Genetic polymorphism and mixed-stock fisheries analysis

Steven T. Kalinowski

Abstract: Genetic data can be used to estimate the stock composition of mixed-stock fisheries. Designing efficient strategies for estimating mixture proportions is important, but several aspects of study design remain poorly understood, particularly the relationship between genetic polymorphism and estimation error. In this study, computer simulation was used to investigate how the following variables affect expected squared error of mixture estimates: the number of loci examined, the number of alleles at those loci, and the size of baseline data sets. This work showed that (*i*) loci with more alleles produced estimates of stock proportions that had a lower expected squared error than less polymorphic loci, (*ii*) highly polymorphic loci did not require larger samples than less polymorphic loci, and (*iii*) the total number of independent alleles examined is a reasonable indicator of the quality of estimates of stock proportions.

Résumé : Les données génétiques peuvent servir à déterminer la composition des stocks dans une pêche commerciale qui englobe plusieurs stocks. La planification de stratégies efficaces pour estimer les proportions du mélange est importante, mais plusieurs aspects du plan d'expérience restent mal compris, particulièrement la relation entre le polymorphisme génétique et l'erreur d'estimation. Une simulation à l'ordinateur a servi dans cette étude à examiner comment les variables ci-dessous affectent l'erreur au carré attendue dans ces estimations des mélanges, soit le nombre de locus examinés, le nombre d'allèles à ces locus et la taille des banques de données de base. La simulation indique que (*i*) les locus avec plus d'allèles donnent lieu à des estimations des proportions des stocks dont l'erreur au carré attendue est plus faible que les locus moins polymorphes, (*ii*) les locus fortement polymorphes ne requièrent pas d'échantillons plus grands que les locus moins polymorphes et (*iii*) le nombre total d'allèles indépendants examinés est un indicateur acceptable de la qualité des estimations des proportions de stocks.

[Traduit par la Rédaction]

Introduction

Management of mixed-stock fisheries requires an accurate description of the composition of the fisheries. Several approaches have been used effectively. Mark and recapture methods provided the first estimates of stock composition (e.g., Gilbert 1924). Variation in morphological traits such as scale patterns (e.g., Hamilton 1947; Mosher 1963) was used extensively for mixture analysis until electrophoretic methods were developed that could identify allozyme variation (e.g., Utter et al. 1974; Grant et al. 1980). Allozymes remain an important tool for mixture analysis but have been eclipsed by DNA markers, including microsatellites, minisatellites, mitochondrial DNA, and MHC genes (Wirgin et al. 1997; Scribner et al. 1998; Shaklee et al. 1999). DNA markers have several advantages over allozymes: sampling is nonlethal, loci are essentially innumerable, and loci are usually more variable. The latest promising class of molecular markers is single-nucleotide polymorphisms (SNPs). SNPs only have two variants per locus but may surpass microsatellites in popularity because genotyping is unambiguous and can be inexpensive (see Moran (2002) for a review).

Received 3 April 2003. Accepted 26 January 2004. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 13 August 2004.
J17441

S.T. Kalinowski. Department of Ecology, 310 Lewis Hall, Montana State University, Bozeman, MT 59717, USA (e-mail: skalinowski@montana.edu).

The wide range of molecular markers available for mixture analysis motivates the question which provide the most cost effective and accurate estimates of mixture proportions? And, within a class of markers, what specific loci are most useful? Empirical comparisons of different marker classes and different sets of loci are accumulating (e.g., Scribner et al. 1998; Allendorf and Seeb 2000; Winans et al. 2004) but are expensive. A more thorough understanding of the general principles of study design for mixture analysis would be useful. One of the most fundamental differences between classes of molecular markers is the amount of polymorphism that they have. For example, SNPs have two alleles per locus, while microsatellite and MHC loci can have over 100. In this study, I used computer simulation to address two questions related to polymorphism and study design: (*i*) how does polymorphism affect the how close estimates of mixture proportions are to parametric proportions and (*ii*) do highly polymorphic loci require larger baseline samples than less polymorphic loci to achieve comparable results?

Estimating mixture proportions

Both maximum likelihood and Bayesian methods are available for estimating mixture proportions within a fishery (e.g., Millar 1987; Pella and Masuda 2001). Although each method is computationally and philosophically different, both methods are built upon estimates of the probability of a fish in the baseline populations having the genotype of a fish observed in the fishery. If mating is random within each population, then the probability of obtaining the genotype A_iA_j

 doi: 10.1139/F04-060

from a population is a simple function of the allele frequencies in the population:

$$\Pr(A_iA_j) = \begin{Bmatrix} f_i f_j \text{ if } i = j \\ 2 f_i f_j \text{ if } i \neq j \end{Bmatrix}$$

where f_i is the frequency of A_i and f_j is the frequency of A_j in the population. These allele frequencies are usually unknown and must be estimated from baseline data. If n_i copies of allele A_i have been observed in a baseline sample of n genes, then n_i/n is a maximum likelihood estimate of f_i, the frequency of A_i in the population. The probability of observing genotype A_iA_j in the population can then be estimated:

$$\Pr(A_iA_j) = \begin{Bmatrix} \left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right) \text{if } i = j \\ 2\left(\frac{n_i}{n}\right)\left(\frac{n_j}{n}\right) \text{if } i \neq j \end{Bmatrix} \quad (1)$$

This approach works well when all of the alleles found in the fishery are found in the baseline population samples. However, if a fish sampled from the fishery has an allele not found in a baseline population sample (i.e., $n_i = 0$), this method of estimating probabilities (eq. 1) excludes that fish from that baseline population. When highly polymorphic loci such as microsatellites are examined, this phenomenon becomes common, and fish are frequently excluded from their population of origin.

