Western Alaska Salmon Stock Identification Program

Technical Document: [1] 8

Version: 1.0

Title: Chum salmon SNP selection process outline

Version: 1.0

Authors: N. DeCovich, J. Jasper, C. Habicht, W. Templin

Date: December 14, 2010

Introduction:

Early in the development process for the Western Alaska Salmon Stock Identification Project (WASSIP) it was clear that the resolution possible for chum salmon spawning in western Alaska regional areas (Norton Sound, lower Yukon and Kuskokwim rivers, and Bristol Bay) was not going to be sufficient to meet the standards set by the Advisory Panel (AP) with available genetic markers, including the recently developed SNP markers (see Technical Document 4 for the current panel of 53 SNPs). These four regional areas define important units for management, yet when treated as separate reporting groups each performed below the 90%-correct-allocation level using the 53-marker set. The Department began the process of discovering additional SNP markers for chum salmon through a contract with the International Program for Salmon Ecological Genetics (IPSEG; http://www.fish.washington.edu/research/ipseg/research.html) at the University of Washington. These efforts were based on cDNA sequences from two chum salmon sampled from the Susitna and Delta rivers. This process has been described in a manuscript that has been published in *Molecular Ecology Resources* (Seeb et al. 2011) which is provided as Technical Document 9. This process added 37 validated SNPs to those already available for chum salmon for use in WASSIP. Subsequent rounds of SNP development at the University of Washington were based on 16 fish from four populations from Western Alaska and increased the total number of described SNPs to 228 (Grau et al. in prep).

[1] This document serves as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and the Western Alaska Salmon Stock Identification Program Technical Committee. As such, these documents serve diverse ad hoc information purposes and may contain basic, uninterpreted data. The contents of this document have not been subjected to review and should not be cited or distributed without the permission of the authors or the Commercial Fisheries Division.

Here we describe the process that we intend to use to select the set of 96 SNPs that maximizes the likelihood of providing the resolution necessary to meet the objectives of WASSIP. A similar process was recently completed with the selection of 96 SNP markers for use with sockeye salmon and is described in Technical Document 6, "Selection of the 96-SNP marker set for sockeye salmon." However, the selection of chum salmon SNPs will be significantly different from that used for sockeye salmon. There are many more SNPs available for chum salmon than were available for sockeye salmon (124 SNPs), and more emphasis is placed on selecting markers to distinguish among regional areas (Norton Sound, Yukon summer, Kuskokwim summer, Western Bristol Bay, and Eastern Bristol Bay) within coastal western Alaska (CWAK).

Method:

I. *Pre-ADF&G selection*: Markers were developed under contract at the IPSEG laboratory.
    a. TaqMan assays were developed or available for a total of 228 SNPs including the original 53 SNPs.
    b. Markers were assayed in 80 - 96 individuals from each of 30 populations (Table 1; Figure 1) chosen from across the species range. Ten of these populations were from CWAK (Figure 2).
    c. Of the 228 markers surveyed, 188 markers have been determined to perform adequately in the laboratory and have a reasonable level of variation. Only these markers will be passed on from IPSEG to ADF&G for further analysis.

II. *Unranked measures*: The measures in this section will be given veto power. Markers will be discarded if they do not pass the following tests.
    1. Hardy-Weinberg Equilibrium. Conformance to HWE will be measured using the program Genetic Data Analysis (GDA; Lewis & Zaykin 2001). GDA uses the methods described in Genetic Data Analysis II (Weir 1996). Markers out of HWE at $\alpha = 0.05$ in more than 5 populations or out of HWE at $\alpha = 0.001$ in more than one population will be dropped.
    2. Linkage Disequilibrium/Phase. Linkage Disequilibrium will be measured with the program GDA.

a. Significant disequilibrium between markers will be determined using the sequential Bonferroni with an overall level $\alpha=0.05$ for each marker set adjusted by the number of populations.

b. For marker sets that exhibit disequilibrium, we will next determine whether combining linked markers or discarding a marker is most useful for MSA. To do this with a pair of linked markers we will set up three treatment files:

   i. Marker A combined with marker B ("composite phenotype"; Habicht et al. 2010);

   ii. Marker A retained and marker B excluded; and

   iii. Marker B retained and marker A excluded.

This can be extended to larger linked groups if necessary. We will use f_{orca} (Rosenberg 2005) and measure correct individual assignment to population for the three treatments. The treatment with the best average correct assignment will be selected for further analyses. This method is similar to the methods outlined in Ackerman et al. (In press) where GENECLASS (Piry et al. 2004) was used for the assignment software.

III. *Ranked or scored measures of population structure and MSA performance*: The measures in this phase of the selection process are either ranked or scored and then weighted. Highest weighting is given to measures associated with variation among CWAK populations. Weights are given as percentages and sum to 100%.

