**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 sequencing data & completed genome** | | |
| **Task 2.1 DNA collection and extraction** | | |
| a. Sample Collection | Collect samples from 2017, 2018 and 2019 birth year cohorts | Completed |
| b. DNA Extraction | Extract DNA from 2017, 2018 and 2019 birth year cohorts | Completed |
| c. RAD sequencing | Sequence 2017, 2018 and 2019 birth year cohorts | Completed |
| **Task 2.2 Estimate Ne in 2017, 2018 and 2019 birth year cohorts** | | |
| a. RAD sequencing data processing | Split, align and quality control all sequencing data | Completed |
| b. Estimation of | Temporal | Completed |
| c. Quantify Genetic Diversity | Theta estimates | Completed |
| **Task 3. Examining domestication selection in the conservation hatchery for delta smelt** | | |
| **Task 3.1 Select individuals for RAD sequencing 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, alignment to the reference genome | Completed |
| a. Identifying regions under domestication selection | Genotype calling, FST analysis between groups | Completed |
| d. Association study | GWAS between DI and reproductive process | N/A |
| **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 four 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 | 35 |
| Long Reads | PacBio HiFi | F | 3 | 29 |
| Hi-C | Phase | F | 1 | n/a |
| Hi-C | Arima | M | 1 | n/a |

**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.** The chromosome counts for delta smelt during the karyotyping experiment indicate that the 2n diploid number in delta smelt is 56. This means that delta smelt have two sets of 28 unique chromosomes. See Figure 6.

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

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

**Table 6.** List of outlier individuals, arranged by birth year, that did not pass filtering and quality control. Outliers listed could be hybrids or data could contain a technical artifact from sequencing and we’re identified as having a PC1 > 0.01 and/or PC2 < -0.01.

|  |
| --- |
| Individual ID |
| Ht02-42\_1995 |
| Ht02-85\_1996 |
| Ht02-86\_1996 |
| Ht01-51\_1998 |
| Ht03-73\_1998 |
| Ht03-76\_1998 |
| Ht03-77\_1998 |
| Ht03-82\_1998 |
| Ht04-91\_2000 |
| Ht11-13\_2006 |
| Ht20-16\_2012 |
| Ht21-31\_2013 |
| Ht21-39\_2013 |
| Ht21-40\_2013 |
| Ht33-39\_2016 |
| Ht41-92\_2018 |
| Ht42-06\_2018 |
| Ht42-40\_2018 |
| Ht42-22\_2019 |

**Table 7.** 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 8.** Comparison of temporal estimates from our study and Finger *et al.* (2017) using two different estimation methods (Nei & Tajima and Pollack) for wild delta smelt from 2011-2013. To provide estimates that can be directly compared, we used a critical value threshold (Pcrit) of 0.02 and reported the corresponding 95% parametric confidence intervals for birth year cohorts found in Finger *et al.* (2017). All alleles with a frequency lower than 0.02 were removed from analysis to produce an unbiased estimate of as replicated from the Finger *et al.* (2017) study. Our study used 911 loci found throughout the genome and the Finger *et al.* (2017) study used 12 microsatellites.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Sample Size | | Nei & Tajima56 *NE* | | Pollack57 *NE* | |
|  | **Finger *et al.* 2017** | **Current Study** | **Finger *et al.* 2017** | **Current Study** | **Finger *et al.* 2017** | **Current Study** |
| 2011 | 995 | 192 | 2,363 (775-∞) | 876 (190-∞) | 2,088 (735-∞) | 630 (176-∞) |
| 2012 | 534 | 191 | 3,454 (789-∞) | 321 (134-∞) | 3,259 (775-∞) | 285 (129-∞) |
| 2013 | 678 | 189 | 2,459 (634-∞) | 183 (99-685) | 2,340 (632-∞) | 169 (96-566) |

**Table 9.** Number of individuals before and after filtration based on their aligned-read quality. The low and medium groups had many individuals removed after filtration, requiring us to standardize the number of aligned reads across groups.

|  |  |  |
| --- | --- | --- |
| Group | Number of samples before filtration | Number of samples after filtration |
| Wild | 48 | 48 |
| Low DI | 48 | 19 |
| Medium DI | 96 | 34 |
| High DI | 48 | 22 |

**Table 10.** Genotypes of individuals at the two loci on Chromosome 5 (Chr05) associated with sex. Color coded sex assignments (female in green and male in red). Female-associated homozygous genotypes are colored in green, heterozygous genotypes are colored in yellow, and male genotypes are colored in red. Diagnostic loci would perfectly correlate at least one allele with sex at a given locus. Here, neither Chr05\_1885249 or Chr05\_1885251 contain alleles that are only found in one sex.

