I. AYK SSI TITLE PAGE:

Project Title: Genomics of maturation age in Yukon Chinook

Investigator(s): Dr. Eric P. Palkovacs
Department of Ecology and Evolutionary Biology
University of California Santa Cruz
Email: epalkova@ucsc.edu

Dr. Peter Westley
College of Fisheries and Ocean Sciences
University of Alaska Fairbanks
Email: pwestley@alaska.edu

John Linderman
Alaska Department of Fish and Game
Email: john.linderman@alaska.gov

James Savereide
Alaska Department of Fish and Game
Email: james.savereide@alaska.gov

Dr. Eric Anderson
National Marine Fisheries Service
Email: eric.anderson@noaa.gov

Dr. John Carlos Garza
National Marine Fisheries Service
Email: carlos.garza@noaa.gov

II. ABSTRACT:

Selective fishing may cause evolutionary changes in harvested fish stocks, but widespread genomic evidence for 'Fishing Induced Evolution' remains lacking. Rapid declines in size and age at maturation in AYK Chinook salmon is a potential harbinger of evolutionary change linked to selective fishing. Here we seek to test the overarching hypothesis that variation in age, and therefore size, of adult Chinook salmon in the Yukon River watershed has an underlying genomic basis associated with specific loci. Testing of this hypothesis is the first step towards addressing whether changes in escapement quality are the result of evolutionary change consistent with selective gill net fishing. Our overarching objective is to characterize the genomic architecture of maturation age in Yukon Chinook using low coverage whole genome sequencing of 454 Chinook spanning age ranges from three to seven years. To accomplish this goal, we had to identify the sex determination region in the Chinook from this study to confirm genomic sex. We identified that ~30% of Chinook being aged and sexed at Pilot station were being misidentified as the incorrect sex. Once we had sexes genetically distinguished, we searched for candidate regions and candidate genes associated with maturation age in male and female fish collected at Pilot Station (Kusilvak, Yukon River) between 2007 and 2020. We identified strikingly different patterns between male and female Chinook salmon. The results show that maturation has a different genomic basis in males versus females. We found a strong association with maturation age in males on Chromosome 17 (the sex determining region or SDY), but several potential candidate regions associated with maturation age in females, non with the same association as in males. This result means that selection against the late-maturation in males is decoupled from selection against late-maturation in females, and fisheries selection is more likely to cause an evolutionary response in males than in females. Since external sex determination at Pilot Station was found to be unreliable, testing this hypothesis in the future will require combining genetic sex determination with age and size data.

II. TABLE of CONTENTS:

Figures:

7. Figure 7. Manhattan plot of associations between genetic variation at 4,888,421 sites throughout the genome and presence or absence of SDY sequences in an individual.

Tables:

1. Table 1: Characteristics of Yukon River Chinook salmon samples and outcome of whole genome sequencing efforts. Sex IDs are field recorded. Numbers of reads and mapped bases are in Millions.
2. Table 2: A summary of the number of individuals of different morphologically reported sex that were found to be males and females based on the genomic data.

III. INTRODUCTION:

We are in the midst of a genomics revolution that is poised to greatly increase our knowledge of the genetic architecture that underpins fitness-related traits in exploited populations. Declines in the size and age at maturity of AYK Chinook salmon may be a harbinger of fishing induced evolutionary trait change, but to date has not been formally tested. This project represents a necessary first step towards directly addressing the genetic components of reproductive phenotypes that could affect the abundance, productivity, and persistence of Chinook salmon populations and in turn well-being of salmon-dependent peoples. By doing so, we squarely contribute to Theme 3 of the AYKSSI Research Priorities that explores the hypothesis that selective fishing has genetically altered the size and age structure of AYK Chinook salmon.

The truncation of size and age structure is one of the most pervasive signals of size-selective exploitation of wild animals (Darimont et al., 2009). Extensive research has sought to understand the consequences of phenotypic shifts towards smaller and younger maturing individuals for long-term population persistence and harvest sustainability. In particular, reductions in size-dependent female fecundity and increased recruitment variability from homogenized age structure are predicted to weaken portfolio effects. Although the conservation implications of losing so called 'Big Fat Old Females' from populations have been the subject of substantial controversy in the literature, there is a growing consensus that truncated size and age structures impacts population growth, productivity, and resiliency. In a compelling example, approximately 20% of observed population growth in a wild population of Soay sheep can be linked to the distribution of body sizes within the population (Pelletier et al., 2007). Within the fisheries realm, the preservation of old-growth age structure has been likened to one of the 10-commandments for sustainable exploitation of marine resources (Francis et al. 2007). Indeed, hindsight has revealed rapid shifts in age structure and age at maturity in northern cod (*Gadus morhua*) were precursors to population collapse (Olsen et al. 2004). Even under modest harvest rates, size-selective fisheries can rapidly shift age distributions toward younger individuals, making it necessary to understand the genomic basis of maturation for achieving population recovery.

Multiple lines of experimental and empirical evidence are consistent with the potential for selective fisheries to result in contemporary evolution with potential consequences for long-term yield and population persistence. A now classic selection experiment using the Atlantic silverside (*Menidia menidia*) revealed nearly twice the yield in terms of numbers (and even more for biomass) in selected lines where small fish were harvested vs. lines where big fish were harvested, mirroring typical patterns of selection in fisheries (Conover and Munch, 2002). Although the results are compelling, the intense and consistent selection regimes imposed upon the experimental lines are unlikely to occur in

nature. Indeed, it is clear that natural selection varies in both time (Siepielski et al., 2009) and space (Kingsolver et al., 2012). Similarly, size-dependent selection resulting from fisheries is known to be highly variable.

Further complicating the implication of fishing on evolution is the widespread lack of knowledge concerning the genetic architecture underpinning phenotypic expression of size and age at maturity in wild, exploited populations of non-model organisms. Multi-generation common garden studies remain the gold standard for partitioning genetic and environmental effects on phenotypes but are often financially and logistically prohibitive for long-lived organisms. In contrast, recent technological advances in genomics that make analyses more effective and cost efficient are poised to revolutionize our understanding of human drivers of evolutionary change (Palumbi 2001, MacColl 2011). Consistent with predictions of fishing induced evolution, Pacific salmon (*Oncorhynchus* spp.) are returning to spawn at younger ages and smaller sizes throughout the North Pacific basin (e.g. Bigler et al., 1996, Jeffrey et al., 2016, Oke et al., 2020). These trends have been identified and contemplated since at least the late 1970s, when Ricker (1981) published declines in size of Chinook salmon (*O. tshawytscha*) harvested in troll fisheries along the West Coast of North America. In this foundational paper, Ricker proposed several non-mutually exclusive hypotheses to explain the trends in declining size. Of these, he suggested potential changes in the stock composition towards populations with smaller average sizes, perhaps the result of selective removal or elimination of populations composed of large-bodied individuals. Given that he was observing changes reflecting population admixtures captured in mixed-stock fisheries, this was indeed a plausible hypothesis. Further, he proposed that size-selective fisheries would favor early age at maturity since individuals that delayed maturation and stayed longer at sea would be more likely to be caught. In addition, Ricker proposed that hatchery selection favoring larger older fish may have been unintentionally selected for slow growth resulting in declines in size-at-age.