1. CWAK –specific measures [84% of total].

Question addressed: What are the best markers for distinguishing among populations or regions within CWAK? This is the most difficult portion of the range to distinguish population structure, yet resolution within this area is central to the objectives of WASSIP.

a. Among populations (24%)

   i. Overall F_{ST} among the 10 CWAK populations. The F_{ST} values calculated from individual markers will be linearly scaled between 0.0 (lowest) and 1.0 (highest) and used as scores.

b. Among regions (60%)

i. Overall θ_P among the 5 CWAK regions. θ_P for each marker will be calculated via a three-level hierarchical ANOVA (Weir, 1995), in which populations from CWAK are organized into five regions (Table 1; Figure 2). The θ_P values calculated from individual markers will be linearly scaled between 0.0 and 1.0 and used as scores. (See Figure 2; 15%)

ii. f_{orca} (Rosenberg 2005) with backward elimination marker selection algorithm method using the five CWAK regions as reporting groups. This method is similar to BELS (Bromaghin 2008) in that it starts with all markers and then sequentially eliminates the marker that provide the least amount of regional discrimination (Technical Document 10). Each marker is then ranked according to the order in which they were eliminated. To then score each marker, we sequentially add markers according to their rank, starting with the most informative marker, and calculate f_{orca} at each step. The resulting f_{orca} values can then be linearly scaled between 0.0 and 1.0, with one corresponding to the most informative marker. BELS is too time-consuming to be used and relies on a simulation method that may introduce bias. (30%)

iii. $\theta_{S(P)} = \theta_S - \theta_P$ for population pairs from adjacent CWAK regions. $\theta_{S(P)}$ for each marker will be calculated via a three-level hierarchical ANOVA, in which populations from adjacent regions are paired. This quantity is a measure of the differentiations among populations within pairs. The four population pairs from adjacent regions with smallest pairwise F_{ST} will be chosen for these tests. The $\theta_{S(P)}$ values calculated from individual markers will be linearly scaled between 0.0 and 1.0 and used as scores. (15%)

2. Pacific–wide measures [10% of total].

Question addressed: What are the best markers for distinguishing among large-scale regions across the species range? Some of the WASSIP fisheries are known to intercept chum salmon from both the western and southeastern extent of the range.

These measures will ensure that broad-scale regions will be identifiable in WASSIP fishery samples.

a. Principle Component Analysis. The amount of variation explained by each marker will be linearly scaled between 0.0 and 1.0 and used as scores.
    i. The amount of variation associated with each marker in the first principle component (3%)
    ii. The amount of variation associated with each marker in the second principle component (3%)
    iii. The amount of variation associated with each marker over the set of principal components that explain 80% of the variation (4%)

3. Outside Alaska, regional measures [6% of total].

Question addressed: What are the best markers for distinguishing between population pairs within or between certain regions outside of Alaska? This is expected to provide insight into markers important for distinguishing broad-scale population structure and is considered to insure a useable panel of SNPs for research groups outside of Alaska. (See Figure 3)

a. Within Japan. Calculate the F_{ST} between populations selected from Honshu and Hokkaido islands (2%). The F_{ST} values calculated from individual markers will be linearly scaled between 0.0 and 1.0 and used as scores.
b. Between Southeast Alaska and Northern British Columbia. Calculate the F_{ST} between population pairs selected from Southeast Alaska and Northern British Columbia (2%). The F_{ST} values calculated from individual markers will be linearly scaled between 0.0 and 1.0 and used as scores.
c. Between Southern British Columbia and Washington. Calculate the F_{ST} between population pairs selected from Southern British Columbia and Washington (2%). The F_{ST} values calculated from individual markers will be linearly scaled between 0.0 and 1.0 and used as scores.

IV. *Final considerations*: The candidate SNPs will be ordered from best to worst with respect to the measures in Section III above. The measures in this section (IV) will be performed on the top 96 candidates based on the measures in Section III (above). If a marker is discarded

due to laboratory performance, the next highest-rated marker from Section III will be evaluated.

1. Performance at the IPSEG Laboratory. Assay performance will be evaluated on three criteria. High-ranking markers that have poor laboratory performance and lower-ranked markers that are difficult to score will be dropped and replaced with the next highest-ranking marker. The process will continue until 96 markers are selected. We incorporate laboratory performance here to avoid the need to examine assay performance of markers that provide little useful MSA performance. Laboratory performance will be evaluated with the following measures as used in the sockeye selection process (Technical Document 6):
    a. Cluster tightness (See Figure 4)
    b. Cluster alignment (See Figure 5)
    c. Drop-out rates (See Figure 6)
2. Final evaluation using simulations to test for loss of MSA resolution for distinguishable regions generally outlined in Seeb et al. (2011). Simulations will be conducted using the selected markers to ensure that the reporting groups represented in this data set that were distinguishable in Seeb et al. (2011) continue to be distinguishable (> 90% correct allocation). Matching exact reporting groups will not be possible, but reasonable approximations will be tested. These reporting groups will include (corresponding population numbers from Table 1 in parentheses): Japan (1,2), Russia (3,4), Kotzebue Sound (5,6,), CWAK (7,8,9,10,13,14,15,16), Yukon Fall (11,12), Eastern Bristol Bay (17,18), North Alaska Peninsula (19,20), South Alaska Peninsula (21,22), Southcentral Alaska (23,24), Southeast Alaska/BC (25,26,27,28), and Washington (29,30). Mean correct allocations in the Seeb et al. (2011) study ranged from 85% to 99%, with the majority of reporting regions allocating above 90%. The results from our analysis are expected to be optimistic given that regions are represented by only a few populations. Therefore, mean correct allocations to reporting groups below 90% will trigger addition of markers that were highly ranked from sections III.2 and III.3. As markers are added, the lowest-ranked markers from the III.1 process will be dropped. Markers will be added and dropped following these rules until the resolution to these broader reporting groups exceeds 90%.

