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

Title: Technical committee comments and review of the original proposal
Authors: WASSIP Technical Committee (B. Weir, R. Waples, and T. Quinn)
Date: September 2, 2008

Introduction

The Western Alaska Salmon Stock Identification Project (WASSIP) Technical Committee (TC) was convened during May, 2008 at the University of Washington in Seattle, WA. At that time the TC was presented by the Alaska Department of Fish and Game (Patricia Nelson, Eric Volk and Bill Templin) with an overview and plan for accomplishing WASSIP. During the Advisory Panel meeting in September 2008 the TC presented its comments on the project and the plan. The complete set of comments from the committee is provided below.

Committee comments (unedited and unabridged)

Western Alaska Salmon Stock Identification Project (WASSIP)
Technical Committee comments

2 September 2008

As requested, we are providing these written comments to the WASSIP Advisory Panel in advance of the 24 September 2008 meeting in Anchorage. We first discuss the 2008 AYK/SSI proposal and provide some general comments on related issues, before turning to a series of specific questions we were asked to address.

2008 AYK/SSI proposal

Although this proposal (which was not ultimately funded by AYK) covers only a part of the overall project, it is the most recent and most detailed description of the proposed methodology, so this is a useful document for focusing technical comments. This proposal emphasizes stock identification of chum salmon in western Alaska, but most of the technical issues apply more broadly to sockeye salmon and other geographic areas.

Performance measures

The proposal clearly outlines a number of quantitative performance measures, which should facilitate evaluation of success of the project.

[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.

A comprehensive, standardized baseline for chum salmon. We believe that the project leaders have the necessary skills, facilities, and resources to accomplish this objective.

Genetic Stock Identification (GSI) estimates within 5% of true value 90% of time. In general, this is not an unreasonable goal, but whether it can be accomplished in all cases will depend on a variety of factors, including sample sizes from source populations and the mixed fisheries, the degree to which these baseline and fishery samples are representative, and the magnitude of underlying genetic differences among populations. This performance measure should be clarified to reduce ambiguity regarding the meaning of "within 5%". For example, if the true mixture fraction for a given stock is 10%, does this performance measure call for estimates that fall in the range (5% to 15%) 90% of the time, or does it require that the estimates fall within the range (9.5% to 10.5%)? [5% of 10 is 0.5.] The latter criterion is obviously much more difficult to accomplish.

Detect a 1% contribution 99% of the time. This is a very ambitious goal. It should be recognized that whether or not this can be achieved is probably more of a statistical problem than a genetic problem. To illustrate, assume for the moment that there is no uncertainty in the genetic analyses, so every fish can be assigned without error to its true stock of origin. A 1% contribution from a particular population of interest (say population A) can be detected only if the sample from the fishery actually contains at least one fish from that population. If we assume that the overall fishery is large enough that the GSI sample can be modeled as a binomial sample, then the problem can be set up as follows. Each fish in the fishery sample has a 1% probability of coming from population A, or, conversely, a 99% probability of coming from a population different than A. If a sample of N fish is taken for analysis, the chances that zero fish in the sample come from stock A (even though stock A is present in the overall fishery at 1%) is 0.99^N. The goal is to have N be large enough that this term (which represents that chance of not detecting presence of stock A) is 1% or less. Using this equation, it is easy to show that the sample size from the fishery must be over 458 fish before the probability of not detecting stock A falls below 1%. So, even with ideal genetic resolution, large samples are needed to meet this performance measure. Furthermore, this simple exercise only shows the probability of failing to detect a *single* stock that contributes at low levels. More generally, managers are probably interested in whether a given mixed-stock fishery analysis fails to detect *any* low-contributing stocks. If multiple stocks contribute at low frequency to a fishery, then the problem becomes more complex and additional increases in sample size are needed to ensure that *none* of the low-contributing stocks are missed. (For example, with two stocks contributing 1%, the sample size would need to be at least 535 to be 99% sure of detecting both.) Furthermore, low-contributing stocks might differ from more abundant ones in run timing, size, spatial distribution, etc. – all of which will make it difficult to ensure a representative sample from the fishery. Therefore, we question whether it is realistic to expect to meet this performance measure. We want to emphasize again that this is more a statistical issue and a sampling issue than a genetic one; the same conclusion would apply to a method (such as coded-wire tags) that can unequivocally identify individuals to population of origin. If some of the low-contributing stocks are genetically similar to other populations that are larger contributors, the whole problem becomes even much more difficult.

