Bayesian methods for analysis of stock mixtures from genetic characters

Jerome Pella

Auke Bay Laboratory
Alaska Fisheries Science Center
National Marine Fisheries Service, NOAA
11305 Glacier Hwy.
Juneau, Alaska 99801-8626
E-mail address: Jerry.Pella@noaa.gov

Michele Masuda

Auke Bay Laboratory
Alaska Fisheries Science Center
National Marine Fisheries Service, NOAA
11305 Glacier Hwy.
Juneau, Alaska 99801-8626

Abstract–An implementation of Bayesian methods to assess general stock mixtures is described. An informative prior for genetic characters of the separate stocks in a mixture is derived from baseline samples. A neutral, low-information prior is used for the stock proportions in the mixture. A Gibbs sampler—the data augmentation algorithm—is used to alternately generate samples from the posterior distribution for the genetic parameters of the separate stocks and for the stock proportions in the mixture. The posterior distribution incorporates the information about genetic characters in the baseline samples, including relatedness of stocks, with that in the stock-mixture sample to better estimate genotypic composition of the separate stocks. Advantages over usual likelihood methods include greater realism in model assumptions, better flexibility in applications, especially those with missing data, and consequent improved estimation of stock-mixture proportions from the contributing stocks. Two challenging applications illustrate the technique and its advantages.

Manuscript accepted 13 September 2000.
Fish. Bull. 99:151–167 (2001).

Fisheries that exploit mixed stocks are very common, and their management oftentimes requires assessment of composition of the mixed catches (Begg et al., 1999). Multilocus genotypes of fish are a natural tag by which to infer their origins. The unknown proportions from stocks comprising a stock mixture, or its stock composition, can be estimated from genotype counts in a random sample of the stock-mixture individuals if relative frequencies (RFs) of the genotypes vary among the contributing stocks. Larger differences in genotypic RFs among stocks result in more accurate and precise stock composition estimates. The conditional maximum likelihood (CML) method (Fournier et al., 1984; Millar, 1987; Pella and Milner, 1987) has most commonly been used for stock-mixture analysis. Baseline samples drawn from the separate contributors are used in estimating the RFs of the observed stock-mixture genotypes in each stock. 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. The baseline multilocus genotype RFs determine the outcome of a stock-mixture analysis. Larger errors in these estimated RFs result in larger stock composition errors. Usually the variation in CML stock composition estimates from baseline and stock-mixture sampling is evaluated by the bootstrap method.

The estimation of the baseline multilocus genotype RFs depends on the mode of inheritance of the observed markers. Among molecular markers developed for fish, allozymes, mitochondrial DNA (mtDNA), minisatellite DNA, and microsatellite DNA are widely known for their utility in stock-mixture analysis. For mtDNA, the entire haplotype passes as a unit from female to offspring and the baseline multilocus haplotype RFs are estimated directly by their observed RFs in the baseline samples. For allozymes, minisatellite DNA, and microsatellite DNA, the multilocus genotypes pass from parents to offspring under the usual rules of diploid inheritance. The expected multilocus genotype RFs for diploids equal the products of the genotypic RFs at the individual loci (or subsets of them) that pass independently from parents to offspring. In the special case when Hardy-Weinberg equilibrium holds at a locus, its expected genotypic RFs are determined by its allele RFs: the homozygote RFs equal the squares of their allele RFs and the heterozygote RFs equal twice the product of their allele RFs. To compute the estimated baseline multilocus genotype RFs for diploids, observed RFs of alleles or genotypes in the baseline samples replace corresponding unknown RFs. The fewer the number of unknown RFs that need to be estimated from the available baseline samples, the more reliable the estimated baseline multilocus genotype RFs become (Altham, 1984). Therefore, under Altham's principle, allele RFs are to be estimated if Hardy-Weinberg equilibrium holds, and genotype RFs at the locus otherwise. If alleles among a linked subset of loci are not inherited independently, the multilocus genotype RFs for the subset have to be estimated directly from their observed RFs in the baseline samples.

The search for genetic variation by which to distinguish among populations of fish and other marine organisms has provided an embarrassment of riches. The numbers of mtDNA haplotypes (Epifanio et al., 1995; Bowen et al., 1996; Rosel et al., 1999) and alleles at minisatellite and microsatellite loci (O'Connell and Wright, 1997) can often be so large that reliable estimation of their RFs in stocks is a concern. Baseline sample sizes are usually limited, and the relative precision of estimated RFs for the numerous haplotypes, alleles, or genotypes (HAGs) declines with the magnitude of their RFs. Resulting stock composition estimates for the stock mixture may suffer. Grouping or binning of HAGs may help to control variation in estimation of their baseline RFs (O'Connell and Wright, 1997). However, for stock-mixture analysis, the practical details of grouping so as to balance the loss of information about the mixture against the gain in precision of the baseline RFs are unresolved. Bayesian methods developed for estimating allele RFs at a locus (sec. 3.7 of Lange, 1997) and for estimating cell probabilities in contingency tables (Bishop et al., 1975; Sutherland et al., 1975; Leonard, 1977) offer another attack on the problem. The Bayesian methods are applied later in estimating the RFs by using genetic similarities of stocks.

Conditional maximum likelihood does not use the information in the stock-mixture sample to improve the estimates of baseline multilocus genotype RFs, and the omission becomes ever more meaningful with the accumulation of mixture individuals from a series of analyses performed on stock mixtures of the same baseline populations. The unconditional maximum likelihood (UML) (Pella and Milner, 1987) or unconditional least squares (ULS) (Xu et al., 1994) methods have been suggested to remedy this shortcoming for analysis of a single mixture. For either approach, estimates are provided both for the stock proportions and baseline genotypic RFs by optimizing a criterion of fit to counts in both the baseline and stock-mixture samples. However, the fitting criteria may have local optima (Smouse et al., 1990), and effective search for the global optimum and corresponding estimates from both methods is unresolved. A practical compromise for either method is to find a particular local optimum by starting the search from the CML estimate of stock proportions and stock genotypic RFs, the latter evaluated from the baseline samples alone.

None of the past approaches—CML, UML, or ULS—makes use of the genetic similarities among stocks to estimate the relative frequencies of haplotypes, alleles, or genotypes more accurately in the separate stocks. Common HAGs are shared universally among stocks; HAGs with moderate RFs are shared at least regionally, and rare HAGs occur only sporadically. Instead, the similarities in HAG RFs are viewed solely as limiting success in distinguishing the origins of the stock-mixture individuals. Improved estimates of HAG RFs, to replace simple observed values, would generally benefit accuracy and precision of stock composition estimation, especially as the number of rare or uncommon HAGs increases (e.g. Xu et al., 1994). Estimation of RFs for rare HAGs in separate stocks from baseline samples is especially problematic. Even when present in a population, they may well be absent from the baseline sample. The Bayesian proposal for stock-mixture analysis will shrink the observed baseline HAG RFs of individual stocks toward better-established grand, regional, or group means in order to control HAG RF estimation error.

All past approaches—CML, UML, or ULS—produce estimates of stock proportions that become increasingly biased as the true stock-mixture proportions become more uneven (Pella and Milner, 1987; Xu et al., 1994). Contributions from abundant stocks are underestimated and those from less common or even absent stocks are overestimated. No effective general solution for this bias has been proposed. The Bayesian proposal results in a probability distribution for the stock composition estimates, the location of which can be characterized by various measures, such as the mean, median, and mode, which differ in their bias when viewed as potential point estimators.

Finally, the previous estimation methods appear limited in capacity to attack practical problems that fail to fit the standard mold of a sampled stock mixture and complete baseline. In particular, missing information for entire stocks from the baseline is very difficult to accommodate (Smouse et al., 1990). Despite their availability for a decade or more, nothing has been accomplished using these methods to incorporate genetic similarities of baseline stocks to deal more realistically with missing data. The Bayesian proposal will initially fill in missing baseline HAG RFs with appropriate grand, regional, or group means, proxies that are revised later during analysis of the stock-mixture sample.

Bayes methods have the potential to correct for these shortcomings better than the likelihood or least squares methods. In our study we describe the rationale for this new approach to stock-mixture analysis, develop the statistical models, and outline the numerical algorithms by which to quantify uncertainty in stock proportions of the mixture as well as in the baseline HAG RFs in the separate contributing stocks. Software developed for performing the computations and summarizing results is available at our anonymous ftp site, with address *ftp://wwwabl.afsc.noaa.gov/sida/mixture-analysis/bayes.* Two applications with special difficulties are used as illustrations. First, a winter stock mixture thought to be composed of four Northwest Atlantic harbor porpoise (*Phocoena phocoena*) populations is assessed. These porpoise populations are characterized by mtDNA haplotypes, the number of which exceeds baseline and stock-mixture sample sizes. Second, a Southeast Alaska steelhead trout (*Oncorhynchus mykiss*) stock mixture is resolved to two populations, only one of which could be sampled separately. Allozymes, microsatellites, and mtDNA were available for independent and confirming assessments of stock-mixture proportions from the two populations.

Methods

The premise of the Bayes method for estimating an unknown quantity, θ, is that some information about θ is available before sampling begins. This information is in the form of a prior probability density, $\pi(\theta)$. After sampling, the new data obtained, $\boldsymbol{Y}$, are used to revise the prior to the posterior probability density for the unknown, $\pi(\theta \mid \boldsymbol{Y})$. The posterior is obtained by application of Bayes's theorem, which states that the posterior is proportional to the product of the prior and the likelihood of the sample, $\pi(\boldsymbol{Y} \mid \theta)$, viz.

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) = k^{-1}\pi(\boldsymbol{\theta})\pi(\boldsymbol{Y} \mid \boldsymbol{\theta}), \tag{1}$$

where $k = \int_{\theta} \pi(\boldsymbol{\theta})\pi(\boldsymbol{Y} \mid \boldsymbol{\theta})d\boldsymbol{\theta}.$

Once the posterior for θ is known, a variety of point estimates—mode, median, or mean—as well as the Bayesian posterior probability interval (the apparent counterpart of frequentist confidence intervals, but which actually is a direct probability statement about the unknown) for θ can be derived from it. In stock-mixture analysis, the unknowns separate into two evident blocks [$\theta = (\boldsymbol{p},\boldsymbol{Q})$]: 1) the stock proportions of the mixture, $\boldsymbol{p}$; and 2) the parameters, $\boldsymbol{Q}$, needed to determine the genetic composition—stock-mixture haplotype or multilocus genotype RFs—of the baseline stocks. For haploids, $\boldsymbol{Q}$ represents the baseline haplotype RFs. For diploids, $\boldsymbol{Q}$ represents the array of baseline allele and genotype RFs that are needed under Altham's principle to compute the stock-mixture genotype RFs in the baseline stocks. The new information comes from the baseline and stock-mixture samples for which the likelihood functions are unchanged from earlier likelihood methods. The stock-mixture sample provides counts of haplotypes or multilocus genotypes, and the baseline samples provide counts from the separate stocks of the haplotypes or alleles and genotypes at the loci comprising the mixture multilocus genotypes.

