A Bayesian method for identification of stock mixtures from molecular marker data

Jukka Corander (contact author)
Pekka Marttinen
Samu Mäntyniemi
Department of Mathematics and Statistics
P.O. Box 68
Fin-00014
University of Helsinki
Helsinki, Finland
Email address for J. Corander: jukka.corander@helsinki.fi

Abstract—Molecular markers have been demonstrated to be useful for the estimation of stock mixture proportions where the origin of individuals is determined from baseline samples. Bayesian statistical methods are widely recognized as providing a preferable strategy for such analyses. In general, Bayesian estimation is based on standard latent class models using data augmentation through Markov chain Monte Carlo techniques. In this study, we introduce a novel approach based on recent developments in the estimation of genetic population structure. Our strategy combines analytical integration with stochastic optimization to identify stock mixtures. An important enhancement over previous methods is the possibility of appropriately handling data where only partial baseline sample information is available. We address the potential use of nonmolecular, auxiliary biological information in our Bayesian model.

Manuscript submitted 13 February 2005 to the Scientific Editor's Office.
Manuscript approved for publication 14 December 2005 by the Scientific Editor.
Fish. Bull. 104:550–558 (2006).

Stock mixture analysis using multilocus genotypes of fish is recognized as a versatile tool in fisheries management. The efficiency of combining polymorphic molecular markers, such as microsatellites, with a model-based approach to estimate stock mixtures, has been clearly demonstrated in the literature (Kalinowski, 2004; Reynolds and Templin, 2004). Since the beginning of the 21st century, Bayesian methods have largely replaced the earlier applied maximum likelihood approach based on latent class mixture models (Pella and Masuda, 2001). A similar trend has been true for the estimation of genetic population structure in general (e.g., Pritchard et al., 2000; Corander et al., 2003, 2004; Beaumont and Rannala, 2004). For an earlier approach to mixture analysis with incomplete information about source populations, see Smouse et al. (1990).

Bayesian methods for estimation of stock mixtures has generally been based on exploitation of data augmentation through Markov chain Monte Carlo (MCMC), where latent origins of caught individuals and values of the other model parameters are successively simulated from the corresponding posterior distributions. Such an approach is capable of avoiding certain estimation problems caused by missing data and rare alleles, which severely affect the maximum likelihood method. However, because of numerical deficiencies, there are situations where the MCMC based method for the latent class mixture model may easily fail to provide appropriate estimates. First, in the presence of very small groups of individuals representing some stock sources, the posterior distribution of the origins for these particular individuals and the corresponding posteriors of the source allele frequencies will typically comprise a high level of uncertainty. Consequently, the resulting MCMC simulation error in the estimates may be considerable. Second, when there are baseline samples available only for a subset of potential stock sources, estimation of origins is not feasible (Pella and Masuda, 2001). Use of the standard approach with a fixed number of sources, based on the available baseline samples, may easily lead to spurious estimates when there are individuals representing several additional sources in the data. Similarly, it is difficult to detect outlier individuals with the latent class approach with a fixed number of sources (Pritchard et al., 2000) because they are unlikely to be identified in the MCMC simulation for data sets of moderate to large size. Third, under partial baseline information, it is difficult to appropriately infer a suitable number of stock sources to represent a particular data set.

Partition-based Bayesian alternatives to latent class models for identification of genetic mixtures without baseline samples have recently been introduced (Dawson and Belkhir 2001; Corander et al., 2003, 2004). Corander et al. (2003, 2004) used an analytical integration strategy combined with stochastic search methods to make Bayesian estimation

more feasible when the number of genetically diverged sources contributing to the observed data is unknown. A wide variety of applications of this approach can be found in the literature (e.g., Heuertz et al., 2004, Seppä et al., 2004, Mäki-Petäys et al., 2005). The approach by Dawson and Belkhir (2001) is similar to that of Corander et al. (2004) in spirit; however, it is subject to two important limitations that prevent an efficient use of this approach in the current context. First, there are no readily available informative forms of the family of prior distributions used by Dawson and Belkhir (2001), which would be necessary for representing the baseline information. Second, their model formulation does not allow for missing alleles in the molecular marker data, which are present in most real data sets.

In our study, we extend the partition-based approach to incorporate *a priori* baseline information, making it suitable for identification of stock mixtures, either under complete or partial baseline sample information. Our focus is on the identification of the putative genetic mixture in the catch sample data, provided by the maximum *a posteriori* estimate of the assignment of the individuals into an unknown number of sources. Given the estimate, the proportions of the stocks in the population can be readily inferred by using the standard multinomial-Dirichlet model (e.g., Pella and Masuda, 2001) and generic Bayesian software, such as BUGS (Spiegelhalter et al., 2003), which has been widely used for fish population modeling (e.g., Meyer and Millar, 1999; Mäntyniemi and Romakkaniemi, 2002; Mäntyniemi et al., 2005).

