Table of Contents

[Chapter 3 – Investigation in identifying sex-specific markers in delta smelt 1](#_Toc113440558)

[Introduction 1](#_Toc113440559)

[Methods 2](#_Toc113440560)

[Sample collection & DNA extraction 2](#_Toc113440561)

[Genome-wide association study 3](#_Toc113440562)

[Genome-wide association study 3](#_Toc113440563)

[K-mer analysis 3](#_Toc113440564)

[Results 4](#_Toc113440565)

[Sample collection & DNA extraction 4](#_Toc113440566)

[Genome-wide association study 4](#_Toc113440567)

[Depth analysis 5](#_Toc113440568)

[K-mer analysis 5](#_Toc113440569)

[Discussion & Conclusion 5](#_Toc113440570)

[Tables & Figures 7](#_Toc113440571)

Chapter 3 – Investigation in identifying sex-specific markers in delta smelt

Introduction

Sex determination in fish is a highly variable and often plastic trait (Kobayashi et al., 2013; Nakamura et al., 1998; Volff, 2005) and understanding its mechanisms is essential for understanding the biology of a species, and for gaining insight into the evolution of sex chromosomes and genetic mechanisms underlying sex determination (Mei & Gui, 2015). Fish represent the oldest and most diverse group of vertebrates on earth with over 30,000 described species (Carroll, 1997; Long, 2011; Nelson et al., 2016). With this diversity and constant exposure to variable environments comes a vast array of morphological, physiological, behavioral, developmental and sexual mechanisms (Baroiller et al., 1999; Kikuchi & Hamaguchi, 2013; Nagahama, 2005; Nakamura et al., 1998). In teleost fishes, sex determination can be genetic, environmental or both. The determination mechanism varies between closely related species so investigation into the causative mechanisms need to happen at a species by species level (Conover & Kynard, 2013; Devlin & Nagahama, 2002; Mank & Avise, 2009; Volff, 2005). Delta smelt are born the sex they will be throughout their entire life cycle and do not appear to have environmental regulation of sex determination which suggests sex may be determined genetically. Within teleost fishes, endogenous genetic sex determination mechanisms can occur at the chromosomal level, where heterogametic males (XY) or females (ZW) exist, or mechanisms can occur at the genic level where single or multiple genes or regions influence sex determination (Devlin & Nagahama, 2002). Clarifying the mechanism of sex determination in delta smelt will contribute to the scientific body of knowledge for understanding biological variation and allow managers to identify and develop diagnostic non-lethal markers for the practical management of an endangered species declining towards extinction.

Sex-ratio bias within populations can occur at all stages of life for reasons such as environmental conditions (Korpelainen, 1990), temperature changes (Baroiller & D’Cotta, 2016; Geffroy & Wedekind, 2020), dispersal patterns (Hutchings & Gerber, 2002), parental condition (Trivers & Willard, 1973), and harvesting (Robertson et al., 2006) to name a few. Additionally, sex-ratio bias within small, isolated populations can arise through demographic stochasticity and contribute to an increased risk of extinction of a species (Lande, 1993). Once wild sex-ratios are understood, breeding programs controlling sex can reduce the loss of genetic diversity within vulnerable populations (Wedekind, 2002). Skewed sex ratios can have discrete consequences within populations. Of most concern, male sex-bias within wild populations, especially small populations, can result in lead to positive feedback loops where populations can no longer meet minimum viability thresholds and enter extinction vortexes (Gilpin & Soule, 1986; Rankin et al., 2011). Knowing population demographic information within delta smelt will lead to informed management decisions to best support recovery efforts within the imperiled fish.

The ability to determine sex through genetics and non-invasively capture population level demographic information in delta smelt would mark a large step forward in management of the species both in the wild and in captivity. Currently, wild fish can only be sexed by the expression of gametes from ripe adult fish or through dissection, both of which sacrifice the life of the fish or gametes. When sex is identified by expression, pressure is put on the abdomen of fish until eggs (in females) or running milt (in males) are excreted (Lindberg et al., 2013). Because sexual identification of wild fish depends on the physiological status of an individual fish, only about two-thirds of wild adult delta smelt sampled can be sexed (Hammock pers. comm.). Within the wild population, the ability to sex fish without culling or relying on gametic expression in fish will allow ecologists to reliably sex fish at all stages of their lifecycle without take. Genetic identification of sex in the captive refuge population would allow for fish to be sexed as juveniles and fewer resources contributing to controlling sex ratios as fish will not have to be reared until adulthood to know the composition of the population. While state and federal agencies conduct annual abundance and distribution sampling throughout the San Francisco Estuary (SFE) at different stages of delta smelt development, knowledge of sex-ratios throughout their lifecycle is currently a gap in a basic piece of population demography. Since sex-bias has the potential to significantly alter the success of the species identifying the genetic underpinnings of sex determination within delta smelt is a necessary asset to better management of the refuge population, acquire basic knowledge of life history characteristics of the species, conduct ecological surveys, and inform management of the wild population through population metrics, species modeling, demographic inference, and sex-based survival (Martínez et al., 2014).