3. Laboratory performance in ADF&G. All 188 SNPs will be assayed in the Gene Conservation Laboratory on 3,032 chum salmon originating from Prince William Sound as part of a Pacific Coast Salmon Recovery Fund project. This will allow us to confirm assay performance in our lab.

Literature Cited:

Ackerman, M. W., C. Habicht, and L. W. Seeb. In Press. SNPs under diversifying selection provide increased accuracy and precision in mixed stock analyses of sockeye salmon from Copper River, Alaska and nearby coastal areas. Transactions of the American Fisheries Society. XX:XXX-XXX

Bromaghin, J. F. 2008. BELS: backward elimination locus selection for studies of mixture composition or individual assignment. Molecular Ecology Resources 8: 568-571.

Habicht, C., L. W. Seeb, K. W. Myers, E. V. Farley, and J. E. Seeb. 2010. Summer-fall distribution of stocks of immature sockeye salmon in the Bering Sea as revealed by single-nucleotide polymorphisms. Transactions of the American Fisheries Society 139(4):1171-1191.

Lewis P.O., and D. Zaykin. 2001. GENETIC DATA ANALYSIS: computer program for the analysis of allelic data, version 1.0 (d16c) Free program distributed by the authors from http://lewis.eeb.uconn.edu/lewishome/software.html.

Piry, S., A. Alapetite, J. M. Cornuet, D. Paetkau, L. Baudouin, and A. Estoup. 2004. GENECLASS2: A software for genetic assignment and first-generation migrant detection. Journal of Heredity 95(6):536-539.

Rosenberg, N.A. 2005. Algorithms for selecting informative marker panels for population assignment. Journal of Computational Biology 12(9):1183-1201.

Seeb, J.E., C.E. Pascal, E.D. Grau, L.W. Seeb, W.D. Templin, S.B. Roberts, and T. Harkins. 2011. Transcriptome sequencing and high-resolution melt analysis advance SNP discovery in duplicated salmonids. Molecular Ecology Resources doi: 10.1111/j.1755-0998.2010.02936.x.

Seeb, L.W., W.D. Templin, S. Sato, S. Abe, K.I. Warheit, and J.E. Seeb. 2011. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. Molecular Ecology Resources xxx: xx-xx

Weir, B. S. 1996. Genetic Data Analysis II. Sinauer Associates, Inc., Sunderland, Mass.

Specific questions for the Technical Committee:

1. Is our approach to linkage disequilibrium and HWE reasonable?
2. Is our method to determine the relative value of different treatments of linked markers advisable? Is the use of f_{orca} as a measure appropriate?
3. Are the tests appropriately structured to provide a set of SNPs that will perform well for WASSIP?
4. Does the weighting applied to each set of tests seem reasonable?
5. Are there other measures that would be more appropriate?

General comments: The approach proposed here borrows useful ideas from the approach used for sockeye salmon (described in Technical Document 6) but appears to be more streamlined and efficient. The text is a bit confusing about how the laboratory screening will occur. At line 41, the report states, "Of the 228 markers surveyed, 188 markers have been determined to perform adequately in the laboratory and have a reasonable level of variation. Only these markers will be passed on from IPSEG to ADF&G for further analysis." This implies that data quality issues in the laboratory have already been evaluated prior to screening loci for power to discriminate populations. However, at line 140 another process is described that seems to involve iterative consideration of discriminatory power and laboratory performance.

Responses to specific questions:

1. *Is our approach to linkage disequilibrium and HWE reasonable?*

For the most part, but we have several comments to consider.

1) For both types of analyses, it is important to ensure that the baseline populations represent single panmictic populations. If not, a Wahlund effect could cause both HW and LD departures that appear to be data quality issues but actually reflect population mixture.
2) For both types of analyses, be careful about only using results of tests of statistical significance. You are really interested in the magnitude of the effect size here, but P values also depend heavily on sample sizes. Also, the direction of departure (e.g., heterozygotes excess or deficiency) can be informative about potential causes.
3) The LD analyses will consider pairs of loci, of which there are n(n-1)/2 possible comparisons for n loci. Since n could be 200 or more, this represents a huge number of pairwise comparisons, each of which could be conducted for many different populations. Using the Bonferroni correction here would require consideration of tiny P values, which could lead to unpredictable results. It is probably more useful to screen for pairs of loci that are consistently out of equilibrium (using the nominal alpha level) in multiple populations. Some consideration of effect size (the magnitude of LD) would also be useful in evaluating how serious a problem any deviations are likely to cause.