The standard stock-mixture analysis for complete baseline and stock-mixture samples by Bayes methods will be outlined here, with details given in following sections. Extension to nonstandard applications will be indicated by example. First, a prior for $\theta = (\boldsymbol{p},\boldsymbol{Q})$ is developed, which is a product of block priors for its components, $\boldsymbol{p}$ and $\boldsymbol{Q}$. The prior proposed for $\boldsymbol{p}$, which will be called "uninformative," allows any substantive stock-mixture sample information regarding $\boldsymbol{p}$ to overwhelm that from the prior. The prior for $\boldsymbol{Q}$, used to analyze the stock-mixture sample, is informative and will be derived from the baseline samples to quantify uncertainty in the genotypic composition of the contributing stocks. Second, the standard likelihood function for the haplotype or multilocus genotype counts seen in the stock-mixture sample is described. Third, and last, the data augmentation algorithm, a Gibbs sampler, is used to alternately generate a sequence of samples from the posterior distributions for $\boldsymbol{p}$ and $\boldsymbol{Q}$. The stock identities of the mixture-sample individuals are reassigned at each sampling cycle by using a chance mechanism that reflects their uncertainty. The stock identities simplify greatly the revision of the prior distributions for $\boldsymbol{p}$ and $\boldsymbol{Q}$ to account for the stock-mixture sample information; just as with the baseline samples, counts of mixture individuals and their HAGs by stock are available at each cycle. Assignment of individuals to stock origin contrasts with their fractional allocation by the CML method (Pella and Milner, 1987). These samples from the posterior distributions are used to quantify the final uncertainty in $\boldsymbol{p}$ and $\boldsymbol{Q}$ after observing the stock-mixture sample.

Prior for stock-mixture proportions, $\pi(\boldsymbol{p})$

The prior for $\boldsymbol{p}$ can incorporate information about the stock-mixture composition other than that in the stock-mixture sample if such is available. More commonly, however, such information is either unavailable or else the researcher prefers to let that of the stock-mixture sample dominate, just as it does with the earlier likelihood or least squares methods. Therefore, the prior proposed will be restricted, providing no useful information about the stock-mixture composition. Such an uninformative prior for the stock proportions of a c-stock mixture must be defined over the stock composition simplex,

$$S(\mathbf{p}) = \left\{ \mathbf{p}\!: 0 < p_i < 1, \sum_{i=1}^{c} p_i = 1 \right\},$$

and have negligible effect on the posterior distribution. The Dirichlet probability density can accommodate these requirements, and it is natural to use it as a prior with compositional count data both for computational convenience and for its interpretation as additional data. Prior draws of $\boldsymbol{p}$ from the Dirichlet probability density,

$$\pi(\mathbf{p}) = D\left(\mathbf{p} \mid \alpha_1, \alpha_2, \ldots, \alpha_c\right) = \frac{\Gamma\left(\sum_{i=1}^{c} \alpha_i\right)}{\prod_{i=1}^{c} \Gamma(\alpha_i)} \prod_{i=1}^{c} p_i^{\alpha_i - 1}, \tag{2}$$
$$\alpha_i > 0, i = 1, \ldots, c,$$

have means, variances, and covariances given by

$$E(p_i) = \alpha_i / \alpha_0, \ \mathrm{var}\,(p_i) = \alpha_i(\alpha_0 - \alpha_i) / \left(\alpha_0^2(\alpha_0 + 1)\right),$$

$$\mathrm{cov}\,(p_i, p_k) = -\alpha_i\alpha_k / \left(\alpha_0^2(\alpha_0 + 1)\right), i, k = 1, 2, \ldots, c, \text{ and} \tag{3}$$
$$\alpha_0 = \sum_{i=1}^{c} \alpha_i.$$

If a prior draw of $\boldsymbol{p} \sim D(\alpha_1,\alpha_2, \ldots, \alpha_c)$ (the notation "$x \sim f$" means "x is distributed as the probability density or probability function f") was obtained for the stock proportions of a stock mixture, and then a stock-mixture sample of size M was drawn such that the individuals could be correctly identified to stock origin, their counts, $\mathbf{Z}=(z_1,z_2, \ldots, z_c)$, would have a conditional multinomial distribution,

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

or $\boldsymbol{Z} \mid \boldsymbol{p} \sim Mult(M, \boldsymbol{p})$. The posterior for $\boldsymbol{p}$, given $\boldsymbol{Z}$, would be the Dirichlet (computational convenience), $\mathbf{p} \mid \mathbf{Z} \sim D(z_1 + \alpha_1, \ldots, z_c + \alpha_c)$. Notice that the prior parameters enter the posterior density in parallel with the sample counts and therefore could be viewed as counts obtained before the stock mixture was sampled (additional data) (sec. 3.5, Gelman et al., 1995). In fact, the mixture individuals are identified to stock origin (with unavoidable random error) during each cycle of the data augmentation algorithm later when samples are generated from the posterior. With the stock origins identified at a cycle, the uncertainty in $\boldsymbol{p}$ is described by the Dirichlet posterior with parameters equal to the sums of stock counts and prior parameters $(z_i+\alpha_i)$.

With equal values summing to 1 assigned to its parameters or "prior counts," $\alpha_1 = \alpha_2 = \ldots = \alpha_c = c^{-1}$, the Dirichlet prior meets our initial requirements. Specifically, the density is defined over the stock composition simplex, and the additional data, which is neutral in the sense of favoring equal stock proportions (mean stock proportions are c^{-1}), would be equivalent to adding just a single individual to the stock-mixture sample. Means, variances, and covariances (substitute $z_i + c^{-1}$ for α_i in Eq. 3) of the resulting posterior distribution of $\mathbf{p} \mid \mathbf{Z}$, $D(z_1 + c^{-1}, \ldots, z_c + c^{-1})$, approximate closely with increase of stock-mixture sample size, the observed stock proportions, their estimated variances, and their estimated covariances, respectively, from standard frequentist analysis of the multinomial sample, $\boldsymbol{Z}$. Therefore, given the stock assignments of the mixture individuals, the posterior distribution for $\boldsymbol{p}$ will be a reasonable description of its uncertainty for both Bayesian and frequentist statisticians.

Prior for genetic parameters given baseline samples, $\pi(\boldsymbol{Q}|\boldsymbol{Y})$

The genetic compositions—haplotype or multilocus genotype RFs—of the separate stocks are determined by their RFs of haplotypes, alleles, or genotypes, $\boldsymbol{Q}$. An estimate of $\boldsymbol{Q}$ from the baseline samples must be used in place of the unknown $\boldsymbol{Q}$ to estimate the stock genetic compositions. When baseline samples are large, the observed and unbiased value of $\boldsymbol{Q}$, together with measures of precision, may be sufficient to anchor the stock-mixture analysis. Commonly, baseline sample sizes are more limited and some tradeoff between bias and precision (sec 1.4.2 of Carlin and Louis, 1996; Bishop et al., 1975) in estimation of $\boldsymbol{Q}$ may well be advisable. The essential idea is to shrink the observed RFs of HAGs for individual stocks toward central values that are more reliably determined and are consistent with the genetic similarity of the stocks. An informative Bayes prior distribution for these unknown genetic parameters underlying the stock-mixture sample can be derived from the baseline samples and would provide for such shrinkage. The statistical modeling begins with the allele RFs at a single locus but applies equally to haplotypes, alleles, or genotypes. Later, the modeling is extended to cover multiple loci.

The Bayesian scenario begins with an imaginary experiment in which the RFs of the T distinct alleles for a single locus are drawn for each of the c baseline stocks (sec. 3.7 of Lange, 1997). Denote the resulting unobserved RFs for the ith stock by $\boldsymbol{q}_i=(q_{i1}, q_{i2},\ldots, q_{iT})$. The draws from the stocks are independent and from a common Dirichlet probability density, which is the Bayes prior for baseline sampling, $\pi(\boldsymbol{q}_i) = D(\beta_1,\beta_2, \ldots, \beta_T)$. The justification for the Dirichlet prior for baseline sampling parallels that for the earlier stock-mixture composition prior, $\pi(\boldsymbol{p})$, that is, computational convenience and its simple interpretation as additional data. 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—$\boldsymbol{y}_i=(y_{i1},y_{i2}, \ldots, y_{iT})$ for the ith stock—have the multinomial distribution, $Mult(n_i, \boldsymbol{q}_i)$, and therefore the baseline posterior for the unknown allele RFs in each stock is also a Dirichlet distribution, $\boldsymbol{q}_i \mid \boldsymbol{y_i} \sim D(\beta_1+y_{i1},\beta_2+y_{i2}, \ldots, \beta_T+y_{iT})$. The posterior mean of $\boldsymbol{q}_i \mid \boldsymbol{y}_i$ can be written as a weighted average of the observed and prior mean RFs (Bishop et al., 1975; Sutherland et al., 1975),

$$E(q_{it} \mid \beta, \boldsymbol{y_i}) = (y_{it} + \beta_t) / (n_i + \beta_\bullet) =$$

$$\frac{n_i}{n_i + \beta_\bullet}\left(\frac{y_{it}}{n_i}\right) + \frac{\beta_\bullet}{n_i + \beta_\bullet}\left(\frac{\beta_t}{\beta_\bullet}\right), \qquad t = 1,2,\ldots,T, \tag{4}$$

where the observed RF is y_{it}/n_i, its prior mean is $\beta_t/\beta_\bullet$, and $\beta_\bullet = \sum_{t=1}^{T} \beta_t$. If the baseline sample is missing ($n_i=0$), the posterior mean equals the prior mean. Otherwise, the posterior mean ranges between the observed and prior mean RFs (as a function of $\beta_\bullet \geq 0$). Shrinkage from the usual estimator of $\boldsymbol{q}_i$, the observed allele RFs, toward the prior mean increases with the prior "sample size," $\beta_\bullet$, but so does bias in estimates of the allele RFs. Therefore, the magnitude of the prior parameters should be no larger than necessary to satisfactorily control estimation error.

Although the choice of a Dirichlet baseline prior was partly for convenience, the resulting posterior density has good properties. The posterior mean is a reliable estimator for the unknown allele RFs: it is strongly consistent, becomes unbiased for large baseline sample size, and moderates the extremes of the usual estimates—the observed RFs—among baseline stocks. 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 it is nonexistent.

The values for the baseline prior parameters, $\beta_1,\beta_2, \ldots, \beta_T$, have been arbitrary. To complete the specification of the baseline posterior, which will serve as the stock-mixture prior for the allele RFs, their values must be assigned. Two approaches—empirical Bayes and pseudo-Bayes—are considered in which the prior parameters are functions of allele counts in the baseline samples. The empirical Bayes method was previously developed for geneticists to estimate allele RFs (Lange, 1997). In this method, the values assigned to the βs are those which maximize the Bayes prior predictive distribution (Gelman et al., 1995) for the allele counts in the baseline samples. This distribution is the marginal distribution of the allele counts, which results from averaging their multinomial distribution, $Mult(n_i, \boldsymbol{q}_i)$, weighted by the prior probability of $\boldsymbol{q}_i$, $D(\beta_1, \beta_2, \ldots, \beta_T)$. The prior predictive distribution is parameterized by the βs alone, and the optimizing values can be computed from the allele counts (Appndx. 1). Limited experience during this study indicates that, with large baseline samples of a few baseline stocks, or lesser baseline samples for large numbers of baseline stocks, the empirical Bayes method can provide values for the prior parameters, which result in sensible weighting of the observed sample and prior means. Commonly, baseline sampling is more limited, and pragmatism requires a less-demanding alternative method.

