**Sequencing the Delta Smelt genome: improved annual monitoring of Ne and further understanding of the wild and hatchery Delta Smelt populations**

June 30, 2019 Annual Report

Shannon E.K. Joslin, Ensieh Habibi, and Amanda J. Finger

University of California, Davis

This document will serve as the 2019 annual report for A19-1844 contract between State Water Contractors and the Regents of the University of California.

**Introduction**

Recently, genome sequencing technologies have become both more cost effective and efficient. The availability of a high-quality reference genome, hereafter called reference genome, and use of next generation sequencing (NGS) technologies in population genetic studies increase the power and precision of inferences made on populations of interest.1 The power and precision of genetic analyses are increased through knowledge of locations and linkage groups throughout the genome as well as surveying a high density of markers that are spread throughout the genome.2,3

The primary purpose of this project is to sequence and assemble a reference genome for the federally threatened Delta Smelt, provide a reference genome as a resource for genetic work on Osmerids, and use the reference genome to better inform select downstream studies on the wild population and the captive population at the Fish Conservation and Culture population (FCCL) in Discovery Bay, CA. More specifically, we will use the reference genome for 1) more precise estimation of contemporary effective population sizes (NE) for the 2017 to 2019 wild cohorts, 2) an examination of domestication selection at discrete locations in the genome, and 3) identification of sex-specific markers and/or chromosomes in Delta Smelt. The creation of a reference genome for Delta Smelt will enable scientists to more fully take advantage of the decreasing samples available and use archival samples more effectively.

**Methods, Preliminary Results, and Next Steps**

**Task 1: Genome assembly**

A reference genome is a near-complete chromosome-scale scaffolding of the genomic DNA found within a species of interest. It has a high accuracy of base calls, correctly oriented DNA segments (contigs), and contains few gaps. Currently, there is no Osmeridae reference genome available to use in the studies of wild and captive Delta Smelt. By knowing the detailed genomic architecture of a Delta Smelt we can identify regions of the genome that have undergone or are currently under selection, identify genomic regions specific to sexes, carry out association studies, define specific haplotypes, and avoid false-positive findings in misassembled regions.4 We will be utilizing 10X Genomics pseudo-long read sequencing and Hi-C Chromatin Conformation Capture to generate a reference genome for the species.

*DNA Sampling*

To construct a reference genome, high molecular weight (HMW) genomic DNA (gDNA) was extracted from sampled specimens. Since the mechanism for sex determination in Delta Smelt is unknown, one male and one female fish from the FCCL captive population were sampled for extraction and subsequent sequencing. After collection, samples were brought to the UC Davis DNA Technologies Core for HMW gDNA extraction and sequencing. HMW gDNA was extracted the protocol described in Wasko et al. (2003)5 and the size range of extracted DNA fragments were determined using a pulse field gel (Figure 1). Genomic DNA was adjusted to a concentration of 0.91 ng/µl, and 1.14 ng of template gDNA was loaded on a Chromium Genome Chip for subsequent sequencing.

*10X Genomics Sequencing*

10X library preparation takes extracted HMW gDNA, shears the DNA into 50kb long segments, inserts each fragment into an oil coated gel emulsion bead (GEM), further shears DNA within each bead, and attaches unique barcodes 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 with one modification. The library was sequenced on a NovaSeq6000 sequencer (Illumina, San Diego, CA) with paired-end 150 bp reads to an estimated depth of 120x for a 1.2Gb genome calculated from the number of *Sbf1* derived loci generated from previous restriction site associated (RAD) sequencing data.

*Post-sequencing Quality control*

After sequencing, bioinformatics quality control steps are necessary at multiple stages of assembling reference genomes and working with NGS data in general. This is because 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. We conducted three bioinformatics quality control steps to examine contamination and look for sequence bias. All bioinformatics work was conducted on the UC Davis farm compute cluster (the farm).

