Regional Information Report 5J12-15

Western Alaska Salmon Stock Identification Program Technical Document 8: Chum Salmon SNP Selection Process Outline

by

Nick DeCovich,

James R. Jasper,

Christopher Habicht,

and

William D. Templin

October 2012

Alaska Department of Fish and Game **Divisions of Sport Fish and Commercial Fisheries**

Symbols and Abbreviations

The following symbols and abbreviations, and others approved for the Système International d'Unités (SI), are used without definition in the following reports by the Divisions of Sport Fish and of Commercial Fisheries: Fishery Manuscripts, Fishery Data Series Reports, Fishery Management Reports, Special Publications and the Division of Commercial Fisheries Regional Reports. All others, including deviations from definitions listed below, are noted in the text at first mention, as well as in the titles or footnotes of tables, and in figure or figure captions.

Weights and measures (metric)

centimeter	cm
deciliter	dL
gram	g
hectare	ha
kilogram	kg
kilometer	km
liter	L
meter	m
milliliter	mL
millimeter	mm

Weights and measures (English)

cubic feet per second	ft^3/s
foot	ft
gallon	gal
inch	in
mile	mi
nautical mile	nmi
ounce	oz
pound	lb
quart	qt
yard	yd

Time and temperature

day	d
degrees Celsius	°C
degrees Fahrenheit	°F
degrees kelvin	K
hour	h
minute	min
second	s

Physics and chemistry

all atomic symbols	
alternating current	AC
ampere	A
calorie	cal
direct current	DC
hertz	Hz
horsepower	hp
hydrogen ion activity (negative log of)	pH
parts per million	ppm
parts per thousand	ppt, ‰
volts	V
watts	W

General

Alaska Administrative Code	AAC
all commonly accepted abbreviations	e.g., Mr., Mrs., AM, PM, etc.
all commonly accepted professional titles	e.g., Dr., Ph.D., R.N., etc.
at	@
compass directions:	
east	E
north	N
south	S
west	W
copyright	©
corporate suffixes:	
Company	Co.
Corporation	Corp.
Incorporated	Inc.
Limited	Ltd.
District of Columbia	D.C.
et alii (and others)	et al.
et cetera (and so forth)	etc.
exempli gratia (for example)	e.g.
Federal Information Code	FIC
id est (that is)	i.e.
latitude or longitude	lat. or long.
monetary symbols (U.S.)	$, ¢
months (tables and figures): first three letters	Jan,...,Dec
registered trademark	®
trademark	™
United States (adjective)	U.S.
United States of America (noun)	USA
U.S.C.	United States Code
U.S. state	use two-letter abbreviations (e.g., AK, WA)

Mathematics, statistics

all standard mathematical signs, symbols and abbreviations	
alternate hypothesis	H_A
base of natural logarithm	*e*
catch per unit effort	CPUE
coefficient of variation	CV
common test statistics	(F, t, χ^2, etc.)
confidence interval	CI
correlation coefficient (multiple)	R
correlation coefficient (simple)	r
covariance	cov
degree (angular)	°
degrees of freedom	df
expected value	*E*
greater than	>
greater than or equal to	≥
harvest per unit effort	HPUE
less than	<
less than or equal to	≤
logarithm (natural)	ln
logarithm (base 10)	log
logarithm (specify base)	$\log_2$, etc.
minute (angular)	'
not significant	NS
null hypothesis	H_O
percent	%
probability	P
probability of a type I error (rejection of the null hypothesis when true)	α
probability of a type II error (acceptance of the null hypothesis when false)	β
second (angular)	"
standard deviation	SD
standard error	SE
variance	
population	Var
sample	var

REGIONAL INFORMATION REPORT 5J12-15

WESTERN ALASKA SALMON STOCK IDENTIFICATION PROGRAM TECHNICAL DOCUMENT 8: CHUM SALMON SNP SELECTION PROCESS OUTLINE

by
Nick DeCovich, James R. Jasper, Christopher Habicht, and William D. Templin
Alaska Department of Fish and Game, Division of Commercial Fisheries, Gene Conservation Laboratory, Anchorage

Alaska Department of Fish and Game
Division of Sport Fish, Research and Technical Services
333 Raspberry Road, Anchorage, Alaska, 99518-1565

October 2012

The Regional Information Report Series was established in 1987 and was redefined in 2006 to meet the Division of Commercial Fisheries regional need for publishing and archiving information such as project operational plans, area management plans, budgetary information, staff comments and opinions to Board of Fisheries proposals, interim or preliminary data and grant agency reports, special meeting or minor workshop results and other regional information not generally reported elsewhere. Reports in this series may contain raw data and preliminary results. Reports in this series receive varying degrees of regional, biometric and editorial review; information in this series may be subsequently finalized and published in a different department reporting series or in the formal literature. Please contact the author or the Division of Commercial Fisheries if in doubt of the level of review or preliminary nature of the data reported. Regional Information Reports are available through the Alaska State Library and on the Internet at http://www.adfg.alaska.gov/sf/publications/.

