Chromosome-scale genome assembly and investigation of the *Hypomesus transpacificus* genome for sex-specific markers, and association of the lactase persistence haplotype block with disease risk in populations of European descent.

By

SHANNON ERICA KENDAL JOSLIN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Dr. Andrea Schreier, Chair

Dr. C. Titus Brown

Dr. Benjamin Sacks

Committee in Charge

2023

To Philia, forever.

Acknowledgements

This dissertation is the synthesis of so many wonderful people's hard work. First, I would like to extend gratitude toward my PI's Dr. Amanda Finger, Dr. Andrea Schreier, and Dr. Michael Miller. I appreciate your time, energy, discussions, dedication, compute resources, and confidence in me. A tremendous thank you to Dr. C. Titus Brown for inspiring, not expecting, me to do better all the time. I'd not be where I am without having been a part of your lab. I would like to thank Dr. Danielle Lemay for being the catalyst for me coming back to UC Davis, teaching me persistence to publication, for consistently checking up on me during the wildfires, being a powerful and supportive role model and inspiration within our LGBTQIA+ community. I have so much gratitude towards all my colleagues and collaborators within the Genomic Variation Lab, Miller Lab, DIB Lab, Delaney Lab, 2016 IGG Cohort, 2017 PBGG Cohort, and DIBSI Instructors, Learners and Volunteers. There's nothing like bashing your head against the virtual wall when you're doing it with other dweebs.

I would like to thank my family: my father, Terry Joslin, for almost always being curious about what I am up to; my uncle, Russ Levy, for reminding me sometimes all you need to reset is a walk out in nature; and my grandfather, Alan Levy, who passed during my studies, but established a deep grit within my character. Thanks to Grace Kerfoot for having the patience to listen to me repetitively practicing saying the same words in front of my whiteboard for weeks in preparation for my qualifying exam. I would like to express my gratitude and appreciation to Sarah Stock, my current supervisor at the National Park Service, for supporting & encouraging me to take time to complete this dissertation, and providing me my dream job with the ability to live in my favorite place in the world – Yosemite Valley.

Thanks to my beautiful friends that believed in and supported me in all the stages of this process, especially when I did not, my climbing family who figuratively – and literally – give me the softest catches and my Yosemite community for sending support and practicing patience in my slog to the finish line: Yoni Ackerman, Natasha Barnes, Kei-san Bautista, William “Boz” Bazargani, Molly Beard, Nik Berry, Shira Biner, Anna Carney, Mary Clapp, Nick Cornwell, Ian Cotter-Brown, Hannah Donnely, Dr. Ally Dorey, Stacey Dorais, Julia Ersan, Maddy Farmer, Emily Fong, Eric Gabel, Dr. Ryan Gallagher and fam, Dustin Garrison, Horacio Gratton, Patrick Grof-Tisza, Fidel Guirado, Juliana Haber, Hannah Hall, Dr. Ed Hartouni, Cody Hays, Jeremy Hemberger, Dr. Breezy Jackson, Little Juni, Cameron King, Lydia King, Sarah Linder, Dr. Lisa Linville, Elliot Lozano, Heather Mackey, Rylan Marshall, Meagan Martin, Mike McDonald, Mary Mecklenberg, Carl Miler, Brianna McGuire, Kimbrough Moore and the Roses, Si Moore, Eileen Morley, Dr. Felicity Muth, Emerson Patton, Gabe Reyes, Eamon Schneider, Chad Shepard, Chris Sinatra, Sean Smith, Tessa Smith, PJ Soloman, Mel Steller, Cheyenne Sukalski, Nick Sullens, Steve Schwortz, Adam Tait, Ian Texeira, Carlo Traversi, Julien Udea, Katharina Ullmann, Hope Vince, Austin Waag, Ian Walters, Erin Wang, Zac Warren, Jimmy Webb, Dave Wetmore, and Roman Yalowitz.

Lastly, Ann Holmes and Shannon Kieran for helping me to the finish line, and Michelle Yalowega for providing Squamish’s best couch and the best bouldering breaks a distressed writer could hope for.

This dissertation, like all my published work, is dedicated to my dog/forever child, Philia Butter Joslin.

Finally, I would like to acknowledge the adorably derpy *Hypomesus transpacificus,* I am sorry humanity has failed you and I thank you for the few individuals I sacrificed to make this work possible. Perhaps in another universe you are thriving.

Abstract

Delta smelt, *Hypomesus transpacificus* (McAllister, 1963), is a federally threatened and California State endangered fish endemic to the San Francisco Estuary and Sacramento-San Joaquin Delta of North America (SFE). The species is a small, pelagic, mostly annual fish with freshwater resident, migratory, and semi-migratory life histories (Campbell et al., 2022; Hobbs et al., 2019). They have historically been considered an indicator species for water quality in the SFE. Over the last few decades, the species has undergone a population collapse associated with drought and anthropogenic disturbances, and it is now believed stochastic processes may push the species to extinction (Fisch et al., 2011; Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, 2016). Meaningful conservation management of the species must encompass gaining a better understanding of the life history, ecology, demography, and physiology of the species so biological components contributing to success in the wild can be preserved. Because genetics, in combination with the environment, influence many aspects of individual and population level phenotypes, building a framework to better understand the species requires the development of genetic resources and monitoring of genetic diversity. Chapter one of this dissertation presents two chromosome-level genome assemblies — one male and one female — which are necessary resources for current and ongoing evolutionary and conservation genetics research concerning delta smelt and other declining and vulnerable species in the Osmeridae family, such as longfin smelt. Chapter two investigates three methods for identifying sex marker(s) within the assembled female and male delta smelt reference genomes. While ultimately no diagnostic sex-specific sequences were found in our RAD-sequencing dataset, abundance discrepancies in k-mers from female and male linked-read sequence data were identified. Chapter three is a first author paper

I wrote titled “Association of the lactase persistence haplotype block with disease risk in populations of European descent” published in Frontiers in Genetics. This chapter switches organisms and investigates the potential for deleterious mutations to hitchhike in haplotype blocks which were heavily selected for in humans. This paper is a result of the work I completed in the first year and a half of my doctoral studies. Together this work contributes to the fields of evolutionary, comparative and conservation genomics. This work specifically contributes to delta smelt monitoring, management and research, and human disease risk studies.

In summary my doctoral work has provided a novel delta smelt genome assembly which is the first chromosome-level and least fragmented publicly available male and female reference genomes within the Osmeridae (smelt) family; an examination of female and male delta smelt sequencing data showing a discrete difference between sexes and establishes a framework for further investigation; and results suggesting that despite the fact that the human lactase persistence haplotype block harbors increased deleterious mutations compared to the rest of the genome, they seem to have little effect on prostate cancer, cardiovascular disease, and bone mineral density disease phenotypes.

Table of Contents

CHAPTER 1 – GENOME ASSEMBLY OF *HYPOMESUS TRANSPACIFICUS* (DELTA SMELT)

Introduction and Background

The San Francisco Estuary (SFE) drains 40% of California's surface area and is a dynamic, radically anthropogenically altered ecosystem currently encompassing 1,240 square kilometers of open water and wetlands in Northern California (Conomos, 1979). Large-scale abiotic and biotic human-induced alteration of the SFE has been documented since the mid 1800s and has transformed the way water is distributed throughout the estuarine environment. Abiotic drivers of change include hydraulic mining, watershed modification, subsidence, and direct sediment removal resulting in increased pollutants such as mercury and alterations in the abundance and timing of freshwater inflow (Barnard et al., 2013). Biotic drivers of change include introduced species affecting competition for resources such as habitat and food availability (Glibert et al., 2011; Nichols et al., 1986). State and federal agencies routinely monitor the relative abundance of fish in the SFE with initiatives like the California Department of Fish and Wildlife (CDFW) Fall Midwater Trawl Survey which began in 1967. Similar to many other estuarine ecosystems throughout the globe (Belarmino et al., 2021; Cottingham et al., n.d.; James et al., 2018), within the SFE many once abundant endemic pelagic fishes, such as delta smelt, have undergone broad declines in population size (Moyle, 2002; Moyle et al., 2018; Sommer et al., 2007).

Delta Smelt

Delta smelt (*Hypomesus transpacificus*) (phylum: Chordata, class: Actinopterygii, order: Osmeriforme, family: Osmeridae) is a small (6 – 9 cm), translucent, semi-anadromous species which migrates between fresh and saline water, reproduces annually and is endemic to the to the

Sacramento-San Joaquin River Delta of the San Francisco Estuary (SFE) in California (Sommer et al., 2011). Delta smelt are part of the Osmeridae family which represents an abundant food source for human consumption in Japan, Europe, and North America and have experienced declining populations worldwide (McAllister, 1963; Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, 2016; Rosenfield & Baxter, 2007). Historically thought to be solely semi-anadromous (Sommer et al., 2011), recent otolith analyses have shown that delta smelt exhibit three migratory phenotypes: semi-anadromous, freshwater resident, and low-salinity brackish-water resident (Hobbs et al., 2019). Because of their annual life cycle and relatively rapid response to the conditions of their habitat, delta smelt are considered an indicator of the overall health of the SFE ecosystem. The species was once one of the most abundant and widely distributed fish species in the SFE, but the delta smelt population abundance has been declining since the 1980s (Moyle et al., 1992). The species was listed as threatened under the federal Endangered Species Act (ESA) in 1993 and endangered under the California ESA in 2009. As a result of their waning population, resource management agencies, such as CDFW, actively monitor the distribution and abundance of the wild population, and the Genomic Variation Laboratory at the University of California Davis (UC Davis) genetically manages a captive breeding program to maintain a refuge population at the UC Davis Fish Conservation and Culture Laboratory (FCCL).

Despite protections under the ESA and CESA delta smelt have continued to decline throughout the SFE (State of California FMWT, 2022; State of California SKT, 2022) (Figure 1.1). It is now believed stochastic processes may push the species to extinction (Fisch et al., 2011; Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, 2016). Decreased pelagic productivity and increased water temperature in the SFE have been shown to be primary drivers of condition

indices negatively affecting delta smelt's fitness (Hammock et al., 2022). However, how genetics influences condition indices or have been affected from the species' decreasing population size remains unknown.

Genomic resources such as reference genomes contribute to two broad categories of study: medicine and biodiversity. Medicine has benefitted from genomic resources using comparative methods to identify conserved locations of the genome (loci) essential to life for different classes of organisms as well as identify genetic variants associated with disease, disease susceptibility and other phenotypic traits (Claussnitzer et al., 2020). Biodiversity relies on having genetically diverse organisms within and between species, as genetic diversity is related to the evolutionary capacity to adapt to environmental change. As such, being able to compare and quantify genetic diversity with high resolution is essential to understanding the genetic underpinnings associated with biodiversity. Rapid development of high-throughput sequencing technologies over the past few decades has led to an era of genomic research for non-model organisms, and much genomic research begins with reference genomes.

Recent large-scale genome assembly initiatives focused on creating high quality (i.e. chromosome-level) reference genomes of species across the tree of life, such as the Earth BioGenome Project and the Vertebrate Genome Project, have allowed scientists to carry out high resolution comparative genomics studies which have distinguished genomic motifs associated with a species' risk of extinction, identified signals of evolutionary selection, and provided population level insights from individual reference genomes of non-model organisms (Feng et al., 2020; Zoonomia Consortium, 2020). In both medicine and biodiversity, research involving reference genomes is limited by the completeness of the assembled resource used in

the analysis. If a genome is highly fragmented or contains large gaps, studies cannot accurately quantify the extent of conserved loci or may fail to probe relevant regions of the genome. As such, it is essential for genomic resources to be as complete as possible. One way of increasing an assembly's contiguity is through combining multiple sequencing technologies to carry out a "hybrid" method of genome assembly, as each technology expands the capacity for capturing and assembling increasingly more of an organism's actual genome.

Next generation sequencing (NGS) and third generation sequencing (TGS) technologies have allowed biologists to generate, low cost, high-throughput sequencing data with relative ease. Of the numerous new methods to generate sequencing data, a few (long-read sequencing, linked-read sequencing, and hi-c chromatin confirmation capture) have been found to be extremely useful for assembling highly contiguous reference genomes. PacBio HiFi sequencing generates long reads tens of thousands of base pairs in length; 10X Genomics linked-reads are sequences of Illumina short reads made into pseudo long-reads through barcoding (linking) short segments of DNA contained on the same a long fragment; and Phase Genomics hi-c chromatin confirmation capture (hi-c) are sequences of Illumina short reads generated from crosslinked physically interacting DNA. There are limitations to the newer sequencing technologies: many TGS technologies requires high molecular weight (HMW) DNA, which can be difficult to generate in sufficient quantities, especially for exceedingly small or rare species; and each of the three technologies (long-read, linked-read, and hi-c) have different biases, errors, and limitations. Long-read based sequencing technologies are generally more error prone, while short-read based technologies, such as linked-reads and hi-c, cannot span large repetitive genomic motifs or only capture short segments of DNA around physical interactions,

respectively. However, through using multiple technologies in sequential steps "hybrid" assemblies can overcome contiguity or accuracy limitations found within single-technology assemblies.

A recent major publication to utilize and highlight current methods in the hybrid approach to *de novo* assembly pertained to the domestic goat (Bickhart et al., 2017) which used long-read sequencing, linked-read sequencing, and interaction mapping to increase the previous goat genome's contiguity by over two orders of magnitude. At the time of publication, it proposed their hybrid genome assembly was the most continuous *de novo* mammalian assembly of its time. Since the publication of the *de novo* goat assembly, hybrid-assembly publications and the resulting reference genomes have become commonplace. As such, hybrid assembly is an accepted and reliable way to achieve a chromosome-scale high-quality reference genome (Bickhart et al., 2017; Rhie et al., 2021). Since 2017, over half of all vertebrate chromosome-level assemblies submit to GenBank have implemented a hybrid assembly approach to genome assembly and the time and resources required for creating a reference genome is continually decreasing (Hotaling et al., 2021). From sampling to annotation current *de novo* genome assemblies can take less than a month to obtain completeness standards which took decades for the first human genome to reach at a fraction of the cost and with vastly fewer contributing biologists (International Human Genome Sequencing Consortium et al., 2001).

We carried out a hybrid method of assembly using linked-read, long read and Hi-C sequencing data to create one female and one male delta smelt reference genome. Linked-read sequencing data was provided through 10X Genomics. It is a useful technology to incorporate into a hybrid genome assembly, as it requires small amounts (nanograms) of HMW DNA and creates highly

accurate pseudo-long reads. Linked-read technology combines the benefits of long and short reads through encapsulating individual long DNA fragments (50-200 kb) into droplets of water-in-oil emulsion, or Gel Bead-in-Emulsions (GEMs), shearing long fragments into short fragments (400-600 bp), attaching a unique barcode to all DNA fragments in one GEM, sequencing paired end (PE) reads on an Illumina sequencer, and using bioinformatics to create pseudo-long reads by reassociating short reads previously contained in one GEM. We generated PacBio HiFi read sequencing data to generate long reads. HiFi reads, which are generally moderate in length (10-30 kb), can span highly repetitive regions of moderate size and can reach base calling accuracy of 99.9%. HiFi reads are generated by creating circularized DNA from the two strands of DNA (5' and 3' ends). Within the circularized DNA the 5' and 3' strands are separated by a known sequence motif. Primers and a polymerase are annealed to synthesize long subreads where the circularized DNA sequence is transcribed repetitively onto one long sequencing read. The repeating sequences separated by a known motif on one read is then used to generate a highly accurate consensus sequence for the captured region. Hi-C chromatin conformation capture data was generated by Phase Genomics. Hi-C data is useful as it measures the frequency in which two pieces of DNA physically interact to derive information on their physical distance. Hi-C technology uses formaldehyde to crosslink pieces of DNA physically interacting in a nucleus. Crosslinked DNA is fragmented into smaller segments, biotinylated bases are attached to 5' overhands, the ends of opposite strands are ligated to one another, and crosslinking is reversed to create a circular segment of DNA. DNA is then fragmented and biotinylated DNA, which contains sequence data from physically interacting DNA is pulled down for highly accurate

paired-end Illumina sequencing. Thus, Hi-C data gives long-range assembly information and is often a highly useful step for scaffolding linked and long read assemblies into chromosomes.

A high-quality delta smelt reference genome provides a valuable resource for conservation research. At the broadest level, a chromosome level genome will provide an additional resource for evolutionary studies spanning the tree of life. Narrowing focus to the Osmeridae family, the delta smelt reference genome can be used for genetic studies in other listed, vulnerable, or declining species, such as longfin smelt (*Spirinchus thaleichthys*). Specific to the delta smelt species, a reference genome can be used to identify associations between non-neutral genetic variants and life history phenotypes (Campbell et al., 2022) and used to investigate domestication occurring in the propagated refuge population (Finger et al., 2018). Since the captive population of delta smelt is being used as a source for supplementing the wild population and currently represents the majority of all living individuals of the species, this genome provides a powerful resource to develop hatchery management strategies to best support adaptive genetic variation contributing to survival within the delta smelt species. Thus, a delta smelt reference genome provides a widely useable tool for active genetic research aiding in the conservation and management of delta smelt and in global efforts to preserve biodiversity and understand genomic commonalities and differences in the branches across the tree of life.

Prior to our work, no reference genome for delta smelt had been assembled. Although reference genomes of evolutionarily closely related species may be used to study genetic diversity, adaptations, and vulnerabilities when a species-specific reference genome is unavailable, the three draft genome assemblies from the Osmeridae family that were available on GenBank were highly fragmented: 1) *Hypomesus nipponensis* (wakasagi smelt, N50 = 0.46

Mb); 2) *Osmerus eperlanus* (European smelt, N50 = 6.8 Kb); and 3) *Thaleichthys pacificus* (euchlon, N50 = 3.1 Kb) (Table 1.1). Understanding the differences between the delta smelt genome and wakasagi genome, estimated to be 464 Mb in size and 2n=56 chromosomes, may be an important factor in conservation efforts as wakasagi were introduced into the SFE in 1959 and are known to hybridize with delta smelt in the wild (Dill & Cordone, 1997; Kitada et al., 1980; Xuan et al., 2021). Despite more contiguous metrics listed in the Xuan et al. (2021) wakasagi genome publication, the associated reference assembly hosted on GenBank presents a highly fragmented contig assembly (L50 = 477) and is less than 7.5% (34.4 Mb) of the total sequence length for their estimated genome size of 464 Mb (Table 1.1). This discrepancy poses limitations when using the closely related wakasagi assembly as a reference genome in delta smelt analyses, as it only represents a relatively small portion of the genome.

Here I describe the development of a chromosome level delta smelt reference genome through a hybrid method of assembly using linked-reads, long reads and chromosome conformation capture.

Materials and Methods

Sample collection & DNA extraction

To obtain high molecular weight (HMW) genomic DNA (gDNA) for long read sequencing, I used tissue samples from two female (T1F02_BM_FF and T3F02_SC_FF) and two male (T3M02_BM_FF and T3F02_SC_FF) adult delta smelt 600 days post hatch reared within the refuge population at the FCCL. I euthanized fish in MS-222 according to the UC Davis IACUC protocol #21533. After euthanasia, I dissected the fish, sampled muscle, internal organs (heart, liver and spleen) and gill tissue, immediately flash froze all tissue samples, and stored samples at -80°C isolated or suspended in propylene glycol – two storage methods known to be conducive to HMW gDNA extraction and long read sequencing in different organisms (Patrick et al., 2016; Wasko et al., 2003).

HMW gDNA was isolated by the UC Davis DNA Technologies & Expression Analysis Core Laboratory (Genome Center) following the protocol described by Wasko et al. (2003). Briefly, ~25-50 mg of flash frozen back muscle tissue and scales from a male and female were homogenized using liquid nitrogen grinding. Tissue was lysed in a buffer containing 10 mM Tris-HCl pH 8.0, 125 mM NaCl, 10 mM EDTA pH 8.0, 0.5% SDS, 4 M urea and 10 mg/mL Proteinase K. The lysate was cleaned with equal volumes of phenol/chloroform using phase lock gels (Quantabio Cat # 2302830). The DNA was precipitated by adding NaCl to the final concentration of 0.3 M and 2X volume of ice-cold ethanol. The DNA pellet was twice washed with 70% ethanol and resuspended in an elution buffer (10 mM Tris, pH 8.0). The integrity of the HMW DNA was verified on a Pippin Pulse gel electrophoresis system (Sage Sciences, Beverly, MA). Purity of the DNA was determined by measuring 260/280 and 260/230 absorbance ratios on a NanoDrop

1000 Spectrophotometer (Thermo Scientific, Wilmington, DE). Extractions with an average read fragment length of 50 kb were used for library prep and sequencing at the UC Davis DNA Technologies and Expression Analysis Core.

If the first library prep and sequencing run from an individual sample did not obtain sufficient coverage to carry out subsequent assembly for each sex, we selected a second high or low-input library prep based on the amount of extracted HMW gDNA remaining for each individual. If no further HMW gDNA was available, we resampled tissue, extracted HMW gDNA, library-prepped and sequenced a new individual of the same sex. Muscle tissue from one female individual (T1F02_BM_FF) was used for generating 10X Genomics linked-reads and Phase Genomics Proxiomo hi-c sequencing data, and gill tissue from a second female (T3F02_SC_FF) was used for generating PacBio HiFi long-read sequencing data to increase the total depth of coverage. One male individual (T3M02_BM_FF) was used for 10X Genomics linked-read and PacBio HiFi long read sequencing data (Table 1.2).