We used the software program kat6 for our initial QC steps. First, to see if there is evidence of contamination, we observed the number of distinct k-mers at different frequencies for k-mers of length 16, 21, 27 and 33 (Figure 1A-D). We found single peaks for each histogram indicating contamination is unlikely. Following this step, to further look for signs of contamination, we plotted observed GC counts against the frequency of k-mers of length 16, 21, 27 and 33. All plots had a single peak of distinct k-mers, meaning there does not appear to be signs of bacterial or organelle contamination (Figure 2A-D). Finally, to observe if there was sequencing bias in the forward (R1) or reverse (R2) read sequences, we plotted the number of distinct k-mers at different frequencies for the R1 and R2 files against one another and does not appear to have major sources of sequencing bias (Figure 3A-D).

*10X Draft Assembly*

An initial 10X genome assembly will string together raw Illumina sequencing reads based on the degree of nucleotide overlap, which will produce a draft genome. Draft genomes have sequences of contiguous known bases (contigs) and sequences of contigs known to associate but in an unknown order (scaffolds). We used the software Supernova to assemble raw 10X sequence reads into a draft genome comprised of contigs and scaffolds.7 Currently, Supernova is optimized to assemble diploid human genomes at 38-56x coverage with higher coverage occasionally being advantageous in non-human species. In order to assemble reads at roughly 60x coverage, we randomly divided raw fastq files in half using the Proc10xG python scripts developed by the UC Davis Bioinformatics Core (<https://github.com/ucdavis-bioinformatics/proc10xG>), and custom bash scripts. After sequence data division, two independent assemblies (A and B), were run at estimated sequencing depths of 60x coverage per assembly. Assembly results were measured in the contig and scaffold parameter N50, where the N50 is the size of the contig or scaffold which 50% of the assembly is of that size or larger. Our initial 10X assembly showed an estimated genome size of 0.9 Gb with read coverage of ~91x in both assemblies. Assembly A produced a contig N50 of 20.11Kb and a scaffold N50 of 1.17Mb. Assembly B produced a contig N50 of 20.05Kb and a scaffold N50 of 1.59Mb. Both assemblies resulted in scaffold N50s over 1Mb, indicating the smallest size of the scaffold containing half of the amount of assembly sequence data is over 1Mb, so we can move on to hi-c chromatin conformation capture.

*Phase Genomics Hi-C Chromatin Conformation Capture*

Due to the highly repetitive nature of genomic DNA in eukaryotes the initial 10X draft assembly is expected to have large, unresolved gaps surrounding repetitive areas such as the centromere. In order to resolve these gaps, we will use hi-c chromatin conformation capture. Tissue from the same female individual used for 10X sequencing was sent to Phase Genomics in Seattle, WA for Hi-C Chromatin Conformation Capture library prep and subsequent deep sequencing.

*Next steps*

After receiving sequencing data from Phase Genomics (expected in July 2020), we will use the software LACHESIS8 to perform proximity guided assembly to scaffold 10X derived contigs into chromosome length scaffolds. After, we will run BUSCO9 to assess the completeness of the assembly.

**Task 2: Estimation of Effective Population Size**

*DNA Sampling*

We have received adipose fin clips of wild BY2017 and BY2018 cohorts. Currently, the samples are being stored at the Genomic Variation Lab (GVL).

*Next Steps*

In order to reduce batch effects between years, BY2017 and BY 2018 samples will be stored until receiving BY2019 cohort samples. Upon receiving BY2019 samples DNA will be extracted and sequenced according to Ali et al. (2016).10

**Task 3: Examining domestication selection in the conservation hatchery for Delta Smelt**

Delta Smelt have been cultivated at the FCCL, a conservation hatchery since 2008. The pedigree-based management applied at the FCCL aims to both minimize average co-ancestry in the refuge population and maintain the genetic diversity of the captive population similar to the wild population. Despite this intense management, Finger et al. (2018)11 showed that there is a strong evidence of genetic adaptation to captivity.

To further investigate domestication selection, we used restriction-site associated DNA (RAD) sequencing to collect genomic data on archived FCCL samples. First, we grouped 240 individuals for sequencing based on their domestication index (levels of hatchery ancestry; DI) and recovery rate. We selected wild, early, medium, and late generation fish with low and high recovery rates to represent varying levels of hatchery ancestry. Next, we prepared libraries for RAD sequencing using 100ng DNA from each individual. RAD library preparations were performed with the *Pst1* enzyme based on the protocol described in Ali et al. (2016).10 All libraries were pooled into a single lane for paired end 150bp sequencing on an Illumina HiSeq4000.

