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)

Project Background and Introduction

Sex Determination in Fish

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 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 is a highly variable and often plastic trait driven by genetic and/or environmental mechanisms. Individuals may be gonochoristic or hermaphroditic or and can switch sexes within a lifecycle (Bachtrog et al., 2014; Baroiller & D’Cotta, 2016; Kobayashi et al., 2013; Nakamura et al., 1998; Volff, 2005). Known mechanisms for environmental sex determination (ESD) include population or social dynamics, temperature, sex ratio, pH, and salinity, and sex reversal may occur throughout the lifespan of a fish. Within genetic sex determination (GSD), sex is resolved upon the fusion of gametes where chromosomal, where heterogametic males (XY) or females (ZW) exist, or genic, with female-specific or male-specific master sex determining regulators, mechanisms drive the primary sexual development and gonadal output or individuals (Bhattacharya & Modi, 2021; Devlin & Nagahama, 2002). Co-occurring sex determining pathways may utilize any combination of ESD and GSD mechanisms where environmental factors influencing epigenetics may alter the sex of GSD individuals through environmental sex reversal (ESR) (Devlin & Nagahama, 2002; Shao et al., 2014). Understanding how sex is determined in a species allows for more effetive management practices such as the ability for aquaculture management to utilize ESR strategies to produce desired sex ratios in captive colonies or being able to non-lethally sex fish at all life stages, regardless of gametic expression (Stelkens & Wedekind, 2010).

Sex Ratios in Small Populations

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. Skewed sex ratios can have discrete consequences within populations. Of most concern, sex-ratio bias within small, isolated populations can arise through demographic stochasticity and contribute to increased risk of extinction of a species (Lande, 1993), and 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). Once wild sex-ratios are understood, breeding programs controlling sex can reduce the loss of genetic diversity within vulnerable populations (Wedekind, 2002). Because delta smelt have a small estimated population size, understanding sex ratios throughout the life-cycle of the annual species would allow for a better understanding of population dynamics in the wild.

While understanding sex determination mechanisms is essential to understanding the evolution of sex chromosomes and the effects of the environment on genetic expression of sex (Mei & Gui, 2015), the ability to identify the sex of individual fish without lethal sampling provides a less invasive strategy for population level studies of wild fish—a crucial aspect for threatened and endangered species—and aquaculture management. Aquaculture management may utilize ESR strategies to produce desired sex ratios in captive colonies (Stelkens & Wedekind, 2010).

Despite the State of California’s active monitoring of the wild delta smelt population abundance in distribution, the inability to identify the sex of fish at all life stages leaves an important metric of population dynamics unknown. Due to the mechanisms for sex determination varying between closely related species, an investigation into causative mechanisms and diagnostic markers must happen at the individual species level (Conover & Kynard, 2013; Devlin & Nagahama, 2002; Kobayashi et al., 2013; Mank & Avise, 2009; Nakamura et al., 1998; Volff, 2005).

Chapter 2 – Investigation in identifying sex-specific markers in delta smelt

Introduction

Currently the sex of a delta smelt can only be determined in mature fish expressing reproductive cells, or through dissection. This presents a hurdle in studying the wild population and for rearing fish within a captive colony. Delta smelt are a gonochoristic species where individuals do not display ESR nor appear to have environmental regulation of sex determination, which leads to the hypothesis that sex may be determined through genetics alone. This chapter investigates the assembled genomes of female and male delta smelt to probe for and define the extent of sex determining region within delta smelt. Through utilizing different techniques for identifying associative markers with sex, we sought to develop markers diagnostic of sex to provide managers a non-lethal method of sexing individual in the wild for the practical management of a listed species.

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 assign the sex of wild and captive delta smelt through identifying genetic markers. We sought to identify potential candidate loci which could be used as genetic diagnostics for classifications of sex using three methods: 1) a genome-wide association study, 2) depth analysis, and 3) k-mer analysis.

Methods

Sample collection, DNA extraction & sequencing

To obtain sequencing data, we sampled adipose fin clips from 24 female and 24 male captive-bred individuals reared within the refuge 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 (Cat No/ID: 69504) as per the manufacturer’s protocol with a modification of elution in 100uL of H2O rather than the proprietary AE Buffer included with the kit. Because prior analyses attempting 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.