Linked-read library prep, sequencing & quality control

Genomic DNA for the male and female extractions were adjusted to a concentration of 0.9 ng/µl and 1.1 ng of template gDNA was loaded on a Chromium Genome Chip. 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. Briefly, gDNA was combined with Master

Mix, a library of Genome Gel Beads, and partitioning oil to create GEMs on a Chromium Genome Chip. The GEMs were isothermally amplified with primers containing an Illumina Read 1 sequencing primer, a unique 16-bp 10X barcode and a 6-bp random primer sequence, and barcoded DNA fragments were recovered for Illumina library construction. The amount and fragment size of post-GEM DNA was quantified by running 1 µl of sample on a Bioanalyzer 2100 with an Agilent High sensitivity DNA kit (Agilent, cat. 5067-4626). Prior to Illumina library construction, the GEM amplification product was sheared on an E220 Focused Ultrasonicator (Covaris, Woburn, MA) to approximately 350 bp (50 seconds at peak power = 175, duty factor = 10, and cycle/burst = 200). Then, the sheared GEMs were converted to a sequencing library following the 10X standard operating procedure.

The sequencing library was quantified by qPCR with a Kapa Library Quant kit (Kapa Biosystems-Roche) and sequenced on NovaSeq6000 sequencer (Illumina, San Diego, CA) to generate paired-end 150 bp reads. We used a previous inhouse RAD-sequencing-based estimate of a haploid delta smelt genome size of 0.6 Gb to sequence the first sample to an estimated 80x coverage. Because we successfully extracted HMW gDNA from a female first, we used the female linked-read sequencing data to improve our estimate of delta smelt genome size through a more accurate k-mer based approach using Genomescope2 (Vurture et al., 2017). After, we used the updated genome size estimate to adjust the amount of all subsequent sequencing data generated for assembly.

We evaluated linked-read sequencing data by looking at base quality metrics and for signs of contamination and sequencing bias. We used FastQC v0.11.9 (Andrews, 2010) to obtain raw sequencing data metrics such as per sequence quality scores, GC content, total number of

reads, average read length and number of bases. To quality control linked-read files for contamination and sequencing bias errors, we conducted three computational quality control steps (`kat hist`, `kat gcp`, and `kat comp`) using the software program KAT (Mapleson et al., 2017a). Each step splits sequencing data into sub-sequences of a given length (k-mers) to plot out frequencies or comparisons to visually inspect the data for quality issues. We looked for signs of bacterial and organellar DNA contamination using `kat hist` and `kat gcp` within the male and female sequencing data. First, we used `kat hist` to plot a histogram of the observed number of distinct k-mers at different frequencies for lengths k=21 and 31. Second, we used `kat gcp` to plot the GC content of each k-mer against the k-mer's frequency in the sequencing data and the number of distinct k-mers for a given GC count vs. frequency. We plotted GC counts against the frequency of k-mers of length k=21 and 31. Lastly, we evaluated the data for sequencing bias between the forward (R1) and reverse (R2) reads. We used `kat comp` to plot the frequency of a given k-mer in each of the paired-end sequence data files (R1 and R2) for k-mers of length k=21 and 31.

Long-read library prep, sequencing & quality control

Extracted HMW gDNA was sheared to roughly 17 kb using Diagenode's Megaruptor's (Diagenode, cat B06010001) long hydropores (Diagenode hydropores, cat E07010002). Sheared DNA was quantified by Quantus Fluorometer (Promega, cat #E6150) using a QuantiFluor® ONE dsDNA Dye assay (Promega, cat #E4871) and size distribution was checked by Agilent Femto Pulse (Agilent Technologies, cat P-0003-0817). Sheared gDNA was then

concentrated using AMPure PB beads (Pacific Biosciences, cat 100-265-900). Concentrated, sheared gDNA was quantified by Quantus Fluorometer (Promega, cat #E6150) using a QuantiFluor® ONE dsDNA Dye assay (Promega, cat #E4871). Low- or high-input PacBio HiFi library construction was carried out based on the amount of available concentrated sheared gDNA present for each sample.

High-input HiFi libraries were constructed using the SMRTbell® Express Template Prep Kit v2.0 (Pacific Biosciences, cat #100-938-900) with protocol "Procedure & Checklist - Preparing HiFi SMRTbell® libraries using SMRTbell® Express Template Prep Kit 2.0 v3, January 2020". We used sheared DNA as input for removal of single-strand overhangs at 37 °C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 minutes, End Repair and A-tailing at 20 °C for 10 minutes and 65 °C for 30 minutes, ligation of overhang adapter v3 at 20 °C for 1 hour and 65 °C for 10 minutes, and nuclease treatment of SMRTbell® library at 37 °C for 1 hour to remove damaged or non-intact SMRTbell® templates (SMRTbell® Enzyme Cleanup Kit, Pacific Biosciences, cat #107-746-400). The resulting SMRTbell® libraries were purified and concentrated by 0.45X AMPure PB beads (Pacific Biosciences, cat #100-265-900) then pooled for size selection using the SageELF system (Sage Science, cat #ELF0001). Input of the purified SMRTbell® library was used to load into the SageELF 0.75% Agarose Cassette (Sage Science, cat ELD7510) using cassette definition 0.75% 1-18 kb v2 for the run protocol. Fragments roughly 16 kb to 18 kb were collected from elution wells and the size-selected SMRTbell® library was purified and concentrated with 0.5X AMPure beads (Pacific Biosciences, cat 100-265-900).

When there were insufficient quantities of extracted DNA, low-input HiFi libraries were constructed using the SMRTbell® Express Template Prep Kit v2.0 (Pacific Biosciences, cat #100-

938-900) with protocol “Procedure & Checklist - Preparing HiFi SMRTbell® libraries from Low DNA Input using SMRTbell® Express Template Prep Kit 2.0 v6, June 2020”. We used sheared DNA as input for removal of single-strand overhangs at 37 °C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 minutes, End Repair and A-tailing at 20 °C for 10 minutes and 65 °C for 30 minutes, ligation of overhang adapters v3 at 20 °C for 1 hour. Low Input HiFi SMRTbell® library was purified and concentrated twice first by 1.8X AMPure PB beads (Pacific Biosciences, cat #100-265-900) and 40% diluted AMPure beads to remove < 3 kb SMRTbell® templates. Each high and low-input library was subsequently loaded onto a single 8M SMRT Cells and sequenced using a Sequel II sequencing plate 2.0 on Pacific Biosciences Sequel II sequencer.

We used PacBio’s CCS v3.3.0 (https://github.com/PacificBiosciences/ccs) statistical model on raw reads to generate base quality called circular consensus (CCS) reads and convert binary data to fastq format for downstream analysis. Reads with quality scores over Q20, denoting an error probability of 0.01% or less, were accepted and used for subsequent assembly.

Hi-C chromatin conformation capture library prep, sequencing & quality control

Female chromatin conformation capture sequencing data was generated by Phase Genomics (Seattle, WA) using the Proximo Hi-C 2.0 Kit, a commercially available version of the Hi-C protocol. Following the manufacturer's instructions for the kit, intact cells from the female sample were crosslinked using a formaldehyde solution, digested using the SAUIII restriction enzyme (cut site GATC), end repaired with biotinylated nucleotides, and proximity ligated to

create chimeric molecules composed of fragments from different regions of the genome that were physically proximal in vivo, but not necessarily proximal in DNA sequence. Continuing with the manufacturer's protocol, molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Finally, 150 bp paired-end reads were generated on an Illumina HiSeq sequencer. Raw sequencing data and an initial scaffolding report were received for the female sample. Due to the COVID-19 pandemic, we were unable to acquire male Hi-C sequencing reads, so we used the female Hi-C sequencing data for both male and female scaffolding.

We evaluated Hi-C sequencing data by looking at base quality and mapping metrics. We used FastQC v0.11.9 (Andrews, 2010) to obtain raw sequencing data metrics such as per sequence quality scores, GC content, total number of reads, average read length and number of bases. To assess if the Hi-C sequencing data would be useful in linking scaffolds, we 1) looked at a percentage of high-quality reads (minimum mapping quality of greater than or equal to 20, a maximum edit distance of less than or equal to 5, and no duplications) that mapped to our draft assembly created using only PacBio reads; and 2) observed the number of reads which aligned to each contig (> 600 desired) and the number of high-quality reads greater than 10 kb apart (1-15% expected).

Genome assembly

We generated an initial draft assembly (A_1) purged of duplicate haplotigs using the IPA HiFi Genome Assembler (ipa) v1.3.1 (https://github.com/PacificBiosciences/pbipa), with purge_dups

v1.2.3 (Guan et al., 2020) and Racon v1.4.13 (Vaser et al., 2017) wrappers enabled to generate phased primary and alternative assembly files polished of errors. A linked- and long read- (A_2) draft assembly was created using scaff10x (Ning, n.d.) with the following parameters: `-longread 1 -gap 100 -matrix 2000 -reads 10 -link 8 -score 20 -edge 50000 -block 50000` to first break the assembly at locations that were incorrectly joined and scaffold the assembly into larger, more contiguous sequencing segments. After linked-read scaffolding, we prepared the Hi-C data following the Arima mapping protocol (https://github.com/ArimaGenomics/mapping_pipeline) so interaction mapping information could be used to further scaffold the A_2 draft assembly. To prep the sequencing data for further scaffolding, we independently aligned paired-end Hi-C reads as single-ended reads to the A_2 assembly using BWA v0.7.17-r1188 (Li & Durbin, 2009) and samtools v1.7 (Li et al., 2009). Next, we retained the 5' end of the read to eliminate chimeric reads using a custom Arima perl script. Then, we paired the Hi-C reads to produce paired-end BAM files and used PicardCommandTools (https://github.com/broadinstitute/picard) to add read groups and remove PCR duplicates. After filtering the Hi-C sequencing data following parameters described in the sequencing and quality control sections, we converted BAM files to sorted BED files with BEDtools v2.29.2 (Quinlan & Hall, 2010). The A_2 draft assembly and BED files were input into SALSA2 (Ghurye et al., 2019) with non-default parameters (`-i 5 -x GATC -m yes`) to scaffold the A_2 assembly with the filtered Hi-C data to produce a linked, long, and Hi-C read (A_3) assembly. Finally, to anchor the A_3 assembly into chromosome-scale scaffolds we used chromonomer v1.13 (Catchen et al., 2020) in combination with a previously published delta smelt linkage map (Lew et al., 2015) to produce a chromosome-level reference genome (A_4) assembly.

Assembly quality assessment

After each step generated a draft assembly (A_1 - A_4), we evaluated the contiguity, content, and composition of the resulting fasta file. To assess each assembly's completeness, we used the evolutionarily informed Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.6 (Simão et al., 2015) Actinopterygii lineage (actinopterygii_odb10) dataset in genome mode.

To acquire assembly metrics, such as N50, L50, number of contigs, and assembly length, we used GenomeTools v1.5.10 (Gremme et al., 2013). Read length histograms were generated using jellyfish (Marçais & Kingsford, 2011). All assemblies within each sex and between sexes were compared using QUAST v5.2.0 (Gurevich et al., 2013).

Karyotypic chromosome validation

We carried out two rounds of organ harvesting for cytogenetic chromosome evaluation by karyotyping. All delta smelt were obtained from the FCCL and were transported and handled according to a UC Davis IACUC-approved animal care protocol (#21533) and standard operating procedures at the UC Davis Center for Aquatic Biology and Aquaculture (CABA). The first harvest involved fifteen 240 day post hatch subadult delta smelt and the second harvest used phenotypically sexed female (N = 13) and male (N = 15) adults (360 day post hatch). Prior to colchicine injections, fish were held in 140 L aerated tanks at 12 °C. Delta smelt (N = 28) were removed from aerated tanks at CABA, anesthetized, measured (total length), injected (i.p.) with

colchicine (10 µL of 1 mg/mL stock) to arrest cells in metaphase, and immediately returned to saline water (0.4 ppt) in 5-gallon temperature-controlled buckets held at 12°C. Adult fish were separated according to sex. Fish were euthanized and a pool of organs (spleen, head kidney, heart and/or gonads) were collected post colchicine injection. For the first harvest, where sex could not be visually determined, organs were collected 4 hours post colchicine injection. For the second harvest, organs were sampled 2 hours post colchicine injection. All organs were rinsed, stored in PBS, and processed within 2 hours of dissection at ambient temperature of the CABA environment.

We established three pools of organs: a subadult mixed sex and mixed organ pool, a female spleen pool, and a male spleen and gonad pool. Organ pools were taken out of PBS, gently minced, and pipette-aspirated into single cell suspensions in a hypotonic solution (0.56% KCl) for 15-20 minutes. The cell suspensions were centrifuged at ~1000 rpm for 10 min, supernatant hypotonic solution was removed and a 3:1 fixative (methanol:glacial acetic acid) was added. Cell pellets were resuspended and stored at 4°C. Two to three more fresh fixative treatments (centrifugation, resuspension in new fixative) were conducted, and cells were applied to slides using an air-dry method one week later. Slides were stained using the DAPI-Vectashield fluorescent dye and cell nuclei were examined using an Olympus BX-40 Microscope. Images of mitotic metaphase cells were captured and stored using the CytoVision Software (v7.4) and the number of chromosomes in the species was determined from images with intact cells with clearly defined and nonoverlapping chromosomes. Eight and fifteen cell slides were prepared from the first and second harvest, respectively. Seventy-five cell images were collected from the three pooled sample sets (mixed sex, males-only, females-only).

Results

Sample collection & DNA extraction

HMW gDNA extraction from back muscle tissue flash frozen and stored unsuspended in liquid yielded mixed results, so we expanded our sampling and storage methods through the additional collection of scale and internal organ tissue, and by storing samples of back muscle tissue in propylene glycol. However, we did not find that suspending flash frozen back muscle in propylene glycol provided more success in the yield of HMW gDNA. As such, all HMW gDNA extractions were from tissues not suspended in any kind of solution after flash freezing. We used back muscle, internal organ and/or scale tissue samples from two female individuals and two male individuals to extract roughly 3.4 µg of HMW gDNA at a concentration of 87 ng/µL for subsequent sequencing (Figure 1.2, Table 1.2). NanoDrop 260/280 absorbance ratios of HMW extractions were between 1.80 to 1.91, and NanoDrop 260/230 ratios ranged from 1.73 to 2.22 (Table 1.2)

Linked-read sequencing & quality control

Post-GEM DNA library electropherograms were quantified and displayed expected distributions (Figure 1.3). We generated 94,825,601,818 bp of paired-end linked-read sequencing data from the female specimen. Using the Genomescope2 k-mer based haploid genome size estimation of the female 10X sequencing data, we estimated the delta smelt genome size to be 0.49 Gb. We used the updated k-mer based estimate to reduce the amount of male data generated in linked- read sequencing and in total, 65,806,680,934 bp of male linked-read sequencing data

were generated. The average per sequence base quality was 33 and 32 in the female R1 and R2 fastq files, respectively and 34 and 32 in the male R1 and R2 fastq files, respectively (Table 1.3).

Mapped k-mer histograms for each sample and at each value of k showed discrete, single peaks indicating no sign of contamination (Figure 1.4). All GC count frequency plots show roughly normal circular distributions of distinct k-mers with no aberrant spotting (Figure 1.5). Additionally, the number of distinct k-mers does not appear to be heavily skewed, indicating no sequencing bias, in the male or female sequencing (Figure 1.6). These data together indicate no observable signs of bacterial or organellar DNA contamination or major sources of sequencing bias in the linked-read sequencing data.

Long-read sequencing & quality control

In order to obtain sufficient sequencing data we created two high-input libraries from one male individual (T3M02_BM_FF) and two high-input and one low-input library from a second female individual (T3F02_SC_FF) (Table 1.2). Starting gDNA inputs ranged from 6.5 ug to 20 ug of gDNA. The sheared gDNA input for the removal of single strand overhangs ranged from 1000 ng to 7 ug, and the average length of gDNA for sequencing ranged from 14-18.4 kb.

Five movie collections (150 hours of sequencing data) from two male and one female high-input library, and two low-input female library runs were collected. A total of 3,095,133 male reads

and 2,741,504 female reads representing 35,841,976,770 and 28,549,585,055 base pairs, respectively, passed quality control and were used for subsequent assembly (Table 1.3).

Hi-C sequencing & quality control

Hi-C sequencing files contained 87,444,477 read pairs in total (Table 1.3). The data contained an average of 2,966 read pairs per contig greater than 5 kb, 18.78% of the read pairs mapped to greater than 10 kb apart and 56.38% of reads were considered high quality indicating successful library prep and sequencing. The average per sequence base quality was 38 and 36 in the R1 and R2 fastq files, respectively (Table 1.3).

Assembly quality assessment

We searched raw data and each iteration of the assemblies for 3,640 conserved single-copy orthologs contained within the 05 August 2020 Actinopterygii lineage dataset using BUSCO. The quality filtered female and male HiFi data contained whole genes or sequence fragments of 95.6% (3.3% complete single copy, 89.3% complete double copy, and 3.0% fragmented) and 94.4% (3.4% complete single copy, 87.0% complete double copy, and 4.0% fragmented) of the conserved Actinopterygii gene dataset, respectively (Table 1.4).

After each step of the assembly the total length and N50 increased, and the L50 and total number of contigs decreased (Figures 1.7 & 1.8). Female HiFi sequencing data had an N50 of 15,048 and an L50 of 771,808, while the male HiFi data had an N50 of 11,604 and an L50 of 1,276,120. The

final female assembly contained 89.3% complete (87.7% single copy and 1.6% double copy) genes and fragments of an additional 0.8% of conserved genes, had an N50 of 14,850,352, L50 of 13, and was a total of 437,273,953 bp long with a total of 376 contigs. The final male assembly contained 88.4% complete (81.2% single copy and 7.2% double copy) genes and fragments of an additional 1.0% of conserved genes, had an N50 of 12,200,365, L50 of 15 and was a total of 472,157,411 bp long with a total of 549 contigs (Table 1.4).

Karyotypic chromosome validation

Organs from a total of 43 fish, comprised of subadults (n=15) and adults (n=28), were sampled for cytogenetic analyses. Subadult and adult total body lengths ranged from 5 to 7.6 cm (mean = 6.4 cm) and 7 to 10.2 cm (mean = 8.9 cm), respectively. After quality control filtration to retain only images with intact cells with clearly defined and nonoverlapping chromosomes, 18 cell images were kept for counting analysis. We determined the diploid (2n) chromosome count for the delta smelt to be 56, with 15 cells exhibiting 2n=56 and 3 cells with hypomodal counts (1 cell with 2n=54, 2 cells with 2n=55) (Figure 1.9).

Discussion & Conclusion

The primary objective of this chapter was to create a highly contiguous chromosome-scale *de novo* delta smelt genome assembly for use within and beyond the scope of this dissertation. We have achieved two chromosome-scale reference genomes, one for a female and one for a male delta smelt, and a chromosome count of 2n=56 for the species was independently validated by

sequencing-free karyotyping. Our two reference genomes were published to NCBI with GenBank assembly accession numbers GCA_021917145.1 (female) and GCA_021870715.1 (male) on February 02, 2022, and February 03, 2022, respectively and the more contiguous female genome was annotated by the NCBI Eukaryotic Genome Annotation Pipeline.

The final total lengths for the female and female assemblies were 0.44 Gb and 0.47 Gb, respectively. These total lengths are similar to the wakasagi smelt genome (*Hypomesus nipponensis*) which has a total length of 0.50 Gb (Xuan et al., 2021). Our final female and male assemblies had 376 and 549 scaffolds with N50's of 0.15 Gb and 0.12 Gb, respectively. The first 28 contigs, representing the number of haploid chromosomes confirmed by karyotyping, contained 81.6% and 73.3% of the total assembled sequences in female and males, respectively. The delta smelt reference assemblies are roughly 25-30 times more contiguous than the previously published *H. niponnensis* assembly and our final female contained 89.3% and final male assembly contained 88.4% of core genes expected in the *Actinopterygii* BUSCO database. As such, our reference genomes provide a strong foundation for the future of delta smelt and evolutionary genomic research at this time.

The male assembly is roughly 0.03 Gb, or 8.0% longer than the female assembly and has a 5.6% increase of double-copy genes. These double-copy genes may account for the longer assembly length. Alternatively, or perhaps additionally, the male genome may have male specific sequences, such as a sex chromosome, which we could not detect in our cytogenetic work.

The diploid chromosome number of 56 for delta smelt revealed by our karyotyping aligns with those reported for other smelt species, 2n=54, 56 or 58 for European smelt (Nygren et al., 1971;

Ocalewicz et al., 2007) and 2n=56 for wakasagi (pond) smelt (Kitada et al., 1980). As others have noted, Robertsonian fusions/fissions of chromosomes (acrocentrics fusing to form metacentrics or vice versa) may be a source of the karyotype variation in the Osmeridae, as observed within and among salmonid species (Hartley, 1987; Ocalewicz et al., 2007). No sex chromosomes have been reported to date for any smelts studied cytogenetically, and here we found no evidence for sex-specific chromosomes although a more detailed study is necessary given the small sample size and low resolution of karyotype images. Similar to reports from other smelt species, we note a preponderance of subtelocentric/acrocentric chromosome pairs over metacentric chromosome pairs. Chromosome composition is a descriptive metric, and our findings did not alter or affect our genome assembly process.

