APPLICATION OF BAYESIAN MIXED STOCK ANALYSIS FOR DETECTION OF SMALL CONTRIBUTIONS IN FISH MIXTURES

Report to the Gene Conservation Laboratory, Alaska Department of Fish and Game
June 2003

By Anton Antonovich
and
Bill Templin

Alaska Department of Fish and Game
Division of Commercial Fisheries
333 Raspberry Road
Anchorage, Alaska 99518

June 2003

Contents

I. Introduction

Mixed stock analysis (MSA) is a widely used tool in the management of commercial fisheries of salmon (*Oncorhynchus*) species. In fisheries MSA is used to estimate the composition of a mixture sample consisting of certain (known) populations (stocks). Such a case arises when fishing for salmon in the ocean where fish congregate in schools and mix in search for food. In these circumstances, a catch usually represents a mixture of salmon stocks. The need to know the composition of a mixture arises in the effort to preserve endangered populations that might be present in a mixture. Given the efficiency of modern fishing equipment and the large volume of catches, it is feasible that endangered stocks might be fished out completely during their ocean life phase if not monitored carefully. Management becomes especially important when spawning runs begin. At this time, knowledge of possible and preferred pathways of returning salmon and restricting fishing efforts in some regions may well be crucial in protecting endangered stocks. In addition, management of international fisheries often requires estimation of stock composition. MSA provides estimates of the relative contribution of stocks in a mixture of individuals. One of the important questions often posed by managers and researchers is whether a specific population, or a group of populations, is present or absent in a given mixture. An answer to this question determines the management strategy and may yield restrictive decisions. For a number of years MSA based on the conditional maximum likelihood approach has been successfully utilized by the Gene Conservation Lab, Alaska Department of Fish and Game (ADF&G) and by other resource management organizations. This study investigates the Bayesian approach to MSA.

Genetic markers, such as multilocus genotypes, are widely used in mixed stock analysis (Pella and Milner, 1987). These genotypes serve as natural tags and allow the identification of fish origin. The unknown proportions of stocks comprising the mixture sample can be estimated from allele counts of individuals in the mixture if allele relative frequencies (RFs) differ among contributing stocks. The larger the differences in allele RFs among contributing stocks, the more accurate estimates of stock proportions can be obtained. In the case when a population has a private (i.e. unique) allele, it will be perfectly identifiable. However, even in this case, the bootstrap method often used to obtain confidence intervals for stock composition estimates, can significantly lower the statistical power of detecting small stock contributions (Reynolds and Templin, 2003).

The Bayesian approach to the mixed stock analysis has become very appealing lately and it is believed to have a number of advantages over the conditional maximum likelihood (CML) method. For example, the CML stock composition estimate maximizes a likelihood function of the stock-mixture genotypes as if their RFs in the baseline stocks were known without error. In practice, however, the baseline allele RFs are determined from samples of limited sizes and thus have some uncertainty associated with them. Therefore, the resulting CML estimates are usually biased and their variability is underestimated. The Bayesian method treats the baseline allele RFs as unknowns with a specified prior distribution. Later, when the baseline and mixture samples become available, the prior is updated to a posterior distribution according to Bayes' rule (see section II for details).

Another powerful argument in favor of Bayesian analysis has its roots in the principal difference between the two approaches. Namely, CML does not use the information from a mixture sample to improve the estimates of baseline allele RFs, whereas the Bayesian method allows for that. As Pella and Masuda (2001) indicate, this omission becomes ever more meaningful with accumulation of mixture individuals from a series of analyses performed on the stock mixtures of the same baseline populations.

Treatment of RFs for rare alleles (alleles with RFs < 0.005, Hartl and Clark, 1997) is also quite different between the two approaches. In CML, if a baseline sample does not have a rare allele present, it is assumed to be absent, even though it might be present in a population. The Bayesian model, described in the next section, shrinks the observed baseline RFs of individual stocks toward genetically better-established grand, regional or group means (Pella and Masuda, 2001). Thus, a rare allele absent from a baseline sample will be assigned a relative frequency from the corresponding prior.

Finally, the Bayesian approach allows for better handling of missing data. In particular, missing baseline allele RFs are easily filled in with appropriate grand, regional or group means – similarly to treatment of rare alleles (Pella and Masuda, 2001). They are revised later during analysis of the mixture sample.

The objective of this study is to investigate and assess the use of Bayesian mixed stock analysis for the detection of specific populations (or groups of populations) in a mixture. Of particular interest is evaluating sensitivity of the Bayesian method to small contributions.

The data set for this study consists of genetic samples from 63 baseline populations of sockeye salmon (*Oncorhynchus nerka*); 51 from Bristol Bay drainages (Alaska) (see Figure 1 and 2) and 12 from Russian rivers (Kamchatka peninsula). These populations are further grouped into 12 reporting regions (RR) according to geographic location and genetic similarities (Table 1). In general, a reporting region is defined through simulations with 100% contribution from that region and it is required to

Figure 1. Map of the region. Courtesy: ADF&G

demonstrate 90% or better estimated mean contribution for simulated mixtures (Seeb et al. 2000, Reynolds and Templin 2003).

Figure 2. Bristol Bay with its main river systems. (Courtesy of Chris Habicht and Bill Templin, ADF&G, Anchorage, AK)

Bristol Bay (Figure 2) with its eight major river systems is home to the largest commercial salmon fishery in the world. Sockeye salmon are by far the most abundant salmon species in Bristol Bay. Abundance of some stocks is known to be declining (personal communication with Bill Templin, ADF&G) and therefore the main management objective is to achieve desired escapement for these stocks. Thus, detection of the specific populations in a mixture is especially important for management.

In this setting, if we had a mixture of fish from these 63 populations, we would like to determine the composition of stocks in the mixture, or as described before - in the case of small contributions - the presence or absence of specific stocks or groups of stocks. The four major questions raised by the geneticists from ADF&G for this study can be formulated as follows:

1) How many fish from Kvichak RR (22 populations) would have to be in a mixture to be detectable (Kvichak and Naknek baseline only)?
2) How many fish from Lake Clark RR (5 populations) would have to be in a mixture to be detectable (Kvichak and Naknek baseline only)?
3) How many fish from Bristol Bay stocks (51 populations) would have to be in a mixture to be detectable (full baseline)?

4) How many fish from Russian stocks (12 populations) would have to be in a mixture to be detectable (full baseline)?

To answer these questions, a series of simulation experiments, with mixtures of size 200 fish, was designed and implemented as described in section III. A Bayesian mixture model developed and described in Pella and Masuda (2001) and Masuda (2002) was used to perform simulations and estimate the contribution of each baseline stock in a mixture sample. The computations were conducted using the free software package BAYES available at the Alaska Fisheries Science Center anonymous ftp site *ftp://www.abl.noaa.gov/sida/mixture-analysis/bayes*.