Ideal situations to assess the genetic hypotheses suggested by Ricker would be found in locations where i) size-selective fisheries are known or suspected that might drive evolutionary change, ii) information exists on sizes and ages from individuals of distinct spawning populations rather than complex stock conglomerates, and iii) hatchery propagation is non-existent. Observations of on-going shifts towards smaller and younger adult Chinook salmon continue to mount (Lewis et al., 2015; Ohlberger et al., 2018; Oke et al. 2020). Yet to date, the major knowledge gap remains whether the widespread changes observed in sizes and ages of Pacific salmon are an evolutionary response to selective fishing. This gap currently impedes our ability to understand population responses to past management decisions and precludes an informed prediction of how salmon may respond to future alternative selective harvest scenarios. Chinook salmon are returning younger and smaller to the Yukon River of western Alaska, with changes in this ‘escapement quality’ having profound socio-ecological-cultural consequences in the region. Western science has largely corroborated the observations of local subsistence fishermen that fish in the Yukon River have gotten smaller in recent decades.

The Yukon River is an ideal region to explore the genetic underpinnings of shifts in age structure given the pressing applied management and cultural implications as well as a lack of confounding sources of variation first noted by Ricker and colleagues decades ago. Specifically, long-term high quality data on age and size (ASL) are available from individual spawning populations where hatchery propagation does not occur. Moreover, Chinook salmon are harvested in size-selective gill net fisheries, and previous quantitative genetic simulation modeling supported by AYKSSI has suggested that changes are consistent with fishing induced evolution (Bromaghin et al. 2011), though direct tests through common garden experiments or genetic analyses have yet to be done.

Here we investigated the underlying genetic architecture of age at maturation in Yukon Chinook salmon. To address this question, we conducted an experimental design and genomics approach which has recently been used to understand the genetic basis of life history traits in several salmonid species. A genomic region of major effect has been shown to be highly associated with early and late run-timing in both steelhead trout (*Oncorhynchus mykiss*) and Chinook (*O. tshawytscha*) (Hess et al. 2016; Narum et al., 2018; Micheletti et al., 2018; Thompson et al. 2020). Another region of large effect controlling variation for age of maturation has also been identified in Atlantic salmon (*Salmo salar*) explaining ~39% of variation (Allyon et al., 2015; Barson et al., 2015). A recent study in the Columbia River basin (Micheletti & Narum, 2017) failed to associate this region with age of maturation in Chinook salmon but did identify several other candidate genes. Hatchery studies on Chinook have pointed to high heritability of age at maturation (0.35 – 0.42; Hard, 2004; Hankin et al., 2009) as well as potential sex dependent heritability (Hankin et al., 1993). We used fine-scale mapping from resequencing data of Yukon Chinook genomes to identify if there are genomic regions of major effect associated with maturation age. We followed our initial screening by characterizing candidate regions of highest association across the variation of age classes to validate the candidate loci. Our study provides insight into the genomic regions underlying age at maturation, as well as the implications for future management. For example, a few regions of large effect would be expected to lead to a higher vulnerability to selective fishing and a slower recovery time compared to many regions of small effect.

IV. OBJECTIVES:

Objective I: Characterize the underlying genetic architecture of age of maturation in AYK Chinook salmon. **We accomplished this goal.**

Objective II: Develop amplicons for candidate regions and apply to a 10 year (two generation) dataset of preserved tissue samples linked to ASL information for the Yukon River to serve as a proof of concept prior to use on longer-archived time series. **Because the genomic basis of maturation differs between males and females (see below), we did not develop maturation markers at this stage of the project. Instead, we did almost 50% more whole-genome sequencing than was anticipated, in place of the amplicon sequencing, so made good use of the funds.**

V. METHODS:

Study area:

Alaska, Kusilvak, Yukon river

Study design:

The initial study design centered around collecting ~340 Chinook from the Salcha and Chena rivers in 2020 and 2021 (if necessary) from different sexes and age classes. Due to warm water temperatures, flooding and covid, only 45 fish were collected in 2020 and no fish in 2021. Therefore an alternative strategy was needed to complete the project. Aged and sexed specimens collected from Pilot Station on the Yukon river (Kusilvak, Alaska) between 2007 to 2020 were then used to complete the study. One caveat to this modification was that these fish would then be a mixed stock sample, leading to the complication of having structure within the study. We can also not presently identify which populations

the individuals used in the study come from, however we adjusted our analysis approach to account for this structure.

Data generation:

We generated high quality genomic libraries for 454 Chinook salmon from the Yukon River, captured at the Pilot Station test fishery in the lower basin. Fish for analysis were chosen to represent the extremes of the age distribution for both females and males. Details of sample distribution, with field-recorded sex, are in Table 1.

DNA extraction, library preparation and sequencing:

Genomic DNA was extracted from fin tissue material using the DNEasy 96 Blood and Tissue Kit protocol, modified with a preliminary RNAse treatment, on a BioRobot 3000 (Qiagen Inc). We used NEBNext Ultra II FS (New England Biolabs) reagents to successfully prepare genomic libraries from all samples received. Libraries were then sequenced in a NovaSeq with an S4 flowcell, paired-end 150 reads, at a targeted coverage of 2-3x.

Table 1: Characteristics of Yukon River Chinook salmon samples and outcome of whole genome sequencing efforts. Sex IDs are field recorded. Numbers of reads and mapped bases are in Millions.

SEX	AGE	Sample Size	Mean (S.D.) no. of reads (M)/individual	Mean (S.D.) no. of bases mapped (M)/individual
Female	4	65	45.7 (12.6)	7.1 (2.0)
Female	6	100	46.5 (9.2)	7.1 (1.4)
Female	7	58	48.5 (12.1)	7.4 (1.9)
Male	3	13	49.9 (12.1)	7.7 (1.9)
Male	4	89	48.8 (11.5)	7.6 (1.8)
Male	6	103	48.6 (11.6)	7.5 (1.8)
Male	7	26	52.2 (16.5)	8.1 (2.5)

Bioinformatic Processing of Genomic Sequence Data:

The sequencing center returned paired-end sequence data in the form of gzipped FASTQ files for 454 Yukon Chinook salmon. These data were processed using a fairly standard bwa-mem (Li and Durbin, 2009) and the GATK pipeline (McKenna et al., 2010) as follows:

- **Trimming**: Adapters not previously removed from the sequence were removed using Trimmomatic version 3.6 (Bolger et al., 2014). At the same time, the ends of the reads were clipped from the point that a base quality score below 3 was found. Finally, a sliding window

trim was done such that the remainder of a read was discarded when the average base quality score in a window of size 4 bases dropped below 15. The latter trimming step has been shown to be effective in removing PolyG tails from the ends of sequencing reads produced on the Illumina NovaSeq platform (Lou and Therkildsen 2022).