One way to deal with this problem is to modify estimates of allele frequencies so that they will not be zero. This can be accomplished by assigning a frequency of $1/(n + 1)$ to an allele not found in a population. An alternative approach, with more justification, is to use a Bayesian method developed by Rannala and Mountain (1997). Rannala and Mountain (1997) have shown that the probability of obtaining the genotype A_iA_j from a population can be estimated by

$$\Pr(A_iA_j) = \begin{Bmatrix} \frac{(n_i + 1/k + 1)(n_j + 1/k)}{(n+1)(n+2)} \text{ if } i = j \\ \frac{2(n_i + 1/k)(n_j + 1/k)}{(n+1)(n+2)} \text{ if } i \neq j \end{Bmatrix} \quad (2)$$

where k equals the total number of alleles observed at the locus. Inspection of eq. 2 shows that it is similar to eq. 1, especially when sample size is large.

Once the probability of observing each fish has been calculated, maximum likelihood estimates of the mixture proportions can be found using established techniques such as the expectation-maximization algorithm (e.g., Millar 1987).

Sources of error

The goal of study design for mixed-stock fisheries is to identify sampling methods that minimize the expected difference between estimates of mixture proportions and the actual stock proportions. There are several potential sources of estimation error. A few include nonrandom sampling of the fishery, baseline data that do not include all of the populations in the fishery, and estimates of baseline allele frequencies that differ from the actual allele frequencies in the populations. Efficient study design requires knowing which sources of error are likely to be largest.

Estimating mixture proportions is a two-step process: first, fish are randomly sampled from a mixed-stock fishery; second, genetic data are used to estimate the mixture proportions within this sample. (By genetic data, I mean both the loci genotyped from the fish sampled from the fishery and the fish sampled from the baseline population.) Both steps contribute to estimation error. Consider the ideal case where all fish in the fishery are tagged so that their stock can be recognized unambiguously. Even in this enviable circumstance, sampling error from the fishery will cause estimates of stock proportions to differ from the actual mixture proportions. This error is minimized by sampling many fish from the fishery but can never be eliminated altogether. I call this fishery sampling error. When genetic data are used to estimate mixture proportions within a sample, additional error is expected. This is the second source of estimation error. I call this genetic estimation error. Understanding the difference between these two sources of error is essential. If the sample from the fishery does not represent the mixture proportions in the fishery, perhaps because the sample is small, genetic data cannot compensate, and estimates of stock proportions will be poor.

Quantifying the quality of estimates of mixture proportions

A useful statistic to describe the quality of an estimate of a stock proportion is the expected squared error (ESE) of that estimate. I represent the parametric proportion of the *i*th stock in the fishery as π_i and an estimate of that proportion as $\hat{\pi}_i$. The ESE of $\hat{\pi}_i$ is defined as

$$\text{ESE}_i = E(\hat{\pi}_i - \pi_i)^2 \quad (3)$$

where E denotes expectation.

Both bias and variance contribute to estimation error. ESE_i is equal to the bias (of estimates of π_i) squared plus the variance of $\hat{\pi}_i$ (Appendix A). There is no analytic method available to calculate ESE_i, but it can be estimated in simulations.

The ESE for the *i*th stock can be partitioned into two components: fishery sampling error, $\text{ESE}_{i,\text{fishery}}$, and genetic estimation error, $\text{ESE}_{i,\text{genetic}}$ (Appendix A):

$$\text{ESE}_i = \text{ESE}_{i,\text{fishery}} + \text{ESE}_{i,\text{genetic}} \quad (4)$$

$\text{ESE}_{i,\text{fishery}}$ is reduced by sampling more fish from the fishery. $\text{ESE}_{i,\text{genetic}}$ is reduced by collecting more genetic data (increasing the number of loci genotyped or increasing the number of individuals in the baseline data).

Equation 4 is useful because it can be used to compare the relative magnitude of both sources of estimation error. There is no formula available to calculate $\text{ESE}_{i,\text{genetic}}$, but it can be estimated in simulations. $\text{ESE}_{i,\text{fishery}}$ can be calculated when the mixture proportions in the fishery are known:

$$\text{ESE}_{i,\text{fishery}} = \frac{\pi_i(1-\pi_i)}{N_{\text{fishery}}} \quad (5)$$

To summarize the quality of estimates of all stocks being estimated, the expected squared errors for each stock can be summed:

$$\text{ESE}_{\text{fishery}} = \sum_{i=1}^{s} \text{ESE}_{i,\text{fishery}} \quad (6)$$

where s is the number of stocks in the mixture.

