Western Alaska Salmon Stock Identification Program **Technical Document**:[1] 7

Title: Regional Allocation Model **Version:** 1.0
Authors: J. Jasper, C. Habicht, A. R. Munro, and W. Templin
Date: April 30, 2010

Introduction

Mixed stock analysis methods for estimating stock (population) compositions in fisheries have evolved over time from conditional maximum-likelihood (Fournier et al. 1984) to Bayesian (Pella and Masuda, 2001) approaches. The Pella-Masuda model (a Bayesian approach; Pella and Masuda, 2001) has been the "gold standard" since 2001. In these methods, however, bias is inevitable because the estimation of the stock proportions is constrained to be non-negative and sum to one, meaning that rare or absent stocks in the mixture are overestimated while common stocks are under estimated (Pella and Milner, 1987). Stocks are usually grouped into regional stock groupings (regions) for reporting.

Recent observations in our laboratory indicate that disproportionate numbers of stocks within a region can lead to significant bias in regional composition estimates when regional stock structure is shallow. We have observed that regions represented by large numbers of stocks seem to acquire higher misallocations than regions represented by fewer stocks (Figure 1). This bias can be reduced at the regional level by grouping stocks with similar genetic attributes into regions, then summing estimated proportions across stocks within the regions (Wood et al. 1987). Here we present a rationale for why we think the observed non-uniform bias occurred and a method that appears to improve allocation at the regional level as well as distribute the misallocation more evenly among regions.

In the Pella-Masuda model, the data augmentation algorithm is used to generate from the posterior distribution the stock identities of each of the mixture individuals, and then generate the stock proportions and baseline allele frequencies based on summaries of these identities. At each cycle of the algorithm, the stock identity of mixture individual *m* is stochastically assigned to stock *i* with probability proportional to the product of stock *i*'s contribution to the mixture and the relative frequency of individual *m*'s genotype in stock *i*. This means that individual *m* has a finite probability of belonging to each and every stock in the baseline. We will refer to these probabilities as the identity probabilities.

The chances that individual *m* is assigned to the correct stock at a particular iteration is a function of not only the genetic distinction of its stock, but also, theoretically, the number of stocks in the baseline. Fortunately, fisheries managers often are not interested in the proportion of individual stocks, but rather in

[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.

the contribution made by all stocks within regions. If the stocks within a region are genetically more similar to each other than to stocks in other regions (strong regional structure), then the chances of correctly assigning an individual to a stock within the correct region each cycle greatly improves estimation (Wood et al. 1987). However, with weak regional structure, the chances of assigning an individual to a stock within the correct region may be significantly influenced by the number of stocks in each region. This may be because the probability of assigning an individual to a particular region is the sum of the identity probabilities across all the stocks in the region, such that adding stocks adds probability. If the amount of misallocation to a region is a function of the number of stocks within that region, an inherent non-uniform bias in regional contribution estimates can occur simply due to differing numbers of stocks among regions.

The purpose of this paper is to illustrate that unequal numbers of stocks among regions leads to unequal biases in misallocation and to determine if a new analytical method may mitigate this bias. We anticipate an upward misallocation bias toward regions that are represented by larger numbers of stocks than regions represented by fewer stocks using the Pella-Masuda model. We present a new analytical model that appears to diminish this bias.

Methods

We considered three methods to examine the assertion that unequal numbers of stocks within regions do not affect bias in misallocation. We selected baseline data for chum salmon stocks from Western Alaska. These data were chosen because these stocks represent weak regional structure (Figure 5).

The first two methods use the Pella-Masuda model but differ in how the priors are assigned. The first method is the widely used True Flat Prior (TFP; Pella and Masuda, 2001). This model provides no *a priori* information about the regional structure and gives an equal prior “count” of $1/C$ to each of the stocks in the baseline, where C is the number of stocks. This is the model that provided the recent observations in our laboratory that suggested that disproportionate numbers of stocks within a reporting group can affect the regional composition estimates.

The second method, termed the Regional Flat Prior (RFP), is a method currently in use at ADF&G’s Gene Conservation Laboratory (Dann et al. 2009). The structure of the prior for stock proportions is an *ad hoc* alternative to the TFP. Under the RFP, for each of the stocks within the gth region, we give a prior “count” equal to $1/G/C_g$, where G is the number of regions and C_g is the number of stocks within the gth region. Therefore, equal prior “count” is given to each region, but the prior “count” given to a stock is dependent upon the number of stocks within its region.

The third method, termed the Regional Allocation Model (RAM), is currently under development at the Gene Conservation Laboratory. This model is very similar to the Pella-Masuda model in that it is based on the data augmentation algorithm that alternates between generating of the parameters of the model. The difference is that in the RAM, we first generate the regional identity of each individual, and then produce regional contributions based on summaries of these regional identities. For individual m, the regional identity probability of belonging to region g is proportional to region g’s contribution to the mixture times a weighted average relative frequency of individual *m*'s genotype across all C_g stocks within the region. The weights are simply the within-region stock proportions, and they sum to one. Because the weights do sum to one, the genetic component of the regional identity probabilities remain on the same scale regardless of the number of stocks within the region, which should presumably moderate the non-uniform bias due to the unequal distribution of stocks among the regions. There is actually a second stage to the data augmentation algorithm in which, after an individual is assigned to a region, it is then allocated to a stock within that region. This is done exactly as is done in the Pella-Masuda model except that it is done with respect to a baseline that is reduced to only that region.

General Bayesian Methods

For estimating parameters $\boldsymbol{\theta}$ from data $\boldsymbol{X}$ using Bayesian methods, we aim at the evaluation of the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{X}) = L(\boldsymbol{X}|\boldsymbol{\theta})\ P(\boldsymbol{\theta})/m(\boldsymbol{X})$, where $L(\boldsymbol{X}|\boldsymbol{\theta})$ is the likelihood of the data given the parameters, $P(\boldsymbol{\theta})$ is the prior distribution of the parameters, which must be specified, and $m(\boldsymbol{X})$ is the constant marginal distribution of the data. From this distribution, summary statistics for $\boldsymbol{\theta}$ can be derived. However, these distributions are rarely soluble in closed form for multidimensional parameter vector $\boldsymbol{\theta}$, and we must rely on drawing samples from it via a Gibbs sampling routine, from which the summary statistics can be calculated. For mixed stock analysis, $\boldsymbol{\theta}$ represents the stock proportions and the baseline allele frequencies while $\boldsymbol{X}$ corresponds to the mixture genotypes and the baseline allele counts. As mentioned previously, a prior distribution must be specified for the parameters. In the forthcoming models, the mathematically convenient Dirichlet distribution is used for the stock proportions as well as the baseline allele frequencies. A Dirichlet distribution with parameter vector $\boldsymbol{\lambda}$ is a distribution on a vector $\mathbf{W}$ whose sum is constrained to one. It has the form:

$$P(\boldsymbol{W}|\boldsymbol{\lambda}) = \frac{\Gamma(\sum_{i=1}^{n}\lambda_i)}{\prod_{i=1}^{n}\Gamma(\lambda_i)}\prod_{i=1}^{n} W_i^{\lambda_i}$$

The Pella-Masuda Model

We denote the count of the *j*th (j=1,2,…,J_d) allele of the *d*th (d=1,2,…,D) locus for mixture individual *m* as x_{mdj}, and let $\boldsymbol{X}_m$ signify the entire multi-locus genotype for this individual. The array $\boldsymbol{X}$ represents the multi-locus genotypes for all *M* mixture individuals. Similarly, we let y_{idj} denote the count of the *j*th allele for the *d*th locus of the *i*th baseline stock, and $\boldsymbol{Y}$ denotes the entire baseline. This describes the data.