- **Mapping**: Trimmed reads were mapped (aligned) to the recently assembled Otsh_v2.0 version of the Chinook salmon genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_018296145.1). This reference genome was made from a male Chinook salmon, so that it should include sequences from the sex-determination on the Y (SDY) locus, which is only found in males. Reads were mapped using bwa mem version 0.7.17 with default settings. Alignments were converted to binary format and sorted into coordinate order using samtools version 1.9 (Li et al., 2009). We provide a summary of the mean and standard deviation of the number of properly paired reads and the total number of bases mapped for individuals of different sexes and ages (Table 1)
- **Marking of duplicates**: We used the PicardTools (version 2.27) program MarkDuplicates to identify and mark PCR duplicates and optical duplicates found in the BAM files created in the mapping step. Marking of such duplicates is necessary that only the information in one member of a duplicated pair is used when calling genotypes.
- **Calling of individual sequencing variation**: We summarized the information about sequence variation in each individual into a gVCF file using the GATK HaplotypeCaller with default settings. In this, and all further instances of GATK, we used version 4.2.6.1.
- **Joint genotype calling**: To call genotypes jointly across all 454 samples, we loaded the data in all 454 gVCF files into a genomics database using GATK GenomicsDBImport. Subsequently the GATK program GenotypeGVCFs was used to jointly call genotypes at all 454 samples and record them in a VCF (variant call format) file.
- **Hard filtering**: Because insufficient genomic and variant databases are available for Chinook salmon to allow for the AI-powered variant filtration methods used with the GATK for human data, we adopted a hard filtering strategy—removing variant sites that did not pass our filters. The filters we adopted were as follows:
    - For insertions/deletions:
        - -filter 'QD < 2.0' --filter-name 'QD2' "
        - -filter 'QUAL < 30.0' --filter-name 'QUAL30' "
        - -filter 'FS > 200.0' --filter-name 'FS200' "
        - -filter 'ReadPosRankSum < -20.0' --filter-name 'ReadPosRankSum-20'
    - For single nucleotide polymorphisms:
        - -filter 'QD < 2.0' --filter-name 'QD2' "
        - -filter 'QUAL < 30.0' --filter-name 'QUAL30' "
        - -filter 'SOR > 3.0' --filter-name 'SOR3' "
        - -filter 'FS > 60.0' --filter-name 'FS60' "
        - -filter 'MQ < 40.0' --filter-name 'MQ40' "
        - -filter 'MQRankSum < -12.5' --filter-name 'MQRankSum-12.5' "
        - -filter 'ReadPosRankSum < -8.0' --filter-name 'ReadPosRankSum-8' "
    - Finally, both types of variants were filtered to include only those with a minor allele frequency greater than 0.0225.
- **Base quality score recalibration**: The VCF file made from the above hard-filtering step was inspected. In particular the distributions of the variant quality scores (QUAL) and the variant quality score normalized by read depth (QD) were examined. From this examination, we chose a rule of QUAL > 500 and QD > 9 for retaining a set of variants that would be considered known variants for the purposes of base quality score recalibration. With such a “known” variant set,

we used the GATK programs BaseRecalibrator and ApplyBQSR to create a new BAM file of alignments with updated base quality scores for each individual.

- **Final variant calling**: From the BAM files with recalibrated base quality scores we repeated the steps *calling of individual sequencing variation*, *joint genotype calling*, and *hard filtering*, described above to obtain our final VCF. It is important to note that this VCF file includes the genotype likelihoods for all the genotype calls, and it was these genotype likelihoods that were used in downstream analyses.

Genomic Sex Identification:

Because fish sampled at the Pilot Station might not have developed pronounced secondary sexual characters, it is difficult to accurately assign sex to them on the basis of morphology. Although we were provided with estimates of the sex (hereafter "reported sex") of all the samples, we found these to be inaccurate, so we assigned sex to samples using the presence or absence of sequence reads from the sex determing gene known as SDY (hereafter "genomic sex").

To locate the SDY region in the Otsh_v2.0 genome, we mapped the 6854 base-pair sequence of the SDY available on Genbank (https://www.ncbi.nlm.nih.gov/nuccore/KC756279.2) to the Otsh_v2.0 genome. All but the last 104 bp of that sequence mapped in a contiguous fashion, starting from position 2431 on the unplaced scaffold NW_024608692.1 in the Otsh_v2.0 genome. Accounting for single-base-pair insertions and deletions, and the 104 base pairs that were soft clipped from the SDY, we concluded that the SDY maps to the genomic coordinates NW_024608692.1:2431-9131 in the Otsh_v2.0 genome. Therefore, we inferred the presence of SDY sequences in a sample by comparing the average read depth in the genomic interval NW_024608692.1:2431-9131 to the total number of bases mapped throughout the genome in each sample. We obtained the average read depth in the NW_024608692.1:2431-9131 region with *samtools depth*(version 1.15.1), and we calculated the total number of bases mapped in from each sample with *samtools stats*(version 1.9).

Population Structure:

Using *pcangsd* (version 1.10) (Meisner and Albrechtsen, 2018), we conducted a principal components analysis to investigate the population-genetic structure amongst the samples. We also investigated the relationship between this apparent population structure and several variables that might vary by population—notably date of collection and age at maturity. Finally, we investigated the relationship between the principal components and library preparation group, sequencing flow cell, and lane to identify if there were any notable artifacts and batch effects of those experimental differences between samples on their population genetic features. There were not (results not shown).

Genome-wide association study (GWAS) for age at maturity:

Previous studies have found certain male-specific haplotypes (those on the same chromosome as the SDY) to be associated with age at maturity in male Chinook salmon. This finding, as well as the fact that males tend to mature at an earlier age than females, argues for analyzing genomic associations with age separately in males and females. We have done that here.

Starting with either all the males or all the females in our study, our workflow for GWAS was as follows:

1. Perform a PCA with all the members of the same sex using *pcangsd*(version 1.1.0), recording the principal component values for all samples. During this step we filtered variants to have an estimated minor allele frequency greater than 0.05.

2. Calculate genotype posteriors for each sample from the results of the PCA and each sample's genotype likelihoods. This is currently the recommended practice for low-coverage whole genome sequencing data (Jørsboe and Albrechtsen 2022), where it has been described as using *individual allele frequency priors.* We did this using a customized version of *pcangsd* 1.10, (available from: https://github.com/eriqande/pcangsd).
3. Use the genotype posteriors in the program *angsd* (version 0.937) (Korneliussen et al., 2014) with the -doAsso option to perform a GWAS using a generalized linear model with age considered as a binary trait (ages 3 and 4 in one category, and 6 and 7 in another), while including cohort (year that the parents of the sample spawned) as a discrete covariate trait and the first four principal components (PCs) of the PCA as continuous covariates (to account for population structure). We chose to use the first four PCs because the first four eigenvalues were considerably larger than the rest, they accounted for much of the variation in the data, and the eigenvalues higher than 4 were all roughly of the same size. Additionally, we performed GWAS including the first 12 PCs as covariates and found no appreciable differences
4. For each SNP site in the genome, the GWAS method returns a likelihood ratio statistic score that assesses the strength of an association between genetic variation at the site and the phenotype of interest (in this case, age at maturity).

Under the null hypothesis of no association, the likelihood ratio statistics are distributed as a Chi-square random variable with one degree of freedom. Accordingly, we converted the likelihood ratio statistics to p-values for the test of association. We summarized these p-values in Manhattan plots showing the log (base 10) of the p-value on the y-axis and the position of the marker in the genome on the x-axis.

Assessing the accuracy of the reference genome assembly using associations with SDY:

During an early phase of our work on this project it appeared to us that there were strong associations between sex and the several parts of the genome. While some degree of association might be expected among genes that function specifically or primarily in one sex or the other, the degree of association we saw was much larger than expected. We followed up those early findings by performing an association test between the genetic variation at millions of sites across the genome with sex, as defined by the presence (males) or absence (females) of sequences at the SDY to identify other regions of the genome that are part of SDY or closely associated to SDY but misassembled in the Chinook genome.

Identifying significantly associated sites via the False Discovery Rate:

We used a false discovery rate correction to identify sites that may have a statistically significant association with age at maturity. In females, we used a desired false discovery rate 0.1 and in males we used a desired false discovery rate of 0.25. We then used *bedtools*(version 2.30.0) to identify which of the 11,950 named genes in the Otsh_v2.0 reference genome are within 10 Kb of each of the associated sites.

VI. RESULTS:

Data quality and final SNP dataset:

All sequenced genomes were retained in the study and coverage varied from 1x - 3x coverage. When filtered to a minor allele frequency greater than 0.0225, this VCF contained records for: 8,744,215 SNPs and 2,436,754 indels.

Assigning Genomic sex:

After interrogating the SDY region, it was clearly evident which samples carried sequences from the SDY regions, and which ones did not (Figure 1).