Interpreting the magnitude of ESE_i is difficult (this is also true for the variance of $\hat{\pi}_i$). I introduce a more meaningful statistic that I call "effective sample size" with a thought experiment.

Consider a mixed-stock fishery in which a stock of interest (stock j) composes 40% of the fish in the fishery, i.e., $\pi_j = 0.4$, and genetic baseline data are available so that mixture proportions from the fishery can be estimated. Let us assume that when 50 fish are sampled from the fishery and genotyped, the expected squared error for estimate π_i equals 0.01. To evaluate whether 0.01 is high or low, let us assume that all of these fish were tagged so that their origin could be identified unambiguously. Equation 5 shows that the ESE_j for tagged fish would be 0.0048. This is substantially less than the value of 0.01 obtained from the hypothetical genetic data. This indicates that the genetic data is only moderately effective in estimating mixture proportion. Equation 5 also shows that if 20 fish were sampled from the fishery and their origin identified by reading their tags, ESE_j would equal 0.012. This shows that (in this hypothetical example) reading tags from 20 fish is not as effective as sampling 50 fish and using genetic data to estimate mixture proportions. In this hypothetical example, genotyping 50 fish would produce the same ESE_j as genotyping reading tags from 24 fish. I interpret this as indicating that the effective sample size of 50 genotyped fish is 24.

More formally, the effective sample size N_{es} is related to the size of the sample taken from the fishery by

$$N_{\text{es}} = N_{\text{fishery}}\left(\frac{\text{ESE}_{\text{fishery}}}{\text{ESE}_{\text{total}}}\right) \quad (7)$$

Effective sample sizes will always be less than the number of fish sampled from the fishery (i.e., $N_{\text{es}} < N_{\text{fishery}}$). This is because genetic data cannot estimate the mixture proportions in a sample without error. Comparing N_{es} and N_{fishery} is useful. If N_{es} is only slightly less than N_{fishery}, genetic data are producing good estimates of mixture proportions. If N_{es} is only a small fraction of N_{fishery}, then the genetic data are not producing good estimates of mixture proportions.

Deciding whether an effective sample size is sufficiently large is facilitated by graphing the distribution of estimates of stock proportions for different sample sizes and stock proportions (e.g., Fig. 1).

Methods

Simulated genetic data were used to examine how effective sample size was affected by: the number of fish sampled from baseline populations (N_{baseline}), the number of loci examined (N_{loci}), the number of alleles per locus (N_{alleles}), and the amount of genetic divergence between the baseline populations (as measured by F_{ST}). The general procedure was as follows. Genetic data for baseline data sets and mixture samples was simulated from a specified evolutionary history, stock proportions were estimated from the simulated mixture sample, and the source of the observed error was partitioned into its two component parts. This was done as follows. Multinomial sampling was used to determine the number of fish in the sample observed from each stock. Fishery sampling error, $\varepsilon_{\text{fishery}}$, was calculated for each trial by subtracting the parametric proportion of each stock within the fishery from the realized proportion within the sample. Stock proportions within the fishery sample were then estimated using the simulated data. Genetic estimation error, $\varepsilon_{\text{genetics}}$, was calculated as the difference between these estimates and the actual stock proportions within the sample. The expected value of these errors and their squares and products were obtained by averaging their values across thousands of simulations. The effective sample size was then calculated for each combination of variables.

Fig. 1. Binomial distributions for events with varying probability (0.05, 0.1, 0.2, and 0.4) and varying number of trials: (*a*) N = 50, (*b*) N = 100, (*c*) N = 200, and (*d*) N = 400. These distributions approximate the distribution of the proportion of fish in a sample taken from a mixed-stock fishery that belong to stocks having frequencies of 0.05, 0.1, 0.2, and 0.4 in the fishery.