To describe the parameters, let the stock proportion for the *i*th stock be denoted as P_i, and let $\boldsymbol{P}$ be the vector of all stock proportions. We place a Dirichlet prior distribution on the stock proportions with prior parameters $\boldsymbol{\alpha}$, where α_i is determined by our choice of prior structure discussed earlier (RFP or TFP).

We let q_{idj} denote the relative frequency of the *j*th allele for the *d*th locus in the *i*th baseline stock and let $\boldsymbol{Q}$ denote the entire array of baseline relative frequencies. We place a Dirichlet prior distribution on $\boldsymbol{Q}_{id}$ with prior parameters $\boldsymbol{\beta}_d$, where $\beta_{dj} = 1/J_d$, with J_d being the number of alleles for locus *d*.

Finally, let z_{mi} be the stock identity for the *m*th mixture individual in the *i*th stock, where z_{mi} is equal to one if individual *m* belongs to the *i*th stock and zero otherwise. We denote $\boldsymbol{Z}_m$ as the vector of stock identities for individual *m*, and ***Z*** as the matrix of stock identities for the entire mixture. We place a multinomial prior on $\boldsymbol{Z}_m$ with size 1 and probabilities equal to the stock proportions ***P***.

The genotypic likelihood of the *m*th individual would be greatly simplified if we knew the stock identity of that individual. In other words, if $z_{mi} = 1$, then the likelihood of observing individual *m* is simply the relative frequency of this individual’s multi-locus genotype in the *i*th stock, which we denote by $f(\boldsymbol{X}_m|\boldsymbol{Q}_i)$, where:

$$f(\boldsymbol{X}_m|\boldsymbol{Q}_i) \propto \prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{idj}^{x_{dj}}$$

Because $z_{mi'} = 0$ for all $i' \neq i$, the full genotypic likelihood may be expressed as:

$$L(\boldsymbol{X}|\boldsymbol{Q},\boldsymbol{Z}) = \prod_{m=1}^{M}\prod_{i=1}^{C} f(\boldsymbol{X}_m|\boldsymbol{Q}_i)^{z_{mi}}$$

In addition to the genotypic data, we need to consider the likelihood of the baseline data, which can be written as:

$$L(\boldsymbol{Y}|\boldsymbol{Q}) \propto \prod_{i=1}^{C}\prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{idj}^{y_{idj}}$$

The full likelihood, $L(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Q},\boldsymbol{Z})$, is simply the product of these two components.

Multiplying this likelihood by the prior distributions leads to the following posterior distribution:

$$\begin{aligned}
\mathrm{P}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y}) &\propto L(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Q},\boldsymbol{Z})\mathrm{P}(\boldsymbol{Z}|\boldsymbol{P})\mathrm{P}(\boldsymbol{P}|\boldsymbol{\alpha})\mathrm{P}(\boldsymbol{Q}|\boldsymbol{\beta}) \\
&\propto \left(\prod_{m=1}^{M}\prod_{i=1}^{C} f(\boldsymbol{X}_m|\boldsymbol{Q}_i)^{z_{mi}}\right)\left(\prod_{i=1}^{C}\prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{idj}^{y_{idj}}\right) \\
&\times \left(\prod_{m=1}^{M}\prod_{i=1}^{C} P_i^{z_{mi}}\right)\left(\prod_{i=1}^{C} P_i^{\alpha_i}\right)\left(\prod_{i=1}^{C}\prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{idj}^{\beta_{dj}}\right)
\end{aligned}$$

The benefit of using the chosen prior distributions is that the conditional posterior distribution for each of the parameters given the data and the remaining parameters is of the same form as the prior distribution (conjugacy). This property makes them easy to sample from within a Gibbs sampler, which proceeds as follows: first, starting with initial values for ***P*** and ***Q***, we draw stock identities for each of the mixture individuals from:

$$\boldsymbol{Z}_m|\boldsymbol{P},\boldsymbol{Q},\boldsymbol{X}_m \sim \text{multinomial}\left(1, \left\{\frac{P_i f(\boldsymbol{X}_m|\boldsymbol{Q}_i)}{\sum_{k=1}^{C} P_k f(\boldsymbol{X}_m|\boldsymbol{Q}_k)}\right\}_{i=1,2,\ldots,C}\right)$$

Next, given these stock identities, ***P*** is drawn from:

$$\boldsymbol{P}|\boldsymbol{Z},\boldsymbol{\alpha} \sim \text{Dirichlet}\left(\left\{\sum_{m=1}^{M} z_{mi} + \alpha_i\right\}_{i=1,2,\ldots,C}\right)$$

Finally, for each stock and for each locus, we generate $\boldsymbol{Q}_{id}$ from:

$$\boldsymbol{Q}_{id}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\beta} \sim \text{Dirichlet}\left(\left\{\sum_{m=1}^{M} \mathrm{z}_{mi}\mathrm{x}_{mdj} + \mathrm{y}_{idj} + \beta_{dj}\right\}_{j=1,2,\ldots,J_d}\right)$$

This process is repeated for several thousand iterations, typically with multiple chains starting from different initial values, and the first few thousand iterations are discarded as “burn-in” to remove the influence of the initial values. Multiple chains are run to assess convergence via the Gelman-Rubin shrink factor (Gelman and Rubin, 1992). By convergence, we mean convergence in distribution rather than convergence to a point.

Regional Allocation Model

The data for this model are exactly the same as for the Pella-Masuda model, except the baseline is framed within a hierarchy in which regions are defined and stocks are assigned to them. Denote y_{gkdj} as the count of the jth allele for the dth locus of the kth stock in the gth region, and denote $\boldsymbol{Y}$ as the entire baseline. The mixture genotype data $\boldsymbol{X}$ remains the same.

The structure of the stock proportions in the RAM is similar to that proposed by Okuyama and Bolker (2005). Let R_g be the regional contribution made by the gth region, and denote $\boldsymbol{R}$ as the vector of these contributions—notice that $\boldsymbol{R}$ must sum to one. We place a Dirichlet prior distribution on $\boldsymbol{R}$ with parameters $\boldsymbol{\gamma}$ such that $\gamma_g = 1/G$, with G being the number of regions.

Denote S_{gk} as the within-region stock proportion for the kth stock in the gth region, and denote $\boldsymbol{S}_g$ as the vector of all C_g stock proportions within the gth region—again, notice that $\boldsymbol{S}_g$ must sum to one. We place a Dirichlet prior distribution on $\boldsymbol{S}_g$ with parameters $\boldsymbol{\delta}_g$, with $\delta_{gk} = 1/C_g$. The ragged matrix of all stock proportions is represented by $\boldsymbol{S}$.