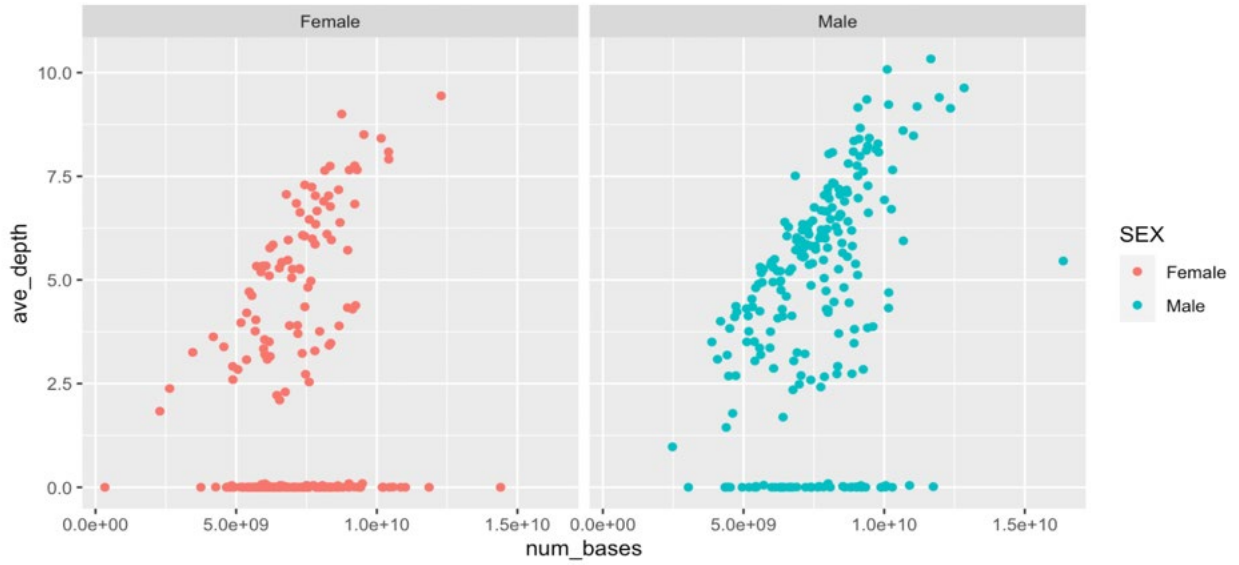

Figure 1. In the scatterplot, each point represents a sample. The y-axis shows average read depth within the SDY region of Otsh_v2.0. The x-axis shows the total number of base pairs mapped. Individuals that do not carry the SDY fall on or near the y=0 line.

Using Figure 1, it was simple to assign genomic sex to each sample. Males were defined as individuals with an average read depth greater than 0.5 in the NW_024608692.1:2431-9131 interval. Females were defined as individuals having read depth less than 0.5 in that region. Unsurprisingly, this did not align well with the reported sexes (Table 2).

Table 2: A summary of the number of individuals of different morphologically reported sex that were found to be males and females based on the genomic data.

	Genomic Females	Genomic Males
Reported Females	133	90
Reported Males	49	182

For subsequent analyses in which sex was needed as an input, we used the genomic sex of the individuals.

Population Structure:

Figure 2 shows a biplot of the first two principal components when all samples were analyzed together. It shows at least three clusters, which likely correspond to population structure in the samples. This would not be surprising, since spawners returning to many different populations across large areas of the watershed are sampled at Pilot Station. However, at this time we cannot say which populations they originate from.

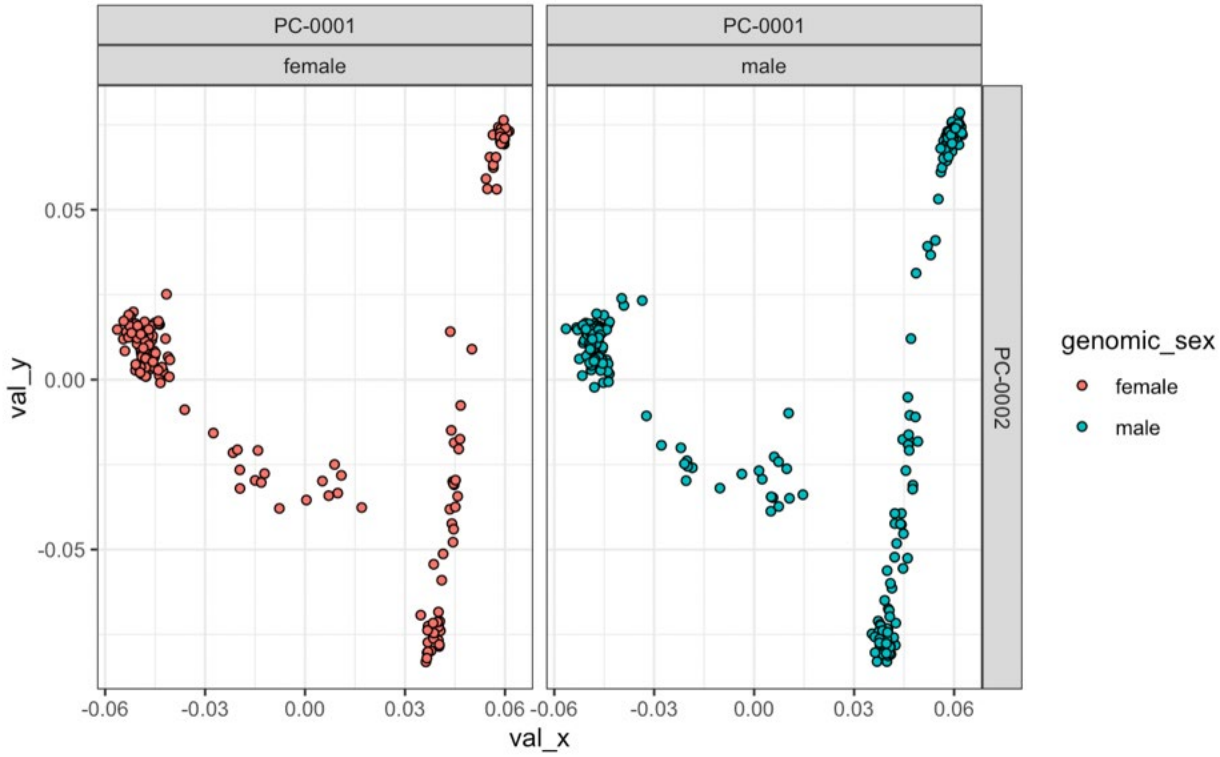

Figure 2. Scatterplot of first two principal components, colored and faceted by genomic sex.

Assessing this population variationmby sex in terms of collection date there appears visually no clear signal of variation in collection date across the different clusters in the principal components plot (Figure 3).

