium and high | Completed |
| c. Determine recovery rate of each individual | Calculate recovery rate of each individual | Completed |
| d. Group individuals based on their recovery rate | Assign individuals to two groups: Low and high recovery rate | Completed |
| **Task 3.2 Preparation and sequencing of individuals** | | |
| a. Sample Collection | Collect samples from BY2008 to BY2015 cohorts | Completed |
| b. DNA Extraction | Extract DNA from BY2008 to BY2015 cohorts | Completed |
| c. RAD-sequencing | Sequence BY2008 to BY2015 cohorts | Completed |
| **Task 3.3 Population Genetic Analysis** | | |
| a. RAD-sequencing data processing | De-multiplexing  Quality filtering | Completed |
| a. Identifying SNPs of domestication selection | Genotype calling  Fst analysis between groups with completed reference genome assembly | Completed |
| d. Association study | GWAS between DI and recovery rate | Completed |
| **Task 4. Identification of sex specific markers** | | |
| **Task 4.1 Investigation of delta smelt genome** | | |
| a. RAD-sequencing data processing | De-multiplex | Completed |
| QC Sequencing data | Completed |
| b. Identification of sex marker(s) | GWAS in M vs F | Completed |
| Depth comparison | Completed |
| K-mer analysis | Completed |
| **Task 4.2 Fluidigm assay creation** | | |
| a. Submit sequences | Submit sex specific sequences to Fluidigm for SNP Type Assay design | N/a |
| b. Validation | Validation of SNP Assay | N/A |

**Table 2.** Table of tissue type and storage method of sampled delta smelt from three different trips. 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 Name** | **Tissue Type** | **Storage Method** | **Used for Sequencing** |
| 1 | F | T1F01 | Back muscle | Flash frozen | - |
| T1F02 | yes |
| T1F03 | - |
| M | T1M01 | Back muscle | Flash frozen | - |
| T1M02 | - |
| T1M03 | - |
| 2 | M | T2M01 | Back muscle | Flash frozen | - |
| T2M02 | - |
| T2M03 | - |
| 3 | F | T3F01\_BM\_PG | back muscle | propylene glycol | - |
| T3F01\_BM\_FF | back muscle | flash frozen | - |
| T3F01\_IO\_FF | internal organ | flash frozen | - |
| T3F01\_SC\_FF | scales | flash frozen | - |
| T3F02\_BM\_PG | back muscle | propylene glycol | - |
| T3F02\_BM\_FF | back muscle | flash frozen | - |
| T3F02\_IO\_FF | internal organ | flash frozen | - |
| T3F02\_SC\_FF | scales | flash frozen | yes |
| M | T3M01\_BM\_PG | back muscle | propylene glycol | - |
| T3M01\_BM\_FF | back muscle | flash frozen | - |
| T3M01\_SC\_FF | scales | flash frozen | - |
| T3M02\_BM\_PG | back muscle | propylene glycol | - |
| T3M02\_BM\_FF | back muscle | flash frozen | yes |
| T3M02\_SC\_FF | scales | flash frozen | - |
| 4 | M | T4M01\_BM\_FF | back muscle | flash frozen | - |
| T4M01\_IO\_FF | internal organ | flash frozen | yes |
| T4M01\_SC\_FF | scales | flash frozen | - |

**Table 3.** Table of raw data sequencing statistics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sequencing Technology** | **Sequencing Company** | **Sex** | **Number of Runs** | **Coverage** |
| Linked Reads | 10X | M | 1 | 120 |
| Linked Reads | 10X | F | 1 | 80 |
| Long Reads | PacBio HiFi | M | 2 |  |
| Long Reads | PacBio HiFi | F | 3 |  |
| Hi-C | Phase | F | 1 |  |
| Hi-C | Arima | M | 1 |  |