Analyze numerous fishery samples. We believe that the project leaders have the necessary skills, facilities, and resources to accomplish this objective.

Methods

The project outlines a number of impressive quality control measures for collection of genetic data. For example, double scoring all genotypes independently by two researchers should ensure a high degree of accuracy. Similarly, reanalyzing a random 8% of samples for all loci should help minimize various types of errors associated with sample handling and data recording. Nevertheless, the proposal somewhat overstates the case for SNP technology. Although SNPs have some definite advantages over microsatellites in terms of repeatability and susceptibility to scoring errors, it is not really accurate to say that SNPs "are automatically standardized across laboratories." Recent experience with human whole-genome SNP scans has revealed the need for considerable attention to data quality control. The project leaders will need to be aware of issues of batch effects on missingness rates, for example, and will need to include standard and blind duplicates to monitor quality. With the relatively small numbers of SNPs proposed, it will be valuable to study the cluster plots of called genotypes for each SNP.

The authors of the proposal estimate that data from 48 independent SNPs will be adequate for stock composition, but it is not clear how they arrived at this number. Certainly this number has been shown to be sufficient in some human studies (Paschou et al., 2008), but these highly discriminating markers were found only after whole-genome scan data were available. Whether any particular number of loci will be adequate depends on a number of variables that were not quantitatively analyzed in the proposal. This point is discussed further below under "Reviewer comments."

The authors briefly discuss the question, "What is a population?" and offer the following definition: "A group of individuals of the same species living in close enough proximity that members of the group can potentially mate with any other member." However, as pointed out in the cited reference (Waples and Gaggiotti, 2006), this (and most other definitions one can find in the literature) is not quantitative and is not operational; that is, you could not give this definition to separate groups of scientists and expect that they would arrive at the same answer, even if they were working with the same data. As one of the major goals of this project is to define population structure of Western Alaska salmon, this topic needs more attention. Note that although scientific input on this question is important, it cannot be answered by science alone. For example, consider samples from two different areas that might or might not represent different "populations" or "stocks." Scientific methods can test the null hypothesis that both samples were drawn from a single population, and can also estimate the true underlying genetic differences between the areas. But science alone cannot say whether a particular level of difference is large enough to represent a separate "population"; that determination should be made in the context of the management goals one is trying to achieve.

The Cavalli-Sforza and Edwards' chord distance has been widely used but it has one important drawback: it does not include a correction for sample size, so all estimates are upwardly biased. This will generally not be a problem for studies of population differentiation, provided either of two conditions are met: 1) the units being compared are different species or strongly divergent populations, in which case this source of bias will be small compared to the overall genetic distance; or 2) sample sizes are identical, so all genetic distances are inflated to the same degree. In this study, neither condition is met. With closely related populations, the sampling bias can be as large or larger than the true signal; furthermore, small samples will tend to look like genetic outliers because their allele frequencies are distorted most by sampling error. The authors also mention using pairwise F_{ST} values, which is a better choice provided an unbiased estimator is used, such as Weir and Cockerham (1984). Once again, better estimates will result from higher numbers of SNPs.

The proposal mentions a number of newer software programs for population genetic data analysis, and these are generally appropriate. However, little detail is provided regarding exactly how these programs will be used to accomplish program objectives.

The project leads have considerable experience in salmon GSI and the available software programs. However, it was not entirely clear from the proposal whether they plan to use the method that Koljonen et al. (2005) found to produce the greatest precision in GSI estimates; this method combines both individual assignments and mixture modeling in an iterative fashion.

Simulations are routinely used to evaluate accuracy and precision of GSI estimates, and they will be an important part of this study. The primary focus will be on what is often referred to as “100% simulations”, in which a simulated ‘mixture’ is actually composed entirely of individuals from one population or one population group. Analysis of the actual allocations thus allows identification of directional biases associated with particular populations. Although this procedure is widely used within the GSI community and can be very informative, it should be recognized that the performance of GSI in practical situations involving complex mixtures of many populations might differ in important ways from the behavior demonstrated in 100% simulations. Therefore, it would be useful to conduct some more complex simulations to verify that the 100% simulations are producing results of practical relevance.