Like the baseline data, the baseline relative frequencies are also broken up, with q_{gkdj} being the relative frequency of the jth allele for the dth locus of the kth stock in the gth region, and $\boldsymbol{Q}$ as the entire array of baseline relative frequencies. We place the same Dirichlet prior distribution on $\boldsymbol{Q}_{gkd}$ as we placed on $\boldsymbol{Q}_{id}$ in the previous model.

We let r_{mg} denote the regional identity for the gth stock for the mth mixture individual, where r_{mg} equals 1 if individual m belongs to the gth region, and zero otherwise. The vector of regional identities for the mth individual is denoted as $\boldsymbol{r}_m$, and the matrix of all regional identities is represented as $\boldsymbol{r}$. A multinomial prior distribution is placed on $\boldsymbol{r}_m$ with size one and probabilities equal to the regional contributions $\boldsymbol{R}$.

Finally, let z_{mgk} be the within-region stock identity for the kth stock in the gth region for the mth mixture individual, where z_{mgk} equals one if individual m belongs to the kth stock of the gth region, and zero otherwise. Denote $\boldsymbol{z}_{mg}$ as the vector of stock identities for the *g*th region for the *m*th individual, and let $\boldsymbol{z}_{m}$ be the ragged matrix of stock identities for this individual. The ragged array of all stock identities is denoted as **z**. We place a multinomial prior distribution on $\boldsymbol{z}_{mg}$ with size $\boldsymbol{r}_{mg}$ and probabilities equal to $\boldsymbol{S}_g$. Because $\boldsymbol{r}_{mg}$ equals 1 if individual *m* belongs to the *g*th region, and zero otherwise, the only way the prior distribution of $\boldsymbol{z}_{mg}$ can have positive size is if $\boldsymbol{r}_{mg}$ equals one. In other words, the *m*th individual cannot belong to a stock that is outside that individual's region.

If we knew both the region and stock of origin for each mixture individual, the full genotypic likelihood can be expressed as:

$$L(\boldsymbol{X}|\boldsymbol{Q},\boldsymbol{S},\boldsymbol{r},\boldsymbol{z}) = \prod_{m=1}^{M}\prod_{g=1}^{G}\left(\prod_{k=1}^{C_g} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk}\right)^{z_{mgk}} I\left(\sum_{k=1}^{C_g} z_{mgk} = r_{mg}\right)\right)$$

Here, we use *I*() as an indicator function that is equal to one if the argument is true, and zero otherwise. Similar to the previous model, the baseline likelihood can be written as:

$$L(\boldsymbol{Y}|\boldsymbol{Q}) \propto \prod_{g=1}^{G}\prod_{k=1}^{C_g}\prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{gkdj}^{y_{gkdj}}$$

The full likelihood, $L(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Q},\boldsymbol{r},\boldsymbol{z})$, is simply the product of these two components. Multiplying the likelihood by the priors gives the posterior distribution:

$$\begin{aligned}
\mathrm{P}(\boldsymbol{R},\boldsymbol{S},\boldsymbol{Q},\boldsymbol{r},\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y}) &\propto L(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Q},\boldsymbol{r},\boldsymbol{z})\mathrm{P}(\boldsymbol{z}|\boldsymbol{r},\boldsymbol{S})\mathrm{P}(\boldsymbol{r}|\boldsymbol{R})\mathrm{P}(\boldsymbol{S}|\boldsymbol{\delta})\mathrm{P}(\boldsymbol{R}|\boldsymbol{\gamma})\mathrm{P}(\boldsymbol{Q}|\boldsymbol{\beta}) \\
&\propto \left\{\prod_{m=1}^{M}\prod_{g=1}^{G}\left(\prod_{k=1}^{C_g} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk}\right)^{z_{mgk}} I\left(\sum_{k=1}^{C_g} z_{mgk} = r_{mg}\right)\right)\right\}\left\{\prod_{g=1}^{G}\prod_{k=1}^{C_g}\prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{gkdj}^{y_{gkdj}}\right\} \\
&\times \left\{\prod_{m=1}^{M}\prod_{g=1}^{G}\left(\prod_{k=1}^{C_g} S_{gk}^{z_{mgk}} I\left(\sum_{k=1}^{C_g} z_{mgk} = r_{mg}\right)\right)\right\}\left\{\prod_{m=1}^{M}\prod_{g=1}^{G} R_g^{r_{mg}}\right\} \\
&\times \left\{\prod_{g=1}^{G}\prod_{k=1}^{C_g} S_{gk}^{\delta_{gk}}\right\}\left\{\prod_{g=1}^{G} R_g^{\gamma_g}\right\}\left\{\prod_{g=1}^{G}\prod_{k=1}^{C_g}\prod_{d=1}^{D}\prod_{j=1}^{J_d} q_{gkdj}^{\beta_{dj}}\right\}
\end{aligned}$$

From this distribution, we need to isolate the conditional distribution of each of the parameters. However, $\boldsymbol{r}_{\boldsymbol{m}}$ and $\boldsymbol{z}_{\boldsymbol{m}}$ are closely linked and separating them is somewhat difficult. Jointly, their conditional distribution is:

$$P(\boldsymbol{r}_m,\boldsymbol{z}_m|\boldsymbol{X},\boldsymbol{R},\boldsymbol{S},\boldsymbol{Q}) \propto \prod_{g=1}^{G} R_g^{r_{mg}}\left(\prod_{k=1}^{C_g}\left(S_{gk} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk}\right)\right)^{z_{mgk}} I\left(\sum_{k=1}^{C_g} z_{mgk} = r_{mg}\right)\right)$$

To find the conditional distribution for $\boldsymbol{r}_{\boldsymbol{m}}$, we need to marginalize over $\boldsymbol{z}_{\boldsymbol{m}}$ by recognizing that:

$$P\left(r_{mg}=1|\boldsymbol{X},\boldsymbol{R},\boldsymbol{S},\boldsymbol{Q}\right)=\sum_{z_{mg}} P\left(r_{mg}=1,z_{mg}|\boldsymbol{X},\boldsymbol{R},\boldsymbol{S},\boldsymbol{Q}\right)$$
$$\propto \sum_{z_{mg}} R_g \prod_{k=1}^{C_g}\left(S_{gk} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk}\right)\right)^{z_{mgk}} = R_g \sum_{k=1}^{C_g} S_{gk} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk}\right)$$

Therefore, we can draw $\boldsymbol{r_m}$ from:

$$\boldsymbol{r}_m|\boldsymbol{X},\boldsymbol{R},\boldsymbol{S},\boldsymbol{Q} \sim \text{multinomial}\left(1,\left\{\frac{R_g \sum_{k=1}^{C_g} S_{gk} f\left(X_m|Q_{gk}\right)}{\sum_{j=1}^{G} R_j \sum_{k=1}^{C_j} S_{jk} f\left(X_m|Q_{jk}\right)}\right\}_{g=1,2,\dots,G}\right)$$