2. *Is our method to determine the relative value of different treatments of linked markers advisable? Is the use of fORCA as a measure appropriate?*

The general procedure described at lines 56-68 of Document 8 seems reasonable, as does the logic for using a procedure that assigns entire individuals rather than making fractional

assignments. With the caveats noted below, *fORCA* should be ok as a means to assess *relative* power for correct assignment.

3. *Are the tests appropriately structured to provide a set of SNPs that will perform well for WASSIP?*

The proposed methods should produce a set of SNPs with high power to resolve stock identification problems in Western Alaska.

4. *Does the weighting applied to each set of tests seem reasonable?*

The weights chosen are obviously somewhat arbitrary but do not appear to be unreasonable. Because of the applied focus of this project, it is appropriate to assign greater weight to markers that have high power for the local areas of interest. However, we were pleased to see that the criteria include non-trivial weight to markers with wider geographic relevance (10% weight for Pacific Rim individual populations, plus 6% for major non-Alaska groups). This will help ensure that the considerable efforts here to develop markers will have much broader application to the scientific and fishery management communities.

Minor comments:

In the proposed PCA analysis for Pacific-wide assessments, part (iii) is partially redundant as it will include information already used for (i) and (ii)

Outside Alaska: we don't necessarily disagree with the particular comparisons proposed, but the rationale for choosing them is not given.

5. *Are there other measures that would be more appropriate?*

Can't think of any offhand.

<u>General comments about bias and *fORCA*</u>

It is important to distinguish between two different types of biases that can potentially arise in evaluations such as those proposed here.

The first type of bias, described by Anderson et al. (2008), occurs when one is interested in assessing the power of a particular set of markers to resolve the composition of a mixture comprised of individuals from a specified group of source populations. The ideal way to do this is to create simulated mixtures of individuals, with the genotype of each individual being chosen based on actual allele frequencies in one of the (randomly chosen) source populations. The bias arises because we never know the actual allele frequencies—we only have samples. Because of random sampling error, allele frequencies in samples from the baseline populations will on average be more divergent than are the true population allele frequencies. On average, this factor inflates Fst among baseline samples by the magnitude 1/(2S), where S is the baseline sample size. When simulated mixtures are constructed using

these baseline allele frequencies (which appear more different than the populations actually are), the population assignments will tend to be overly optimistic. Furthermore, the relative importance of sampling error (and hence the bias) will be larger when true genetic differences among populations are very small—as occurs with Western Alaska chum salmon. Anderson et al. (2008) described a simple leave-one-out procedure that eliminates the bias, but the routine described at lines 41-50 of Document 10 would be subject to this type of bias.

The second type of bias, described by Anderson (2010), applies to locus-selection programs. The bias is not in the locus selection *per se*, but rather in the evaluation of power of the resulting set of loci for population assignment. Anderson (2010) showed that the bias arises because none of the commonly-used software programs for locus selection (including BELS) use proper cross validation. Instead, some of the information used to select the panel of loci is also used to evaluate its performance, and this leads to an overly optimistic assessment of assignment power. We did not see any indication that the combined *fORCA*-BELS approach proposed in Document 10 would *not* be subject to this type of bias. Also, although the authors list 4 methods Rosenberg (2005) evaluated for selecting subsets of loci, they don't explain why they did not consider any of them for the current project.

One reason that proper cross-validation is often not done is that it is costly in terms of information content. The "gold standard" of cross validation is to split the data in half: the first half is used to develop the algorithm, the second half to evaluate its performance. However, doing this means that the algorithm is likely to be less precise because it is based on less data. Researchers are thus typically faced with a trade-off between precision in developing the best algorithm (use all the data in the first step) and the downstream consequences (subsequent assessments of performance using the same data will tend to be overly optimistic). Anderson (2010) suggested a simple modification to the cross-validation procedure that retains most of the information without leading to appreciable bias in assessing performance.

In summary, both types of biases can lead to overly optimistic assessments of power, which should be a concern given the stated goals of the project. For applications that only consider relative power, these biases might not be important. Also, it might be the case that the proposed locus-selection approach is perfectly fine for selecting an optimal panel of loci, but that the estimates of power to be expected when that panel is applied to real data are biased upwards.

Text at lines 84-91 of Document 10 seems to acknowledge at least the bias problem identified by Anderson et al. (2008), but it is not clear that both of the potential sources of bias described above have been fully considered in the documents we reviewed. This topic merits closer scrutiny to determine the optimal way to proceed given project goals.

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

Anderson, E.C. 2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources* 10:701-710.

Table 1. Population set used in this analysis. Map numbers correspond to numbers in Figure 1.