The results of simulation experiments are discussed in section VI. It is shown that the Bayesian method under consideration has relatively high sensitivity. In other words, it is able to detect small contributions of selected stock groups with reasonably high statistical power. Tables of summary statistics based on the posterior distributions for each contribution level and each simulated mixture sample are presented in Appendix 1. Histograms of posterior mean contributions for each scenario (i.e., reporting region by contribution level) are presented in Appendix 2.

II. Description of the Bayesian Model (adapted from Pella & Masuda, 2001)

According to Bayes' rule, the posterior distribution of an unknown parameter θ given the data Y, $f(\theta \mid Y)$, is proportional to the product of its prior density $f(\theta)$ and the likelihood of the sample, $g(Y \mid \theta)$, (Gelman et al., 1995)

$$f(\theta \mid Y) \propto f(\theta) g(Y \mid \theta). \qquad (1)$$

The main idea behind this rule is that some information about θ, which is expressed through its prior density, is available before the experiment begins. This information might be based upon previous study, or just be a researcher's best guess. If nothing is known a priori about θ, an uninformative prior can be chosen in a sense that its influence on the posterior density will be minimized. In this case, knowledge about θ will only come from the data, *Y*. The likelihood function, $g(Y \mid \theta)$, reflects the probability of observing the data, *Y*, given the parameter θ and it is based on the probabilistic model defining the distribution of Y as a function of θ. After sampling, the observed data, *Y*, are used to revise the prior to the posterior probability density of the unknown parameter.

Once the posterior for θ is obtained, a variety of point estimates such as mean, median, or mode can be derived along with corresponding credibility intervals (the Bayesian counterpart of frequentist confidence intervals). In mixed stock analysis with genetic markers, the unknown parameters are partitioned into two parts [$\theta = (\mathbf{p}, \mathbf{Q})$]: (1) the stock proportions of the mixture, **p**; and (2) the allele relative frequencies (RF) of the baseline stocks, **Q**. The mixture sample provides multilocus genotypes of the mixture individuals whereas the baseline samples provide allele RFs at the loci measured on the mixture genotypes.

Prior for stock proportions, $f(\mathbf{p})$

Building a Bayesian model starts with specifying a prior for the unknown set of parameters $\theta = (\mathbf{p}, \mathbf{Q})$, which is a product of block priors for its components, **p** and **Q**. An uninformative uniform prior is chosen for **p** in this study, to allow the mixture sample information to dominate the prior. A prior for a mixture of *c* possible stocks must comply with two restrictions on vector **p**. First, each individual stock proportion must lie between zero and one; and, second, their sum over all stocks should equal one,

$$0 < p_i < 1, \quad \sum_{i=1}^{c} p_i = 1.$$

The Dirichlet probability density accommodates these requirements and is commonly used as a prior with compositional count data both for its computational convenience and for it straightforward interpretation as additional data. The computational convenience of the Dirichlet prior density lies in the fact that paired with multinomial likelihood function it forms the conjugate family. In other words, if the sampling distribution of the data is multinomial, then choosing the Dirichlet prior will automatically yield a Dirichlet posterior. Prior draws of **p** from the Dirichlet probability density,

$$f(\mathbf{p}) = D(\mathbf{p} \mid \alpha_1, \alpha_2, ..., \alpha_c) = \frac{\Gamma\left(\sum_{i=1}^{c} \alpha_i\right)}{\prod_{i=1}^{c} \Gamma(\alpha_i)} \prod_{i=1}^{c} p_i^{\alpha_i - 1}, \tag{2}$$

$$\alpha_i > 0, \; i = 1, 2, ..., c,$$

have means, variances, and covariances given by (see Gelman et al, 1995)

$$E(p_i) = \frac{\alpha_i}{\alpha_0}, \quad Var(p_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad Cov(p_i, p_k) = \frac{-\alpha_i \alpha_k}{\alpha_0^2(\alpha_0 + 1)}, \text{ for } i \neq k \;,$$

where $i, k = 1, 2, ..., c$ and $\alpha_0 = \sum_{i=1}^{c} \alpha_i$.

If a prior draw of $\mathbf{p} \sim D(\alpha_1, \alpha_2, ..., \alpha_c)$ was obtained for the stock proportions of a stock mixture, and then a mixture sample of size *M* was drawn such that the individuals could be correctly identified to stock origin, their counts, $\mathbf{Z} = (z_1, z_2, ..., z_c)$, would have a conditional multinomial distribution,

$$f(\mathbf{Z} \mid \mathbf{p}, M) = \frac{M!}{z_1! z_2! \cdots z_c!} p_1^{z_1} p_2^{z_2} \cdots p_c^{z_c} \,,$$

or **Z** | **p** ~ *Mult(M,* **p***)*. Thus the posterior for **p**, given **Z**, would be the Dirichlet,

$$f(\mathbf{p} \mid \mathbf{Z}) \sim D(\alpha_1 + z_1, \alpha_2 + z_2, ..., \alpha_c + z_c) \,. \tag{3}$$

The prior parameters enter the posterior density in the same way as the sample counts and therefore can be viewed as counts obtained before the stock mixture was sampled (thus, having an interpretation as additional data). In practice, however, the mixture individuals are identified to stock origin (with unavoidable random error) during each cycle of the data augmentation algorithm (Gibbs sampler) later when samples are generated from the

posterior. With the stock origins identified at a cycle, the uncertainty in **p** is described by the Dirichlet posterior with parameters equal to the sums of stock counts and the prior parameters ($z_i + \alpha_i$).

Assigning equal values summing to 1 to the prior parameters (as was done in this study), $\alpha_1 = \alpha_2 = \ldots = \alpha_c = c^{-1}$, would be equivalent to adding just a single individual to the mixture sample, thus allowing the information from the mixture sample to dominate that from the prior distribution.

Prior for allele RFs given baseline samples, $f(\mathbf{Q} \mid \mathbf{Y})$

Genetic structures of separate stocks are determined by their allele RFs, **Q**. In the Bayesian approach, **Q** is treated as an unknown quantity. An estimate of **Q** can be obtained from baseline samples to estimate the stock genetic compositions. However, since baseline sample sizes are commonly limited, the observed RFs for individual stocks are shrunk toward central values that are more reliably determined and are consistent with the genetic similarity of the stocks (Pella and Masuda, 2001).