Note: 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.

Note: The appearance of product names or specific company names is not an Alaska Department of Fish and Game recommendation for or implied endorsement. The Alaska Department of Fish and Game, in accordance with State of Alaska ethics laws, does not favor one group over another through endorsement or recommendation.

Nick DeCovich, James R. Jasper, Christopher Habicht, and William D. Templin
Alaska Department of Fish and Game, Division of Commercial Fisheries, Gene Conservation Laboratory
333 Raspberry Road, Anchorage, Alaska, 99518-1565 USA

This document should be cited as:
DeCovich, N., J. R. Jasper, C. Habicht, and W. D. Templin. 2012. Western Alaska Salmon Stock Identification Program Technical Document 8: Chum salmon SNP selection process outline. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-15, Anchorage.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

Uncertainty about the magnitude, frequency, location, and timing of the nonlocal harvest of sockeye and chum salmon in Western Alaska fisheries was the impetus for the Western Alaska Salmon Stock Identification Program (WASSIP). The project was designed to use genetic data in mixed stock analysis (MSA) to reduce this uncertainty. A baseline of allele frequencies is required for use in mixed stock analysis to estimate the stock of origin of harvested fish. The single nucleotide polymorphism (SNP) baseline for chum salmon *Oncorhynchus keta* to be used for MSA in WASSIP is in a state of perpetual improvement. To meet the standards set by the Advisory Panel (AP) for increased resolution more emphasis was placed on selecting markers to distinguish among regional areas within Coastal Western Alaska (CWAK). 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.

Key words: Western Alaska Salmon Stock Identification Program, WASSIP, chum salmon, *Oncorhynchus keta*, mixed stock analysis, MSA, genetic baseline, single nucleotide polymorphism, SNP

INTRODUCTION

Early in the development process for the Western Alaska Salmon Stock Identification Program (WASSIP) it was clear that the resolution possible for chum salmon *Oncorhynchus keta* 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 Jasper et al. 2012 for the current panel of 53 SNPs). These 4 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 Alaska Department of Fish and Game (ADF&G) 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 2 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 in Seeb et al. (2012). 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 4 populations from Western Alaska and increased the total number of described SNPs to 228 (E. L. Petrou et al. *in prep*).

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 Dann et al. (2012). 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).

METHODS

Phase 1: Pre-ADF&G selection

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.[1]

Phase 2: Unranked measures

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 (HWE). 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 exhibiting overall significance, measured across all thirty populations, at $\alpha = 0.01$ will be dropped. An overall p-value will be calculated using the following equation: p =CHIDIST(2*SUM(LN(C3:AE3)), 2*COUNT(C3:AE3).

2. Linkage Disequilibrium. Linkage Disequilibrium will be measured with the program GDA. Marker pairs that exhibit linkage disequilibrium at $\alpha = 0.05$ in more than 50% of populations examined will be considered "associated". For marker sets considered associated,[2] 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 3 treatment files:

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

   b. Marker A retained and marker B excluded; and

   c. Marker B retained and marker A excluded.

We either removed 1 of the associated SNPs or combined the pair into a composite, phenotype marker in further analyses if the pattern of linkage provided information useful for mixed stock analysis. We used f_{ORCA} as our measure of information. f_{ORCA} assesses the rate of correct

[1] This sentence is commented on in the section entitled "Technical Committee Review and Comments."

[2] In the original version of this document, this phrase and the previous sentence were rather: "Significant disequilibrium between markers will be determined using the sequential Bonferroni with an overall level α=0.05 for each marker set adjusted by the number of populations. For maker sets that exhibit disequilibrium. . ." The original sentence is commented on in the section entitled "Technical Committee Review and Comments."

allocation of simulated individuals to defined reporting groups based upon the markers in question (Rosenberg 2005). Because combinations of alleles from 2 or more markers can exist in more forms than single markers (9 possible phenotypes vs. 2 alleles for SNPs), composite markers generally have higher f_{ORCA} values than the single markers that form them. Simple comparisons of these values would often suggest combining linked pairs into composite markers. However, there is a cost associated with composite markers as estimates of 8 phenotype frequencies are less precise than estimates of 1 allele frequency for a given sample size.