*RAD sequencing Data Processing*

After sequencing and quality filtering, we de-multiplexed sequences for each individual.

We produced a RAD sequencing derived partial assembly (RAD assembly) using the Novoalign (2010, http:/www.novocraft.com/ ) and PRICE12 programs. We then aligned sequences to the RAD assembly using BWA software13. The output files from the alignments were Sequence Alignment Map (SAM) files, which were then converted to Binary Alignment Map (BAM) files using SAMtools.14 SAMtools was also used to sort, filter for proper pairs, remove PCR duplicates, and index the BAM files.

After the alignment, we began our preliminary analyses. All population genetic analyses were conducted in ANGSD15 which analyzes raw data based on a probabilistic framework in the form of Genotype Likelihoods (GL). For the analyses, we used the SAMtools genotype likelihood model (-GL 1) with a minimum base quality of 20 (-minQ 20) and minimum mapping quality of 20 (-minMapQ 20).

After genotype analysis, we detected SNPs associated with domestication selection. To do this, first, we performed Fst analysis and observed 90 outliers SNPs on 11 contigs (RADtag) with the highest Fst value (Figure 5). After Fst analysis, we looked at allele frequency differences of the 90 SNPs in each group and compared them with each other. The result shows a large shift in allele frequencies between wild group and the three hatchery groups (Figure 6). These allele frequency differences are evidence that domestication selection may have occurred during early generations.

*Assessing Relatedness Within Wild Group*

We conducted an additional analysis to evaluate the level of relatedness within the wild population compared to low, medium and high DI sample groups using the program NgsRelate,16 which estimates maximum likelihood of pairwise relatedness based on genotype likelihoods. The NgsRelate analysis results of the four groups were evaluated based on a K0 value which represents the fraction of genome in which two individuals share 0 alleles identity-by-descent (Figure 7). The K0 value ranges from 0 to 1, and 1 means unrelated.

The wild group had a K0 above 0.8, which is high enough to be considered a low level of kinship among individuals. Surprisingly, the low and medium DI groups show a lower level of relatedness (average K0= 0.92) compared to the wild group. This may be due to the genetic management at the FCCL.

*Next Steps*

After the reference genome is assembled, we will reanalyze our data with the new 10X and Hi-C derived reference genome. Although we expect to have the same or only slightly different candidate loci, we will be able to evaluate linkage disequilibrium across all SNPS.

Using the reference genome, we will also run a genome-wide association study (GWAS) to assess association of our candidate loci with the recovery rate at the FCCL. If there is an association between domestication selection and recovery rate, we expect to observe the same loci as the domestication selection candidate loci.

**Task 4: Sex Marker**

*DNA Sampling*

For the identification of sex specific markers, we have already sampled 24 female and 24 male individuals sexually identified through either dissection or gametic expression from the FCCL. We have extracted DNA and RAD sequenced sampled individuals according to Ali et al. (2016).10 To maximize the number of restriction enzyme cut sites to produce more data, we digested DNA using the *Pst1* restriction enzyme, which shears DNA sixteen times as often as the *Sbf1* restriction enzyme used in previous analyses. RAD sequencing libraries were prepared at the GVL and sequenced at the UC Davis Sequencing Center with 150 bp paired-end reads on an Illumina HiSeq. Sequencing data was uploaded to the farm compute cluster where all bioinformatic work has taken place. After sequencing, we de-multiplexed individual sequences using a custom perl scripts (<https://github.com/shannonekj/sexMarker_Hypomesus-transpacificus>).