To develop our model we start with allele RFs at a single locus and then extend the reasoning to cover multiple loci. Consider a locus with T distinct alleles. Each of c baseline stocks will have a different set of RFs for these alleles. Denote the resulting unobserved RFs for the i^{th} stock by $\mathbf{q}_i = (q_{i1}, q_{i2}, \ldots, q_{iT})$. The Bayes prior in the form of a Dirichlet probability density is chosen for baseline sampling,

$$f(\mathbf{q}_i) = D(\beta_1, \beta_2, ..., \beta_T) .$$

Next, baseline samples of $n_1, n_2, \ldots, n_c$ alleles of the locus are available from the c stocks. The counts of the different alleles for the i^{th} stock, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$, have the multinomial distribution, $Mult(n_i, \mathbf{q}_i)$, and therefore the baseline posterior for the unknown allele RFs in each stock is also Dirichlet,

$$f(\mathbf{q}_i \mid \mathbf{y}_i) \propto D(\beta_1 + y_{i1}, \beta_2 + y_{i2}, ..., \beta_T + y_{iT}) . \quad (4)$$

The posterior means of $\mathbf{q}_i \mid \mathbf{y}_i$ can be written as a weighted average of the observed and prior mean RFs (e.g., see Rice, 1995). If the baseline sample is missing, the posterior mean equals the prior mean. Otherwise the posterior mean ranges between the observed and prior mean RFs. Note, that all posterior means for the allele RFs are positive, so that absence of an allele from a stock's baseline sample implies it is only rare and was missed in sampling rather than nonexistent (Pella and Masuda, 2001).

The pseudo-Bayes method is used in this study to estimate the values of the baseline prior parameters, $\beta_1, \beta_2, \ldots, \beta_T$ (see Pella and Masuda 2001 for details). Complete analysis of the baseline requires repeated and separate application of the pseudo-Bayes method to each locus. Suppose a total of H loci compose the stock-mixture multilocus genotypes. Let the h^{th} locus have J_h alleles with prior parameters $\boldsymbol{\beta}_h = (\beta_{h1}, \beta_{h2}, ..., \beta_{hJ_h})$, and allele RFs in the i^{th} stock of $\mathbf{q}_{ih} = (q_{ih1}, q_{ih2}, ..., q_{ihJ_h})$. If $\mathbf{Q}_i$ denotes the ith stock's combined arrays, $\mathbf{q}_{i1}, \mathbf{q}_{i2}, \ldots, \mathbf{q}_{iH}$, then the prior for the allele RFs of the complete baseline, $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_c)$ will be

$$f(\mathbf{Q}) = \prod_{i=1}^{c} f(\mathbf{Q}_i) = \prod_{i=1}^{c}\prod_{h=1}^{H} f(\mathbf{q}_{ih}) = \left(\prod_{h=1}^{H} D(\beta_{h1}, \beta_{h2}, ..., \beta_{hJ_h}) \right)^c ,$$

that is, prior draws for allele RFs are independent among stocks and loci.

The baseline samples are drawn independently from the stocks. Denote by $\mathbf{Y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, ..., \mathbf{y}_{iH})$ the H arrays of allele counts in the baseline sample for the i^{th} stock, and by **Y,** the entire baseline collection of $\mathbf{Y}_1$, $\mathbf{Y}_2$, …, $\mathbf{Y}_c$. Then the Bayesian posterior density for the allele RFs of the entire baseline is the product of Dirichlet densities,

$$f(\mathbf{Q} \mid \mathbf{Y}) = \prod_{i=1}^{c} f(\mathbf{Q}_i \mid \mathbf{Y}_i) = \prod_{i=1}^{c}\prod_{h=1}^{H} f(\mathbf{q}_{ih} \mid \mathbf{y}_{ih}) = \prod_{i=1}^{c}\prod_{h=1}^{H} D(\beta_{h1} + y_{ih1}, ..., \beta_{hJ_h} + y_{ihJ_h}), \qquad (5)$$

where each density in the product has a mean vector, for the stock and locus, equal to a weighted average of the observed allele RFs and corresponding prior means.

Likelihood function of a mixture sample, $g(\mathbf{X} \mid \boldsymbol{\theta})$

The mixture sample likelihood function is proportional to the probability of drawing individuals with observed genotypes as a function of the unknown parameters, **p** and **Q**. Let $\mathbf{X_m}$ denote the multilocus genotype of the m^{th} individual and let the array **X** denote the collection of such arrays for the M individuals composing the stock-mixture sample. Next, let the probability of observing individuals with the genotype $\mathbf{X_m}$ in the i^{th} stock, which depends on that stock's allele RFs, $\mathbf{Q}_i$, be denoted as $\pi(\mathbf{X}_m \mid \mathbf{Q}_i)$. Then, the proportion of individuals with the genotype in the stock mixture is the weighted sum, $\sum_{i=1}^{c} p_i \pi(\mathbf{X}_m \mid \mathbf{Q}_i)$, where the weights are represented by the (unknown) population's contributions to the mixture, p_i's. Subsequently, the likelihood function for the entire mixture sample is

$$g(\mathbf{X} \mid \mathbf{p}, \mathbf{Q}) = \prod_{m=1}^{M} \left(\sum_{i=1}^{c} p_i \pi(\mathbf{X}_m \mid \mathbf{Q}_i) \right) \qquad (6)$$

Here **X** is treated as fixed, and the likelihood, $g(\mathbf{X} \mid \boldsymbol{\theta})$**,** is a random function of the unknown parameters **p** and **Q**.

Posterior distribution of the unknowns, $f(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y})$

To complete the Bayesian setting, the posterior distribution of the unknown stock proportions in the mixture and of the baseline allele RFs is proportional to the product of the prior density for the unknown parameters and the likelihood function of the mixture sample, i.e.

$$f(\mathbf{p},\mathbf{Q}\,|\,\mathbf{X},\mathbf{Y}) \propto f(\mathbf{p},\mathbf{Q})g(\mathbf{X}\,|\,\mathbf{p},\mathbf{Q}) \tag{7}$$

The prior density for the stock proportions, $f(\mathbf{p})$, is the uninformative Dirichlet of Equation 2. The baseline posterior of Equation 5 becomes the prior for the allele RFs. Now, since the priors on stock composition and allele RFs are considered independent, the joint prior for the unknowns is the product of its two constituents,

$$f(\mathbf{p},\mathbf{Q}) = f(\mathbf{p})f(\mathbf{Q}\,|\,\mathbf{Y}).$$

The analytical evaluation of the posterior distribution is complicated by the large number of terms in the likelihood function, which makes finding a proportionality constant impractical (Pella and Masuda, 2001). Instead, a sufficient number of samples can be drawn sequentially from the posterior distribution to accurately describe it. A data augmentation algorithm – the Gibbs sampler – is used to draw a sequence of samples (Gelman et al., 1995).

