Improved Resolution of Chum Salmon Genetic Stock Identification

Annual Progress Report to the Pollock Conservation Cooperative Research Center

Reporting Period: 1 Sep 2017 - 31 Dec 2017

Prepared by:
Garrett McKinney
Postdoctoral Researcher

And

Megan V. McPhee (Principal Investigator)
Associate Professor, Fisheries

College of Fisheries and Ocean Sciences
University of Alaska Fairbanks
17101 Point Lena Loop Road
Juneau, AK 99801

Corresponding author:
mvmcphee@alaska.edu
(907) 796-5464

10 January 2017

The widespread genetic similarity of summer-run chum salmon across coastal western Alaska, from Kotzebue southward to Bristol Bay, has challenged genetic stock identification in the region. This challenge has become problematic as greater attention is focused on chum salmon stock management in the face of declining Chinook salmon resources in the region, coupled with periodically high levels of bycatch of both chum and Chinook salmon in the Bering Sea-Aleutian Islands pollock fishery. Recent technological advances allow more efficient genotyping of a greater number of individuals at a larger number of markers across the genome, providing hope that we can achieve better resolution of genetic differences among summer-run chum salmon in coastal western Alaska. In this project, we build on past genetic marker (SNP, or single nucleotide polymorphism) discovery efforts to design a panel of markers (~500-600 SNPs) that can be efficiently genotyped using the "Genotyping-by-Thousands" (GT-seq) protocol. The panel will be used to better define reporting groups for coastal western Alaska chum salmon stocks, allowing more refined stock composition analysis of salmon harvest and bycatch. Below, we provide a brief description of progress to date, attempting to minimize jargon. A more detailed report that includes technical description of methods and results is appended.

Project objectives

1) Develop, test, and optimize a GT-seq marker panel of 500 SNPs for western Alaska chum salmon

2) Provide the optimized panel and protocols to resource managers throughout Alaska and adjacent areas

Approach

We started with efforts from prior project (funded by the Coastal Impact Assistance Program) to conduct marker discovery using the "RAD-seq" method. This method allows for efficient detection of numerous SNPs across the genome in multiple individuals. In consultation among geneticists from ADFG, NOAA, UAF, and the University of Washington (UW), we targeted 48 individuals from each of 6 geographically distinct collections from the four most problematic regions in coastal western Alaska (Norton Sound through Bristol Bay; Table 1) for sequencing. This effort yielded over 135,000 variable sites (SNPs), which we are now in the process of winnowing down to a collection of the ~500-600 SNPs that yield the best information for distinguishing among the regions within coastal western Alaska.

We are using a combination of analytical methods to choose the best markers, including the traditional approaches of genetic distance among collections (F_{ST}) based on single SNPs and multivariate clustering approaches (PCA). However, these sequence data also allow us to identify "haplotypes", or short regions of the genome containing multiple SNPs, where the genotype consists of the allelic identify at each of these multiple SNPs. These haplotypes allow

greater resolution of genetic differences than just would summing up differences across multiple SNPs, and this method has been used by McKinney, Seeb, and Seeb to improve genetic stock identification resolution in Chinook salmon from western Alaska. We are also taking advantage of new marker selection methods based on machine-learning algorithms, which can improve upon F_{ST}-based approaches.

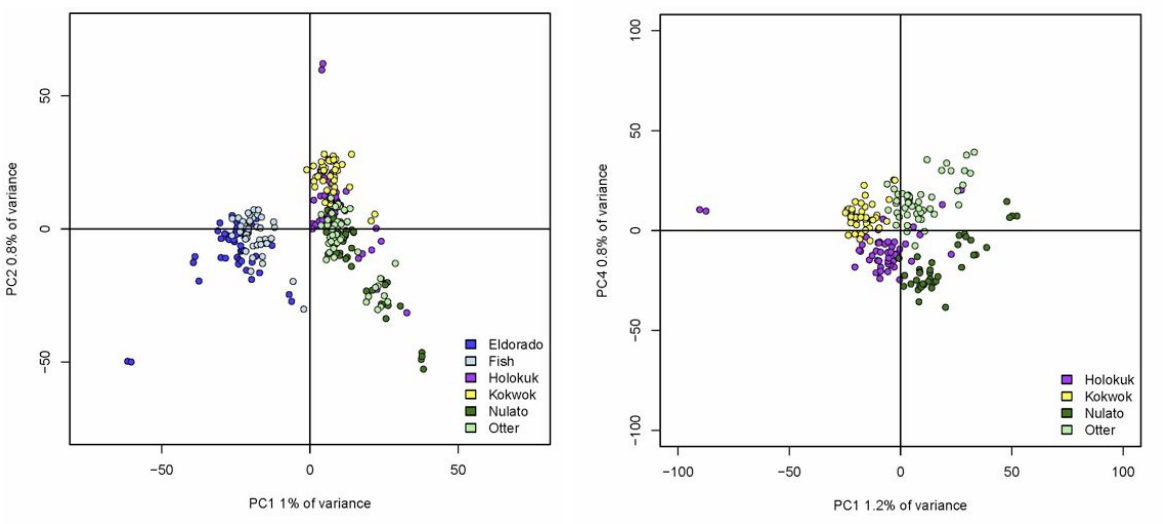
Table 1. Collections used for original SNP discovery by region (desired reporting group) including ADFG collection ID and number of individuals (N) for which quality RAD-seq data was obtained.

Region	Collection	ADFG Collection ID	N
Norton Sound	Eldorado River	CMELD05	48
	Fish River	CMFISH04	43
Lower Yukon	Nulato River	CMNUL03	47
	Otter Creek (Anvik River)	CMOTT93	48
Kuskokwim	Holokuk River	CMHOL08	48
Bristol Bay	Kokwok River	CMKOKW11	48

Results to date

Most of the work to date has been spent filtering the enormous amount of raw sequence data from the original round of SNP discovery. Using a number of different filtering criteria based on sequence data quality, duplicated versus non-duplicated status, and coarse measures of information content, a large number of SNPs (30,006) were found to be suitable for population genetics analysis.