RAD-sequencing data was used for the genome-wide association study, and depth analysis, and female and male 10X Genomics linked-read sequencing data generated for the *de novo* genome assembly were used for k-mer analyses. W aligned raw RAD-sequencing data to each reference genome using bwa v0.7.17-r1188 (Li & Durbin, 2009) and samtools v1.9 (Li et al., 2009) using an inhouse bash script (<https://raw.githubusercontent.com/shannonekj/ngs_scripts/master/align_RAD_2019.sh>). In short, we sorted reads, filled in mate coordinates and insert size fields, and removed duplicate reads to obtain a filtered dataset for subsequent analyses.

Genome-wide association study

We performed two sets of genome-wide association studies (GWAS) using a dominant and recessive model for each of the previously assembled male and female reference genomes. To do this, we tested for case-control differences in allele frequencies of genotype likelihoods spread throughout the genome. Female and male individuals were assigned as cases (1) and controls (0), respectively. Next, we fed individual status into a dominant (-model 2) or recessive (-model 3) model association analysis 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 which 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

We investigated RAD-sequencing data for read depth disparities between sexes expected to occur in digametic species. To do this, we looked for signs of sex-specific sequencing depth differences between female and male RAD-sequencing data. We performed two experiments – the first using our female assembly and second using our male assembly as a reference genome. Each experiment used the 24 female and 24 male 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, we discarded loci with zero coverage in either sex and compared the ratio of the mean depth for both sexes at each loci. To identify locations in the genome where one sex may exhibit consistently high coverage and the opposite sex contained less than or equal to half of the opposite sex’s depth, we sorted and looked for high-fidelity regions greater than 5,000 bp exhibiting a sex coverage ratio greater than or equal to two.

K-mer analysis

We looked for 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 the female and male individual’s linked-read sequence data generated for our 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 product 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. After, we discarded k-mers shared between female and males to obtain sex-specific k-mers and selected high abundance (50-100x) k-mers. The resulting high abundance, single sex k-mers were used in subsequent k-mer analyses.

Initially, we determined if the high abundance male-only k-mers were consistently elevated in a large region of the genome. To do this, we extracted contigs from the A1 assembly containing five or more k-mers. The A1 assembly was used to acquire contigs with moderate contiguity compared to the final reference genome (Table 2.4). Because we scaled down the number of hashes to 1/1,000 in the sourmash compute step, each selected contig was expected to have a minimum length of roughly 5,000 bp. We compared the abundance of female-only and male-only k-mers found within contigs and took the median abundance of k-mers within every contig to find the given contig’s abundance in each sex. Finally, we compared the female contig abundance to the male contig abundance and isolated the male-only contigs to compile a “putative Y” subset of sequences for further validation.

To test our results, we ran a depth analysis on reads contained within the putative Y reads. First, we filtered the putative Y contigs which entirely aligned to one location within the male reference genome using a stringent end-to-end alignment in bowtie2 (Langmead & Salzberg, 2012). Next, we filtered for alignment depth information at loci where both putative Y contigs and RAD-sequencing reads aligned to the male reference genome using the software BEDtools (Quinlan & Hall, 2010). After obtaining depth information across all putative Y regions, we ran the same depth analysis as above.

SexFindR Workflow

We followed the SexFindR workflow, developed and presented by Grayson et al. (preprint 2022 TK), created for identifying sex-linked sequences along the evolutionary gradient of sex chromosome divergence times. The SexFindR workflow goes through a 3-step protocol carrying out coverage-based, variant-based, and combined analyses. Prior to probing sequencing data for sex-linked sequences, we mapped linked and RAD-sequencing reads to the female and male reference genomes using bwa v0.7.17-r1188 (Li & Durbin, 2009) and samtools v1.10 (Li et al., 2009), called variants using Platypus (Rimmer et al., 2014), and removed low quality calls and selected for biallelic sites using bcftools v1.9 and vcftools v0.1.14, respectively (Danecek et al., 2011, 2021).