|  |  |  |
| --- | --- | --- |
| Sex | Locus | |
| Chr05\_1885249 | Chr05\_1885251 |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | AA | TT |
| female | GA | GT |
| female | GA | GT |
| female | GA | GT |
| female | GA | GG |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| female | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GG | GG |
| male | GA | GT |
| male | GA | GT |
| male | GA | GT |
| male | GA | GT |
| male | GA | GT |
| male | AA | TT |
| male | AA | TT |
| male | AA | TT |
| male | AA | TT |

**Table 11.** Regions in the male genome assembly that contain putative Y sequences. Large segments of sequence are contained on multiple chromosomes, indicating sex determination in delta smelt may be polygenic.

|  |  |  |
| --- | --- | --- |
| **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 BUSCO26. 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 BUSCO. 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 BUSCO. 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 BUSCO. 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. Outlying data points (PC1 > 0.01 and PC2<-0.01) show a large degree of variation not contained in the more clustered samples. This indicates that these data points are either hybrids or the result of a sequencing artifact.



**Figure 8**. One-generation temporal estimates––with high and low jackknife confidence intervals of the of wild delta smelt cohorts from birth year 1995 to 2019 and a critical value threshold of 0.05. All alleles with a lower frequency that 0.05 were removed from this analysis. We used three different estimators. A) Jorde & Ryman (pink); B) Nei & Tajima (green); C) Pollak (blue); D) Average stimates taken from all three methods of estimating one-generation temporal . For ease of visualization, estimates above 5,000 were given a value of 5,000, and negative estimates (which indicate the true is too large to be defined) were given a value of 5,500. 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 ease of 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 from all three temporal estimators (Jorde & Ryman, Nei & Tajima, and Pollak) show a decline in in more recent years (roughly after 2013). After 2013 one can observe fewer estimates >10,000, and more estimates centering around or below 1,000. This analysis also indicates that the estimators give more reliable estimates when NE is smaller, which has been documented extensively.



**Figure 10.** Per site genetic diversity (theta, ) for birth years from 1995 to 2020 with regression lines and 0.95 confidence intervals. We used two theta estimates: normalized number of segregating sites (Watterson, ), and average pairwise nucleotide differences (Pi, ). Here we see that genetic diversity as measured by and decline from 1995 to 2020. The more substantial decrease observed in is expected when a population has declined or is currently declining in size, because low frequency alleles are rapidly lost from a population.

*Diagram

Description automatically generated*

**Figure 11.** Experimental design. DI = Domestication Index. Green boxes show the four groups based on their DI value: low DI (0 < DI ≤ 2, 48 individuals), medium DI (3 ≤ DI ≤ 4, 96 individuals), high DI (DI≥7, 48 individuals), wild (DI = 0, 48 individuals). Pink boxes show two groups based on their recovery rate within the medium DI group: high offspring recovery rate (≥ 0.7, 48 individuals) and low offspring recovery rate (≤ 0.3, 48 individuals).

Chart, scatter chart

Description automatically generated

**Figure 12.** Distribution of individuals in the four groups based on their aligned-read and genotype call count (individuals with lower number of aligned-read count have lower number of genotype counts). Each dot represents an individual and color represents their group. X-axis is the number of aligned-read and y-axis is the number of genotype calls for each individual. Based on the distribution, an aligned-read count per individual ranging between 3,000,000bp to 6,000,000bp was chosen.

**Chart, timeline

Description automatically generated**

**Figure 13.** Pairwise FST sliding window analysis. X-axis: position on a chromosome, Y-axis: FST value of each window. Each dot represents a window and a red dot represents significant windows (p≤0.001). A) comparison 1 between wild/low DI group vs. medium/high DI group. Elevated regions on Chromosomes 15 and 22 are significant (lg 15 and lg 22). B) comparison 2 between wild group and hatchery (low/medium/high DI) group. Elevated regions on Chromosomes 4, 10, 13, 15, 16, 19, 22, 23, and 26 are significant. Elevated regions on Chromosomes 15 and 22 present in both comparisons are more likely to be under domestication selection.

Graphical user interface, chart, histogram

Description automatically generated

**Figure 14.** Higher resolution of pairwise FST sliding window analysis of the candidate regions in the two comparisons. X-axis: position on a chromosome, Y-axis: FST value of each window. Red represent comparison 1 and blue represents comparison 2. A) distribution of windows on Chromosome 15 (lg15). B) distribution of windows on Chromosome 22 (lg22). The peaks are located near to the edge of the chromosomes and window distribution follows the same pattern in both comparisons. This strengthens the probability of the region being under selection.

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 sequencing coverage. High coverage whole genome resequencing is recommended to better survey the region in question.