Improved Resolution of Chum Salmon Genetic Stock Identification

Final Report to the Pollock Conservation Cooperative Research Center

Prepared by:

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

And
Garrett McKinney
Postdoctoral Researcher

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

4 January 2019

Executive Summary

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 acute 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 populations across coastal western Alaska. In this project, we built on past genetic marker (SNP, or single nucleotide polymorphism) discovery efforts to design a 'GT-seq' panel, which is comprised of 503 SNP markers that can be genotyped efficiently using the 'Genotyping-in-Thousands' genotyping-by-sequencing protocol. This SNP panel still did not achieve the genetic resolution to break coastal western Alaskan summer chum salmon into the long-desired four regional reporting groups (Norton Sound, Lower Yukon, Kuskoskim, and Bristol Bay). However, with this panel we could erect two new reporting groups within coastal western Alaska with > 90% accuracy: Norton Sound and Lower Yukon/Kuskokwim/Bristol Bay. The GT-seq panel of 448 SNPs was selected from 30,006 high-quality SNPs that were filtered from 222,668 putative SNPs in 94,002 restriction-site-associated tags across the chum salmon genome, suggesting that these two reporting groups have approached the maximum stock resolution for coastal western Alaska that can realistically be achieved with genetic methods alone.

Background

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.* 2011a). 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.* 2011b; 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 (CWAK) group remained largely undifferentiated for GSI applications.

Considerable progress has been achieved in GSI applications for Chinook salmon from western Alaska (Larson *et al.* 2014a,b) by sequencing restriction-site associated DNA (RAD-seq) 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 the results of a focused effort using a similar approach of RAD-seq to genotype >30,000 DNA markers for development into a higher-resolution Genotyping-in-Thousands by sequencing panel (GT-seq; Campbell *et al.* 2015) for better distinguishing among populations of chum salmon from CWAK.

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

Methods & Results[1]

We started with efforts from prior project (funded by the Coastal Impact Assistance Program; PI McPhee) to conduct marker discovery using the "RAD-seq" (restriction-site-associated DNA sequencing) 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 >200,000 variable sites (SNPs), which formed the starting point for the development of the GT-seq panel.

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

SNP filtering

A number of filtering steps were applied in order to arrive at a panel of approximately 500 SNPs that would best distinguish among the stocks within CWAK and that could be genotyped using the GT-seq protocol. We started with 222,668 SNPs in 94,002 RAD tags (short pieces of DNA associated with the specific cutting site of restriction enzyme *SbfI)*. Loci that were successfully genotyped in < 50% of the samples were excluded, as were loci with a minor allele frequency <0.05 (SNPs typically have 2 different variants, or 'alleles'). 'Paralogs' (when alleles appear to be from the same locus but are not inherited as such, due to the remaining effects of an ancient

[1] This report was prepared for a non-specialist audience; more detailed methods and results are contained in a manuscript currently being prepared for submission to a peer-reviewed journal. A draft version of this manuscript is available upon request.

whole-genome duplication in the family Salmonidae) were identified using *HDplot* (McKinney et al. 2017). These paralogs could not be genotyped accurately and were also excluded from subsequent analysis. After these filters were applied, we selected from the remaining loci (45,639 SNPs in 31,919 RAD tags) those that were successfully genotyped in 90% of the samples; these loci (30,006 SNPs in 22,693 RAD tags) were then evaluated for information content related to ability to distinguish among the CWAK collections, as described below.

SNP selection based on information content

We calculated F_{ST} for the remaining 30,006 SNPs; this value represents the degree of genetic divergence among collections, theoretically ranging from 0 (complete lack of genetic differences) to 1 (completely different alleles in different collections). Genetic population structure among the collections within CWAK was visualized over the ~30,000 SNPs with principal component analysis (PCA) in the R package Adegenet (Jombart 2008). This plot showed that the two collections from Norton Sound (Eldorado and Fish rivers) were distinct from the remaining collections to the south (Figure 1).