In the first step of the SexFindR workflow (Step 1) we used DifCover (Smith et al., 2018; Timoshevskaya et al., 2023) to run coverage-based analyses looking for large regions differentiated between sexes within our female and male reference genomes. We performed a total of four analyses since Step 1 uses sequence data from a single female and single male individual – one round of analyses RAD-sequencing data from individuals of both sexes with most alignments either the female and male reference genome and a second set of analyses using female and male individual linked-read sequencing data aligned to each of the reference genomes. For each experiment, we initially calculated library-specific adjustment coefficients (AC) by using samtools v1.10 to calculate modal depths followed by taking the ratio of modal depths between the samples used in each analysis. However, the AC for the RAD-sequencing data samples used for analysis were equal, so we took the ratio of bam file sizes, at the recommendation of the DifCover author (Table 3.1). Once we calculated AC’s for the respective analysis, we ran DifCover using default input parameters.

In the sequence-based analyses of Step 2 in the SexFindR workflow, we investigated RAD-sequencing data to look for candidate regions correlated with sex using four population genetic analyses, 1) SNP density correlation; 2) allelic fixation differences (Fst); 3) a reference genome-based genome-wide association study using SNPs; and 4) a reference free association study using k-mers. To investigate if there were SNP density differences between male and female sequencing data, we obtained the density of SNPs for each individual using vcftools v0.1.14 and obtained results by processing our individual’s density metrics through SexFindR’s R script which uses a 10 kb sliding window to calculate means and differences to carry out a permutation test to generate p-values. Next, we used vcftools v0.1.14 (--weir-fst-pop) to calculate Fst between females (population A) and males (population B). After, we performed a reference genome based GWAS. We used vcftools v0.1.14 to filter data to remove sites with a large proportion of missing data (> 50%), sites containing rare alleles (minor allele frequency less than 5% or greater than 95%), and indels. Files were formatted as plink binary files using plink v1.90b6.21 (Purcell et al., 2007) and fed into GEMMA v 0.98.3 (Zhou & Stephens, 2012) to perform a GWAS using a likelihood ratio test (-lm 2). Finally, we performed a reference-free GWAS using k-mers. To do this, we followed the protocol outlined exactly in the SexFindR workflow by counting k-mers (k=31) using KMC3 (Kokot et al., 2017); adding strand information, creating a list of k-mers found in multiple samples, building a kmer table and converting to plink binary file format with kmerGWAS v0.2 (Voichek & Weigel, 2020); and performing an association analysis to obtain p-values associated with sex using plink v 1.9. We used a conservative Bonferroni cutoff to filter for highly associated k-mers (p < 1e-9) and assembled k-mers into contigs using ABySS 2.2.5 (Jackman et al., 2017). Finally, we identified the location of the associated sequences in the male and female reference assemblies using blastn (Altschul et al., 1990).

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 3.1). Two loci (Chr05:1885249 G/A and Chr05:1885251 G/T) located on Chromosome 5 of the male assembly 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 (Figure 3.1, Table 3.1). Despite being highly associated with sex the genotypes at these loci were not diagnostic of sex (Table 3.2).

Depth analysis

After removal of reads with zero depth, 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 areas greater than 5kb corresponding to given sex having higher or lower depth of coverage compared to the other sex.

K-mer analysis

First pass filtration for distinct k-mers from each sex resulted in a total of 1,284,592 distinct hashes from combined data sets, implying roughly 1.284592e+09 original k-mers. Female and male median k-mer abundance was 13.0 and 7.0, respectively, resulting in a female correction of 0.539. We observed three distinct peaks in the distribution of male to female k-mer abundance (Figure 3.2). After removing k-mers shared between sexes, we obtained 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 and after filtering for k-mers with an abundance level from 50-100 we found 4,964 hashes corresponding to approximately 4,964,000 high abundance k-mers (Figure 3.3).