This chapter focuses on investigating methods to non-invasively identify the sex of wild and captive delta smelt through genetic identification of species. We sought to identify potential candidate loci which could be used as genetic diagnostics for classifications of sex using three different methods: 1) a genome-wide association study, 2) depth analysis, and 3) k-mer analysis.

Methods

Sample collection, DNA extraction & sequencing

We sampled adipose fin clips from captive-bred individuals reared within the captive colony at the UC Davis Fish Conservation and Culture Laboratory (FCCL). Each fish was sexually identified through dissection or gametic expression. DNA was extracted using the Qiagen DNEasy 96 Blood & Tissue Kit with a modification of elution in 100uL of H2O rather than the proprietary AE Buffer included with the kit.

Because prior analyses attempted to identify sex markers using the *Sbf1* restriction enzyme did not produce loci diagnostic of sex, we used the *Pst1* restriction enzyme to increase the breadth of sampled sites. RAD-sequencing libraries were prepared according to Ali *et al.* (Ali et al., 2016) and sequenced with 150 bp paired-end reads on an Illumina HiSeq 4000 sequencer.

Female and male linked-read data generated for the *de novo* genome assembly were used for k-mer analyses below.

Genome-wide association study

We performed two genome-wide association studies (GWAS) using the either the new male or female assemblies as reference genomes. For each GWAS, we aligned raw RAD-sequencing data from 24 male and 24 female sexed fish to the reference genome using bwa v0.7.17-r1188 (Li & Durbin, 2009) and samtools v. Next, we tested for case-control differences in allele frequencies of genotype likelihoods spread throughout. To do this, we first used the male and female classifications into control (0) or case (1), respectively. Next, we fed the case control status into Dominant and Recessive model association analyses (-model 2 or -model 3) using ANGSD v0.921 (Korneliussen et al., 2014) with the following additional specifications -doAsso 1 -GL 1 -doMajorMinor 1 -doMaf 1 -SNP\_pval 1e-6. Allelic association with sex category was reported as a likelihood ratio test (LRT) statistic and is chi-square distributed with one degree of freedom. We applied a conservative significance cutoff with a Bonferroni corrected p-value of 0.05 using the formula where is the number of loci analyzed, is the desired p-value or significance threshold (0.05), and is the adjusted p-value given the number of loci used in the analysis.

Depth analysis

To test for read depth disparities expected between sexes in digametic species, we looked for signs of sex-specific sequencing depth differences between male and female RAD-sequencing data. We performed two experiments – one using the male assembly as a reference genome and second using the female assembly as a reference genome. Each experiment used the 24 male and 24 female alignment files from the prior GWAS. First, we acquired the depth of aligned reads at every nucleotide in the reference genome using samtools v1.9 (Li et al., 2009). Next, using custom bash and perl scripts (TKTKT INSERT GITHUB LINK 2 SCRIPTS) we discarded loci where both female or male reads had zero coverage. We then summed the number of alignments from each individual to get the total coverage for female and males at each loci, took the mean coverage and sorted by the per sex coverage ratio. We then looked for high fidelity regions which exhibited a sex coverage ratio greater than or equal to two to identify locations in the genome where one sex had consistently high coverage and the opposite sex had consistently low or half the amount of coverage.

K-mer analysis

We sought to identify unique differences of sequence content in males versus females using a k-mer based approach. To do this, we created and filtered sex-specific sequence signatures from one male and one female individual’s linked-read data generated for the prior genome assemblies. First, we created MinHash sketches of 21-mers for each sequencing data file (sourmash compute -k 21, 31, 51, --scaled 100 --track-abundance) and merged the resulting signature files together (sourmash sig merge -k 21) using sourmash c3.5.0 (Brown & Irber, 2016). Next, we eliminated k-mers likely to be the produce of sequencing errors by purging signature files of k-mers with abundances less than five (sourmash sig filter -m 5). We extracted all unique k-mers from the dataset, normalized abundances for each sex and observed the ratios of male to female abundances. Finally, we discarded k-mers shared between female and males to obtain sex-specific k-mers. The resulting high abundance, single sex k-mers were used in subsequent analyses.