To account for this cost, and to ensure that we combined only SNP pairs that provided significantly more information than the single SNPs in question, we compared the difference between f_{ORCA} values of the composite marker and the single SNP with the greater f_{ORCA} value in the pair ($\Delta = f_{ORCA\text{-pair}} - \max(f_{ORCA\text{-single1}}, f_{ORCA\text{-single2}})$). This difference ($\Delta$) was our test statistic. Since we did not know the distribution of Δ, we conducted a sampled randomization test (Sokal and Rohlf, 2005). We randomly selected 1,000 SNP pairs, calculated Δ for each pair to empirically define the test statistic distribution, and set the 90th quantile of the distribution as a critical value (Δ_{90}). We then combined linked SNPs into composite, phenotype markers if Δ was greater than this critical value and dropped the SNP with the lower f_{ORCA} value if Δ was less than the critical value.[3]

Phase 3: Ranked or scored measures of population structure and MSA performance

III. 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. Measures were linearly scaled between 0.0 (lowest) and 1.0 (highest) and used as scores using the equation: $\theta=\theta + |\theta\min| / \theta\max + |\theta\min|$ for cases where $\theta\min < 0$, and $\theta=\theta + \theta\min / \theta\max + \theta\min$ for cases where $\theta\min > 0$. Weights were calculated as percentages and sum to 100%, and are given in parentheses below. 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 3-level hierarchical ANOVA (Weir, 1995), in which populations from CWAK are organized into 5 regions (Table 1; Figure 2). The θ_P values

[3] Instead of this paragraph and the previous one, the original version of the document had rather: "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. (2011) where GENECLASS (Piry et al. 2004) was used for the assignment software." The original paragraph is commented on in the section entitled "Technical Committee Review and Comments."

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 all 15 regions as reporting groups. Individuals will be sampled from the Norton Sound, Lower Yukon, and Kuskokwim regions because these have been the hardest to differentiate in prior analyses. 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 (Jasper and Templin 2012). 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 1 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. F_{ST} for population pairs from adjacent CWAK regions. F_{ST} for each marker will be calculated in which populations from adjacent regions are paired. The 4 population pairs from adjacent regions with smallest pairwise F_{ST} will be chosen for these tests. The F_{ST} 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 for the first 2 components. 2Weight will be divided between these 2 components in proportion to the variance each component explains.

      i. The amount of variation associated with each marker in the first principle component (0-10%)

      ii. The amount of variation associated with each marker in the second principle component (0-10%)

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 (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 θ_P between population pairs selected from Southern British Columbia and Washington (2%). The θ_P values calculated from individual markers will be linearly scaled between 0.0 and 1.0 and used as scores.

Final Considerations

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.[4]

1. Performance at the IPSEG Laboratory. Assay performance will be evaluated on 3 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 on DNA templates extracted from various tissue types. Further, assay performance will be evaluated on DNA template that has been both pre-amplified and not pre-amplified with the expectation that the pre-amplified templates will produce better results. Doing this is expected to give an indication assay robustness across varying template qualities. Assays will be given a rating of poor, acceptable, good, or great. Those assays given a poor rating will be dropped. As with the sockeye selection process, the following indicators of performance will be considered:

   a. Cluster tightness (Figure 4)

   b. Cluster alignment (Figure 5)

   c. Drop-out rates (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

[4] This paragraph is commented on in the section entitled "Technical Committee Review and Comments."

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 those reported in Seeb et al. 2011 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.

ACKNOWLEDGMENTS

The Technical Document series served as a record of communication between the Alaska Department of Fish and Game Commercial Fisheries Division and the Western Alaska Salmon Stock Identification Program (WASSIP) Technical Committee during the implementation of the program. The authors would like to thank the WASSIP Technical Committee and Advisory Panel for their constructive input on each of the documents throughout the project. The authors would also like to thank Erica Chenoweth who coordinated and prepared the Technical Document series for publication and Publication Specialists Amy Carroll and Joanne MacClellan for implementing the series into Regional Information Reports.