The pseudo-Bayes method is based on several practical considerations to determine values for the baseline prior parameters. First, the baseline prior parameters, $\beta_1, \beta_2, \ldots, \beta_T$, have no intrinsic value, other than as tuning parameters by which to perform stock-mixture analysis. Nonetheless, a sound rationale and simple computational formulas for their values are desirable. Second, the prior mean should reflect the similarity of the allele RFs among the baseline stocks. Third, the weights assigned to the prior and observed allele RFs should allow a realistic evaluation of the uncertainty in the genetic composition of the stock, yet not cause misleading bias in the estimated stock composition. Loci with large variation among stocks have more effect on estimated stock composition than those with small variation. Therefore, shrinkage from observed allele RFs toward prior means for loci with large variation should be less than for loci with small variation. If the prior parameter sum, $\beta_{\bullet}$, is substantially smaller than the baseline sample sizes, the bias will be limited. However, with $\beta_{\bullet}=0$, all weight goes to the observed RFs. Then, when a baseline sample misses an allele that is present, sampling error will be underestimated (as it is with bootstrapping under the CML method). Fourth, and last, the weight assigned to the observed RFs for a stock should be positively related to its baseline sample size.

The pseudo-Bayes method of this proposal is original to estimating allele RFs and appears in practice to satisfy the aforementioned criteria. The prior mean will be centered within the observed allele RFs for the stocks of the baseline samples with

$$\beta_t = \tilde{\beta}_{\bullet} \cdot \bar{y}_t, \quad t = 1, 2, \ldots, T,$$

where $\tilde{\beta}_{\bullet}$ = is an estimate (Appndx. 2) of the value for $\beta_{\bullet}$ that minimizes the baseline risk, or expected squared-errors between the posterior means at Equation 4 and the unknown allele RFs of all baseline stocks, and

$\bar{y} = \frac{1}{c}\sum_{i=1}^{c}\frac{y_{it}}{n_i}$ = is the baseline center, or unweighted arithmetic mean, of the observed RFs for the t-th allele among stocks.

With this definition for the βs, the prior mean equals the baseline center. The posterior mean for any stock is the weighted average of its observed allele RFs and the baseline center as at Equation 4. Although the central allele RFs for the entire set of baseline stocks anchors the estimation of $\boldsymbol{Q}$ in this description, extensions to accommodate regional or other groupings of stocks could be accomplished as simply by anchoring on regional or group centers.

Complete analysis of the baseline requires repeated and separate application of the empirical Bayes or pseudo-Bayes methods to each locus. Suppose a total of H loci compose the stock-mixture multilocus genotypes. Let the hth locus have J_h alleles with prior parameters $\boldsymbol{\beta_h} = (\beta_{h1}, \beta_{h2}, \ldots, \beta_{hJ_h})$ and allele RFs in the ith stock of $\boldsymbol{q_{ih}} = (q_{ih1}, q_{ih2}, \ldots, q_{ihJ_h})$. If $\boldsymbol{Q_i}$ denotes the ith stock's combined arrays, $\boldsymbol{q_{i1}}, \boldsymbol{q_{i2}}, \ldots, \boldsymbol{q_{iH}}$, then the prior for the allele RFs of the complete baseline, $\boldsymbol{Q}=(\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_c)$, will be

$$\pi(\mathbf{Q}) = \prod_{i=1}^{c}\pi(\mathbf{Q_i}) = \prod_{i=1}^{c}\prod_{h=1}^{H}\pi(\mathbf{q_{ih}}) = \left(\prod_{h=1}^{H} D(\beta_{h1}, \beta_{h2}, \ldots, \beta_{hJ_h})\right)^{c},$$

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

The baseline samples are drawn independently from the stocks. Denote by $\boldsymbol{Y_i} = (\boldsymbol{y_{i1}}, \boldsymbol{y_{i2}}, \ldots, \boldsymbol{y_{iH}})$ the H arrays of allele counts in the baseline sample for the ith stock, and by $\boldsymbol{Y}$, the entire baseline collection of $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_c}$. Then the Bayesian posterior density for the allele RFs of the entire baseline is the product of Dirichlet densities,

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

and 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 (as at Eq. 4). Although the statistical modeling of the baseline samples has been described with alleles and loci, it applies equally to any combination of independent components: alleles at loci, haplotypes at mtDNA, and genotypes at loci in Hardy-Weinberg disequilibrium.

Stock-mixture sample likelihood function for unknowns, $g(\boldsymbol{X}|\boldsymbol{\theta})$

The stock-mixture sample likelihood function is proportional to the probability of drawing the observed stock-mixture genotypes as a function of the unknowns, $\boldsymbol{p}$ and

$\boldsymbol{Q}$. Denote the count of the jth allele of the hth locus for the mth mixture individual by x_{mhj}. Let the collection of such counts, $\boldsymbol{X_m}$, denote the multilocus genotype of the mth individual, and let the array $\boldsymbol{X}$ denote the collection of such arrays for the M individuals composing the stock-mixture sample. Further, let the RF of individuals with the genotype $\boldsymbol{X_m}$ in the ith stock, which depends on that stock's allele RFs, $\boldsymbol{Q_i}$, be denoted as $f(\boldsymbol{X_m} \mid \boldsymbol{Q_i})$. The RF of the genotype in the stock mixture is the weighted sum, $\sum_{i=1}^{c} p_i f(\boldsymbol{X_m} \mid \boldsymbol{Q_i})$, and so the likelihood function for the stock-mixture sample is

$$g(\boldsymbol{X} \mid \mathbf{p},\mathbf{Q}) = \prod_{m=1}^{M}\left(\sum_{i=1}^{c} p_i f(\mathbf{X_m} \mid \mathbf{Q_i})\right). \qquad (6)$$

In the Bayesian view, $\boldsymbol{X}$ is fixed and $g(\boldsymbol{X} \mid \boldsymbol{p},\boldsymbol{Q})$ is a random function of the unknowns, $\boldsymbol{p}$ and $\boldsymbol{Q}$. Again, although the likelihood function for the stock-mixture genotypes has been described with alleles and loci, it applies equally to stock-mixture genotypes of any combination of independent components: alleles at loci, haplotypes at mtDNA, and genotypes at loci in Hardy-Weinberg disequilibrium.

Posterior distribution of the unknowns, $\pi(\theta \mid X,Y)$

The Bayesian assessment of the unknown stock proportions in the stock mixture and of the baseline RFs of haplotypes, alleles, or genotypes is provided by their joint posterior distribution. This posterior distribution is proportional to the product of the prior density for the unknowns and the likelihood function of the stock-mixture sample, given the unknowns. The prior density for the stock-mixture proportions is the uninformative Dirichlet of Equation 2. The baseline posterior at Equation 5 becomes the stock-mixture prior for the HAG RFs. Prior information on stock-mixture composition and the HAG RFs is reasonably considered independent, so the joint prior for the unknowns is the product (Eqs. 2 and 5),

$$\pi(\mathbf{p},\mathbf{Q}) = \pi(\mathbf{p})\pi(\mathbf{Q} \mid \mathbf{Y}). \qquad (7)$$

The posterior distribution for $\boldsymbol{p}$ and $\boldsymbol{Q}$ with the stock-mixture sample observed, $\pi(\boldsymbol{p},\boldsymbol{Q} \mid \boldsymbol{X},\boldsymbol{Y})$, is proportional to the product of their likelihood at Equation 6 and their prior at Equation 7. Analytic evaluation of the posterior distribution is impractical because of the prodigious computation required, caused by the combinatorial explosion of terms in the likelihood function with increase in stock-mixture sample size (Bernardo and Girón, 1988). Instead, a sufficient number of samples are drawn sequentially from the posterior distribution to accurately describe it. The data augmentation algorithm (Tanner and Wong, 1987; Diebolt and Robert, 1994) can be used to draw the sequence of samples. The idea underlying the algorithm is that the estimation problem would be much simplified if the stock identities of the mixture individuals were known. Given the stock identities, the posterior distribution for the stock proportions and HAG RFs in the baseline stocks simply requires updating of the Dirichlet priors with multinomial counts from the stock mixture.

The stock identities of the mixture individuals are determined by chance in the data augmentation algorithm. Let $\boldsymbol{z_m}=(z_{m1},z_{m2}, \ldots , z_{mc})$ indicate the stock origin of the mth mixture individual by a single "1" at the coordinate of the contributing stock, and c–1 "0"s at the remaining coordinates. For later reference, let $\boldsymbol{Z}=(\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots , \boldsymbol{z_M})$ denote the stock origins of all the mixture individuals. If $\boldsymbol{p}$ and $\boldsymbol{Q}$ were known, the proportion of mixture individuals with genotype $\boldsymbol{X_m}$ that came from the ith stock could be calculated as

$$w_{mi} = p_i f(\boldsymbol{X_m} \mid \boldsymbol{Q_i}) / \sum_{k=1}^{c} p_k f(\boldsymbol{X_m} \mid \boldsymbol{Q}_k), \qquad i = 1,2,\ldots c. \quad (8)$$

Equivalently, the probability that a randomly drawn mixture individual with genotype $\boldsymbol{X_m}$ came from the ith stock is w_{mi} of Equation 8. The data augmentation algorithm draws the missing stock identity, $\boldsymbol{z_m}$, for each mixture individual from the multinomial distribution, $\boldsymbol{z_m} \sim Mult(1,\boldsymbol{w_m})$, where the probabilities for the stocks listed by $\boldsymbol{w_m}=(w_{m1}, w_{m2}, \ldots ,w_{mc})$ are computed from the current samples of $\boldsymbol{p}$ and $\boldsymbol{Q}$. Colloquially, the stock identity of each stock-mixture individual is randomly assigned with the probability for any stock equal to the stock-mixture fraction of the genotype contributed by the stock.

In broad outline, the data augmentation algorithm used to draw posterior samples is straightforward. After the initial sample is obtained (as described later), a sequence of samples is drawn with each sample dependent only on the preceding sample, that is, the algorithm is a Markov chain Monte Carlo (MCMC) method. At the kth sample, two steps are performed:

1 Draw stock identities of the mixture individuals, $\boldsymbol{z}^{(\boldsymbol{k})}_{\boldsymbol{m}} \sim$ Mult$(1,\boldsymbol{w}^{(\boldsymbol{k})}_{\boldsymbol{m}})$, using Equation 8 for genotype $\boldsymbol{X_m}$ and the current values $\boldsymbol{p}=\boldsymbol{p}^{(k)}$ and $\boldsymbol{Q}=\boldsymbol{Q}^{(\boldsymbol{k})}$, m=1,2, . . . , M.

2 Draw $\boldsymbol{p}^{(k+1)}$ and $\boldsymbol{Q}^{(k+1)}$ from their respective posterior densities, $\pi(\boldsymbol{p} \mid \boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}^{(k)})$, and $\pi(\boldsymbol{Q} \mid \boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}^{(k)})$.

