Table of Contents

[Chapter 3 – Investigation in identifying sex-specific markers in delta smelt 1](#_Toc113355006)

[Introduction 1](#_Toc113355007)

[Methods 1](#_Toc113355008)

[Sample collection & DNA extraction 1](#_Toc113355009)

[Genome-wide association study 2](#_Toc113355010)

[Genome-wide association study 2](#_Toc113355011)

[K-mer analysis 3](#_Toc113355012)

[Results 3](#_Toc113355013)

[Sample collection & DNA extraction 3](#_Toc113355014)

[Genome-wide association study 3](#_Toc113355015)

[Depth analysis 4](#_Toc113355016)

[K-mer analysis 4](#_Toc113355017)

[Discussion & Conclusion 4](#_Toc113355018)

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

Introduction

Sex determination in fish is a highly variable trait67 and understanding its mechanisms is crucial for understanding both the biology of a species and for gaining insight into the evolution of sex chromosomes and genetic mechanisms underlying sex determination68. Fish represent the most diverse group of vertebrates on earth with over 30,000 described species69. With this diversity and constant exposure to variable environments comes a vast array of morphological, physiological, behavioral, developmental and sexual mechanisms70–73. In teleost fishes, sex determination can be genetic or environmental and varies even between closely related species67,74–76. 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. In teleost fishes, endogenous genetic sex determination mechanisms can occur at the chromosomal level, where heterogametic males (XY) or females (ZW) exist, or mechanisms can be at the genic level where single or multiple genes influence sex determination75. While clarifying the mechanism of sex determination in delta smelt will increase our biological knowledge, it will also allow us to identify and develop diagnostic markers for the practical management of the species.

The ability to non-invasively identify sex in delta smelt will assist in management of the captive colony and develop knowledge of the biology of wild delta smelt. 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 excreted63. 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.). Knowledge of the genetic underpinnings of sex determination in fishes is a vital asset to the better management of captive populations, basic knowledge of life history characteristics of the species, ecological surveys and management regarding population metrics, species modeling, demographic inference, and sex-based survival77.

In order to non-invasively identify the sex of wild and captive delta smelt, we sought to identify potential candidate allele(s) which could be used as genetic diagnostics for classifications of sex

Methods

Sample collection & DNA extraction

To identify a sex specific marker or markers for delta smelt, we sampled adipose fin clips from 24 female and 24 male captive-bred individuals taken from the FCCL and sexually identified each fish through either 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.

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. RAD sequencing libraries were prepared at the GVL according to Ali *et al.* (2016)52 and sequenced at the UC Davis Sequencing Center with 150 bp paired-end reads on an Illumina HiSeq 4000 sequencer.

Genome-wide association study

We performed two rounds of a GWAS using the new male and female reference genomes. For each GWAS we first aligned raw RAD sequencing data from 24 male and 24 female sexed fish to the reference genome. Next, we looked for the association of an allele at any location in the genome with sex classification by carrying out a dominant model case-control GWAS in the program ANGSD53 using males or females as controls (0) and the opposite sex as cases (1). The goal for this GWAS was to find alleles associated with a particular sex. The association of a particular allele 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.

Genome-wide association study

If sex determination in delta smelt is caused by chromosomal differences, it would be expected that the heterogametic sex (e.g., XY) would have roughly half the sequencing depth (the number of sequences that cover a given locus) of the homogametic sex (e.g., XX) at the large region that determines sex in the fish. Additionally, the heterogametic sex would be expected to have novel sequence content. To look for signs of sex specific sequencing depth differences, we looked for male and female differences in the presence and depth of RAD markers spread throughout the genome. To do this we performed two experiments, one using the male genome as a reference and another using the female genome as a reference. Each experiment used the 24 male and 24 female alignment files from the aforementioned GWAS. First, we acquired the depth of aligned reads at each location in the reference genome using samtools depth. Next, using custom bash and perl scripts we discarded all genomic locations where no male or female RAD sequencing data aligned, and we totaled the number of male alignments 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.

K-mer analysis

In our k-mer (where k is equal to a specified sequence length) analysis we sought to identify unique differences of sequence content in males versus females. To do this, we used 10X Genomics linked-read data from one male and one female. First, we used the software sourmash78 to create a signature of all k-mers belonging to each sex with sourmash compute -scaled 1000 to sample one k-mer from ever 1000 base pairs. The signature compute step was followed by sourmash signature merge to incorporate all data from the R1 and R2 files for each sex. We purged the signature files of low abundance k-mers (abundance < 5) to eliminate k-mers that are more likely sequencing errors and discarded k-mers that were shared between male and females to leave only sex-specific k-mers. The resulting high abundance, single sex k-mers were used in our analysis.

First, we plotted and compared k-mer abundance for each sex to determine if there were observable differences between sexes. A difference in k-mer abundance could correspond to a sex determining regions within one sex (i.e., sex chromosome). Next, we extracted contigs containing five or more k-mers, which corresponds to a contig length of roughly 5,000 bp. We then compared the abundance of male and female k-mers found within those contigs. After, we took the median abundance of k-mers within a contig to find its 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. 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. 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 and custom bash and perl scripts. After we obtained depth information across all of the putative Y regions, we ran the same depth analysis as above.

Results

Sample collection & DNA extraction

TKTK

Genome-wide association study

We analyzed 848,444 and 922,975 loci spread across the male and female genome assemblies, respectively. These loci correspond to a Bonferroni corrected p-value cutoff of 5.893141e-08 in loci found within the male reference genome and 5.417265e-08 in loci analyzed within the female reference genome. 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). The two SNPs most associated with sex in delta smelt had LRT scores of 37.854854 and 35.802804 which correspond 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

We found no large areas that corresponded to one particular sex having higher depth of coverage compared to the other sex and therefore did not identify any sex-specific loci in this analysis.

K-mer analysis

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. Upon filtering k-mers for those found on long contigs (contigs containing 5 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. We found 44 contigs with k-mer mean abundance above 5 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 the genome and two RAD sequencing data sets in multiple ways to identify sex-specific markers, but we did not find diagnostic SNPs. This means that delta smelt 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 our RAD sequencing dataset and also via k-mer analysis using the linked-read sequencing data generated for the genome assembly. 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 (determined by several genes or locations scattered throughout the genome rather than a single region as in the human Y chromosome) 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 just a fraction. 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.

A picture containing tree, day

Description automatically generated

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

Chart, histogram

Description automatically generated

**Figure 16**. Histogram of male-only and female -only k-mer abundances of sequencing data. The male sequencing data appears to have 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, 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.