Another novelty in our method is the possibility of using available biologically relevant information to preassign catch data into groups that can be considered as sampling units in the model. For instance, when the behavior of the investigated species is such that individuals obtained simultaneously at a single catch location are known to represent the same (yet unknown) stock, they can be allocated as a single unit to an origin. Such use of auxiliary information enhances the statistical power to detect the correct origin when the number of molecular marker loci available is limited. To illustrate our modeling approach, and to investigate its performance under various biological settings, we present results from several simulation experiments based partly on real molecular data for the Baltic Sea stock mixture of Atlantic salmon (*Salmo salar*).

Methods

Bayesian stock mixture model

In stock mixture estimation, there are typically available in samples two types of individuals, which are genotyped. One type consists of individuals with known origin (baseline data), and the other type represents a catch sample, which may have been pooled from several sources. Let m be the number of potential stocks, such that for each stock $i = 1,\ldots, m$, there are n_i baseline individuals available. Furthermore, there may be an additional number of potential stocks contributing to the catch population; however, these are not represented by any baseline samples. We let K ($m \leq K$) denote the total number of potential stocks, which can have contributed to a catch sample of n individuals, whose origins are unknown. Notice that K is typically determined from the relevant biological information about the species under consideration. The target for our estimation is to infer the number of stocks, say k, having actually contributed to the catch sample, from the multilocus genotypes of both the baseline and catch individuals.

Under the assumption that the genetic information consists of N_L molecular marker loci, where at each locus $j = 1,\ldots, N_L$, there are $N_{A(j)}$ different alleles distinguished among all baseline and catch samples. Pella and Masuda (2001) introduced a rather complicated empirical Bayes procedure to determine the prior distribution for the allele frequencies in the potential stocks through the observed genotypes of the baseline individuals (all stocks were assumed to be represented by baseline samples). Here we consider a simpler approach, by suitably modifying the standard Dirichlet prior used in Corander et al. (2003, 2004). We assume that the allele frequencies between marker loci are conditionally independent given the stock origins and consider the potential stocks to be in Hardy-Weinberg equilibrium (HWE).

Let p_{ijl} be the unknown frequency (or probability) of allele l in the stock i at locus j, given that k ($k \leq K$) stocks are considered. Further, for each locus $j = 1,\ldots, N_L$, let α_{ijl} be a hyperparameter for a Dirichlet prior distribution of the allele frequencies of stock i ($i = 1,\ldots, k$;$l = 1,\ldots, N_{A(j)}$). Given the baseline information, we may partially update our beliefs about the allele frequencies using the posterior distribution derived from an initially vague reference prior. For each of the m stocks, where baseline samples are available, we set $\alpha_{ijl} = n_{ijl} + 1/N_{A(j)}$, where n_{ijl} is the observed number of copies of allele l at locus j among individuals in the baseline sample of size n_i. This hyperparameter updating procedure is standard in Bayesian analysis with the multinomial-Dirichlet model (Gelman et al., 2004). Correspondingly, for the other potential stocks, not represented by any baseline information, the count n_{ijl} is zero, and the hyperparameter is determined as $\alpha_{ijl} = 1/N_{A(j)}$.

A putative assignment of the catch data to the potential stocks is represented in our study by a partition-valued parameter $S = (s_1,\ldots, s_k)$, which allocates the n individuals into k non-empty clusters. A cluster is labeled as either the corresponding baseline sample or, alternatively, as a group of unknown geographical origin. The prior distribution of the allocation parameter $P(S)$ is defined according to

$$P\left(S = (s_1,\ldots, s_k)\right) = \begin{cases} c, \text{ if } k \leq K \\ 0, \text{ otherwise} \end{cases}, \quad (1)$$

which corresponds to a uniform distribution over the possible allocations of the catch individuals under the

restriction $k \leq K$. This particular prior specification allows a convenient estimation algorithm to be constructed for identifying those allocations associated with high posterior probabilities, given the molecular marker data. Also, our prior is considerably more informative from the biological perspective than the uniform prior with $K = n$ used in Corander et al. (2004) because it assigns a zero probability to most allocations that are very unrealistic for a moderate or large sample size n.