REFERENCES CITED

Ackerman, M. W., C. Habicht, and L. W. Seeb. 2011. 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 140(3):865-881.

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

Dann, T. H., J. R. Jasper, H. A. Hoyt, H. Hildebrand, and C. Habicht. 2012. Western Alaska Salmon Stock Identification Program Technical Document 6: Selection of the 96 SNP marker set for sockeye salmon. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-11, Anchorage.

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.

Jasper, J. R., and W. D. Templin. 2012. Western Alaska Salmon Stock Identification Program Technical Document 10: Optimal rate of correct assignment with backward elimination locus selection. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-19, Anchorage.

Jasper, J. R., N. DeCovich, W. D. Templin. 2012 . Western Alaska Salmon Stock Identification Program Technical Document 4: Status of the SNP baseline for chum salmon. Alaska Department of Fish and Game, Regional Information Report 5J12-09, Anchorage.

Lewis P.O., and D. Zaykin. 2001. GENETIC DATA ANALYSIS: computer program for the analysis of allelic data, version 1.0 (d16c) http://darwin.eeb.uconn.edu/eeb348/archives/000359.html. (Accessed April 20, 2012).

Petrou, E. L., D. Gomez-Uchida, J. E. Seeb, W. D. Templin, R. S. Waples, and L. W. Seeb. *In prep.* Secondary contact, colonization, and natural selection in lineages of chum salmon following Pleistocene glacial retreat. *To be submitted to Molecular Ecology*.

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, J. E., C. E. Pascal, E. D. Grau, L. W. Seeb, W. D. Templin, T. Harkins, and S. B. Roberts. 2012. Western Alaska Salmon Stock Identification Program Technical Document 9: Chum salmon SNP discovery – First method. Alaska Department of Fish and Game, Division of Commercial Fisheries, Regional Information Report 5J12-14, Anchorage.

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 11: 195-217.

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

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?

RESPONSES FROM THE TECHNICAL COMMITTEE

Note: This technical document, particularly the Methods section, underwent revision subsequent, and often in response to the Technical Committee's comments. This report represents the final version and includes all comments, including those on the original document. Any text altered from the original version and commented on by the Technical Committee is reproduced in lettered footnotes.

General Comments

The approach proposed here borrows useful ideas from the approach used for sockeye salmon (described in Dann et al., 2012) but appears to be more streamlined and efficient. The text is a bit confusing about how the laboratory screening will occur. 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 (see Methods section II.1, p. 2, note 2)." This implies that data quality issues in the laboratory have already been evaluated prior to screening loci for power to discriminate populations. However, in the opening statement of section IV (p. 5, note 4) another process is described that seems to involve iterative consideration of discriminatory power and laboratory performance.

Responses to Questions

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

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

   a) 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.

   b) 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.

c) 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 f_{ORCA} as a measure appropriate?

The general procedure described in section II.2 seems reasonable, as does the logic for using a procedure that assigns entire individuals rather than making fractional assignments. With the caveats noted below, f_{ORCA} 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.

Comments about Bias and f_{ORCA}

It is important to distinguish between 2 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 1 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 F_{st} 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 f_{ORCA}-BELS approach proposed in Jasper and Templin (2012) 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.

The final paragraph of Jasper and Templin (2012) 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.

TABLES

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
[a]Norton Sound	Kwiniuk River	95	7
	Unalakleet River	95	8
[a]Yukon summer	Andreafsky River - East Fork weir	95	9
	Nulato River	95	10
Yukon fall	Fishing Branch	95	11
	Kluane River	95	12
[a]Kuskokwim summer	Salmon River	95	13
	Kanektok River weir	95	14
[a]Western Bristol Bay	Osviak River	95	15
	Iowithla River	95	16
[a]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