Once we know which region the *m*th individual belongs to, we can draw $\boldsymbol{z_{mg}}$ from:

$$\left(\boldsymbol{z}_{mg}|\boldsymbol{r}_{mg}=1,\boldsymbol{X},\boldsymbol{S},\boldsymbol{Q}\right) \sim \text{multinomial}\left(1,\left\{\frac{S_{gk} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk}\right)}{\sum_{k'=1}^{C_g} S_{gk'} f\left(\boldsymbol{X}_m|\boldsymbol{Q}_{gk'}\right)}\right\}_{k=1,2,\dots,C_g}\right)$$

Next, given the regional identities, $\boldsymbol{R}$ is drawn from:

$$\boldsymbol{R}|\boldsymbol{r},\boldsymbol{\gamma} \sim \text{Dirichlet}\left(\left\{\sum_{m=1}^{M} r_{mg}+\gamma_g\right\}_{g=1,2,\dots,G}\right)$$

Then, given the stock identities for each region, $\boldsymbol{S_g}$ is drawn from:

$$\boldsymbol{S}_g|\boldsymbol{z},\boldsymbol{\delta} \sim \text{Dirichlet}\left(\left\{\sum_{m=1}^{M} z_{mgk}+\delta_{gk}\right\}_{k=1,2,\dots,C_g}\right)$$

Finally, for each stock within each region and for each locus, we generate $\boldsymbol{Q}_{gkd}$ from:

$$\boldsymbol{Q}_{gkd}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{z},\boldsymbol{\beta} \sim \text{Dirichlet}\left(\left\{\sum_{m=1}^{M} z_{mgk}x_{mdj}+y_{gkdj}+\beta_{dj}\right\}_{j=1,2,\dots,J_d}\right)$$

This completes one cycle of the Gibbs algorithm for the RAM.

Simulations

Analyzing multiple simulated mixtures with Bayesian methods is somewhat challenging because no “canned” software is available to conduct automated analyses. For this reason, we were limited in the number of mixtures that could be analyzed. To simulate each fish, we randomly selected the stock of origin from the appropriate region, then, for each locus, we drew a genotype from the multinomial distribution using the observed baseline allele relative frequencies. We simulated 100 mixtures of 200 fish that were each composed of 100% Norton Sound chum, and analyzed them with a Western Alaska baseline. The baseline was composed of 53 SNPs and included 60 stocks representing 6 regions, including: Kotzebue Sound (5 stocks), Seward Peninsula (2 stocks), Norton Sound (12 stocks), Lower Yukon River (18 stocks), Kuskokwim River/Bay (17 stocks), and Bristol Bay (6 stocks). The mixtures were analyzed in three ways: 1) Pella-Masuda Model with the True Flat Prior; 2) Pella-Masuda Model with the Regional Flat Prior; and 3) Regional Allocation Model. The Pella-Masuda analyses were conducted in the R programming language utilizing the package BRUGS. The RAM analyses were also conducted within an R program, but the program called upon a C++ function that was developed at the Gene Conservation Laboratory to speed up analysis. For each mixture, one chain was run for 30,000 iterations, discarding the first 5,000 as burn-in. From the 25,000 iterations that were retained, posterior means of the stock proportions and the regional proportions were calculated. Also calculated were the means, central 90% quantiles, and root mean square errors of the 100 posterior means.

Results

The mean and central 90% of the Norton Sound proportions for the Pella-Masuda model TFP, the Pella-Masuda model RFP, and the RAM were 0.831 (0.686-0.929), 0.834 (0.696-0.932), and 0.880 (75.7-0.949), respectively (Table 1; Figures 2-4), and the root mean square errors were 0.091, 0.088, and 0.063, respectively (Table 1). For the Pella-Masuda model, while both the TFP and the RFP showed very similar amounts of misallocation, the RFP tended to shift some of the misallocation away from the regions with the most stocks and into regions with fewer stocks (Figures 2-3). The RAM showed less misallocation than both prior structures of the Pella-Masuda model in terms of point estimate and tightness of the central 90% quantiles, and tended to flatten out the amount of misallocation more evenly across the remaining regions (Figure 4).

Discussion

The RAM appeared to be moderately successful in reducing the non-uniform bias due to the unequal distribution in the number of stocks among the regions, much more so than the Pella-Masuda model with the RFP. Comparing Figure 4 with Figures 2 and 3 shows that the misallocation to the regions represented by larger numbers of stocks (i.e. Yukon and Kuskokwim) was somewhat reduced. We suspect that the larger misallocation to these regions that persisted with the RAM were due to the fact that these are more genetically similar to Norton Sound than the other regions, and less due to failure of the RAM to reduce the non-uniform bias. The dendrogram shown in Figure 5 supports this suspicion. Another improvement of the RAM was that the width of the central 90% quantiles was somewhat narrower. This reduction in variation about the expected value, in addition to the reduced bias, equates to an improvement of the estimator’s mean square error (Table 1). While the RAM still failed to achieve the 90% mark that the Gene Conservation Laboratory strives to attain, overall it performed better than either of the Pella-

Masuda models in this tough situation. The addition of new SNP markers to the RAM may provide the resolution to meet the 90% mark.

The rationale for why the RAM was expected to reduce the non-uniform bias can be seen by inspecting the regional identity probability:

$$P(\boldsymbol{r}_{mg} = 1|\boldsymbol{X}, \boldsymbol{R}, \boldsymbol{S}, \boldsymbol{Q}) \propto \; R_g \sum_{k=1}^{c_g} S_{gk} f(X_m|Q_{gk})$$

This probability is a product of the regional contribution and a weighted average genotypic frequency, with the weights summing to one. Because the weights sum to one, the genetic component of this probability, i.e. the weighted average genotypic frequency, remains comparable regardless of the number of stocks within the region, which levels the playing field. The effect of this was seen in our simulation results. In our simulations, every mixture individual belonged to Norton Sound. Under the Pella-Masuda model, when allocating the *m*th fish at each cycle, all 60 stocks competed for allocation of this fish. As can be seen in Figures 2 and 3, the larger regions were more successful at gaining this allocation simply because they have more stocks to compete with. However, under the RAM, when allocating the fish, only 6 regions were competing for allocation, each acting a single unit.

A further benefit is that the regional proportions are directly given a prior distribution, which allows the transmission of prior information at the regional level in a straight forward manner. This has great potential for modeling prior information in hierarchical models where there is often not enough information to adequately estimate hyperparameters for each of the individual stocks.

The RAM presented here is extended to only two levels of hierarchy of stocks within regions. However, it is conceivable to expand this model to further levels of hierarchy, such as sub-stocks within stocks, and stocks within regions. Such a model may be useful in situations where multiple levels of structure exist.

Literature Cited

Dann, T. H., C. Habicht, J. R. Jasper, H. A. Hoyt, A. W. Barclay, W. D. Templin, T. T. Baker, F. W. West, and L. F. Fair. 2009. Genetic stock composition of the commercial harvest of sockeye salmon in Bristol Bay, Alaska, 2006-2008. Fishery Manuscript Series No. 09-06.