Upon filtering for k-mers in the A1 assembly, a total of 2,067 hashes, or 2,067,000 k-mers, were found on A1 contigs containing five or more hashes. Both female and male sequencing data had broad distributions of k-mers with an abundance of 90-140, however, a male specific k-mer abundance peak was observed from 30-70 at roughly half of the female abundance level (Figure 3.4). Of these, we found 44 putative Y contigs with a k-mer mean abundance above five in the male sequencing data which displayed zero k-mer abundance in the female sequencing data (Figure 3.5). We mapped the putative Y data back to the male reference genome and found the reads mapped to multiple regions within the genome. Upon repeating a depth analysis within the putative Y contigs, we did not find a significant difference in male versus female read depth.

Discussion & Conclusion

In an attempt to identify sex-specific markers within the genome of delta smelt, our experiments thoroughly probed Illumina data, utilizing linked-read and two RAD-sequencing datasets in multiple ways. We did not find SNPs completely diagnostic of sex in any of our experiments. Our results indicate the species may not have straightforward chromosomal sex-determination, though we cannot yet completely rule it out for reasons explained below. While we did not find diagnostic sequences, we did find indicators warranting further analysis—our GWAS identified candidate loci using RAD-sequencing data, and k-mer analysis found unique male-specific k-mers in the linked-read sequencing data.

Analyses using RAD-sequencing data alone showed mixed results. Our association analysis revealed two statistically significant SNPs on Chromosome 5 (chr5:1885249 and chr5:1885251) of the male reference assembly. While GWAS results identified two SNPs highly associated with sex, neither was perfectly correlated with sex and cannot be used as a diagnostic marker for applications in the field. The region containing Chromosome 5 may be a good candidate region for future investigation as it contains genes (such as TENM1 and smarca1) which are found on mammalian X chromosomes. Furthermore, we did not specifically sequence genes in this region and may not have captured adequate variation with our RAD data. Depth analysis using RAD-sequencing data revealed no markers with consistent depth disparities between sexes. Our inability to identify markers diagnostic of sex in the GWAS and depth analyses could be due to inadequate coverage of the delta smelt sex determining region. Since RAD-sequencing data only samples at specific sequences, our data may not have adequately sampled genetic material in delta smelt’s sex determining region or may have missed the region entirely. Lack of adequate coverage in the area of interest could easily result in inconclusive results as other fish species genomes have been shown to contain only a single sex-linked SNP when performing a similar analysis with whole genome resequencing data (Grayson et al., 2022; Kamiya et al., 2012). Thus, if the sex determining region in delta smelt is particularly small or is in an area without regular *PstI* cut sites, we would not pick up a signature in our analyses.

Interestingly, k-mer analysis using linked-read data 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 male-specific peak at roughly half the abundance of the female-specific peak shown in Figure 3.4 and the abundance of k-mers only contained within the male sequencing data shown in Figure 3.5 may be an indications 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.

Additionally, 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 sex-specific markers within this region using the RAD-sequencing data generated for this project further suggesting RAD data provides insufficient sampling of the delta smelt genome.

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 survey the genome more evenly in hopes of identifying sex-specific markers. While RAD-sequencing data provide an adequate distribution of discrete locations throughout the genome of individuals, it unevenly samples the genome as it is dependent upon using specific sequences at cut sites. Analyses performed using RAD-sequencing data may have insufficient coverage over sex determining or diagnostic regions of the genome. Insufficient coverage of sequencing data throughout the genome has previously been documented to mask diagnostic markers in fish (Narum et al., 2018; Prince et al., 2017). Using high-coverage WGS data would comprehensively survey the entire genome of individuals, as its sampling locations are not dependent upon the sequence of the individual. 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 and Figures