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

FIGURES

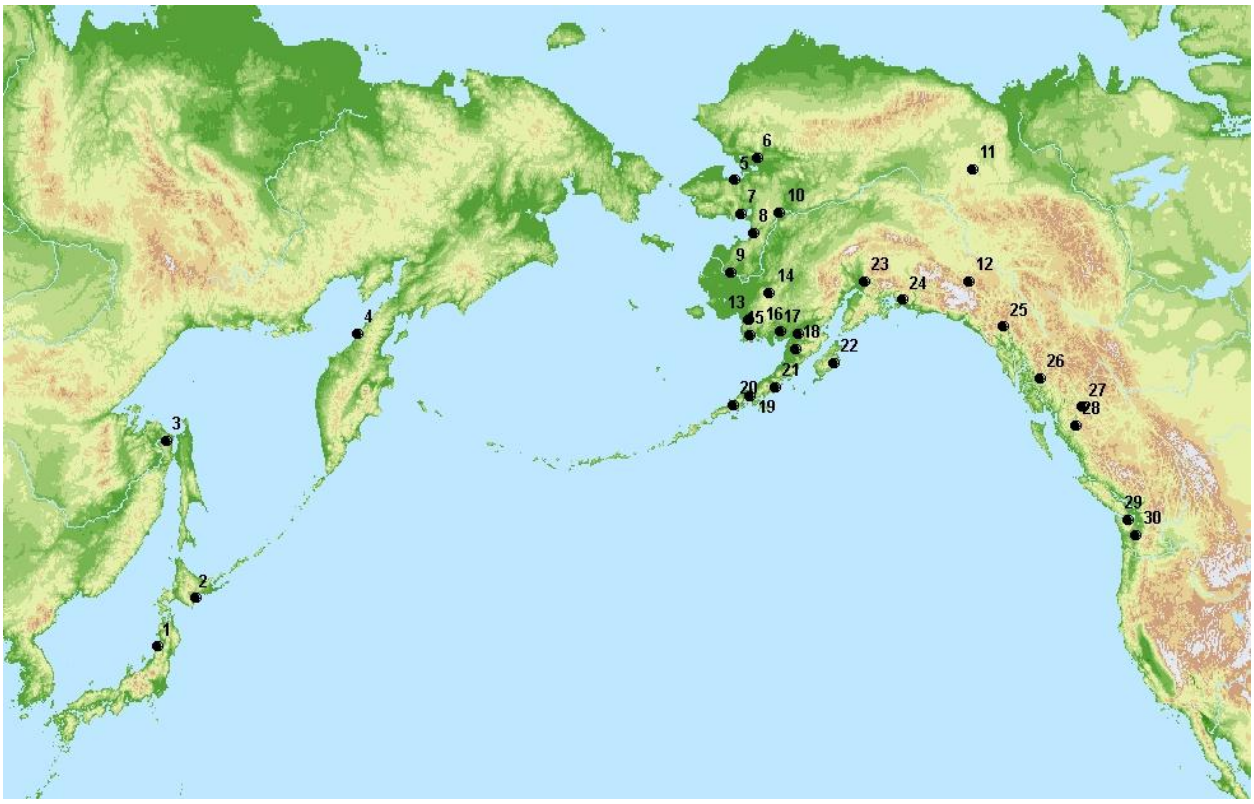

Figure 1.–Map of chum salmon populations used in SNP selection process.