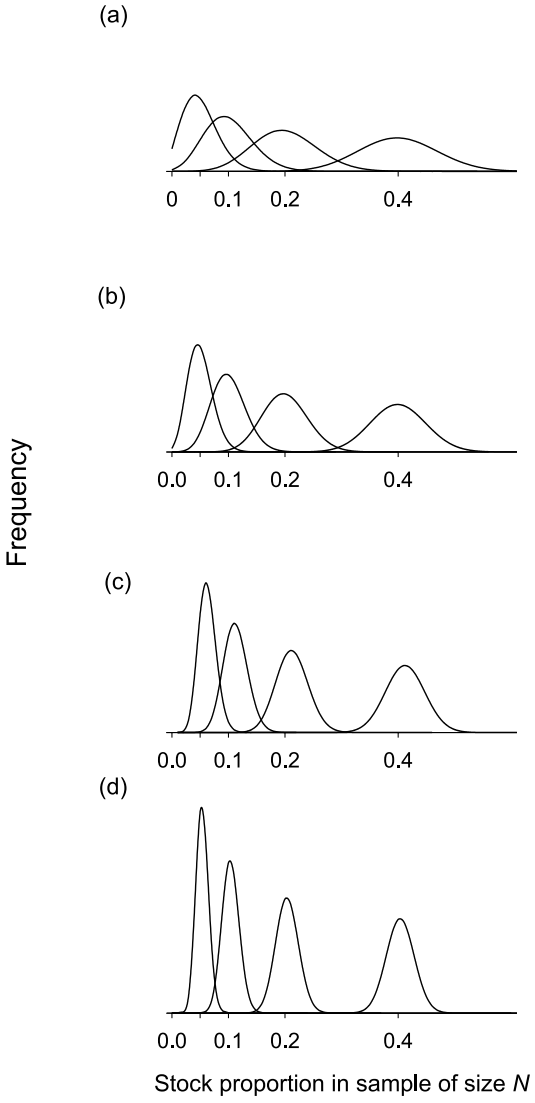
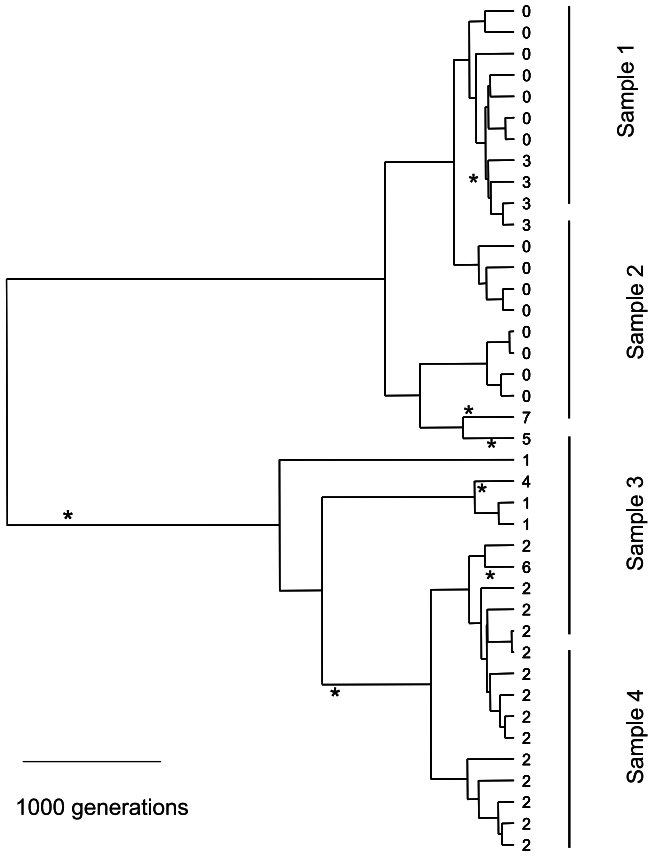


Fig. 2. Example of a gene tree with mutations produced by coalescent simulation for a simple model of population divergence. In this example, the ancestry of four samples of 10 genes was simulated from an ancestral population that split into two populations 200 generations before sampling (which each in turn split into two more populations 50 generations before sampling). Asterisks indicate the location of simulated mutations on the gene tree. The ancestral allele was type "0".



A bifurcating model of population fragmentation was used in all simulations. I assumed that the eight potential source populations that were sampled had an effective population size of 1000 individuals. These populations were descended from a single ancestral population of 8000 individuals that instantaneously split into two populations, each having an effective size of 4000 individuals. Subsequent fragmentation events split these two populations of 4000 into four populations of 2000, which later split into eight populations of 1000 individuals. Three degrees of population differentiation were examined. The timing of the population fragmentation events was 12, 25, and 50 generations in the past for the least differentiated populations (F_{ST} = 0.01), 50, 100, and 200 generations in the past (F_{ST} = 0.04) for the second set of populations, and 200, 400, and 800 generations for populations with the most differentiation (F_{ST} = 0.16) (F_{ST} for each evolutionary history was estimated by calculating F_{ST} across 10 000 biallelic loci).

Genotypic data were simulated using the standard coalescent approach (e.g., Hudson 1990; Hartl and Clark 1997; Avis 2000). This method has not been widely used in the fisheries literature, but it is common in the human genetics literature (e.g., Nordborg 2001; Rosenberg and Nordborg 2002; Felsenstein 2003). Coalescent simulation produces genotypic data with allele counts in the proportion specified by neutral theory and the evolutionary history of the populations. This technique proceeds by first simulating the ancestral gene tree of the sample backwards in time until all of the genes sampled share a common ancestor (Fig. 2). The shape of this gene tree is influenced by the effective size and fragmentation history of the populations. Once a gene tree was simulated for the sample, mutations were placed on the tree. Longer tree branches have a higher probability of being assigned a mutation than short branches. Loci having 2–512 alleles were simulated by placing mutations on the gene tree until the desired number of alleles in the sample was obtained. Samples for each locus were simulated independently.

Table 1. Gene diversities (H) at loci with different numbers of alleles ($N_{alleles}$) for eight simulated populations with divergence times of 50, 100, and 200 generations.