This high-quality delta smelt reference genome is the most contiguous assembled Osmeridae genome, and provides a valuable resource for smelt conservation research. It is the most contiguous assembled smelt genome publicly available with timely broad and specific uses. It has been immediately useable for identifying polygenic adaptive genetic variants such as the species' complex migratory phenotypes (Campbell et al., 2022), informing hatchery management strategies, and investigating mechanisms driving domestication of the FCCL population (Habibi, 2022; Habibi et al., in prep). Further this genomic resource will be useful for investigating hybridization with wakasagi smelt, the genetic basis for sex determination, and domestication effects of the refuge population. On a broad level, this genomic resource allows for future evolutionary theory development by preserving a record of an imperiled species' genome and current diversity in light of all time high rates of species loss and the persistence of the principal drivers of current and future biodiversity loss: increasingly severe drought,

temperatures, and habitat loss (Caro et al., 2022; East & Sankey, 2020; Strona & Bradshaw, 2022; Ullrich et al., 2018).

Chapter 2 – Investigation in Identifying Sex-Specific Markers in Delta Smelt

Introduction

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 can switch sexes within a life cycle (Bachtrog et al., 2014; Baroiller & D'Cotta, 2016; Kobayashi et al., 2013; Nakamura et al., 1998; Volff, 2005). Known influences for environmental sex determination (ESD) include population or social dynamics, temperature, sex ratio, pH, background color, and salinity, and sex reversal can occur throughout the lifespan of a fish (D. R. Robertson, 1972; Shen & Wang, 2018; Tenugu & Senthilkumaran, 2022; Uhlenhaut et al., 2009). Within genetic sex determination (GSD), sex is resolved upon the fusion of gametes. Chromosomal (heterogametic males (XY) or females (ZW)) or genic (female- or male-specific master sex determining regulators) mechanisms drive the primary sexual development and gonadal output of individuals with GSD (Bhattacharya & Modi, 2021; Devlin & Nagahama, 2002; Guiguen et al., 2018). 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 effective management practices such as the ability to utilize ESR strategies to produce desired sex ratios in captive populations or to non-lethally sex fish at all life stages, regardless of gametic expression (Stelkens & Wedekind, 2010).

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), sex-specific dispersal patterns (Hutchings & Gerber, 2002), parental condition (Trivers & Willard, 1973), and sex-biased harvesting (B. C. Robertson et al., 2006), to name a few. Sex-ratio bias within small, isolated populations can arise through demographic stochasticity and contribute to increased risk of extinction (Lande, 1993). Skewed sex ratios can have discrete consequences for various mating systems, as they contribute to deviations from Hardy-Weinberg Equilibrium thereby influencing genotype frequency change within populations. For polygynous species males that mate with many females have larger reproductive success than any individual female, while within monogamous species a deficit in one sex results in reduced opportunities for reproduction in the other sex. Such non-random mating contributes to a decrease in the number of breeding individuals (i.e. N_e) and a higher susceptibility to the effects of inbreeding depression, genetic drift and reductions in fitness (Frankham, 2005; Hedrick & Garcia-Dorado, 2016; Kardos et al., 2016). Additionally, male sex-bias within wild populations, especially small populations, can 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). Because delta smelt have a small 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. Despite the State of California's active monitoring of the wild delta smelt population abundance and distribution, the inability to identify the sex of fish at all life stages leaves an important metric of population dynamics unknown. Because mechanisms for sex determination vary between closely related species and within different populations of a single species, an investigation into causative mechanisms and a search for diagnostic markers must be performed at the individual species level (Conover & Kynard, 2013; Devlin & Nagahama, 2002; Kobayashi et al., 2013; Mank & Avise, 2009; Nakamura et al., 1998; Volff, 2005; Volff & Schartl, 2001).

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(s) 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 individuals in the wild for the practical management of a listed species.

This chapter focuses on investigating methods to assign the sex of wild and captive delta smelt relatively non-invasively through identifying genetic markers. We sought to identify candidate loci associated with sex using three methods: 1) a genome-wide association study, 2) read depth analysis, and 3) k-mer analysis. The genome-wide association study uses a reference genome and RAD-sequencing data to look for SNPs diagnostic of sex. Our read depth analysis investigates RAD-sequencing data to probe for read depth disparities between females and males which would be expected in chromosome-based sex determination. And finally, our k-mer analysis is a reference genome free investigation into sequence differences between female and male individual's linked-read data.

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 one modification, eluting in 100 uL of deionized water rather than the proprietary AE Buffer included with the kit. To sample a broad distribution of loci throughout

the genome, we digested DNA using the *Pst*I restriction enzyme. RAD-sequencing libraries were prepared according to 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. We 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` (https://raw.githubusercontent.com/shannonekj/DS_sex-

marker/master/scripts/doAssoc_LRT.sh). 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 using the formula $p = \frac{\alpha}{n}$ where α is the desired p-value or significance threshold (α = 0.05), n is the number of loci analyzed (n varies depending on number of RAD tags post-filtration), and p 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-tags. We performed two experiments – the first using our female reference genome and second using our male reference genome 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 (`samtools depth -aa`) using samtools v1.9 (Li et al., 2009). Next, we discarded loci with zero coverage in both sexes and compared the ratio of the mean depth for females and males at each locus (https://raw.githubusercontent.com/shannonekj/DS_sex-marker/master/scripts/get_depth_24v24.pl). To identify locations in the genome where one sex exhibited consistently high coverage and the opposite sex exhibited less than or equal to half of the opposite sex's depth, we looked for high-fidelity regions greater than 5,000 bp exhibiting a sex coverage ratio greater than or equal to two.

K-mer analysis

Next, we used a k-mer based approach to look for unique differences of sequence content in males versus females. First, we created and filtered sex-specific sequence signatures from the female and male individuals' linked-read sequence data generated for our genome assemblies. Next, 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). After, 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. Finally, 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 containing five or more k-mers from the A_1 version of the delta smelt assembly (Joslin et al., in prep). We used the A_1 assembly to acquire contigs with moderate contiguity compared to the final reference genome (Table 1.4). 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. Lastly, 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.

We ran a depth analysis on RAD-seq reads found within the putative Y contigs to look for RAD-tags which confirmed the presence of male-specific sequences in individuals beyond the single female and male used in our k-mer analysis. 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 v2.5.0 (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 v2.29.2 (Quinlan & Hall, 2010). After obtaining depth information across all putative Y regions, we ran the same depth analysis described above.

Results

Sample collection, DNA extraction & sequencing

We acquired RAD-sequencing data from a total of 48 (24 female and 24 male) captive-bred delta smelt. The average Phred score for all reads was 39 and mean 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 for RAD-seq reads 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. With these loci we calculated a Bonferroni corrected p-value cutoff of 5.417265e-08 and 5.893141e-08 required for significance of associations found within the female and male reference genome, respectively. No loci were found to be significantly associated with sex in the female reference genome (Figure 2.1). 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 2.1, Table 2.1).However, the genotypes at these loci were not diagnostic of sex (Table 2.2). Despite individual alleles at each locus being highly associated with sex, female and male sequencing data contained individuals that were homozygous for the major and minor alleles and heterozygous in significant proportions (Table 2.3)

Depth analysis

After removal of loci with zero coverage, we carried out depth analyses using 92,808 and 92,735 RAD loci aligned to the female and male reference genome, respectively. In both analyses we found no areas longer than 5 kb with 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 1.284592e9 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 2.2). 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 of 50-100 we found 4,964 hashes corresponding to approximately 4,964,000 high abundance k-mers (Figure 2.3).

Upon filtering for k-mers in the A_1 assembly, a total of 2,067 hashes, or 2,067,000 k-mers, were found on A_1 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 2.4). 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 2.5). 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

Knowing population demographic information will lead to informed management decisions to best support recovery efforts for imperiled delta smelt. 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 the sex of a delta smelt can only be determined non-lethally by the expression of reproductive cells (i.e. eggs or sperm), where pressure is put on the abdomen of fish until eggs (in females) or running milt (in males) are excreted (Lindberg et al., 2013). Dissection and visual inspection of gonads represents a lethal method to identify sex. 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.). This presents a hurdle in studying the wild population and for rearing fish in captivity. Within the wild population, the ability to genetically sex fish through non-lethal fin clip sampling, without culling or relying on gametic expression, will allow ecologists to reliably sex fish at all stages of their life cycle without reducing the species' population size. Genetic identification of sex in the captive refuge population would allow for fish to be sexed at all stages of their lifecycle, and allowing managers to factor population demographics into breeding decisions. While state and federal agencies conduct annual abundance and distribution monitoring throughout the San Francisco Estuary (SFE) at different stages of delta smelt development, knowledge of sex-ratios throughout their lifecycle is currently unknown. Since sex ratio bias has the potential to significantly alter the success of the species, identifying the genetic underpinnings of sex determination within delta smelt and developing a genetic sex marker to identify sex would allow managers to make more informed decisions and better understand the influences affecting the fate of the wild

population. Such information would inform decisions on how to best utilize the captive breeding program to reduce genetic bottlenecks associated with skewed sex bias. The ability to non-lethally sex fish opens up the door for managers to better understand current and past population dynamics, carry out species modeling, and test the association of sex-specific behavior, geographic location, salinity, and life stage to identify vulnerable subgroup(s) which can then be classified as high priority for conservation efforts, protection, and future research (Marchi et al., 2021; Martínez et al., 2014).

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. While GWAS results identified two SNPs highly associated with sex on Chromosome 5, neither was perfectly correlated with sex and cannot be used as a diagnostic marker for applications in the field. 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 sequences near *Pst*I restriction enzyme cut sites, our data may not have adequately sampled genetic material in delta smelt's sex determining region entirely as the cleavage ability for *Pst*I to cut at its recognition site is influenced by the surrounding sequences (i.e. repetitive sequences or GC content), and proximity to ends in linear DNA as well as other cut sites. Lack of adequate coverage in the area of interest could easily result in inconclusive results as the genomes of other fish species 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 *Pst*I 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 2.4 and the abundance of k-mers only contained within the male sequencing data shown in Figure 2.5 may indicate 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 the 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. Interestingly, sex differentiation in *Plecoglossus altivelis* (sweetfish/ayu), another fish species in the Osmeriformes order which has XX females and XY males, is controlled by an *amhr2* paralog (*amhr2bY*) located on the Y chromosome (Nakamoto et al., 2021). The current published and annotated version of the delta smelt genome, which is the female assembly, does not contain copies of the *amhr2* or *amhr2bY* genes. However, the published delta smelt genome does

contain the *amh* gene, but it is located on an unplaced scaffold (chrUn_NW_025813713v1). Notably the unplaced scaffold also contains several genes (*fkbp8, ELL, DOT1L*) known to be in close proximity, within or linked to sex determining regions of other fish (Eshel et al., 2012; Rodríguez-Marí et al., 2005; Triay et al., 2020).

Our work shows a need for further investigation using high-coverage whole-genome resequencing (WGS) data from a large cohort of male and female delta smelt to survey the genome more evenly in hopes of identifying sex-specific markers. While RAD-sequencing data interrogates hundreds of thousands of discrete locations throughout individuals' genomes, it unevenly samples genomes as it is dependent upon the location of restriction enzyme 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 sequencing coverage is not dependent upon the sequence of the individual. Furthermore, including a larger 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.

CHAPTER 1 & 2 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

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* (0.11.9) [Computer software]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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

Barnard, P. L., Schoellhamer, D. H., Jaffe, B. E., & McKee, L. J. (2013). Sediment transport in the San Francisco Bay Coastal System: An overview. *Marine Geology*, *345*, 3–17. https://doi.org/10.1016/j.margeo.2013.04.005

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

Belarmino, E., Nóbrega, M. F. de, Grimm, A. M., Copertino, M. da S., Vieira, J. P., & Garcia, A. M. (2021). Long-term trends in the abundance of an estuarine fish and relationships with El Niño climatic impacts and seagrass meadows reduction. *Estuarine, Coastal and Shelf Science*, *261*, 107565. https://doi.org/10.1016/j.ecss.2021.107565

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

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., Ponce de León, F. A., … Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, *49*(4), 643–650. https://doi.org/10.1038/ng.3802

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

Campbell, M. A., Joslin, S. E. K., Goodbla, A. M., Willmes, M., Hobbs, J. A., Lewis, L. S., & Finger, A. J. (2022). Polygenic discrimination of migratory phenotypes in an estuarine forage fish. *G3 Genes|Genomes|Genetics*, *12*(8), jkac133. https://doi.org/10.1093/g3journal/jkac133

Caro, T., Rowe, Z., Berger, J., Wholey, P., & Dobson, A. (2022). An inconvenient misconception: Climate change is not the principal driver of biodiversity loss. *Conservation Letters*, *15*(3), e12868. https://doi.org/10.1111/conl.12868

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

Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: A Tool Set for Repairing and Enhancing Assembled Genomes Through Integration of Genetic Maps and Conserved

Synteny. *G3: Genes|Genomes|Genetics*, *10*(11), 4115–4128. https://doi.org/10.1534/g3.120.401485

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., North, K. N., Plon, S. E., Rehm, H. L., Risch, N., Rotimi, C. N., Shendure, J., Soranzo, N., & McCarthy, M. I. (2020). A brief history of human disease genetics. *Nature*, *577*(7789), 179–189. https://doi.org/10.1038/s41586-019-1879-7

Conomos, T. J. (1979). *San Francisco Bay—The urbanized estuary*. American Association for the Advancement of Science- Pacific Division.

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.

Cottingham, A., Huang, P., Hipsey, M. R., Hall, N. G., Ashworth, E., Williams, J., & Potter, I. C. (n.d.). *Growth, condition, and maturity schedules of an estuarine fish species change in estuaries following increased hypoxia due to climate change*. 20.

Covaris. (2017). *E/LE220 Series SonoLab 7 USER MANUAL*. https://cdn.shopify.com/s/files/1/0377/2163/6908/files/pn_010277.pdf?v=1632901354

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

East, A. E., & Sankey, J. B. (2020). Geomorphic and Sedimentary Effects of Modern Climate Change: Current and Anticipated Future Conditions in the Western United States. *Reviews of Geophysics*, *58*(4), e2019RG000692. https://doi.org/10.1029/2019RG000692

Eshel, O., Shirak, A., Weller, J. I., Hulata, G., & Ron, M. (2012). Linkage and Physical Mapping of Sex Region on LG23 of Nile Tilapia (*Oreochromis niloticus*). *G3 Genes|Genomes|Genetics*, *2*(1), 35–42. https://doi.org/10.1534/g3.111.001545

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., … Zhang, G. (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature*, *587*(7833), 252–257. https://doi.org/10.1038/s41586-020-2873-9

Finger, A. J., Mahardja, B., Fisch, K. M., Benjamin, A., Lindberg, J., Ellison, L., Ghebremariam, T., Hung, T.-C., & May, B. (2018). A Conservation Hatchery Population of Delta Smelt Shows Evidence of Genetic Adaptation to Captivity After 9 Generations. *Journal of Heredity*, *109*(6), 11.

Fisch, K. M., Henderson, J. M., Burton, R. S., & May, B. (2011). Population genetics and conservation implications for the endangered delta smelt in the San Francisco Bay-Delta. *Conservation Genetics*, *12*(6), 1421–1434. https://doi.org/10.1007/s10592-011-0240-y

Frankham, R. (2005). Genetics and extinction. *Biological Conservation*, *126*(2), 131–140. https://doi.org/10.1016/j.biocon.2005.05.002

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

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, *15*(8), e1007273. https://doi.org/10.1371/journal.pcbi.1007273

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

Glibert, P. M., Fullerton, D., Burkholder, J. M., Cornwell, J. C., & Kana, T. M. (2011). Ecological Stoichiometry, Biogeochemical Cycling, Invasive Species, and Aquatic Food Webs: San Francisco Estuary and Comparative Systems. *Reviews in Fisheries Science*, *19*(4), 358–417. https://doi.org/10.1080/10641262.2011.611916

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

Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *10*, 645–656. https://doi.org/10.1109/TCBB.2013.68

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Guiguen, Y., Fostier, A., & Herpin, A. (2018). Sex Determination and Differentiation in Fish: Genetic, Genomic, and Endocrine Aspects. In H. Wang, F. Piferrer, S. Chen, & Z. Shen (Eds.), *Sex*

Control in Aquaculture (1st ed., pp. 35–63). Wiley. https://doi.org/10.1002/9781119127291.ch2

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Habibi, E. (2022). *Conservation genomics of two California endangered native species: Delta smelt and Upper McCloud River Redband trout.* [PhD]. University of California Davis.

Hammock, B. G., Hartman, R., Dahlgren, R. A., Johnston, C., Kurobe, T., Lehman, P. W., Lewis, L. S., Van Nieuwenhuyse, E., Ramírez-Duarte, W. F., Schultz, A. A., & Teh, S. J. (2022). Patterns and predictors of condition indices in a critically endangered fish. *Hydrobiologia*, *849*(3), 675–695. https://doi.org/10.1007/s10750-021-04738-z

Hartley, S. E. (1987). THE CHROMOSOMES OF SALMONID FISHES. *Biological Reviews*, *62*(3), 197–214. https://doi.org/10.1111/j.1469-185X.1987.tb00663.x

Hedrick, P. W., & Garcia-Dorado, A. (2016). Understanding Inbreeding Depression, Purging, and Genetic Rescue. *Trends in Ecology & Evolution*, *31*(12), 940–952. https://doi.org/10.1016/j.tree.2016.09.005

Hobbs, J. A., Lewis, L. S., Willmes, M., Denney, C., & Bush, E. (2019). Complex life histories discovered in a critically endangered fish. *Scientific Reports*, *9*(1), 16772. https://doi.org/10.1038/s41598-019-52273-8

Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences*, *118*(52), e2109019118. https://doi.org/10.1073/pnas.2109019118

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

Illumina. (2022). *Illumina NovaSeq 6000 Sequencing System Guide*.

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

James, N. C., Cowley, P. D., & Whitfield, A. K. (2018). The marine fish assemblage of the East Kleinemonde Estuary over 20 years: Declining abundance and nursery function? *Estuarine, Coastal and Shelf Science*, *214*, 64–71. https://doi.org/10.1016/j.ecss.2018.09.010

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

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, *9*(10), 1205–1218. https://doi.org/10.1111/eva.12414

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

Kitada, J., Tatewaki, R., & Tagawa, M. (1980). Chromosomes of the pond smelt Hypomesus transpacificus nipponensis. *Chrom. Inform. Serv.*, *28*, 8–9.

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

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

Lew, R. M., Finger, A. J., Baerwald, M. R., Goodbla, A., May, B., & Meek, M. H. (2015). Using Next-Generation Sequencing to Assist a Conservation Hatchery: A Single-Nucleotide Polymorphism Panel for the Genetic Management of Endangered Delta Smelt.

Transactions of the American Fisheries Society, *144*(4), 767–779. https://doi.org/10.1080/00028487.2015.1037016

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

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

Mapleson, D., Accinelli, G. G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017b). *Sequence analysis KAT : a K-mer analysis toolkit to quality control NGS datasets and genome*

assemblies. *33*(November 2016), 574–576. https://doi.org/10.1093/bioinformatics/btw663

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Marchi, N., Schlichta, F., & Excoffier, L. (2021). Demographic inference. *Current Biology*, *31*(6), R276–R279. https://doi.org/10.1016/j.cub.2021.01.053

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

McAllister, D. E. (1963). *Revision of the smelt family, Osmeridae.* National Museum of Canada, Biological Series 191.

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

Moyle, P. B. (2002). *Inland Fishes of California* (2nd ed.). University of California Press.

Moyle, P. B., Herbold, B., Stevens, D. E., & Miller, L. W. (1992). Transactions of the American Fisheries Society Life History and Status of Delta Smelt in the Sacramento-San Joaquin Estuary , California. *Transactions of the American Fisheries Society*, *121*(1), 67–77. https://doi.org/10.1016/j.fertnstert.2012.04.042

Moyle, P. B., Hobbs, J. A., & Durand, J. R. (2018). Delta Smelt and Water Politics in California. *Fisheries*, *43*(1), 42–50. https://doi.org/10.1002/fsh.10014

Moyle, Peter B., Brown, Larry R., Durand, John R., Hobbs, J. A. (2016). Delta Smelt: LIfe History and Decline of a Once-Adundant Species in the San Francisco Estuary. *San Francisco Estuary and Watershed Science*, *14*(2), 1–40.

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

Nakamoto, M., Uchino, T., Koshimizu, E., Kuchiishi, Y., Sekiguchi, R., Wang, L., Sudo, R., Endo, M., Guiguen, Y., Schartl, M., Postlethwait, J. H., & Sakamoto, T. (2021). A Y-linked anti-Müllerian hormone type-II receptor is the sex-determining gene in ayu, Plecoglossus altivelis. *PLOS Genetics*, *17*(8), e1009705. https://doi.org/10.1371/journal.pgen.1009705

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.