The posterior probability of an allocation S over the class of all putative candidates is defined as

$$p(S \mid data) = p(data \mid S)P(S) / \sum_{S \in \mathfrak{S}} p(data \mid S)P(S), \quad (2)$$

where $p(data \mid S)$ is the marginal likelihood of having the allele frequency parameters of each class s_i in S integrated out analytically (formula A1 in Corander et al., 2003) according to

$$p(data \mid S) = \prod_{i=1}^{k} \prod_{j=1}^{N_L} \left[\frac{\Gamma(\sum_l \alpha_{ijl})}{\Gamma(\sum_l \alpha_{ijl} + n_{ijl})} \prod_{l=1}^{N_{A(j)}} \frac{\Gamma(\alpha_{ijl} + n_{il})}{\Gamma(\alpha_{ijl})} \right], \quad (3)$$

where n_{ijl} is now the observed number of copies of allele l at locus j among catch individuals allocated to cluster i. The analytical integration approach was used earlier in a related genetic context by Balding and Nichols (1997), Rannala and Hartigan (1996), Rannala and Mountain (1997).

As can be seen from the above expression, the parameter that needs to be estimated in our stock mixture model is the allocation partition S. In situations where auxiliary biological information provides a pre-assignment of some catch individuals into *a priori* sampling units, each known to represent a single unknown origin, the above formula still applies. However, the prior distribution of the allocation parameter needs to be modified accordingly, to exclude the values of S that assign individuals in the same sampling unit to distinct origins.

Algorithm for allocation estimation

In the sequel, we consider the general situation where the catch data are represented by a number of *a priori* sampling units, each containing potentially more than a single individual (see the previous section). When no auxiliary information for pre-assignment to sampling units is available, each sampling unit will correspond to a single individual in the catch data.

The two main aspects of our stock mixture model that need to be estimated from marker data are the number of clusters (k) suitable for a particular data set, and the allocation of sampling units to the clusters. Corander et al. (2003, 2004) used various stochastic search strategies for estimation of a comparable model without prior baseline samples. In our study, we develop an alternative method to enable analyses of data sets ranging from moderate to challenging because the MCMC algorithms introduced by Corander et al. (2003, 2004) are computationally time and memory intensive. The central idea is to use a "greedy" stochastic search algorithm (Fletcher, 1987) to find a partition S with the highest posterior probability (Eq. 2). Repeated runs of the algorithm enable investigation of the stability of the estimation procedure, in a way that is similar to the parallel MCMC approach of Corander et al. (2004).

An initial partition for the algorithm is determined by assigning sampling units one by one to such a cluster, so that the resulting partition has the highest possible posterior probability. After specifying the initial partition, the greedy algorithm proceeds by performing the following operations repeatedly on the current partition:

1 The algorithm moves sampling units from one cluster to another. In a stochastic order, for each sampling unit, it calculates the change imposed to the posterior probability $P(S \mid data)$ by moving the particular sampling unit to any of the other clusters, including even an empty cluster, unless that would lead to $k > K$. It moves the sampling unit to the cluster that increases the value of $P(S \mid data)$ most (if no increase is possible, the sampling unit is not moved).
2 It joins clusters. For each pair of clusters, the algorithm calculates the change to $P(S \mid data)$ imposed by joining them. It joins the two clusters that cause the maximal increase (if no increase is possible, no clusters are joined).
3 It splits clusters. Using the Kullback-Leibler divergence between sampling units (Corander et al., 2003), the algorithm splits each cluster into maximally 20 subclusters and calculates the change to $P(S \mid data)$ imposed by keeping one of the introduced subclusters as a separate new cluster, or by joining it to any of the previously existing clusters. It keeps the new configuration that improves the value of $P(S \mid data)$ most (if no improvement is possible the split is not made).
4 It splits clusters into exactly two maximally homogeneous subclusters with the Kullback-Leibler divergence, otherwise analogously as in step 3.
5 It re-allocates several sampling units from a cluster. In a stochastic order, for each cluster, the algorithm orders the sampling units of the cluster such that the first sampling unit is the one whose removal from the cluster would improve the marginal likelihood of the cluster most, and so on. A putative candidate for a new partition is formed by moving sampling units one by one to some other cluster, such that the $P(S \mid data)$ of the resulting partition is as high as possible (these moves are performed even if a single move results in a worse solution). If at some point the total change in $P(S \mid data)$ is positive, the putative candidate is accepted and operation 5 is completed. When the total change remains negative, even after re-allocation of all the sampling units in a single cluster, the putative candidate is rejected and another cluster is chosen as a target until all clusters have been considered.

The optimization operations are repeated in a varying order, until none of them improves the posterior probability $P(S|data)$ of the current partition. The allocation of sampling units to different clusters is based on the obtained partition, and the suitable number of clusters k is estimated from the partitions visited during the simulation.