GSI simulations can be of two general types: 1) those that model mixtures involving hypothetical populations with specified levels of divergence (e.g., as measured by F_{ST} values or genetic distance), and 2) those that model mixtures of actual salmon populations from which baseline samples have been collected. The latter type of simulations can provide information of direct practical relevance but are tricky to implement because of sampling error in the baseline samples. Because of this factor, allele frequency differences between samples from baseline populations will, on average, be larger than the true underlying differences between populations. When simulated mixtures are created using the baseline allele frequencies, the simulated fish in the mixture are more genetically divergent than are fish from the real populations, and this can cause an overly optimistic assessment of precision. Furthermore, this effect is most pronounced with closely related populations, for which the sampling bias can be as large or larger than the true signal. Anderson et al. (2008) showed that this problem is not solved by resampling the baseline populations. Fortunately, a simple modification is available that fixes this bias problem, and it is implemented in two freely available software packages: ONCOR, a Windows-based program: http://www.montana.edu/kalinowski; and *gsi_sim*, with a command line interface suitable for Unix-like operating systems: http://swfsc.noaa.gov/staff.aspx?id=740. Although it was not apparent from the proposal, based on discussions with the project leads we believe they understand this potential problem and have a strategy to deal with it.

Reviewer comments on the AYK/SSI proposal

Reviewers of the AYK/SSI proposal acknowledged that it addressed high-priority questions within the region, but they had two major criticisms. First, they felt that the proposal was not well coordinated with existing, multi-agency efforts at stock identification and mixed-stock fishery analysis in Alaska. Although this is an important issue, it is beyond the expertise of the Technical Committee to comment on. The second major comment was that, because of the biology of chum salmon, many (perhaps most) populations are characterized by relatively low levels of genetic differentiation, and as a consequence it is not realistic to expect large increases in accuracy and precision of GSI estimates with development of new markers. We believe this concern has some merit. To the extent that this comment is accurate for the geographic areas considered in this project, the mixed-stock fishery problem will become very challenging. In theory, arbitrarily small (but real) differences among populations can be successfully resolved if one can collect arbitrarily large amounts of data (many individuals scored for many genetic markers). In practice, this will not often be feasible; in addition, as the true genetic signal becomes weaker and weaker, various sources of noise in the analysis (non-random sampling, data scoring or recording errors) assume a relatively larger importance. It seems quite possible, therefore, that in at least some cases the genetic differences among populations will be too small to allow satisfactory resolution of units that managers would like to be able to distinguish.

Other comments

Ascertainment bias

Ascertainment bias refers to biases that result from the process of discovering or ascertaining new genetic markers. Although this issue can apply to any type of marker, SNPs are particularly prone to ascertainment bias, and this topic has been the focus of a number of recent publications. Two types of ascertainment bias are relevant. Within-population biases arise when the markers that are discovered are not representative of all markers with respect to allele frequency distribution. This occurs, for example, if the markers used have, on average, more intermediate allele frequencies (and hence higher heterozygosities) than would markers selected at random. Among-population biases can arise when markers are selected for characteristics exhibited in a single population (or a few nearby populations) and then applied to populations from different areas. For example, a suite of markers selected specifically for the ability to discriminate stocks in geographic region X might have low levels of variability and little power to distinguish stocks in region Y.

Based on the description in the AYK/SSI proposal of the SNP discovery process, it appears there are substantial opportunities for both types of ascertainment bias. According to the proposal, a single chum salmon from Susitna River will be used for SNP discovery. This means that SNP markers that are at intermediate frequencies (around 50%) have a much higher chance of being detected as polymorphic than do SNPs where the alternate allele occurs at low frequency. (The probability that one individual is heterozygous and a SNP detected is 50% when the population allele frequencies are 0.5 but the probability drops to 18% when the minor allele has a frequency of 0.1.) Similarly, SNPs that are highly variable in the Susitna River might not be variable throughout the range of chum salmon in Alaska, and markers that could be particularly useful in other areas might be missed by focusing discovery on a single population.