Preliminary analyses of these 30,006 SNPs indicate promise for better resolution of western Alaskan chum salmon populations. A principal components analysis (PCA) on variation among all 268 individuals in the marker discovery panel shows clear separation of collections from Norton Sound from the other samples (Fig. 1a). When we repeated this analysis on only the Yukon, Kuskokwim, and Bristol Bay collections, we found weaker structure although separation among the regions, and even among collections within regions, was still apparent (Figure 1b).



(a) (b)

Figure 1. Results of principal components analysis (PCA) of allelic variation at 30,006 SNPs among individuals from (a) all six collections from coastal western Alaska; and (b) four collections representing the Lower Yukon, Kuskokwim, and Bristol Bay.

Preliminary analyses suggested that using haplotype data would improve GSI resolution over single-SNP data. Self-assignment rates exceeded 90% accuracy for Norton Sound using haplotype data, and approached that for Yukon and Kuskokwim/Bristol Bay reporting groups (85 - 88% accuracy). Interestingly, results thus far also suggested that F_{ST}-based marker selection outperformed machine learning-based selection, although this could change with the use of different parameters for the algorithm.

Next steps

We will continue to refine the parameters of the machine learning-based marker selection and compare to F_{ST}-based approaches, and then pick a marker-selection method to choose the final panel for optimization. The selected markers will then be tested with the GT-seq panel to identify a marker panel that 1) best distinguishes among the desired reporting groups for coastal western Alaska; and 2) performs well using the GT-seq genotyping protocol.

APPENDIX A - Detailed Progress Report

Introduction

Summer-run chum salmon stocks in western Alaska have been an ongoing challenge for applications of genetic stock identification (GSI) (Wilmot *et al.* 1994; Smith & Seeb 2008; Seeb *et al.* 2011b). There is widespread genetic similarity among summer-run chum salmon populations originating from Bristol Bay, north to Kotzebue Sound, resulting in a single reporting unit for genetic assignments in all of Coastal Western Alaska (e.g., Seeb *et al.* 2004; Decovich *et al.* 2012). This observed genetic similarity could reflect high levels of contemporary gene flow or, alternatively, recent common ancestry of these populations as a result of regional hydrological dynamism such as stream capture events and movement of the mouth of the Yukon River (Seeb & Crane 1999; McPhee *et al.* 2009; Olsen *et al.* 2011; Garvin *et al.* 2013).

Initial efforts to increase resolution to provide finer-scale reporting units for chum salmon focused on adding loci for panels of 100-200 single nucleotide polymorphisms (SNPs) (Seeb *et al.* 2011a; Decovich *et al.* 2012; Petrou *et al.* 2013; Garvin *et al.* 2016) using medium-density arrays and TaqMan assays (Seeb *et al.* 2009). Although some additional resolution was obtained with increasing numbers of SNPs in chum salmon (Jasper *et al.* 2013; Petrou *et al.* 2014), the Coastal Western Alaska group remained largely undifferentiated for GSI applications.

Recently, considerable progress has been achieved in GSI applications for Chinook salmon from Western Alaska (Larson *et al.* 2014a; Larson *et al.* 2014b) by sequencing restriction-site associated DNA (RADseq) to genotype large numbers of SNPs (>10,000) in representative populations; the resulting data were used to identify subsets of higher-resolution SNPs to use in screening multiple populations. Here we report a short-term focused effort using a similar approach of RADseq to genotype >10,000 DNA markers for development into a higher-resolution Genotyping-by-Thousands by sequencing panel (GT-seq, Campbell *et al.* 2015) for help in distinguishing populations of chum salmon from Coastal Western Alaska. This effort will take advantage of recent developments in GSI including the use of haplotype alleles (McKinney *et al.* 2017a) and random forest marker selection (Sylvester *et al.* 2017) to enhance panel accuracy.

Materials and Methods

SNP Discovery[1]

DNA aliquots for 288 individual fish from six geographically distinct collections (48 per collection) were obtained from Alaska Department of Fish and Game (ADFG; Table A1). These collections were selected in consultation by ADFG, NOAA, UAF, and UW to represent the range of populations contained in the Coastal Western Alaska reporting group.

RADseq libraries were prepared using a modification of the Rapture protocol (Best-RAD, Ali *et al.* 2016); all individuals were uniquely barcoded. Sequencing was conducted on three lanes of an Illumina HiSeq 4000 with 96 individuals per lane and 150 bp paired-end reads. An initial round of sequencing was conducted after which samples were demultiplexed using the *process_radtags* module from *Stacks* (Catchen *et al.* 2013) to obtain the number of sequence reads per individual. A round of resequencing was then conducted where the proportion of DNA for each individual was adjusted for a targeted yield of 4 million total sequence reads per individual. For both rounds of sequencing, default settings were used for *process_radtags* with the following exceptions: -c -q -r --filter_illumina -t 140.

SNP Filtering

Sequence reads for each individual were run through *Stacks* to obtain genotypes. Default settings were used for all *Stacks* modules with the following exceptions: *ustacks:* --model_type bounded --bound_high 0.01 -M 2, *cstacks* -n 2. The *Stacks* catalog of loci and genetic variation was created with six randomly chosen individuals from each population. After running *Stacks*, individuals and loci were filtered to obtain the final data set. An initial filter was used to remove loci with less than 50% genotype rate. Individuals with reduced genotype rate due to insufficient sequence depth were then removed. Following removal of low-quality individuals, data were then run through *Genepop* (Rousset 2008) to obtain allele frequency and F_{ST} estimates. A second filter was applied requiring loci to have a minor allele frequency of at least 5%. Retained loci were then run through *HDplot* (McKinney *et al.* 2017b) to identify paralogs. A 90% genotype rate filter was then applied to singleton (non-paralogous) loci to obtain a final high-quality locus set.