The stock identities, $\boldsymbol{Z}^{(\boldsymbol{k})}$, of the stock-mixture sample are sufficient statistics for $\boldsymbol{p}$ (Pella et al., 1996). With them available, the genetic data of the stock-mixture sample is of no value to estimation of $\boldsymbol{p}$. Therefore, the posterior for $\boldsymbol{p}$ is obtained by updating the Dirichlet prior for $\boldsymbol{p}$ with the counts of stock identities for the mixture individuals,

$$\pi(\boldsymbol{p} \mid \boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}^{(k)}) = \pi(\boldsymbol{p} \mid \boldsymbol{Z}^{(k)}) = D\left(\frac{1}{c}+\sum_{m=1}^{M} z_{m1}^{(k)}, \frac{1}{c}+\sum_{m=1}^{M} z_{m2}^{(k)}, \ldots, \frac{1}{c}+\sum_{m=1}^{M} z_{mc}^{(k)}\right). \qquad (9)$$

The posterior density for HAG RFs of the genetic components, $\pi(\boldsymbol{Q} \mid \boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}^{(k)})$, updates the stock-mixture prior, or baseline posterior, $\pi(\boldsymbol{Q} \mid \boldsymbol{Y})$, at Equation 5 for the HAG counts from the identified mixture individuals as

$$\pi(\mathbf{Q_i} \mid \mathbf{X},\mathbf{Y},\mathbf{Z^{(k)}}) = \prod_{h=1}^{H} \pi(\mathbf{q_{ih}} \mid \mathbf{X},\mathbf{y_{ih}},\mathbf{Z^{(k)}}) = \qquad (10)$$

$$\prod_{h=1}^{H} D\left(\begin{array}{c} \beta_{h1} + y_{ih1} + \sum_{m=1}^{M}(z_{mi}^{(k)}x_{mh1}),\ldots,\beta_{hJ_h} + y_{ihJ_h} + \\ \sum_{m=1}^{M}(z_{mi}^{(k)}x_{mhJ_h}) \end{array}\right), \qquad i = 1,2,\ldots,c. \tag{10 (cont.)}$$

Notice that each of the updated Dirichlet parameters for the HAG RF in the *i*th stock equals the sum of the prior parameter, the HAG count in the baseline sample, and the HAG count for the mixture individuals identified to the stock.

The data augmentation algorithm cycles the two steps and eventually outputs a sequence, or chain, of samples of stock proportions and baseline genetic parameters from the posterior Bayes distribution. However, the early samples of a chain are influenced by the starting values of $\boldsymbol{p}$ and $\boldsymbol{Q}$. To make valid inferences, early burn-in samples must be discarded and a sufficient number of subsequent samples must be generated to accurately describe the posterior. Statistics to determine the number of samples to generate per chain—the Raftery and Lewis (1996) convergence diagnostic—or to monitor convergence of multiple chains to the desired posterior density—the Gelman and Rubin (1992) shrink factor—are widely used in practice and are comparatively inexpensive to compute. Although these two statistics are used in the applications that follow, MCMC research is currently very active (e.g. Brooks and Roberts, 1999), and alternatives may prove to be superior for these purposes. Main interest is usually in the stock composition of a stock mixture and so in the later applications the statistics have been applied only to samples of $\boldsymbol{p}$ even though they can be applied to samples of $\boldsymbol{Q}$ as well. However, convergence of samples for $\boldsymbol{p}$ without corresponding convergence for $\boldsymbol{Q}$ is not thought to be possible.

The diagnostic outlined by Raftery and Lewis (1996) determines the number of samples required for estimating quantiles (q) of posterior distributions with a specified accuracy (r) and probability (s). The FORTRAN implementation of the diagnostic, called GIBBSIT,[1] is applied in later examples (with q=0.975, r=0.02, and s=0.95) to each of several chains of stock proportions generated from different starting values. The diagnostic first requires that an initial pilot sample be generated for each chain, which is used to compute its recommended number of samples. An additional number of samples are generated to satisfy the maximum recommended. The combined samples—original pilot samples and the additional samples—are used with GIBBSIT as pilot samples to compute recommended sample sizes again. Further samples are generated if the maximum recommended sample size for any stock exceeds the number so far generated. This iterative scheme is applied to the first chain, beginning with a pilot sample size of 235 (the initial number recommended from the chosen values of q, r, and s). The other chains are run the length of the first chain and then analyzed separately by GIBBSIT. If GIBBSIT suggests that any of the chains should be longer than the first chain, then all the chains are run for the largest number of samples recommended for all stocks and all chains.

Gelman and Rubin (1992) recommended running a small number of independent chains with dispersed starting points to reduce the possibility that a chain is accepted as representative of the posterior distribution before convergence has occurred. To monitor convergence of the chains to the posterior density, a univariate statistic, called the shrink factor (Gelman and Rubin, 1992), is computed[2] for each of the stock proportions. One chain per stock is started, with most of the stock mixture initially contributed by that stock. Once chains of samples are generated and length determined from the Raftery and Lewis diagnostic, shrink factors for all stocks are computed to verify that the chains have converged. The shrink factor is computed from the second halves of the chains and compares the variation within a single chain for a given parameter to the total variation among the chains. Estimates of shrink factors close to one indicate convergence, and acceptable values are less than 1.2 (see Kass et al., 1998). Because the shrink factor is computed from the second halves of chains, the first halves of chains are discarded as burn-in samples. The purpose of discarding initial burn-in samples is to remove dependence on the starting values. Samples subsequent to the burn-in samples may be thought of as coming from the desired posterior distribution. Once convergence of chains has been verified, the MCMC samples (after burn-in discard) of stock composition estimates are combined and summarized with various statistics (equivalent to parameters because of the large samples): means, standard deviations, and empirical percentiles (2.5, 50, and 97.5). Baseline haplotype, allele, or genotype RFs can be summarized similarly.

Checking the fit of the stock-mixture model

Current stock-mixture modeling presumes that a stock mixture composed of random genotypes from the contributing stocks occurs and that a simple random sample of the mixture individuals has been drawn. Further, it is assumed that simple random samples of HAGs from all contributing stocks are available by which to estimate, with an appropriate and known genetic model, the stock-mixture genotype RFs in each of the separate stocks. Samples are considered small in relation to the populations sampled, so that the multinomial distribution can be used to describe sampling variation in counts. These assumptions may be plausible in many applications, but violations can also occur. For example, baseline samples of juvenile salmon, drawn before their families have mixed, would have extra-multinomial variation (Waples

[1] FORTRAN program GIBBSIT (version 2.0) can be obtained without cost from the general archive at *http: //lib.stat.cmu.edu/*.

[2] Our current FORTRAN program for the diagnostic is a translation of Gelman's S function itsim (free from the Statlib S archive at *http://lib.stat.cmu.edu/*), and its modification (version 0.4) in CODA (Best et al., 1995) (free from the MRC Biostatistics Unit, University of Cambridge, at *http://www.mrc-bsu.cam.ac.uk/bugs/*).

and Teel, 1990), as would stock-mixture samples if the populations segregated (McKinnell et al., 1997). The Bayes method opens the way for checking that the models fit. Lack of fit is indicated if the observed samples are unusual realizations of the Bayes posterior predictive distribution (Gelman et al., 1995). Test statistics should be designed to detect suspected problems in stock-mixture analysis: unrepresentative samples, unsatisfactory priors, presence of extra stocks, etc. In particular, the haplotype, allele, or genotype counts in the actual baseline samples should not be outliers of their corresponding predictive distribution. When violations to the assumptions are detected, the posterior distribution of stock proportions and baseline genetic parameters would be misleading. New samples drawn by improved design, or alternate sampling models, could be needed to make the stock-mixture analysis trustworthy.

Samples are easily drawn from the posterior predictive distribution. The kth predictive baseline sample for the HAG counts in a stock is simply a multinomial sample with size equal to that of the actual sample and with probabilities equal to the HAG RFs in the kth posterior sample. The kth predictive stock-mixture sample is obtained in two steps. First, a multinomial sample of M individuals identified to stock is drawn, with probabilities equal to the stock proportions in the kth posterior sample. Second, the stock-mixture genotype of each individual is generated by sequentially drawing the HAGs of the multiple characters by using the HAG RFs for its stock from the kth posterior sample.

Applications

Two applications are considered next to illustrate use of the Bayesian method. In the first application, large numbers of mtDNA haplotypes are present in the baseline and stock-mixture samples and pose special difficulty in analysis. The fairly common availability of mtDNA data makes this application of general interest. In the second application, only one of two populations in a stock mixture could be sampled separately. The Bayesian solution for the missing baseline samples from the second population should be of special interest to biologists concerned with assessing stock mixtures of anadromous and resident populations in streams (Busby et al., 1996; Michael, 1983), and of general interest for extensions to the standard stock-mixture analysis.

Example 1: mtDNA samples from harbor porpoise (*Phocoena phocoena*) of the northwest Atlantic Ocean (Rosel et al., 1999) Rosel et al. (1999) obtained mtDNA sequence data for samples from four summer breeding populations—Gulf of Maine-Bay of Fundy, Gulf of St. Lawrence, Newfoundland, and West Greenland—of harbor porpoises in the northwest Atlantic and from a wintering group along the mid-Atlantic states. The authors were reasonably certain that the wintering group comprised one or more of the summer populations. Because of special conservation concerns for the Gulf of Maine-Bay of Fundy population, the authors wished to determine if it alone could have been the wintering group. Contingency table analysis of the mtDNA haplotype frequencies indicated only that the Gulf of Maine-Bay of Fundy population was almost surely not alone (P<0.06), if at all present. Rosel et al. (1999) used a stock-mixture analysis by the CML method to attempt to delimit the population contributions better with the mtDNA data. Here the Bayesian method is applied to the same data for comparison. Summer sample sizes for each of the populations were between 40 and 80 individuals, and the winter sample size was 41. A total of 67 distinct haplotypes was observed in the summer samples, and the winter sample of 41 individuals included an additional 8 singleton haplotypes previously unseen. Among the total of 253 individuals of all samples, the five most numerous haplotypes were represented by 45 (18%), 42 (17%), 15 (6%), 9 (4%), and 7 (3%) individuals. Most haplotypes were sporadic in samples; the most common counts in the summer and winter samples being 0 and 1. The occurrence of a few fairly common and many scarce haplotypes is characteristic of mtDNA data (Xu et al., 1994) and poses special difficulty in estimation. For example, under the CML method, stock-mixture haplotypes contributed by a particular stock have another apparent source if absent from its baseline sample.

