Western Alaska Salmon Stock Identification Program **Technical Document**:[1] 10

Title: Optimal Rate of Correct Assignment with backward elimination locus selection **Version:** 1.0

Authors: J. Jasper and W. Templin
Date: December 14, 2010

Introduction

As part of the locus selection process proposed for chum salmon in WASSIP, we propose using f_{ORCA} (Rosenberg et al. 2003; Rosenberg 2005) with backward elimination as one of the marker selection methods for choosing SNPs for the chum salmon baseline (Tech Doc 8). 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. 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 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 each iteration.

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. Randomly generate a multi-locus genotype based on the allele frequencies of the population chosen in the first step.

[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.

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.

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 four 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 Backward Elimination Locus Selection (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 this is a relevant measure, unlike f_{ORCA}, it suffers from being prohibitively slow and may be biased in some circumstances (Anderson 2008).

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 sub-sets 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 sub-set with maximum f_{ORCA}.
5. Record which locus was removed.
6. Return to Step 1 using the sub-set 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 Technical Document 8, "Chum salmon SNP selection process outline."

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.

Citations

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

Rosenberg, NA, LM Li, R Ward, & JK Pritchard. 2003. Informativeness of Genetic Markers for Inference of Ancestry. *American Journal of Human Genetics* 73 (1421):1402-1422

Rosenberg, NA. 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:

Line 13: "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?

Line 29: "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?

Line 63: "While we agree that this is a relevant measure, unlike *fORCA*, it suffers from being prohibitively slow and may be biased in some circumstances (Anderson 2008)." After "unlike *fORCA",* two attributes are listed but only one (being slow) is unlike *fORCA.* The bias described by Anderson et al. (2008) is equally applicable to *fORCA.* See below for more on this point.

<u>Responses to specific questions:</u>

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.