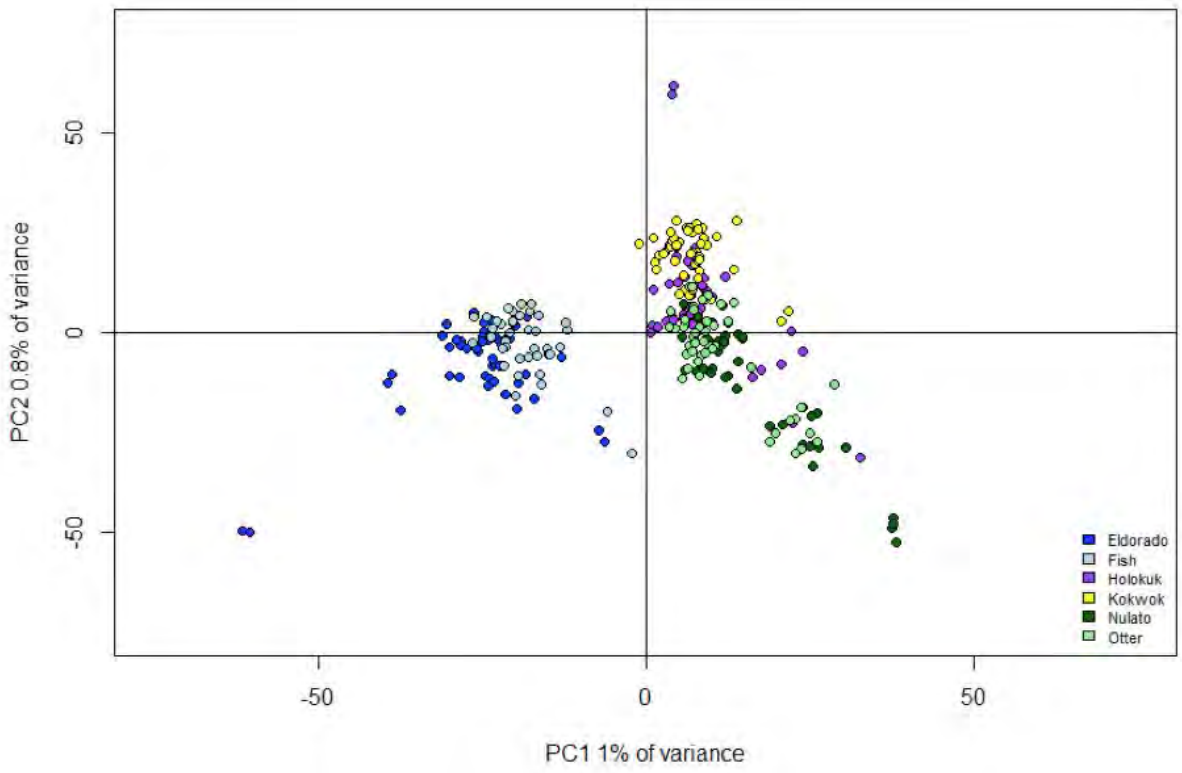

Figure 1. First two principal components of genetic variation among the six collections for all analyzed SNPs (each point represents an individual chum salmon).

We then performed mixture analyses in *GSIsim* (Anderson et al. 2008, Anderson 2010) using the 100% simulation method of Seeb et al. (2000). We simulated individuals from the allele frequencies of the six collections across CWAK and evaluated the accuracy at which these individuals could be allocated back to specific collections or region (Norton Sound, Lower Yukon, Kuskokwim, and Bristol Bay; see Table 1). For regional assessments, collections within regions contributed equally to the simulated mixtures. Because we had a single collection each for Kuskokwim and Bristol Bay, these were combined into a single region for the regional mixture analyses.

To evaluate the maximum resolution achievable with this dataset, we first performed the mixture analyses on simulated genotypes at all 30,006 SNPs. In order the assess the resolution that could be achieved with a marker panel that is practical to genotype, we then conducted the mixture analyses on sets of simulated genotypes at 500 SNPs. We evaluated three different 500-SNP panels that differed slightly based on method of marker selection: 1) highest individual F_{ST} values; 2) highest F_{ST} values including single and linked SNPs; and 3) random forest marker selection of Sylvester et al. (2018). Samples used for marker evaluation were not used in the generation of simulated mixtures used to evaluate those marker selection methods in order to minimize upward bias of performance estimates.

When all 30,006 SNPs were used, accuracy of allocation back to individual collections was >98%. However, these estimates are unrealistically high for two reasons. First, it is not yet feasible to efficiently genotype fish at ~30K SNPs for genetic stock analysis. Second, the accuracy of allocation to specific collections (i.e., spawning populations) is biased high in these kinds of 100% simulation evaluations; real-world applications will show lower accuracy as individual fish in mixtures originate from multiple spawning populations within a region. Accuracy of the 500-SNP panels in 100% simulations at the regional level are given in Table 2; these results indicated the marker selection based on single SNP F_{ST} values yielded the most accurate panel across the 3 regions.