Paralogs (loci united in common ancestry via genome duplication, but now behaving as diploid loci) were excluded from analyses of population structure due to genotyping uncertainty. Allele frequencies are currently being estimated for paralogs using *Polyfreqs* (Blischak *et al.* 2016). Allele frequency estimates will be used to calculate F_{ST} to determine if any paralogs look

[1] work conducted under previous project, funded by Coastal Impact Assistance Program; PIs M. McPhee (UAF) and J. Guyon (NOAA); subcontracted to J. Seeb and L. Seeb, University of Washington (UW)

particularly informative for GSI. Informative paralogs will be included in the GSI panel if necessary to increase panel accuracy.

Population Genetics Analyses

Population structure was examined using individual-based Principal Component Analysis (PCA) performed with *Adegenet* (Jombart 2008). A threshold of F_{ST} >0 was applied to singleton loci for PCA input. Paralogs were not used for PCA due to uncertainty in genotyping. An initial PCA was conducted using all populations. Further PCAs were conducted using subsets of populations to identify patterns of finer-scale population differentiation.

Loci were aligned to the rainbow trout genome (v. Omyk_1.0) using bowtie2 (Langmead & Salzberg 2012) using default settings to determine if loci of interest concentrated in specific regions of the genome. Pairwise linkage disequilibrium (LD) was estimated for all loci using the r-squared method in plink (Purcell *et al.* 2007). Patterns of LD for each chromosome were visualized by plotting heatmaps in R.

GSI panels have traditionally examined a single SNP per locus and have used F_{ST} to choose informative markers for discriminating stocks. Loci with multiple SNPs are common in salmonid RADseq data and GT-seq allows these multiple SNPs to be combined into haplotypes. Haplotype data increased accuracy for discriminating closely related populations of Chinook salmon in Western Alaska (McKinney *et al.* 2017a) and is likely to show similar gains for chum salmon in this region. F_{ST} is commonly used to identify informative markers; however, random forest marker selection was recently shown to yield similar increases in accuracy for genetic stock identification as achieved using haplotype data (Sylvester *et al.* 2017). Combining haplotype data with random forest marker selection may yield gains in accuracy beyond either method independently.

We are currently testing F_{ST} and random forest marker selection with single-SNP and haplotype data to determine the optimal method for marker selection. For each method, the 500 loci with the highest ranking are evaluated using mixture analysis in GSI_sim (Anderson *et al.* 2008). A training set consisting of half the individuals from each population is used to choose markers and the accuracy of the selected panel is evaluated by assigning the remaining individuals from each population (holdout set) back to the baseline (all samples). This training-holdout approach minimizes bias when evaluating GSI accuracy (Anderson 2010). Reporting groups for mixture analysis consist of Norton Sound (Eldorado and Fish Rivers), Yukon (Nulato River and Otter Creek), and Kuskokwim/Nushagak (Holokok and Kokwok Rivers). Once the optimal method for marker selection is determined, loci for the final panel will be selected using genotype data from all samples.

Results

Over 900 million sequence reads were retained across all samples with the average number of reads per sample ranging from 1.8 million to 5.1 million across populations (Table A1), substantially above the minimum-required threshold of 1.5 million reads per individual used for successful genotyping in RADseq studies (for example Waples *et al.* 2016). A total of 282 individuals passed genotype rate filters. The initial filter of 50% genotype rate resulted in ~60,000 RAD tags and ~135,000 SNPs retained. Filtering for a 5% minor allele frequency reduced these numbers to ~36,000 RAD tags and ~55,000 SNPs (Table A2).

Thresholds for *HDplot* were set at $H \geq 0.55$ and $|D| \geq 10$ to differentiate duplicates from singleton loci (Figure 1). A threshold of $H \geq 0.77$ was used to identify diverged duplicates (duplicate loci inherited as independent disomic loci). Approximately 32,000 RAD tags and ~46,000 SNPs were classified as singletons, ~4,900 RAD tags and ~8,600 SNPs were classified as duplicates, and ~400 RAD tags and ~500 SNPs were classified as diverged duplicates (Table A2). The final filter of 90% genotype rate for singleton loci resulted in 22,693 RAD tags and 30,006 SNPs retained for population genetic analysis.

PCA of individuals from all collections revealed two distinct clusters, one cluster containing the two collections from Norton Sound and a second cluster containing collections from the Kuskokwim, Nushagak, and Yukon rivers (Figure A2a). Individuals from the Kuskokwim, Nushagak, and Yukon rivers were then examined to determine if there was evidence of differentiation among these collections. Individuals from each of these collections formed distinct clusters (Figure 2B); however, there was little separation between these clusters. A final PCA was conducted where collection pairs from each region (Norton Sound, Kuskokwim/Nushgak rivers, and Yukon River) were analyzed separately. Collections within each region formed distinct clusters when regions were examined separately (Figure A3).

Population substructure was apparent within the Otter Creek and Nulato River collections. Individuals were divided into quadrants based on substructure (Figure A4a); F_{ST} was then estimated for each locus, treating each quadrant as a population. The pattern of within-collection differentiation was primarily driven by 39 loci with near-perfect linkage disequilibrium (LD). These 39 loci are inherited as a single disomic locus; individuals are either homozygous or heterozygous for nearly all loci in this LD block. Discrepancies in this pattern primarily result from heterozygous individuals having a homozygous genotype. The identification of individuals that were largely heterozygous for the LD block, but homozygous at a single SNP within the block, most likely resulted from genotyping error but may represent true genotypes. The lines on Figure A4b show how the genotype patterns for the LD block correspond to the within-population structure.