Nichols, F. H., Cloern, J. E., Luoma, S. N., & Peterson, D. H. (1986). The Modification of an Estuary. *Science*, *231*(4738), 567–573. https://doi.org/10.1126/science.231.4738.567

Ning, Z. (n.d.). *Scaff10X: Pipeline for scaffolding and breaking a genome assembly using 10x genomics linked-reads* [Computer software].

Nygren, A., Nilsson, B., & Jahnke, M. (1971). Cytological studies in the smelt (Osmerus eperlanus L.). *Hereditas*, *67*(2), 283–286. https://doi.org/10.1111/j.1601-5223.1971.tb02381.x

Ocalewicz, K., Hliwa, P., Krol, J., Rábová, M., Stabinski, R., & Ráb, P. (2007). Karyotype and chromosomal characteristics of Ag–NOR sites and 5S rDNA in European smelt, Osmerus eperlanus. *Genetica*, *131*(1), 29–35. https://doi.org/10.1007/s10709-006-9110-9

Patrick, H. J. H., Chomič, A., & Armstrong, K. F. (2016). Cooled Propylene Glycol as a Pragmatic Choice for Preservation of DNA From Remote Field-Collected Diptera for Next-Generation Sequence Analysis. *Journal of Economic Entomology*, *109*(3), 1469–1473. https://doi.org/10.1093/jee/tow047

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*.

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

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

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

Robertson, D. R. (1972). Social Control of Sex Reversal in a Coral-Reef Fish. *Science*, *177*(4053), 1007–1009. https://doi.org/10.1126/science.177.4053.1007

Rodríguez-Marí, A., Yan, Y.-L., BreMiller, R. A., Wilson, C., Cañestro, C., & Postlethwait, J. H. (2005). Characterization and expression pattern of zebrafish anti-Müllerian hormone (amh) relative to sox9a, sox9b, and cyp19a1a, during gonad development. *Gene Expression Patterns*, *5*(5), 655–667. https://doi.org/10.1016/j.modgep.2005.02.008

Rosenfield, J. A., & Baxter, R. D. (2007). Population Dynamics and Distribution Patterns of Longfin Smelt in the San Francisco Estuary. *Transactions of the American Fisheries Society*, *136*(6), 1577–1592. https://doi.org/10.1577/T06-148.1

Sage Sciences. (2014). *User Manual Pippin Pulse Electrophoresis Power Supply*. https://www.sagescience.com/wp-content/uploads/2014/01/Pippin-Pulse-User-Manual-RevH.pdf

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

Shen, Z., & Wang, H. (2018). Environmental Sex Determination and Sex Differentiation in Teleosts – How Sex Is Established. In H. Wang, F. Piferrer, S. Chen, & Z. Shen (Eds.), *Sex Control in Aquaculture* (1st ed., pp. 85–115). Wiley. https://doi.org/10.1002/9781119127291.ch4

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sommer, T., Armor, C., Baxter, R., Breuer, R., Brown, L., Chotkowski, M., Culberson, S., Feyrer, F., Gingras, M., Herbold, B., Kimmerer, W., Mueller-Solger, A., Nobriga, M., & Souza, K. (2007). The Collapse of Pelagic Fishes in the Upper San Francisco Estuary: El Colapso de los Peces Pelagicos en La Cabecera Del Estuario San Francisco. *Fisheries*, *32*(6), 270–277. https://doi.org/10.1577/1548-8446(2007)32[270:TCOPFI]2.0.CO;2

Sommer, T., Mejia, F., Nobriga, M., Feyrer, F., & Grimaldo, L. (2011). The Spawning Migration of Delta Smelt in the Upper San Francisco Estuary. *San Francisco Estuary and Watershed Science*, *9*(2), 1–44.

State of California FMWT. (2022). *2022 Fall Midwater Trawl annual fish abundance and distribution summary* (pp. 1–16) [Annual Survey Results Memorandum]. State of California Department of Fish and Wildlife. https://wildlife.ca.gov/Conservation/Delta/Fall-Midwater-Trawl

State of California SKT. (2022). *2022 Spring Kodiak Trawl Delta Smelt Index* (pp. 1–2) [Annual Survey Results Memorandum]. State of California Department of Fish and Wildlife. https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentId=204731&inline

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

Strona, G., & Bradshaw, C. J. A. (2022). Coextinctions dominate future vertebrate losses from climate and land use change. *Science Advances*, *8*(50), eabn4345. https://doi.org/10.1126/sciadv.abn4345

Tenugu, S., & Senthilkumaran, B. (2022). Sexual plasticity in bony fishes: Analyzing morphological to molecular changes of sex reversal. *Aquaculture and Fisheries*, *7*(5), 525–539. https://doi.org/10.1016/j.aaf.2022.02.007

Thermo Scientific. (2010). *NanoDrop 1000 Spectrophotometer User's Manual*. https://assets.thermofisher.com/TFS-Assets/CAD/manuals/nd-1000-v3.8-users-manual-8%205x11.pdf

Triay, C., Conte, M. A., Baroiller, J.-F., Bezault, E., Clark, F. E., Penman, D. J., Kocher, T. D., & D'Cotta, H. (2020). Structure and Sequence of the Sex Determining Locus in Two Wild Populations of Nile Tilapia. *Genes*, *11*(9), 1017. https://doi.org/10.3390/genes11091017

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

Uhlenhaut, N. H., Jakob, S., Anlag, K., Eisenberger, T., Sekido, R., Kress, J., Treier, A.-C., Klugmann, C., Klasen, C., Holter, N. I., Riethmacher, D., Schütz, G., Cooney, A. J., Lovell-Badge, R., & Treier, M. (2009). Somatic Sex Reprogramming of Adult Ovaries to Testes by FOXL2 Ablation. *Cell*, *139*(6), 1130–1142. https://doi.org/10.1016/j.cell.2009.11.021

Ullrich, P. A., Xu, Z., Rhoades, A. M., Dettinger, M. D., Mount, J. F., Jones, A. D., & Vahmani, P. (2018). California's Drought of the Future: A Midcentury Recreation of the Exceptional Conditions of 2012–2017. *Earth's Future*, *6*(11), 1568–1587. https://doi.org/10.1029/2018EF001007

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

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

Volff, J. N., & Schartl, M. (2001). Variability of genetic sex determination in poeciliid fishes. *Genetica*, *111*, 101–110.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

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

Xuan, B., Kim, E. M., Song, M.-Y., Shin, Y., Jeon, J.-H., & Bae, E. (2021). Draft Genome of the Korean smelt Hypomesus nipponensis and its transcriptomic. *G3: Genes|Genomes|Genetics*, 35. https://doi.org/10.1093/g3journal/jkab147

Zoonomia Consortium. (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, *587*(7833), 240–245. https://doi.org/10.1038/s41586-020-2876-6

Chapter 1 & 2 Tables and Figures

Table 1.1. Assembly metrics from Osmeridae genome assemblies listed in peer reviewed papers or publicly available on NCBI's GenBank prior to public release of the delta smelt genome on February 04, 2022.

Species	Sex	Release Date (YYYY.MM.DD)	Accession Number	Assembly Level	Coverage	Total sequence length	Number of scaffolds	Scaffold N50	Scaffold L50	Number of contigs	Contig N50	Contig L50	Number of chr. and plasmids	Final number of scaffolds
Hypomesus transpacificus (female)	F	2022.02.04	GCA_021917145.1	chromosome	120x	437,273,953	376	14,850,352	13	1,850	412,669	267	26	376
Hypomesus transpacificus (male)	M	2022.02.04	GCA_021870715.1	chromosome	137x	471,985,164	548	12,200,365	15	2,127	347,532	333	26	548
Thaleichthys pacificus (eulachon)	N/A	2021.03.09	GCA_017311245.1	scaffold	210x	416,131,685	324,311	3,050	34,112	330,739	2,918	35,367	0	324,311

Osmerus eperlanus (European smelt)	N/A	2018.03.18	GCA_900302275.1	scaffold	19x	342,758,722	73,274	6,820	13,139	99,348	4,524	21,105	0	73,274
Hypomesus nipponensis (Japanese smelt/wakasagi)	N/A	2021.05.12	GCA_018346875.1	contig	126x	34,375,595	N/A	460,000	477	20,639	2,124	4,887	0	20,639
Hypomesus nipponensis (Japanese smelt/wakasagi)*	N/A	2021.09.06	N/A	contig	51x	498,930,205	1,987	464,585	300	4,106	316,684	477	N/A	1,987

* metrics taken from publication

Table 1.2. Table of tissue type and storage method of sampled delta smelt from four sampling events. Included are the names referred to in the text. T= sampling trip, F= female, M=male, BM = back muscle, SC = scales, IO = internal organ, FF = flash frozen, and PG = propylene glycol.

Sex	Sample ID	Tissue Type	Storage Solution	NanoDrop 260/280	NanoDrop 260/230	Used for Sequencing
F	T1F01_BM_FF	back muscle	no solution			-
	T1F02_BM_FF	back muscle	no solution	1.8	1.73	10X Genomics linked-reads, Phase Genomics Proxiomo Hi-C
	T1F03_BM_FF	back muscle	no solution			-
M	T1M01_BM_FF	back muscle	no solution			-
	T1M02_BM_FF	back muscle	no solution			-
	T1M03_BM_FF	back muscle	no solution			-
M	T2M01_BM_FF	back muscle	no solution			-
	T2M02_BM_FF	back muscle	no solution			-
	T2M03_BM_FF	back muscle	no solution			-
F	T3F01_BM_PG	back muscle	propylene glycol			-
	T3F01_BM_FF	back muscle	no solution			-
	T3F01_IO_FF	internal organ	no solution			-
	T3F01_SC_FF	scales	no solution			-
	T3F02_BM_PG	back muscle	propylene glycol			-
	T3F02_BM_FF	back muscle	no solution			-
	T3F02_IO_FF	internal organ	no solution			-
	T3F02_SC_FF	scales	no solution	1.84	2.22	Pac Bio HiFi long reads
M	T3M01_BM_PG	back muscle	propylene glycol			-
	T3M01_BM_FF	back muscle	no solution			-
	T3M01_SC_FF	scales	no solution			-
	T3M02_BM_PG	back muscle	propylene glycol			-
	T3M02_BM_FF	back muscle	no solution	1.91	2.02	10X Genomics linked-reads, Pac Bio HiFi long reads
	T3M02_SC_FF	scales	no solution			-

Table 1.3. Raw sequencing data metrics for the delta smelt genome assembly.

Sex	Sequencing Platform	Individual ID	Run ID	Number of Bases	Number of Reads	Mean Read Length	GC%	Average Read Quality*
F	PacBio HiFi	T3F02_SC_FF	m64069_201002_215024	7,617,422,156	1,275,836	5,970	45%	36
F	PacBio HiFi	T3F02_SC_FF	m64069_200830_055940	6,404,937,329	624,944	10,248	45%	33
F	PacBio HiFi	T3F02_SC_FF	m64069_200603_183739	13,962,511,851	840,724	16,607	45%	30
M	PacBio HiFi	T3M02_BM_FF	m64069_200220_045555	23,993,220,246	2,054,534	11,678	44%	35
M	PacBio HiFi	T3M02_BM_FF	m64069_200211_020731	11,151,984,598	1,040,599	10,716	44%	33
F	10X Illumina	T1F02_BM_FF	10X_R1_F	94,825,601,818	627,984,118	151	51%	33
F	10X Illumina	T1F02_BM_FF	10X_R2_F	94,825,601,818	627,984,118	151	51%	32
M	10X Illumina	T3M02_BM_FF	10X_R1_M	65,806,680,934	435,805,834	151	49%	34
M	10X Illumina	T3M02_BM_FF	10X_R2_M	65,806,680,934	435,805,834	151	49%	32
F	Phase Genomics Hi-C	T1F02_BM_FF	hic_R1_F	13,116,671,550	87,444,477	150	46%	38
F	Phase Genomics Hi-C	T1F02_BM_FF	hic_R2_F	13,116,671,550	87,444,477	150	46%	36

*Illumina based sequencing (linked read and hi-c) calculated from the output of fastqc (Supplemental Table 1).

Table 1.4. Table of assembly steps with corresponding metrics. A_0 = Metrics for unassembled, filtered PacBio HiFi reads; A_1 = draft resulting from initial long-read assembly step; A_2 = draft resulting from scaffolding A_1 assembly using linked-reads; A_3 = draft resulting from scaffolding A2 assembly using Hi-C data; A_4 = final assembly metrics resulting from anchoring chromosomes with a linkage map. Continuity metrics created from genometools, BUSCO scores from comparison to August 05, 2020 Actinopterygii lineage gene (n=3640) dataset.

Metrics		Male					Female				
		A_0	A_1	A_2	A_3	A_4	A_0	A_1	A_2	A_3	A_4
Continuity Metrics	N50 (bp)	11,604	353,581	1,188,596	2,749,144	12,200,365	15,048	418,614	1,392,224	4,383,157	14,850,352
	L50	1,276,120	324	106	38	15	771,808	264	80	26	13
	# contigs (bp)	3,095,133	2,086	1,106	705	549	2,741,504	1,805	1,012	515	376
	total length	35,145,204,844	471,831,811	471,929,811	472,145,811	472,157,411	27,984,871,336	436,920,153	436,999,453	437,264,453	437,273,953
BUSCO Scores	complete	90.4%	88.0%	88.5%	88.2%	88.4%	92.6%	89.0%	85.9%	89.5%	89.3%
	single	3.4%	79.5%	80.5%	80.5%	81.2%	3.3%	87.4%	84.4%	88.0%	87.7%
	double	87.0%	8.5%	8.0%	7.7%	7.2%	89.3%	1.6%	1.5%	1.5%	1.6%
	fragmented	4.0%	1.5%	1.1%	1.1%	1.0%	3.0%	1.1%	3.1%	0.8%	0.8%

Table 2.1. 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 2.2. 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 (red= homozygous for major allele; purple=heterozygous; blue=homozygous for minor allele).

position	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

Table 2.3. Frequency of genotypes in female and male sequencing data from GWAS using the male reference assembly.

Locus	Genotype	Frequency	
		Female	Male
Chr5:1885249	GG	62.50%	29.17%
	GA	20.83%	16.67%
	AA	16.67%	54.16%
Chr5:1885251	GG	62.50%	33.33%
	GT	20.83%	12.50%
	TT	16.67%	54.17%

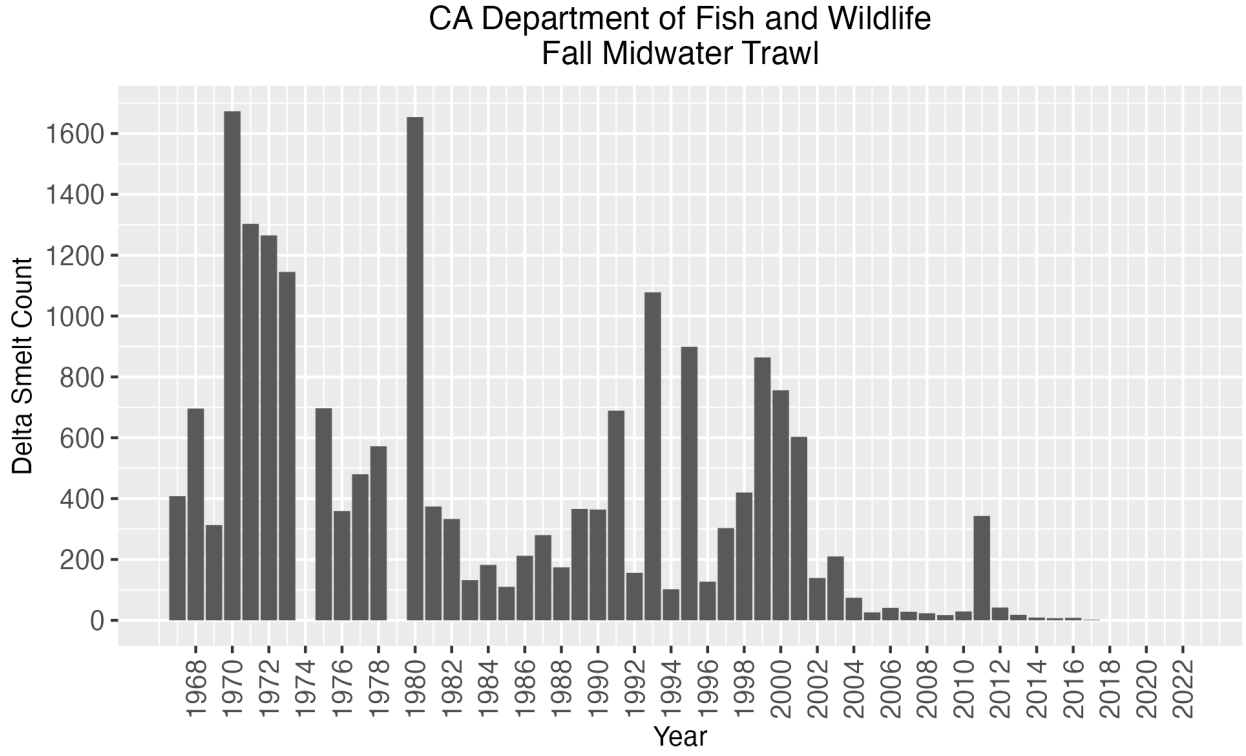


Figure 1.1. CDFW annual Fall Midwater Trawl delta smelt catch numbers (indices) from 1967 to 2022. CDFW did not sample in 1974 and 1979. Indices taken from CDFW publicly hosted dataset (https://www.dfg.ca.gov/delta/data/fmwt/indices.asp).

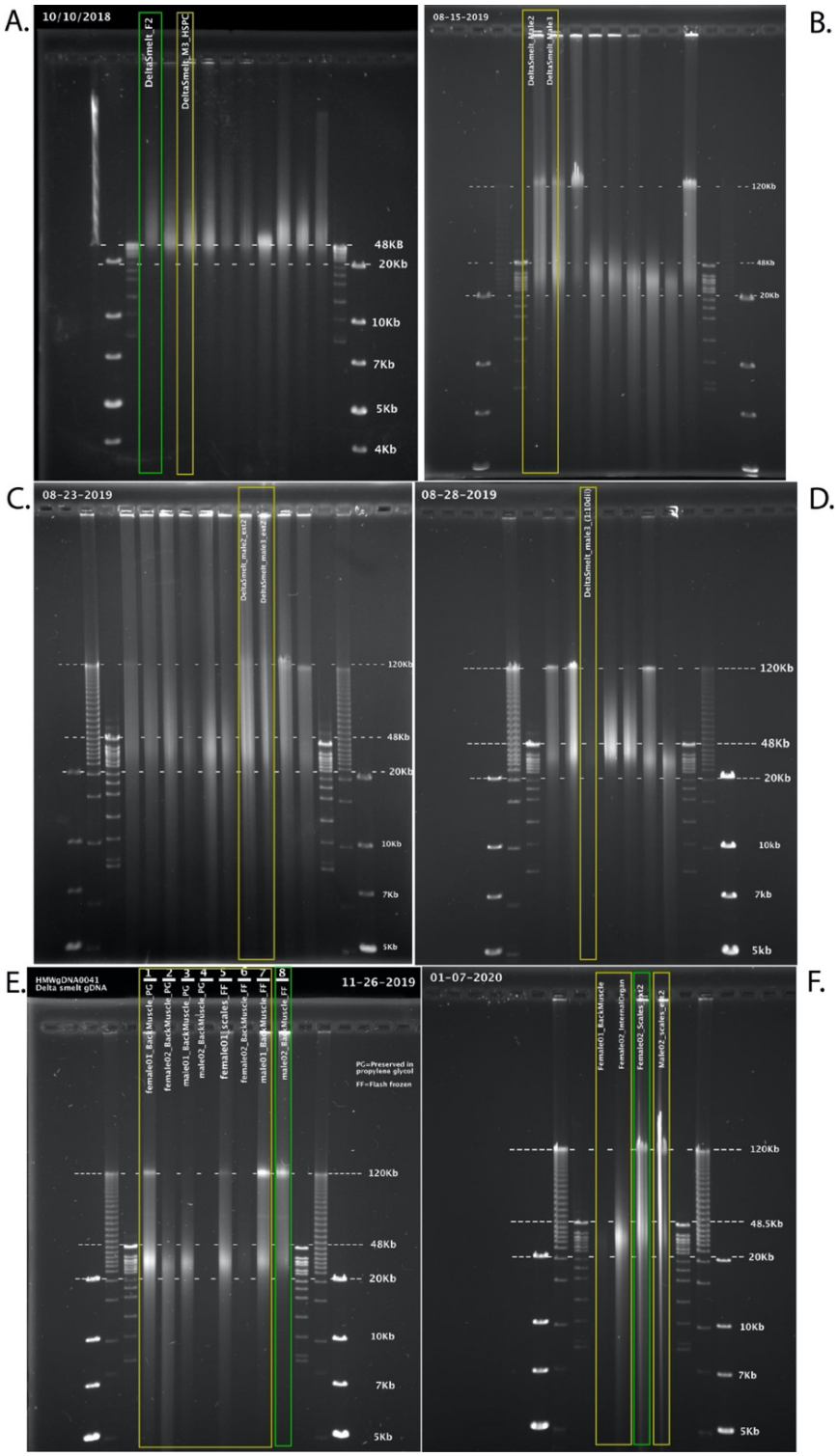


Figure 1.2. Pulse field gel images of extracted HMW gDNA from six rounds (A-F) of extractions. Green boxes surround lanes from extracted samples usable for long-read and linked-read

sequencing (extraction distribution centered ~50 kb) by the UC Davis DNA Technologies and Expression Analysis Core, yellow boxes surround lanes from samples with insufficient extract lengths or concentration. A) Extraction #1: usable HMW gDNA from female back muscle tissue (T1F02_BM_FF), B) Extraction #2: no usable DNA, C) Extraction #3: no usable DNA, D) Extraction #4: No usable DNA, E) Extraction #5: usable HMW gDNA from male back muscle tissue (T3M02_BM_FF); F) Extraction #6: usable HMW gDNA from female gill tissue (T3F02_SC_FF).