Four chains of samples were generated by data augmentation with both the empirical Bayes and pseudo-Bayes methods for specifying the baseline prior “count” parameters. The total prior “sample size,” $\beta. = \sum \beta_t$, computed by the methods was 22 (pseudo-Bayes) and nearly 2000 (empirical Bayes). An initial pilot chain of 235 samples was analyzed by using the FORTRAN implementation of GIBBSIT, which indicated that chains of 2012 samples should be run (given q=0.975, r=0.02, and s=0.95). The four chains were begun with diverse values for population proportions: one chain was begun for each population, with it composing 0.95 of the stock mixture and the other three populations composed equal parts (thirds) of the remainder (0.05). The four chains had mixed sufficiently, or converged, by their second halves so that the Gelman-Rubin shrink factors were less than 1.03 for any one population. The samples from the second halves were pooled to represent 4024 draws from the posterior distribution. Predictive baseline samples were generated from the posterior samples for haplotype RFs, and indicated lack of fit only for the empirical Bayes method (Fig. 1). Therefore, only the posterior distribution from the pseudo-Bayes method will be described further. Parameters for population proportions computed from the posterior sample (Fig. 2) include the mean, mode, median, standard deviations, and equal-tail bounds of posterior intervals (Table 1). Conditional maximum likelihood estimates for the winter sample were computed for comparison, along with bootstrap evaluation of their precision from 1000 resamplings. Corresponding statistics of the bootstrap sample for the CML method are the means, standard errors, and 95% confidence bounds (Table 2). This CML analysis differs from that of Rosel et al. (1999) by using 1) the counts of all individual haplotypes instead of pooling to form subsets with larger counts, and 2) an alternate method for constructing confidence bounds. Rosel et al. (1999) used the

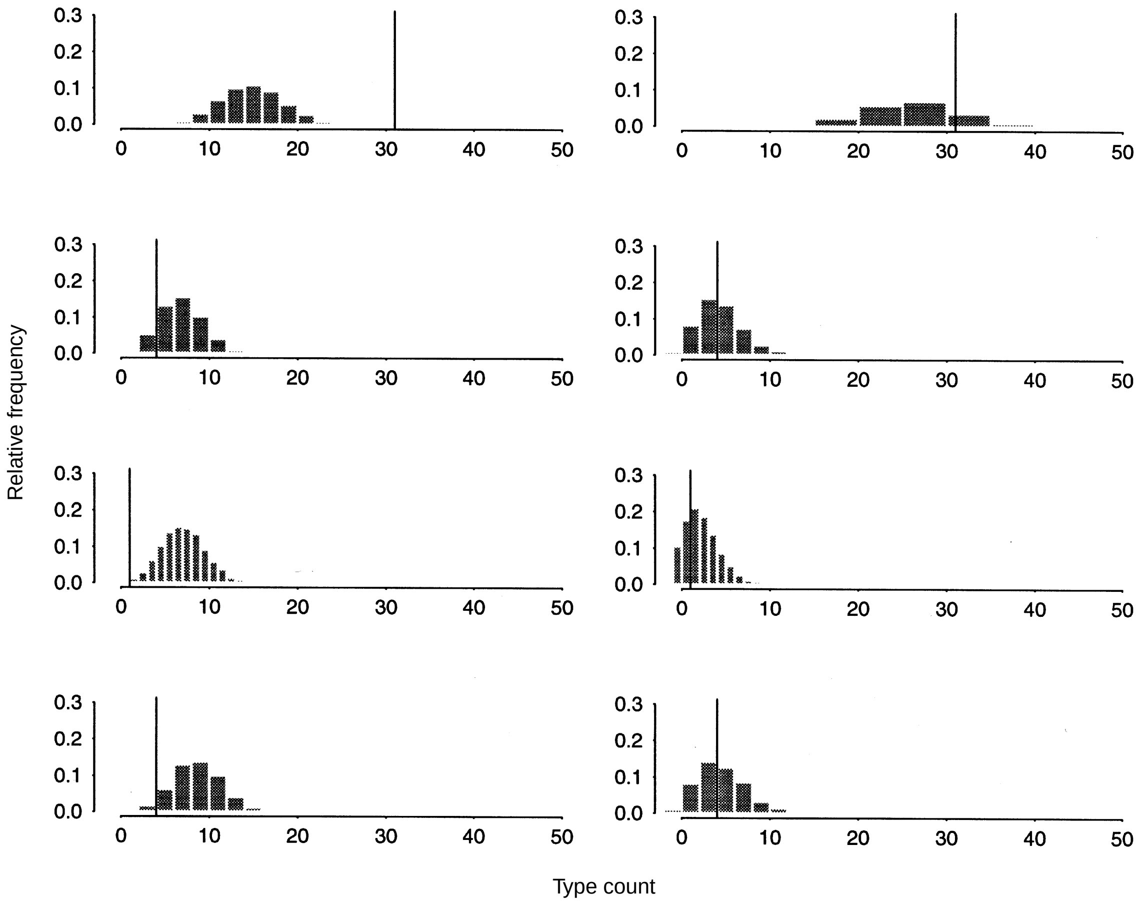

Figure 1

Histograms of predictive baseline population (Gulf of Maine-Bay of Fundy, top; Gulf of St. Lawrence, second; Newfoundland, third; and West Greenland, bottom) sample counts for the most common haplotype in the pooled summer and winter samples, by empirical Bayes (left) and pseudo-Bayes (right) methods. The actual count of the haplotype in each baseline sample is shown as a spike.

Table 1

Parameters of the posterior density for harbor porpoise population proportions composing the winter stock mixture. Reported proportions do not necessarily sum to 1.0 because they are rounded.

				Posterior quantiles		
Population	Mean	Mode[1]	SD	2.5%	Median	97.5%
Gulf of Maine-Bay of Fundy	0.12	0.02	0.13	0.00	0.08	0.46
Gulf of St. Lawrence	0.48	0.69	0.19	0.14	0.48	0.84
Newfoundland	0.15	0.02	0.16	0.00	0.10	0.52
West Greenland	0.24	0.26	0.18	0.00	0.22	0.66

[1] The mode is computed by 4-dimensional binning of the Markov chain Monte Carlo samples for stock proportions, each bin with sides of 0.05, and then normalizing the bin center having maximum count.

percentile interval (Efron and Tibshirani, 1993) for confidence bounds. The alternate method, called the nonsymmetric percentile bootstrap (Lunneborg, 2000), is expected to have superior coverage properties to the standard percentile method for the usual skew distributions of stock-mixture composition estimates. The confidence

Table 2

The conditional maximum likelihood point estimate for harbor porpoise population proportions composing the winter stock mixture, and its bootstrap standard error and 95% confidence bounds (nonsymmetric percentile method). Reported proportions do not necessarily sum to 1.0 because they are rounded.

			95% Confidence bounds	
Population	Point estimate	SE[1]	Lower	Upper
Gulf of Maine-Bay of Fundy	0.19	0.14	0.00	0.37
Gulf of St. Lawrence	0.40	0.16	0.13	0.77
Newfoundland	0.18	0.15	0.00	0.35
West Greenland	0.24	0.15	0.00	0.48

[1] These standard errors are reduced by 30% to 50% from those reported by Rosel et al. (1999). At our earlier recommendation, the authors pooled subsets of haplotypes without a well-grounded basis in order to avoid the small counts of individual haplotypes used here. The point estimate is unchanged, but the confidence intervals differ mainly because a new method was used in their computation.

bounds were computed with an update of the program, SPAM 3.2 (Debevec et al., in press), available on the internet at *http://www.cf.adfg.state.ak.us/ geninfo/research/ genetics/Software/SpamPage.htm*.

Both the pseudo-Bayes and conditional maximum likelihood methods are in agreement on the population composition of the winter sample in five respects (Tables 1 and 2). First, any of the populations could be involved in the stock mixture and comprise much (upper posterior bounds range from 0.46 to 0.84, and upper confidence bounds, from 0.35 to 0.77) of it. Second, the contributions by any of the populations are very imprecisely determined from the mtDNA counts (widths of all interval estimates exceed 0.35, and standard deviations or standard errors range from 0.13 to 0.19). Third, more than one population seems to be present, given that none of the interval estimates includes 1.0. Fourth, the most frequent estimates, or modes, of the Bayes posterior imply that the mix is almost entirely composed of the Gulf of St. Lawrence and West Greenland populations (Fig. 2). Fifth, and last, the Gulf of St. Lawrence population was almost certainly present (lower 95% posterior bound for the proportion, 0.14; corresponding lower 95% confidence bound, 0.13).

The claim that the Gulf of St. Lawrence population was wintering along the mid-Atlantic coast is important to the conservation issue. Is its presence conspicuous to a direct examination of the genetic samples? The answer is yes if one knows where to look. The Bayes method actually identifies the stock origins of the stock-mixture individuals during generation of each sample from the posterior distribution, and so the posterior identity distribution—the relative frequency of assignment to each population—for each stock-mixture individual is available. The identity distributions of mixture individuals showed 4 of the 33 (12%) winter porpoise were more likely than not (posterior probabilities=0.60, 0.64, 0.67, 0.72) to be from the Gulf of St. Lawrence, and 4 more were more certainly (posterior probabilities=0.87, 0.88, 0.88, and 0.89) from that population. Corresponding probabilities (Eq. 5, Pella and Milner, 1987) from the CML method for the first (0.62, 0.69, 0.70, 0.81) and second groups (1, 1, 1, 1) agreed reasonably. The summer samples contained the following numbers of the same 8 haplotypes: Gulf of Maine-Bay of Fundy, 2 of 80 (3%); Gulf of St. Lawrence, 14 of 40 (35%); Newfoundland, 4 of 42 (10%); and West Greenland, 6 of 50 (12%). Except for the Gulf of St. Lawrence population, in which these haplotypes were fairly common, their observed RFs in the other populations were half or less of that (24%) in the winter sample. With the observed haplotype RFs assumed to be accurate, the probability is less than 0.05 that 8 of 33 individuals with the haplotypes came from any population other than Gulf of St. Lawrence. The conjunction of necessary sampling errors—higher frequencies of the 8 haplotypes in the other populations or lower frequency in the stock mixture—without the presence of Gulf of St. Lawrence is deemed highly improbable from the Bayes computations. Without the posterior probabilities of stock identities, a search for direct evidence of the presence of particular populations in the winter sample would have been futile.

A total of 25 sets of simulated baseline and stock-mixture samples of harbor porpoise mtDNA haplotypes was generated for each of four experiments. Sizes of the simulated samples equaled those of the actual data. The experimental conditions that were controlled include the proportions from the four populations in the stock mixtures and their haplotype RFs. In three of the experiments, the Gulf of St. Lawrence population comprised 0.95 of the stock mixture, and the other stocks comprised equal thirds of the remaining 0.05. In the fourth experiment, the four populations contributed equal parts (0.25) to the stock mixture. The haplotypes of the samples were drawn with replacement from either the original baseline samples (0% addition) or augmented baseline samples for which half (50% addition) or all (100% addition) of the missing haplotypes were replaced by singletons. The conditional maximum likelihood method was applied to each set of simulated samples just as it had to the actual samples. The Bayes method was similarly applied, but with a single exception—a long fixed sequence of 5000 samples (first 2500 discarded as burn-in) was generated for all sets to reduce processing labor. Average point estimates among the 25

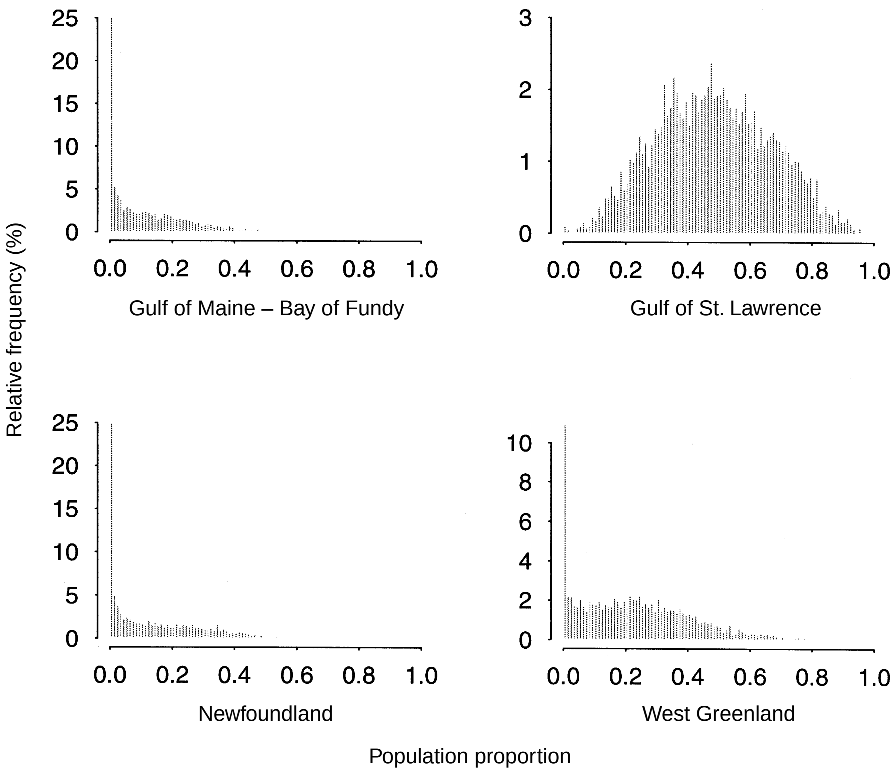

Figure 2

Histograms of samples from the posterior distribution of unknown proportions of summer populations of harbor porpoise (Gulf of Maine-Bay of Fundy, Gulf of St. Lawrence, Newfoundland, and West Greenland) in the wintering group.

sets of samples for each experiment—Bayes mode, Bayes mean, and conditional maximum likelihood (CML) estimate—and their standard errors were computed (Table 3).