Thirty-three of the 39 high-LD loci had alignments to the RBT genome and 23 of these aligned to a 30Mb block of Omy28. There is a gap of ~10Mb between the LD block and the last locus with high LD suggesting misalignment to the RBT genome or structural differences between species. The true span of the LD block is likely closer to 20Mb which represents half the total length of the current Omy28 assembly (41MB). The pattern of differentiation within populations is consistent with an inversion on the chum ortholog of Omy28. Re-examination of genotype patterns in the other populations revealed the presence of this inversion in all populations but at lower frequencies. Holokuk River has seven individuals that are heterozygous for the inversion and one individual that is homozygous (corresponds to the lower right samples in Nulato River). Eldorado River, Fish Creek, and Kokwok River each have two individuals that are heterozygous for the inversion.

Initial GSI results shows greater accuracy with haplotype data relative to single-SNP data and greater accuracy for F_{ST} based marker selection relative to random forest marker selection. The Norton Sound reporting group shows high accuracy (>90%) for haplotype data regardless of marker selection method. Yukon and Kuskokwim/Nushagak reporting groups show 85-88% accuracy with F_{ST} based marker selection but only 71-79% accuracy with random forest marker selection.

Discussion

This study took advantage of several recent developments in RADseq technology and methodology to identify informative markers for GSI. The best-RAD library preparation method incorporates a biotinylated barcode adapter which results in substantial increases in the proportion of retained sequence reads. We were also able to use the Ilumina HiSeq 4000 that was recently installed at the genomics core facility at the University of Oregon. The HiSeq 4000 instrument generates twice as many sequence reads as the previous HiSeq 2000 (400 vs. 200 million read output) for the same total cost. By taking advantage of these developments, we have approximately doubled the average sequence reads per individual relative to previous RADseq projects of the same scale. We also saw a substantial increase in the number of SNPs detected as a result of increased sequencing length in this project. Many SNP discovery projects using RADseq have sequenced to 100 bp length (Larson *et al.* 2014b; Benestan *et al.* 2015; Candy *et al.* 2015) whereas we increased sequencing length to 150 bp for this project. The distribution of SNPs is approximately uniform throughout the length of the RAD tags, so an increase in RAD tag length of 50% should yield a corresponding increase in the number of SNPs detected.

Paralogs were excluded from PCAs due to genotyping uncertainty. Applying diploid genotyping algorithms to paralogs results in systemic mis-genotyping of heterozygous individuals as homozygous (McKinney *et al.* 2018). This introduces errors in allele frequency estimates but also tends to homogenize allele frequencies across populations, reducing the observed F_{ST} relative to the true F_{ST}. Panels of informative loci for GSI in Atlantic salmon have been observed to be enriched for paralogs (Gilbey *et al.* 2016), suggesting that these loci may contain more information on average than singleton loci. Paralogs can be accurately genotyped with amplicon sequencing methods that are currently being used for GSI; however, greater read depth is necessary to genotype these loci and genotype uncertainty makes it difficult to assess their utility prior to panel construction (McKinney *et al.* 2018).

Principal component analysis revealed signals of population structure for chum salmon throughout Western Alaska (Figures A2, A3). The largest degree of separation occurred between populations from Norton Sound and populations from Kuskokwim, Nushagak, and Yukon rivers (Figure A2a). The clear separation between Norton Sound and the rest of the populations suggests that Coastal Western Alaska chum salmon can be split into at least two reporting groups for genetic stock identification. There was less separation between population clusters within the Kuskokwim, Nushagak, and Yukon rivers and separation of populations between regions showed no difference relative to the separation of populations within each region (Figure A2b). Panels that prioritize markers that differentiate specific regions may be able to further resolve these populations into reporting groups but the success of any finer scale subdivision is uncertain. Intriguingly, PCA of population pairs within each region showed distinct clustering of populations (Figure A3). This suggests that even if these populations cannot be distinguished in Bering Sea mixtures that include populations from other regions, individual assignment to population of origin may be possible within drainages. It is important to keep in mind that these signals of population structure were apparent with 30,000 SNPs; however, practical and technological considerations limit marker panels for management to hundreds of loci.

The presence of a putative inversion in chum salmon provides intriguing opportunities for further research. Inversions have been noted in other salmonid species, notably the Omy5 inversion associated with life history in steelhead/rainbow trout (Pearse *et al.* 2014). The inversion is present in all chum salmon populations examined in this study but is only at high frequency in the Yukon River populations and Holokuk River. The restricted regional distribution of the inversion may be a signal of adaptation but could also be the result of genetic drift. Further study into the distribution of the inversion along with tests of association with environmental variables or phenotypes may shed light on any evolutionary importance of this genomic region. The high degree of linkage disequilibrium within this region also means that the inversion type of an individual fish (homozygous for inversion, heterozygous, or homozygous for normal

chromosome) can be effectively determined with just a few markers. Markers to type the inversion can be included in the chum salmon panel to allow inversion typing for future studies.

Haplotype analysis outperformed single-SNP analysis for all marker panels tested. Results from McKinney *et al.* (2017a) suggested that restricting panels to only include markers with haplotype data yielded greater accuracy than panels that included a mix of single-SNP and haplotype data. This study showed mixed results; the accuracy of the Yukon reporting group increased when panels were restricted to markers with haplotype data but the accuracy of the Kuskokwim/Nushagak reporting group decreased. The final panel will include a mix of single-SNP and haplotype markers. F_{ST} marker selection outperformed random forest marker selection in all tests conducted so far. Possible explanations for this reduced performance include the inherent stochasticity in random forest and use of suboptimal parameters. A well-known characteristic of random forest classification is variance in parameter ranking when the same dataset is examined multiple times. One method to account for this is to take the intersect of the top N rankings after several runs of random forest which is the approach taken by Sylvester *et al.* (2017). Running random forest with inappropriate parameters can lead to biased or uninformative results. Our initial testing of random forest examined the results from single runs, we are now performing multiple runs of random forest to examine if the intersect of these runs results in a more powerful locus set. We are also examining parameter optimization methods for random forest to determine if performance can be increased through model tuning.

References

Ali OA, O'Rourke SM, Amish SJ*, et al.* (2016) RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics* **202**, 389-400.

