fodder\_doc

Intro

Genetic underpinnings may contribute to any aspect of life history, ecology and physiology. Therefore developing genomic resources and maintenance of genetic diversity within the species are key components to

genetic diversity = closely tied to the evolutionary capacity for adaptation to environmental change

phenotypic variation to allow for a broad respond to climatic factors deeply influencing the Delta. Three key components to

of basic ecology is phenotypic variation; thus, maintenance of genetic diversity is essential as it is tied to the evolutionary capacity for adaptation to environmental change. To assess genetic diversity within delta smelt, contemporary and historical effective population size estimates were made using 2318[tk] samples from 25[tk] generations.

Conservation of genetic diversity within the species will require several components, including an assembled genome to allow for fine-scale genomic change, quantifying genetic diversity through time, and identifying loci to assist in non-lethal monitoring methods of population dynamics, such as sex ratios, of the species.

genetic resources for and monitoring genetic diversity within delta smelt is a critical step to aid in research efforts for making informed decisions to manage this imperiled species. Resource development requires genomic tools for

The abundance of delta smelt (*Hypomesus transpacificus*) has been in decline for decades1,2. Genetic tools have been useful for several management concerns: monitoring overall diversity and effective population size () in the wild population, genetic management in the captive population, developing assays to detect hybrids between delta smelt, wakasagi, and longfin smelt, and developing assays for eDNA sampling or species identification3–5. Attempts at quantifying in the wild population have been hampered by several interacting factors including a lack of power from using microsatellite markers and a very large historical 6,7. In order to improve estimates of , we assembled the delta smelt genome. We then leveraged the genome to estimate contemporary and the loss of genome-wide diversity between 1993 and 2020, interrogate domestication selection, and search for a sex marker. While is useful for monitoring purposes, we strongly advice against the use of alone for making management decisions.

For this project our tasks were to: 1) assemble a high-quality reference genome for delta smelt that is publicly available for all researchers, 2) estimate contemporary effective population sizes () for the wild 2017 to 2019 birth year cohorts of delta smelt, 3) search for genetic evidence of domestication selection across the genome, and 4) search for sex-specific markers and/or chromosomes in delta smelt. In order to obtain a more comprehensive understanding of through time, we have expanded Task 2 to include analysis of historical datasets of delta smelt dating back to 1995 and samples collected from 2020. The status of each deliverable is listed in Table 1. We have divided this final report into 4 sections, each with its own background, methods, results and discussion, followed by a final conclusion integrating our findings.

GENOME

Genome Intro: Why?

* innovation (theoretical or medical)
* preserve biodiversity
  + species are going extinct faster than ever
  + 5th great extinction?
  + need to know the genetic composition of the speices that sustain ecological communities
  + “the genome is the core entitiy to life”
  + previously focused on a single gene approach but genes do not exist in isolation,
  + genetics have influenced
  + “genomics starts from a reference genome”
  + model genomic research has led to innumerous medical advances for humans
  + humans are radically altering the natural world
    - we need to understand the effects of this
  + can understand what genetics lead to making a mammal a mammal and a reptile a reptile, or bats live for fucking ever
  + from just one mammal that currently doesn’t ha
    - know about blindness,
  + disease susceptibility

What could happen?

* scientific innovation
* theories stemming from observations intangible to the naked eye
* contributing to a larger and ever growing resource
* can better understand things like aging (insert bats),

Genome Intro: old

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).

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.

genome assembly contiguity can be further increased by incorporating linkage maps, which uses the frequency of recombination between markers to generate relative distances throughout the genome.

METHODS

*Sample Collection*

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 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.

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.

*Hi-C chromatin conformation capture prep & sequencing*

We received sequence files with 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).

CHAPTER 3

Prior analyses that attempted to determine sex markers in delta smelt used the *Sbf1* restriction enzyme, which cuts DNA approximately every 65,000 base pairs, but no sex markers were identified. For this library preparation we sought to maximize the number of restriction enzyme cut sites and acquire reads from more locations throughout the genome. To do this we digested extracted DNA using the *Pst1* restriction enzyme, which shears DNA sixteen times more often than the *Sbf1* restriction enzyme, or roughly once every 4,100 base pairs, providing more coverage of the genome than Sbf1.

Next, we looked for the association of alleles at any location in the genome with sex classification by carrying out a dominant model case-control GWAS using the program ANGSD v0.921 (Korneliussen et al., 2014) using sex as the case-control classification.