The main lesson of these simulations is that the Bayes method, as configured, performs reasonably well in the frequency sense, that is, under repeated sampling. The Bayes posterior mode seems to be a practical point estimator for population proportions in stock mixtures: it was less biased than the Bayes posterior mean and the conditional maximum likelihood estimate when the experimental conditions caused bias. As is characteristic of stock-mixture composition bias, uneven population contributions combined with large variation in estimated contributions were aggravating. When the populations contributed equally (stock mixture 4), bias was negligible for any estimator, given the large variation of estimated population proportions, but with unequal contributions (stock mixtures 1–3), the bias became increasingly severe because the haplotypes were added to the populations and increased the variation in estimated contributions. Lower bias of the Bayes mode was not without cost because its variation among sets of samples was generally larger than that of the Bayes mean or conditional maximum likelihood estimate for the more-difficult third and fourth stock mixtures.

Example 2: Sashin Creek steelhead (*Oncorhynchus mykiss*) stock mixture Sashin Creek on Baranof Island in Southeast Alaska contains a population of anadromous rainbow trout, or steelhead, in its lower portion. In addition, a self-sustaining population above a barrier waterfall was created in 1926 by a transplant from the lower to the upper portion (which includes two lakes). Although the falls was a barrier to upstream migration, migrating juveniles from the upper portion apparently survived the plunge to the lower river. Samples of mature adults returning from the ocean, obtained from the lower portion, were compared with similar samples from the upper population for allozymes (21 loci with 2–6 alleles per locus), microsatellites (10 loci with 3–26 alleles per locus), and mtDNA (5 haplotypes). An excess of homozygotes at loci (Wahlund effect) provided evidence that the samples came from a mixture of both populations. In particular, the allozyme, PGK2, appeared to be fixed (100%) in the upper population, yet the fixed allele represented less than 50% of the PGK2 alleles in the stock-mixture sample from the lower portion. Biologists[3] were able to infer that roughly 25% of the stock mixture probably originated from the upper popula-

[3] Thrower, F. 2000. NMFS, Auke Bay Laboratory, Juneau, AK 99801-8626.

Table 3

Average point estimates—Bayes mode, Bayes mean, and conditional maximum likelihood (CML) estimate—and their standard errors (in parentheses) for 25 simulated samplings of four stock mixtures composed of harbor porpoises from the Gulf of Maine-Bay of Fundy, Gulf of St. Lawrence, Newfoundland, and West Greenland. The haplotypes of stock mixtures were drawn from the original baseline samples (0%) or augmented baseline samples for which half (50%) or all (100%) of the missing haplotypes were replaced by singletons. Baseline and stock-mixture sample sizes were those reported by Rosel et al. (1999) and analyzed earlier in this section. Reported proportions do not necessarily sum to 1.0 because they are rounded.

Stock mixture and estimator	Gulf of Maine-Bay of Fundy	Gulf of St. Lawrence	Newfoundland	West Greenland
Stock mixture 1: 0%	0.95	0.01666	0.01666	0.01666
Bayes mode	0.91 (0.07)	0.03 (0.02)	0.02 (0.00)	0.04 (0.07)
Bayes mean	0.76 (0.17)	0.07 (0.11)	0.09 (0.12)	0.08 (0.12)
CML mean	0.82 (0.09)	0.04 (0.06)	0.07 (0.08)	0.06 (0.08)
Stock mixture 2: 50%	0.95	0.01666	0.01666	0.01666
Bayes mode	0.85 (0.15)	0.04 (0.07)	0.07 (0.11)	0.04 (0.07)
Bayes mean	0.71 (0.20)	0.09 (0.13)	0.11 (0.13)	0.09 (0.13)
CML mean	0.68 (0.13)	0.09 (0.10)	0.14 (0.09)	0.10 (0.09)
Stock mixture 3: 100%	0.95	0.01666	0.01666	0.01666
Bayes mode	0.72 (0.30)	0.08 (0.19)	0.04 (0.06)	0.16 (0.27)
Bayes mean	0.59 (0.24)	0.12 (0.17)	0.12 (0.16)	0.17 (0.20)
CML mean	0.56 (0.13)	0.12 (0.13)	0.13 (0.11)	0.19 (0.14)
Stock mixture 4: 0%	0.25	0.25	0.25	0.25
Bayes mode	0.25 (0.26)	0.18 (0.29)	0.21 (0.22)	0.36 (0.34)
Bayes mean	0.27 (0.22)	0.23 (0.22)	0.24 (0.21)	0.26 (0.23)
CML mean	0.30 (0.14)	0.21 (0.14)	0.24 (0.13)	0.25 (0.15)

tion. Their method depended on the fixed condition of the locus in the upriver population: removal of about 25% of such mixture individuals resulted in the remainder meeting Hardy-Weinberg equilibrium. This approach was difficult to generalize to the other loci, most of which were highly variable. Further, the approach could not provide a complete description of the genetic composition of the lower-river population. To use the information better, all loci available for each type of genetic data were analyzed to provide a Bayes posterior distribution of the population proportions and their allele (allozymes and microsatellites) or haplotype RFs (mtDNA). Each type of genetic data was treated separately in order to examine the consistency of population composition estimates from independent data.

The stock-mixture prior distributions for the baseline characters required some change to accommodate the single-population baseline. As is routine, loci were assumed to have been inherited independently, and their alleles were in Hardy-Weinberg equilibrium for either population. However, the baseline prior parameters (βs) for each genetic character of the upstream population were uninformative: their sum, the baseline prior "sample size," equaled just 1, and each equaled the inverse of the number (J_h) of HAGs. (empirical Bayes or pseudo-Bayes methods for computing the prior parameters were not applicable with a single baseline population.) With the baseline sample counts for locus h denoted as $\mathbf{y_h} = (y_{h1}, \ldots, y_{hJh})'$, the stock-mixture prior (or baseline posterior) of HAG RFs for the upstream population was

$$\pi(\mathbf{Q}_{up} \mid \mathbf{Y}_{up}) = \prod_{h=1}^{H} D(y_{h1} + J_h^{-1}, \ldots, y_{hJ_h} + J_h^{-1}).$$

Notice that the stock-mixture prior "sample size" was the unit-augmented actual sample size, n_h+1, where $n_h = \sum y_{hj}$.

The corresponding downstream population stock-mixture prior reflected the even greater uncertainty in that population's characteristics by a downstream stock-mixture prior "sample size" equal to only 1, yet with average HAG RFs closely approximating those from the counts, $\mathbf{x_h} = (x_{h1}, \ldots, x_{hJh})'$, seen in the stock-mixture sample, viz.

$$\pi(\boldsymbol{Q}_{low} \mid \boldsymbol{X}) = \prod_{h=1}^{H} D\left(\frac{x_{h1} + J_h^{-1}}{\sum_j x_{hj} + 1}, \ldots, \frac{x_{hJ_h} + J_h^{-1}}{\sum_j x_{hj} + 1} \right).$$

The prior for population proportions, $\pi(\boldsymbol{p})$, was the standard Dirichlet, D(0.5,0.5). Three chains were generated, beginning from diverse upriver population proportions of 0.95, 0.50, and 0.05 (Fig. 3). The chain lengths for allozymes and microsatellites were 10,000 samples with the first 5000 discarded as burn-in. The posterior sample comprised the 15,000 samples from their second halves. The chain lengths for mtDNA were 100,000 samples and the posterior sample comprised the 150,000 samples from their second halves.

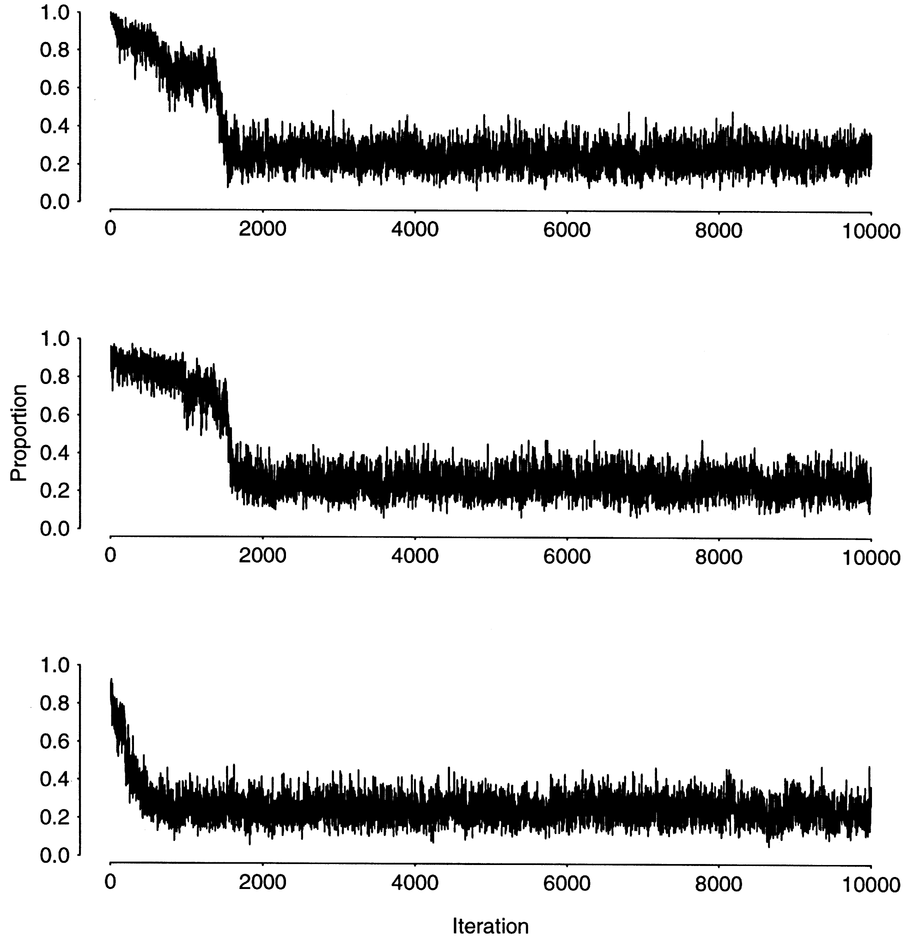

Figure 3

The three chains of posterior samples for the unknown proportion from the upper Sashin Creek steelhead population present in the lower Sashin Creek stock mixture, based on microsatellite data. Each chain was initiated with a different value for the unknown upper proportion, 0.95, 0.50, and 0.05. All chains tracked to high proportions during the early phase and later stabilized at the equilibrium posterior distribution.

Table 4

Parameters of the marginal posterior density for the upriver steelhead proportion in the lower-river stock mixture.