ADF&G region	Population	Sample size	Map Number
Japan	Tokachi River	80	1
	Gakko River late	80	2
Russia	Amur River summer	95	3
	Palana River	95	4
Kotzebue Sound	Kiana River	95	5
	Inmachuk River	95	6
[1]Norton Sound	Kwiniuk River	95	7
	Unalakleet River	95	8
[1]Yukon summer	Andreafsky River - East Fork weir	95	9
	Nulato River	95	10
Yukon fall	Fishing Branch	95	11
	Kluane River	95	12
[1]Kuskokwim summer	Salmon River	95	13
	Kanektok River weir	95	14
[1]Western Bristol Bay	Osviak River	95	15
	Iowithla River	95	16
[1]Eastern Bristol Bay	Whale Mountain Creek	95	17
	Alagnak River	95	18
North Alaska Peninsula	Frosty Creek	95	19
	Sapsuk - Nelson River	95	20
South Alaska Peninsula	Portage Creek	95	21
Kodiak	Rough Creek	95	22
Southcentral Alaska	Little Susitna River weir	95	23
	Beartrap Creek	95	24
Southeast Alaska	Chilkat River - 24Mile	95	25
	North Arm Creek	95	26
British Columbia	Kitimat River	95	27
	Kitwanga River	95	28
Washington	Nisqually River Hatchery	95	29
	Elwha River	95	30

[1] Populations in the Coastal Western Alaska (CWAK) Region.

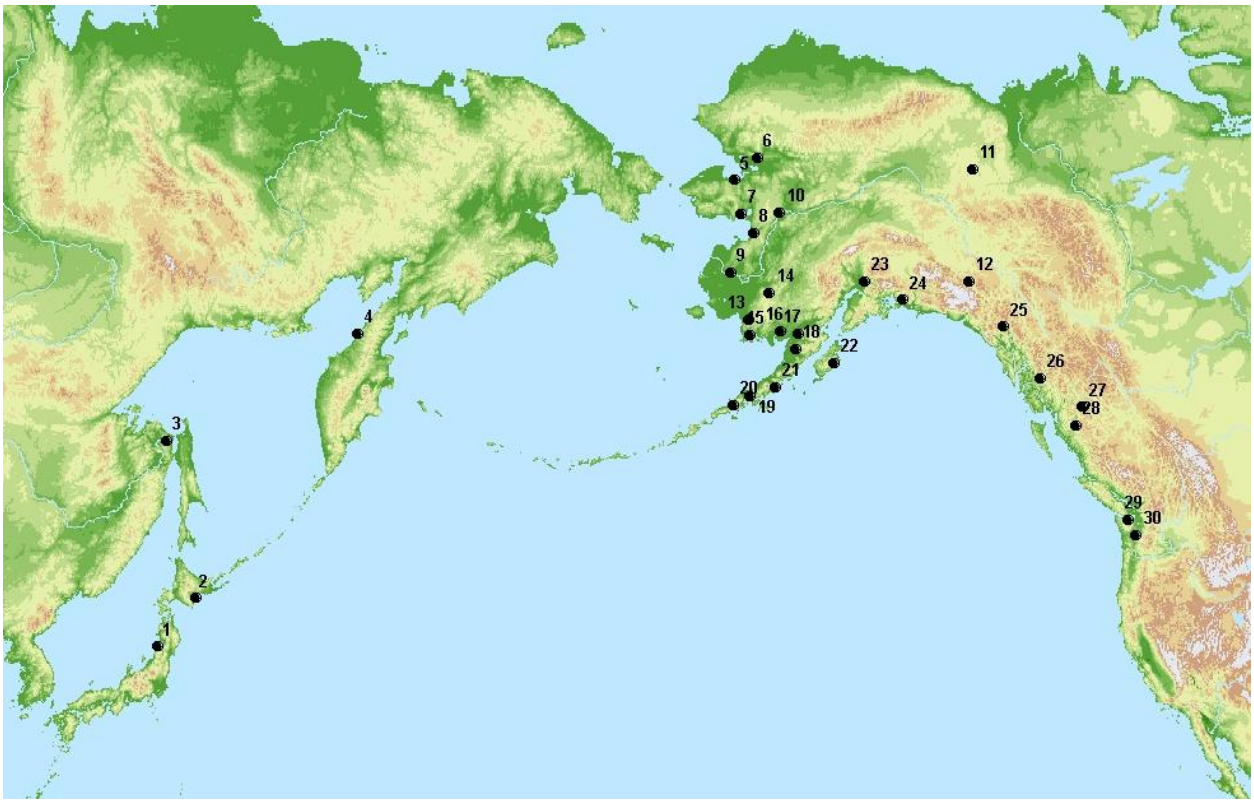

Figure 1. Map of chum salmon populations used in SNP selection process.