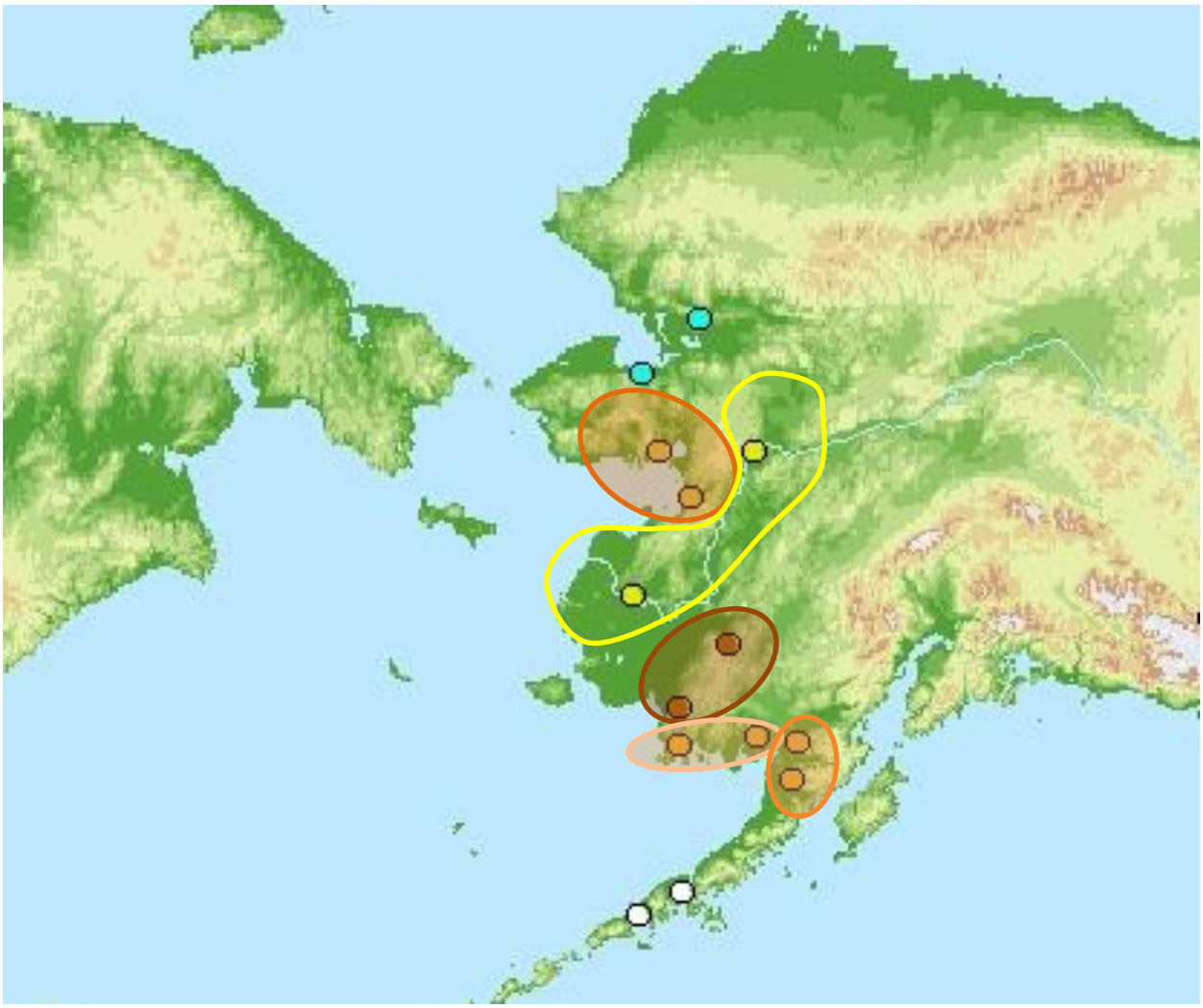

Figure 2.–Locations of chum salmon collections within western Alaska. The 5 regions within Coastal Western Alaska to be measured using overall F_{ST} are indicated by the ellipses.

