Regional Information Report 5J12-19

Western Alaska Salmon Stock Identification Program Technical Document 10: Optimal Rate of Correct Assignment with Backward Elimination Locus Selection

by

James R. Jasper

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 5J12-19

WESTERN ALASKA SALMON STOCK IDENTIFICATION PROGRAM TECHNICAL DOCUMENT 10: OPTIMAL RATE OF CORRECT ASSIGNMENT WITH BACKWARD ELIMINATION LOCUS SELECTION

by
James R. Jasper 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.

James R. Jasper 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:
Jasper, J. R., 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.

TABLE OF CONTENTS

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, particularly to increase resolution among regional areas within Coastal Western Alaska (CWAK). This document discusses our proposal to determine the best set of loci to provide separation among reporting groups while taking advantage of potential synergy among loci. This analysis is proposed to provide 30% of the locus-selection weight, the most of any of the marker selection methods for choosing SNPs for the chum salmon baseline. The backward elimination locus selection (BELS) algorithm will be used to select marker panels to evaluate using the optimal rate of correct assignment (f_{ORCA}) to measure the marker set's ability to assign individuals back to the correct region. This will avoid the prohibitively slow analysis required to evaluate performance in the software BELS. We suggest that marker-selection applications with large numbers of populations and loci should employ the BELS algorithm for selecting marker panels to evaluate, but use the f_{ORCA} function to do the evaluation. We believe this method will improve the likelihood of providing resolution 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, backward elimination locus selection, BELS, optimum rate of correct assignment, f_{ORCA}

INTRODUCTION

As part of the locus selection process proposed for chum salmon in the Western Alaska Salmon Stock Identification Program (WASSIP), we propose using f_{ORCA} (Rosenberg et al. 2003; Rosenberg 2005) with backward elimination as one of the marker selection methods for choosing single nucleotide polymorphisms (SNPs) for the chum salmon *Oncorhynchus keta* genetic baseline (DeCovich et al. 2012). Results from this analysis are proposed to provide 30% of the locus-selection weight, the most of any analysis. The information measure, f_{ORCA}, returns the Optimal Rate of Correct Assignment (ORCA) for a particular locus set with respect to a specific baseline. At each iteration of the routine, a randomly drawn individual is assigned to a population for which its genotypic probability is a maximum.[1] We propose adapting f_{ORCA} to allow us to determine the best set of loci to provide separation among reporting groups taking advantage of potential synergy among loci. To do this we propose implementing a backward elimination algorithm similar to that described in the backward elimination locus selection (BELS; Bromaghin 2008). However, we opted not to use the program BELS because it is too time consuming. Even though the Gene Conservation Laboratory does proportional allocation (as does BELS) rather than individual assignment (as does f_{ORCA}), we feel that f_{ORCA} with backward elimination has merit under a Bayesian mixed stock analysis routine because it attempts to select a suite of markers that optimizes the genotypic probabilities of potential mixture individuals, and BAYES (Pella and Masuda 2001) uses these probabilities to stochastically assign the mixture individuals at each iteration.

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

Current f_{ORCA} Algorithm

While a closed form solution of f_{ORCA} is available (Rosenberg et al. 2003), it becomes impractical for large locus sets. Therefore, Rosenberg (2005) provided an iterative algorithm for estimating f_{ORCA}. This algorithm can be explained as follows:

1. Uniformly draw a population at random from the baseline.[2]
2. Randomly generate a multi-locus genotype based on the allele frequencies of the population chosen in the first step.
3. Assign that genotype to the population for which its genotypic probability is a maximum.
4. Repeat steps 1–3 10,000 times.
5. After repeating this process multiple times, f_{ORCA} is calculated as the proportion of times that the assignment in Step 3 is the same population drawn in Step 1.

While f_{ORCA} is typically used to evaluate how well a marker set can assign individuals back to the correct population, it could also be adapted for evaluating how well a marker set can be used to assign individuals back to the correct region. With this application the algorithm would be as follows.