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Mol Ecol Resour* **10**, 701-710.

Anderson EC, Waples RS, Kalinowski ST (2008) An improved method for predicting the accuracy of genetic stock identification. *Can J Fish Aquat Sci* **65**, 1475-1486.

Benestan L, Gosselin T, Perrier C*, et al.* (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Mol Ecol* **24**, 3299-3315.

Blischak PD, Kubatko LS, Wolfe AD (2016) Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Mol Ecol Resour* **16**, 742-754.

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Mol Ecol Resour* **15**, 855-867.

Candy JR, Campbell NR, Grinnell MH*, et al.* (2015) Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (*Thaleichthys pacificus*, Osmeridae), an anadromous Pacific smelt. *Mol Ecol Resour* **15**, 1421-1434.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**, 3124-3140.

Decovich NA, Dann TH, Olive SDR*, et al.* (2012) Chum Salmon Baseline for the Western Alaska Salmon Stock Identification Program. *Alaska Department of Fish and Game, Division of Commercial Fisheries* **Alaska Department of Fish and Game, Special Publication No. 12-26, 110p**.

Garvin MR, Kondzela CM, Martin PC*, et al.* (2013) Recent physical connections may explain weak genetic structure in western Alaskan chum salmon (*Oncorhynchus keta*) populations. *Ecology and Evolution* **3**, 2362-2377.

Garvin MR, Templin WD, Gharrett AJ*, et al.* (2016) Potentially adaptive mitochondrial haplotypes as a tool to identify divergent nuclear loci. *Methods in Ecology and Evolution*.

Gilbey J, Cauwelier E, Coulson MW*, et al.* (2016) Accuracy of assignment of Atlantic Salmon (*Salmo salar* L.) to rivers and regions in Scotland and northeast England based on single nucleotide polymorphism (SNP) markers. *PLoS One* **11**, e0164327.

Jasper JR, Habicht C, Moffitt S*, et al.* (2013) Source-sink estimates of genetic introgression show influence of hatchery strays on wild chum salmon populations in Prince William Sound, Alaska. *Plos One* **8**.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359.

Larson WA, Seeb JE, Pascal CE, Templin WD, Seeb LW (2014a) Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. *Canadian Journal of Fisheries and Aquatic Sciences* **71**, 698-708.

Larson WA, Seeb LW, Everett MV*, et al.* (2014b) Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evol Appl* **7**, 355-369.

McKinney GJ, Seeb JE, Seeb LW (2017a) Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification. *Can J Fish Aquat Sci* **74**, 429-434.

McKinney GJ, Waples RK, Pascal CE, Seeb LW, Seeb JS (2018) Resolving allele dosage in duplicated loci using genotyping by sequencing data: a path forward for population genetic analysis. *Mol Ecol Resour (Accepted)*.

McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017b) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol Ecol Resour* **17**, 656-669.

McPhee MV, Zimmerman MS, Beacham TD*, et al.* (2009) A hierarchical framework to identify influences on Pacific salmon population abundance and structure in the Arctic-Yukon-Kuskokwim Region. In: *Pacific salmon:ecology and management of western Alaska's populations* (eds. Krueger CC, Zimmerman CE), pp. 1177-1197. American Fisheries Society, Symposium 70, Bethesda, Maryland.

Olsen JB, Crane PA, Flannery BG*, et al.* (2011) Comparative landscape genetic analysis of three Pacific salmon species from subarctic North America. *Conservation Genetics* **12**, 223-241.

Pearse DE, Miller MR, Abadía-Cardoso A, Garza JC (2014) Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings of the Royal Society B: Biological Sciences* **281**.

Petrou EL, Hauser L, Waples RS*, et al.* (2013) Secondary contact and changes in coastal habitat availability influence the nonequilibrium population structure of a salmonid *(Oncorhynchus keta*). *Mol Ecol* **22**, 5848-5860.

Petrou EL, Seeb JE, Hauser L*, et al.* (2014) Fine-scale sampling reveals distinct isolation by distance patterns in chum salmon (*Oncorhynchus keta)* populations occupying a glacially dynamic environment. *Conservation Genetics* **15**, 229-243.

Purcell S, Neale B, Todd-Brown K*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575.

Rousset F (2008) Genepop'007: A complete re-implementation of the Genepop software for Windows and Linux. *Mol Ecol Resour* **8**, 103-106.

Seeb JE, Pascal CE, Grau ED*, et al.* (2011a) Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources* **11**, 335-348.

Seeb JE, Pascal CE, Ramakrishnan R, Seeb LW (2009) SNP genotyping by the 5'-nuclease reaction: advances in high throughput genotyping with non-model organisms. In: *Methods in Molecular Biology, Single Nucleotide Polymorphisms, 2d Edition* (ed. Komar A), pp. 277-292. Humana Press.

Seeb LW, Crane PA (1999) High genetic heterogeneity in chum salmon in Western Alaska, the contact zone between northern and southern lineages. *Transactions of the American Fisheries Society* **128**, 58-87.

Seeb LW, Crane PA, Kondzela CM*, et al.* (2004) Migration of Pacific Rim chum salmon on the high seas: Insights from genetic data. *Environmental Biology of Fishes* **69**, 21-36.

Seeb LW, Seeb JE, Habicht C, Farley EV, Utter FM (2011b) Single-nucleotide polymorphic genotypes reveal patterns of early juvenile migration of sockeye salmon in the Eastern Bering Sea. *Transactions of the American Fisheries Society* **140**, 734-748.

Smith CT, Seeb LW (2008) Number of alleles as a predictor of the relative assignment accuracy of short tandem repeat (STR) and single-nucleotide-polymorphism (SNP) baselines for chum salmon. *Transactions of the American Fisheries Society* **137**, 751-762.

Sylvester EVA, Bentzen P, Bradbury IR*, et al.* (2017) Applications of random forest feature selection for fine-scale genetic population assignment. *Evolutionary Applications*.