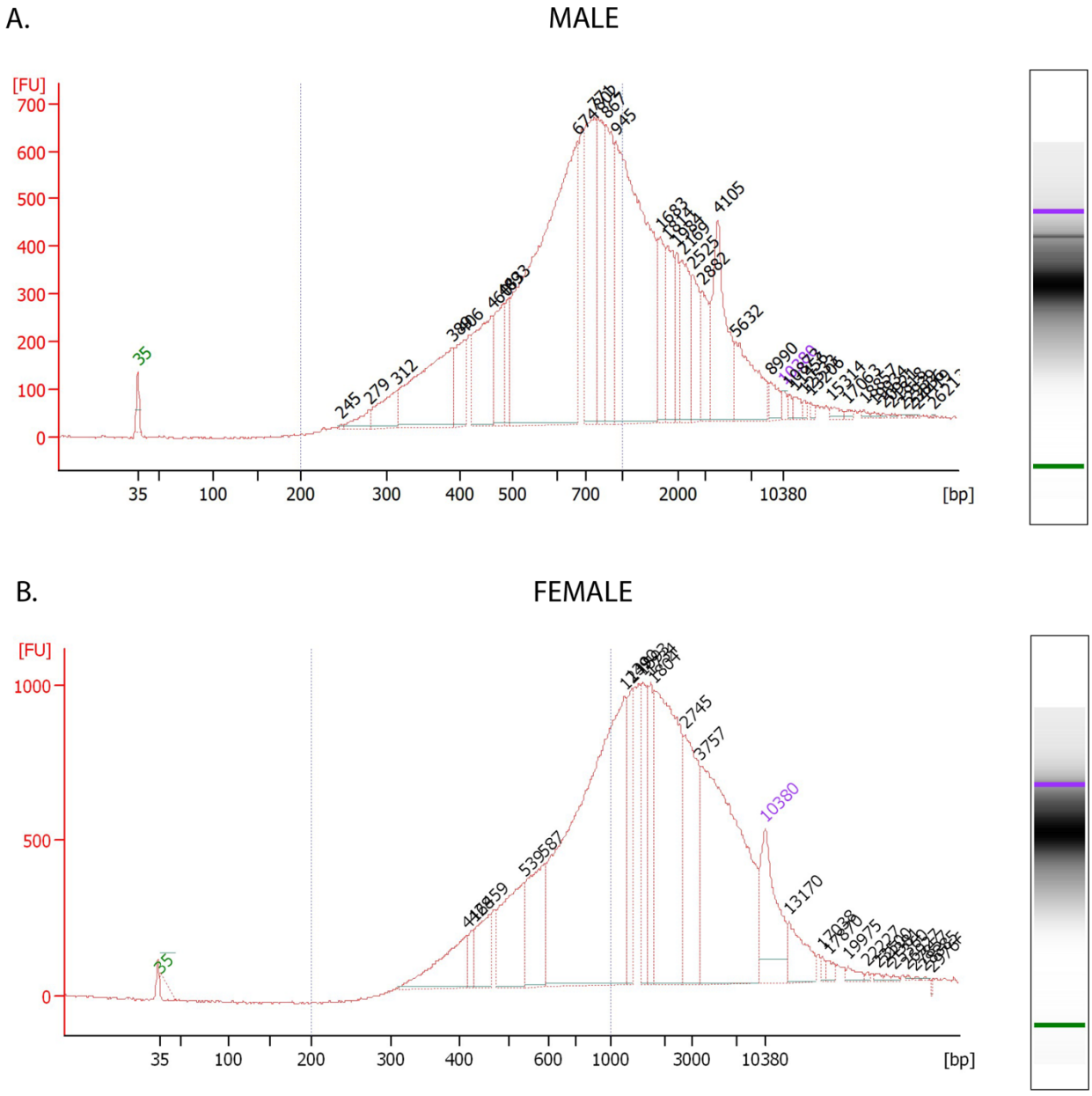


Figure 1.3. Bioanalyzer 2100 electropherogram read out of sheared female and male post-GEM DNA used for linked-read library prep. Lower and upper Agilent High Sensitivity DNA Kit (Agilent, cat. 5067-4626) ladder DNA Markers denoted in green (35 bp) and purple (10,380 bp),

respectively. Y-axis denotes fluorescence intensity (FU) and x-axis denotes fragment size in base pairs (bp).

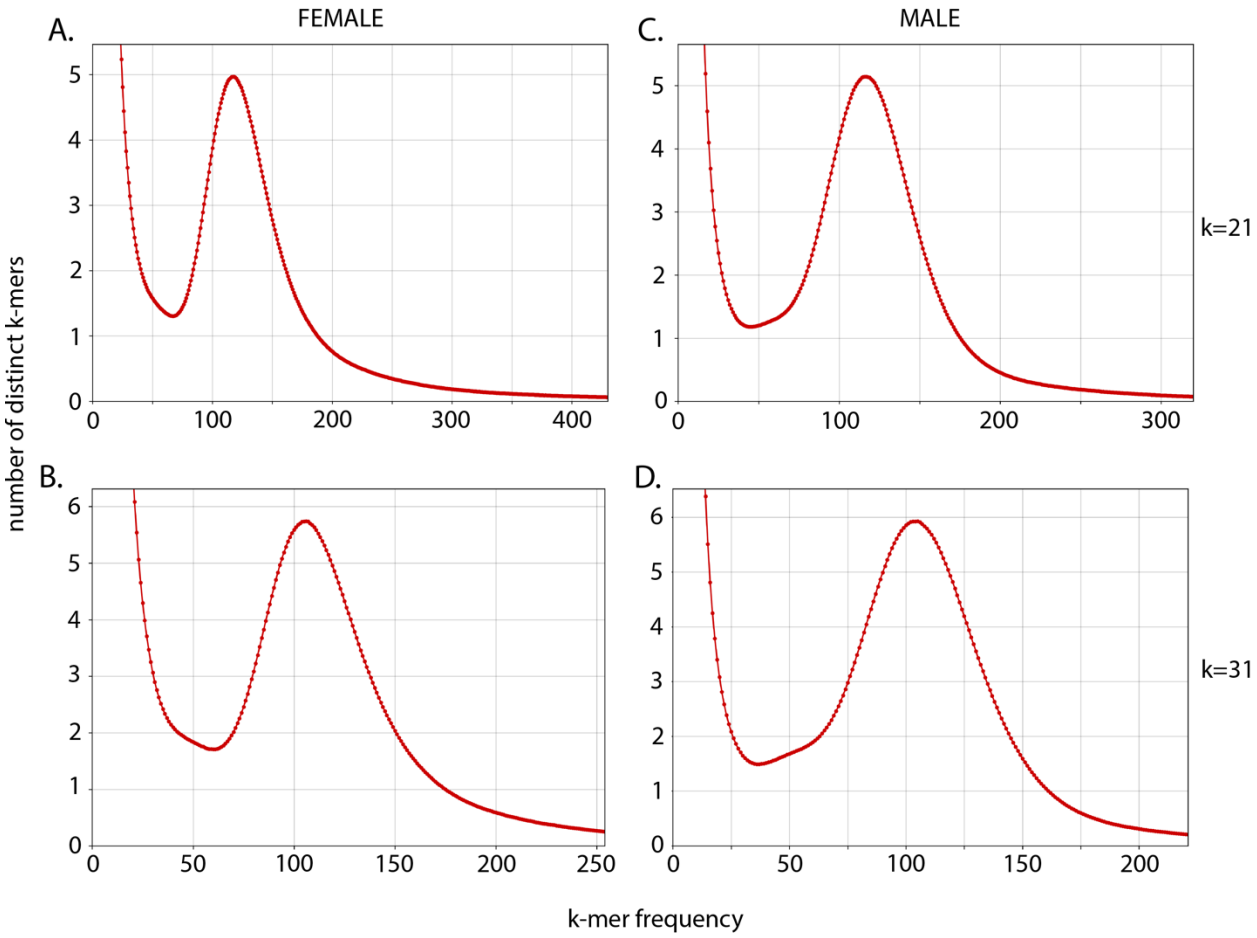


Figure 1.4. Linked-read k-mer spectra histograms created from `kat hist` function in KAT (Mapleson et al., 2017b). Each plot shows the number of distinct k-mers at different frequencies from female (A & B) and male (C & D) sequencing data. Histograms using k=21 (A & C), and k=31 (B & D). The high abundance of low frequency k-mers are expected as a product of sequencing and base calling errors.

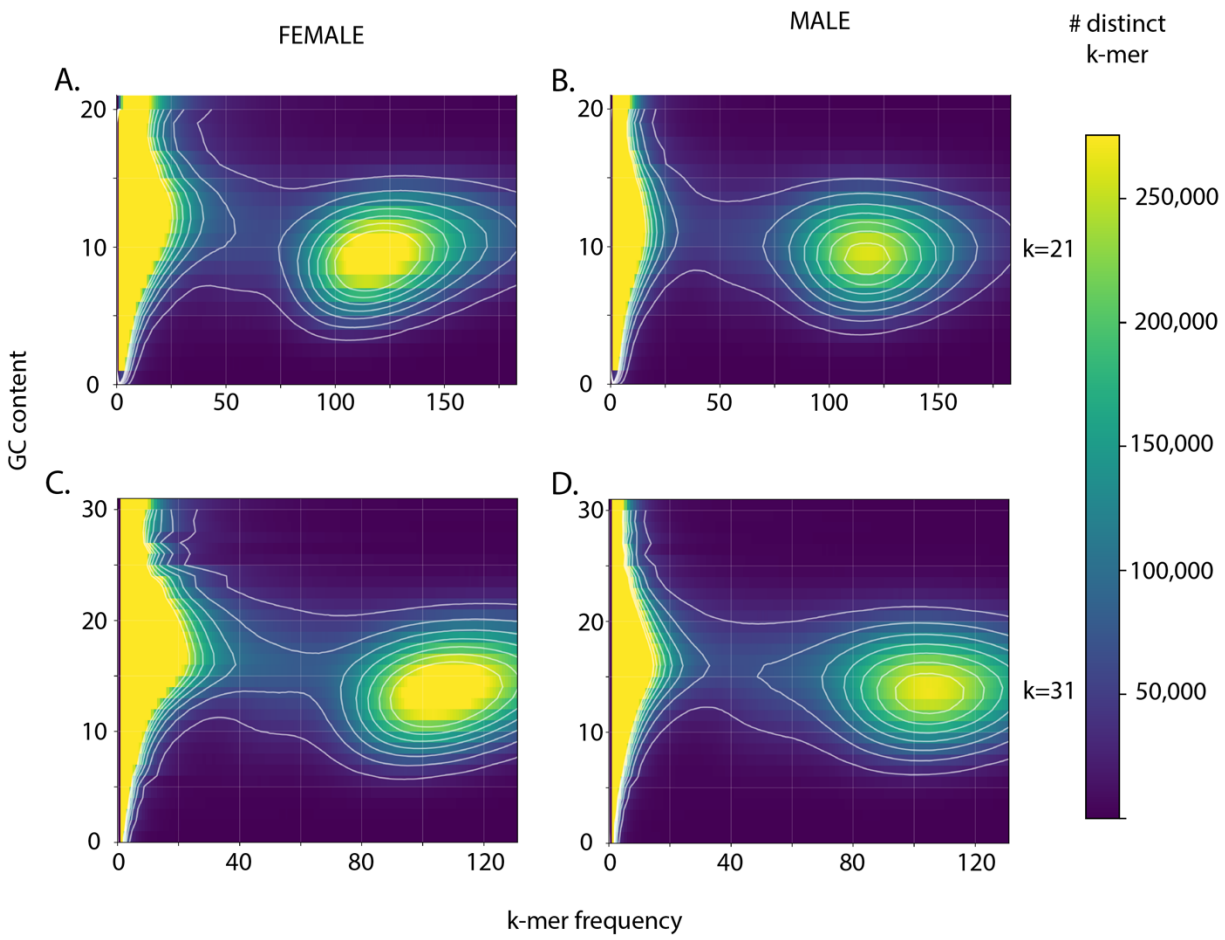


Figure 1.5. Heatmaps of k-mer frequency (x-axis) vs GC count (y-axis) colored by the number of distinct k-mers created by with the `kat gcp` function in KAT (Mapleson et al., 2017b). Blue indicates fewer distinct k-mers with a given GC count and frequency, while yellow indicates more distinct k-mers. Plots using k=21 (A & C), and k=31 (B & D). No indication of contamination was detected in female (A & B) and male (C & D) sequencing data. Low frequency k-mers with a broad distribution of GC content, not observed here, would be expected from sequencing and base calling errors.

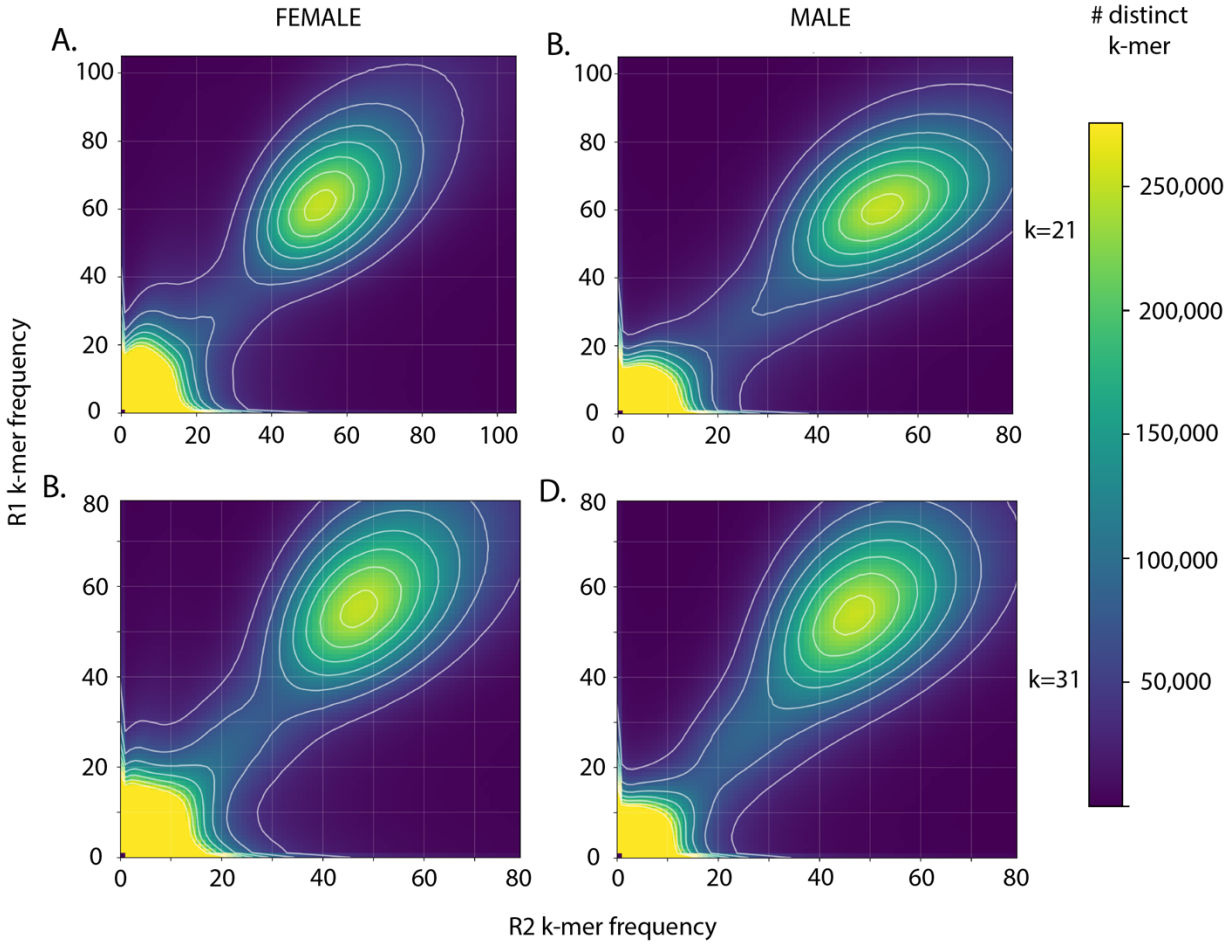


Figure 1.6. Plot comparing the number of distinct k-mers at different frequencies in linked-read sequence data using the `kat comp` function in KAT (Mapleson et al., 2017b) with female (A & B) or male (C & D) samples. Plots using k=21 (A & C), and k=31 (B & D). For all plots the R1 (x-axis) and R2 (y-axis) capture slightly different information, but no major sources of sequencing bias are observed. Sequencing bias in either of the two files would result in an irregular pattern in the number of distinct k-mers.

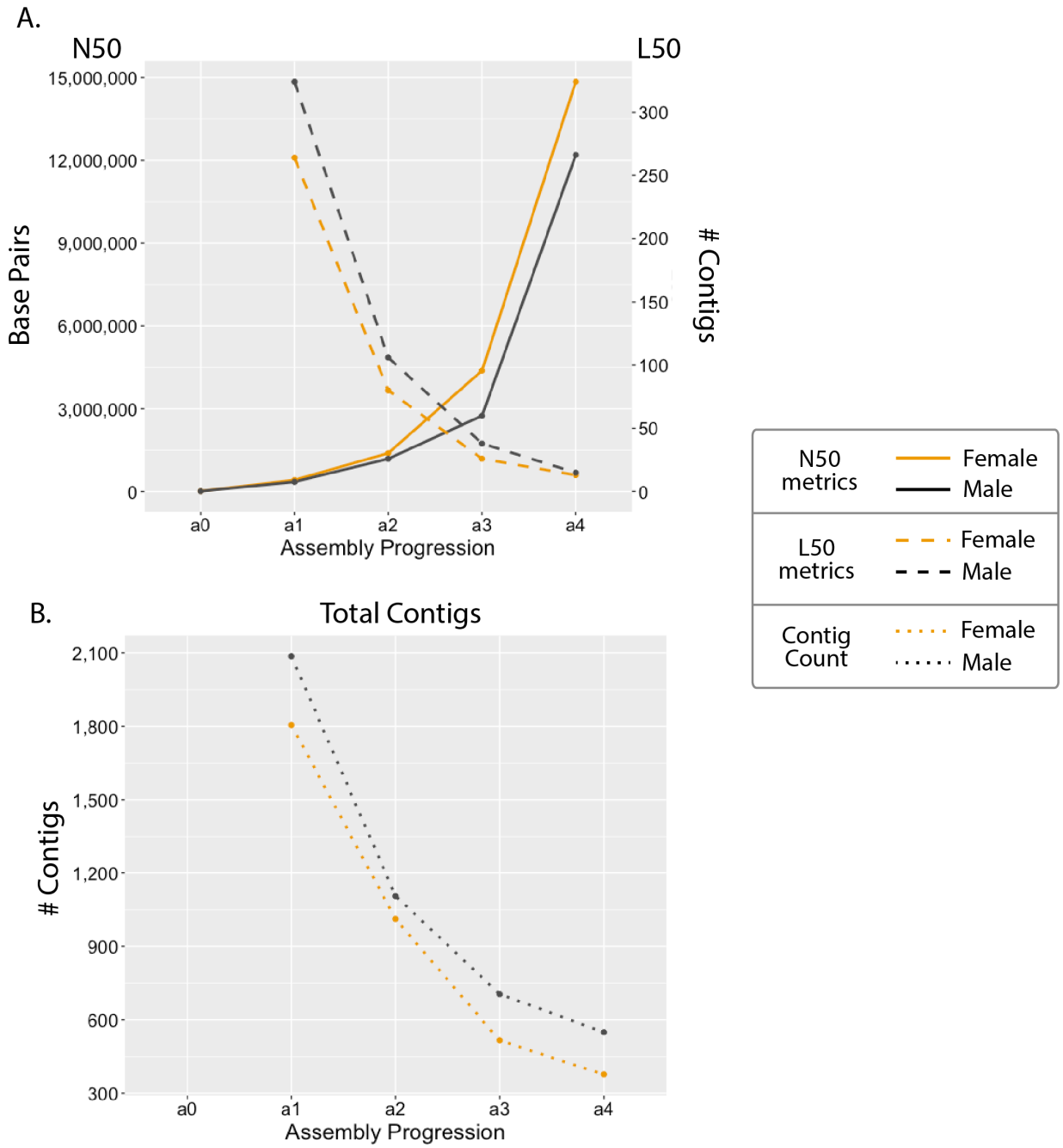


Figure 1.7. Metrics from each assembly step. A) N50 and L50 metrics and B) The total number of contigs resulting from each assembly step. Metrics show an increase in contiguity after each step of the assembly pipeline. Data points containing the total number of "contigs" from the A_0 version of the assembly, which only contained unassembled raw reads (female=2,741,504 and

male=3,095,133; Table 1.4), have been omitted to better visualize differences in the number contigs between steps incorporating difference sequencing types.