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3.1.** Genome-wide association study results from loci meeting Bonferroni corrected p-value cutoff. | | | | |
|  | Reference Genome | Male | Male |  |
|  | Chromosome | 5 | 5 |  |
|  | Position Number | 1885249 | 1885251 |  |
|  | Major Allele | G | G |  |
|  | Minor Allele | A | T |  |
|  | LRT | 37.854854 | 35.802804 |  |
|  | p-value | 7.62E-10 | 2.18E-09 |  |
|  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3.2,** Genotypes of female (1) and male (0) individuals at the two loci on Chromosome 5 of the male reference genome found to be significantly associated with sex in delta smelt. | | | | |
|  |  | Chr05:1885249 | Chr05:1885251 |  |
|  | major | G | G |  |
|  | minor | A | T |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GG | GG |  |
|  | 1 | GA | GT |  |
|  | 1 | GA | GT |  |
|  | 1 | GA | GT |  |
|  | 1 | GA | GT |  |
|  | 1 | GA | GT |  |
|  | 1 | AA | TT |  |
|  | 1 | AA | TT |  |
|  | 1 | AA | TT |  |
|  | 1 | AA | TT |  |
|  | 0 | GG | GG |  |
|  | 0 | GG | GG |  |
|  | 0 | GG | GG |  |
|  | 0 | GG | GG |  |
|  | 0 | GG | GG |  |
|  | 0 | GG | GG |  |
|  | 0 | GG | GG |  |
|  | 0 | GA | GG |  |
|  | 0 | GA | GT |  |
|  | 0 | GA | GT |  |
|  | 0 | GA | GT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  | 0 | AA | TT |  |
|  |  |  |  |  |

A picture containing tree, day

Description automatically generated

**Figure 3.1.** Manhattan plots of each of 28 chromosomes from the male assembly. Location on the x axis and significance on the y axis. Significant SNPs on Chromosome 5 are marked in blue.

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, histogram

Description automatically generated

**Figure 3.4.** Distribution of female-only (red) and male-only (blue) k-mer abundances of k-mers located on contigs containing five hashes (approximately 5,000 bp in length) from the male A1 assembly.

Chart, scatter chart

Description automatically generated

**Figure 3.5.** Female (y-axis) and male (x-axis) median k-mer abundance on contigs containing 5 or more hashes, corresponding to roughly 5,000 bp. k-mers. Dashed line shows slope of one-to-one abundance between female and male abundances. Right plot show all contigs, while the left plot is zoomed in to better visualize 44 contigs with k-mer abundances contained in male sequencing data but zero k-mer abundance in female sequencing data. This indicates the male sequencing data contains sex-specific sequences in high abundance that are not contained in the female sequencing data.

References

Ali, O. A., O’Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bachtrog, D., Perrin, N., Ming, R., Valenzuela, N., Mayrose, I., Peichel, C. L., Hahn, M. W., Ashman, T.-L., Vamosi, J. C., Ross, L., Kirkpatrick, M., Kitano, J., Otto, S. P., & Mank, J. E. (2014). Sex Determination: Why So Many Ways of Doing It? *PLoS Biology*, *12*(7), e1001899–e1001899. https://doi.org/10.1371/journal.pbio.1001899

Baroiller, J.-F., & D’Cotta, H. (2016). The Reversible Sex of Gonochoristic Fish: Insights and Consequences. *Sexual Development*, *10*, 242–266.

Baroiller, J.-F., Guiguen, Y., & Fostier, A. (1999). Endocrine and environmental aspects of sex differentiation in fish. *Cellular and Molecular Life Sciences*, *55*, 910–931. https://doi.org/10.1007/978-3-0348-7781-7\_9

Bhattacharya, I., & Modi, D. (2021). Sex Determination in Teleost Fish. In J. K. Sundaray, M. A. Rather, S. Kumar, & D. Agarwal (Eds.), *Recent updates in molecular Endocrinology and Reproductive Physiology of Fish* (pp. 121–138). Springer Singapore. https://doi.org/10.1007/978-981-15-8369-8\_9

Brown, C. T., & Irber, L. (2016). sourmash: A library for MinHash sketching of DNA. *The Journal of Open Source Software*, *1*(5), 27. https://doi.org/10.21105/joss.00027

Carroll, R. L. (1997). *Vertebrate Paleontology and Evolution* (7th ed.). W.H. Freeman and Company.