Waples RK, Seeb LW, Seeb JE (2016) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Mol Ecol Resour* **16**, 17-28.

Wilmot RL, Everett RJ, Spearman WJ*, et al.* (1994) Genetic Stock Structure of Western Alaska Chum Salmon and a Comparison with Russian Far East Stocks. *Canadian Journal of Fisheries and Aquatic Sciences* **51**, 84-94.

Appendix A Tables & Figures

Table A1. Summary information for reads retained from each collection after sequence filtering.

Region	Location	Collection ID	N	Total Reads	Average Reads per Sample
Norton Sound	Eldorado River	CMELDO05	48	246,442,515	5,124,219
Norton Sound	Fish River	CMFISH04	43	90,528,515	1,886,011
Yukon River	Nulato River	CMNUL03	47	88,415,415	1,841,988
Yukon River	Otter Creek (Anvik)	CMOTT93	48	222,085,276	4,626,777
Kuskokwim River	Holokuk River	CMHOL08	48	148,918,235	3,102,463
Bristol Bay	Kokwok River	CMKOKW11	48	178,158,672	3,711,639

Table A2. Number of RAD tags and SNPs retained after initial filtering for genotype rate, minor allele frequency, and paralog status.

	Filtering Criteria		Paralog Status		
	50% genotype rate	5% MAF	Singletons	Duplicate	Diverged Duplicate
RAD tags	60,541	36,118	31,919	4,931	359
SNPs	135,822	54,842	45,639	8,606	540

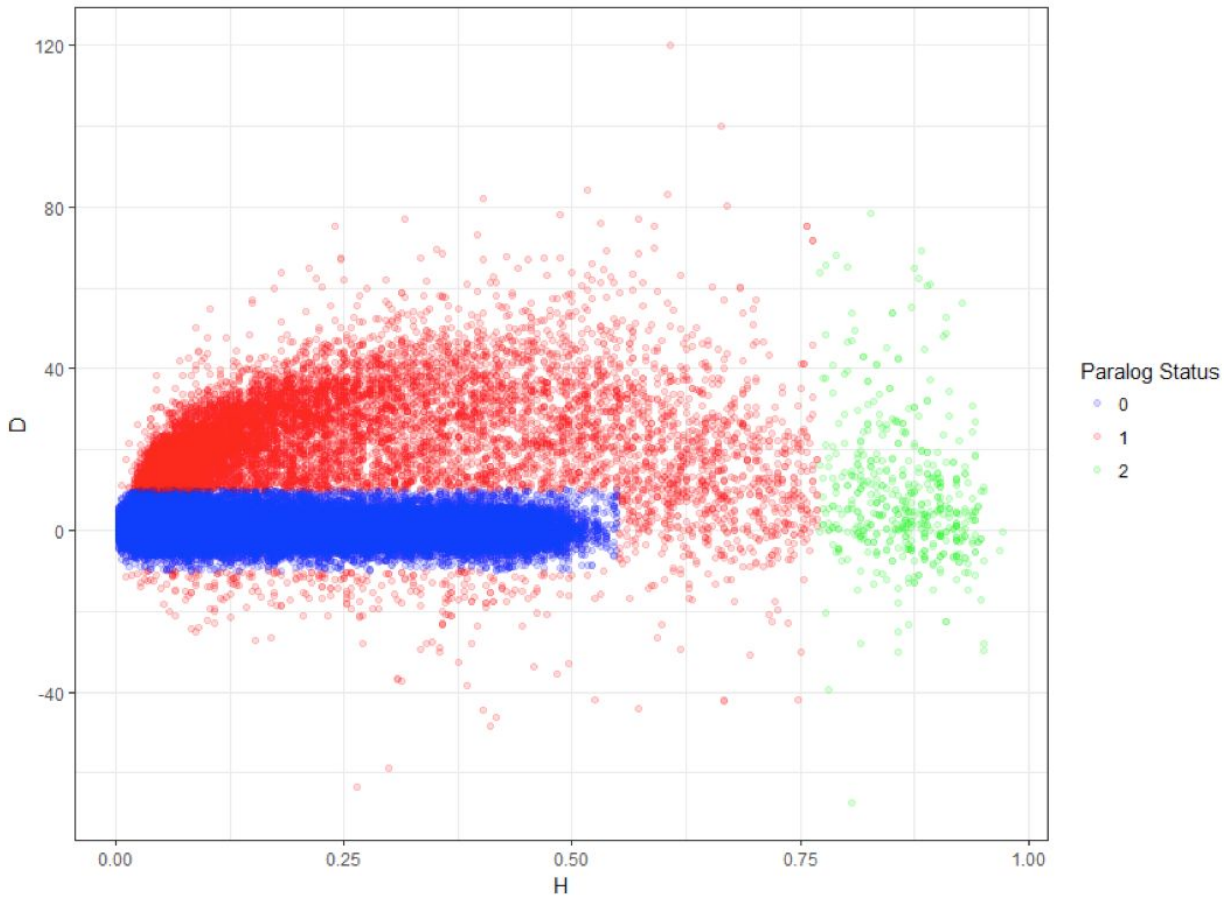
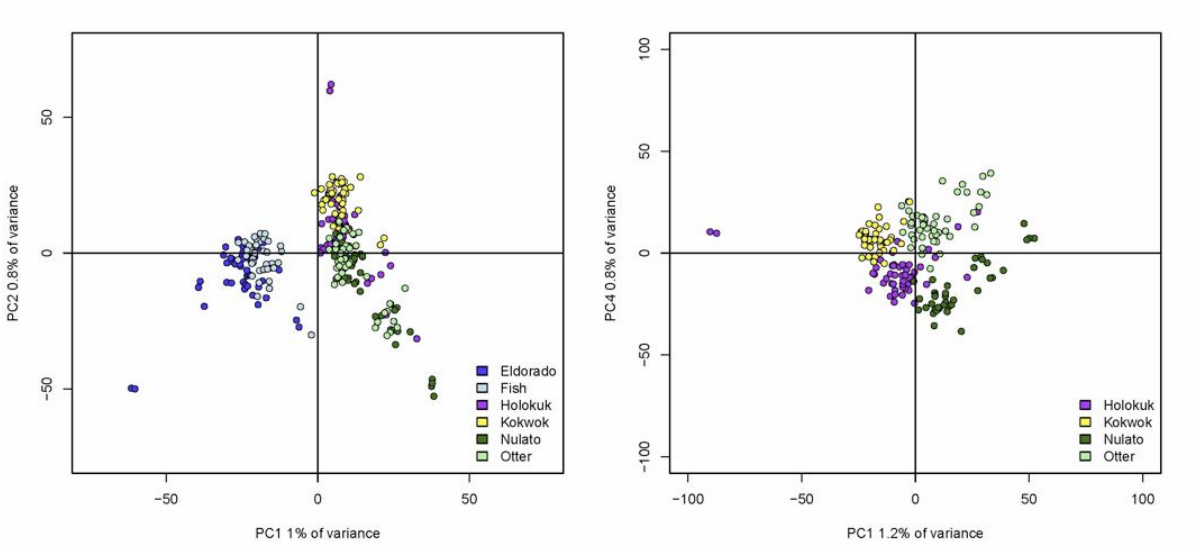