The data augmentation algorithm

The main idea behind the algorithm lies in the fact that the estimation of parameters would be greatly simplified if the stock identities of the mixture individuals were known. Then, the posterior distribution for stock proportions and allele RFs can be directly obtained by updating Dirichlet priors of Equations 2 and 5 with the multinomial counts from the mixture sample (Pella and Masuda, 2001).

The stock identities of the mixture individuals are determined by chance in the data augmentation algorithm. Let $\mathbf{z}_m = (z_{m1}, z_{m2}, ..., z_{mc})$ indicate the stock origin of the m^{th} mixture individual by a single "1" at the coordinate of the contributing stock, and $(c-1)$ "0" at the remaining coordinates. Also, let $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_M)$ denote the stock origins of all mixture individuals. If $\mathbf{p}$ and $\mathbf{Q}$ were known, the probability that a randomly drawn individual with genotype $\mathbf{X}_m$ came from the i^{th} stock would be

$$w_{mi} = \frac{p_i \pi(\mathbf{X}_m\,|\,\mathbf{Q}_i)}{\sum_{j=1}^{c} p_j \pi(\mathbf{X}_m\,|\,\mathbf{Q}_j)}, \qquad i = 1, 2, ..., c \tag{8}$$

Then, $\mathbf{w}_m = (w_{m1}, w_{m2}, ..., w_{mc})$ would represent a vector of probabilities of the origin of m^{th} mixture individual. The data augmentation algorithm draws the missing stock identity, $\mathbf{z}_m$, for each mixture individual from the multinomial distribution, $\mathbf{z}_m \sim Mult(1, \mathbf{w}_m)$, where the probabilities $\mathbf{w}_m$ are computed from the current values of $\mathbf{p}$ and $\mathbf{Q}$.

Given stock identities of mixture individuals, $\mathbf{Z}^{(k)}$, obtained at the k^{th} step of the algorithm, the posterior density for the unknown stock proportions, $f(\mathbf{p}\,|\,\mathbf{Z}^{(k)})$, is just an updated Dirichlet prior

$$f(\mathbf{p} \mid \mathbf{Z}^{(k)}) = D\left(\frac{1}{c}+\sum_{m=1}^{M} z_{m1}^{(k)},\ \frac{1}{c}+\sum_{m=1}^{M} z_{m2}^{(k)},\ \ldots,\ \frac{1}{c}+\sum_{m=1}^{M} z_{mc}^{(k)}\right).$$

The posterior distribution for allele RFs given the baseline and mixture samples, is also a Dirichlet

$$f(\mathbf{Q}_i \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}^{(k)}) = \prod_{h=1}^{H} f(\mathbf{q}_{ih} \mid \mathbf{X}, \mathbf{y}_{ih}, \mathbf{Z}^{(k)})$$

$$= \prod_{h=1}^{H} D\left(\beta_{h1} + y_{ih1} + \sum_{m=1}^{M}\left(z_{mi}^{(k)} x_{mh1}\right),\ \ldots,\ \beta_{hJ_h} + y_{ihJ_h} + \sum_{m=1}^{M}\left(z_{mi}^{(k)} x_{mhJ_h}\right)\right),$$

$$i = 1, \ldots, c.$$

As one can see, each term in the last expression is the sum of the prior parameter, allele counts from the baseline sample, and allele counts from the mixture sample.

To start drawing samples from the posterior distributions, one needs to specify the initial (or starting) values of $\mathbf{p}^{(0)}$ and $\mathbf{Q}^{(0)}$. Consecutively, after the initial sample is obtained, a sequence of samples is drawn with each sample dependent only on the preceding sample, thus making the algorithm Markov Chain Monte Carlo (MCMC). At the k^{th} cycle of the algorithm, two steps are performed:

1) Determine stock identities of the mixture individuals, $\mathbf{z}_m^{(k)} \sim Mult(1, \mathbf{w}_m^{(k)})$, using Equation 8 to obtain probabilities of origin, $\mathbf{w}_m^{(k)}$, for m^{th} individual with genotype $\mathbf{X}_m$, and the current values $\mathbf{p} = \mathbf{p}^{(k)}$ and $\mathbf{Q} = \mathbf{Q}^{(k)}$, $m = 1, 2, \ldots, M$.
2) Draw $\mathbf{p}^{(k+1)}$ and $\mathbf{Q}^{(k+1)}$ from their respective posterior densities, $f(\mathbf{p} \mid \mathbf{Z}^{(k)})$, and $f(\mathbf{Q} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}^{(k)})$.

The data augmentation algorithm cycles these two steps and outputs a sequence, or chain, of samples of stock proportions and baseline RFs from the posterior distribution. Since the early samples of a chain are influenced by the starting values of $\mathbf{p}$ and $\mathbf{Q}$, they are discarded for the purpose of making valid inferences. Usually, as a rule of thumb, half the length of a chain is considered as burn-in samples that are discarded; the rest of the chain should accurately represent the posterior distribution, given that convergence of a chain is reached (Gelman et al., 1995). The Raftery and Lewis (1996) diagnostic is used to establish convergence to desired posterior density. This convergence diagnostic determines the number of samples required for estimating quantiles (q) of posterior distributions with a specified accuracy (r) and probability (s). In this study we used the following values for these parameters, recommended by Pella and Masuda (2001): q = 0.975, r = 0.02, and s = 0.95. An initial pilot sample size is first generated, which is used by the diagnostic to compute the recommended number of samples. After an additional number of samples are generated, the diagnostic uses combined samples – the original pilot samples and the first recommendation – to compute the recommended sizes again. This iterative scheme is applied to each chain beginning with a pilot sample size of 235 (the initial number suggested from the chosen values of q, r, and s).

Once convergence of a chain has been reached, the MCMC samples (after burn-in discard) of stock composition estimates are summarized with various statistics: means, standard deviations, and empirical percentiles (5, 50, and 95).

III. Experimental setting

A complete list of 63 baseline populations of sockeye salmon is presented in Table 1. Populations are grouped into twelve reporting regions based on their genetic similarities and geographic location.

Table 1. List of baseline populations and their affiliation to reporting regions.