Table 2. Accuracy (with 95% confidence intervals) with which a simulated mixture from a single region was allocated back to that region, by three different marker-selection methods.

500-SNP panel	Norton Sound	Lower Yukon	Kuskokwim & Bristol Bay
Individual SNP F_{ST}	0.90 (0.87-0.94)	0.92 (0.89-0.95)	0.83 (0.78-0.88)
Linked-SNP F_{ST}	0.85 (0.82-0.89)	0.92 (0.89-0.95)	0.83 (0.78-0.88)
Random forest marker selection	0.92 (0.89-0.95)	0.82 (0.78-0.87)	0.68 (0.63-0.74)

GT-seq panel development

We initially selected 700 SNPs for the initial round of primer development, ultimately aiming for a GT-seq panel of ~ 500 SNPs. While these SNPs were chosen primarily by individual F_{ST} (based on results of mixture analyses reported above), we eliminated SNPs that would be problematic for genotyping (e.g., likely transposable elements or loci with primers that aligned to multiple places in the chum salmon genome). 503 of the SNPs developed with RAD-seq passed the initial filters. We also included 31 of the SNPs, ranked by F_{ST}, currently in use by ADF&G in their chum salmon baseline (Seeb et al. 2011b, DeCovich et al. 2012). 533 resulting SNPs were then assessed in two rounds of panel optimization conducted on 48 individual chum salmon from CWAK. SNPs that did not perform well during these optimization rounds were removed, resulting in a final panel of 448 SNPs (which included 29 loci currently used by ADF&G).

GT-seq panel evaluation

The previously described mixture analyses are likely to be overly optimistic, given that limited individuals and collections were included in the evaluation and based on the same collections used for SNP discovery. We therefore expanded the sample set with which to more thoroughly evaluate the performance of the GT-seq panel, by adding more individuals to the collections previously used in SNP discovery and by adding more collections per region (Table 3, Figure 2). We genotyped 871 samples at 448 SNPs; after removing poor genotypes or duplicate individuals, we had genotypes for 798 individuals from 10 collections with which to evaluate the panel.

Table 3. Number of individuals per collection and region used in evaluation of GT-seq panel.

Region	Collection	Number of individuals
Norton Sound	Eldorado River	82
	Fish River	77
	Kwiniuk River	74
Lower Yukon	Nulato River	80
	Otter Creek	92
	E. Fork Andreafsky River	83
Kuskokwim	Holokuk River	79
	Aniak River	79
Bristol Bay	Kokwok River	79
	Mulchatna River	73