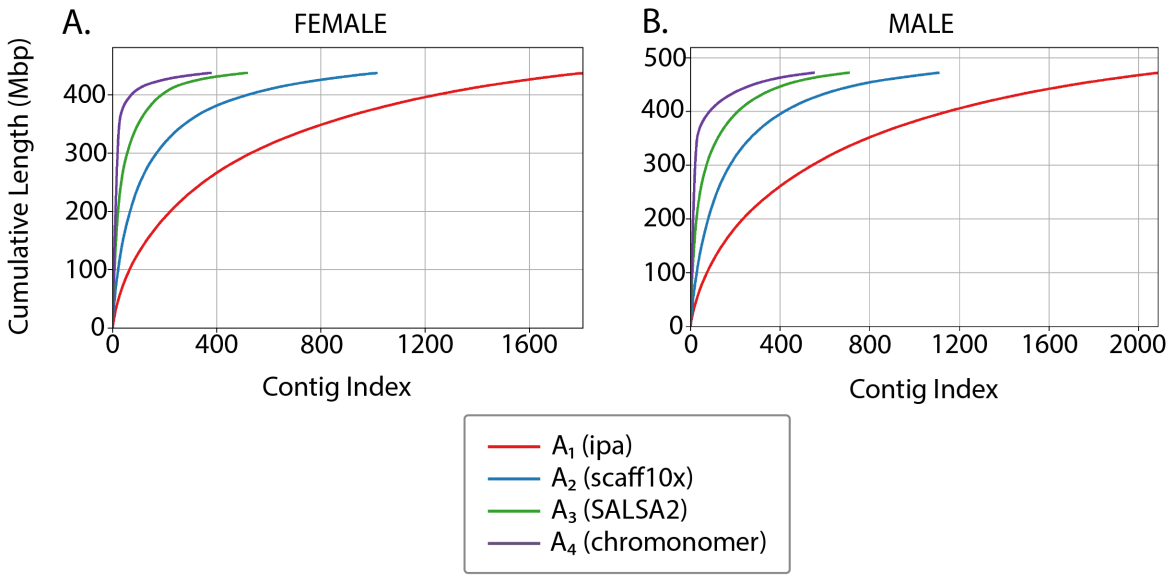
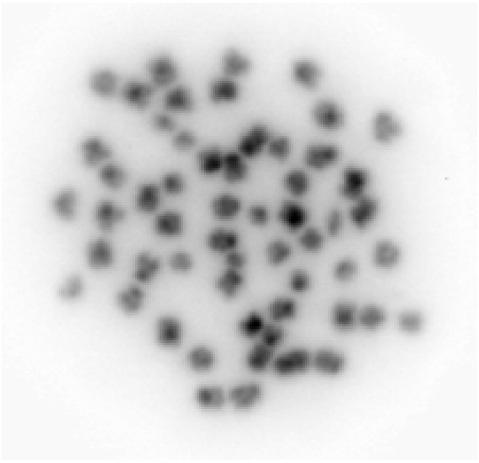
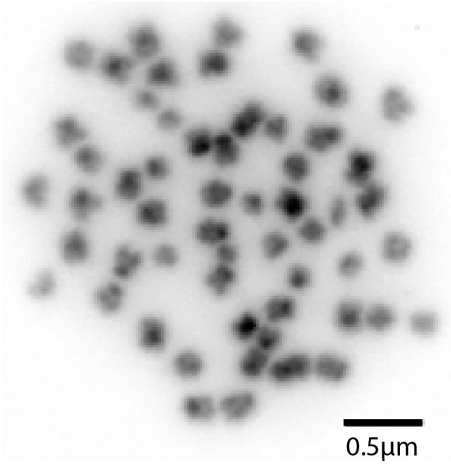


Figure 1.8. Cumulative read length plots of each iteration (A_1 – A_4) of the female and male genome assemblies.

A.



B.



C.

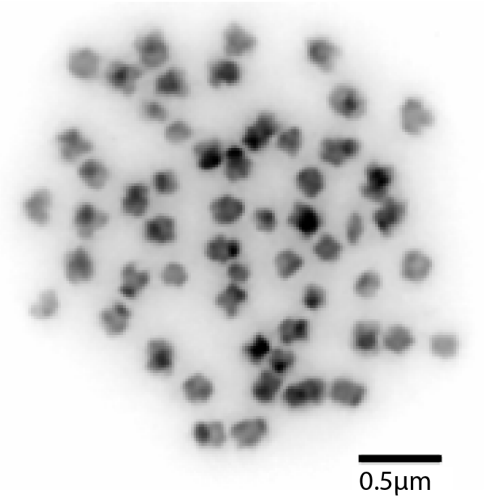


Figure 1.9. Karyotype of metaphase stage mitotic cell from a male delta smelt showing 2n = 56 chromosomes; A) unmodified image, no scale bar; B) Adobe Photoshop modified image, plus scale bar; C) Adobe Photoshop focused image, plus scale bar.

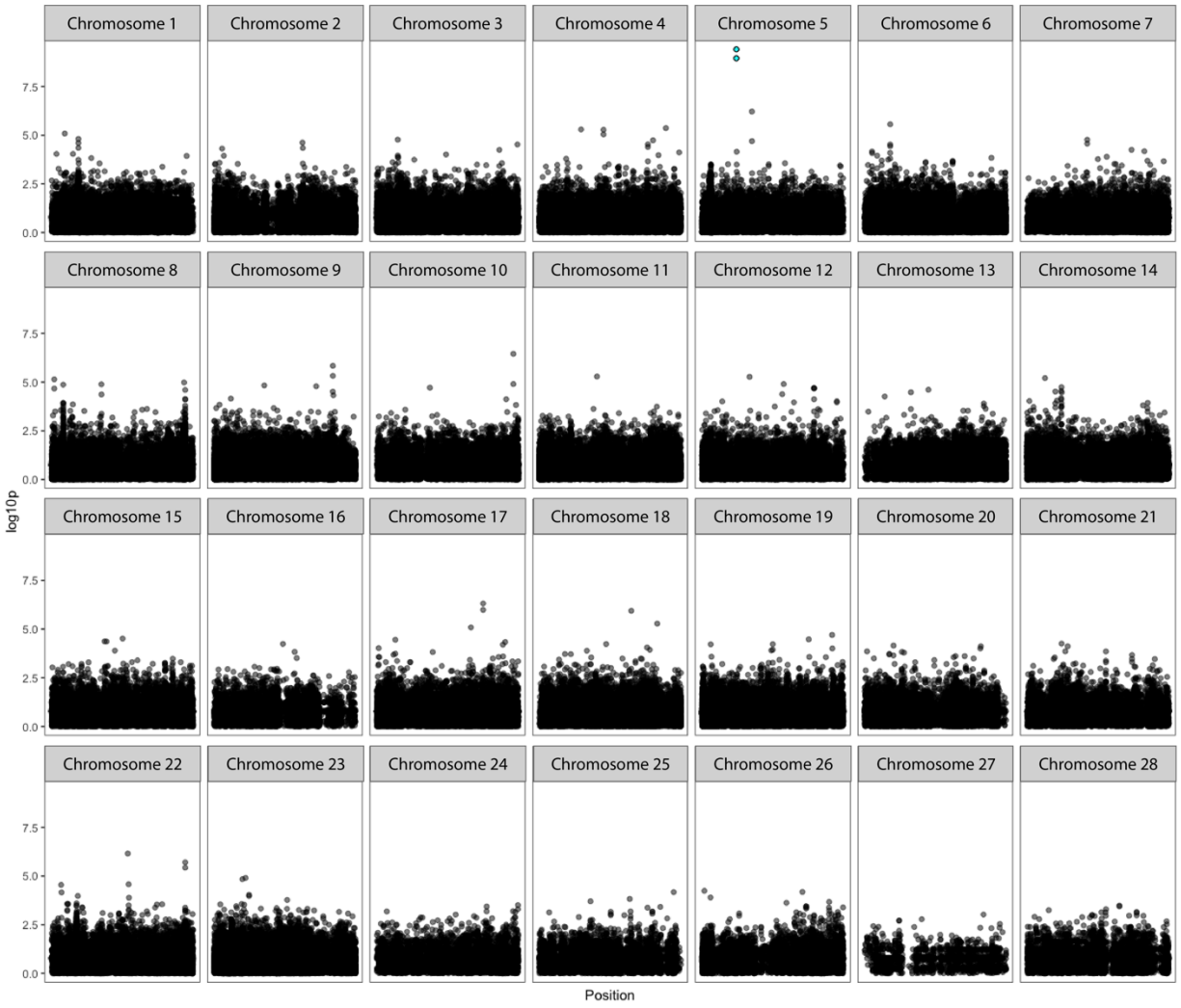


Figure 2.1. Manhattan plots of each of 28 chromosomes from the final male delta smelt reference genome (Joslin in prep.). Location on the x axis and $\log_{10} P$ significance on the y axis. Significant SNPs on Chromosome 5 are marked in blue.

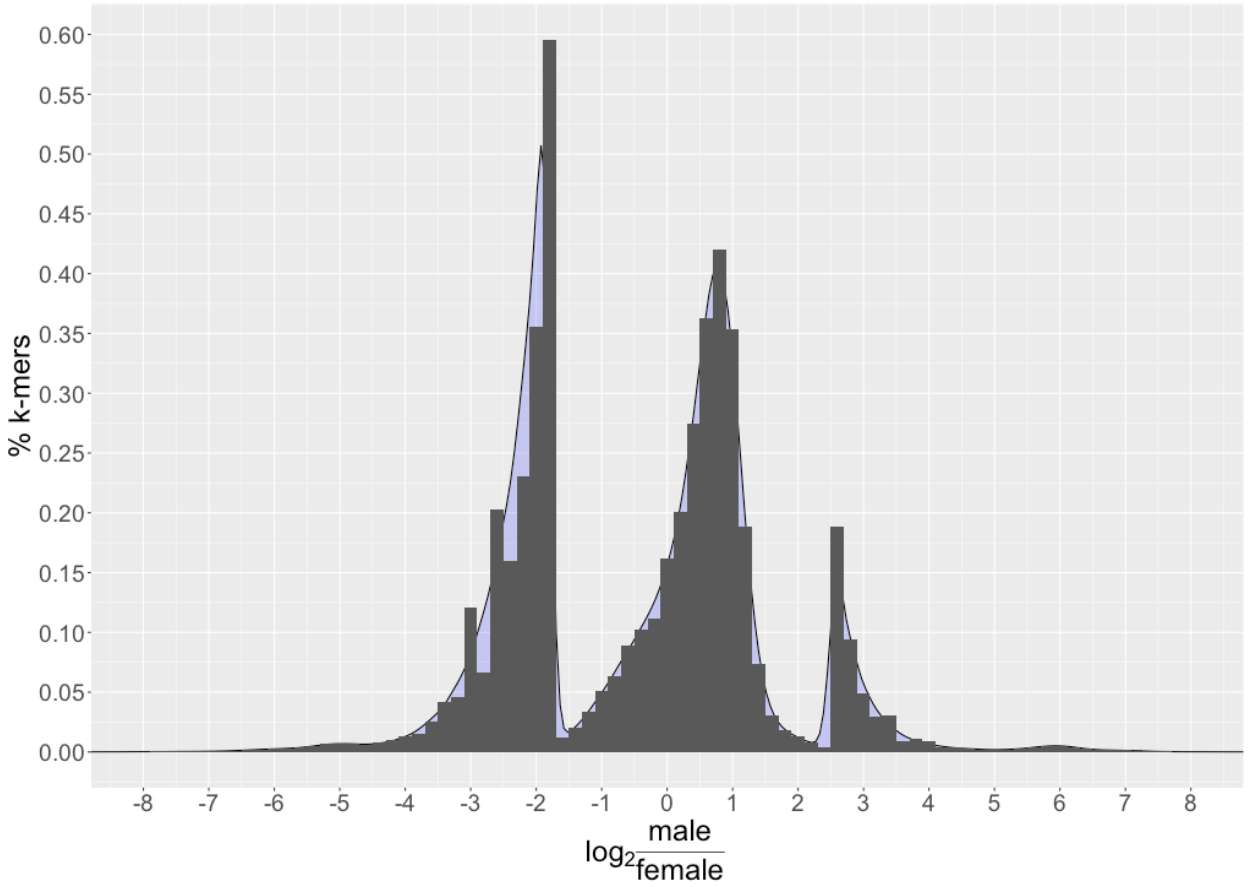


Figure 2.2. Distribution of the proportion of change from $\log_2 \frac{(m+1)}{(f_{cor}+1)}$ where m = male k-mer abundance and f_{cor} = corrected female abundance ($f_{cor} = f * A_{cor}$, where f = female k-mer abundance and A_{cor} = male to female abundance correction of 0.538 resulting from the ratio of male to female median abundances $A_{cor} = \frac{Med(m)}{Med(f)}$).

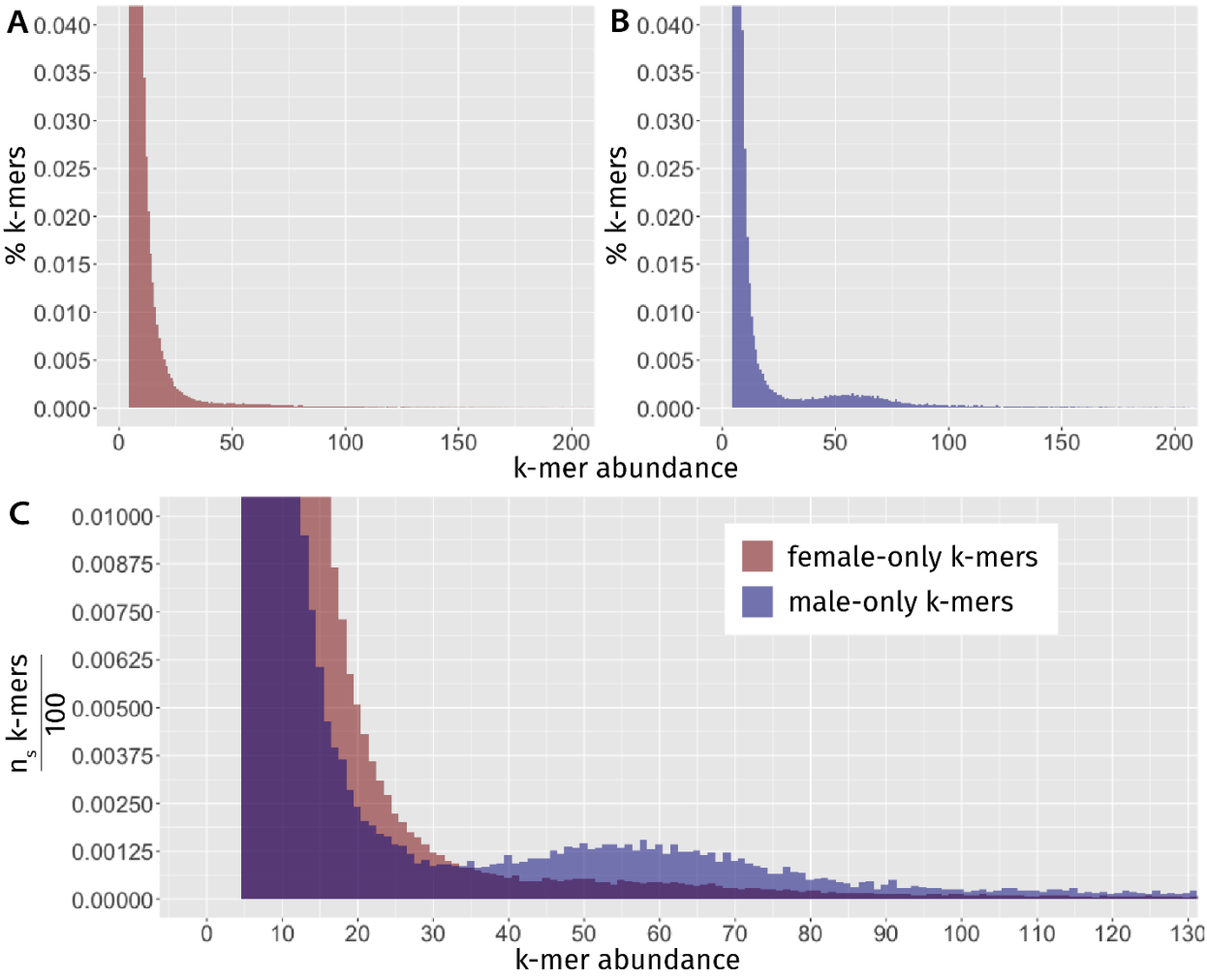


Figure 2.3. Distributions of k-mer abundances in female and male linked-read sequencing data. A.) Corrected female-only k-mer abundances (A_{cor} = 0.538). B.) Male-only k-mer abundances. C.) Overlay of the of corrected female-only and male-only k-mer abundances and corresponding percent (n) of k-mers for each sex where s = sex.

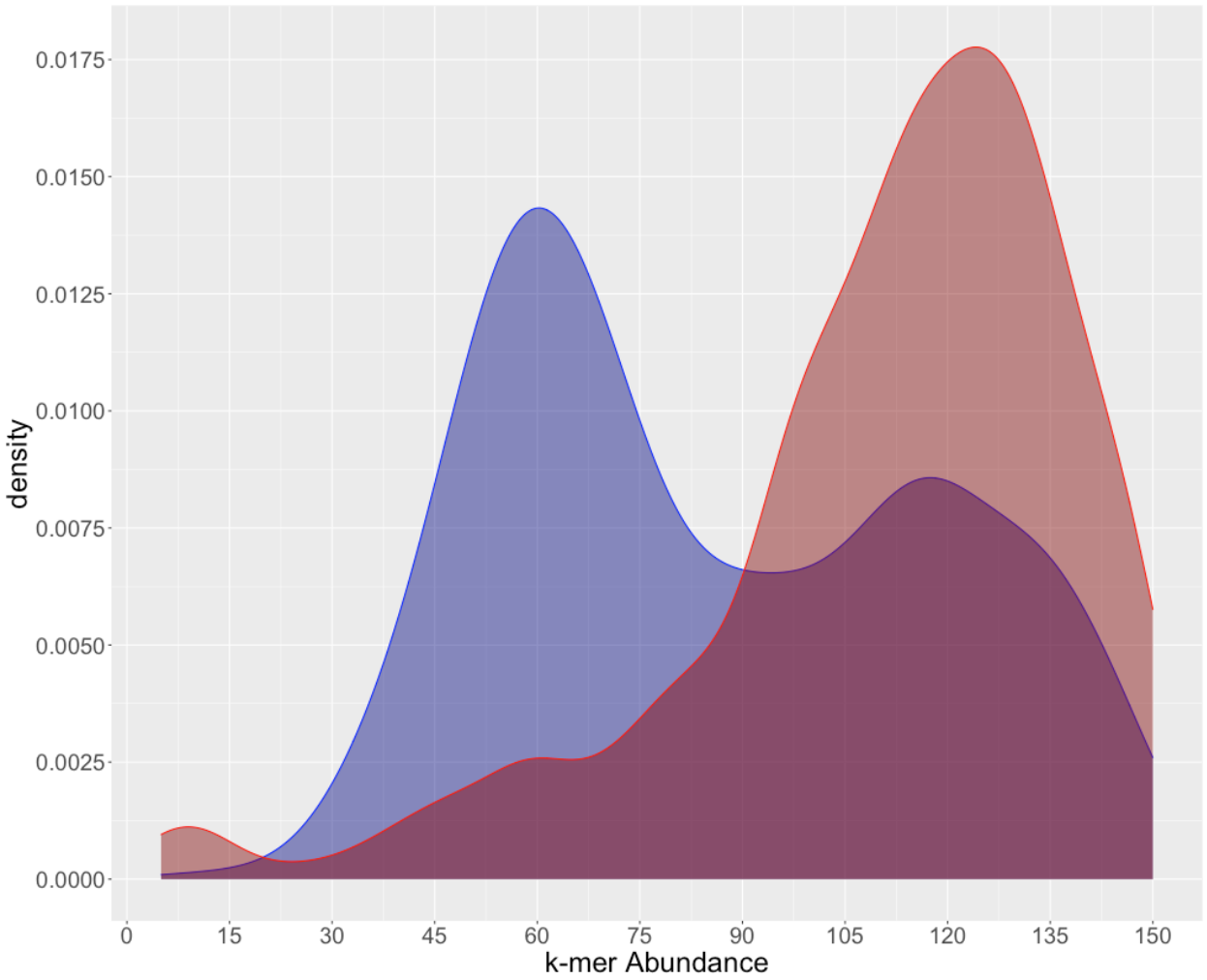


Figure 2.4. Distribution of female-only (red) and male-only (blue) k-mer abundances for k-mers located on contigs containing five hashes (approximately 5,000 bp in length) from the male A_1 assembly.