**Table 4.** Table of assembly steps with corresponding metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Software/Step | Metrics | | Male | Female |
| ipa | assembly metrics | N50 | 353,581 | 418,614 |
| L50 | 324 | 264 |
| # contigs | 2086 | 1805 |
| assembly length | 471,831,811 | 436,920,153 |
| BUSCO | complete | 88.00% | 89.00% |
| single | 79.50% | 87.40% |
| double | 8.50% | 1.60% |
| fragmented | 1.50% | 1.10% |
| scaff10x | assembly metrics | N50 | 1,188,596 | 1,392,224 |
| L50 | 106 | 80 |
| # contigs | 1106 | 1012 |
| assembly length | 471,929,811 | 436,999,453 |
| BUSCO | complete | 88.50% | 85.90% |
| single | 80.50% | 84.40% |
| double | 8.00% | 1.50% |
| fragmented | 1.10% | 3.10% |
| SALSA2 | assembly metrics | N50 | 2,749,144 | 4,383,157 |
| L50 | 38 | 26 |
| # contigs | 705 | 515 |
| assembly length | 472,145,811 | 437,264,453 |
| BUSCO | complete | 88.20% | 89.50% |
| single | 80.50% | 88.00% |
| double | 7.70% | 1.50% |
| fragmented | 1.10% | 0.80% |
| chromonomer | assembly metrics | N50 | 12,200,365 | 14,850,352 |
| L50 | 15 | 13 |
| # contigs | 549 | 376 |
| assembly length | 472,157,411 | 437,273,953 |
| BUSCO | complete | 88.40% | 89.30% |
| single | 81.20% | 87.70% |
| double | 7.20% | 1.60% |
| fragmented | 1.00% | 0.80% |

**Table 5.** Chromosome counts of delta smelt (*Hypomesus transpacficus*) inidicated 2n diploid number is 56

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| **2n Count:** |  | **54** | **55** | **56** | **Total # Cells** |  |
| **# Cells:** |  | 1 | 2 | 15 | 18 |  |
|  |  |  |  |  |  |  |
|  | | | | | |  |

**Table 6.** Distribution of the total samples acquired for effective population size estimations.

|  |  |
| --- | --- |
| **Year** | **Number of Individuals** |
| 1993 | 9 |
| 1995 | 74 |
| 1996 | 111 |
| 1997 | 65 |
| 1998 | 126 |
| 1999 | 45 |
| 2000 | 5 |
| 2002 | 192 |
| 2004 | 192 |
| 2006 | 189 |
| 2007 | 1 |
| 2008 | 191 |
| 2009 | 183 |
| 2010 | 35 |
| 2011 | 192 |
| 2012 | 191 |
| 2013 | 189 |
| 2014 | 217 |
| 2015 | 108 |
| 2016 | 83 |
| 2017 | 167 |
| 2018 | 167 |
| 2019 | 147 |
| 2020 | 66 |

**Table 7.** Genotypes of individuals

**Table 8.** Genotypes of individuals at loci associated with sex.

|  |  |  |
| --- | --- | --- |
|  | lg05\_1885249 | lg05\_1885251 |
| male | GG | GG |
| male | GG | GG |
| male | AA | TT |
| male | GG | GG |
| male | AA | TT |
| female | AA | TT |
| female | GG | GG |
| female | AA | TT |
| female | GG | GG |
| male | GA | GT |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | AA | TT |
| male | GG | GG |
| male | GG | GG |
| female | GA | GT |
| female | AA | TT |
| female | GA | GG |
| female | AA | TT |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| female | AA | TT |
| female | AA | TT |
| female | GA | GT |
| male | AA | TT |
| male | GA | GT |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GA | GT |
| male | GA | GT |
| male | GA | GT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | GG | GG |
| female | AA | TT |
| female | AA | TT |
| female | GA | GT |
| male | GG | GG |
| male | GG | GG |