Are these probable sources of bias likely to be problem for the WASSIP? That depends. GSI models do not depend on any particular assumptions about heterozygosity levels or distributions of allele frequency. Therefore, there might be little or no direct effect of these biases on GSI estimates, apart from some loss of efficiency caused, for example, by having to screen many markers discovered in the lone Susitna chum salmon to find a few that prove to be informative in other geographic regions. However, documents we have reviewed indicate that the goals of the project include broader objectives such as describing population structure and levels of connectivity among populations. As pointed out by one of the reviewers, these applications are more sensitive to ascertainment bias, and the ability to accomplish those objectives could be compromised in some situations. Therefore, we recommend that the project leaders carefully consider this issue and find an appropriate balance between efficiency in SNP discovery and related ascertainment biases that could affect the ability to accomplish some program goals. One option to consider is to broaden the number of individuals used for SNP discovery. Although this creates some technical challenges for the 454 sequencing process, it can reduce both types of bias. For example, in a large European Union project (designed to allow tracing of fish products in the marketplace) that one of us (RSW) is involved with, two individuals from each of four geographic areas are used in the discovery panel.

Temporal stability

GSI estimates are based on the implicit assumption that baseline samples represent the population allele frequencies over the entire period during which fisheries are analyzed. The extent to which this is true will depend on the number of years involved and the rate of change of allele frequencies by genetic drift in the baseline populations. The rate of genetic drift is inversely related to effective population size (N_e), and previous work has showed how standard population genetics theory of genetic drift and effective size can be modified to account for the life history of Pacific salmon (Waples 1990a). In the current project, Table 1 (in the 2008 proposal) indicates that the SNP baseline for chum salmon includes samples collected over a period of nearly 20 years (1989-2006) and these will be used to analyze fisheries in 2006 and subsequent years. This represents as much as 4-5 salmon generations, and during this period alleles at moderate frequencies can drift on average by 2-3% in populations with N_e = 1000 and by 5-10% in populations with N_e = 100. Whether this source of random noise will be consequential for GSI estimates depends on the relative magnitude of these temporal changes compared to the genetic differences among populations. If the latter differences are large, then this source of noise might pose little or no practical problem. On the other hand, for closely related populations, random temporal changes can be comparable in size to differences among populations, in which case resolution of mixed-stock fisheries becomes very difficult.

This important topic merits more thorough consideration and quantitative evaluation. Table 1 indicates that samples from more than one year are available for perhaps 10% of the populations. At a minimum, these should be evaluated to determine 1) the magnitude of allele frequency change over time; and 2) the relative magnitude of temporal and geographic differences in allele frequency. Whenever possible, it would be prudent to resample populations sampled over a decade ago to ensure that the baseline data still reflect contemporary genetic profiles for these populations. Waples (1990b) provides some guidance about how to deal with temporal variation in GSI analyses.

Miscellaneous

The results of Smith and Seeb (2008) are a little surprising. In Table 1 of their 2008 paper they show F_{ST} estimates – there is little point in showing locus-specific values because of the high sampling variance of such estimates. Estimates should be combined over loci to reduce this variance. The overall agreement between microsatellite and SNP values is striking and not a common finding.

The various multivariate techniques for population distinction do not mention principal component analysis. Much attention has been given recently to PCA, especially for individual-level analyses of Patterson (2006) and Price (2006) and by the recent work of Paschou (2008) on selecting a small subset of PCA-correlated SNPs for population subdivision. Are PCA methods worth pursuing for WASSIP?

Reviewer 3 mentions the lack of error bars in Figure 1. The criticism is probably too harsh given Figures 3 (with error bars) and 4 (without) in Smith and Seeb (2008). This reviewer also points to possibly high gene flow between local populations. Does that matter? Is regional attribution sufficient for the study?

Responses to specific questions (posed by Eric Volk in email to TC dated 16 June 2008)

Is the sampling design reasonable to achieve our goal of estimating stock compositions in these fisheries?