Figure A1. Duplicated loci. Results from HDplot showing distribution of *H* and *D* with each locus represented by a dot. Singleton loci (green, Paralog Status 0) are concentrated in the dense cloud of loci with *H* up to ~0.55 and D centered around 0; duplicate loci (red, Paralog Status 1) form a ring around this cloud, exhibiting greater *H* and/or *D* than singleton loci. Diverged duplicate loci (green, Paralog Status 2) are concentrated in the cloud with $H > 0.78$.



(a) (b)

Figure A2. Individual-based principal component analysis based on 30,006 SNPs for (a) for all populations and (b) populations outside Norton Sound. The primary separation is between populations from Norton Sound and all other populations (A); based on the separation seen here it is likely that Coastal Western Alaska chum salmon populations can be split into a minimum of two reporting groups. There is less differentiation between populations outside of Norton Sound; however, these populations show somewhat distinct clusters when analyzed on their own (B). The clustering of individuals within populations raises the possibility of discriminating more reporting groups, but the differentiation between these groups may be weak.

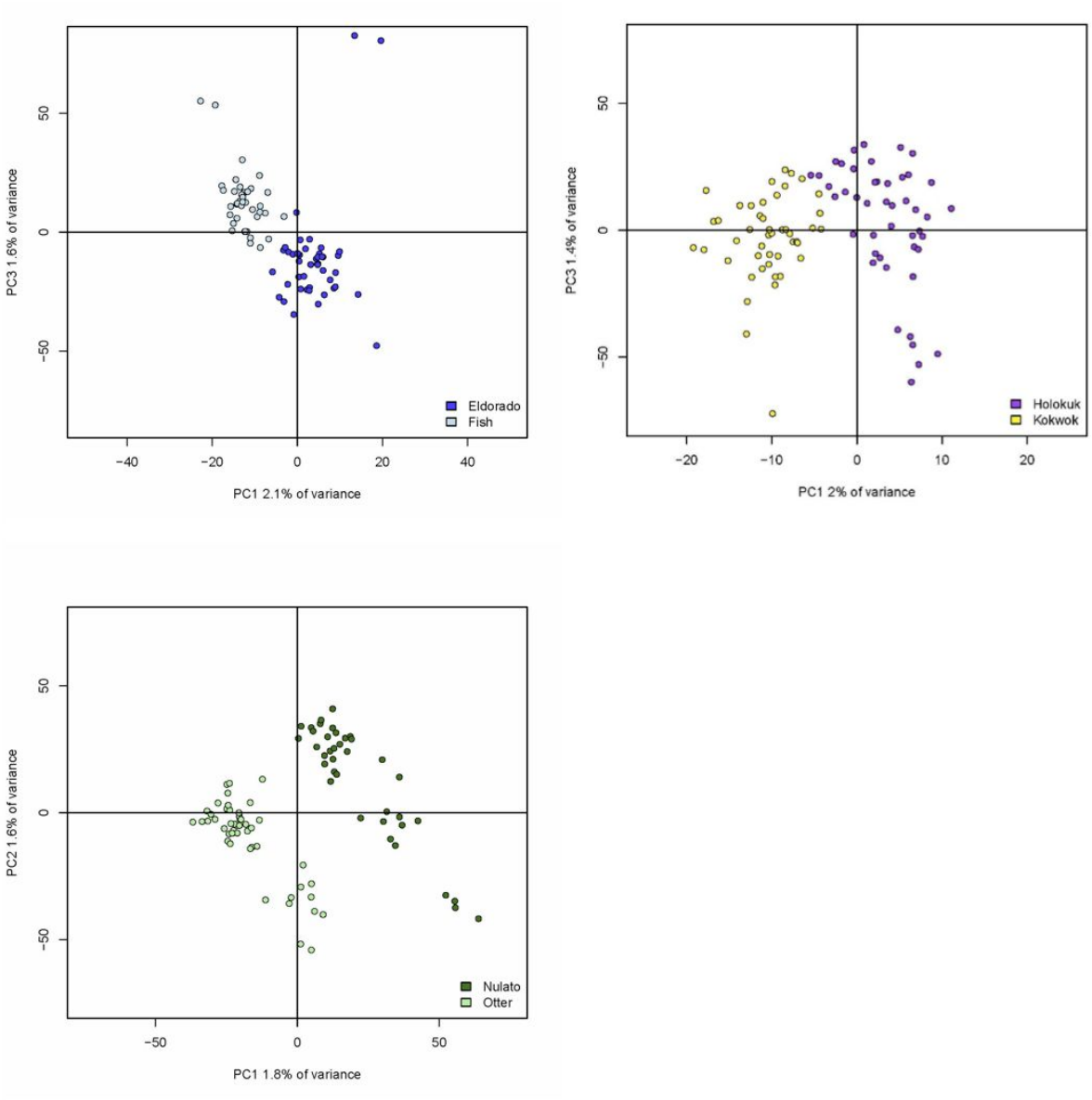
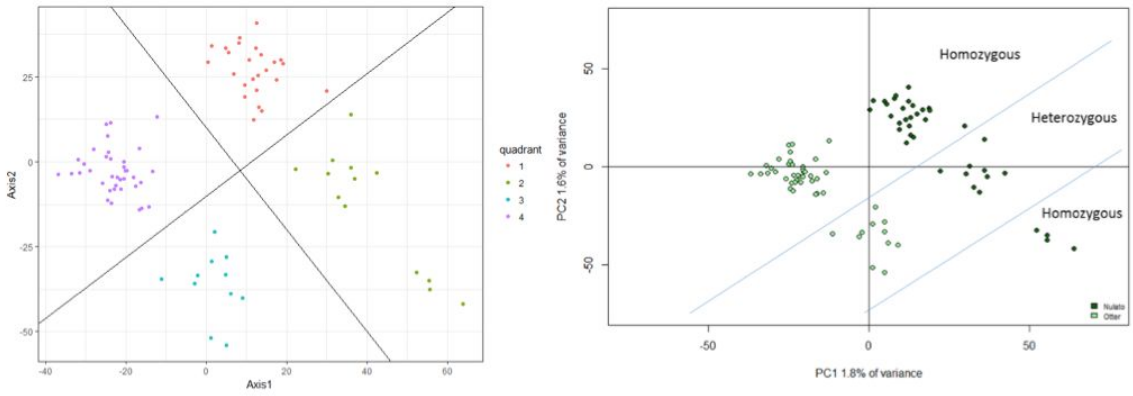