Conover, D. O., & Kynard, B. E. (2013). *Environmental Sex Determination: Interaction of Temperature and Genotype in a Fish Environmental Sex Determinaffon: Interaction of Temperature and Genotype in a Fish*. *213*(4507), 577–579.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Devlin, R. H., & Nagahama, Y. (2002). Sex determination and sex differentiation in fish: An overview of genetic, physiological, and environmental influences. *Aquaculture*, *208*(3–4), 191–364. https://doi.org/10.1016/S0044-8486(02)00057-1

Geffroy, B., & Wedekind, C. (2020). Effects of global warming on sex ratios in fishes. *Journal of Fish Biology*, *97*(3), 596–606. https://doi.org/10.1111/jfb.14429

Gilpin, M. E., & Soule, M. E. (1986). Viable Populations for Conservation. In *Minimum viable populations: Processes of species extinction*. Cambridge University Press.

Grayson, P., Wright, A., Garroway, C. J., & Docker, M. F. (2022). *SexFindR: A computational workflow to identify young and old sex chromosomes*. https://doi.org/10.1101/2022.02.21.481346

Hutchings, J. A., & Gerber, L. (2002). Sex–biased dispersal in a salmonid fish. *Proceedings of the Royal Society B: Biological Sciences*, *269*, 2487–2493. https://doi.org/10.1098/rspb.2002.2176

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., & Birol, I. (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777. https://doi.org/10.1101/gr.214346.116

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., Tohari, S., Brenner, S., Miyadai, T., Venkatesh, B., Suzuki, Y., & Kikuchi, K. (2012). A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, Takifugu rubripes (Fugu). *PLoS Genetics*, *8*(7), e1002798. https://doi.org/10.1371/journal.pgen.1002798

Kikuchi, K., & Hamaguchi, S. (2013). Novel sex-determining genes in fish and sex chromosome evolution. *Developmental Dynamics*, *242*(4), 339–353. https://doi.org/10.1002/dvdy.23927

Kobayashi, Y., Nagahama, Y., & Nakamura, M. (2013). Diversity and Plasticity of Sex Determination and Differentiation in Fishes. *Sexual Development*, *7*(1–3), 115–125. https://doi.org/10.1159/000342009

Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: Counting and manipulating *k* -mer statistics. *Bioinformatics*, *33*(17), 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

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

Korpelainen, H. (1990). SEX RATIOS AND CONDITIONS REQUIRED FOR ENVIRONMENTAL SEX DETERMINATION IN ANIMALS. *Biological Reviews*, *65*(2), 147–184. https://doi.org/10.1111/j.1469-185X.1990.tb01187.x

Lande, R. (1993). Risks of Population Extinction from Demographic and Environmental Stochasticity and Random Catastrophes. *The American Naturalist*, *142*(6), 911–927. https://doi.org/10.1086/285580

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindberg, J. C., Tigan, G., Ellison, L., Rettinghouse, T., Nagel, M. M., & Fisch, K. M. (2013). Aquaculture methods for a genetically managed population of endangered delta smelt. *North American Journal of Aquaculture*, *75*(2), 186–196. https://doi.org/10.1080/15222055.2012.751942

Long, J. A. (2011). *The Rise of Fishes: 500 Million Years of Evolution* (2nd ed.). Johns Hopkins University Press.

Mank, J. E., & Avise, J. C. (2009). Evolutionary diversity and turn-over of sex determination in teleost fishes. *Sexual Development*, *3*(2–3), 60–67. https://doi.org/10.1159/000223071

Martínez, P., Viñas, A. M., Sánchez, L., Díaz, N., Ribas, L., & Piferrer, F. (2014). Genetic architecture of sex determination in fish: Applications to sex ratio control in aquaculture. *Frontiers in Genetics*, *5*(SEP), 1–13. https://doi.org/10.3389/fgene.2014.00340

Mei, J., & Gui, J. F. (2015). Genetic basis and biotechnological manipulation of sexual dimorphism and sex determination in fish. *Science China Life Sciences*, *58*(2), 124–136. https://doi.org/10.1007/s11427-014-4797-9

Nagahama, Y. (2005). Molecular mechanisms of sex determination and gonadal sex differentiation in fish. *Fish Physiology and Biochemistry*, *31*, 105–109. https://doi.org/10.1007/s10695-006-7590-2

Nakamura, M., Kobayashi, T., & Chang, X. (1998). Gonadal sex differentiation in teleost fish. *Journal of Experimental Zoology*, *281*, 362–372.