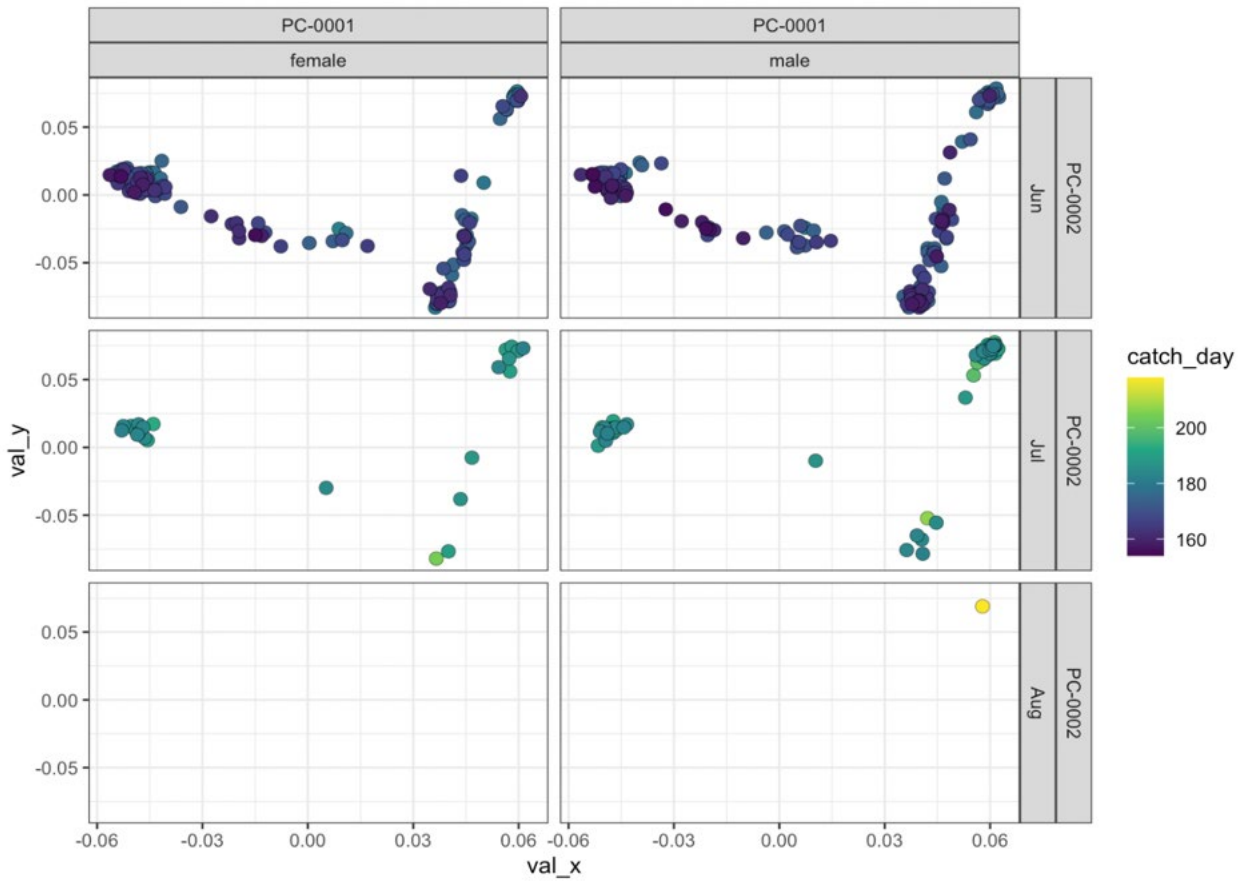

Figure 3. The first two principal components and their relationship to collection date across all years of samples. Left column are females, the right column males. The rows correspond to different months of collection and colors denote the day of collection (catch_day).

On the other hand, there does appear to be some variation in the distribution of age at return across the different clusters in the plot (Figure 4). Most notably, among males, the cluster in the lower right of the plot includes a very large number of four-year-olds with nearly no 7 year-olds in the same cluster, while the cluster on the left of the plot appears to include considerably more 6- and 7-year-olds, relative to the number of four-year-olds. This finding indicates that we will certainly need to account for population structure when performing a genome-wide association study for age at maturity.

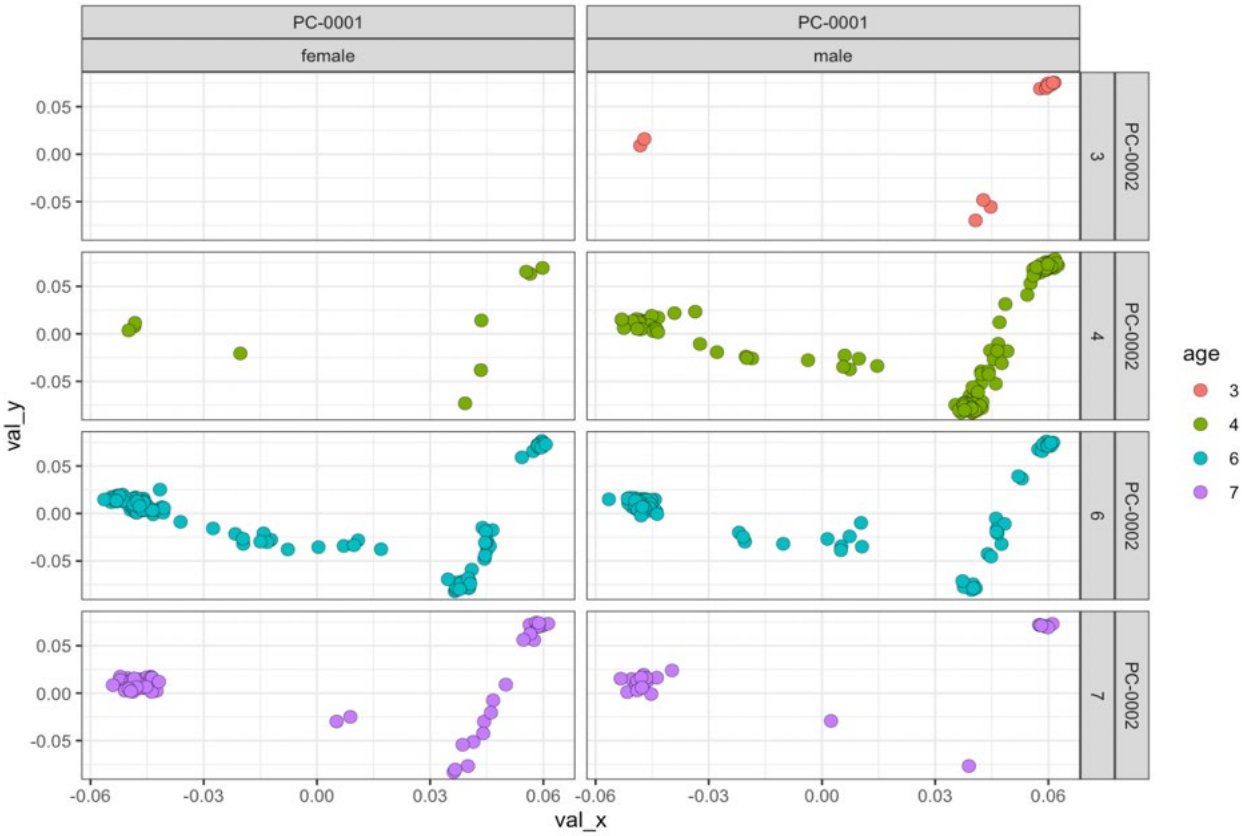

Figure 4. The first two principal components and their relationship to age and sex across all years of samples. Left column holds females, right column, males. The rows correspond to different ages.

Genome-wide Association Study:

Inspection of the Manhattan plot for the association in males shows a large cluster of elevated negative log10-p-values (thus, very small p-values) upon Chromosome 17, which typically also holds the SDY in male salmon (Figure 5). There are additionally, a small number of variants throughout the genome with moderate to high values of the negative log10 p-value (> 5).