To determine if the high abundance male-only k-mers were consistent elevated a larger region of the genome, we extracted contigs containing five or more k-mers, corresponding to a contig length of roughly 5,000 bp.

We compared the abundance of male-only and female-only k-mers found within those contigs. Then, we took the median abundance of k-mers within every contig to find the given contig’s abundance in each sex. We compared the male contig abundance to the female contig abundance and isolated male-only contigs deemed “putative Y” sequences for further validation.

To determine if there were observable differences between sexes we compared normalized k-mer abundance for each sex, as differences in k-mer abundance may correspond to candidate sex determining regions within a particular sex. Next, we extracted contigs containing five or more k-mers, corresponding to a contig length of roughly 5,000 bp. We compared the abundance of male and female k-mers found within those contigs. Then, we took the median abundance of k-mers within every contig to find the given contig’s abundance in each sex. We compared the male contig abundance to the female contig abundance and isolated male-only contigs deemed “putative Y” sequences for further validation.

To validate our results, we mapped RAD-sequencing data to the putative Y sequences and ran a depth analysis. In order to ensure the putative Y reads were indeed mapping to one location in the genome, we performed a stringent end-to-end alignment of the putative Y sequencing data using bowtie2 (Langmead & Salzberg, 2012). Only reads that entirely aligned to regions in the genome were used in the subsequent analysis. To find depth locations, we aligned the RAD-sequencing data to the male reference genome using samtools (Li et al., 2009). We then extracted RAD alignment depth information from all of the locations where the putative Y sequencing data had also aligned using the software BEDtools (Quinlan & Hall, 2010) and custom bash and perl scripts. After obtaining depth information across all putative Y regions, we ran the same depth analysis as above.

Results

Sample collection, DNA extraction & sequencing

We acquired sequencing data from a total of 48 (24 female and 24 male) captive-bred individuals. The average Phred score for all reads was 39 and number of reads captured per individual was 10,644,266 and 9,698,327 in female and male sequencing data, respectively.

Genome-wide association study

Post filtration alignment scores were 92.64% and 91.90% to the female and male reference genome, respectively. We analyzed 922,975 and 848,444 loci spread across the female and male reference genome, respectively. These loci correspond to a Bonferroni corrected p-value cutoff of 5.417265e-08 and 5.893141e-08 required for significance of in associations found within the female and male reference genome, respectively. No significant association in sex was found using the female reference genome. Two loci located in the male reference genome on Chromosome 5 were significantly associated with sex in delta smelt (Figure 15). Two lg05\_ 1885249G/A and lg05\_ 1885251G/T located on Chromosome 5 were highly associated with sex in delta smelt and had LRT scores of 37.854854 and 35.802804, corresponding to p-values of 7.621e-10 and 2.183e-9, respectively. Despite being highly associated with sex the genotypes at these loci were not diagnostic of sex (Table 10).

Depth analysis

After removal of reads, we carried out depth analyses using 92,808 and 92,735 loci aligned to the female and male reference genome, respectively. In both analyses we found no large areas corresponding to one sex having higher or lower depth of coverage compared to the other sex indicating

Next, using custom bash and perl scripts we discarded loci with zero coverage in either sex, totaled the number of individuals with coverage for each sex, totaled the absolute difference of coverage between sexes, and calculated the ratios of the mean depth of coverage between sexes at each loci. We looked for loci with low difference in coverage and a sex coverage ratio equal to or greater than two.

Next, using custom bash and perl scripts we totaled the number of individuals of each sex we discarded loci where neither male or female RAD-sequencing data aligned, then we totaled the number of alignments for each sex and gathered the total coverage for each sex. Finally, we totaled the difference of coverage between male and females and sorted the output to look for locations in the genome where one sex had high and consistent coverage and the other sex had low or no coverage.

\*\*could be due to sex determining area being in a region where markers did not sample or assemble

K-mer analysis

First pass filtration resulting in distinct k-mers for each sex resulted in a total of 1,284,592 distinct hashes from combined data sets, implying roughly 1.284592e+09 original k-mers. We observed three distinct peaks in the distribution of male to female k-mer abundance (Figure 3.2) female. After removing k-mers shared between sexes we found 494,251,000 female-only and 118,191,000 male-only k-mers. We observed a distinct increase of high abundance male-only k-mers (Figure 3.3).

Female and male median k-mer abundance was 13.0 and 7.0, respectively, resulting in a female correction of 0.5384615384615384.