Measurement of the strength of evidence for any particular value of the partition S, given the marker data, is an intricate process, in particular for large data sets from complex stock mixtures. Theoretically, the unknown largest posterior probability may be extremely small, even in situations where a particular model provides an adequate fit to the observations. An important factor explaining such a feature is the large number of possible allocations, which all have positive posterior probabilities by definition. This feature is of general concern in a Bayesian analysis that comprises vast model spaces, see, e.g., the discussion in Madigan and Raftery (1994). Because the actual estimated value of the posterior probability may be an intuitively misleading goodness-of-fit measure, we use an alternative strategy for characterization of the uncertainty in relation to the estimated allocation.

Bayes factors (e.g., Kass and Raftery, 1995) provide a computationally efficient approach to local assessment of the amount of the peak of the posterior distribution around an estimate of S. Let S^* denote an alternative allocation obtained from an estimate S by moving any particular sampling unit to another putative stock. The strength of evidence in favor of placing that sampling unit in the original stock against placement in the new stock is provided by the Bayes factor

$$B_{S,S^*} = \frac{p(data \mid S)P(S)}{p(data \mid S^*)P(S^*)}, \quad (4)$$

which measures how many times more plausible the allocation S is for the particular sampling unit. When the value of Equation 4 is small, say $B_{S,S^*} < 10$ (or $\log_e B_{S,S^*} < 2.3$, Kass and Raftery, 1995), the data do not strongly support a single origin for the particular sampling unit. Because calculation of these Bayes factors is computationally inexpensive, they can be easily provided for every possible sampling unit or stock combination.

In addition to Bayes factors, conditional posterior probabilities for the allocation of each individual over the range of different putative stocks identified through S can be used to characterize the uncertainties in the Bayesian estimate. The conditional posterior probability distribution is defined for each individual by

$$P(S_i \mid data) = \frac{p(data \mid S_i)P(\ _i)}{\sum_{i=1}^{k} p(data \mid S_i)P(S_i)}, \quad (5)$$

where S_i denotes that the particular individual is allocated to the ith class of S (over the k possible alternatives). When only a single stock has a high conditional posterior probability, the allocation is made on a firm basis. However, when at least two sources are identified with reasonably high posterior probabilities, the genetic evidence is not conclusive enough for a classification of the particular individual to a single source. The advantage of the conditional posterior probabilities over Bayes factors in characterization of the classification uncertainty for each individual is that the former compares simultaneously all putative sources, whereas the latter provides only a pairwise judgement.

The correct number of clusters needed to describe the data can be estimated from the partitions that were visited during the simulation. During the simulation the algorithm stores the marginal likelihoods and the sizes of the 30 best visited partitions, and the posterior probabilities for the different numbers of clusters can then be estimated analogously to those estimated by Corander et al. (2004). Usually, if there is a lot of molecular data available (e.g., hundreds of loci have been observed) only a few of the best partitions have influence on the computed posterior probabilities because the relation of marginal likelihoods between different partitions can be up to ~exp(1000). If the data are sparse (e.g., only about 10–20 loci have been observed) and only partial baseline information is available, the uncertainty related to the correct number of clusters can be considerable because many partitions with differing sizes and approximately equal marginal likelihood may be found. In these cases, to obtain a more reliable estimate of the correct number of clusters, the algorithm should be run multiple times with different upper bounds (K) in order to facilitate the identification of those partitions that have real influence in the posterior probabilities. In our implementation of the estimation algorithm, we have included the possibility to automatically process information from multiple runs.

Empirical illustration of the partition-based approach

The Bayesian estimation algorithm described in the previous section is implemented in BAPS software.[1] The examples considered here are produced by BAPS analyses of data simulated by using the real data from Koljonen et al. (2002), who assessed allele frequencies for nine microsatellite markers in Atlantic salmon within the Baltic Sea region. We have experimented with several simulation configurations to investigate how our method would be expected to perform under a variety of biological conditions.

The five wild stocks of Atlantic salmon considered in Koljonen et al. (2002) correspond to five different rivers draining into the Baltic Sea: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva. The pairwise genetic distances (Nei's D_A, Nei et al., 1983) between these stocks underlying our simulations are

[1] BAPS software is freely available at URL http://www.rni.helsinki.fi/~jic/bapspage.html. Results presented here were calculated with version 3.1 (release date 5 March 2005).

given in Table 1 (reproduced from Koljonen et al., 2002). An estimate of total F_{ST} (Weir and Cockerham, 1984) equal to 0.07 was obtained in Koljonen et al. (2002) for these stocks on the basis of the nine microsatellite loci. The magnitude of the genetic differentiation in the underlying population is fairly small, and the pairwise distances vary considerably. Thus, we may conclude that these stocks represent a biologically challenging setting for inference about the genetic mixture in a population sample. Using the individual stock allele frequencies, we have simulated baseline individuals and catch samples under the assumptions of HWE and no linkage between the loci. A wide variety of configurations, with complete and partial baseline information and different sample sizes, were tested. In the analyses involving five underlying stocks, we used $K = 10$ as the prior upper bound, and the estimation algorithm was run 12 times for each replicate data set. For cases with only two underlying stocks, the upper bound was set as $K = 6$.