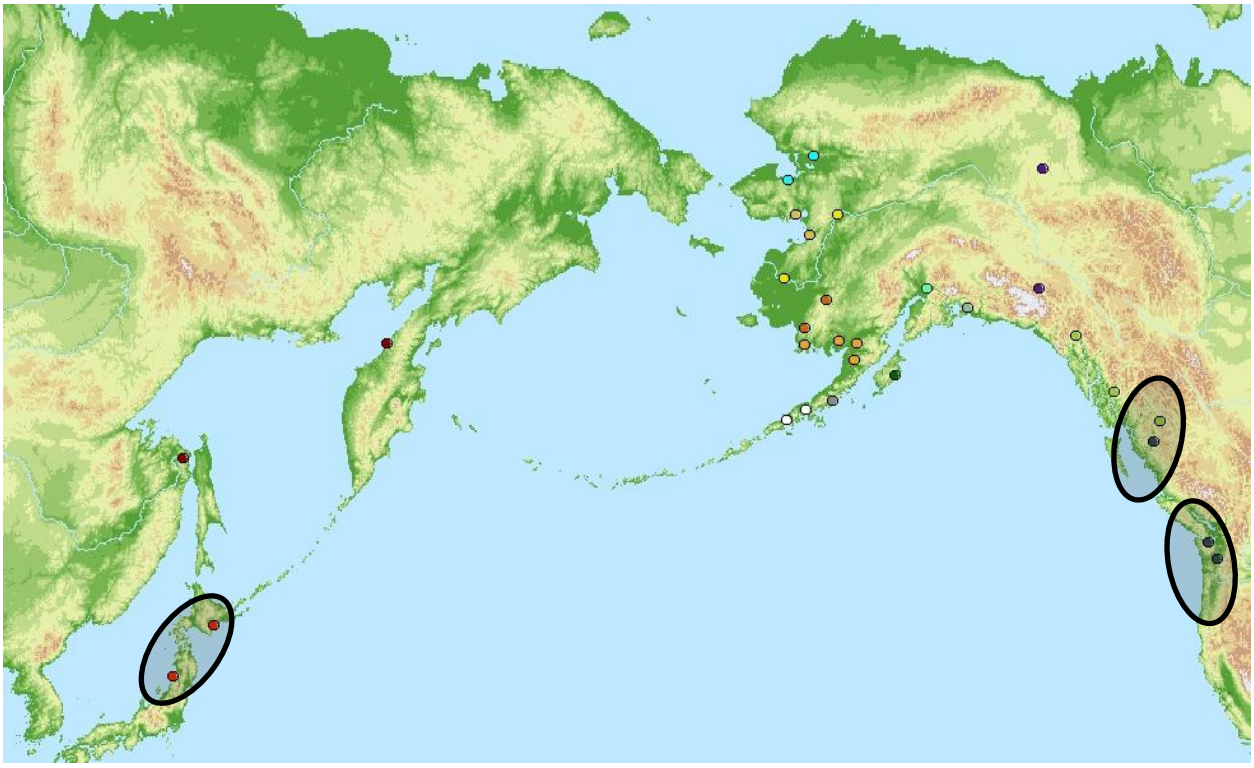

Figure 3.–Chum salmon populations used in SNP selection process highlighting the 3 population pairs (in ovals) of chum salmon chosen to measure F_{ST} within regions of interest to research groups outside of Alaska.

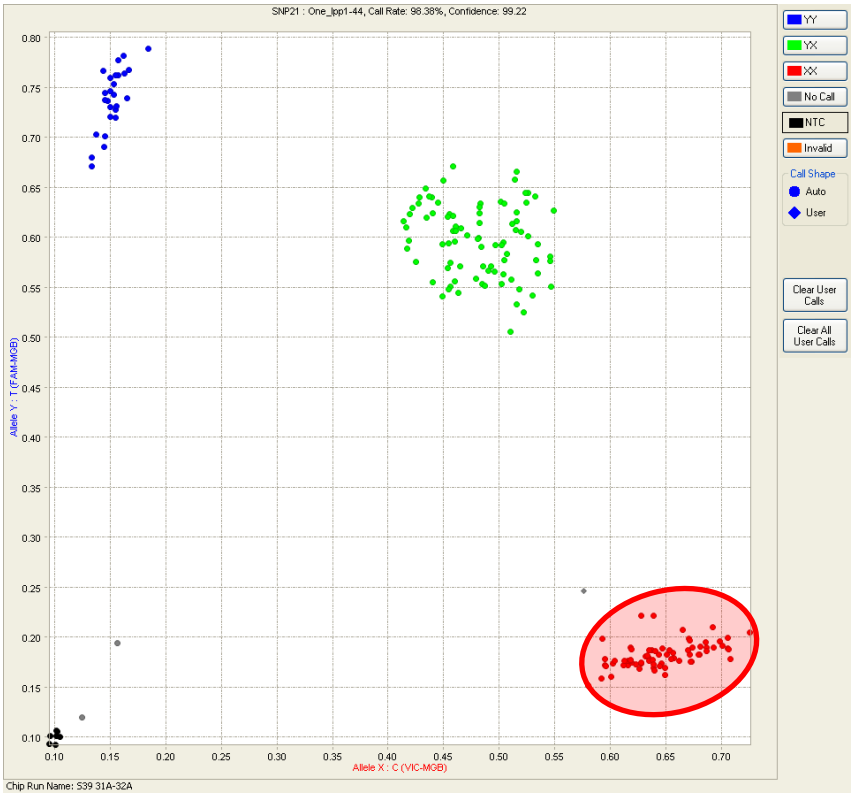

Figure 4.–Screen capture of a scatter plot from genotyping software. Each point represents a single fish. The 3 clusters represent each possible genotype (TT homozygote - blue, TC heterozygote - green, and CC homozygote - red). The size of the shaded area for the CC homozygote distribution is an indication of cluster tightness.