Pop #	Population Name	Group	Reporting Region
1	Bear River Weir Early 2000	1	N. Peninsula
2	Bear River Weir Late 2000	1	N. Peninsula
3	Nelson River 2000	1	N. Peninsula
4	Ugashik Outlet 2000	2	Ugashik
5	Ugashik Narrows 2000	2	Ugashik
6	Ugashik Creek 2001	2	Ugashik
7	Becharof Creek 2000	3	Bacharof Lake, Egegik
8	Cabin Creek 2000	3	Bacharof Lake, Egegik
9	Headwaters Creek early 2000	4	Brooks Lake, Naknek
10	Headwaters Creek late 2001	4	Brooks Lake, Naknek
11	Up-a-tree Creek 2000	4	Brooks Lake, Naknek
12	Margot Creek 2000	5	Naknek Lake, Naknek
13	Idavain Creek 2000	5	Naknek Lake, Naknek
14	American Creek 2000	5	Naknek Lake, Naknek
15	American Creek 2001	5	Naknek Lake, Naknek
16	Kulilk River 2001	6	Nonvianuk Lake, Iliamna L Early
17	Moraine Creek 2001	7	Kukaklek Lake, Kvichak
18	Battle River 2001	7	Kukaklek Lake, Kvichak
19	Illiamna Lake outlet late 1999	8	Iliamna Lake Late, Kvichak
20	Flat Island 2000	6	Iliamna Lake Early, Kvichak
21	Woody Island 2001	6	Iliamna Lake Early, Kvichak
22	Triangle Island 2000	6	Iliamna Lake Early, Kvichak
23	Finger Beach 2000	6	Iliamna Lake Early, Kvichak
24	Knutson Bay Beach 2000	6	Iliamna Lake Early, Kvichak
25	Lower Talaric Creek 2000	6	Iliamna Lake Early, Kvichak
26	Lower Talaric Creek 2001	6	Iliamna Lake Early, Kvichak
27	Dennis Creek 2000	6	Iliamna Lake Early, Kvichak
28	Gibraltar River 2000	6	Iliamna Lake Early, Kvichak
29	Southeast Creek 2000	6	Iliamna Lake Early, Kvichak
30	Dream Creek 2001	6	Iliamna Lake Early, Kvichak
31	Nick N. Creek 2000	6	Iliamna Lake Early, Kvichak
32	Copper River 2000	6	Iliamna Lake Early, Kvichak
33	Tommy Creek 2000	6	Iliamna Lake Early, Kvichak
34	Chinkelyes Creek 2000	6	Iliamna Lake Early, Kvichak
35	Tazimina River 2001	6	Iliamna Lake Early, Kvichak
36	Chulitna Bay Beaches 1999	9	Lake Clark, Kvichak
37	Kijik Lake Beach 2000	9	Lake Clark, Kvichak
38	Kijik River 2001	9	Lake Clark, Kvichak
39	Little Kijik River 2001	9	Lake Clark, Kvichak
40	Upper Tlikakila River 2001	9	Lake Clark, Kvichak
41	Allen River Beach 2000	10	Upper Nushagak River

42	Mulchatna River, site A 2001	10	Upper Nushagak River
43	Koktuli River 2000	10	Upper Nushagak River
44	Upper Nushagak-slough 2001	10	Upper Nushagak River
45	Nuyakuk Lake Beaches 2000	10	Upper Nushagak River
46	Tikchik River 2001	10	Upper Nushagak River
47	Bear Creek, L. Aleknagik 2001	11	Lower Nushagak River
48	Agulowok River 2001	11	Lower Nushagak River
49	Agulukpak River 2001	11	Lower Nushagak River
50	Lake Kulik, Wood R. 2001	11	Lower Nushagak River
51	Gechiak Lake 2000	11	Lower Nushagak River
52	Kamchatka River late 1998	12	Russia
53	Kamchatka River early 1998	12	Russia
54	Hapiza River early 1998	12	Russia
55	Hapiza River late 1998	12	Russia
56	Kitigina River 1998	12	Russia
57	Kirushutk River 2000	12	Russia
58	Olada Bay 2000	12	Russia
59	Ozernaya Bay 2000	12	Russia
60	Ozernaya River 2000	12	Russia
61	Vichenkiya River 2000	12	Russia
62	Bistraya River 1998	12	Russia
63	Bolshaya River 1998	12	Russia

Baseline samples contain information on individual genotypes for eight microsatellite loci (genetic markers) named: *Omy77*, *One102*, *One108*, *One109*, *One111*, *Ots107*, *Ots3*, and *uSat60*. Each locus has a specific number of alleles associated with it. The number of alleles varies from 14 for locus *One107* to 43 for locus *One111*. Allele relative frequencies determined from the baseline samples are used for generating mixtures with known contributions. They are also used in determining the baseline posterior (Equation 5), which serves as a prior for allele RFs.

To test the sensitivity of the Bayesian model, the following simulation study has been designed. For each group/region of populations, mixtures of 200 fish were generated with specific (known) contribution from that group. Seven contribution levels of 20%, 10%, 5%, 4%, 3%, 2%, and 1% were considered for each group. A contribution from a given population was generated by randomly simulating genotypes based on the population's allele relative frequencies, obtained from the baseline samples. Thirty mixtures were generated at each contribution level. There were 28 scenarios in all (four groups times seven contribution levels). The mixtures with known contribution, say θ_A, from a group/region of interest were created using the statistical software package S-Plus 2000 and locally written functions (S+Genetics, ADF&G).

Mixtures of size $M = 200$ were simulated so that each population in, say, group A would have contributed equally to the mixture and the sum of these individual contributions would equal θ_A. The remaining baseline populations contribute evenly to the mixture to make up for the rest ($1 - \theta_A$) part of the mixture. Then, a number of individuals from group A, m_A, can be found as a product of the group contribution, θ_A, and the size of the mixture sample, M, i.e., $m_A = 200 * \theta_A$. For example, for the 5% contribution level ($\theta_A = 0.05$), a simulated mixture will have 10 individuals from the populations of group A. If the number of populations comprising group A is less than or

equal to the cumulative number of individuals that should come from the entire group, then each population will supply at least one individual to the mixture. In the case when the number of populations in a group, say n, is greater than the number of individuals allocated to that group, say m_A, then the distribution of individuals to populations within the group is simulated as multinomial with equal proportions, i.e.,

$$(z_1, z_2, ..., z_n) \sim Mult(m_A, p_1, p_2, ..., p_n), \text{ where } p_1 = p_2 = ... = p_n = \frac{1}{n}.$$

The distribution of individuals among populations of group A, $(z_1, z_2, ..., z_n)$, generated in this way would be even across a relatively large number of mixtures. Thus, in the long run, this procedure maintains equal contributions to the mixtures from all populations in the group.