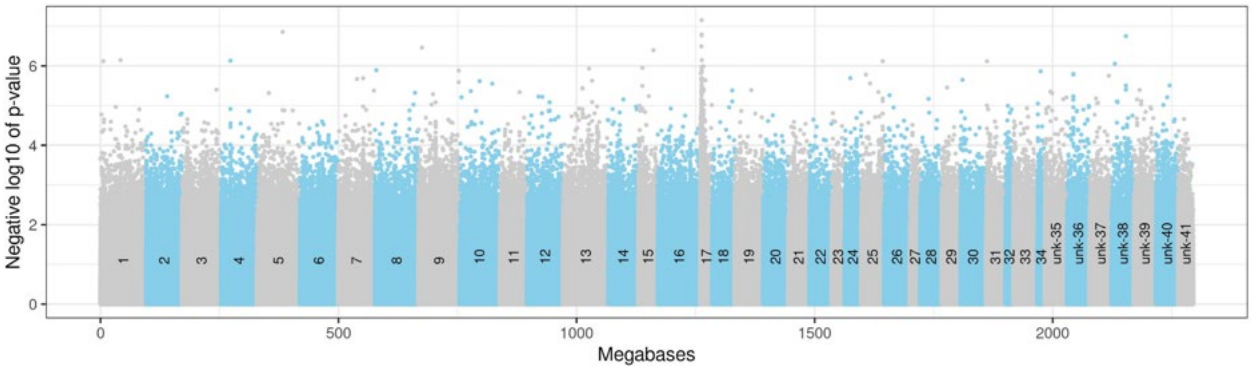

Figure 5. Manhattan plot of significance of association tests (negative log10 of p-value) for 4,679,605 variants across the genome. The values measure the significance of association between the variant and age at maturity in males. Color of points on adjacent chromosomes alternates gray and blue. Chromosome numbers are listed on top of each chromosome's band of points. Unplaced scaffolds are listed in genome coordinate order, but have been placed into several different groups labeled unk-XX.

The Manhattan plot for females does not show the same peak on Chromosome 17 found in males; however, there are quite a few variants that register negative log10 p-values greater than 6 (thus p-values < 0.000001) (Figure 6). There also appears to be a contiguous section of genome with elevated negative log10 p-values on Chromosome 26.

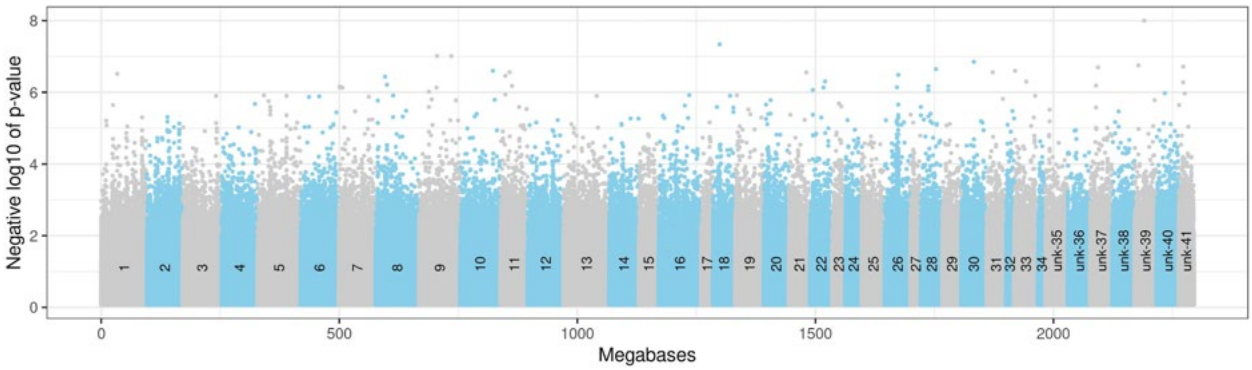

Figure 6: Manhattan plot of significance of association tests (negative log10 of p-value) for age at maturity at 3,972,970 variants across the genome in females.

From such an experiment, one would expect to find very high values of association at or near the region of the genome where the SDY resides, and nowhere else. Instead, while we found a preponderance of associated variants on Chromosome 17, which is the typical location in the genome at which to find the SDY, we also found a number of comparably high peaks of association in other regions of the genome—both on unplaced scaffolds as well as on the assembled chromosomes (Figure 7).

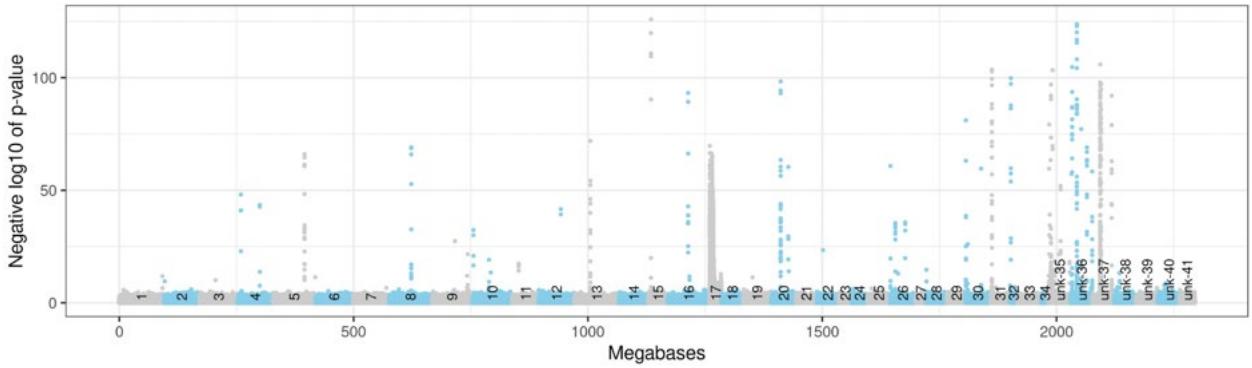

Figure 7. Manhattan plot of associations between genetic variation at 4,888,421 sites throughout the genome and presence or absence of SDY sequences in an individual.

The most parsimonious explanation for this observed pattern is that the reference genome has not been assembled accurately. The peaks outside of Chromosome 17 (and especially those that are higher than anything on Chromosome 17 are likely regions that include the SDY, or are adjacent to the SDY, but have not been properly assembled onto Chromosome 17. In the next sections, where we examine the most highly associated SNPs in both males and females, we will note those SNPs that are also on or near the peaks in the above figure (and which are, hence, likely to be truly on Chromosome 17). To prepare for this, we first noted that almost all adjacent SNPs on Chromosome 17 with negative log10 p-values > 6 for association with sex are within 5 Kb one one another. Accordingly, in the following sections, we look

for sites within 5 Kb of any of the sites throughout the genome that are associated with sex (at negative log10 p-values > 6) and flag them as possibly being within a section of Chromosome 17 genome that has been misassembled elsewhere in the genome.

After controlling for false discoveries, 31 sites were identified in females and 46 sites in males. The genes associated with these sites are listed in the google doc table below:

assoc-sites-and-nearby-genes

The column headings in that file are as follows:

- **sex**: the sex of the samples used in the association study
- **fdr**: the desired false discovery rate used
- **chromosome**: the number of the chromosome. Unplaced scaffolds listed as NA
- **ref_seq**: the RefSeq name for the chromosome.
- **position**: position of the site along the chromosome or scaffold
- **near_non_chr17_sex_assoc**: TRUE if within 5 Kb of a site in a peak of association for sex (indicating, possibly, a piece of Chromosome 17 that was incorrectly assembled into another chromosome or scaffold).
- **p_value**: the p-value of the association test at the site.
- **neg_log10_p**: -log10(p_value).
- **ref**: the base in the reference genome at this site.
- **alt**: the alternate allele at this site.
- **alt_freq**: the frequency of the alternate allele in all the samples at this site.
- **N**: the number of samples.
- **LRT**: the likelihood ratio test value.
- **beta**: the effect size.
- **SE**: the standard error of the effect size.
- **high_WT/HE/HO**: the number of homozygous reference (WT), heterozygous (HE), or homozygous alternate (HO) genotypes amongst the samples with genotype posterior probability exceeding 0.90.
- **emIter**: number of EM algorithm iterations to maximize the likelihood at this site.
- **nearby_named_genes**: a comma-separated list of named genes within 10 Kb of the site. Each gene is included in the following format: name|GeneID:XXXXXX|Chr:start-stop.