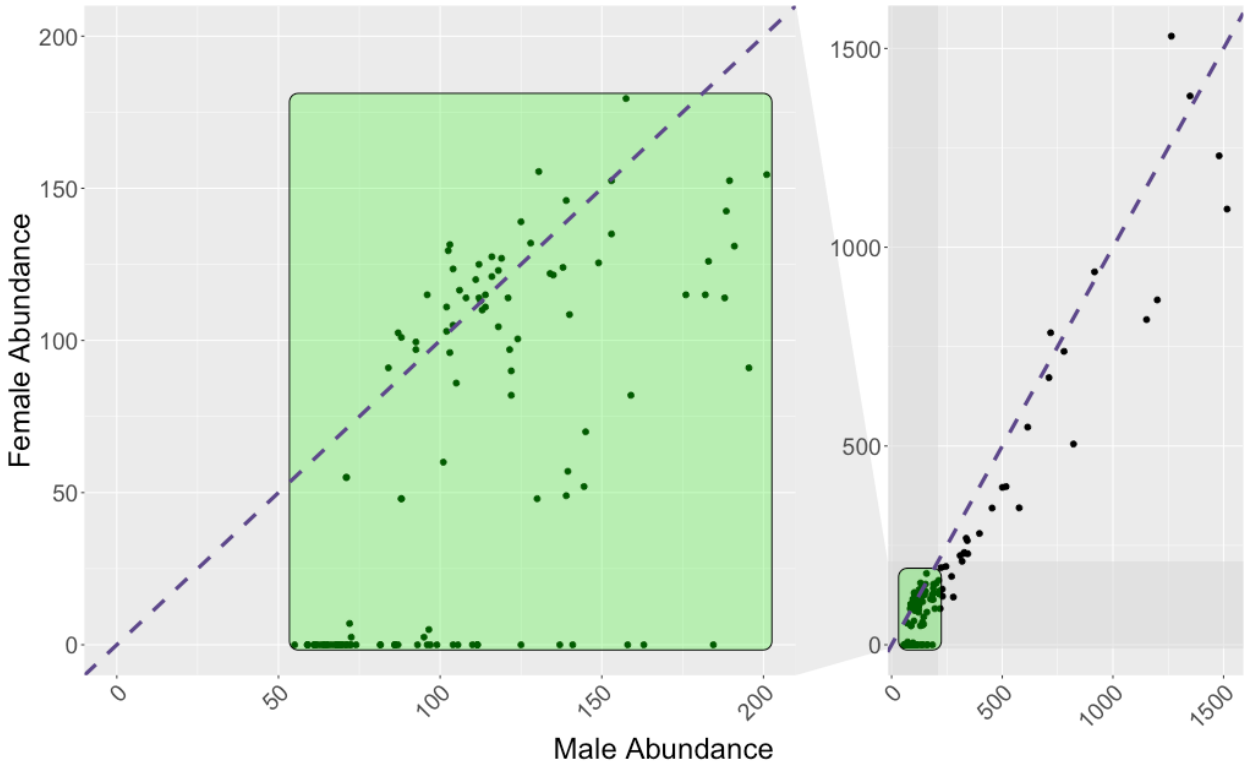


Figure 2.5. Female (y-axis) and male (x-axis) median k-mer abundance on contigs containing five or more hashes, corresponding to roughly 5,000 bp k-mers. Dashed lines show the slope of a one-to-one ratio between female and male abundances. The right plot shows all analyzed contigs, while the left plot is zoomed in to better visualize the 44 contigs with k-mers present in male sequencing data but absent in female sequencing data. Contigs present in males with zero abundance in females indicate the male sequencing data contains sex-specific sequences in high abundance that are not contained in the female sequencing data.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 29 October 2020
doi: 10.3389/fgene.2020.558762

Association of the Lactase Persistence Haplotype Block With Disease Risk in Populations of European Descent

*Shannon E. K. Joslin[1], Blythe P. Durbin-Johnson[2,3], Monica Britton[2], Matthew L. Settles[2], Ian Korf[2,4] and Danielle G. Lemay[2,5,6]**

[1] Department of Animal Science, UC Davis, Davis, CA, United States, [2] UC Davis Genome Center, Davis, CA, United States, [3] Department of Public Health Sciences, UC Davis School of Medicine, Davis, CA, United States, [4] Department of Molecular and Cellular Biology, UC Davis, Davis, CA, United States, [5] USDA ARS Western Human Nutrition Research Center, Davis, CA, United States, [6] Department of Nutrition, UC Davis, Davis, CA, United States

OPEN ACCESS

Edited by:
L. Joseph Su,
University of Arkansas for Medical Sciences, United States

Reviewed by:
Kui Zhang,
Michigan Technological University, United States
Raja Amir Hassan Kuchay,
Baba Ghulam Shah Badshah University, India
Luana Caroline Oliveira,
Federal University of Paraná, Brazil

***Correspondence:**
Danielle G. Lemay
Danielle.Lemay@usda.gov

Specialty section:
This article was submitted to Human Genomics, a section of the journal Frontiers in Genetics

Received: 04 May 2020
Accepted: 08 October 2020
Published: 29 October 2020

Citation:
Joslin SEK, Durbin-Johnson BP, Britton M, Settles ML, Korf I and Lemay DG (2020) Association of the Lactase Persistence Haplotype Block With Disease Risk in Populations of European Descent. Front. Genet. 11:558762. doi: 10.3389/fgene.2020.558762

Among people of European descent, the ability to digest lactose into adulthood arose via strong positive selection of a highly advantageous allele encompassing the lactase gene. Lactose-tolerant and intolerant individuals may have different disease risks due to the shared genetics of their haplotype block. Therefore, the overall objective of the study was to assess the genetic association of the lactase persistence haplotype to disease risk. Using data from the 1000Genomes project, we estimated the size of the lactase persistence haplotype block to be 1.9 Mbp containing up to 9 protein-coding genes and a microRNA. Based on the function of the genes and microRNA, we studied health phenotypes likely to be impacted by the lactase persistence allele: prostate cancer status, cardiovascular disease status, and bone mineral density. We used summary statistics from large genome-wide metanalyses—32,965 bone mineral density, 140,306 prostate cancer and 184,305 coronary artery disease subjects—to evaluate whether the lactase persistence allele was associated with these disease phenotypes. Despite the fact that previous work demonstrated that the lactase persistence haplotype block harbors increased deleterious mutations, these results suggest little effect on the studied disease phenotypes.

Keywords: human evolution, population genetics, diet, physiological traits, phenotype, selective sweep, lactose, lactose tolerance

INTRODUCTION

Lactose is the main carbohydrate found in milk. The enzyme lactase, encoded by the *LCT* gene, allows for the breakdown of lactose in infant mammals. Various human populations continue to express *LCT* post weaning and can digest lactose into adulthood, a trait known as lactase persistence (LP) (Swallow, 2003). The rs4988235 (−13910 C > T) transition variant, or LP allele, in the promoter of the *LCT* gene allows for LP in populations of European descent. The allele frequency of this advantageous mutation rapidly rose in groups with milk and dairy production and consumption relatively recently, as seen by

the signature of a relatively large haplotype block surrounding the LP allele (Bersaglieri et al., 2004; Itan et al., 2009; Ségurel and Bon, 2017). Therefore, lactose tolerant and intolerant individuals' genetic backgrounds differ in the alleles surrounding the LP allele. Positive selection for the trait of LP can hold slightly deleterious alleles that are in linkage disequilibrium (LD) with the LP allele at a higher frequency than expected under balancing selection alone (Smith and Haigh, 1974; Fay and Wu, 2000; Chun and Fay, 2011; Cutter and Payseur, 2013). Prior work by Chun and Fay (2011) found European samples harbored multiple deleterious or neutral non-synonymous SNPs within the *LCT* gene and two other genes in the surrounding the region. However, it is unclear whether mutations found within the LP haplotype block give rise to unfavorable phenotypes. Determining the differential risk of disease based on individual genetic backgrounds with the indirect phenotype of lactase persistence may help resolve contrasting epidemiological findings and improve public health. Therefore, the objective this study was to determine the size of the LP haplotype block and its impact on disease risk in humans with and without the LP allele.

MATERIALS AND METHODS

Data Sets

Genotype

The 1000 Genomes Phase 3 datasets were accessed on March 15, 2016 from the organization's data portal[1]. Unrelated individuals from the 1000 Genomes ethnic classifications of Northern and Western European Ancestry (CEU), Finish (FIN), English or Scottish (GBR), Iberian (IBS) or Italian (TSI) were used to define the lactase persistence haplotype block.

Osteoporosis

Data used in our osteoporosis analyses were downloaded from the 2015 data release of the GEFOS single variant bone mineral density (BMD) meta-analysis dataset comprised of individuals of European ancestry (Zheng et al., 2015). We accessed the genome-wide meta-analysis summary statistics from the genome-wide association meta-analysis (GWAMA) of femoral neck BMD, forearm BMD and lumbar spine BMD using 32,965 individuals in the publicly accessible GEFOS database[2] in July 2018. To date, the GEFOS study is the largest publicly accessible femoral neck, lumbar spine and forearm bone mineral density genome-wide meta-analysis dataset.

Prostate Cancer

In order to carry out our prostate cancer analyses we accessed summary association statistics data from the publicly accessible Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium Oncoarray database[3] in July 2018. Genome-wide meta-analysis statistics of prostate cancer from 79,194 prostate cancer cases and 61,112 controls of European ancestry were made publicly available in 2018 from the PRACTICAL dataset and were subsequently used in our analyses (Schumacher et al., 2018).

Coronary Artery Disease

Meta-analysis summary data for individuals with coronary artery disease were accessed from the CARDIoGRAMplusC4D 1000 Genomes-based GWAS dataset[4] in October 2017. The dataset is a meta-analysis of GWAS studies of mainly European, South Asian, and East Asian descent imputed using the 1000 Genomes phase 1 v3 training set with 38 million variants and was made publicly available in 2015. The study interrogated 9.4 million variants and involved 60,801 CAD cases and 123,504 controls (Nikpay et al., 2015). To date, this dataset is the most comprehensive GWAS of coronary artery disease in populations of European descent.

Analysis

Characterizing Lactase Persistence LD

We determined the pattern of LD with the lactase persistence T allele at rs4988235 in populations of European descent from the 1000 Genomes Phase 3 datasets. Unrelated subjects of CEU, FIN, GBR, IBS or TSI ethnicity were considered for the purpose of our analysis. Downloaded data were converted from vcf to PLINK formatted ped/map files using the fcgene software (Roshyara and Scholz, 2014). Genome-wide pairwise associations with marker rs4988235 (2:136608646_G_A) were calculated in PLINK (Purcell et al., 2007). Pairwise calculations were conducted using 1,000,000 markers and 1,000,000,000 base pair windows. Thus, the upper limit of distance or pairwise comparisons for the marker of interest shall not exceed more than 1 M markers nor 100 Mb. LD was measured using r^2 and calculated with the equation: $r^2\left(p_a, p_b, p_{ab}\right) = \frac{(p_{ab}-p_a p_b)^2}{p_a(1-p_a)p_b(1-p_b)}$, where p_a, p_b and p_{ab} denote the frequencies of the a and b alleles and the ab haplotype, respectively. A SNP with $r^2 > 0.2$ was considered in LD with rs4988235 for the purpose of this study. The upper and lower boundaries of European LP haplotype block were determined by selecting the most distant SNPs with an $r^2 > 0.2$ with rs4988235.

Omnibus Effect

To test if SNPs in high LD with the LP SNP are also associated with a disease of interest, class-level genetic association tests (GenCAT) were performed for individual phenotypes within each disease using the GenCAT package in R (Qian et al., 2016). These tests use a user defined "class" of SNPs, in this case SNPs in high LD ($r^2 > 0.2$) with the LP SNP, to see if either the number of SNPs in the class or correlations of the SNPs in the class are statistically meaningful when compared to the SNPs associated with a disease of interest. Log odds ratios for all effect alleles were aligned to be in phase with the lactase persistence haplotype. SNPs in high LD ($r^2 > 0.2$) were considered a class and tested to determine if class had an effect on all phenotypes for each disease. A total of five phenotypes corresponding to three different diseases were evaluated for a class effect of LP haplotype block on the phenotype of interest (**Table 1**). We investigated

[1] ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

[2] http://www.gefos.org/sites/default/files/README.txt

[3] http://practical.icr.ac.uk/blog/?page_id=8164

[4] www.CARDIOGRAMPLUSC4D.ORG

TABLE 1 | Disease phenotypes, measurements and corresponding test statistics.

Phenotype	Measure	Fisher's exact test *P*-value	Wilcoxon *P*-value, test on two-sided *P*-values	Wilcoxon *P*-value, test on upper *P*-values	Wilcoxon *P*-value, test on lower *P*-values	Mean *B*eta for SNPs in *P*hase
Prostate cancer	Affected	1	0.001393	1	1.611e-7	1.021982
Coronary artery disease	Affected	1	2.2e-16	2.545e-11	1	0.995843
BMD	Femoral neck	1	0.03217	1	2.2e-16	1.006456
BMD	Lumbar spine	1	1	1	2.2e-16	1.005034
BMD	Forearm	1	2.2e-16	1	2.2e-16	1.009849

class associations of the lactase persistence haplotype block on the following phenotypes: femoral neck bone mineral density, forearm bone mineral density, lumbar spine bone mineral density, individuals with prostate cancer, and individuals with coronary artery disease.

Enrichment Analyses

To determine if disease-associated markers were enriched in the lactase persistence haplotype, the distribution of *p*-values associated with SNPs in linkage disequilibrium and equilibrium were compared. A Fisher's Exact Test was used to test if the proportion of markers with an adjusted *p*-value less than 0.05 (Bonferroni correction) in each phenotype's meta-analysis results were higher among SNPs in high LD ($r^2 > 0.2$) with the lactase persistence SNP than expected under homogeneity. Wilcoxon rank sum tests were used to test if markers in high LD ($r^2 > 0.2$) with the lactase persistence marker (rs4988235) had significantly smaller *p*-values in a meta-analysis than those not in high LD. For each meta-analysis, this test was performed on three different sets of *p*-values:

(1) Two-sided *p*-value (i.e., the raw *p*-value provided in the disease phenotype meta-analysis results), to test if the log odds ratio beta does not equal zero, i.e., an effect of a SNP on the case status in either direction. A significant enrichment test based on these *p*-values means that the SNPs most associated with case status (e.g., CAD, prostate cancer), or BMD (regardless of direction) are enriched for SNPs in LD with the lactase persistence SNP.

(2) One-sided, upper, *p*-value to test if beta is greater than zero (derived from the Z statistic in the disease phenotype meta-analysis table), i.e., if a SNP is associated with a higher odds of case status or higher BMD. A significant enrichment test based on these *p*-values means that the SNPs most associated with higher odds of case status (e.g., CAD, prostate cancer), or higher BMD, are enriched for SNPs in LD with the lactase persistence SNP.

(3) One-sided, lower, *p*-value to test if beta is less than zero (derived from the Z statistic in the disease phenotype meta-analysis table), i.e., if a SNP is associated with a lower odds of case status or lower BMD. A significant enrichment test based on these *p*-values means that the SNPs most associated with lower odds of case status (e.g., CAD, prostate cancer), or lower BMD, are enriched for SNPs in LD with the lactase persistence SNP.

Software

The software versions used in this paper are R versions 3.3.1, 3.3.2, and 3.4.0, fcgene-1.0.7 (Roshyara and Scholz, 2014), METAL version 2011-3-25 (Willer et al., 2010), and PLINK v1.90p 64-bit (Purcell et al., 2007). Analyses were conducted using the following R packages: GenCAT version 1.0.3 (Qian et al., 2016), SNPRelate version 1.8.0 (Zheng et al., 2012), snpStats version (Solé et al., 2006).

RESULTS

Patterns of Linkage Disequilibrium

We determined the pattern of LD with rs4988235 in populations of European descent from the 1000 Genomes Phase 3 datasets. Unrelated subjects of CEU, FIN, GBR, IBS or TSI ethnicity were considered for the purpose of our analysis. These individuals (n = 407) had a minor allele frequency (MAF) of 0.497852 for the lactase non-persistent (ancestral) C allele at rs4988235 (2:136608646_G_A). The European lactase persistence haplotype block ($r^2 > 0.2$) was 1.89 Mb, spanned from rs3791298 (135,163,759) to rs6706934 (137,052,317) on chromosome 2 and contained 1,187 SNPs in high LD with rs4988235 (**Figure 1**, **Supplementary Figures S2, S3**). This linkage block of contains at least nine protein-coding genes and one micro-RNA. The protein-coding genes include RAB3GAP1, ZRANB3, R3HDMI, UBXN4, *LCT*, MCM6, DARS, CXCR4, THS7B, and possibly MAP3K19. The microRNA is MIR128-1. These genes and their known functions are listed in **Supplementary Table S1**. The microRNA, MIR-128-1, and two genes—CXCR4 and THSD7B—in the *LCT* locus all have putative roles in prostate cancer (**Supplementary Table S2**). We also manually reviewed the 100 SNPs most highly correlated with the lactase persistence SNP for known health associations in the literature. Numerous SNPs in the *LCT* locus were associated with total cholesterol or cardiovascular disease (**Supplementary Table S3**). Thus, the health phenotypes of prostate cancer and cardiovascular disease where chosen for further study. Bone mineral density phenotypes were also chosen for study due to the association of milk consumption with fracture risk in a cohort of people of European descent in which there was no assessment of genetics (Michaëlsson et al., 2014).

Health Phenotypes

For the health phenotypes of interest, we collected publicly available summary statistics from genome-wide meta-analyses

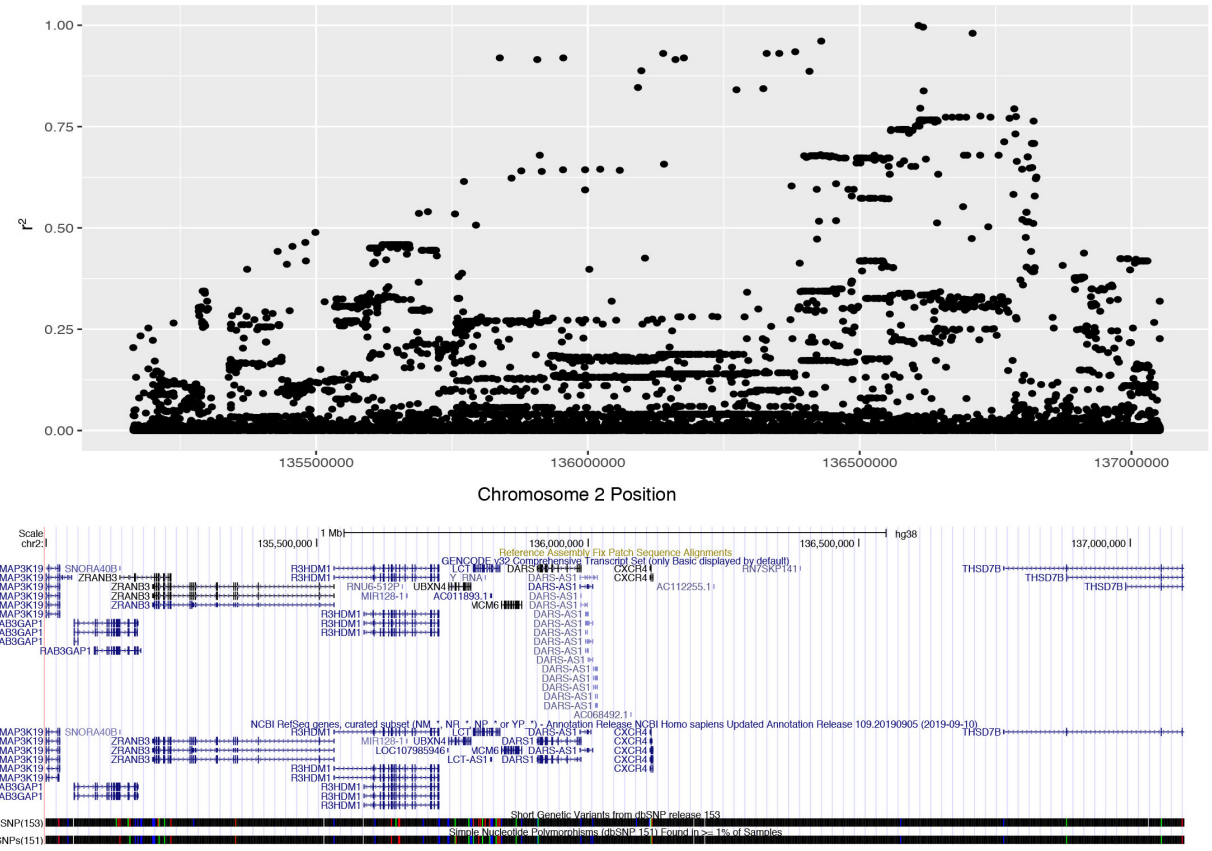


FIGURE 1 | Expanded Manhattan plot of lactase persistence haplotype block on chromosome 2 encompassing nine genes and one microRNA found in the human GRCh38 assembly. SNPs associated ($r^2 > 0.2$) with rs4988235 (chr2:135,851,076) span from rs3791298 (chr2:135,163,759) to rs6706934 (chr2:137,052,317).

and applied two types of statistical tests. For the first type of test—a class-level genetic association test (GenCAT)—used SNP-level meta-analysis test statistics across the class of all SNPs in high LD with the LP SNP ($r^2 > 0.2$) to determine whether that class of SNPs (the *LCT* locus) was statistically meaningful, given the size of the class and its unique correlation structure (Qian et al., 2016). The second type of statistical tests were enrichment tests to determine whether the SNPs in the *LCT* locus were enriched for phenotype-specific significance (based on SNP-level meta-analysis test statistics) relative to SNPs not in the locus (see Methods). In addition to the statistical tests, the mean beta for each phenotype was computed across all SNPs in the *LCT* locus as measure of the association of that phenotype with the *LCT* locus.