Fournier, D. A., T. D. Beacham, B. E. Ridell, and C. A. Busack. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characheristics. Can. J. Fish. Aquat. Sci. 52:1688-1702.

Gelman, A., and D. Rubin. 1992. Inferences from iterative simulation using multiple sequences. Stat. Sci. 7:457-511.

Okuyama, T. and B. M. Bolker. 2005. Combining genetic and ecological data to estimate sea turtle origins. Ecol. App. 15(1):315-325.

Pella, J. and M. Masuda. 2001. Bayesian methods for analysis of stock mixtures from genetic characters. Fish. Bull. 99:151-167.

Pella, J. J., and Milner, G. B. (1987), "Use of genetic marks in stock composition analysis," in *Population Genetics and Fishery Management*, eds. N. Ryman and F. Utter, Seattle, WA: Washington Sea Grant Program, pp. 247-276.

Wood, C. C., S. McKinnell, T. J. Mulligan, and D. A. Fournier. 1987. Stock identification with the maximum-likelihood mixture model: sensitivity analysis and application to complex problems. Can. J. Fish. Aquat. Sci. 44:866-881.

Technical Committee review and comments

WASSIP Technical Document 7 Regional Allocation Model (RAM)

This documents outlines and tests the performance of two modifications of the Pella-Masuda stock composition estimation algorithm, applying them to 100% single stock samples from the Western Alaska chum salmon genetic baseline. One approach (the Regional Flat Prior) modifies the prior probabilities assigned to the model, while another (the Regional Allocation Model) modifies the model structure to incorporate the regional identities. Both approaches reduce the overallocation of samples to regions comprising many stocks, but the RAM performs better than the RFP.

Overall, this is a very nice exposition and test of an extension of the Pella-Masuda model, and convincingly demonstrates that, at least under some conditions, this extension will improve performance of regional allocations from stock mixtures. The TC was encouraged to see this interesting idea developed into a form that could easily be modified as a journal submission. We think the novel approach will provide useful options for conducting GSI. For publication in a journal (and this paper merits it), it would be nice to generalize the results beyond Western AK chum by drawing genetic samples from simulated stocks. In simulations, the genetic similarity among stocks could be controlled, and the effects of the number of stocks sampled from a region isolated from the effects of similarity of stocks within and among regions.

Although we did not identify any major flaws in the analyses, there are some issues regarding ghost populations and the appropriate priors that need further consideration. The general problem the RAM is intended to address is cumulative upward bias in estimated contributions of stocks that in reality contribute very little, or nothing, to the mixture. The bias is a type of edge effect that arises because individual stock estimates are constrained to the biologically plausible range 0-1; if the true value for a particular stock is 0, there is no possibility of balancing the occasional over-estimate by a negative one, and the result is upward bias (and hence downward bias in estimating contributions of stocks that actually do contribute substantially to the mix). Empirically, the bias is known to increase with the number of non-contributing stocks in a baseline. The bias is also positively correlated with uncertainty; if source populations are very divergent genetically (and assuming adequate sample sizes from the fishery), stock contributions can be determined with high precision and the resulting bias is small. With poorly differentiated stocks, cumulative mis-assignments to stocks that actually do not contribute to the mix can be substantial. Also, in the case of uncertain stock assignments, priors used in the Bayesian analysis can assume a relatively greater importance and can significantly influence results.

The general scenario that the RAM is appropriate to address is the following.

- Stocks are organized hierarchically into 2 or more regions or Reporting Groups (RGs).

- The RGs have the same number of actual populations but different numbers of populations that have been sampled for the baseline.
- A flat prior of stock contribution is computed as 1/n, where n is the total number of populations in the baseline.
- In this scenario, the RGs that have the most populations in the baseline will tend to attract the most spurious contribution assigned to low- or non-contributing stocks.

The solution to this problem proposed by Technical Document 7 is two-fold:

1. Ensure that each RG has the same overall prior, and within each RG ensure that each stock has an equal prior. This means that stocks in RGs with different numbers of populations in the baseline have different priors.
2. First determine which RG a fish is from, then which stock within the RG.

The second item in the list above is the novel feature of this document, and we think it merits publication. However, we question whether the idea of forcing each RG to have an equal overall prior is a general solution to the problem described. In fact, we can find little support for the idea that, in general, different RGs should have the same prior. Rather, we think the priors for each RG should reflect the relative probability that a given fish in the mix can be expected to come from the RG. The appropriate prior should reflect, among other things, the actual number of populations in each RG, the size of each population, the proximity to the location of the fishery sample, and things such as migration routes.

Consider the following scenario:

- Stocks are organized hierarchically into 2 or more regions or RGs.
- The RGs have different numbers of actual populations, and each actual population has been sampled for the baseline.
- Each population has the same size and productivity.

Under this scenario, the appropriate priors for each RG are proportional to the number of stocks in the baseline, and enforcing equal RG priors as in item 1) above could be expected to reduce accuracy of the estimates.

We therefore believe that the issue of appropriate priors needs more careful consideration, and these considerations should include not only the number of populations in the baseline but also the number of actual populations and perhaps information about each population. Real populations that are not sampled in a population genetics study are called ghost populations (Beerli 2004), and it is known that they can profoundly affect results of statistical analyses. Based on results obtained by Slatkin (2005), it likely will be difficult or impossible to develop a general formula that captures the effects of ghost populations on GSI estimates. This suggests that the most appropriate priors for use in GSI should be evaluated on a case-by-case basis.

For the particular case of separating stocks in mixtures taken from the WASSIP study area, the authors might think about the potential for using semi-informative priors, and investigate whether the priors have an appreciable effect on the results. For example, abundance varies greatly among the stocks/regions investigated; proximity of these stocks to the WASSIP area varies as well, and there is some rudimentary oceanic distribution information from tagging studies. Hopefully, the results aren't too sensitive to the priors on stock composition, but if they are, these priors should receive careful attention. In case of sensitivity, priors should be chosen based on the best biological information and possibly partially on management priorities. The effects of priors on estimates for small stocks should get particularly careful consideration. If the priors weight each region equally, and some of these small stocks get treated like a region, the priors could potentially dominate the results and strongly overweight their contributions.

Specific comments keyed to line number:

28: this is true only if some method has been used to account for unsampled alleles

51: isn't this a null hypothesis rather than an assertion?

150: is ragged matrix a real term?

185: "once we know …" … do you mean, "once we have estimated"?

208: what exactly did the C++ routine do?

247: we agree that in the example chosen, the new method helps to "level the playing field." However, as discussed above, forcing equal RG priors is not a sound general strategy for leveling the playing field.

Figure 1: how was the individual stock of origin for each Norton Sound fish in the simulated mixtures chosen?

How does the new method perform with different sampling fractions? And more realistic mixtures?

For publication in a journal, more context needs to be provided. For instance, the type of genetic characteristics comprising the baseline isn't specified.