**Table 9.** Table of the regions in the male genome assembly that contain putative Y sequences.

|  |  |  |
| --- | --- | --- |
| **Chromosome** | **Start** | **End** |
| lg09 | 21021012 | 21058853 |
| lg09 | 21028673 | 21060036 |
| lg09 | 21032415 | 21061475 |
| lg09 | 21041512 | 21085263 |
| lg09 | 21050628 | 21088126 |
| lg09 | 21055337 | 21076797 |
| lg17 | 8138840 | 8173464 |
| lg17 | 8150980 | 8182769 |
| lg17 | 8160327 | 8186257 |
| lg21 | 9824718 | 9865735 |
| lg21 | 9829062 | 9876304 |
| lg21 | 9829267 | 9852703 |
| lg21 | 9830489 | 9861772 |
| lg21 | 9830548 | 9877539 |
| lg21 | 9835162 | 9870510 |
| lg23 | 10574145 | 10610523 |
| scaffold\_9 | 6162370 | 6193388 |
| scaffold\_9 | 6164418 | 6207798 |
| scaffold\_9 | 6166727 | 6189675 |
| scaffold\_9 | 6190157 | 6225909 |
| scaffold\_9 | 6198039 | 6236101 |
| scaffold\_9 | 6207685 | 6244021 |
| scaffold\_9 | 6207917 | 6235633 |
| scaffold\_9 | 6234942 | 6302173 |
| scaffold\_9 | 6273953 | 6369734 |
| scaffold\_9 | 6288767 | 6323264 |
| scaffold\_9 | 6293089 | 6319738 |
| scaffold\_9 | 6308372 | 6343980 |
| scaffold\_9 | 6311690 | 6398470 |
| scaffold\_9 | 6336293 | 6369942 |
| scaffold\_9 | 6342838 | 6380273 |
| scaffold\_9 | 6349134 | 6399774 |
| scaffold\_190 | 85866 | 106849 |

Chart, diagram

Description automatically generated

**Figure 1.** Workflow diagram for completion of assembling reference genome for delta smelt (Task 1). Both male and female delta smelt samples went through the pipeline separately to produce two independent reference genomes––one male and one female. All work was completed at the GVL unless otherwise noted. A) Tissue from male and female delta smelt are sampled from the captive population at the FCCL. B) HMW DNA is extracted from the fish tissue samples. C) Extracted DNA of sufficient length undergoes a library prep, subsequent sequencing and raw data quality control for each of the three NGS technologies (linked-read, long-read and hi-c). D) Long-read sequencing data are assembled into individual draft genome. The draft genome is quality assessed by the software BUSCO13. E) The long-read assembly is scaffolded using linked-read sequencing data to produce a consensus assembly. The quality of the consensus assembly is then assessed by the software BUSCO13. F) Hi-c data connects long-range gaps in the linked- & long-read consensus assembly. The quality of the consensus assembly is then assessed by the software BUSCO13. G) Linkage map data further connects the hi-c, linked- & long-read consensus assembly. The quality of the consensus assembly is then assessed by the software BUSCO13. H) Manual curation to produce a reference genome.

A picture containing text, electronics, computer

Description automatically generated

**Figure 2.** Pulse field gel images of extracted HMW gDNA from three sampling trips (“T”) and six rounds of extractions (“E”). Green boxes surround lanes from extracted samples usable for long-read and linked-read sequencing (extraction distribution centered ~ 50kb), yellow boxes surround lanes from samples with insufficient extraction lengths. A) T1, E1: one usable female sample (T1\_F02); B) T2, E2: no usable samples; C) T2, E3: no usable samples; D) T2, E4: no usable samples; E) T3, E5: one usable male sample (T3M02\_BM\_FF); F) T3, E6: one usable female (T3F02\_SC\_FF).



**Figure 3.** Linked-read k-mer spectra histogram of 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). Uncontaminated samples are expected to have a single peak with a high abundance of k-mers at a very low frequency due to sequencer errors.



**Figure 4.** Plots of k-mer frequency (x-axis) vs GC count (y-axis) colored by the number of distinct k-mers used to detect bacterial and organelle contamination in linked-read sequence data. Blue indicate 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.



**Figure 5.** K-mer comparison plot of the number of distinct k-mers at different frequencies in linked-read sequence data from the 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.