The Bayesian model was run on each mixture sample. Point estimates of contributing stock proportions and 90% credibility intervals were obtained from each resulting posterior distribution (Appendix 1). A contribution of a group was determined as the sum of contributions for populations comprising a group. Thus, for each scenario with pre-determined contribution level, we would have 30 estimates of a group contribution obtained by running the Bayes model on 30 random mixtures with a priori known contribution from that group. The statistical power to detect a group of populations is defined as the proportion of times (out of 30 runs) that a group has been successfully detected. Detection of a group is equivalent to a hypothesis testing for a non-zero contribution: $H_0 : \theta = 0$ vs. $H_a : \theta > 0$. A decision to reject the null hypothesis is made if the lower limit of the 90% credibility interval for θ is not zero when rounded to two significant digits. With the sample size of 200 fish in the mixture, this would be equivalent to having detected just a single individual from the group in question. In Bayesian terms an interpretation of this rule can be expressed as a request that posterior probability of $\theta > 0.005$ be greater than or equal to 0.95, i.e., $\Pr(\theta > 0.005) \geq 0.95$. If this happens, the null hypothesis is rejected and the group is believed to be present in the mixture. In the opposite case, the data provide insufficient evidence to reject H_0 and the group's contribution, θ, is accepted to be not significantly different form zero.

Detectability of a stock group is naturally expected to be dependent on the group identifiability relative to the rest of the baseline, level of the group's contribution to the mixture, and the size of the mixture sample. The higher identifiability of the populations in a group the better detectability of the group should be seen in the mixed stock analysis. With 100% identifiability (i.e. presence of private allele/s), one should expect perfect detectability for any contribution level down to a single individual in the mixture. Usually, the power to detect small contributions decreases with the contribution level and the mixture sample size. For example, with a mixture size of 100 fish, contribution of 1% will bring just a single individual from a population of interest, which will be really difficult to identify if it does not have any fixed alleles. Increasing sample size to 1000 fish will provide 10 individuals from the population in question, which should make detection easier. In this study mixture size is fixed at 200 fish. So, the primary concern is how fast does the statistical power to detect small contributions from a specific group of stocks decline with decrease in the contribution level.

Before we can begin discussion of the results, it is important to mention that such terms as "estimate", "statistical power", and "hypothesis testing" are pure frequentist

terms and do not make much sense in the strictly Bayesian context. Despite this fact, it is often convenient to adapt and use these classical concepts for interpretation of Bayesian results. First, even though posterior distribution fully describes the state of our knowledge concerning a parameter θ, a single estimate is usually needed for practical purposes. Thus, managers may want to know an estimate of stock contribution in a mixture, rather than its posterior distribution, in order to reach a certain executive decision. As mentioned earlier, posterior mean, median and/or various quantiles can serve as point estimates for an unknown parameter θ. Secondly, the power calculations, as described above, will enable a direct comparison of the Bayesian mixed stock analysis with similar studies (e.g., Reynolds and Templin, 2003) based on the classical maximum likelihood methods.

IV. Results and discussions

The population detectability was analyzed for the four regional groups: Kvichak reporting region (RR), Lake Clark RR, a group of 12 Russian populations, and finally a group of all Bristol Bay stocks combined. The detectability of the first two groups was tested relative to a truncated baseline consisting of the Kvichak and Naknek populations only (32 populations), whereas the last two groups were tested relative to the full baseline of 63 available populations.

The power to detect nonzero contributions for the four regions of interest is shown in Table 2 as well as on Figures 3 and 4. The first two rows of Table 2 and the corresponding plots in Figure 3 show the detection power for populations from Lake Clark and Kvichak regions respectively. The simulations for these two regions were conducted based on the truncated baseline of 32 populations (Kvichak & Naknek RRs). As one can see, the detection power decreases progressively with the decline in contribution level, however it does not drop to zero even at the lowest contribution of 1%. At high contributions of 20% and 10% both groups are detectable with power 1.00. Power declined rapidly below 10%, however it is still relatively high at 3%, 4%, and 5% contributions.

Table 2. Power to detect non-zero contributions for the four reporting regions.

	1%	**2%**	**3%**	**4%**	**5%**	**10%**	**20%**
LakeClark	0.03	0.17	0.60	0.70	0.73	1.00	1.00
Kvichak	0.07	0.17	0.27	0.63	0.63	0.97	1.00
Bristol B.	0	0.23	0.43	0.70	0.87	1.00	1.00
Russia	0.03	0.13	0.33	0.40	0.83	1.00	1.00

Simulations based on the full baseline show that the power to detect the group of Russian stocks and the group of all Bristol Bay stocks declines somewhat slower (Table 2, Figure 4). At the 5% contribution level the two groups can still be detected with high power (0.83 and 0.87 respectively). Even though after the 5% mark a steady decline in power is observed, at the 3% contribution level the Russian group exhibits the moderate detection power of 0.33 and the Bristol Bay group of 0.43.

Figure 3. Power to detect non-zero contributions (Kvichak & Naknek baseline)