Results of our simulation experiments are summarized in Tables 2–8. As a summary, we highlight the following aspects. Uneven proportions of stock presence in the samples do not seem to affect the inference notably, even when the baseline information is only partial. The results in Table 3 are produced under a particularly challenging situation, where the baseline information comprises 40 individuals from a single stock only. The sample configuration then contains 40 individuals from this stock and 10 individuals from another, *a priori* unknown stock. The results show that our method performs surprisingly well in the identification of the outgroup, given that the genetic difference between the two underlying stocks is not negligible. However, as the results in Table 4 illustrate, the presence of putative stocks not represented by baseline information may also be masked by the baseline available for a genetically similar stock. Identification of putative stocks without using any baseline information may *de facto* be more successful under such circumstances (compare Tables 4a–4d). Therefore, we suggest that in practice both types of analyses are performed and the results compared, since this is computationally inexpensive with our method. Our results indicate that incomplete baseline information is expected to be most fruitful for the identification task when there are baseline samples available from the stocks that are genetically most similar. The baseline configurations in Table 4 can be

Table 1

Pairwise genetic distances (Nei's D_A, Nei et al., 1983) between different Atlantic salmon stocks within the Baltic Sea region (reproduced from Table 2 in Koljonen et al., 2002). Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva.

Stocks	TornW	Simo	Ii	Oulu
Simo	0.129			
Ii	0.068	0.125		
Oulu	0.110	0.164	0.131	
Neva	0.261	0.285	0.284	0.261

Table 2

Allocation average percentages over 20 simulations, when 25 individuals from each stock were present in the sample data, and the number of baseline individuals available from each stock was (**A**) 30, (**B**) 15, and (**C**) 5. The column with the heading "Other" refers to additional stocks inferred by the method. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva.

		Allocation					
Origin		TornW	Simo	Ii	Oulu	Neva	Other
A	TornW	0.80	0.01	0.15	0.03	0.00	0.00
	Simo	0.03	0.95	0.02	0.00	0.00	0.00
	Ii	0.11	0.02	0.84	0.03	0.00	0.00
	Oulu	0.03	0.01	0.01	0.95	0.00	0.00
	Neva	0.00	0.00	0.00	0.00	1.00	0.00
B	TornW	0.72	0.03	0.18	0.06	0.00	0.00
	Simo	0.02	0.94	0.02	0.02	0.00	0.00
	Ii	0.11	0.03	0.84	0.02	0.00	0.00
	Oulu	0.04	0.01	0.03	0.91	0.00	0.00
	Neva	0.00	0.00	0.00	0.00	0.99	0.00
C	TornW	0.59	0.06	0.23	0.09	0.01	0.02
	Simo	0.04	0.90	0.04	0.01	0.00	0.00
	Ii	0.20	0.06	0.70	0.03	0.00	0.01
	Oulu	0.03	0.01	0.05	0.91	0.00	0.00
	Neva	0.00	0.00	0.00	0.00	0.98	0.02

Table 3

Origin identification performance when the sample consists of 40 individuals from the Iijoki River (Ii) and 10 individuals from another stock. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva.Baseline information was in each case available only from Ii (40 simulated individuals). The numbers are based on 50 replicates of each configuration.

Stock without baseline (outgroup)	Percentage of correct recognition for individuals from Ii			Percentage of correct recognition for individuals from the outgroup			Percentage of replicates where at least 50% of the outgroup was recognized
	min.	max.	avg.	min.	max.	avg.	
Neva	0.95	1.00	0.9955	0.9	1.00	0.998	1.00
Oulu	0.875	1.00	0.9665	0.0	1.00	0.7440	0.86
Simo	0.85	1.00	0.9705	0.0	1.00	0.6820	0.90
TornW	0.90	1.00	0.9820	0.0	0.80	0.3660	0.34

categorized in this respect as neutral (no biasing effect; Table 4, A and B), positive (strengthens the inference; Table 4C), negative (biases the inference; Table 4, D and E).

Our results indicate that commonly occurring levels (<5%) of missing marker data do not inhibit the ability of our method to detect the correct stocks, assuming that the missing values are randomly distributed over loci and individuals (Table 5). As an overall conclusion from the simulations, it is clear that the genetic dissimilarities of the stocks matter most for identification performance. When baseline samples are available for all stocks, most individuals can be correctly assigned to their origin even when the genetic distance between the stocks is negligible (such as between Tornionjoki and Iijoki rivers). Usefulness of the conditional posterior probabilities for characterization of the allocation uncertainty is exemplified in Table 6.