3. 118191000 male only k-mers (est); 494251000 female only k-mers (est)

4. hashes unique to males w elevated male abundance vs female = selected 4964 hashes total

We sought to identify unique differences of sequence content in males versus females using a k-mer based approach. To do this, we used male and female individual’s linked-read data generated for the prior genome assemblies to create sex-specific sequence signature files using sourmash compute with options -k 21, 31, 51, --scaled 100 and --track-abundance in sourmash v3.5.0 (Brown & Irber, 2016). Next, we created MinHash sketches of 21-mers sampled, purged signature files of low abundance k-mers (abundance < 5) to eliminate k-mers likely to be the product of sequencing errors. Then, we discarded k-mers shared between male and females to leave only sex-specific k-mers. The resulting high abundance, single sex k-mers were used in our subsequent analyses.

After abundance filtration, there were approximately 118,191,000 male-only k-mers and 494,251,000 female-only k-mers. There was a clear distinction between the distribution of male and female abundances, where males had more high abundance k-mers compared to females (Figure TK). Upon filtering k-mers for those found on long contigs (contigs containing five k-mers or more) there was a clear increase of male-specific k-mers at half the abundance of the main distribution of female and male k-mers. Of these, we found 44 contigs with a k-mer mean abundance above five in the male sequencing data that had zero abundance in the female sequencing data.

We mapped the putative Y data back to the male reference genome and found the reads mapped to multiple regions within the genome (Table 11). However, we did not find a significant difference in male versus female read depth at locations across the putative Y regions.

Discussion & Conclusion

Our study thoroughly probed all available Illumina data, utilizing the linked-read and two RAD-sequencing data sets in multiple ways in an attempt to identify sex-specific markers. We did not find SNPs diagnostic of sex within delta smelt. The species may not have straightforward chromosomal sex-determination, though we cannot yet completely rule it out. While we did not find diagnostic SNPs, we did find paths forward for further analysis that may result in understanding delta smelt sex determination. For example, we found candidate loci via GWAS using RAD-sequencing data, and via k-mer analysis using the linked-read sequencing data generated for the genome assemblies. The GWAS found two markers on Chromosome 5 that were significantly associated with sex but did not have alleles diagnostic of sex. Interestingly, k-mer analysis detected DNA sequences only found within the male individual’s linked-read sequencing––one or more of these loci could contain a sex determining region or SNPs diagnostic of sex. The post k-mer analysis depth analysis showed that the observed increase in male specific k-mers at roughly 50% abundance of the normally distributed peak of the female k-mer abundance is consistent with the male sequencing data potentially having heterogametic (male sex-specific) regions in its genome (such as the 50:50 ratio between Y chromosomes paired with X chromosomes in human males). However, we could not identify or confirm sex-specific markers with the RAD-sequencing data generated for this project.

Additionally, many contigs containing male-specific k-mers were located on Chromosome 9. While there is a clear increase in associated SNPs on Chromosome 9, none met the significance threshold or were found to be diagnostic of sex. An additional important observation is that the k-mer analysis revealed male-specific linked-read sequencing data from an individual male aligned to multiple regions throughout the genome. This may indicate that sex determination in delta smelt is polygenic but further sequencing and analysis is needed to test this hypothesis.

Our work shows a need for further investigation using high-coverage whole-genome resequencing (WGS) data from a large cohort of male and female fish to more evenly survey the genome in hopes of identifying sex-specific markers. While RAD-sequencing data provide an adequate distribution of discrete locations throughout the genome of individuals, high-coverage WGS data more comprehensively survey the entire genome of individuals rather than sequence dependent restriction enzyme-based loci. Furthermore, including a large number of individuals (e.g., 500) in this analysis would provide more statistical power to detect loci with a modest effect on sex, as would be expected with polygenic sex determination.

Tables & Figures

A picture containing tree, day

Description automatically generated

**Figure 3.1.** 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 3.2.** Distribution of the proportion of change from where m = male k-mer abundance and fcor = corrected female abundance (, where f = female k-mer abundance and Acor = male to female abundance correction of 0.538 resulting from the ratio of male to female median abundances ).

Graphical user interface

Description automatically generated

**Figure 3.3.** Distributions of k-mer abundances in female and male linked-read sequencing data. A.) Corrected female-only k-mer abundances (Acor = 0.538). B.) Male-only k-mer abundances. C.) Overlay of the of corrected female-only and male-only k-mer abundances and corresponding percent (n) of k-mers for each sex where s = sex.

Chart, scatter chart

Description automatically generated

**Figure 3.5**. Histogram of male-only and female -only k-mer abundances of sequencing data. The male sequencing data appears to have more 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 (blue) peak can be seen from 30-70 abundance. This male-specific peak at roughly half the abundance of the female-specific peak indicates that the male genome contains a large amount of sequencing data not contained in the female genome (potentially a Y or male-specific chromosome) and provides evidence that the male delta smelt may be a heterogametic sex.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

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.