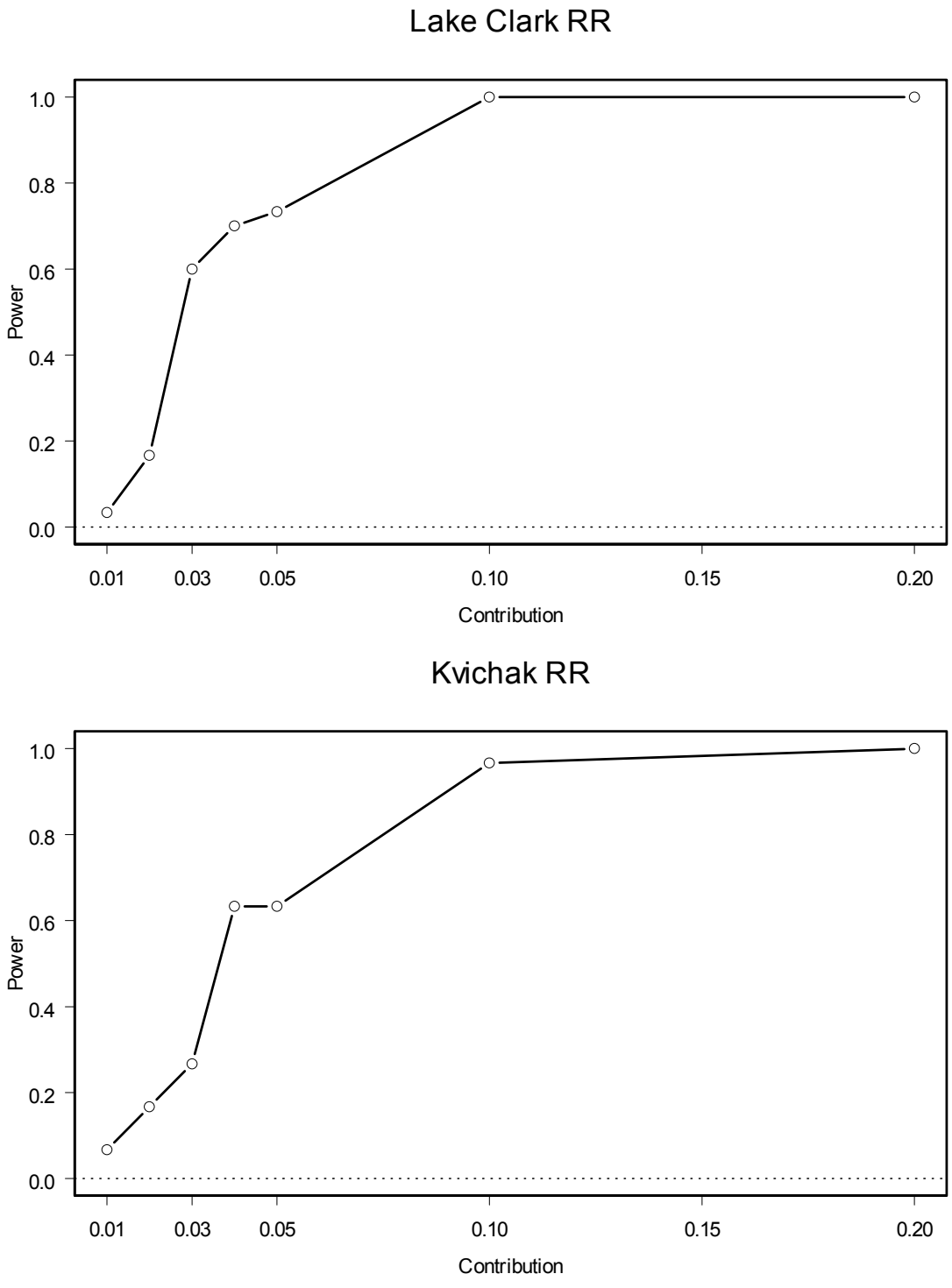
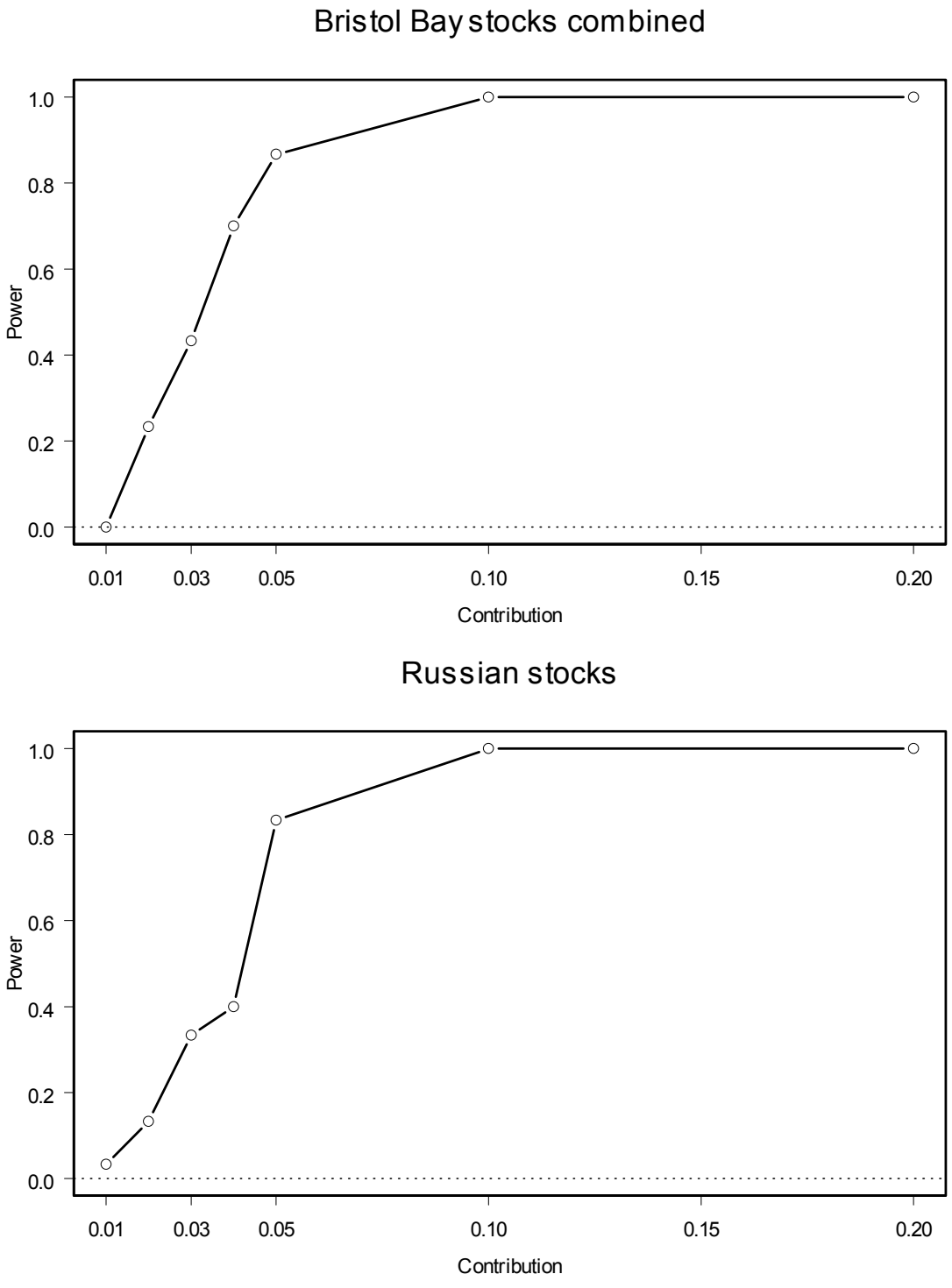


Figure 4. Power to detect non-zero contributions (full baseline)



High detection power down to the 5% contribution level observed for combined Bristol Bay stocks and for the group of Russian stocks can be attributed to large genetic differences between these two groups. In other words, larger differences in

allele relative frequencies allow better identifiability between populations and hence lead to the high detection power by reducing the variability of the posterior distribution of θ.

Simulations for all four groups at the two lowest contribution levels of 1% and 2% yielded detection power of less than 0.25. With the mixture size of 200 fish, 1% and 2% contributions will supply 2 and 4 individuals respectively in a mixture sample from the group of interest. Clearly, with existing genetic relationships between the populations of the four regions and their corresponding baseline stocks, this low number of individuals in a mixture is not enough to consistently detect the presence of the specific group. Since the populations comprising a group are not perfectly identifiable, the few individuals contributed from that group can be adequately explained as having originated from similar populations (see discussion in Reynolds and Templin, 2003).