Black dots on a white background

Description automatically generated with medium confidence

**Figure 6.** Karyotype of metaphase stage mitotic cell from a male delta smelt showing 2n = 56 chromosomes. A.) unmodified image, no scale bar; B.) Adobe Photoshop modified image, plus scale bar (most journals want a scale bar; C.) Adobe Photoshop focused image, plus scale bar.

****

**Figure 7.** Multi-dimensional scaling of individual delta smelt from corresponding birth years for hybrid identification analysis.





**Figure 8**. One-generation temporal estimates––with high and low parametric confidence intervals of the of wild delta smelt cohorts from birth year 1995 to 2019 and a PCrit threshold of 0.05––using three different estimators. A) Jorde and Ryman (pink); B) Nei and Tajima (green); C) Pollak (blue); D) Average estimates taken from all three methods of estimating one-generation temporal . Lower estimates of occur more frequently in more recent years and show narrower confidence intervals.



**Figure 9.** All-by-all generation scatter plot of temporal estimates. Points are located at the mean generation of and . For visualization estimates above 10,000 were given a value of 10,000 and negative estimates, which indicate the true is too large to be defined, were given a value of 10,300. Estimations in all three temporal method estimators (Jorde & Ryman, Nei & Tajima, and Pollak) show a broad decline in estimates in more recent generations. This is exhibited by an observed decrease in the number of estimates above 10,000, and decrease in the number of negative estimates in more recent generations.



**Figure 10.** Per site genetic diversity (theta) for birth years from 1995 to 2020 estimated for the normalized number of segregating sites (Watterson’s, ) and average pairwise nucleotide differences (Pi, ) with regression lines and 0.95 confidence intervals. Genetic diversity broadly declines from the early 1990’s to 2020. The more substantial decrease observed in Watterson’s theta over the time period is expected when a population has declined or is currently declining in size because low frequency alleles are rapidly lost from a population. The slight uptick in genetic diversity shown birth years 2018 to 2020 is more likely due to a technical artifact create from a mandatory change in the number of base pairs attached to a given RAD-tag. Previous years (1995 to 2017) had two sets of 100 base pairs associated with each tag, whereas 2018 to 2020 at 150 base pairs associated––an increase of 50%. Thus, more base pair substitutions were likely captured on each tag.

**Figure 11.** Ensi

**Figure 12.** Ensi

**Figure 13** Ensi

**Figure 14.** Ensi

A picture containing tree, day

Description automatically generated

**Figure 15.** Manhattan plots of each of the 28 male chromosomes. Location on the x axis and significance on the y axis. Significant SNPs on Chromosome 5 are marked in blue. This region is worthy of continued exploration as many times significant SNPs will indicate a region is associated with a given trait, but the region may not have adequate coverage. High coverage whole genome resequencing is recommended to better survey the region in question.

Chart, histogram

Description automatically generated

**Figure 16**. Histogram of male-only and female -only k-mer abundances of sequencing data. The male sequencing data appears to have a higher abundance k-mers while the female sequencing data has more lower abundance k-mers. Low abundance sex-specific k-mers are likely the result of sequencing errors, while the higher abundance male-only k-mer peak indicates the male sample contains real variation only contained within the male sample.

Chart, histogram

Description automatically generated

**Figure 17.** All k-mer abundances filtered through contigs containing five or more k-mers to acquire contigs that span roughly 5,000bp. Both female and male have a broad distribution of k-mers with 90-140 abundance, while a male specific peak (in blue) can be seen from 30-70 abundance. This the male-specific peak at roughly half the abundance of the female-specific peak indicates the male genome contains a large amount of sequencing data not contained in the female genome and provide evidence that male delta smelt may be a heterogametic sex.

Chart, scatter chart

Description automatically generated

**Figure 18.** Male (x-axis) versus female (y-axis) median k-mer abundance on contigs with 5 or more k-mers. A) All contigs containing 5 or more k-mers B) Zoomed in view to show clear line of contigs with zero abundance in female sequencing data. There are 40 k-mers that show abundance in males but not females. This indicates the male sequencing data contains sex-specific sequences in high abundance that are not contained in the female sequencing data.