In general, the scale of the project is impressive and should produce a great deal of valuable information. However, as discussed above, a number of biological and analytical factors conspire to make the problem of obtaining reliable GSI estimates for closely related populations a challenging one. Before the adequacy of the study design can be rigorously evaluated, more work is needed to quantify how power of discrimination is affected by sample sizes (markers, baselines fisheries), the representativeness of the samples, and underlying differences among populations.

Are methods for sample collection, DNA extraction, genotype assays, data acquisition and quality control appropriate and robust?

This project has established high standards for quality control throughout the various steps of the project. However, the processes of standardizing SNPS scores and ensuring accuracy are not as simple as implied in the proposal.

Given the performance of existing baselines and sample demands of the study, are SNPs an appropriate DNA marker choice?

SNP genotyping is rapidly becoming cheaper and is consistent with imminent whole-genome sequencing. The Smith and Seeb (2008) paper suggests SNPs can be effective for relatively closely related chum salmon populations, although reviewers of the AYK/SSI proposal were not convinced that use of SNPs would substantially improve resolution. The experience of human geneticists suggests that, ultimately, SNPs will improve resolution. The number of SNPs needed to ensure this increase, however, will need to be answered empirically.

Are there specific study elements that require modification?

Some topics that were mentioned above merit careful evaluation to determine whether modifications are needed: 1) expectations for detecting low-contributing stocks; 2) consequences of ascertainment bias; 3) effects of temporal variability; 4) developing an operational definition of “population” or “stock”. In addition, the analytical plans could be expanded to include relatedness and an evaluation of PCA.

Final comment: It is apparent that, apart from the many logistical challenges, success of this project will depend to a large degree on whether GSI can adequately resolve the contributions from stocks that are genetically similar. Numerous comments above detail the challenges associated with doing that. Another important factor to consider in this context is that closely related populations can sometimes be grouped together in a management unit or reporting unit. In that case, although it might be impossible to provide reliable estimates for individual stocks within a reporting group, it still might be possible to provide highly accurate and precise estimates for the reporting group as a whole. It is clear that some of this is contemplated as part of the WASSIP project, but the information available to the Technical Committee did not allow us to evaluate how likely this is to completely resolve the problem. We expect that cases might arise where managers want to distinguish the contributions of two or more populations that are too similar genetically to allow reliable discrimination. It should be possible to identify these situations from careful analysis of the baseline collections and comparing those results with proposed reporting units. We recommend that this exercise be conducted as soon as feasible so that expectations for the project can be adjusted (if necessary) in a timely fashion.

References cited

Anderson, E. C., R. S. Waples, and S. T. Kalinowski. 2008. An improved method for predicting the accuracy of genetic stock identification. Can. J. Fish. Aquat. Sci. 65:1475–1486.

Koljonen, M.-L., Pella, J.J., Masuda, M. 2005. Classical individual assignments vs. mixture modeling to estimate stock proportions in Atlantic salmon (*Salmo salar*) catches from DNA microsatellite data. Can. J. Fish. Aquat. Sci. 62:2143-2158.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron,W., Mahoney, M.W., and Drineas, P. 2007. PCA-correlated SNPs for structure identification in world wide human populations. PLoS Genetics 3(9) e160.

Patterson, P., Price, A.L., and Reich, D. 2006. Population structure and eigenanalysis. PLoS Genetics 2:(12) e190.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E, Shadick, N.A. and Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: (8).

Smith, C.T., and L.W. Seeb. 2008. Number of alleles as a predictor of the relative assignment accuracy of short tandem repeat (STR) and single-nucleotide-polymorphism (SNP) baselines for chum salmon. Trans. Am. Fish. Soc. 137:751–762.

Waples, R.S. 1990a. Conservation genetics of Pacific salmon. II. Effective population size and the rate of loss of genetic variability. J. Heredity 81:267-276.

Waples, R. S. 1990b. Temporal changes of allele frequency in Pacific salmon populations: implications for mixed-stock fishery analysis. Can. J. Fish. Aquat. Sci. 47:968-976.

Waples, R. S., and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology 15:1419-1439.

Weir, B.S., and Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

Robin S. Waples (Chair)
Northwest Fisheries Science Center, NOAA, Seattle

Bruce S. Weir
University of Washington, Seattle

Thomas P. Quinn
University of Washington, Seattle