Table 1. Simulation results and root mean square error (rMSE) for 100 mixtures of 100% Norton Sound chum for the Pella-Masuda Model with the True Flat Prior (P-M TFP), the Pella-Masuda Model with the Regional Flat Prior (P-M RFP), and the Regional Allocation Model (RAM).

Region	P-M TFP	P-M RFP	RAM
Kotzebue Sound	0.012	0.018	0.014
Seward Pen	0.004	0.011	0.010
Norton Sound	0.831	0.834	0.880
Lower Yukon	0.064	0.049	0.036
Kuskokwim	0.076	0.065	0.041
Bristol Bay	0.012	0.022	0.019
rMSE	0.091	0.088	0.063

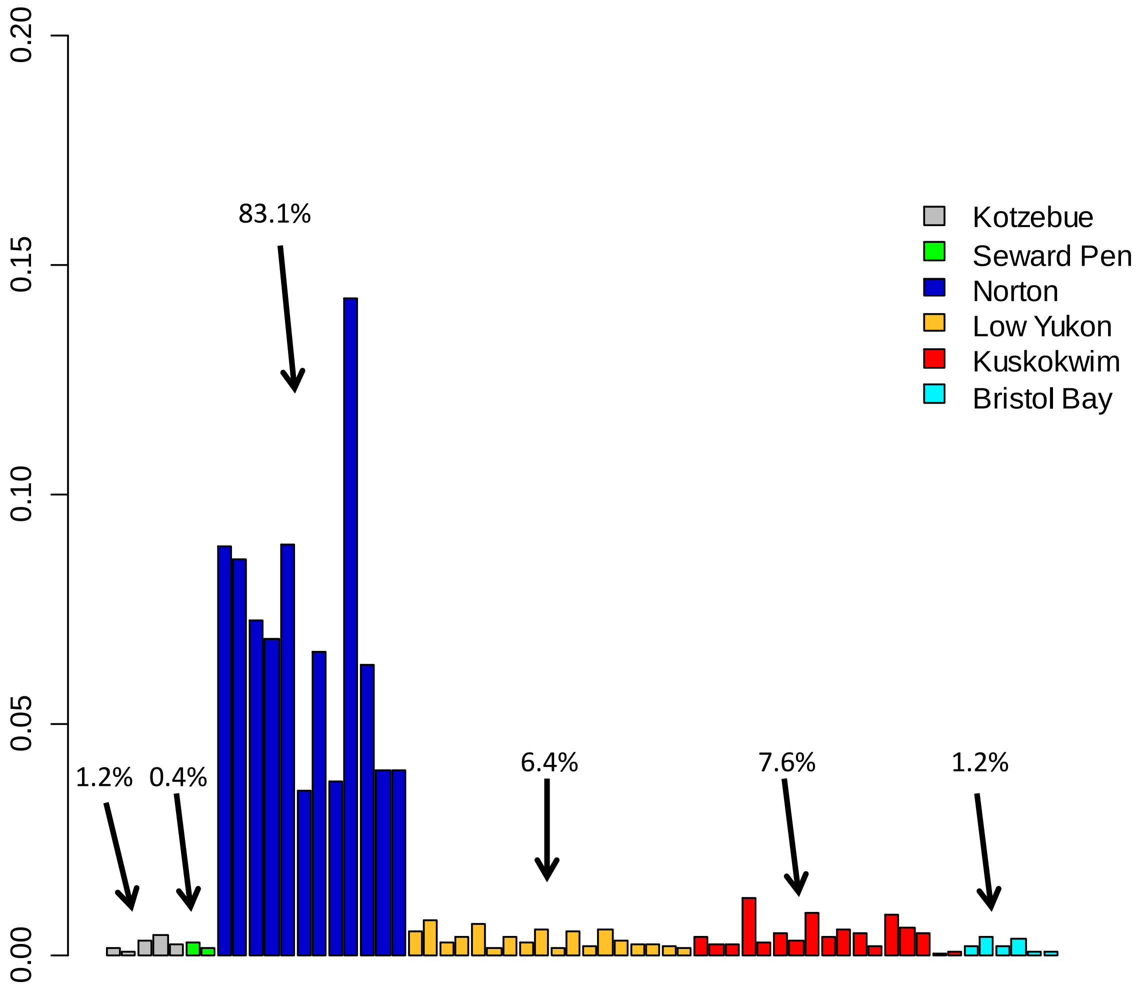

Figure 1. Simulation results for 100 mixtures of 100% Norton Sound chum for the Pella-Masuda Model shown at the individual stock level. The height of the bars represents the mean of 100 repetitions. An equal prior “count” of one divided by the number of stocks was given to each stock.