VI. DISCUSSION:

We sought to characterize the genomic basis of maturation age in Yukon Chinook salmon to better understand if size-selective fishing is causing evolution towards smaller size and younger maturation age. Our first major result is that external sex determination at Pilot Station is not reliable. We found that reported sex differed from genomic sex in about one third of cases. This result corroborates the growing recognition that analyses based on external sex determination of Yukon Chinook are unreliable (Bradley and Brown 2021).

Secondly, we found significant population structure in our samples that appears to be related to patterns of maturation, which we interpret to reflect stock-specific migration schedules. That our sample was composed of multiple genetic stocks was expected, as Pilot Station samples are known to be

from mixed stock sources (West and Prince, 2018). Although these signals are controlled for in the GWAS analysis, the genomic data generated in this study will be useful for future genomic research on Chinook in Alaska (particularly on the Yukon river) and the wider geographic region, so identifying which populations these fish originate from would be a priority. One potential solution is to cross-validate our samples with the ADFG stock identification markers to determine if it is possible to assign stock of origin. However this may not be feasible with the current Chinook baseline, but recently a new panel of dense SNP markers has been developed to resolve relationships between closely related Chinook populations in Alaska on the Kuskokwim river (McKinney et al., 2020).

Thirdly, we found that the genomic basis of maturation in Yukon Chinook differs between males and females. We found a strong association with maturation age in males on Chromosome 17 (the sex determining region or SDY), but several candidate loci with strong associations with maturation age in females. This result means that fisheries selection against late-maturation in males is decoupled from selection against late-maturation in females. A study by Micheleti and Narum (2018) using a pooled sequencing approach of male Chinook of age three, four and five and female Chinook of four and five sampled at a weir of Johnson Creek (Idaho) found differences in candidate regions in each sex with a candidate on chr 14 in males and chr 24 in females, but did not identify the large SDY region on chr 17 associated with age in maturity of males. Recent work by McKinney et al. (2020, 2021) using natural and hatchery fish from the Wenatchee River (Washington) also found a strong genomic association on SDY and male maturation age using GWAS of reduced representation sequencing (RAD-seq) data based on 40,180 SNPs. Our results are based on a whole genome sequencing approach, so we have interrogated a larger proportion of the genome (~200x more loci than previous studies) and found no other regions of strong association. The flip-side of this result is that we found several loci with strong associations with maturation age in female Chinook salmon, but no single region of the genome associated as strongly as in males. This indicates a likely polygenic basis of female maturation age and the first study to identify genomic regions associated with maturation age in female Chinook salmon. The more complex genomic basis of maturation age in females may mean that fisheries selection is less likely to result in an irreversible evolutionary response in females compared to males. This could be good news, as the evolutionary loss of larger females will have negative effects on fecundity and population productivity.

Future aims:

Most phenotypic traits have variation that is affected by multiple genomic regions, most of which have small effects. As such, large samples of individuals and in-depth genomic coverage are a necessary first step in elucidating their heritable basis. By generating this large amount of genomic data for Yukon Chinook salmon we now have a sufficiently large dataset on variation in different regions of the genome to accurately evaluate likely contributions to determining the age of maturation in Yukon River Chinook salmon. Unfortunately, a complete and accurate evaluation of the genomic basis of maturation age will require a better genome assembly than is currently available. We are investigating options for generation of a chromosome-level assembly reference genome for Chinook salmon to pursue this goal. However, this is just the start and basis of studies focused on understanding age and size declines in Yukon Chinook. Ultimately we would like to evaluate the consequences of fisheries-induced pressure in Chinook salmon using a time-series of samples. From our results, it is evident that this can be done immediately for males but not females. A potential strategy will be to develop use a genetic sex ID marker for Yukon Chinook salmon that can be used to genetically sex fish from scales, so that sex-specific patterns of size declines may be further studied. We anticipate that males and females likely responded differently to fishery-induced size selection, due to the differences in the underlying genomic basis of maturation age in males and females.

The second goal of this project was to develop amplicons for any genomic regions of strong association to look for changes through time that would indicate an evolutionary response to selection, as has been done for Atlantic salmon (Czorlich et al., 2018; Mobley et al., 2021). That the genomic basis for maturation differs between males and females, and there is no single locus of strong effect in females, means that this goal is more complicated than anticipated, however this could be done in males targeting variation specifically associated with different age classes.

Although female maturation age has a more complex genetic architecture than in males, there are still several genomic regions that could be evaluated further, particularly in long-term time series,, where the age of return is determined by scale analysis or parentage and for which population of origin information is available.

VII. REFERENCES:

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

Ayllon, F., Kjærner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., ... & Edvardsen, R. B. (2015). The vgll3 locus controls age at maturity in wild and domesticated Atlantic salmon (*Salmo salar* L.) males. PLoS Genetics, 11(11), e1005628.

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., ... & Kent, M. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature, 528(7582), 405.

Bigler, B. S., Welch, D. W., & Helle, J. H. (1996). A review of size trends among North Pacific salmon (*Oncorhynchus* spp.). Canadian Journal of Fisheries and Aquatic Sciences, 53(2), 455-465.
Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Bradley, C.A. & Brown, R.J., 2021. Development of a Simple Morphometric Model to Identify Sex in Chinook Salmon Returning to Spawn in the Yukon River. North American Journal of Fisheries Management, *41*(5), 1538-1548.

Brown, R. J., von Finster, A., Henszey, R. J. & Eiler, J. H. (2017). Catalog of Chinook salmon spawning areas in Yukon River Basin in Canada and United States. Journal of Fish and Wildlife Management 8, 558–586.

Bromaghin, J.F., Nielson, R.M. and Hard, J.J. (2011). A model of Chinook salmon population dynamics incorporating size-selective exploitation and inheritance of polygenic correlated traits. Natural Resource Modeling, 24(1),1-47.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources, 15(4), 855-867.

Christensen, K. A., Leong, J. S., Sakhrani, D., Biagi, C. A., Minkley, D. R., Withler, R. E., ... & Devlin, R. H.

(2018). Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. PloS one, 13(4), e0195461.

Conover, D. O., & Munch, S. B. (2002). Sustaining fisheries yields over evolutionary time scales. Science, *297*(5578), 94-96.

Czorlich, Y., Aykanat, T., Erkinaro, J., Orell, P., & Primmer, C. R. (2018). Rapid sex-specific evolution of age at maturity is shaped by genetic architecture in Atlantic salmon. *Nature Ecology & Evolution*, *2*(11), 1800-1807.

Darimont, C. T., Carlson, S. M., Kinnison, M. T., Paquet, P. C., Reimchen, T. E., & Wilmers, C. C. (2009). Human predators outpace other agents of trait change in the wild. Proceedings of the National Academy of Sciences, pnas-0809235106.

Francis, R. C., Hixon, M. A., Clarke, M. E., Murawski, S. A., & Ralston, S. (2007). Ten Commandments for ecosystem-based fisheries scientists. Fisheries, 32(5), 217-233.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907v2, 9.

Hankin, D. G., Nicholas, J. W., & Downey, T. W. (1993). Evidence for inheritance of age of maturity in Chinook salmon (*Oncorhynchus tshawytscha*). Canadian Journal of Fisheries and Aquatic Sciences, 50(2), 347-358.

Hankin, D. G., Fitzgibbons, J., & Chen, Y. (2009). Unnatural random mating policies select for younger age at maturity in hatchery Chinook salmon (*Oncorhynchus tshawytscha*) populations. Canadian Journal of Fisheries and Aquatic Sciences, 66(9), 1505-1521.