				Posterior quantiles		
Data	Mean	Mode	SD	2.5%	Median	97.5%
Allozymes	0.25	0.22	0.07	0.12	0.24	0.40
Microsatellites	0.24	0.22	0.06	0.13	0.24	0.37
MtDNA	0.41	0.00	0.19	0.01	0.46	0.70

The posterior marginal distributions of the upriver proportion from allozymes and microsatellites were nearly identical and their modes (0.22) or means (0.24, 0.25) were supportive of the earlier approximate assessment of 0.25 (Table 4). Either data set points with high probability (0.95) to an upriver population presence between 0.1 and 0.4 of the lower-river stock mixture. The mtDNA was not nearly as informative in regard to population composition, judging from the resulting posterior marginal distribution. Its standard deviation roughly equaled threefold, and its posterior interval length, twofold, that from allozymes or microsatellites.

Discussion

Analysts, accustomed to using likelihood or least squares methods for stock-mixture problems, should not be deterred by the novelty of the proposed Bayes method. Instead, the Bayesian implementation should be seen as eminently practical and sensible. The data augmentation algorithm recognizes each mixture individual as an entity and labels it with a stock origin. Given the stock assignments, the observed stock proportions are obvious estimates of the stock composition. In concurrence, the means, variances, and covariances of the Bayes posterior distribution for stock proportions approximate closely the observed stock proportions, their estimated variances, and their estimated covariances, respectively, from frequentist methods. Given the current stock proportions and genetic parameters of an MCMC chain, the random labeling accurately reflects the uncertainty in stock origins. Each mixture individual is assigned to one of the baseline stocks, by using probabilities of stocks proportional to each stock's contribution of its genotype to the mixture. Stock proportions and genetic parameters of the MCMC chain gravitate toward their true values because draws from the posterior, which integrates the baseline and stock-mixture information, are more probable nearby.

One goal in developing a Bayesian method of stock-mixture analysis was to replace the conditional maximum likelihood assumption of ignorable baseline sampling error by modeling that acknowledged the uncertainty in genetic composition of the baseline stocks. Ignorable baseline sampling error is especially unrealistic in applications for which uncommon genotypes are present. Stock-mixture individuals, particularly those with uncommon genotypes, may be contributed by stocks whose baseline samples imply their absence. Current bootstrap resampling of the baseline samples does not accommodate reasonably this mismatch between stock-mixture presence and apparent absence of a genotype in a baseline stock. The individual is presumed to come from a stock different from that of the contributor. Such mismatches become frequent when many rare and uncommon genotypes occur, such as in mtDNA data. The simulations for harbor porpoise showed that as greater numbers of rare haplotypes occurred in the populations, the bias of the CML estimator became severe. When none of the baseline samples can explain presence of a stock-mixture genotype, the CML assumption leaves only an outside source. Data sets generated during bootstrap resampling can require an apparent outside source even when the original samples did not. Pooling of uncommon types to circumvent their effects on estimation should be preceded by careful study to assure information useful to stock-mixture composition is not lost.

Some potential to improve stock-mixture assessment by Bayesian methods arises from the prior for θ=($\boldsymbol{p}$,$\mathbf{Q}$), which has no counterpart in the likelihood approach. The Bayes proposal for stock-mixture analysis emulates the objectivity of likelihood methods by letting stock-mixture sample information dominate that of the neutral low-information prior for stock proportions. If information about $\boldsymbol{p}$ is truly unavailable, or the researcher prefers to withhold it and let the "data do the talking," the neutral low-information prior will be adequate. However, the resulting composition estimates may be so imprecise as to be of limited practical value. If additional information is available, either customizing the prior to include it, or updating the posterior (it becomes the prior) with the additional information may improve precision. As an example of updating, the three independent data sets for Sashin Creek steelhead trout could be integrated sequentially into a single posterior for population proportions.

In attempting to maintain objectivity for description of the uncertainty in genotypic composition of the separate stocks, the empirical Bayes method for specifying the baseline prior parameters was examined. The empirical Bayes method for choosing prior parameters for the haplotype, allele, or genotype RFs provided excessive weight to the prior mean for harbor porpoise, with the prior "sample size" parameter, $\beta_{\bullet}$, for harbor porpoise of nearly 2000, many-fold the total of actual sample sizes. In addition, the empirical Bayes method consistently weighted prior means heavily on several applications examined and not reported. Information in these typical baseline samples is evidently inadequate for estimation of the prior parameters. A full Bayesian analysis, which views them as random variables (sec. 5.3 of Gelman et al., 1995), would require an informative prior for them. Because such an informative prior was not evident, the pseudo-Bayes approach (Bishop et al., 1975) was adopted.

Under the pseudo-Bayes approach, the posterior mean for HAG RFs of stocks interpolates between observed values for individual stocks and a baseline central value for all stocks, with the shrinkage, or weighting, determined by the values of the baseline prior parameters. The best choice for values of the prior parameters remains an open question. Possibly, the prior parameter values could be chosen for their performance in experiments of simulated stock-mixture analyses. However, the computations involved would be extensive and without guarantee beforehand of a clear solution. This proposal included an objective criterion—minimum squared-error risk of baseline allele RFs—by which to determine weighting between the prior and observed HAG RFs from the baseline samples alone. Researchers who find choice of weighting a deterrent to application of the Bayes method can set the baseline prior parameters to zero with the qualification that variation of stock proportions may be understated as with the CML method. In return, the Bayes algorithm easily includes the information in the stock-mixture and baseline samples in assessing stock proportions. The simulations for harbor porpoise showed that the weighting from the pseudo-Bayes method resulted in good frequency performance.

In many practical applications, fisheries managers require a point estimate of stock composition. The well-known bias of the conditional maximum likelihood estimate has been troublesome for this reason. Any corrections for its bias have referred estimated stock proportions to simple one-dimensional graphical relationships between simulation averages and known stock proportions. The simulations for harbor porpoise showed that the Bayesian posterior mode had considerably less bias in situations for which the CML estimate was severely biased. The mode has intuitive appeal as the most frequent estimate from the posterior. Its promise for situations requiring point estimates needs to be explored by simulation in further applications. In addition, computation of the multidimensional mode requires some smoothing of the posterior samples, such as the binning used here (see footnote, Table 1). Stock grouping, followed by summing of individual stock proportions for group totals, may also be necessary because finding posterior modes becomes more problematic with large numbers of stocks.

The Dirichlet distribution was the basis for probability modeling because it is a natural choice. First, it is defined for random compositions (i.e. arrays with nonnegative elements that sum to one). Second, the posterior for multinomial data can be written in closed form and is also Dirichlet. Third, the prior parameters can be interpreted as additional data. Fourth, and last, it is easy to sample by computer. However, compositional data are also nicely modeled with the logistic normal density (Aitchison, 1986; Billheimer et al., 1998), whose flexibility and relation to normal theory may have advantages in stock-mixture analysis. Use of geographical structure for the stock proportions in complex stock mixtures comprising many stocks is an area for exploration with the logistic normal and Bayesian hierarchical methods.

Acknowledgments

We much appreciated the opportunity to examine the genetic samples of harbor porpoises and steelhead trout. The harbor porpoise samples were made available by Patricia E. Rosel of the National Ocean Service, NOAA, 219 Fort Johnson Road, Charleston, SC 29412, and the steelhead trout samples were made available by Frank Thrower, National Marine Fisheries Service, Auke Bay Laboratory, Juneau, AK 99801-8626. The manuscript was revised after many insightful comments received from colleagues. Our special thanks for their reviews and ideas go to Eric Anderson of the University of Washington, Joel Reynolds of the Alaska Department of Fish and Game, Peter Smouse of Rutgers University, Jim Murphy and Richard Wilmot of the Auke Bay Laboratory, and three anonymous reviewers of the *Fishery Bulletin*.

Literature cited

Aitchison, J.
1986. The statistical analysis of compositional data. Chapman & Hall, New York, NY, 416 p.

Altham, P. M. E.
1984. Improving the precision of estimation by fitting a model. J. Roy. Statist. Soc. B 46:118–119.

Begg, G. A., K. D. Friedland, and J. B. Pearce.
1999. Stock identification-its role in stock assessment and fisheries management: a selection of papers presented at a symposium of the 128th annual meeting of the American Fisheries Society in Hartford, Connecticut, USA, 23–27 August 1998. Fish. Res. 43/1–3, 249 p.

Bernardo, J. M., and F. J. Girón.
1988. A Bayesian analysis of simple mixture problems. *In* Bayesian statistics 3 (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.), p. 67–78. Oxford Univ. Press, Oxford.

Best, N. G., M. K. Cowles, and S. K. Vines.
1995. CODA: convergence diagnosis and output analysis software for Gibbs sampling output, version 0.3. MRC Biostatistics Unit, Cambridge, UK, 41 p. + addendum.

Billheimer, D., P. Guttorp, and W. F. Fagan
1998. Statistical analysis and interpretation of discrete compositional data. NRCSE Tech. Rep. 11, Univ. Washington, Seattle, WA, 29 p. + appendix.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland.
1975. Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, MA, 557 p.

Bowen, B. W., A. L. Bass, A. Garcia-Rodriguez, C. E. Diez, R. van Dam, A. Bolten, K. A. Bjorndal, M. M. Miyamoto, and R. J. Ferl.
1996. Origin of hawksbill turtles in a Caribbean feeding area as indicated by genetic markers. Ecol. Applications 6:566–572.

Brooks, S. P., and G. Roberts.
1999. On quantile estimation and Markov chain Monte Carlo convergence. Biometrika 86:710–722.

Busby, P. J., T. C. Wainwright, G. J. Bryant, L. J. Lierheimer, R. S. Waples, F. W. Waknitz, and I. V. Lagomarsino.
1996. Status review of west coast steelhead from Washington, Idaho, Oregon, and California. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-NWFSC-27, 261 p.

Carlin, B. P., and T. A. Louis.
1996. Bayes and empirical Bayes methods for data analysis. Chapman & Hall, New York, NY, 399 p.

Debevec, E. M., R. B. Gates, M. Masuda, J. Pella, J. Reynolds, and L. W. Seeb.
In press. SPAM (version 3.2): statistics program for analyzing stock mixtures. J. Heredity.

Diebolt, J., and C. P. Robert.
1994. Estimation of finite stock-mixture distributions through Bayesian sampling. J. Roy. Statist. Soc. B 56:363–375.

Efron, B., and R. J. Tibshirani.
1993. An introduction to the bootstrap. Chapman & Hall, New York, NY, 436 p.

Epifanio, J. M., P. E. Smouse, C. J. Kobak, and B. L. Brown.
1995. Mitochondrial DNA divergence among populations of American shad (*Alosa sapidissima*): how much variation is enough for mixed-stock analysis? Can. J. Fish. Aquat. Sci. 52:1688–1702.

Fournier, D. A., T. D. Beacham, B. E. Riddell, and C. A. Busack.
1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Can. J. Fish. Aquat. Sci. 41:400–408.

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

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

Kass, R.E. (Moderator), B. P. Carlin, A. Gelman, and R. M. Neal (Panelists).
1998. Markov chain Monte Carlo in practice: a roundtable discussion. Am. Statist. 52:93–100.

Lange, K.
1997. Mathematical and statistical methods for genetic analysis. Springer, New York, NY, 265 p.

Leonard, T.
1977. Bayesian simultaneous estimation for several multinomial distributions. Commun. Statist.-Theor. Meth. A6: 619–630.