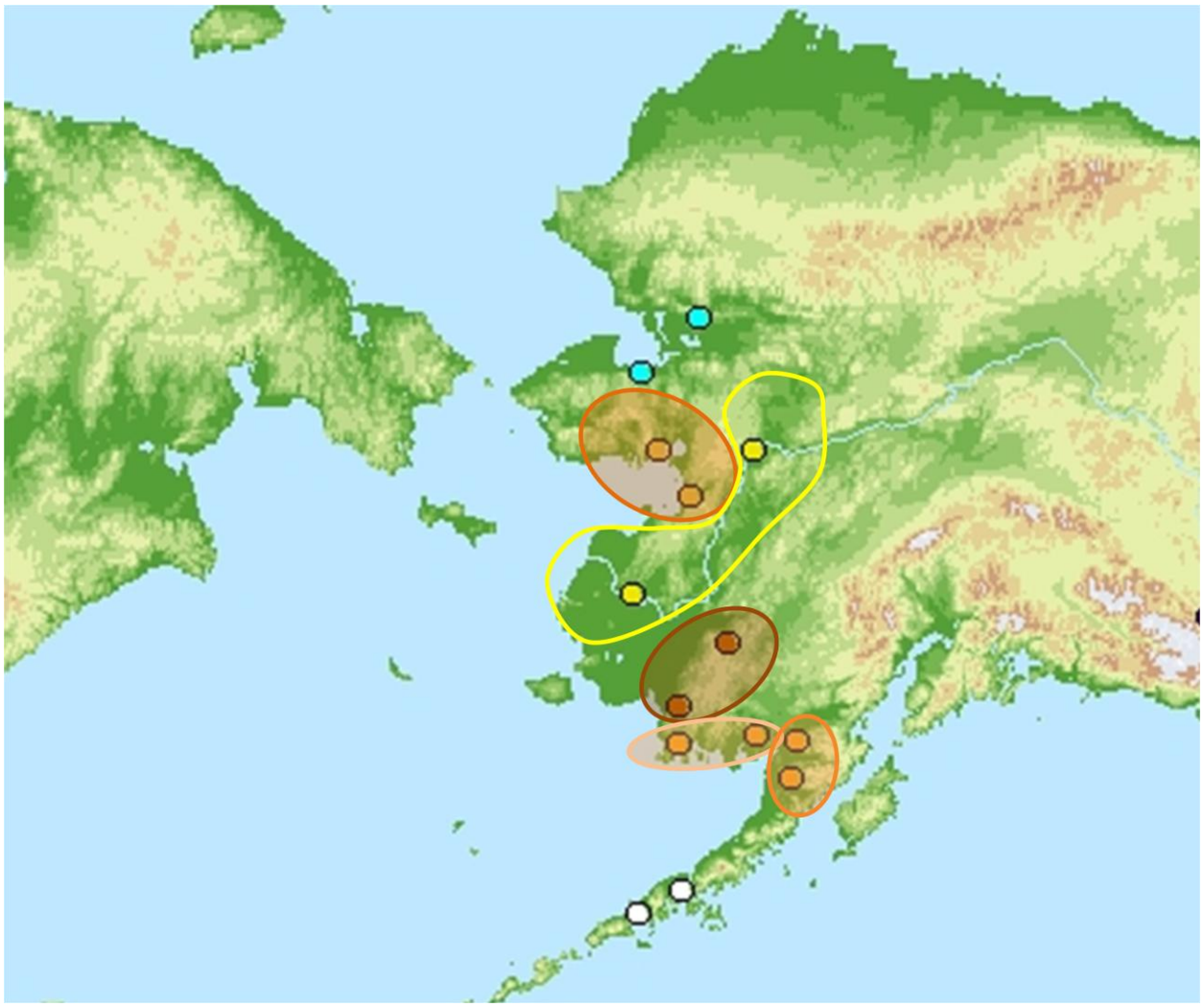

Figure 2. Locations of chum salmon collections within western Alaska. The five regions within Western Coastal Alaska to be measured using overall F_{ST} are indicated by the ellipses.