Prostate Cancer

We investigated the association of the European lactase persistence haplotype block with prostate cancer risk using genome-wide meta-analysis statistics of prostate cancer from 79,194 prostate cancer cases and 61,112 controls of the PRATICAL Consortium Oncoarray database (Schumacher et al., 2018). A Manhattan plot of the LP haplotype block in the context of chromosome 2 for the GWAS study of prostate cancer shows a SNP above -log10(P) > 3, but no SNPs rise to the level of significance of a standard GWAS (**Supplementary Figure S3**). The odds ratio of the LP SNP itself for prostate cancer risk was −0.01518. An omnibus analysis via a class-based association test (GenCAT) was conducted to determine if the lactase persistence haplotype block was associated with prostate cancer risk. The omnibus analysis revealed no significant association of markers in high LD with the LP SNP ($p = 1$, **Supplementary Table S4**). Next, enrichment analyses were conducted to determine if the distribution of *p*-values associated with SNPs in linkage disequilibrium with the LP allele differed from those associated with SNPs in equilibrium with the LP allele. Fisher's exact test revealed no significant difference in the proportion of prostate cancer-associated SNPs being higher among SNPs in high LD with the lactase persistence SNP than would be expected under homogeneity. The Wilcoxon rank sum tests revealed that *p*-values from GWAS were smaller for SNPs in LD than for those not in LD, regardless of the sign of beta, $p = 0.001393$). A one-sided Wilcoxon rank sum test revealed SNPs in high LD with the lactase persistence SNP had smaller *p*-values for one-sided tests of beta < 0, compared to SNPs not in high LD with lactase persistence ($p = 1.611e\text{-}7$, **Figure 2A**) suggesting a possible negative association of the *LCT* locus with prostate cancer disease relative to SNPs not in the locus. However, the mean beta for all SNPs in LD with LP was 1.02, indicating that the odds of the LP allele being associated with prostate cancer participants is not

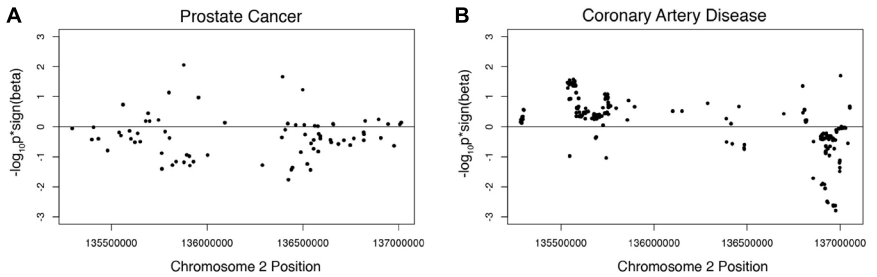


FIGURE 2 | Wilcoxon rank sum tests showing beta signed *p*-values for **(A)** prostate cancer associated SNPs in high LD with the lactase persistence SNP and **(B)** CAD associated SNPs in high LD with the lactase persistence SNP.

lower, but is nearly the same as association with controls. The former test (using Wilcoxon) is a comparison of the odds ratios in the *LCT* locus compared to the odds ratios outside of the *LCT* locus while the latter measurement (mean beta) is a statement about the odds ratios in the *LCT* locus taken by themselves. While the results are equivocal on whether there is a lower risk of prostate cancer with the LP allele, one can safely conclude that the LP allele does not significantly increase prostate cancer risk.

Coronary Artery Disease

Given the known SNPs associated with total cholesterol in the *LCT* locus (**Supplementary Table S3**), we next investigated the association of this locus with the risk of coronary artery disease (CAD) using summary data from CARDIoGRAMplusC4D 1000 Genomes-based GWAS dataset which included 60,801 cases and 123,504 controls (Nikpay et al., 2015). The odds ratio of the LP allele was 0.0141738 in the coronary artery disease dataset. A Manhattan plot of the LP haplotype block in the context of chromosome 2 suggests that no individual SNPs in this block are significant for CAD (**Supplementary Figure S4**). The omnibus analysis revealed no significant association of markers in high LD with the LP allele with coronary heart disease (p = 1, **Supplementary Table S4**). The Fisher's exact test revealed no significant difference in the proportion of CAD case-associated SNPs being higher among SNPs in high LD with the lactase persistence SNP than would be expected under homogeneity. A one-sided Wilcoxon rank sum test revealed CAD-associated SNPs in high LD with the lactase persistence SNP had smaller p-values from the one-sided test that beta > 0, compared to SNPs not in high LD with lactase persistence (p = 2.545e-11, **Figure 2B**, **Table 1**), suggesting a possible increase in odds of coronary artery disease associated with SNPs in LD with the LP allele, relative to those not in LD. However, the mean beta for all SNPs in LD with LP was 0.99 (**Table 1**), suggesting that the odds of the LP allele being associated with coronary artery disease is nearly the same as with controls. Taken together, there is limited support for an effect of the LP locus on the risk of CAD.

Bone Mineral Density

While most studies have shown a positive association of dairy consumption with bone mineral density, the paradoxical findings of increased fracture risk with increased milk consumption in a cohort of European descent suggested the potential for a genetic risk shared among those with the LP allele. We therefore investigated associations of the LP locus with bone mineral density (BMD), using genome-wide meta-analysis summary statistics of BMD of the femoral neck, the forearm, and the lumbar spine from 32,965 individuals in the Genetic Factors for Osteoporosis Consortium (GEFOS) consortium. Manhattan plots of the LP haplotype block in the context of chromosome 2 suggests that no individual SNPs in this block are significant for any of the three BMD traits (**Supplementary Figures S5–S7**). Like the other phenotypes, the GenCAT analyses revealed no significant association of markers in high LD with the LP allele for any of the BMD measurements (p = 1, **Supplementary Table S4**). A one-sided Wilcoxon rank sum test revealed femoral neck, lumbar spine and forearm associated SNPs in high LD with the lactase persistence SNP had smaller p-values for one-sided tests of beta < 0, compared to SNPs not in high LD with lactase persistence (p = 2.2e-16, **Figures 3A–C**, **Table 1**), suggesting a possible negative association of the *LCT* locus with bone mineral density relative to SNPs not in the locus. However, the mean beta for femoral neck, lumbar spine, and forearm BMD for all SNPs in phase with the LP SNP was approximately 1.01 for all three BMD sites (**Table 1**), suggesting little effect of the LP locus on BMD.

DISCUSSION

In the current study we investigated a locus in humans that encompasses the lactase gene (LCT) and shows signatures of strong, recent, positive selection with a relatively high deleterious mutation load, indictive of genetic hitchhiking. We identified the LP haplotype block in populations of European descent to be 1.9 Mb with at least nine protein-coding genes and one micro-RNA. Next, we investigated the association of the haplotype block with disease risk, prioritizing phenotypes based on known associations with SNPs in the region or with dairy consumption. Class-level (e.g., the class of all SNPs in the locus) associations were not significant for any of the phenotypes. Some of the enrichment tests were significant, but in the opposite direction of the sign of the average beta. Thus, despite recent selection for the lactase persistence allele there is little evidence that the LP haplotype block is associated with bone mineral density nor the risk of prostate cancer or coronary heart disease.

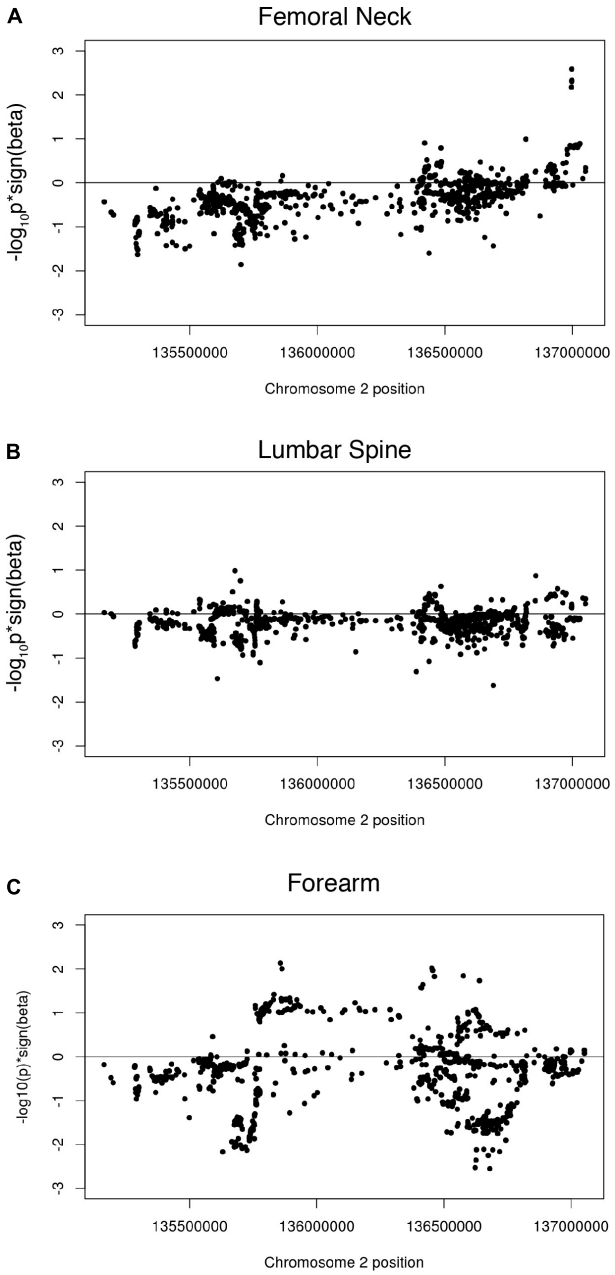


FIGURE 3 | Wilcoxon rank sum tests showing beta signed *p*-values for **(A)** femoral neck BMD associated SNPs in high LD with the lactase persistence SNP; **(B)** lumbar spine BMD associated SNPs in high LD with the lactase persistence SNP and **(C)** forearm BMD associated SNPs in high LD with the lactase persistence SNP.

Between people who harbor the LP allele and those who don't, there are potential differences due to both genetics (e.g., DNA hitchhiking with the LP SNP in those with the LP allele) and dairy consumption. There was no information about dairy consumption for the participants included in the current analysis. However, a recent study of healthy United States adults demonstrated that the LP genotype was only a very weak predictor of recent dairy intake as measured using 24-h recalls and was not significantly associated with overall habitual dairy intake measured using a food frequency questionnaire (Chin et al., 2019). Nevertheless, we are not able to quantify the contribution of dairy consumption to health phenotypes in the current study.

Prior studies have indicated that dairy consumption is associated with improved bone mineral density (Chan et al., 1995; Tai et al., 2015) and protective against cardiovascular disease (Alexander et al., 2016; Gholami et al., 2017; Dehghan et al., 2018), but potentially increases the risk of prostate cancer (Song et al., 2013; Aune et al., 2015). The relationship between dairy consumption and prostate cancer risk is controversial; a recent prospective investigation of 49,472 men in the United States found no association between dairy consumption and risk of prostate cancer (Preble et al., 2019). The fact that the lone microRNA in the *LCT* haplotype block has previously been shown to have a role in prostate cancer (Jin et al., 2014; Sun et al., 2015) implicated a potential for the genetics of the *LCT* locus to affect prostate cancer. However, we did not find evidence of this. MicroRNAs are highly redundant; for example, there is not a known knockout experiment of microRNAs in which their function has been shown to be essential. Thus, it is possible that deleterious mutations in or near the microRNA would have no effect on the phenotype.

Overall, across all studied phenotypes, there does not appear to be a notable effect of the LP locus on disease risk. This is surprising given the accelerated selection of the LP allele with increased potential for genetic hitchhiking and the fact that deleterious mutations have occurred in this locus at an increased rate (Chun and Fay, 2011). Our results suggest that these deleterious mutations have not given rise to unfavorable phenotypes for the studied diseases. Although it is possible that we have not studied a relevant phenotype, our manual review of the locus found SNPs or genes related to total cholesterol and prostate cancer, suggesting that cardiovascular disease and prostate cancer were the two phenotypes most likely to be affected by the genetics of the locus. Separately, the paradoxical findings of high milk consumption associated with increased fracture rate in a large cohort of European descent (Michaëlsson et al., 2014) might be explained by a genetic contribution given that dairy consumption is known to be associated with improved bone mineral density (Chan et al., 1995; Gholami et al., 2017; Dehghan et al., 2018). Thus, we covered the phenotypes most likely to be affected by the SNPs at the *LCT* locus and yet did not find an association of the locus with the phenotypes suspected to be affected. Additionally, we did not analyze other SNPs associated with LP, such as rs182549 (−22018G > A), due to having weaker associations of the derived allele and the LP phenotype (Enattah et al., 2002; Ségurel and Bon, 2017). Finally the rs4988235 T allele is not strictly predictive of the LP phenotype due to epigenetic effects (Leseva et al., 2018). Overall, our results suggest that rapid positive selection of the *LCT* locus and an increase in deleterious mutations have not translated into unfavorable disease phenotypes in humans.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz; http://www.gefos.org/sites/default/files/README.txt; http://practical.icr.ac.uk/blog/?page_id=8164; www.CARDIOGRAMPLUSC4D.ORG.

AUTHOR CONTRIBUTIONS

SJ carried out manual reviews, contributed to the collection of the datasets, carried out bioinformatic analyses, interpreted results, and wrote the manuscript. BD-J conducted statistical analysis and interpretation. MB conducted bioinformatics analyses. MS and IK provided bioinformatics guidance. DL conceived and directed the study, coordinated the collection of the datasets, interpreted results, and wrote the manuscript. All authors contributed to study design, reviewed and edited the manuscript.

FUNDING

This work was supported by California Dairy Research Foundation. This research was also funded in part by United States Department of Agriculture 2032-51530-026-00D. The United States Department of Agriculture is an equal opportunity provider and employer. For the PRACTICAL consortium data, we thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign United Kingdom (now Prostate Action), The Orchid Cancer Appeal, The National Cancer Research Network United Kingdom, The National Cancer Research Institute (NCRI) United Kingdom. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. The Prostate cancer genome-wide association analyses are supported by the Canadian Institutes of Health Research, European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), Cancer Research United Kingdom Grants C5047/A7357 C1287/A10118, C1287/A16563, C5047/A3354, C5047/A10692, C16913/A6135, and the National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148537-01 (the GAME-ON initiative). Genotyping of the OncoArray was funded by the United States National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I] and by Cancer Research United Kingdom grant A8197/A16565. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher).

ACKNOWLEDGMENTS

We thank Fotios Drenos and Brendan J. Keeting for making earlier cardiovascular disease SNP data available to us. We thank Zeynep Alkan for comments on the manuscript. We thank the PRACTICAL consortium, CARDIoGRAMplusC4D Consortium, and GEFOS consortium for public access to their respective data sets. The data on coronary artery disease/myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.558762/full#supplementary-material

REFERENCES

Alexander, D. D., Bylsma, L. C., Vargas, A. J., Cohen, S. S., Doucette, A., Mohamed, M., et al. (2016). Dairy consumption and CVD?: a systematic review and meta-analysis. *Br. J. Nutr.* 115, 737–750. doi: 10.1017/S0007114514995000

Aune, D., Navarro Rosenblatt, D. A., Chan, D. S., Vieira, A. R., Vieira, R., Greenwood, D. C., et al. (2015). Dairy products, calcium, and prostate cancer risk?: a systematic review and meta-analysis of cohort studies. *Am. Soc. Nutr.* 101, 87–117. doi: 10.3945/ajcn.113.067157.Prostate

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111–1120. doi: 10.1086/421051

Chan, G. M., Hoffman, K., and McMurry, M. (1995). Effects of dairy products on bone and body composition in pubertal girls. *J. Pediatr.* 126, 551–556.

Chin, E. L., Huang, L., Bouzid, Y. Y., Kirschke, C. P., Durbin-Johnson, B., Baldiviez, L. M., et al. (2019). Association of lactase persistence genotypes (Rs4988235) and ethnicity with dairy intake in A. *Nutrients* 11, 1–23. doi: 10.3390/nu11081860

Chun, S., and Fay, J. C. (2011). Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet.* 7:e1002240. doi: 10.1371/journal.pgen.1002240

Cutter, A. D., and Payseur, B. A. (2013). Genomic signatures of selection at linked sites : unifying the disparity among species. *Nat. Rev. Genet.* 14, 262–274. doi: 10.1038/nrg3425

Dehghan, M., Mente, A., Rangarajan, S., Sheridan, P., Mohan, V., Iqbal, R., et al. (2018). Association of dairy intake with cardiovascular disease and mortality in 21 countries from five continents (PURE): a prospective cohort study. *Lancet* 392, 2288–2297. doi: 10.1016/S0140-6736(18)31812-9

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30, 233–237. doi: 10.1038/ng826

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive darwinian selection. *Genetics* 155, 1405–1413.

Gholami, F., Khoramdad, M., Esmailnasab, N., Moradi, G., Nouri, B., Safiri, S., et al. (2017). The effect of dairy consumption on the prevention of cardiovascular diseases: a meta-analysis of prospective studies. *J. Cardiovasc. Thorac. Res.* 9, 1–11. doi: 10.15171/jcvtr.2017.01

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., and Thomas, M. G. (2009). The origins of lactase persistence in Europe. *PLoS Comput. Biol.* 5:e1000491. doi: 10.1371/journal.pcbi.1000491

Jin, M., Zhang, T., Liu, C., Badeaux, M. A., Liu, B., Liu, R., et al. (2014). MiRNA-128 suppresses prostate cancer by inhibiting BMI-1 to inhibit tumor-initiating cells. *Cancer Res.* 74, 4183–4195. doi: 10.1158/0008-5472.CAN-14-0404

Leseva, M. N., Grand, R. J., Klett, H., Boerries, M., Busch, H., Binder, A. M., et al. (2018). Differences in DNA methylation and functional expression in lactase persistent and non-persistent individuals. *Sci Rep*. 8:5649. doi: 10.1038/s41598-018-23957-4

Michaëlsson, K., Wolk, A., Langenskiöld, S., Basu, S., Warensjö Lemming, E., Melhus, H., et al. (2014). Milk intake and risk of mortality and fractures in women and men: cohort studies. *BMJ* 349, 1756–1833. doi: 10.1136/bmj.g6015

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A Comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47, 1121–1130. doi: 10.1038/ng.3396

Preble, I., Zhang, Z., Kopp, R., Garzotto, M., Bobe, G., Shannon, J., et al. (2019). Dairy product consumption and prostate cancer risk in the United States. *Nutrients* 1–12. doi: 10.3390/nu11071615

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Qian, J., Nunez, S., Reed, E., Reilly, M. P., and Foulkes, A. S. (2016). A simple test of class-level genetic association can reveal novel cardiometabolic trait loci. *PLoS One*:e0148218. doi: 10.1371/journal.pone.0148218

Roshyara, N. R., and Scholz, M. (2014). FcGENE?: a versatile tool for processing and transforming SNP datasets. *PLoS One* 9:e97589. doi: 10.1371/journal.pone.0097589

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., et al. (2018). association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* 50, 928–936. doi: 10.1038/s41588-018-0142-8

Ségurel, L., and Bon, C. (2017). On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.* 18, doi: 10.1146/annurev-genom-091416-35340

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Res.* 23, 23–35. doi: 10.1017/S0016672308009579

Solé, X., Guinó, E., Valls, J., Iniesta, R., and Moreno, V. (2006). SNPStats : a web tool for the analysis of association studies. *Bioinformatics* 22, 1928–1929. doi: 10.1093/bioinformatics/btl268

Song, Y., Chavarro, J. E., Cao, Y., Qiu, W., Mucci, L., Sesso, H. D., et al. (2013). Whole milk intake is associated with prostate cancer-specific mortality among U.S. Male Physicians. *J. Nutr.* 143, 189–196. doi: 10.3945/jn.112.168484

Sun, X., Li, Y., Yu, J., Pei, H., Luo, P., and Zhang, J. (2015). MiR-128 modulates chemosensitivity and invasion of prostate cancer cells through targeting ZEB1. *Jap. J. Clin. Oncol.* 45, 474–482. doi: 10.1093/jjco/hyv027

Swallow, D. M. (2003). Genetics of lactase persistence and lactose. *Annu Rev Genet.* 37, 197–219. doi: 10.1146/annurev.genet.37.110801.143820

Tai, V., Leung, W., Grey, A., Reid, I. R., and Bolland, M. J. (2015). Calcium intake and bone mineral density: systematic review and meta-analysis. *BMJ* 351:h4183. doi: 10.1136/bmj.h4183

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Zheng, H. F., Forgetta, V., Hsu, Y. H., Estrada, K., Rosello-Diez, A., Leo, P. J., et al. (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* 526, 112–117. doi: 10.1038/nature14878

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.