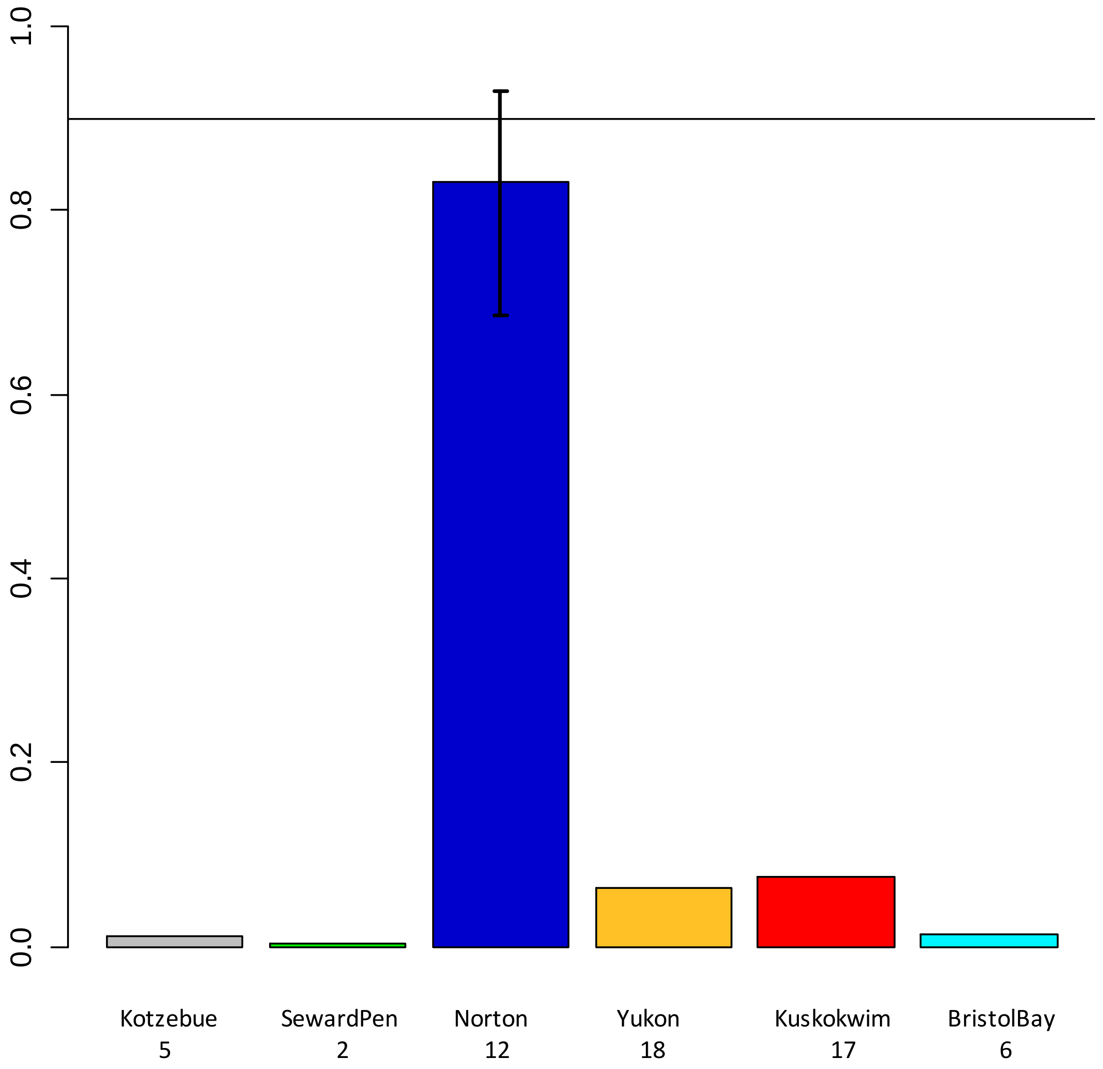

Figure 2. Simulation results for 100 mixtures of 100% Norton Sound chum for the Pella-Masuda Model using the True Flat Prior. The height of the bars represents the mean of 100 repetitions. Vertical bar represents the central 90%. Horizontal bar is the 90% line. Numbers under labels are the number of stocks within the region. These results are the same as shown in Figure 1 with the stock proportions summed into regions.

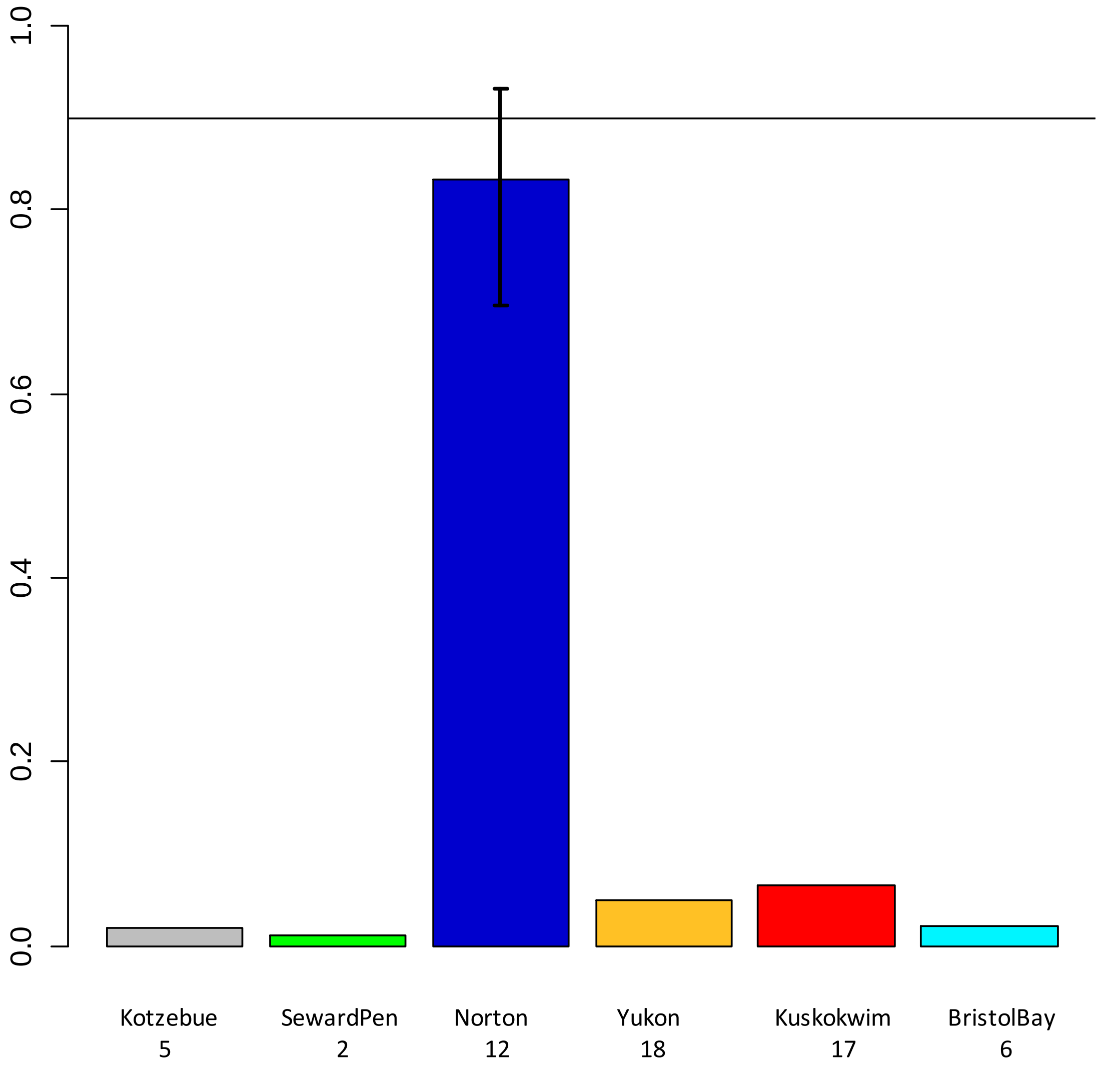

Figure 3. Simulation results for 100 mixtures of 100% Norton Sound chum for the Pella-Masuda Model using the Regional Flat Prior. The height of the bars represents the mean of 100 repetitions. Vertical bar represents the central 90%. Horizontal bar is the 90% line. Numbers under labels are the number of stocks within the region.