Hard, J. J. (2004). Evolution of Chinook salmon life history under size-selective harvest. Evolution illuminated: Salmon and their relatives, 315-337.

Hess, J.E., Zendt, J.S., Matala, A.R., & Narum, S.R. (2016). Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. Proceedings of the Royal Society B 283: 20153064.

Jeffrey, K. M., Côté, I. M., Irvine, J. R., & Reynolds, J. D. (2016). Changes in body size of Canadian Pacific salmon over six decades. Canadian Journal of Fisheries and Aquatic Sciences, 74(2), 191-201.
Jørsboe, E., & Albrechtsen, A. (2022). Efficient approaches for large-scale GWAS with genotype uncertainty. *G3, 12*(1), jkab385.

Kingsolver, J. G., Diamond, S. E., Siepielski, A. M., & Carlson, S. M. (2012). Synthetic analyses of phenotypic selection in natural populations: lessons, limitations and future directions. Evolutionary Ecology, 26(5), 1101-1118.

Korneliussen, T. S, Albrechtsen, A., & Nielsen, R. (2014)."ANGSD: analysis of next generation sequencing data." BMC Bioinformatics 15, no. 1: 356.

Lewis, B., Grant, W. S., Brenner, R. E., & Hamazaki, T. (2015). Changes in size and age of Chinook salmon *Oncorhynchus tshawytscha* returning to Alaska. PLoS One, 10(6), e0130184.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Lou, R. N., & Therkildsen, N. O. (2022). Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. Molecular Ecology Resources, 22(5), 1678-1692.

MacColl, A. D. (2011). The ecological causes of evolution. Trends in Ecology & Evolution, 26(10), 514-522.

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27(21), 2957-2963.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20:1297-303. DOI: 10.1101/gr.107524.110.

McKinney, G. J., Nichols, K. M., & Ford, M. J. (2021). A mobile sex-determining region, male-specific haplotypes and rearing environment influence age at maturity in Chinook salmon. Molecular Ecology, 30(1), 131-147.

McKinney, G. J., Seeb, J. E., Pascal, C. E., Schindler, D. E., Gilk-Baumer, S. E., & Seeb, L. W. (2020). Y-chromosome haplotypes are associated with variation in size and age at maturity in male Chinook salmon. Evolutionary Applications, 13(10), 2791-2806.

McKinney, G. J., Pascal, C. E., Templin, W. D., Gilk-Baumer, S. E., Dann, T. H., Seeb, L. W., & Seeb, J. E. (2020). Dense SNP panels resolve closely related Chinook salmon populations. Canadian Journal of Fisheries and Aquatic Sciences, 77(3), 451-461.

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. Genetics, 210(2), 719-731.

Micheletti, S. J., & Narum, S. R. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. Molecular Ecology Resources, 18, 825-837.

Micheletti, S. J., Hess, J. E., Zendt, J. S., & Narum, S. R. (2018). Selection at a genomic region of major effect is responsible for evolution of complex life histories in anadromous steelhead. BMC Evolutionary Biology, 18(1), 140.

Mobley, K. B., Aykanat, T., Czorlich, Y., House, A., Kurko, J., Miettinen, A., ... & Primmer, C. R. (2021). Maturation in Atlantic salmon (*Salmo salar*, Salmonidae): a synthesis of ecological, genetic, and molecular processes. Reviews in Fish Biology and Fisheries, *31*(3), 523-571.

Ohlberger, J., Ward, E. J., Schindler, D. E., & Lewis, B. (2018). Demographic changes in Chinook salmon across the Northeast Pacific Ocean. Fish and Fisheries, 19(3), 533-546.

Oke, K. B., Cunningham, C. J., Westley, P. A. H., Baskett, M. L., Carlson, S. M., Clark, J., ... & Palkovacs, E. P. (2020). Recent declines in salmon body size impact ecosystems and fisheries. Nature Communications, 11(1), 1-13.

Olsen, E. M., Heino, M., Lilly, G. R., Morgan, M. J., Brattey, J., Ernande, B., & Dieckmann, U. (2004). Maturation trends indicative of rapid evolution preceded the collapse of northern cod. Nature, 428(6986), 932.

Czorlich, Y., Aykanat, T., Erkinaro, J., Orell, P., & Primmer, C. R. (2018). Rapid sex-specific evolution of age at maturity is shaped by genetic architecture in Atlantic salmon. Nature Ecology & Evolution, *2*(11), 1800-1807.

Palumbi, S. R. (2001). Humans as the world's greatest evolutionary force. Science, 293, 1786-1790.

Pelletier, F., Clutton-Brock, T., Pemberton, J., Tuljapurkar, S., & Coulson, T. (2007). The evolutionary demography of ecological change: linking trait variation and population growth. Science, 315(5818), 1571-1574.

Ricker, W. E. (1981). Changes in the average size and average age of Pacific salmon. Canadian Journal of Fisheries and Aquatic Sciences, 38(12), 1636-1656.

Siepielski, A. M., DiBattista, J. D., & Carlson, S. M. (2009). It's about time: the temporal dynamics of phenotypic selection in the wild. Ecology Letters, 12(11), 1261-1276.

Thompson, N. F., Anderson, E. C., Clemento, A. J., Campbell, M. A., Pearse, D. E., Hearsey, J. W., Kinziger, A. P. &, Garza, J. C. (2020) A complex phenotype in salmon controlled by a simple change in migratory timing. Science, 370:609–613

West, F., & Prince, D. (2019). Genetic stock identification of Pilot Station Chinook salmon, 2018. *Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 3A19-09, Anchorage, Alaska*.

DELIVERABLES:

Deliverables from this project include this final report and the genomic sequence data and analysis code, which will be made publicly available upon publication.

IX. PROJECT DATA:

Genomic sequence data and analysis code will be made publicly available upon publication.

X. ACKNOWLEDGEMENTS:

Tyler Dann (ADFG), Sara Gilk-Baumer (ADFG), and Jennifer Hoey (UCSC) provided critical contributions to study design and data analysis. Cassie Columbus, Ellen Campbell and Elena Correa (all UCSC) provided assistance with laboratory work.

XI. PRESS RELEASE:

Provide a press release pertaining to the results of your project and potential applications of your results that would capture the interest of AYK salmon fishers, community residents, or fishery managers. Press release should not exceed 500 words.

XII. APPENDICES:

Use appendices to report supplementary information that illustrates, enlarges on, or otherwise supports the text, but which is not needed to directly support results and conclusions.

Attachment E - AYK SSI Title Page Template

Arctic-Yukon-Kuskokwim Sustainable Salmon Initiative
Project Final Product[1]

Genomics of maturation age in Yukon Chinook

by:

John Carlos Garza[2], **Eric C. Anderson**[2], **Kerry Reid**[3], **Peter Westley, Eric P. Palkovacs**[3]

[2] Institute of Marine Sciences, University of California Santa Cruz & NOAA Southwest Fisheries Science Center, 110 McAllister Way Santa Cruz CA 95060

[3] Department of Ecology and Evolutionary Biology, University of California Santa Cruz, 130 McAllister Way Santa Cruz CA 95060

[4] Department of Fisheries, University of Alaska, Fairbanks, 150 Koyukuk Drive Fairbanks, Alaska 99775

September 30, 2022

1 Final products of AYK Sustainable Salmon Initiative-sponsored research are made available to the Initiatives Partners and the public in the interest of rapid dissemination of information that may be useful in salmon management, research, or administration. Sponsorship of the project by the AYK SSI does not necessarily imply that the findings or conclusions are endorsed by the AYK SSI.