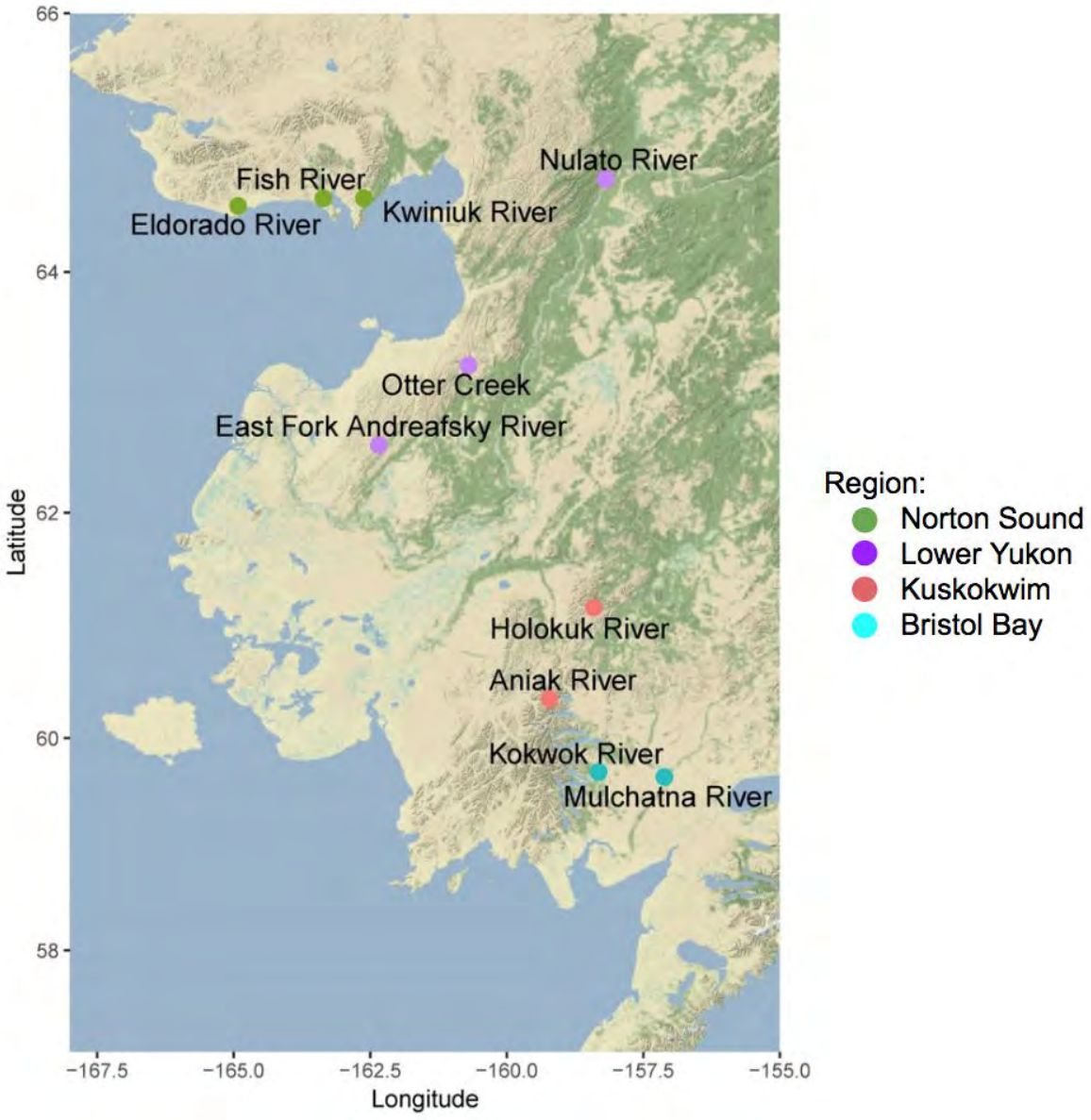

Figure 2. Collections by region used in evaluation of the GT-seq panel.

We assessed the performance of the GT-seq panel, using the 100% simulation and *GSIsim* methods described previously, with three iterative scenarios for reporting groups (i.e., the stock groupings desired by management for genetic stock identification). First, we evaluated each region as an individual reporting group (i.e., Norton Sound, Lower Yukon, Kuskokwim, and Bristol Bay). This scenario yielded accuracies < 90% for all groups but Norton Sound (Table 4), indicating that the GT-seq panel does not achieve the number of reporting groups and accuracy (>0.90) initially set out as the goal of the WASSIP project.

Table 4. Proportion (with 95% confidence intervals) of each collection allocated under the four regional reporting groups scenario. Correct allocations are shaded to facilitate easier viewing.

Collection	Norton Sound	Lower Yukon	Kuskokwim	Bristol Bay
Eldorado River	0.98 (0.96 - 1.0)	0.01 (0-0.02)	0.01 (0-0.02)	0 (0-0.01)
Fish River	0.97 (0.94-0.99)	0.02 (0-0.04)	0.01 (0-0.03)	0 (0-0.01)
Kwiniuk River	0.92 (0.89-0.96)	0.04 (0.01-0.07)	0.03 (0-0.05)	0.01 (0-0.03)
Nulato River	0 (0-0.01)	0.84 (0.80-0.89)	0.15 (0.10-0.19)	0 (-0.001-0.002)
Otter Creek	0 (0-0.01)	0.80 (0.75-0.86)	0.18 (0.12-0.23)	0.02 (0-0.04)
E. F. Andreafsky	0.01 (0-0.02)	0.63 (0.57-0.68)	0.29 (0.23-0.35)	0.07 (0.03-0.11)
Holokuk River	0 (0-0.01)	0.36 (0.29-0.42)	0.40 (0.33-0.46)	0.24 (0.18-0.31)
Aniak River	0.01 (0-0.02)	0.29 (0.23-0.36)	0.58 (0.51-0.66)	0.12 (0.07-0.17)
Kokwok River	0 (0-0.01)	0.07 (0.03-0.11)	0.20 (0.15-0.26)	0.72 (0.67-0.78)
Mulchatna River	0 (0-0.01)	0.07 (0.03-0.11)	0.30 (0.23-0.37)	0.63 (0.55-0.70)

We next evaluated three reporting groups, keeping Norton Sound and Lower Yukon as separate groups but combining Kuskokwim and Bristol Bay into a single group. This yielded slightly better results, but there was still considerable misallocation in collections from the Lower Yukon and Kuskokwim regions (Table 5). Finally, we were able to achieve > 90% accuracy when coastal western Alaska was divided into two reporting groups: Norton Sound and Lower Yukon/Kuskokwim/Bristol Bay (Table 6).