Lunneborg, C. E.
2000. Data analysis by resampling: concepts and applications. Duxbury Press, Pacific Grove, CA, 568 p. + xvii.

McKinnell, S., J. J. Pella, and M. L. Dahlberg.
1997. Population-specific aggregations of steelhead trout (*Oncorhynchus mykiss*) in the North Pacific Ocean. Can. J. Fish. Aquat. Sci. 54:2368–2376.

Michael, J. H., Jr.
1983. Contribution of cutthroat trout in headwater streams to the sea-run population. Calif. Fish Game 69(2):68–76.

Millar, R. B.
1987. Maximum likelihood estimation of mixed stock fishery composition. Can. J. Fish. Aquat. Sci. 44:583–590.

O'Connell, M., and J. M. Wright.
1997. Microsatellite DNA in fishes. Rev. Fish Biol. Fish. 7:331–363.

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.

Pella, J., M. Masuda, and S. Nelson.
1996. Search algorithms for computing stock composition of a stock-mixture from traits of individuals by maximum likelihood. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-61, 68 p.

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. Chapman & Hall, London.

Rosel, P. E., S. C. France, J. Y. Wang, and T. D. Kocher.
1999. Genetic structure of harbor porpoise *Phocoena phocoena* populations in the Northwest Atlantic based on mitochondrial and nuclear markers. Mol. Ecol. 8:S41–S54.

Smouse, P. E., R. S. Waples, and J. A. Tworek.
1990. A genetic stock-mixture analysis for use with incomplete source population data. Can. J. Fish. Aquat. Sci. 47:620–634.

Sutherland, M., P. W. Holland, and S. E. Fienberg.
1975. Combining Bayes and frequency approaches to estimate a multinomial parameter. *In* Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage (S. Fienberg and A. Zellner, eds.), p. 585–617. North-Holland Publishing Company, New York, NY.

Tanner, M. A., and W. Wong.
1987. The calculation of posterior distributions by data augmentation (with discussion). J. Am. Statist. Assoc. 82: 528–550.

Waples, R. S., and D. J. Teel.
1990. Conservation genetics of Pacific salmon. I. Temporal changes in allele frequency. Conserv. Biol. 4:144–156.

Xu, S., C. J. Kobak, and P. E. Smouse.
1994. Constrained least squares estimation of mixed population stock composition from mtDNA haplotype frequency data. Can. J. Fish. Aquat. Sci. 51:417–425.

Appendix 1—The empirical Bayes method for maximum likelihood estimation of the Dirichlet prior parameters from the prior predictive distribution (adapted from Lange, 1997).

If the allele RFs at a locus for the ith stock, $\mathbf{q}_i=(q_{i1}, \ldots, q_{iT})'$, are distributed as the Dirichlet density, $D(\boldsymbol{q}_i \mid \beta_1,\beta_2, \ldots, \beta_T)$, and the allele counts, $\mathbf{y}_i=(y_{i1},\ldots,y_{iT})'$, in a random sample of n_i alleles have the multinomial distribution, $Mult(n_i,\boldsymbol{q}_i)$, then the prior predictive distribution (Gelman et al., 1995) for the allele counts, obtained by integrating the product of the probability distributions, $Mult(n_i,\boldsymbol{q}_i)$ and $D(\boldsymbol{q}_i \mid \beta_1,\beta_2, \ldots, \beta_T)$, over the simplex, $S(\boldsymbol{q}_i)$, is

$$\int_{S(\mathbf{q}_i)} \frac{n_i!}{y_{i1}!\ldots y_{iT}!}\prod_{t=1}^{T} q_{it}^{y_{it}} \frac{\Gamma(\beta.)}{\prod_{t=1}^{T}\Gamma(\beta_t)} q_{it}^{\beta_t-1} = \frac{n_i!}{y_{i1}!\ldots y_{iT}!}\frac{\Gamma(\beta.)}{\Gamma(n_i+\beta.)}\prod_{t=1}^{T}\frac{\Gamma(y_{it}+\beta_t)}{\Gamma(\beta_t)}, \qquad (1)$$

where $S(\mathbf{q}_i)=\left\{\mathbf{q}_i : 0 < q_{it} < 1, \sum_{t=1}^{T} q_{it} = 1\right\}$ (Lange, 1997).

Estimation of the β's begins with the logarithm of Equation 1, which can be viewed as the ith component of the loglikelihood or support function for the unknown β's,

$$LogL_i(\beta_1,\ldots,\beta_T) = \log\left(\frac{n_i!}{y_{i1}!\ldots y_{iT}!}\right) + \log\Gamma(\beta.) - \log\Gamma(n_i+\beta.) + \sum_{t=1}^{T}\left(\log\Gamma(y_{it}+\beta_t) - \log\Gamma(\beta_t)\right). \qquad (2)$$

The total support function for all baseline samples is the sum of the individual support functions at Equation 2,

$$LogL(\beta_1,\ldots,\beta_T) = \sum_{i=1}^{c} LogL_i(\beta_1,\ldots,\beta_T).$$

The score function for β_t, s_t, is the first derivative of the total support function,

$$s_t = \frac{\partial LogL}{\partial \beta_t} = \sum_{i=1}^{c}\frac{\partial LogL_i}{\partial \beta_t} = c\psi(\beta.) - \sum_{i=1}^{c}\psi(n_i+\beta.) + \sum_{i=1}^{c}\left(\psi(y_{it}+\beta_t) - \psi(\beta_t)\right), \qquad (3)$$

where $\psi(w) = \frac{d}{dw} \log \Gamma(w)$ is the digamma function.

The elements of the information matrix are

$$-\frac{\partial^2 LogL}{\partial \beta_t \partial \beta_u} = -c\psi'(\beta.) + \sum_{i=1}^{c} \psi'(n_i + \beta.) - \delta_{tu} \sum_{i=1}^{c} \left(\psi'(y_{it} + \beta_t) - \psi'(\beta_t)\right), \tag{4}$$

where the Kronecker delta is defined as δ_{tu}=0 if $t \neq u$, δ_{tu}=1 if t=u, and

$$\psi'(w) = \frac{\delta^2}{\delta^2 w} \log \Gamma(w)$$

is the trigamma function. The observed information matrix, or negative Hessian of the total support function, can be written as

$$\left(-\frac{\partial^2 LogL}{\partial \beta_t \partial \beta_u}\right)_{T \times T} = \mathbf{D} - b\mathbf{11'}, \tag{5}$$

where $\mathbf{D}$ is a diagonal matrix with main diagonal elements

$$d_{tt} = \sum_{i=1}^{c} \left(\psi'(\beta_t) - \psi'(y_{it} + \beta_t)\right), \qquad t = 1, 2, \ldots, T, \tag{6}$$

$\mathbf{1}$ is the unit column vector of T "1"s, and

$$b = c\psi'(\beta.) - \sum_{i=1}^{c} \psi'(n_i + \beta.) \text{ is a scalar.} \tag{7}$$

A quasi-Newton search for the maximum likelihood estimate of $\boldsymbol{\beta}$=(β_1, β_2,…β_T)′ is performed. At the kth step

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + [\mathbf{D}^{(k)} - \tilde{b}^{(k)}\mathbf{11'}]^{-1}\mathbf{S}^{(k)}, \tag{8}$$

where $\boldsymbol{\beta}^{(k)}$ = the approximation of the maximum likelihood estimate at the kth step,

$\mathbf{D}^{(k)}$ = denotes the matrix $\mathbf{D}$ at Equation 6 when evaluated at $\boldsymbol{\beta}^{(k)}$,

$\tilde{b}^{(k)}$ = denotes the minimum of either the scalar, b, at Equation 7 when evaluated at $\boldsymbol{\beta}^{(k)}$ or the ratio, (1–ε)/[$\mathbf{1}'(\mathbf{D}^{(k)})^{-1}\mathbf{1}$] with ε being an arbitrary constant in (0,1), and

$\mathbf{S}^{(k)}$ = the vector of scores at Equation 3 evaluated at $\boldsymbol{\beta}^{(k)}$.

An arbitrary choice for $\boldsymbol{\beta}^{(1)}$ such as the unit column, $\mathbf{1}$, can be used to start the search.

Appendix 2—Minimum squared-error risk estimate of *β.* with the prior mean fixed (an extension of sec. 12.2.3 of Bishop et al., 1975).

Let the baseline risk criterion be the expected value of the squared distance of any matrix estimator of baseline RFs from the true values,

$$R(\hat{\mathbf{Q}}, \mathbf{Q}) = \sum_{i=1}^{c} \sum_{t=1}^{T} n_i E(\hat{q}_{it} - q_{it})^2.$$

Denote the random version of the posterior mean of $\mathbf{q}_i \mid \mathbf{y}_i$ by

$$\hat{\mathbf{q}}_\mathbf{i} = \hat{\mathbf{q}}_\mathbf{i}(\beta., \lambda) = \frac{n_i}{n_i + \beta.}\left(\frac{\mathbf{y}_\mathbf{i}}{n_i}\right) + \frac{\beta.}{n_i + \beta.}\lambda,$$

where $\mathbf{y}_\mathbf{i}$ = (y_{i1}, y_{i2}, … , y_{iT})′ = the array of sample allele counts from the ith stock,

λ = the baseline center, $\bar{\mathbf{y}}$=($\bar{y}_1$, $\bar{y}_2$, . . . $\bar{y}_\mathrm{T}$)′,

$\hat{\mathbf{q}}_\mathbf{i}$ = ($\hat{q}_{i1}$, $\hat{q}_{i2}$, . . . , $\hat{q}_{iT}$)′ and

$\bar{y}_t = \frac{1}{c}\sum_{i=1}^{c} y_{it} / n_i, t = 1, 2, \ldots, T,$ = the arithmetic average of the observed RFs of the tth allele among stocks.

With $\beta.$ and λ viewed as fixed, the baseline risk is

$$R(\hat{\mathbf{Q}}, \mathbf{Q}) = \sum_{i=1}^{c} \left(\frac{n_i}{n_i + \beta.}\right)^2 \left(1 - \sum_{t=1}^{T} q_{it}^2\right) + \sum_{i=1}^{c} \left(\frac{\beta.}{n_i + \beta.}\right)^2 n_i \sum_{t=1}^{T} (q_{it} - \lambda_t)^2.$$

The value of $\beta.$ that minimizes the risk is found by setting the derivative of the risk function (with respect to $\beta.$) equal to zero and solving. The minimizing value of $\beta.$ must satisfy the following equation:

$$\beta. = \sum_{i=1}^{c} n_i^2 \frac{\left(1 - \sum_{t=1}^{T} q_{it}^2\right)}{(n_i + \beta.)^3} \bigg/ \sum_{i=1}^{c} n_i^2 \frac{\sum_{t=1}^{T} (q_{it} - \lambda_t)^2}{(n_i + \beta.)^3}.$$

The equation includes the unknown RFs, $\boldsymbol{q_i}$, and the observed RFs are substituted for their unknown values for estimation. The equation can be iterated to solve for the optimal $\beta.$. Beginning with an arbitrary value, $\beta.$=1, on the right-hand side, the first revised value for optimal $\beta.$ results on the left-hand side. This new approximation for optimal $\beta.$ is used on the right-hand side to compute the next revision, and so on to convergence. If the resulting solution for optimal $\beta.$ is less than 1, setting $\beta.$ equal to 1 seems to be a practical remedy when numerical problems occur in sampling of HAG RFs from their Dirichlet posterior during MCMC computations.