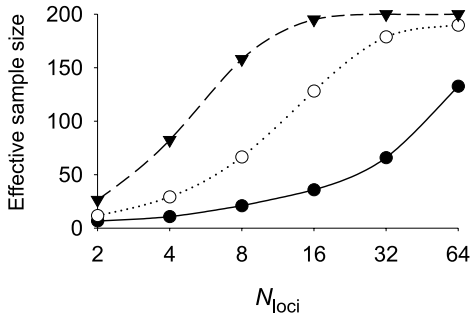
$N_{alleles}$	H
2	0.13
4	0.32
8	0.51
16	0.71
32	0.83
64	0.90
128	0.94
256	0.96
512	0.98

Note: The gene diversities of populations with small divergence times are slightly higher, and the gene diversities of populations with longer divergence times are slightly lower.

Simulating sample sizes from loci with varying numbers of alleles must account for the expectation that, everything else being equal, large samples will have more alleles than small samples. I dealt with this by standardizing the sample size used to count the number of alleles at each simulated locus. I chose 512 individuals, distributed equally among eight populations, as the standard. For example, $N_{alleles}$ = 16 indicates that 16 alleles were observed in 512 randomly selected individuals distributed among the eight baseline populations. Therefore, $N_{alleles}$ is a measure of how much variation there is at a locus (as opposed to a sample). If the baseline sample included more than 512 individuals, then additional alleles might be present in the sample. If the total baseline sample included less than 512 individuals, then not all of the 16 alleles might be found in the baseline used to estimate mixture proportions. This assures that loci with the same number of alleles have the same expected heterozygosity. Heterozygosity ranged from 0.13 for loci with two alleles to 0.98 for loci with 512 alleles (Table 1).

The stock proportions within the fishery were set at {0.4, 0.2, 0.1, 0.1, 0.1, 0.05, 0.05, 0.0} for all simulations. In all simulations, 200 simulated fish were sampled from the fishery.

Fig. 3. Effective sample size of estimates of the mixture proportions of eight stocks in a mixed-stock fishery as a function of the number of loci examined. Three degrees of evolutionary divergence were examined: F_{ST} = 0.01 (solid circles), 0.04 (open circles), and 0.16 (triangles). All samples from the fishery had 200 fish and all loci examined had eight alleles.



The expectation-maximization algorithm as described by Millar (1987) was used to obtain maximum likelihood estimates of stock proportions within the sample of fish taken from the simulated fishery, with the exception that eq. 2 was used to estimate the probability of each genotype from each population. This algorithm uses iteration to find maximum likelihood estimates of stock proportions. At least 50 iterations were performed for each estimate. After that, iteration was stopped when the sum of the absolute value of changes in frequency of stock proportions was less than 10^{-6}. A maximum of 1000 iterations were performed. Slow convergence was only a problem with very small data sets (e.g., two loci with two alleles each).

Results

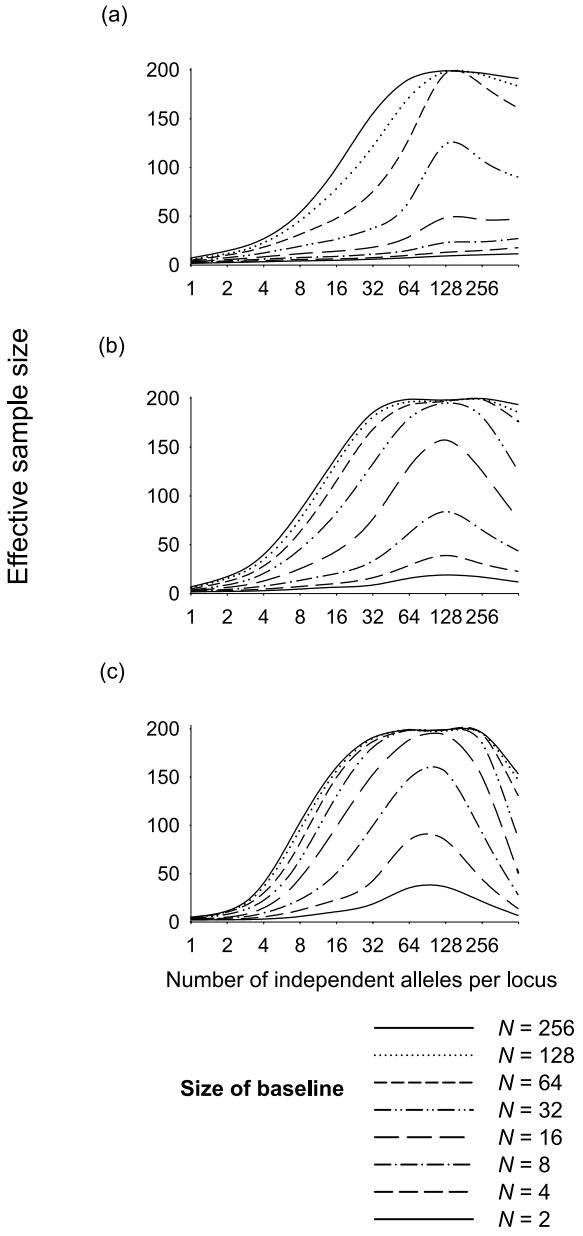
Number of loci and degree of evolutionary divergence