*Next Steps*

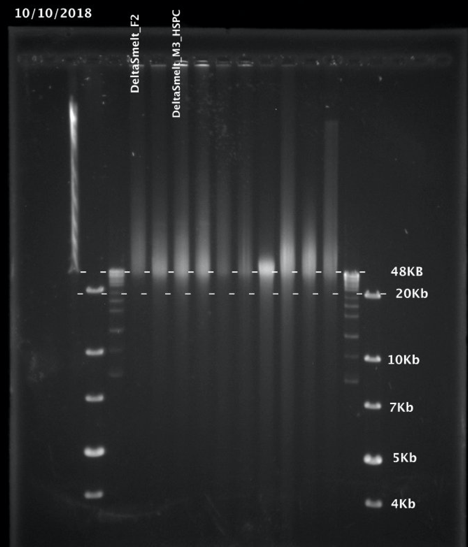
We will align, filter and test for population structure using the reference genome, as we have described in Task 2. We will then carry out a case-control GWAS using the software PLINK.17 We will correct for multiple testing by determining significance using a Benjamini-Hochberg false discovery rate18 and permutation testing through PLINK.

Sequences containing sex associated SNPs will be submitted to Fluidigm for SNP Type assay design as described in Benjamin et al. (2018).19 SNPs will be validated using an additional 24 females and 24 males sexed through either dissection or gametic expression, which were not included in the initial dataset.

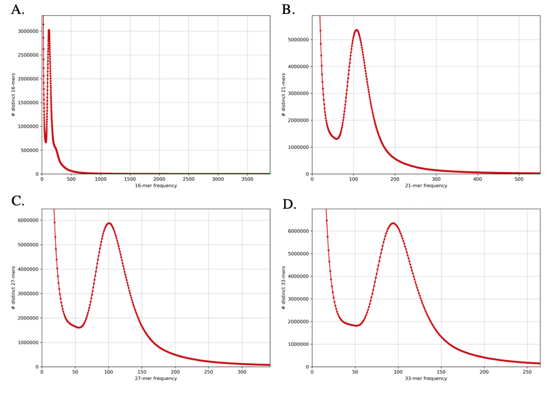
**Tables and Figures**

Table 1. List and description of deliverables with the status of each task.

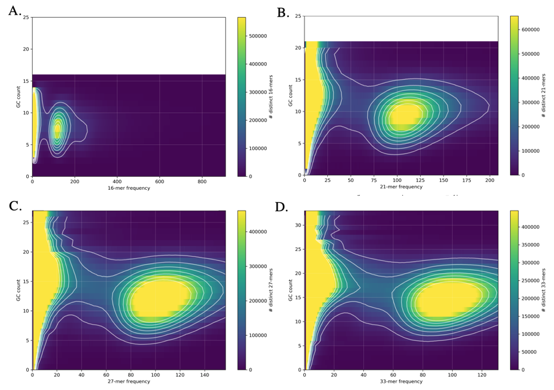
|  |  |  |
| --- | --- | --- |
| **Deliverable** | **Description** | **Status** |
| **Task 1. Sequence and assemble a draft Delta Smelt genome + linkage map** | | |
| **Task 1.1 10X Genomics Sequencing** | | |
| a. Sample acquisition | Sample tissue of male (M) and female (F) fish. | Completed |
| b. Sequencing | 10X Chromium Library Prep | F completed,  M in progress |
| Illumina NovaSeq Sequencing | F completed,  M in progress |
| c. Assembly | Supernova assembly | F completed,  M in progress |
| Assembly QC | In progress |
| tigmint assembly correction | In progress |
| **Task 1.2 PacBio Sequencing** | | This step is not necessary at this time. |
| a. Sample preparation | Sample tissue and organs of male and female fish. | –– |
| b. Sequencing | PacBio Prep | –– |
| Sequel Sequencing | –– |
| c. Assembly | FALCON assembly | –– |
| tigmint assembly correction | –– |
| **Task 1.3 Hi-C Chromatin Conformation Capture & Sequencing** | | |
| a. Sample preparation | Sample tissue and organs of male and female fish. | F completed,  M in progress |
| b. Sequencing | Hi-C Library Prep | In progress |
| Deep Sequencing | In progress |
| c. Assembly | LACHESIS scaffolding | In progress |
| minimap & miniasm assembly software | In progress |
| **Task 1.4 High-density linkage map** | | |
| a. SNP Discovery | Sample tissue and organs of male and female fish. | In progress |
| b. Construction | Construction of a linkage map using HighMap | In progress |
| **Task 1.5 Map linked regions to reference genome** | | |
| a. Map regions | Align linked regions to genome | In progress |
|  |  |  |
| **Task 2. Estimate contemporary Ne using RAD-seq data & draft genome** | | |
| **Task 2.1 DNA collection and extraction** | | |
| a. Sample Collection | Collect samples from BY2017, BY2018 and BY2019 cohorts | BY2017 & BY2018 completed  BY2019 in progress |
| b. DNA Extraction | Extract DNA from BY2017, BY2018 and BY2019 cohorts | In progress |
| c. RAD-sequencing | Sequence BY2017, BY2018 and BY2019 cohorts | In progress |
| **Task 2.2 Estimate Ne in BY2017, BY2018 and BY2019 cohorts** | | |
| a. RAD-sequencing data processing | Split, align and quality control all sequencing data | In progress |
| b. Estimation of Ne | Temporal Ne  LD Ne | In progress |
|  |  |  |
| **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 | a. Determine domestication index (DI) |
| b. Group individuals based on their DI | Assign individuals to three DI group: Low, medium and high | b. Group individuals based on their DI |
| c. Determine recovery rate of each individual | Calculate recovery rate of each individual | c. Determine recovery rate of each individual |
| d. Group individuals based on their recovery rate | Assign individuals to two groups: Low and high recovery rate | d. Group individuals based on their recovery rate |
| **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 | In progress |
| d. Association study | GWAS between DI and recovery rate | In progress |
| **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 | In progress |
| b. Identification of sex marker(s) | GWAS in M vs F | In progress |
| **Task 4.2 Fluidigm assay creation** | | |
| a. Submit sequences | Submit sex specific sequences to Fluidigm for SNP Type Assay design | In progress |
| b. Validation | Validation of SNP Assay | In progress |