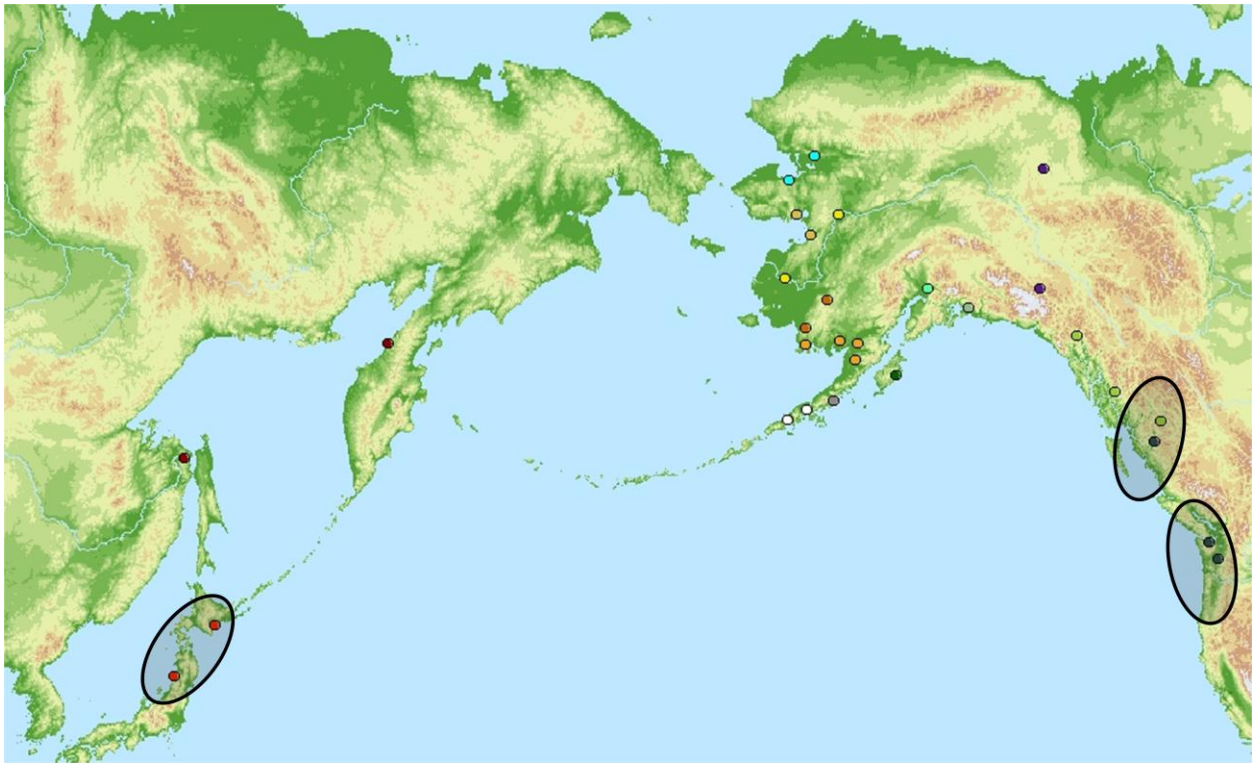

Figure 3. Chum salmon populations used in SNP selection process highlighting the three population pairs (in ovals) of chum salmon chosen to measure F_{ST} within regions of interest to research groups outside of Alaska.

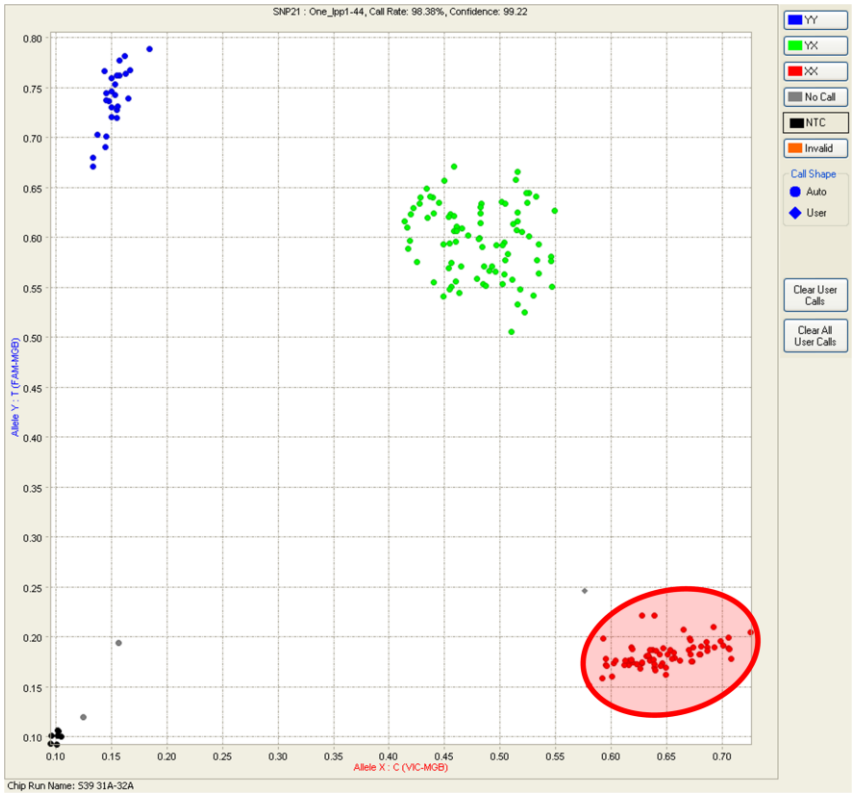

Figure 4. Screen capture of a scatter plot from genotyping software. Each point represents a single fish. The three clusters represent each possible genotype (TT homozygote - blue, TC heterozygote - green, and CC homozygote - red). The size of the shaded area for the CC homozygote distribution is an indication of cluster tightness.