Increasing the number of loci produced better estimates of stock proportion until the effective sample size reached the actual sample size from the fishery (Fig. 3). The degree of evolutionary divergence among the populations also had a strong influence on the effect of increasing the number of loci examined. When populations were relatively genetically similar, estimating mixture proportions was difficult and many loci needed to be examined to accurately estimate the mixture proportion within the sample from the fishery. When populations were strongly genetically differentiated, estimating mixture proportions was much easier and fewer loci were needed (Fig. 3).

Number of alleles per locus

Increasing the number of alleles at loci produced increasingly better estimates of mixture proportions until a turning point was reached (Fig. 4). This point of "too much polymorphism" was approximately 128 alleles for 512 individuals. This number of alleles corresponds to a heterozygosity of 0.94. This turning point appears to be weakly affected by F_{ST}: high levels of polymorphism are slightly less informative when F_{ST} is high than when F_{ST} is low (Fig. 4). Interestingly, the ideal level of polymorphism appeared to be independent of baseline sample sizes (Fig. 4). For example, loci with 128 alleles produced better estimates of stock proportions than loci with fewer alleles, even when baseline sample sizes

Fig. 4. Effective sample of estimates of the mixture proportions of eight stocks in a mixed-stock fishery as a function of the number of alleles per locus and the size of baseline samples. Three degrees of evolutionary divergence were examined: (*a*) F_{ST} = 0.01, 16 loci, (*b*) F_{ST} = 0.04, eight loci, and (*c*) F_{ST} = 0.16, four loci. All samples from the fishery had 200 fish and were genotyped at eight loci.



were small (two, four, eight, etc., individuals per baseline population). Highly polymorphic loci, therefore, do not require larger sample sizes than less polymorphic loci. Nor do they appear to benefit more from large sample sizes than less polymorphic loci.

The accuracy of estimates of mixture proportions can be improved by increasing either the number of loci (Fig. 3) or the number of alleles at loci (Fig. 4). This motivates the following question. Which data produce better results: a few loci with many alleles or many loci with few alleles? The

Fig. 5. Effective sample size of estimates of mixture proportions of eight stocks in a mixed-stock fishery as a function of the number of loci (N_{loci}) examined and the number of independent alleles per locus ($N_{alleles}$). Circles indicate samples with a total of 32 independent alleles (e.g., 16 loci with three alleles each). All samples from the fishery had 200 fish. The baseline data set consisted of 64 fish per stock. The F_{ST} for the eight stocks equaled 0.04.

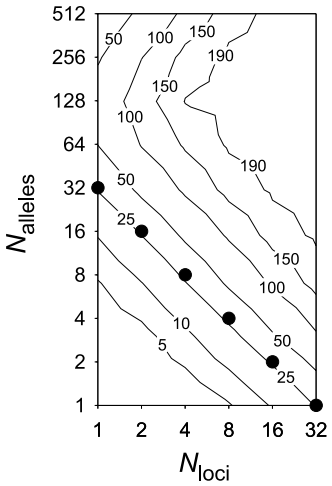
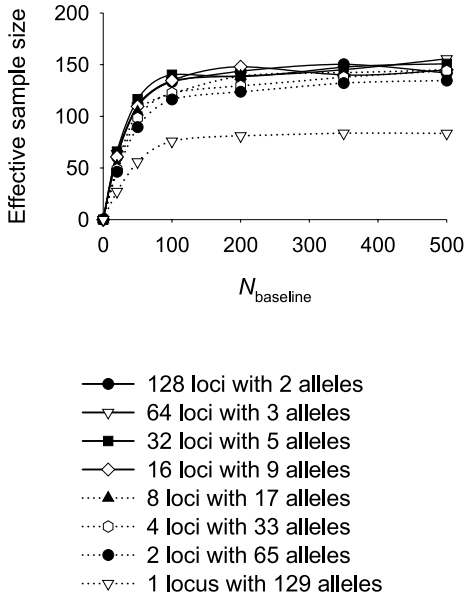


Fig. 6. Effective sample size of estimates of mixture proportions of eight stocks in a mixed-stock fishery as a function of baseline samples sizes. The F_{ST} for the eight stocks equaled 0.04.



answer is that each approach works equally well, as long as the number of alleles per locus is not excessive (Fig. 5). The relevant number of alleles here is the number of independent alleles (a locus with k alleles has $k - 1$ independent alleles). One locus with 32 independent alleles is equivalent to 32 loci with one independent allele each (Fig. 5). This trend breaks down at loci with too many alleles (Fig. 5). For example, one locus with 128 independent alleles is not as effective as 128 loci with one independent allele.