Reynolds and Templin (2003) and Pella and Milner (1987) show that methods based on maximum likelihood produced increasingly biased estimates of stock contributions as proportions of different stocks in the mixtures became more uneven. Contributions from abundant stocks are often underestimated and those from stocks that contribute little or nothing are usually overestimated. In this study, we consider mean and median of posterior distribution as possible point estimates of stock proportions (histograms of posterior means for each scenario, i.e., group * contribution level are presented in Appendix 2). Each histogram represents a distribution of 30 posterior means obtained by running the Bayesian model on 30 simulated mixtures per scenario. Table 3 summarizes these results showing average across 30 mixtures posterior mean contributions and their standard errors for each scenario.

Table 3. Average across 30 mixtures posterior mean and its standard error.

	1%	**2%**	**3%**	**4%**	**5%**	**10%**	**20%**
LakeClark	0.013 (0.011)	0.019 (0.011)	0.033 (0.015)	0.035 (0.014)	0.051 (0.020)	0.094 (0.022)	0.201 (0.018)
Kvichak	0.029 (0.016)	0.035 (0.016)	0.041 (0.026)	0.057 (0.027)	0.063 (0.024)	0.091 (0.033)	0.178 (0.051)
Bristol B.	0.013 (0.005)	0.022 (0.011)	0.025 (0.010)	0.033 (0.013)	0.041 (0.011)	0.081 (0.018)	0.173 (0.020)
Russia	0.013 (0.008)	0.019 (0.011)	0.027 (0.013)	0.029 (0.014)	0.044 (0.015)	0.101 (0.025)	0.205 (0.024)

As one can see from the table, the averaged across simulated mixtures mean contributions are centered fairly accurately at the pre-assigned values. Their standard errors increase along with the contribution level, however at a much slower rate than the mean values themselves. In other words, the coefficient of variation (not shown) decreases with the increase in contribution level. Kvichak reporting region has the highest variation of the posterior means and the average mean contributions are somewhat off the nominal levels. Kvichak reporting region contains 22 populations and when analyzed against the truncated baseline (with 32 populations total) produces consistent overestimates of the group's contribution at 1% and 2% nominal levels.

The overestimation is likely to be a result of genetic closeness between the Kvichak stocks and the rest of the truncated baseline consisting of the Naknek stocks, which makes it more difficult to correctly identify mixture individuals. For the other three reporting regions, even at the small contributions, the mean estimates are quite precise and less variable.

Medians of posterior distributions represent alternative Bayesian point estimates (Table 4 shows the average medians with their standard errors for the four regions and seven contribution levels). In general, the medians behave similarly to the means being slightly more biased-low at the smaller contribution levels.

Table 4. Average across 30 mixtures posterior median and its standard error.

	1%	**2%**	**3%**	**4%**	**5%**	**10%**	**20%**
LakeClark	0.009 (0.011)	0.016 (0.012)	0.030 (0.015)	0.033 (0.015)	0.048 (0.021)	0.092 (0.022)	0.200 (0.018)
Kvichak	0.023 (0.016)	0.029 (0.016)	0.034 (0.027)	0.051 (0.029)	0.057 (0.025)	0.086 (0.034)	0.175 (0.052)
Bristol B.	0.010 (0.005)	0.019 (0.012)	0.022 (0.010)	0.030 (0.013)	0.038 (0.011)	0.079 (0.018)	0.171 (0.020)
Russia	0.008 (0.009)	0.015 (0.012)	0.023 (0.013)	0.025 (0.015)	0.041 (0.016)	0.099 (0.025)	0.204 (0.024)

Overall, the Bayesian method shows adequate potential in application to the mixed stock analysis and is evidently capable of detecting reasonably small contributions in mixtures of large number of baseline populations. Taking into account its advantages over the widely used maximum likelihood methods, such as better handling of missing data, ability to update the baseline RFs based on the information form mixture samples, and shrinking the baseline RFs to the better established grand or group means, the Bayesian approach is a sound alternative to the CML method.

V. Acknowledgements

This study was made possible by support of the Gene Conservation Lab, Alaska Department of Fish and Game, Anchorage, AK and the University of Alaska Fairbanks, Fairbanks, AK. The idea of this project originated at the Gene Conservation Lab and the lab also provided genetic data on the baseline populations. Bill Templin and Chris Habicht of the Gene Conservation Lab were always ready to answer any questions and offer their competent and much needed help. The work was done using facilities of the Department of Mathematical Sciences, UAF. In particular, an enormous amount of simulations was completed in the local computer lab by running several machines simultaneously.

Special thanks goes to Dr. Joel Reynolds of the Fish & Wildlife Services, Anchorage, AK for sharing his insights into mixed stock analysis, for his invaluable suggestions on how to tackle a controversy of the Bayesian and frequentist approaches, and, finally, for the review of this manuscript.

I am in debt to my teachers and statistical faculty at UAF – Prof.s Dana Thomas, Ron Barry and Shunpu Zhang who not only guided me through the work on this project, but also shared a part of their extensive experience and knowledge during the two years of my masters program. Dana Thomas also reviewed this manuscript and made his valuable comments.

VI. References

Gelman, A., J. B. Carlin, H. J. Stern, and D. B. Rubin.
1995. Bayesian data analysis. Chapman & Hall, New York, NY, 526 p.

Hartl, D.L. and A.G. Clark.
1997. Principles of Population Genetics. Sinauer Associates, Inc., 542 p.

Masuda, M.
2002. User's Manual for BAYES (Bayesian Stock-Mixture Analysis Program). Alaska Fisheries Science Center, Auke Bay Laboratory, Juneau, AK, 28 p.

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

Pella, J. J. and G. B. Milner.
1987. Use of genetic marks in stock composition analysis. In Population genetics and fisheries management (N. Ryman and F. Utter, eds.), p. 247-276. Univ. Washington Press, Seattle, WA.

Reynolds, J. H. and W. D. Templin.
2003. Testing component contributions in finite discrete mixtures: detecting specific populations in mixed stock fisheries. Proceedings of the Annual Meeting of the American Statistical Association, Aug 11-15, 2002, NYC, NY. Section on Statistics and the Environment, pages 2873-2878.

Raftery, A. E., and S.M. Lewis.
1996. Implementing MCMC. In Markov chain Monte Carlo in practice (W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds.), p. 115 – 130. Chapnam & Hall, London

Rice, J. A.
1995. Mathematical Statistics and Data Analysis. Duxbury press, 602 p.

Seeb, L.W., C. Habicht, W.D. Templin, K.E. Tarbox, R.Z. Davis, L.K. Brannian & J.E. Seeb.
2000. Genetic diversity of sockeye salmon of Cook Inlet, Alaska, and its application to management of populations affected by the Exxon Valdez oil spill. Trans. Am. Fish. Soc. 129: 1223-1249.

S+Genetics: S-PLUS Data Structures and Functions for Genetic Data Analysis.
Alaska Department of Fish and Game, Gene Conservation Lab, Anchorage, AK.