Figure A3. Individual-based PCA based on 30,006 SNPs with population pairs analyzed within each region: A) Norton Sound (Eldorado River and Fish River), B) Kuskokwim/Bristol Bay (Holokuk River and KokWok River), and C) Yukon River (Nulato River, Otter Creek). Within each region, individuals formed distinct clusters by population. These results suggest that panels of SNPs will be able to discriminate populations in regional collections.



(a) (b)

Figure A4. Individual PCA for Nulato River and Otter Creek chum salmon: A) quadrants used for F_{ST} examination, B) Genotype pattern overlayed onto PCA of Nulato River and Otter Creek chum salmon. Substructure within these populations corresponds with the individual genotypes for the Omy28 inversion.

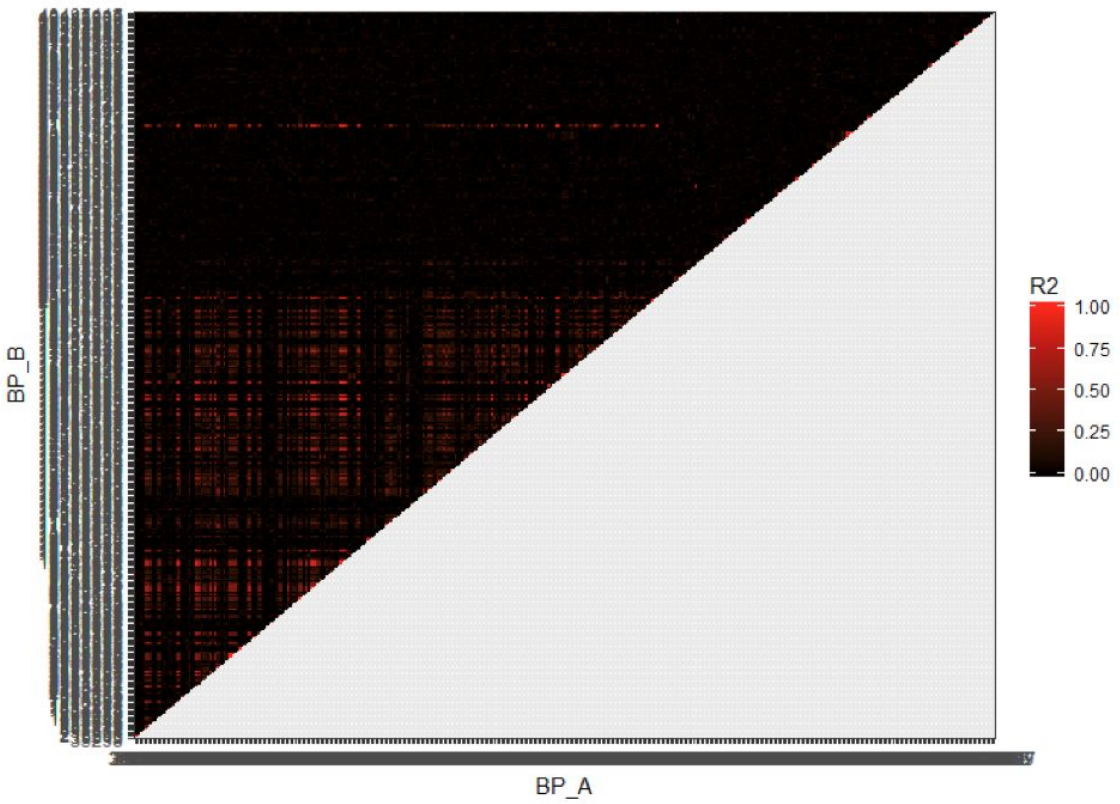


Figure A5. Pattern of linkage disequilibrium (LD) for loci aligned to chromosome Omy28. Loci with high LD ($R^2 \sim 1$) span approximately half of the chromosome. Genotype patterns within Nulato River and Otter creek suggest that this region of the chromosome is inherited as a single disomic block. This pattern of inheritance is consistent with an inversion polymorphism within these populations.