The number of inferred putative stocks was in general well in accordance with the underlying true number and there was no tendency to overestimate k. However, when the number of available marker loci was decreased to five (Table 7), the probability of obtaining additional putative stocks was slightly increased. Because it is widely known that the level of polymorphism of the markers affects their usefulness in origin identification, it is difficult to specify very clear boundaries with respect to the amount of loci necessary for an acceptable performance of any assignment method. It is important to notice that an acceptable characterization of uncertainty inherently depends on the real biological context in a particular modeling situation. As a simple rule of thumb for our method, we would suggest that $N_L \leq 6$ might be regarded as an insufficient value for reliable estimation. However, when auxiliary information is available such that the sample data can be grouped before analysis (as in Table 8), the statistical power to detect correct origins and k increases considerably. This situation would correspond to a geographical sampling scheme where the individuals assigned to the same sampling unit are caught simultaneously at a specific location.

Discussion

We have introduced a novel Bayesian method for an investigation of stock mixtures using molecular marker data by suitably modifying existing partition-based Bayesian models for estimation of genetic population structure. To enable smooth applicability, the implementation is made freely available in a user-friendly software. One particular advantage of our method is the possibility of appropriately analyzing data in a situation where only partial baseline information is available for the potential stocks. Use of an analytical integration approach enhances considerably the numerical performance when the stock mixture structure is challenging (e.g., in the presence of small stocks for which no baseline samples have been collected).

Contrary to the earlier Bayesian methods introduced in Corander et al. (2003, 2004), we have exploited a considerably less computationally intensive strategy that is based on stochastic optimization instead of MCMC simulation. To obtain stable estimates for moderate to large data sets, many long parallel MCMC chains would be needed, but the process for obtaining these chains often is not feasible under a single CPU architecture. Our intelligent search strategy, instead of the random search used in MCMC, seems to resolve this problem very efficiently. A disadvantage of stochastic optimization compared to optimization with MCMC is that a statistically consistent estimate of the number of stocks contributing to the sample cannot be derived. Nevertheless, our novel method has performed satisfactorily in this respect under realistic sampling scenarios. We are currently exploring possibilities for using intelligent proposals in MCMC and an online-based parallel implementation of the method, both of which would provide an ideal framework for biologists using molecular data in stock mixture estimation.

The most relevant biological assumptions used in our approach are HWE and nonlinkage of the marker loci. The latter assumption is generally valid, at least approximately, for the microsatellite markers often used in

Table 4

Allocation of 30 sample individuals from each of five stocks (150 individuals in total) under different baseline settings: (**A**) no baseline, (**B**) 40 baseline individuals from TornW, (**C**) 40 baseline individuals from TornW and Ii, (**D**) 15 baseline individuals from TornW and Ii, (**E**) 15 baseline individuals from Simo. Each "C" refers to an inferred putative stock for which no baseline information was available. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva.

		Allocation				
Origin		C1	C2	C3	C4	C5
A	TornW	3	1	0	3	23
	Simo	1	1	0	25	3
	Ii	0	1	0	4	25
	Oulu	0	28	0	0	2
	Neva	0	1	29	0	0
		TornW	C2	C3	C4	
B	TornW	28	1	1	0	
	Simo	2	27	1	0	
	Ii	26	4	0	0	
	Oulu	2	0	28	0	
	Neva	0	0	0	30	
		TornW	Ii	C3	C4	C5
C	TornW	22	6	1	1	0
	Simo	1	2	26	1	0
	Ii	5	25	0	0	0
	Oulu	3	1	0	26	0
	Neva	0	0	0	0	30
		TornW	Ii	C3	C4	
D	TornW	24	5	0	1	
	Simo	1	2	0	27	
	Ii	5	23	0	2	
	Oulu	29	1	0	0	
	Neva	0	1	29	0	
		Simo	C2	C3	C4	C5
E	TornW	12	2	1	15	0
	Simo	29	1	0	0	0
	Ii	24	0	0	5	1
	Oulu	1	0	28	1	0
	Neva	0	0	0	0	30

Table 5

(**A**) Average numbers of allocations (over 20 replicates) to the different stocks under an uneven sample size distribution: TornW, n=60, Simo, n=20, Ii, n=30, Oulu, n=5, Neva, n=10. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva. The number of baseline individuals available from each of the five stocks was 30. The column with the heading "Other" refers to additional stocks inferred by the method. The results in (**B**) are otherwise based on an analogous configuration, except that 5% of the marker data was randomly set as missing values.