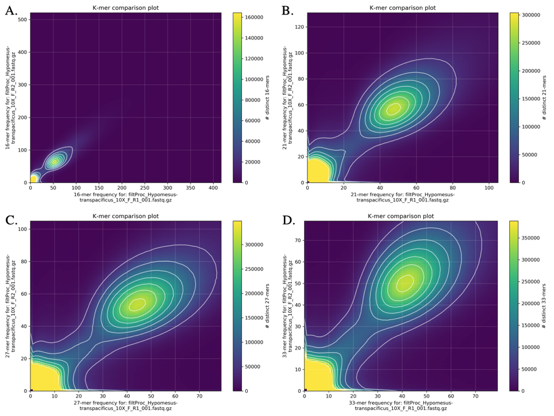
**Figure 1.** A pulse field gel image of the extracted HMW gDNA from female (DeltaSmelt\_F2) and male (DeltaSmelt\_M2\_HSPC) samples. Extracted DNA from the female sample lies above the 48Kb indicating many extracted fragments are larger than 48Kb. However, the male sample contains a larger proportion of DNA below the 48Kb threshold.



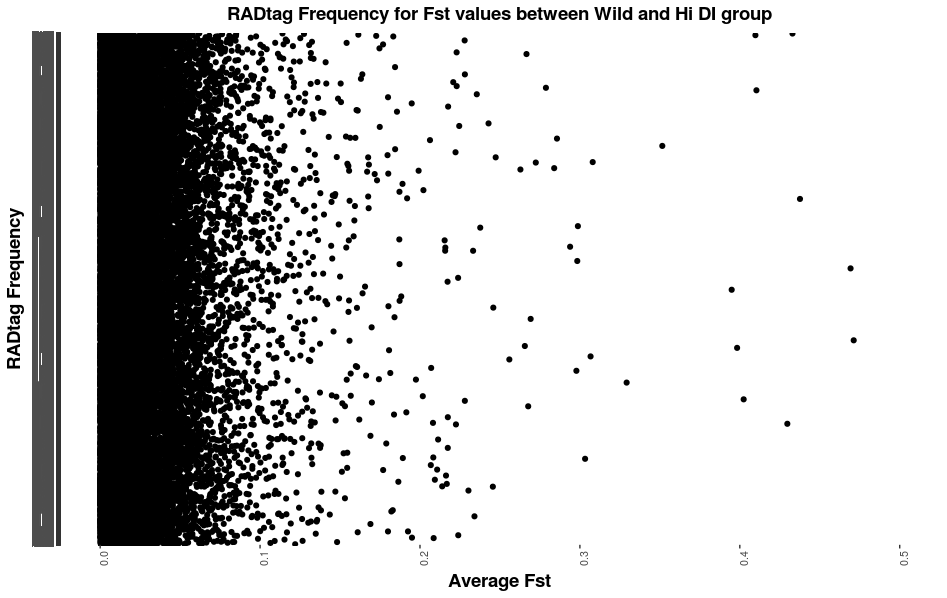
**Figure 2.** K-mer histogram of the number of distinct k-mers at particular frequencies in the genome for: A.) 16-mer, B.) 21-mer, C.) 27-mer, and D.) 33-mer.



**Figure 3.** A GC counts plot for: A.) 16-mer, B.) 21-mer, C.) 27-mer, and D.) 33-mer used to detect bacterial and organelle contamination. No indication of contamination was detected.