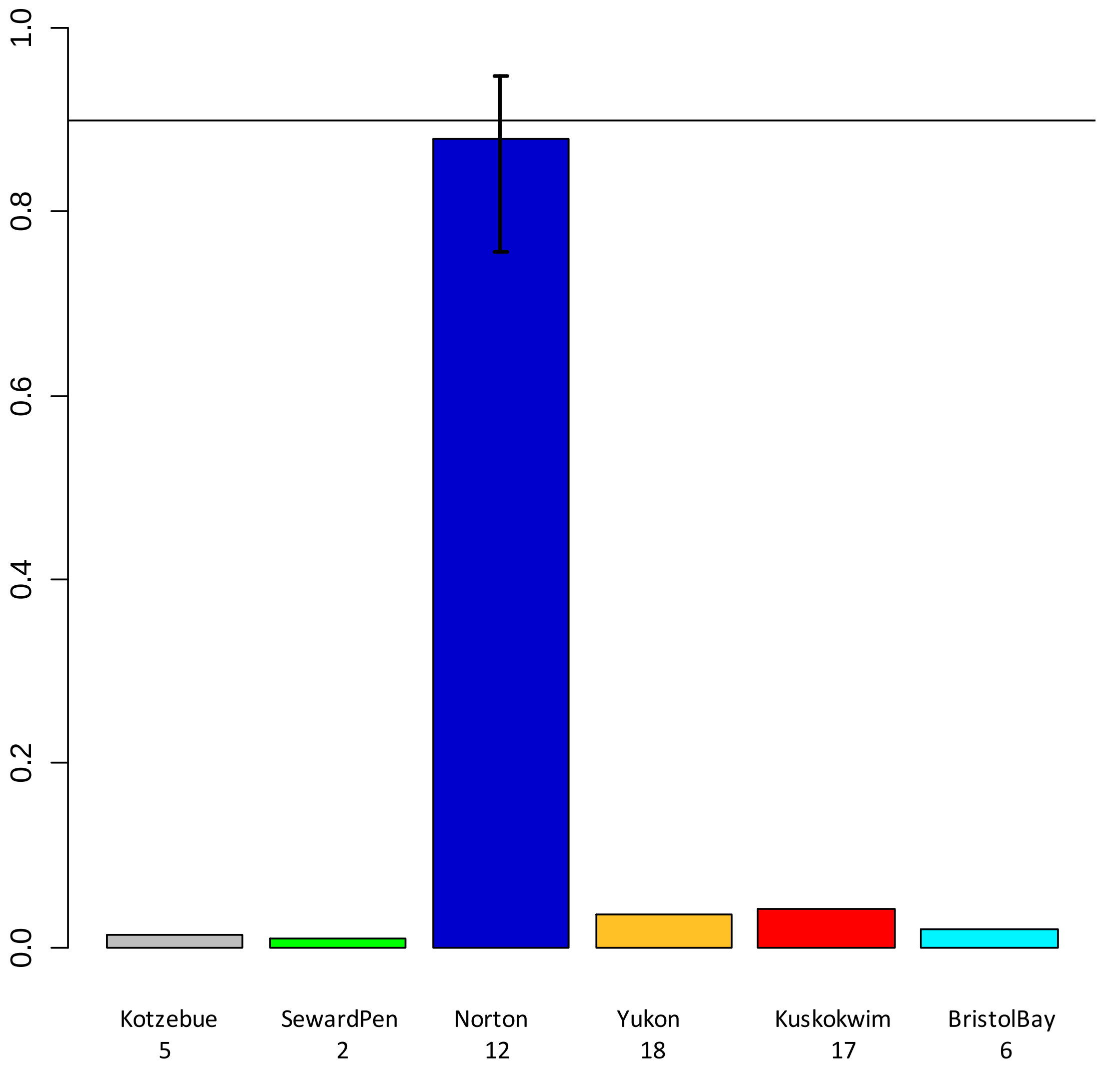

Figure 4. Simulation results for 100 mixtures of 100% Norton Sound chum for the Regional Allocation Model. The height of the bars represents the mean of 100 repetitions. Vertical bar represents the central 90%. Horizontal bar is the 90% line. Numbers under labels are the number of stocks within the region.

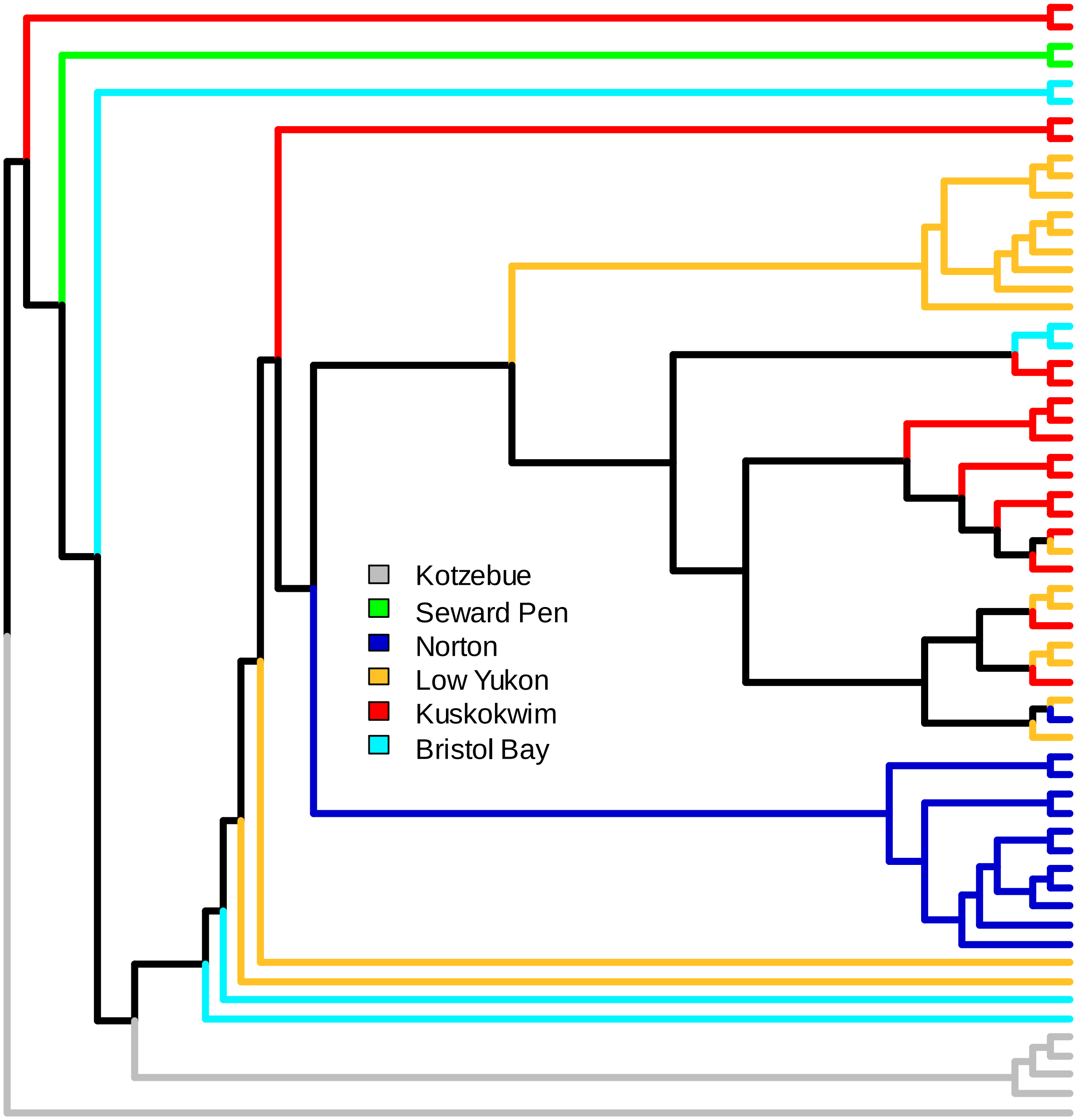

Figure 5. UPGMA tree of pair-wise F_{ST} for 60 stocks of Western Alaska chum demonstrating that Norton Sound chum are more genetically similar to Lower Yukon and Kuskokwim than the other regions.