Baseline size

Large baseline samples produce better estimates of stock proportions than small baseline samples (Figs. 4 and 6). However, increasing baseline sample size produces diminishing results. In all cases examined, increasing baseline sample sizes past 100 individuals resulted in limited improvement (Fig. 6). This occurred for loci with 128 alleles as well as for loci with two alleles.

Discussion

Two significant results were obtained: (*i*) when polymorphism is not extreme (greater than 0.90 heterozygosity or 64 alleles), the total number of independent alleles across loci is a good indicator of how accurate mixture estimates are likely to be and (*ii*) loci with many alleles do not require larger baseline sample sizes than loci with few alleles (and 100 individuals per population should be sufficient).

The results presented here describe how polymorphism is expected to affect estimates of stock proportions. For example, I have shown that, on average, loci with four alleles produce better mixture estimates than loci with two alleles. This does not mean that every locus with four alleles will produce better estimates of stock proportions than every locus with two alleles. The actual allele frequencies will also affect the quality of mixture estimates. A locus with two common alleles could easily work better than a locus with one common allele and three rare alleles.

The algorithm used here to estimate mixture proportions appears to be novel. It was chosen for three reasons: (*i*) it is similar to maximum likelihood approaches with a history of use, (*ii*) it should minimize problems associated with highly polymorphic loci, and (*iii*) it is quick enough to be used on thousands of simulated data sets. I expect that the results discussed here also apply to a complete Bayesian approach (BAYES) recently developed by Pella and Masuda (Masuda 2000; Pella and Masuda 2001). Applying the small sample size correction in eq. 2 appears to be essential for using highly polymorphic loci when baseline sample sizes are modest. If eq. 2 is not used, loci with many alleles do not produce good estimates of mixture proportions unless baseline sample sizes are very large (results not shown). In addition, increasing the number of the loci examined can decrease the accuracy of mixture estimates. This occurs because the probability of individual in the mixture having an allele not present in the baseline data from its population increases when more loci are examined.

Recent empirical work has been concordant with the results presented here. Winans et al. (2004) showed that (*i*) eq. 2 produced better mixture estimates than eq. 1, (*ii*) more loci produced better estimates than fewer loci, (*iii*) five microsatellite loci with 63 independent alleles produced almost as good estimates as 32 allozyme loci with 77 independent alleles, and (*iv*) sampling more than 100 individuals resulted in little benefit.

Several questions deserve further research. First, I do not have an adequate explanation for why highly polymorphic loci do not benefit more from larger samples than less polymorphic loci. This result does not contradict any theory that I am aware of, but it seems paradoxical. Second, more work will be necessary to identify the evolutionary and statistical

variables that determine how much variation is ideal. The results that I present here suggest that a heterozygosity of 0.9 is not too much, but my simulations were not exhaustive. Third, my simulations assumed that all baseline populations that contributed to the mixture were present in the mixture. This will not be true in many applications. Fourth, further work is necessary to understand the effect of temporal variations in allele frequency on mixture estimates (e.g., Jorde and Ryman 1995). Multiple baseline samples or sample sizes greater than 100 individuals may be necessary.

These results suggest that geneticists estimating stock proportions in mixed-stock fisheries have a great deal of flexibility in study design. For example, either a few highly polymorphic loci might be used or many less polymorphic loci, such as SNPs, might be used. Because each of these approaches can work, technical considerations such as ease or expense of scoring can be used to select the strategy most appropriate for a specific laboratory.

References

Allendorf, F.W., and Seeb, L.W. 2000. Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. Evolution, **54**: 640–651.

Avis J.C. 2000. Phylogeography. Harvard University Press, Cambridge, Mass.

Felsenstein, J. 2003. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Gilbert, C.H. 1924. Experiment in tagging adult red salmon, Alaska Peninsula fisheries reservation, summer of 1922. Fish. Bull. U.S. **39**: 39–50.

Grant, W.S., Milner, G.B., Krasnowski, P., and Utter, F.M. 1980. Use of biochemical genetic variants for identification of sockeye salmon (*Oncorhynchus nerka*) stocks in Cook Inlet, Alaska. Can. J. Fish. Aquat. Sci. **37**: 1236–1247.

Hamilton, J.A.R. 1947. Significance of certain scale characters in the recognition of Fraser River sockeye races. M.Sc. thesis, University of British Columbia, Vancouver, B.C.

Hartl, D.L., and Clark, A.G. 1997. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7**: 1–44.

Jorde, P.E., and Ryman, N. 1995. Temporal allele frequencies change and estimation of effective population size with overlapping generations. Genetics, **139**: 1077–1090.

Masuda, M. 2000. User's manual for BAYES: stock-mixture analysis program based on Bayesian methods. DOC/NOAA/NMFS/AFSC, Auke Bay Laboratory, U.S.–Canada Salmon Program, 11305 Glacier Hwy., Juneau, AK 99801-8626, USA.

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

Moran, P. 2002. Current conservation genetics: building an ecological approach to the synthesis of molecular and quantitative genetic methods. Ecol. Freshw. Fish, **11**: 30–55.

Mosher, K. 1963. Racial analysis of red salmon by means of scales. Int. North Pac. Fish. Comm. Bull. **11**: 31–56.