		Allocation					
Origin		TornW	Simo	Ii	Oulu	Neva	Other
A	TornW	47.5	1.2	7.9	3.2	0.0	0.3
	Simo	0.6	18.9	0.3	0.2	0.1	0.0
	Ii	3.1	1.2	25.3	0.5	0.0	0.0
	Oulu	0.3	0.0	0.1	4.7	0.0	0.0
	Neva	0.0	0.1	0.1	0.0	9.9	0.0
B	TornW	47.4	1.7	7.2	3.6	0.2	0.2
	Simo	0.6	18.4	0.8	0.3	0.0	0.1
	Ii	3.5	1.1	25.0	0.5	0.0	0.1
	Oulu	0.3	0.1	0.1	4.6	0.0	0.0
	Neva	0.0	0.1	0.0	0.1	9.9	0.1

Table 6

Average conditional posterior probabilities (over 20 replicates) for allocations of individuals to the different stocks under an uneven sample size distribution: TornW, n=60, Simo, n=20, Ii, n=30, Oulu, n=5, Neva, n=10. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva. The number of baseline individuals available from each stock was 30. The column with the heading Other refers to additional stocks inferred by the method.

	Allocation					
Origin	TornW	Simo	Ii	Oulu	Neva	Other
TornW	0.80	0.03	0.13	0.04	0.00	0.00
Simo	0.02	0.94	0.03	0.01	0.00	0.00
Ii	0.12	0.04	0.82	0.03	0.00	0.00
Oulu	0.06	0.01	0.03	0.90	0.00	0.00
Neva	0.00	0.00	0.00	0.01	0.99	0.00

stock mixture analyses. Minor deviations from HWE are not expected to notably affect our inference method; however, presence of samples from small stocks under strong inbreeding could result in an overestimation of k when there is limited baseline information available. Samples from such stocks would tend to be split into parts by the model if no baseline information about the stock allele frequencies can be used to identify the joint origin.

In addition to the molecular markers, auxiliary information, such as simultaneous catch at a common geo-

Table 7

Average numbers of allocations (over 20 replicates) to the different stocks under an uneven sample size distribution: TornW, n=60, Simo, n=20, Ii, n=30, Oulu, n=5, Neva, n=10. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva. The number of baseline individuals available from each of the five stocks was 30. The column with the heading "Other" refers to additional stocks inferred by the method. The marker loci used for inference were randomly sampled from the original nine microsatellites for each replicate; in (**A**) seven loci were used, in (**B**) five loci were used.

		Allocation					
Origin		TornW	Simo	Ii	Oulu	Neva	Other
A	TornW	44.5	3.2	7.6	4.3	0.2	0.0
	Simo	0.9	17.8	0.9	0.4	0.0	0.1
	Ii	2.8	1.6	24.7	0.9	0.1	0.0
	Oulu	0.3	0.1	0.2	4.5	0.0	0.0
	Neva	0.0	0.0	0.0	0.0	10.0	0.0
B	TornW	37.3	4.5	9.3	5.5	0.5	3.1
	Simo	1.0	16.5	1.4	0.6	0.2	0.4
	Ii	3.3	2.8	21.2	2.2	0.1	0.5
	Oulu	0.5	0.2	0.4	4.0	0.0	0.1
	Neva	0.1	0.0	0.1	0.2	9.6	0.2

graphical location, can be incorporated into the analysis. This information is incorporated by the pre-assignment of individuals in the catch data to *a priori* sampling units, when such are considered to be relevant for the species under investigation. Such prior information is particularly useful if the available molecular data are scarce because it enhances the statistical power to detect correct stock origins, as illustrated in our example analyses.

Although the Bayesian method that we propose seems to be a versatile tool for stock mixture identification, certain modifications of the model would also provide fruitful extensions for a variety of biological settings. Current use of the auxiliary information necessitates that the individuals assigned to the same sampling unit represent with certainty the same origin. However, the existence of such conclusive information cannot be assumed in applications in general. There is still a possibility of incorporating information about a tendency to a geographical clustering among the catch individuals with respect to the stock origin, through a suitable modification of the prior distribution of the partitions. In general, use of biological information concerning the behaviour of a species, in combination with geographical sampling information, provides a rich area for further model development. In particular, this combination of information highlights the potential use of the Bayesian statistical framework because the relevant biological information can often be efficiently incorporated through the prior distributions for the model parameters.

Table 8

Allocation of the 25 sample individuals from each of five stocks (125 individuals in total) under different baseline settings: (**A**) no baseline, (**B**) 15 baseline individuals from each stock, (**C**) 15 baseline individuals from each stock, and auxiliary biological information was introduced by considering the simulated catch data as sampling units of the size of five individuals. Stocks correspond to five different rivers: Tornionjoki (TornW), Simojoki (Simo), Iijoki (Ii), Oulujoki (Oulu), and Neva. Each "C" refers to an inferred putative stock for which no baseline information was available.