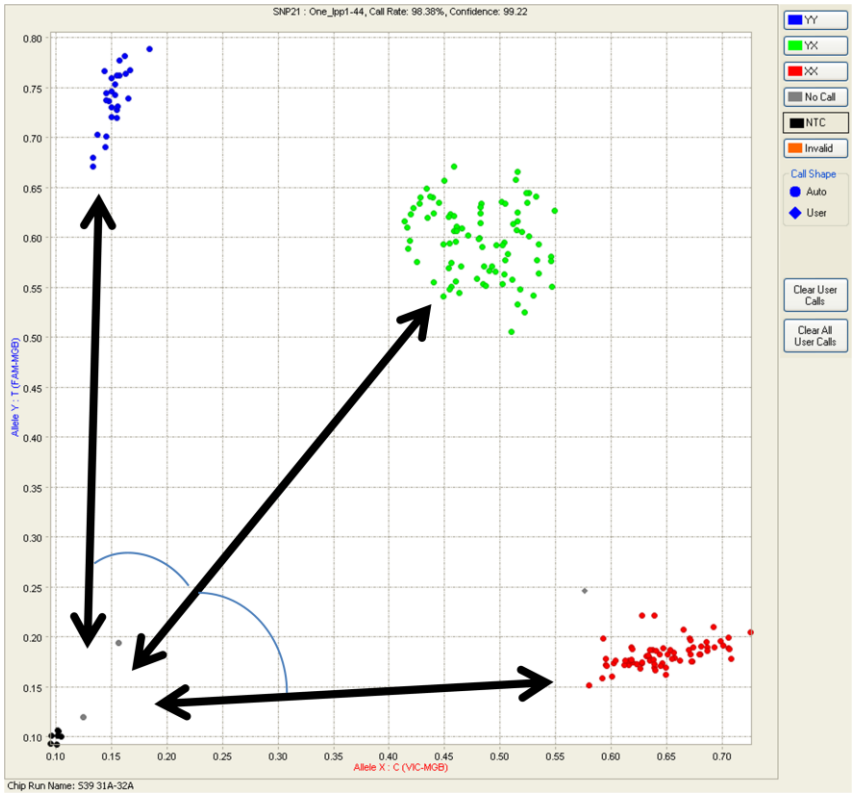

Figure 5. Screen capture of a scatter plot from genotyping software. Each point represents a single fish. The three clusters represent each possible genotype (TT homozygote - blue, TC heterozygote - green, and CC homozygote - red). The angle between the double-ended arrows is an indication of cluster alignment.

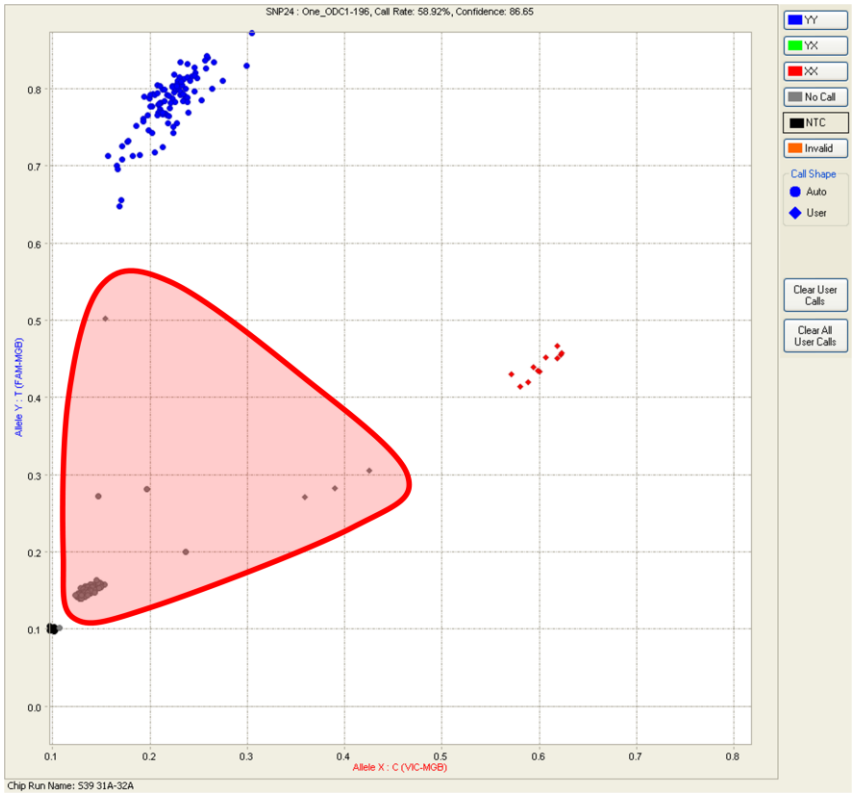

Figure 6. Screen capture of a scatter plot from genotyping software. Each point represents a single fish. The three clusters represent each possible genotype (TT homozygote - blue, TC heterozygote - green, and CC homozygote - red). The red shaded area represents fish for which the assay failed.