Nordborg, M. 2001. Coalescent theory. Chap. 7. *In* Handbook of statistical genetics. *Edited by* D. Balding, M. Bishop, and C. Cannings. Wiley, Chichester, UK.

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

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

Rosenberg, N.A., and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3**: 380–390.

Scribner, K.T., Crane, P.A., Spearman, W.J., and Seeb, L.W. 1998. DNA and allozyme markers provide concordant estimates of population differentiation: analyses of US and Canadian populations of Yukon River fall-run chum salmon (*Oncorhynchus keta*). Can. J. Fish. Aquat. Sci. **55**: 1748–1758.

Shaklee, J.B., Beacham, T.D., Seeb, L., and White, B.A. 1999. Managing fisheries using genetic data: case studies from four species of Pacific salmon. Fish. Res. **43**: 45–78.

Utter, F.M., Hodgins, H.O., and Allendorf, F.W. 1974. Biochemical genetic studies of fishes: potentialities and limitations. *In* Biochemical and biophysical perspectives in marine biology. Vol. 1. *Edited by* D. Malins. Academic Press, San Francisco, Calif. pp. 213–237.

Winans, G.A., Paquin, M.Z., Van Doornik, D.M., Rawding, D., Baker, B., Marshall, A., Moran, P., and Kalinowski, S.T. 2004. Genetic stock identification of steelhead in the Columbia River basin: an evaluation of different molecular markers. N. Am. J. Fish. Manag. **24**: 672–685.

Wirgin, I.I., Waldman, J.R., Maceda, L., Sabile, J., and Vecchio, V.J. 1997. Mixed-stock analysis of Atlantic coast striped bass (*Morone saxatilis*) using nuclear DNA and mitochondrial markers. Can. J. Fish. Aquat. Sci. **54**: 2814–2826.

Appendix A

The estimate of the *i*th stock proportion ($\hat{\pi}_i$) can be modeled:

$$\text{(A1)} \quad \hat{\pi} = \pi + \varepsilon_{\text{estimation}} + \beta$$

where $\varepsilon_{\text{estimation}}$ is a random variable with mean 0 and β is a constant. For simplicity, I have dropped the subscript *i*. By definition, the variance of $\hat{\pi}$ is given by

$$\text{(A2)} \quad \text{Var}(\hat{\pi}) = E(\hat{\pi} - E(\hat{\pi}))^2$$

By inspection of eq. A1, we observe that $E(\hat{\pi}) = \pi + \beta$. When we substitute this and $\hat{\pi} = \pi + \varepsilon_{\text{estimation}} + \beta$ (eq. A1) into eq. A2, we obtain

$$\text{(A3)} \quad \text{Var}(\hat{\pi}) = E(\pi + \varepsilon_{\text{estimation}} + \beta - \pi - \beta)^2$$

which simplifies to

$$\text{(A4)} \quad \text{Var}(\hat{\pi}) = E(\varepsilon^2_{\text{estimation}})$$

This makes sense because $\varepsilon_{\text{estimation}}$ is the only source of variation in $\hat{\pi}$.

Equation A1 shows that the expected squared error for $\hat{\pi}$ is

$$\text{(A5)} \quad E(\pi - \hat{\pi}) = E(\pi - \pi + \varepsilon_{\text{estimation}} + \beta)^2$$

which simplifies to

$$\text{(A6)} \quad E(\pi - \hat{\pi}) = E(\varepsilon^2_{\text{estimation}}) + E(2\beta\varepsilon_{\text{estimation}}) + E(\beta^2)$$

The middle term on the right is equal to $2\beta E(\varepsilon_{\text{estimation}})$, which is equal to zero (because the expected value of $\varepsilon_{\text{estimation}}$ is equal to zero). This leaves us with

$$\text{(A7)} \quad E(\pi - \hat{\pi}) = E(\varepsilon^2_{\text{estimation}}) + E(\beta^2)$$

or

(A8) $E(\pi - \hat{\pi}) = \mathrm{Var}(\hat{\pi}) + E(\beta^2)$

Alternatively, the estimate of the *i*th stock proportion can be modeled:

(A9) $\hat{\pi} - \pi + \varepsilon_{\mathrm{fishery}} + \varepsilon_{\mathrm{genetics}}$

where $\varepsilon_{\mathrm{fishery}}$ is a random variable with an expected value of zero that models how sampling a finite number of fish from the fishery affects mixture estimates and $\varepsilon_{\mathrm{genetics}}$ is a random variable that models how genetic data affects estimates. We have

(A10) $E(\pi - \hat{\pi}) = E(\varepsilon_{\mathrm{fishery}})^2 + E(2\varepsilon_{\mathrm{fishery}}\varepsilon_{\mathrm{genetics}}) + E(\varepsilon_{\mathrm{genetics}})^2$

I define the first term on the right hand side of eq. A10 as $\mathrm{ESE}_{\mathrm{fishery}}$ and the remaining two terms (on the right) as $\mathrm{ESE}_{\mathrm{genetics}}$.