		Allocation				
Origin		C1	C2	C3	C4	
A	TornW	20	0	4	1	
	Simo	2	0	2	21	
	Ii	20	1	3	1	
	Oulu	1	0	24	0	
	Neva	0	24	1	0	
		TornW	Simo	Ii	Oulu	Neva
B	TornW	18	0	7	0	0
	Simo	2	23	0	0	0
	Ii	0	1	23	1	0
	Oulu	2	0	0	23	0
	Neva	0	0	0	1	24
		TornW	Simo	Ii	Oulu	Neva
C	TornW	25	0	0	0	0
	Simo	0	25	0	0	0
	Ii	0	0	25	0	0
	Oulu	0	0	0	25	0
	Neva	0	0	0	0	25

Acknowledgments

The authors thank Marja-Liisa Koljonen for providing data about the microsatellite allele frequencies in Baltic salmon stocks. This work was supported by the Centre of Population Genetic Analyses, University of Oulu, Finland (Academy of Finland, grant no. 53297), and by Research funds of University of Helsinki, Finland.

Literature cited

Balding, D. J., and R. A. Nichols.
1997. Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 78:583–589.

Beaumont, M. A., and B. Rannala.
2004. The Bayesian revolution in genetics. Nat. Rev. Genet. 5:251–261.

Corander, J., P. Waldmann, P. Marttinen, and M. J. Sillanpää.
2003. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20:2363–2369.

Corander, J., P. Waldmann, and M. J. Sillanpää.
2004. Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374.

Dawson, K. J., and K. Belkhir.
2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res. 78:59–77.

Fletcher, R.
1987. Practical methods of optimization, 450 p. Wiley, New York, NY.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin.
2004. Bayesian data analysis, 2nd ed., 668 p. Chapman & Hall/CRC, Boca Raton, LA

Heuertz, M., J. F. Hausman, O. J. Hardy , G. G. Vendramin, N. Frascaria-Lacoste, and X. Vekemans.
2004. Nuclear microsatellites reveal contrasting patterns of genetic structure between western and southeastern European populations of the common ash (*Fraxinus excelsior* L.). Evolution 58:976–988.

Kalinowski, S. T.
2004. Genetic polymorphism and mixed-stock fisheries analysis. Can. J. Fish. Aquat. Sci. 61:1075–1082.

Kass, R., and A. E. Raftery.
1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.

Koljonen, M.-L., J. Tähtinen, M. Säisä, and J. Koskiniemi.
2002. Maintenance of genetic diversity of Atlantic salmon by captive breeding programmes and the geographic distribution of microsatellite variation. Aquaculture 212:69–92.

Madigan, D., and A. E. Raftery.
1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. 89:1535–1546.

Mäki-Petäys, H., A. Zakharov, L. Viljakainen, J. Corander, and P. Pamilo.
2005. Genetic changes associated to declining populations of Formica ants in fragmented forest landscape. Mol. Ecol. 14:733–742.

Mäntyniemi, S., and A. Romakkaniemi.
2002. Bayesian mark-recapture estimation with an application to a salmonid smolt population. Can. J. Fish. Aquat. Sci. 59:1748–1758.

Mäntyniemi, S., A. Romakkaniemi, and E. Arjas.
2005. Bayesian removal estimation of a population size under unequal catchability. Can. J. Fish. Aquat. Sci. 62:291–300.

Meyer, R., and R. Millar.
1999. BUGS in Bayesian stock assessment. Can. J. Fish. Aquat. Sci. 56:1078–1086.

Nei, M., F. Tajima, and Y. Tateno.
1983. Accuracy of estimated phylogenetic trees from molecular data. J. Mol. Evol. 19:153–170.

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

Pritchard, J. K., M. Stephens, and P. Donnelly.
2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Rannala, B., and J. A. Hartigan.
1996. Estimating gene flow in island populations. Genet. Res. 67:147–158.

Rannala, B., and J. L. Mountain.
1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197–9201.

Reynolds, J. H., and W. D. Templin.
2004. Detecting specific populations in mixtures. Environ. Biol. Fish. 69:233–243.

Seppä, P., N. Gyllenstrand, J. Corander, and P. Pamilo.
2004. Coexistence of the social types: Genetic population structure in the ant, *formica exsecta*. Evolution 58:2462–2471.

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

Spiegelhalter, D. L., A. Thomas, N. Best, W. Gilks, and D. Lunn.
2003. BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England. [Available at www.mrc-bsu.cam.ac.uk/bugs/.]

Weir, B. C., and C. C. Cockerham.
1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1350–1370.