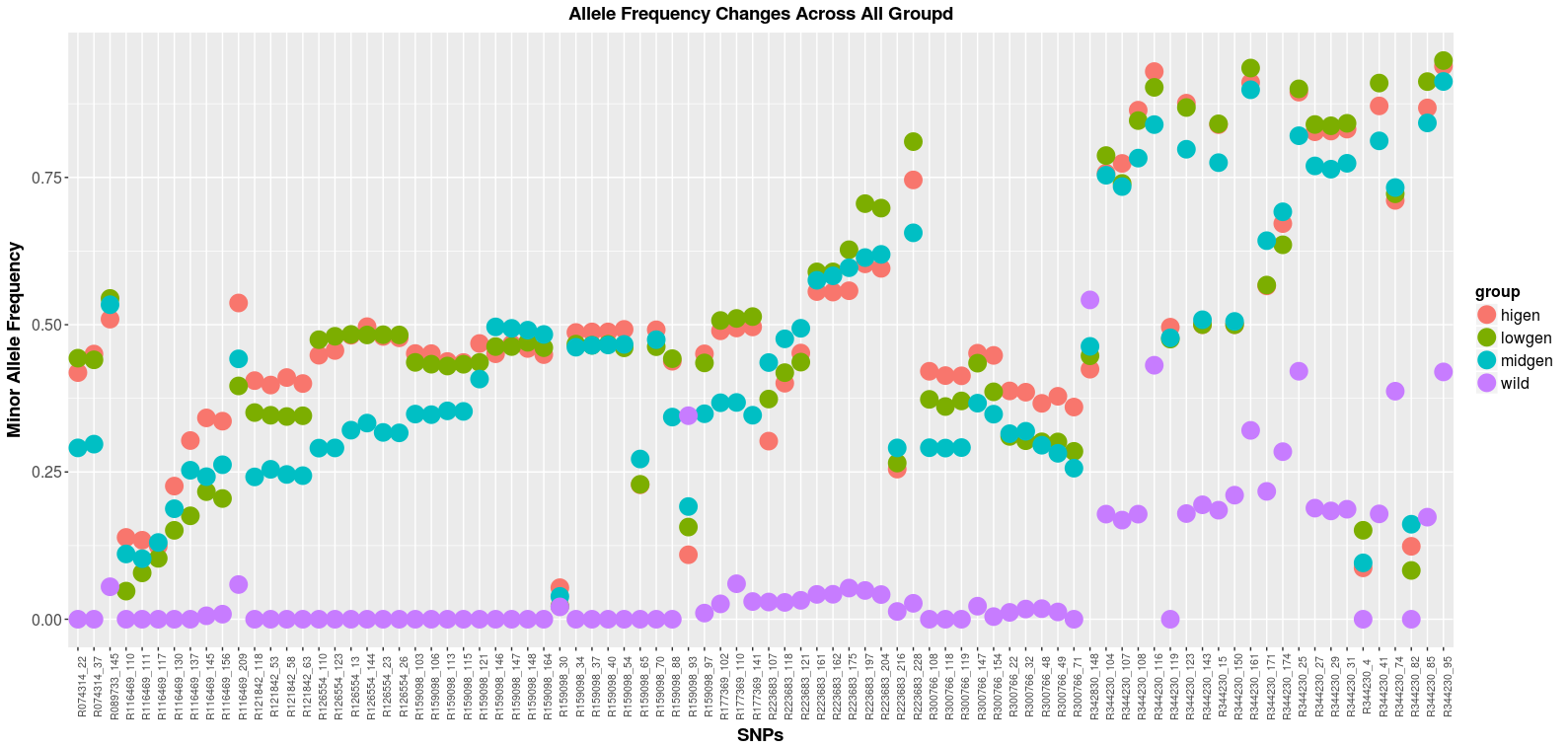


**Figure 4.** A K-mer comparison plot of the number of distinct k-mers at different k-mer frequencies for k-mers of size: A.) 16-mer, B.) 21-mer, C.) 27-mer, and D.) 33-mer. For all plots the R1 (x-axis) and R2 (y-axis) captures