1. Uniformly draw a population at random from the baseline.
2. Determine the region to which the population belongs.
3. Randomly generate a multi-locus genotype based on the allele frequencies of the population chosen in the first step.
4. Assign that genotype to the population for which its genotypic probability is a maximum.
5. Determine the region to which the assignment population belongs.
6. Repeat steps 1–5 10,000 times.
7. After repeating this process multiple times, f_{ORCA} is calculated as the proportion of times that the assignment in Step 5 is the same region drawn in Step 2.[3]

Backward Elimination Locus Selection Algorithm

Rosenberg's f_{ORCA} algorithm provides a means of evaluating the performance of a locus set, but it does not provide us with an algorithm for selecting sets of markers to evaluate. Rosenberg (2005) does provide 4 such algorithms and discusses the advantages and limitations of each: 1) Exhaustive evaluation, 2) Univariate accumulation, 3) Greedy accumulation, and 4) Maxmin accumulation.

One locus selection algorithm that Rosenberg failed to discuss is the method used in the BELS algorithm laid out by Bromaghin (2008). This algorithm has the advantages of being both simple to implement and it exploits synergies among loci. However, Bromaghin (2008) does not use f_{ORCA} to evaluate marker sets; rather he uses actual maximum likelihood mixed stock analysis and bootstrap simulations to evaluate performance in the software BELS. While we agree that

[2] This sentence is commented on in the section entitled "Technical Committee Review and Comments."
[3] This 7-step section is commented on in the section entitled "Technical Committee Review and Comments."

this is a relevant measure, unlike f_{ORCA}, it suffers from being prohibitively slow and may be biased in some circumstances (Anderson 2008).[4]

We suggest that marker selection applications with large numbers of populations and loci should employ the BELS algorithm for selecting marker panels to evaluate, but use the f_{ORCA} function to do the evaluation. For the purposes of WASSIP, we will use the correct assignment to region algorithm described above.

This would be accomplished by the following:

1. Start with entire set of L potential markers.
2. Create L subsets of L-1 markers by removing each marker, in turn, from full the set.
3. Evaluate f_{ORCA} on all L sub-sets using correct assignment to region.
4. Identify subset with maximum f_{ORCA}.
5. Record which locus was removed.
6. Return to Step 1 using the subset identified in Step 4 as the new full set of L-1 loci.

This process is continued until no markers remain. The loci can be ranked according to the order in which they were removed or scored according to their f_{ORCA} value.

This algorithm has been implemented in R for use with the chum salmon SNP selection process described in DeCovich et al. (2012).

The limitations of f_{ORCA} are: 1) it (likely) suffers from providing an optimistic rate of correct assignment, and; 2) spurious differences in allele frequencies can lead to falsely identifying some loci as influential. An extension of f_{ORCA} that may alleviate its limitations would be to implement a "leave-one-out" approach by which we randomly draw an individual from the ascertainment baseline, recalculate the allele frequencies without that individual, then assign the individual based on the recalculated allele frequencies. While more difficult to implement, this version may be a more viable solution. We are currently working on programming this extension.[5]

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.

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

REFERENCES CITED

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.

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

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 5J12-13, Anchorage.

Pella, J., and M. Masuda. 2001. Bayesian methods for analysis of stock mixtures from genetic characters. Fishery Bulletin 99(1):151–167.

Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard. 2003. Informativeness of Genetic Markers for Inference of Ancestry. American Journal of Human Genetics 73 (1421):1402–1422.

Rosenberg, N. A. 2005. Algorithms for Selecting Informative Marker Panels for Population Assignment. Journal of Computational Biology 12 (9):1183–1201.

TECHNICAL COMMITTEE REVIEW AND COMMENTS

General comments

In general the approach seems reasonable, but we have some specific comments as detailed below.

Minor comments

Page 1, 1st ¶, fourth sentence (note 1): "At each iteration of the routine, a randomly drawn individual is assigned to a population for which its genotypic probability is a maximum." How is this individual chosen? What is the pool of candidate individuals?

Page 2, (note 2): "Uniformly draw a population at random from the baseline." What exactly does this mean? Each population has equal weight, and then the draw is random?

Page 3, first sentence (note 4): "While we agree that this is a relevant measure, unlike f_{ORCA}, it suffers from being prohibitively slow and may be biased in some circumstances (Anderson 2008)." After "unlike f_{ORCA}", two attributes are listed but only one (being slow) is unlike f_{ORCA}. The bias described by Anderson et al. (2008) is equally applicable to f_{ORCA}. See below for more on this point.

Comments about bias and f_{ORCA}

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 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 in steps 1-7 (page 2, note 3) of Technical 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 Technical 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.

The final paragraph of Technical Document 10 (Page 3, note 5) 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.

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.