Narum, S. R., Di Genova, A., Micheletti, S. J., & Maass, A. (2018). Genomic variation underlying complex life-history traits revealed by genome sequencing in Chinook salmon. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1883), 20180935. https://doi.org/10.1098/rspb.2018.0935

Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). *Fishes of the World* (5th ed.). John Wiley & Sons.

Prince, D. J., Saglam, I. K., Hotaling, T. J., Spidle, A. P., & Miller, M. R. (2017). The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *SCIENCE ADVANCES*.

Purcell, S., Neale, B., Todd-brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., Bakker, P. I. W. D., Daly, M. J., & Sham, P. C. (2007). PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(September), 559–575. https://doi.org/10.1086/519795

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rankin, D. J., Dieckmann, U., & Kokko, H. (2011). Sexual Conflict and the Tragedy of the Commons. *The American Naturalist*, *177*(6), 780–791. https://doi.org/10.1086/659947

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, Wilkie, A. O. M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918. https://doi.org/10.1038/ng.3036

Robertson, B. C., Elliott, G. P., Eason, D. K., Clout, M. N., & Gemmell, N. J. (2006). Sex allocation theory aids species conservation. *Biology Letters*, *2*(2), 229–231. https://doi.org/10.1098/rsbl.2005.0430

Shao, C., Li, Q., Chen, S., Zhang, P., Lian, J., Hu, Q., Sun, B., Jin, L., Liu, S., Wang, Z., Zhao, H., Jin, Z., Liang, Z., Li, Y., Zheng, Q., Zhang, Y., Wang, J., & Zhang, G. (2014). Epigenetic modification and inheritance in sexual reversal of fish. *Genome Research*, *24*(4), 604–615. https://doi.org/10.1101/gr.162172.113

Smith, J. J., Timoshevskaya, N., Ye, C., Holt, C., Keinath, M. C., Parker, H. J., Cook, M. E., Hess, J. E., Narum, S. R., Lamanna, F., Kaessmann, H., Timoshevskiy, V. A., Waterbury, C. K. M., Saraceno, C., Wiedemann, L. M., Robb, S. M. C., Baker, C., Eichler, E. E., Hockman, D., … Amemiya, C. T. (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nature Genetics*, *50*(2), 270–277. https://doi.org/10.1038/s41588-017-0036-1

Stelkens, R. B., & Wedekind, C. (2010). Environmental sex reversal, Trojan sex genes, and sex ratio adjustment: Conditions and population consequences: ENVIRONMENTAL SEX REVERSAL AND TROJAN SEX GENES. *Molecular Ecology*, *19*(4), 627–646. https://doi.org/10.1111/j.1365-294X.2010.04526.x

Timoshevskaya, N., Eşkut, K. İ., Timoshevskiy, V. A., Robb, S. M. C., Holt, C., Hess, J. E., Parker, H. J., Baker, C. F., Miller, A. K., Saraceno, C., Yandell, M., Krumlauf, R., Narum, S. R., Lampman, R. T., Gemmell, N. J., Mountcastle, J., Haase, B., Balacco, J. R., Formenti, G., … Smith, J. J. (2023). An improved germline genome assembly for the sea lamprey Petromyzon marinus illuminates the evolution of germline-specific chromosomes. *Cell Reports*, *42*(3), 112263. https://doi.org/10.1016/j.celrep.2023.112263

Trivers, R. L., & Willard, D. E. (1973). Natural Selection of Parental Ability to Vary the Sex Ratio of Offspring. *Science*, *179*(4068), 90–92. https://doi.org/10.1126/science.179.4068.90

Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, *52*(5), 534–540. https://doi.org/10.1038/s41588-020-0612-7

Volff, J. N. (2005). Genome evolution and biodiversity in teleost fish. *Heredity*, *94*(3), 280–294. https://doi.org/10.1038/sj.hdy.6800635

Wedekind, C. (2002). Manipulating sex ratios for conservation: Short‐term risks and long‐term benefits. *Animal Conservation*, *5*, 13–20. https://doi.org/10.1017/S1367943002001026

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310