a slightly different information.

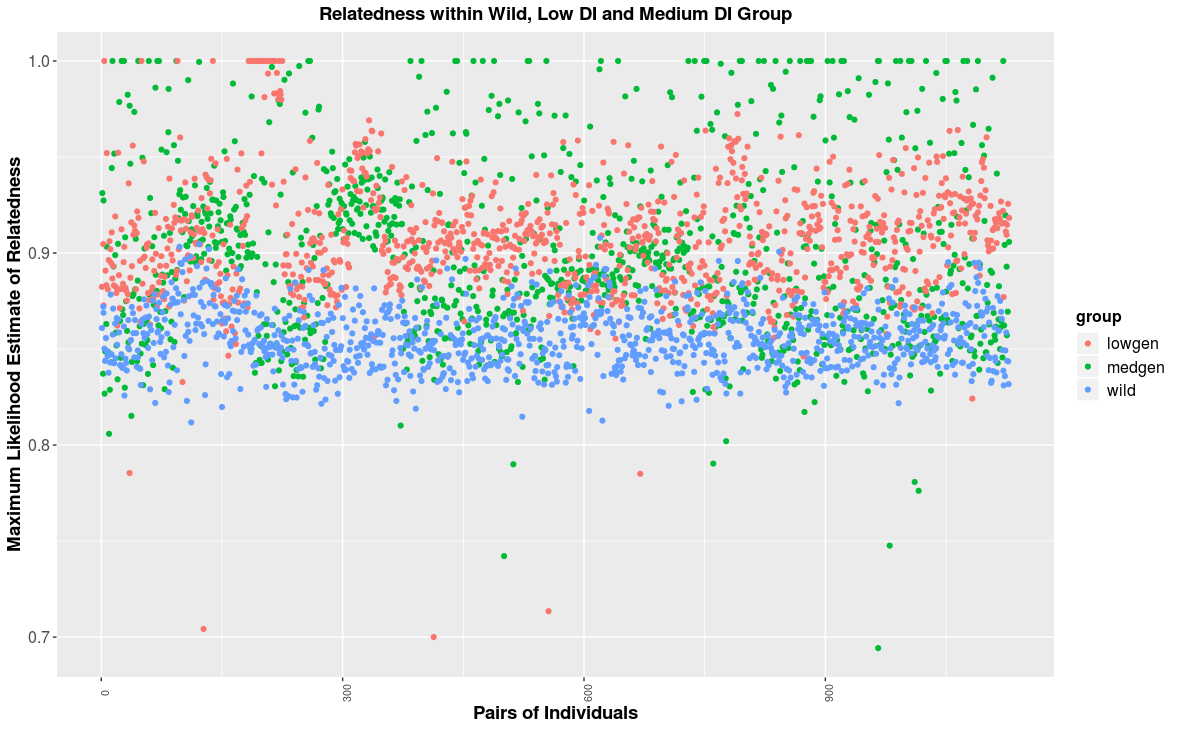


**Figure 5.** The figure shows the frequency of RADtag for each Fst value. Each dot is a RADtag.

The top 11 RADtag are selected (Fst value between 0.4 - 0.5) for outlier SNPs.