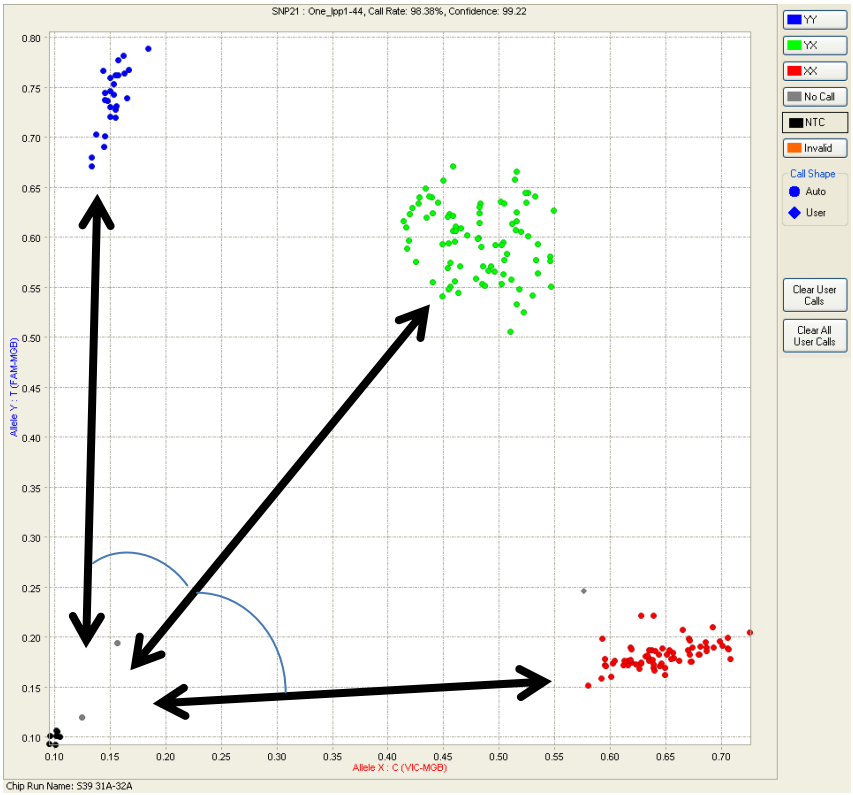

Figure 5.–Screen capture of a scatter plot from genotyping software. Each point represents a single fish. The 3 clusters represent each possible genotype (TT homozygote - blue, TC heterozygote - green, and CC homozygote - red). The angle between the double-ended arrows is an indication of cluster alignment.

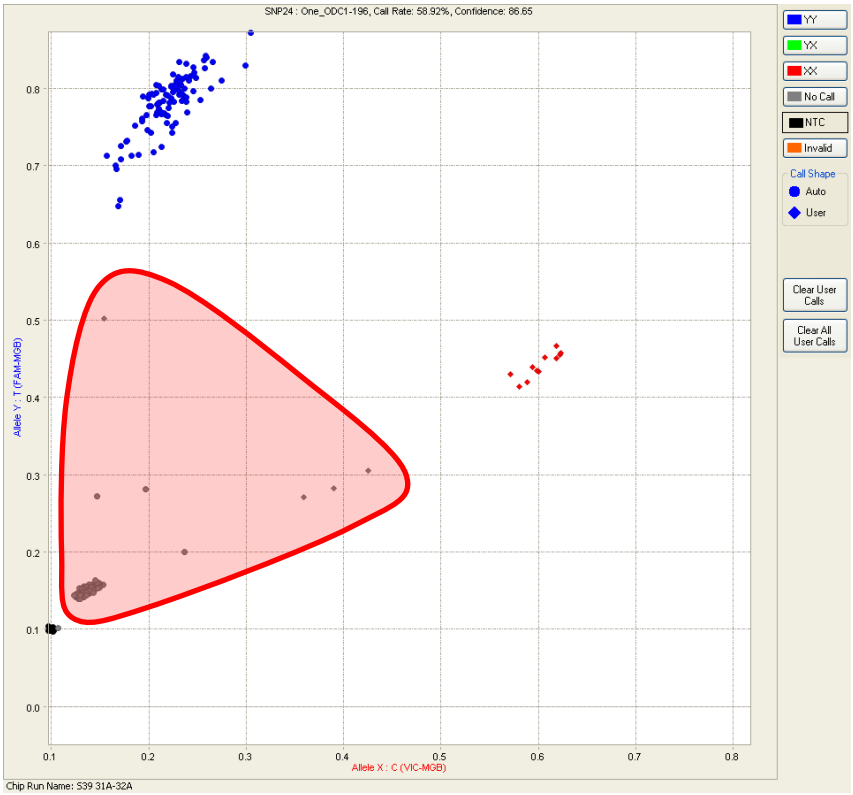

Figure 6.–Screen capture of a scatter plot from genotyping software. Each point represents a single fish. The 3 clusters represent each possible genotype (TT homozygote - blue, TC heterozygote - green, and CC homozygote - red). The red shaded area represents fish for which the assay failed.