Table 5. Proportion (with 95% confidence intervals) of each collection allocated under the three reporting groups scenario (Kuskokwim and Bristol Bay combined). Correct allocations are shaded to facilitate easier viewing.

Collection	Norton Sound	Lower Yukon	Kuskokwim/Bristol Bay
Eldorado River	0.98 (0.96 - 1.0)	0.01 (0-0.02)	0.01 (0-0.02)
Fish River	0.97 (0.94-0.99)	0.02 (0-0.04)	0.02 (0-0.03)
Kwiniuk River	0.92 (0.89-0.96)	0.04 (0.01-0.07)	0.04 (0.01-0.07)
Nulato River	0.01 (0-0.01)	0.85 (0.79-0.91)	0.15 (0.09-0.20)
Otter Creek	0 (0-0.01)	0.80 (0.75-0.86)	0.19 (0.14-0.24)
E. F. Andreafsky R.	0.01 (0-0.02)	0.63 (0.56-0.70)	0.36 (0.29-0.43)
Holokuk River	0 (0-0.01)	0.35 (0.29-0.42)	0.64 (0.58-0.71)
Aniak River	0.01 (0-0.02)	0.29 (0.22-0.36)	0.70 (0.63-0.77)
Kokwok River	0 (0-0.01)	0.07 (0.03-0.11)	0.92 (0.88-0.96)
Mulchatna River	0 (0-0.01)	0.07 (0.03-0.11)	0.93 (0.89-0.97)

Table 6. Proportion (with 95% confidence intervals) of each collection allocated under the two reporting groups scenario (Norton Sound and Lower Yukon/Kuskokwim/Bristol Bay). Correct allocations are shaded to facilitate easier viewing.

Collection	Norton Sound	Lower Yukon/Kuskokwim/Bristol Bay
Eldorado River	0.98 (0.96 - 1.0)	0.02 (0-0.04)
Fish River	0.96 (0.94-0.99)	0.04 (0.01-0.06)
Kwiniuk River	0.92 (0.88-0.96)	0.08 (0.04-0.12)
Nulato River	0.01 (0-0.01)	0.99 (0.99-1.00)
Otter Creek	0 (0-0.01)	1.00 (0.99-1.00)
E. F. Andreafsky River	0.01 (0-0.02)	0.99 (0.98-1.00)
Holokuk River	0 (0-0.01)	1.00 (0.99-1.00)
Aniak River	0.01 (0-0.02)	0.99 (0.98-1.00)
Kokwok River	0 (0-0.01)	1.00 (0.99-1.00)
Mulchatna River	0 (0-0.01)	1.00 (0.99-1.00)

Additional insights into the chum salmon genome

Some of the SNP loci showed an unexpected pattern of three clusters in the PCA regardless of which collection an individual came from. Further examination of these loci revealed a pattern consistent with some individuals carrying an inversion variant (i.e., one segment of the DNA 'flipped' within the chromosome) in linkage group Omy28. Additional structure within the clusters was consistent with the inversion being associated with the sex-determining region (Figure 3); the limited data we had on phenotypic sex (male vs. female) of genotyped individuals supported this inference, but additional data would be needed to confirm. Further interpretation of these results was beyond the scope of this project, but greater detail is available in the manuscript in preparation. Loci that type the inversion and the putative sex linkage are included in the GT-seq panel, which enhances the value of the panel for applications beyond genetic stock identification, including parentage-based tagging.