**Figure 6.** Changes in the allele frequency across wild (purple), high (pink), medium (blue), and low (green) DI groups. The X-axis and Y-axis represent SNPs and Minor Allele Frequency, respectively. Each dot represents frequency of allele in each group. There is a large difference in the allele frequency between wild and hatchery groups. The large shift between wild and low DI group shows domestication selection occurred in the early generations.



**Figure 7.** Relatedness within each of the Wild (blue), Low DI (pink), and Medium DI (green) groups. Each dot represents pairwise comparison of individuals within each group. Y-axis is maximum likelihood of the relatedness (K0). K0 equals 1 means no relatedness. The high K0 value of each group (> 0.8) means the relatedness among individuals in each group is low.

**Works Cited**

1. Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. *Nat Rev Genet*. 2010. doi:10.1038/nrg2844.

2. Steiner CC, Putnam AS, Hoeck PEA, Ryder OA. Conservation Genomics of Threatened Animal Species. *Annu Rev Anim Biosci*. 2013;1(1):261-281. doi:10.1146/annurev-animal-031412-103636.

3. Waples RS. Making sense of genetic estimates of effective population size. *Mol Ecol*. 2016;25(19):4689-4691. doi:10.1111/mec.13814.

4. Carbonell-Caballero J, Amadoz A, Alonso R, et al. Reference genome assessment from a population scale perspective: An accurate profile of variability and noise. *Bioinformatics*. 2017;33(22):3511-3517. doi:10.1093/bioinformatics/btx482.

5. Wasko AP, Martins C, Oliveira C, Foresti F. Non-destructive genetic sampling in fish. An improved method for DNA extraction from fish fins and scales. *Hereditas*. 2003;138(3):161-165. doi:10.1034/j.1601-5223.2003.01503.x.

6. Mapleson D, Accinelli GG, Kettleborough G, Wright J, Clavijo BJ. Sequence analysis KAT : a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. 2017;33(November 2016):574-576. doi:10.1093/bioinformatics/btw663.

7. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Corrigendum: Direct determination of diploid genome sequences. *Genome Res*. 2017;27(4):757-767. doi:10.1101/gr.235812.118.

8. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*. 2013;31(12):1119-1125. doi:10.1038/nbt.2727.

9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351.

10. Ali OA, O’Rourke SM, Amish SJ, et al. Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*. 2016;202(2):389-400. doi:10.1534/genetics.115.183665.

11. Finger AJ, Mahardja B, Fisch KM, et al. A conservation hatchery population of delta smelt shows evidence of genetic adaptation to captivity after 9 generations. *J Hered*. 2018;109(6):689-699. doi:10.1093/jhered/esy035.

12. Ruby JG, Bellare P, DeRisi JL. PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3&amp;#58; Genes|Genomes|Genetics*. 2013;3(5):865-880. doi:10.1534/g3.113.005967.

13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.

14. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.

15. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. 2014;15(1):1-13. doi:10.1186/s12859-014-0356-4.

16. Korneliussen TS, Moltke I. NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*. 2015;31(24):4009-4011. doi:10.1093/bioinformatics/btv509.

17. Purcell S, Neale B, Todd-brown K, et al. PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007;81(September):559-575. doi:10.1086/519795.

18. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 2018;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.

19. Benjamin A, Sağlam İK, Mahardja B, Hobbs J, Chieh T, Amanda H. Use of single nucleotide polymorphisms identifies backcrossing and species misidentifications among three San Francisco estuary osmerids Collection Sites for Unknown Fish Yolo Bypass sites Lower Sacramento River site. *Conserv Genet*. 2018;0(0123456789):0. doi:10.1007/s10592-018-1048-9.