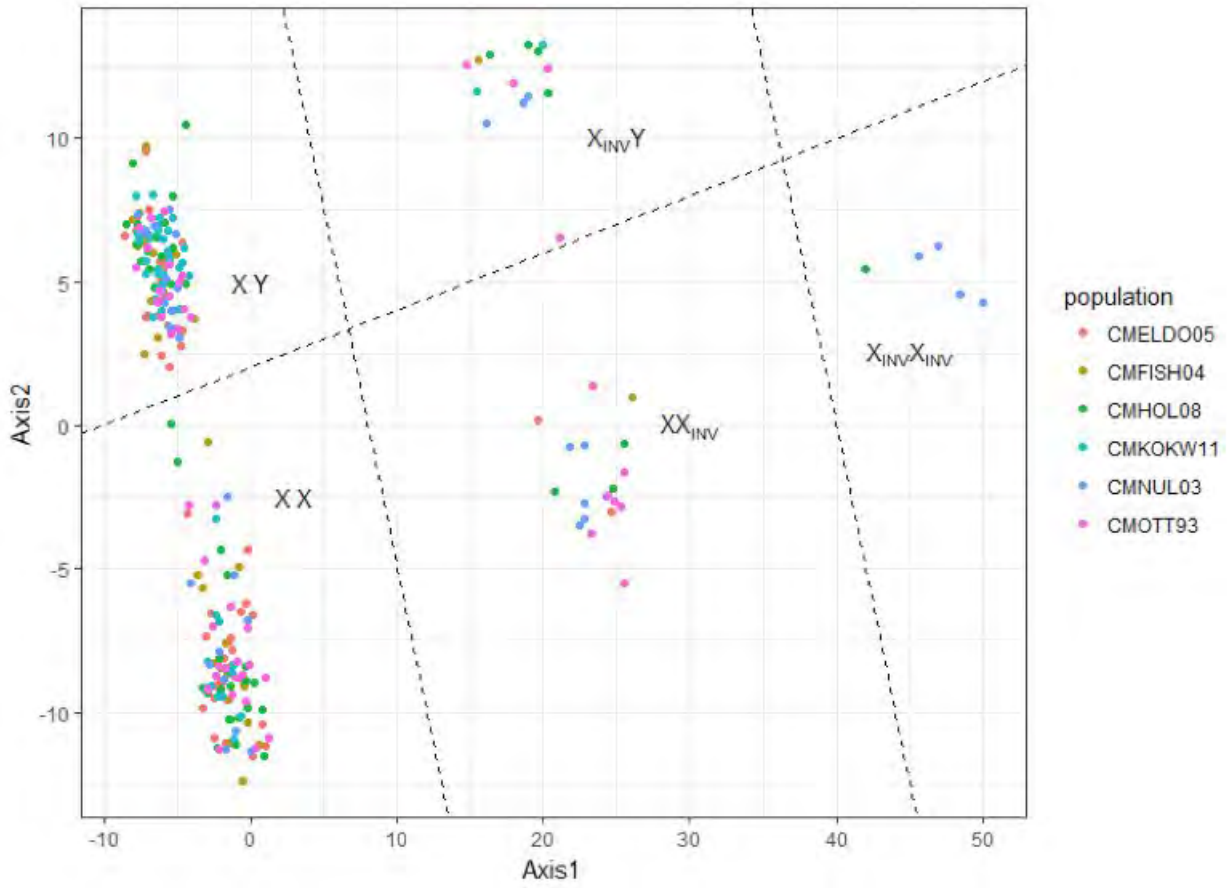

Figure 3. First two principal components of genetic variation among the 6 collections for SNPs on the Omy28 linkage group (each point represents an individual chum salmon). The first axis of variation is driven by loci associated with the chromosomal inversion, while the second axis is associated with putative sex-linked loci.

Literature Cited

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, E. C., R. S. Waples, and S. T. Kalinowski (2008) An improved method for predicting the accuracy of genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences 65:1475-1486.

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.

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, 110 p.

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.* (2017) Potentially adaptive mitochondrial haplotypes as a tool to identify divergent nuclear loci. Methods in Ecology and Evolution 8:821-834.

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.

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, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb (2017) 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. American Fisheries Society Symposium 70: 1177-1197.

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.

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.

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, Li. W., C. Habicht, W. D. Templin, K. E. Tarbox, R. Z. Davis, L. K. Brannian, and J. E. Seeb (2000) Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to management of populations affected by the Exxon Valdez oil spill. Transactions of the American Fisheries Society 129:1223–1249.

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 (2011a) 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.

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

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, E. V. A., P. Bentzen, I. R. Bradbury, M. Clément, J. Pearce, J. Horne, and R. G. Beiko. (2018) Applications of random forest feature selection for fine-scale genetic population assignment. Evolutionary Applications 11:153-165.

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.